MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation

Yuexu Jiang
Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO

Duolin Wang
Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO

Yifu Yao
Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO

Holger Eubel
Leibniz University Hannover

Patrick Künzler
Institute of Plant Genetics, Leibniz University Hannover, Hannover, Germany

Ian Møller
Aarhus University

Dong Xu (xudong@missouri.edu)
Department of Electrical Engineering and Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO https://orcid.org/0000-0002-4809-0514

Article

Keywords: protein localization, mechanisms of localization, MULocDeep, subcellular and suborganellar localization

Posted Date: July 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-40744/v1

Version of Record: A version of this preprint was published at Computational and Structural Biotechnology Journal on August 1st, 2021. See the published version at https://doi.org/10.1016/j.csbj.2021.08.027.

MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation

Yuexu Jiang[1], Duolin Wang[1], Yifu Yao[1], Holger Eubel[2], Patrick Künzler[2], Ian Max Møller[3] and Dong Xu[1]*

[1] Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, Columbia, Missouri, USA
[2] Institute of Plant Genetics, Leibniz University Hannover, Hannover, Germany
[3] Department of Molecular Biology and Genetics, Aarhus University, Forsøgsvej 1, DK-4200 Slagelse, Denmark

* email: xudong@missouri.edu

Abstract

Prediction of protein localization plays an important role in understanding protein function and mechanisms. In this paper, we propose a general deep learning-based localization prediction framework, MULocDeep, which can predict multiple localizations of a protein at both subcellular and suborganellar levels. We collected a dataset with 45 suborganellar localization annotations in 10 major subcellular compartments—the most comprehensive suborganelle localization dataset to date. We also experimentally generated an independent dataset of mitochondrial proteins in *Arabidopsis thaliana* cell cultures, *Solanum tuberosum* tubers, and *Vicia faba* roots and made this dataset publicly available. Evaluations using the above datasets show that overall, MULocDeep outperforms other major methods at both subcellular and suborganellar levels. Furthermore, MULocDeep assesses each amino acid's contribution to localization, which provides insights into the mechanism of protein sorting and localization motifs. A web server can be accessed at http://mu-loc.org.

Introduction

In eukaryotic cells, proteins perform diverse functions governed by the compartments or organelles in which those proteins are located. The aberrant localization of proteins is often associated with diseases, such as Alzheimer's disease, metabolic disorders, and cancers[1,2]. The mechanisms of protein localization are complex. Many protein localizations rely on targeting peptides within protein sequences[3], while others depend on overall protein properties, such as surface charge[4]. Identifying protein localization and studying the mechanisms of localization may help us understand protein sorting and design therapeutic strategies related to protein targeting[2]. Extensive efforts have been made to identify protein localization using technologies such as mass spectrometry and fluorescence-tagging methods[5,6]. However, such methods are time- and labor-consuming and produce false results to a certain degree. Hence, the localization of many proteins is still unknown at the organellar level and only a limited number of proteins—even in humans—have a well-defined localization at the suborganellar level. Thus, computational methods can play an important role in this area.

Computational methods for protein localization prediction can be roughly divided into two approaches. The first approach takes advantage of homology information and Gene Ontology (GO) annotations[7]. Representative methods include LocTree3[8], SherLoc2[9], MultiLoc2[10] , and YLoc[11]. These methods tend to perform well when reliable annotations of homologous proteins are available. However, they may not be applicable when a protein lacks homology to any protein of known localization. The second approach is *ab initio* prediction using patterns learned from training samples. Some traditional machine-learning methods rely on manually extracted features. For example, WoLF PSORT[12] converts a protein's amino acid sequence into features like sorting signals, amino acid composition, and motifs for training a k-nearest neighbor (KNN) classifier. TPpred3[13] detects the targeting signal in the N-terminal region of a protein based on a support vector machine (SVM) classifier. Predotar[14] applies a neural network to identify proteins targeting the endoplasmic reticulum, mitochondria, and plastids in plants by N-terminal targeting signals. TargetP[15,16] also uses neural networks to discriminate proteins destined for mitochondrion, chloroplast, and the secretory pathway based on the N-terminal sequence information. More recently, deep learning methods have been explored in protein localization prediction. For example, DeepLoc[17] uses a convolutional neural network (CNN) and long short-term memory (LSTM) to give a prediction of 10 subcellular protein localizations. Deep neural networks were used in our previous MU-LOC method[18], in which features including amino acid frequency, sequence profile, and gene co-expression were used to predict whether or not a plant protein was mitochondrial. The latest version of TargetP (v2.0)[19] applies bidirectional LSTM to predict thylakoid transit peptides, etc.

Although several methods have achieved good prediction results on specific protein localization cases, these methods still face limitations and many unsolved problems. Most of the methods focus on the prediction of protein localization at the subcellular level. Although there are some predictors for specific suborganelle localizations[19-22], a systematic suborganelle localization prediction tool at the whole-cell scale is still missing. Furthermore, protein localization is a multi-label problem; i.e., one protein may be found in several different compartments in a cell. Some efforts have been made in multi-label prediction[21,23], but current deep learning-based methods simplify the protein localization prediction as a one-label classification problem in which each protein can only be predicted at a unique compartment—which is not the case for 15-20% of proteins (Figure S1). In addition, there is room to increase the deep learning model's interpretability for characterizing localization signals. For example, both TargetP 2.0[19] and DeepLoc[17] attempted to identify strong contributing sequence factors to localization using the attention mechanism[24,25]. However, TargetP 2.0 considers only the first 200 amino acids near the N-terminus of a protein, which cannot detect localization signals in other parts of a protein. DeepLoc addresses this problem by taking as many as 500 amino acids from each terminus of a protein, but the interpretation resolution cannot reach to the single amino acid level.

In this paper, we propose a multi-label protein localization classifier named MULocDeep that covers 10 main subcellular localizations and 45 suborganellar localizations. A matrix data structure captures the intrinsic hierarchical relationships between organelles and their subcompartments, enabling our method

to make predictions at both levels simultaneously. The core of the method consists of Long Short Term Memory (LSTM)[26] and multi-head self-attention[24], which have the ability to extract biological features that contribute to localizations at the single amino acid resolution. Some of these features match the current knowledge of protein sorting signals, while there are novel discoveries that could provide some new insights. This paper also includes an experimental study, in which the mitochondrial proteomes of 3 species, including Arabidopsis cell cultures, Solanum tubers, and Vicia roots, were extracted and identified (*Mito3* dataset). We also systematically collected a dataset from the UniProt database[27], containing proteins of animals and plants in 45 suborganellar compartments in 10 subcellular localizations with experimental evidences (*UniLoc* dataset). Evaluations using the above datasets show that overall, MULocDeep outperforms other major methods at both subcellular and suborganellar levels. The datasets themselves can be used as benchmarks for methods developed by others. The source code and the web server of our method are publicly available.

Results

In this part, we first present the proposed MULocDeep model and compare it with other methods. Then we demonstrate the effectiveness of MULocDeep in interpreting the contribution of each amino acid to localization prediction. Some of these important amino acids can match to well-known protein sorting peptides or signals. Finally, we briefly introduce the key features of the MULocDeep web server.

MULocDeep framework and the workflow

The workflow of our framework is presented in Figure 1. Protein sequences with known localization information were collected, processed, and fed into our deep learning model for training. During the training process, the output of the “attention” layer was extracted separately for sorting signal interpretation and visualization. Finally, the trained model was used to predict localization for new proteins. A description of the MULocDeep model is shown in the right panel in Figure 1. The input layer was composed of encoded protein sequences with a fixed length of 1000 amino acids. Each amino acid was encoded as a 25-dimension vector (see the Methods section for encoding details). The input layer was followed by two layers of bidirectional LSTM[26], which ensured that every amino acid received a signal from both sides. Two such layers were stacked to give the model the ability to fit complex high-order functions. The sequence length remained unchanged after the bidirectional LSTMs, while only the encoding dimension was changed to 180. Then a multi-head self-attention layer[24] was applied (“A” in Figure 1). The embedding matrix (“M” in Figure 1) was derived as the weighted sum by multiplying the attention layer with the output from the bidirectional LSTM. The attention itself was also an output in order to assess the contribution of each amino acid to localization. The embedding matrix was flattened into a 7380 (180X41) long vector, then fully connected with an 80-dimensional dense layer, which was further reshaped into an 8-by-10 matrix. Each column of the matrix represented a major subcellular localization (10 organelles) and each element under the column represented a suborganelle category. According to our suborganellar dataset, one organelle contained up to eight suborganellar localizations, e.g., the

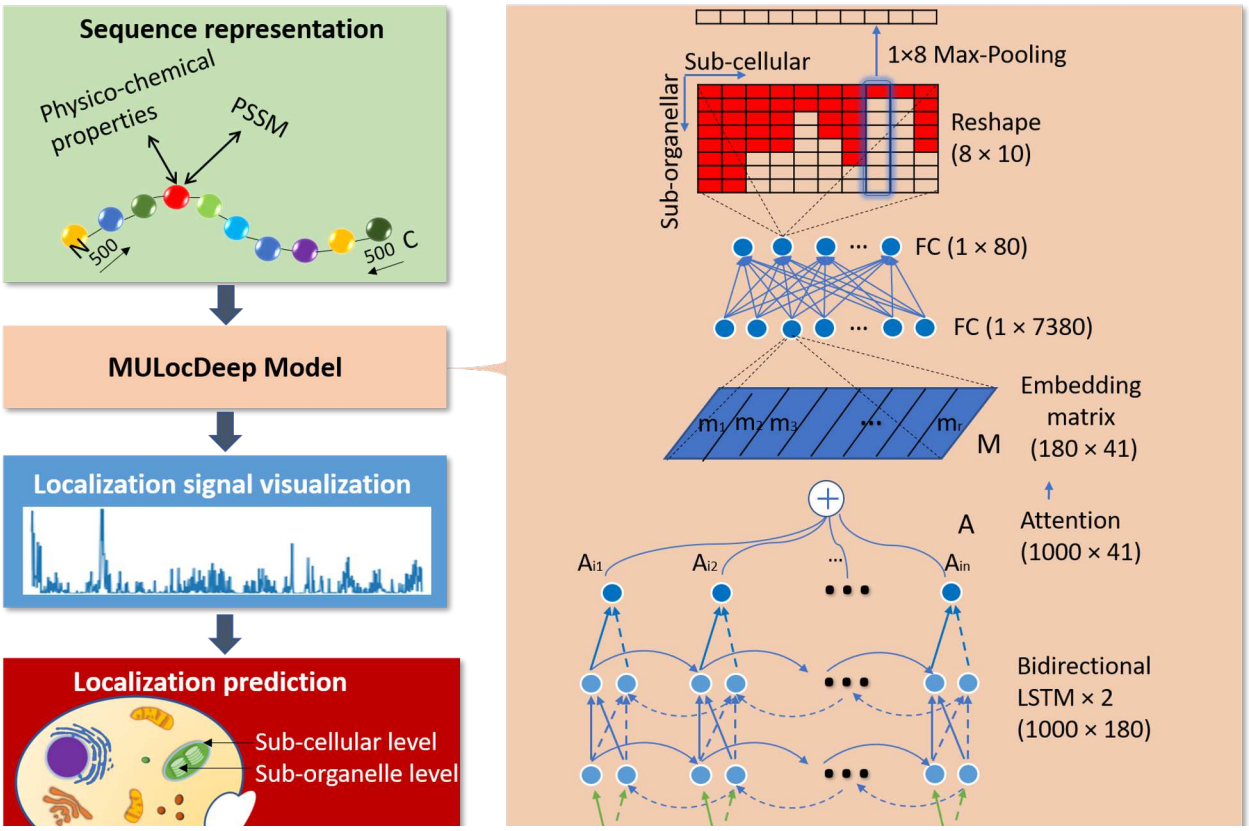


Figure 1. MULocDeep workflow and neural network architecture. The workflow is composed of four steps: (1) Protein sequence representation, (2) training the MuLocDeep model, (3) localization signal visualization, and finally (4) localization prediction. The details of the neural network architecture are displayed in the right panel.

cytoplasm and nucleus. For other organelles that had fewer suborganelle localizations, the empty slots in the matrix were padded with zeros. When processing a new sample, the predicted value in the matrix was used for the suborganelle prediction. Then a 1×8 max-pooling layer was applied to the matrix, so that the highest predicted value of a suborganelle localization was used as the prediction score of the corresponding organelle localization. Only the suborganelle and organelle localizations with prediction scores above the threshold (0.5) were used as output results. This way we could perform multi-label predictions at both subcellular and suborganelle levels and still keep the results consistent. When training the MULocDeep model, we tried different strategies to tune the hyperparameters and tested their impact on its performance. The details of the training process and hyperparameter configuration are introduced in the Methods section.

Comparison of performance in localization prediction

Table 1. Evaluation and comparison of protein localization prediction tools.

Method	Scope	AVAIL	Subcellular	Suborganellar	Assessments					
					ROC_auc	P&R_auc	MCC	Recall	Prec	Acc
Mitochondrion localization prediction using the *Mito3* dataset:										
MULocDeep	1-10	W&L	Mitochondrion	/	0.72	**0.76**	0.36	0.31	0.90	0.64
MULoc[18]	4	W&L		/	**0.74**	0.73	0.33	**0.41**	0.77	0.64
DeepLoc[17]	1-10	W&L		/	0.70	0.58	**0.42**	0.33	0.78	**0.80**
TargetP v5[19]	3,4,7	W&L		/	0.69	0.72	0.29	0.17	**0.96**	0.58
MitoFates[28]	4	W&L		/	0.64	0.67	0.26	0.18	0.88	0.57
SherLoc2[9]	1-11	W		/	0.64	0.66	0.24	0.18	0.85	0.57
MultiLoc2[10]	1-11	W&L		/	0.66	0.68	0.25	0.16	0.88	0.57
YLoc[11]	1-11	W		/	0.38	0.28	0.18	0.20	0.59	0.66
Predotar[14]	4,6,7	W		/	0.63	0.65	0.27	0.22	0.84	0.58
Localizer[29]	1,4,7	W		/	/	/	0.22	0.21	0.77	0.57
Suborganellar localization prediction using the *UniLoc-test* dataset:										
MULocDeep	1-10	W&L	Mitochondrion	Inner membrane	0.67	0.45	0.15	**0.83**	0.30	0.46
				Outer membrane	**0.91**	**0.73**	**0.47**	0.25	**1.00**	**0.90**
				Matrix	**0.86**	**0.71**	**0.72**	**0.79**	**0.82**	**0.88**
			Golgi apparatus	Trans-Golgi	**0.46**	**0.42**	**0.24**	0.08	**1.00**	**0.76**
DeepMito[20]	4	W&L	Mitochondrion	Inner membrane	**0.79**	**0.70**	**0.29**	0.78	**0.38**	**0.61**
				Outer membrane	0.53	0.30	0.01	0.25	0.13	0.67
				Matrix	0.57	0.42	0.13	0.28	0.44	0.65
Sub-Golgi v2[30]	8	W	Golgi apparatus	Trans-Golgi	0.44	0.24	0.07	0.85	0.24	0.29
TetraMito[31]	4	W	Mitochondrion	Inner membrane	0.48	0.52	0.07	0.47	0.39	0.56
				Outer membrane	0.81	0.61	0.30	**0.50**	0.40	0.77
				Matrix	0.51	0.51	0.01	0.35	0.47	0.51
Subcellular localization prediction using the *DeepLoc* dataset:										
MULocDeep	1-10	W&L	Nucleus	/	**0.95**	**0.90**	**0.71**	**0.85**	0.77	0.87
			Cytoplasm	/	**0.88**	**0.64**	**0.50**	**0.58**	**0.60**	**0.85**
			Extracellular	/	0.99	**0.98**	**0.94**	**0.96**	0.94	**0.98**
			Mitochondrion	/	0.97	**0.92**	0.80	0.80	**0.85**	0.96
			Cell membrane	/	0.96	0.29	0.28	0.37	0.24	0.96
			ER	/	0.93	**0.69**	**0.60**	0.65	**0.59**	**0.96**
			Plastid	/	0.99	0.90	**0.79**	0.81	**0.78**	**0.99**
			Golgi apparatus	/	**0.94**	**0.39**	**0.40**	0.33	**0.50**	0.98
			Lysosome	/	**0.91**	**0.45**	**0.41**	**0.21**	**0.82**	**0.97**
			Peroxisome	/	**0.98**	**0.63**	0.41	0.21	**0.80**	**0.99**
DeepLoc	1-10	W&L	Nucleus	/	0.94	0.89	0.69	0.73	**0.83**	0.87
			Cytoplasm	/	0.80	0.51	0.41	0.56	0.50	0.81
			Extracellular	/	0.99	0.97	0.91	0.91	**0.95**	0.97
			Mitochondrion	/	0.97	0.91	**0.83**	**0.88**	0.83	0.96
			Cell membrane	/	0.96	**0.30**	**0.43**	**0.66**	**0.30**	0.96
			ER	/	0.93	0.62	0.56	**0.66**	0.52	0.95
			Plastid	/	0.99	0.90	0.75	**0.87**	0.65	0.98
			Golgi apparatus	/	0.90	0.28	0.35	0.33	0.38	0.98
			Lysosome	/	0.86	0.18	0.23	0.18	0.33	0.96
			Peroxisome	/	0.94	0.39	**0.48**	**0.31**	0.75	0.98

The upper part of the table uses our *Mito3* dataset to evaluate the performance of the mitochondrial protein prediction; the middle part uses the *UniLoc-test* dataset to evaluate the performance of the suborganelle level prediction; and the lower part uses the test samples in the *DeepLoc* dataset after removing the redundant sequences in our training data to evaluate the performance for subcellular localization prediction. Availability (AVAIL) is either through a web server (W) or a local tool (L). NA means the method is unavailable at the time of our evaluation. The prediction scope includes compartments in: 1. nucleus; 2. Cytoplasm; 3. extracellular; 4. mitochondrion; 5.cell membrane; 6. endoplasmic reticulum; 7. plastid/chloroplast; 8. Golgi apparatus; 9. lysosome/vacuole; 10. peroxisome; 11. plasma membrane. Criteria of assessment include ROC_auc (area under receiver operating characteristic curve), P&R_auc (area under precision & recall curve), MCC (Matthew's correlation coefficient), recall, precision, and accuracy.

We compared the protein localization prediction of MULocDeep to other available tools on three independent benchmark datasets, including the *Mito3* dataset, the *UniLoc-test* dataset, and part of the test samples in the *DeepLoc* dataset. We summarized a list of localization classifiers regarding their scope (target localizations), availabilities (web server or local tool) and the performance on the first two benchmark datasets. We have actually tested many other methods, but they are excluded from the comparison because they were either unavailable, not working properly at the time of test, or only accepts a single sequence for web submission. The *Mito3* dataset was used to evaluate the performance of different classifiers for the prediction of mitochondrial proteins (Table 1, upper part). The *UniLoc-test* dataset was used to test the suborganellar prediction for Golgi and mitochondrion (Table 1, middle part). The *DeepLoc* dataset data was only used to compare with DeepLoc for multiple subcellular level localization prediction (Table 1 lower part). Among the six measurements of the performance, the ROC_auc (area under receiver operating characteristic curve) and P&R_auc (area under precision & recall curve) are the most important criteria as they reflect accuracies at a continuous range of thresholds for a binary prediction. According to Table 1 (upper), MULocDeep is a competitive method for the mitochondrial protein prediction. More than half of the measurements are better than any other method in a pair wisely comparison. Especially, MULocDeep has higher ROC_auc and P&R_auc than others except for the MULoc method. For the suborganellar level prediction (Table 1 middle), MULocDeep is much better than other tools except DeepMito in predicting the inner membrane under mitochondrion. Note that both MULoc and DeepMito were designed specifically for mitochondria. In the lower part of Table 1, 1347 proteins were picked from the original testing samples in the *DeepLoc* dataset after removing the proteins that have homologs (>40% sequence identity) in the suborganelle dataset. According to the comparison results with DeepLoc, except cell membrane, MULocDeep has achieved a higher score in most of the measurements.

The evaluation so far is from the perspective of a tool. Even though the testing datasets are the same and independent from any method, the training datasets are still different. Besides, the number of target classes varies among different methods. Some methods are binary classifiers, like MULoc, while some other methods have more target classes, e.g. 45 classes in the case of MULocDeep. To evaluate different approaches from the method perspective under a fair condition, we created a variant of the MULocDeep model. We used the variant model to compare with different methods individually at both subcellular and suborganellar levels. The details of the variant model are provided in Supplementary Note 1. At the suborganellar level, only a few of methods have provided clearly separated datasets for training and testing, which makes it difficult to have a fair comparison. Here we compared with DeepMito [20], a recently published deep learning method for sub-mitochondrial protein localization prediction. A variant model was trained using the same *DeepMito* datasets provided by the DeepMito paper. The output layer was a 4-dimensional vector representing four target compartments (outer membrane, inner membrane, intermembrane space and matrix) in mitochondria as in DeepMito. Processing the data as in the DeepMito method, the *SM424-18* dataset and the *SubMitoPred* dataset were split into 10 and 5 folds,

respectively. The comparison was based on the MCC of different compartments from the cross validation as claimed in the DeepMito paper[20]. MULocDeep performed better than DeepMito for every mitochondrial compartment in both datasets (Table 2). We also compared with DeepLoc at the subcellular level using a variant model. The comparison results can be found in Supplementary Note 1.

Table 2. Comparison of method effectiveness between MULocDeep and DeepMito.

Dataset	Method	Feature/CV method	MCC(O)	MCC(I)	MCC(T)	MCC(M)
SM424-18	DeepMito	SEQ	0.17	0.15	0.13	0.07
		PROP	0.17	0.07	0.22	0.13
		PSSM	0.51	0.47	0.42	0.57
		SEQ+PROP	0.16	0.07	0.55	0.09
		PSSM+PROP	0.46	0.47	0.53	0.65
	MULocDeep	PSSM+PROP	**0.53**	**0.59**	**0.59**	**0.67**
SubMitoPred	SubMitoPred	RS	0.42	0.34	0.19	0.51
	DeepMito	RS	0.45	0.68	0.54	0.79
	DeepMito	CL	0.42	0.60	0.46	0.76
	MULocDeep	RS	**0.67**	**0.76**	**0.67**	0.79

The comparison is at the sub-organellar level. Four target compartments are “O”: Outer membrane, “I”: Inner membrane, “T”: Intermembrane space, and “M”: Matrix. Features used include one-hot encoding residue (SEQ), physico-chemical properties (PROP), and position specific scoring matrix (PSSM). Cross-validation (CV) methods include randomly splitting the dataset (RS) and confining local similarity into the same cross-validation set (CL). The assessment is based on Mathew’s correlation coefficient (MCC).

Attention weight interpretation and visualization

MULocDeep can not only make accurate localization predictions, but also indicate the contribution of each amino acid in localization and suggest localization motifs. This is achieved by attentive embedding through assigning higher weights to specific parts of a protein sequence. We assume that the regions with higher attention weights are more likely to contribute to the localization. When using a high resolution of attention, it is possible to predict sites and motifs relevant to protein localization. For example, the peptide cleavage site could be predicted directly from an amino acid level attention[19]. Our method provides interpretable results for all the 45 types of suborganelle localizations as far as 500 amino acids from each terminus.

Firstly, we used several proteins with known localization signals as cases to demonstrate the ability of attention weights for indicating the contribution of each amino acid in localization. The proteins are: SV40 large T antigen (P03070) located at the nucleus, with the known signal motif “PKKKRKV” in the middle of

the protein sequence; lactalbumin (P09462) located at the secreted pathway, with a known signal peptide "MMSFVSLLLVGILFWATEAEQLTKCEVFQ" at the N-terminus; and COX4 (P04037) located in the mitochondrial inner membrane, with the known transit peptide "MLSLRQSIRFFKPATRTLCSSRYLL" at the N-terminus[32-34]. When using these proteins as input, MULocDeep predicted the localization correctly for all three proteins. In the meantime, we obtained the attention weights for each protein along the sequence, as shown in Figure 2. The x-axis presents the sequence position from the N-terminus to the C-terminus, and the y-axis presents the value of attention weights. It shows that the high attention regions match the corresponding known motifs of the proteins.

Next, we investigated the attention weights in terms of groups of proteins from the same subcellular compartments and the same suborganelle compartments. Firstly, we visualize the attention weights of

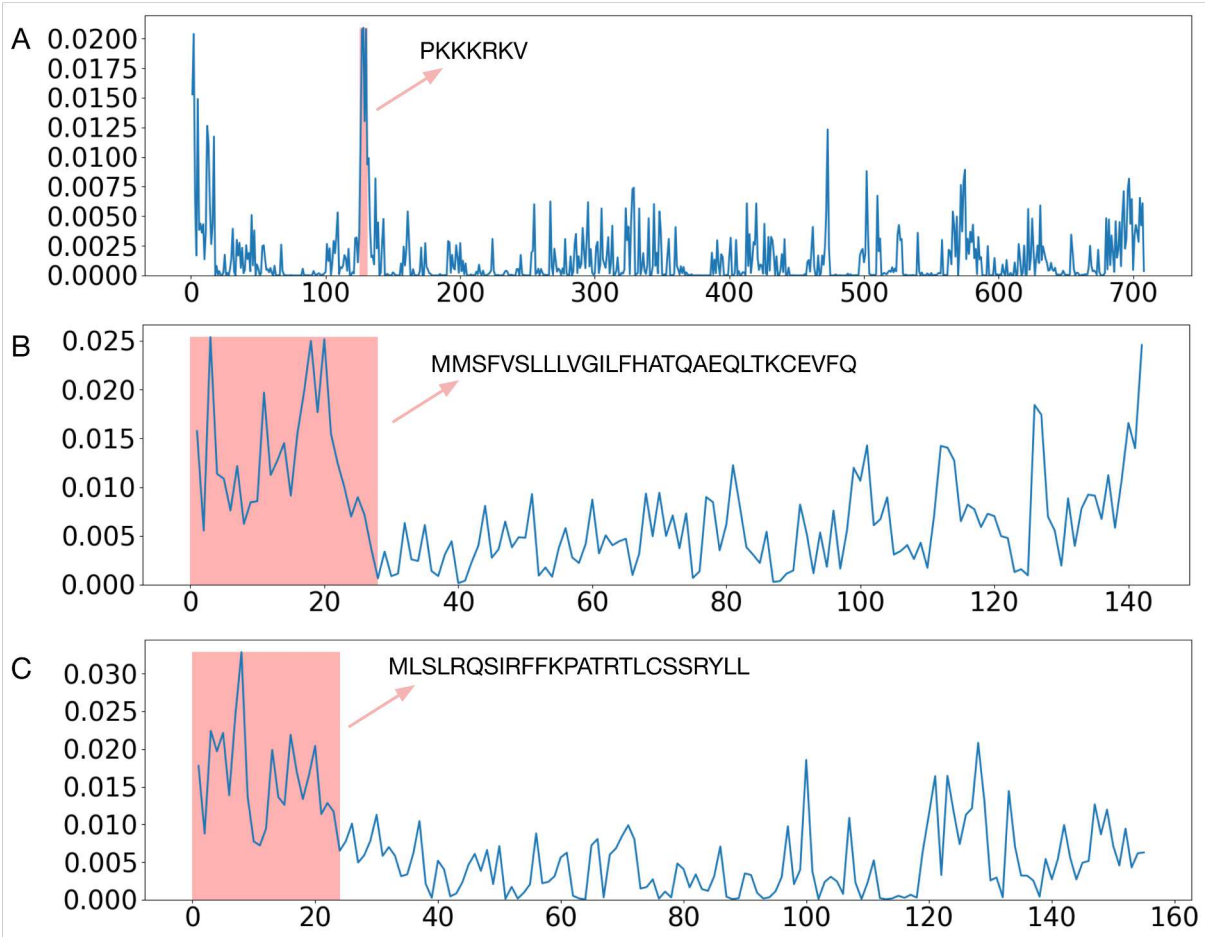


Figure 2. The Visualization of attention weights for (A) SV40 large T antigen (P03070), (B) lactalbumin (P09462), and (C) cytochrome oxidase subunit 4 (P04037). The region of the known sorting motif is highlighted in peach and labelled with the sequence of known localization signals.

proteins from ten subcellular compartments (Figures S2 to S4). Among them, the localization of proteins in extracellular, mitochondrial, plastid, and thylakoid lumen (Figure S2) are believed to be controlled by signal peptides near the protein N-terminus[19]. Comparing to other localizations (Figures S3 and S4), the signals near the N-termini of proteins in Figure S2 have higher attention weights, more over-represented amino acid patterns, and maintain at high levels for longer sequence segments. Our result is consistent with TargetP in detecting the N-terminal sorting signals using attention weights[15,19]. These N-terminal sorting signals are often proteolytically removed at the cleavage cites after the protein reaches the final destination. We then aligned the weighted sequences of these four types of proteins at the cleavage site. The cleavage site annotation was obtained from the UniProt database. The attention visualization result is shown in Figure 3. An immediate decrease in the attention weight is observed after the cleavage site for

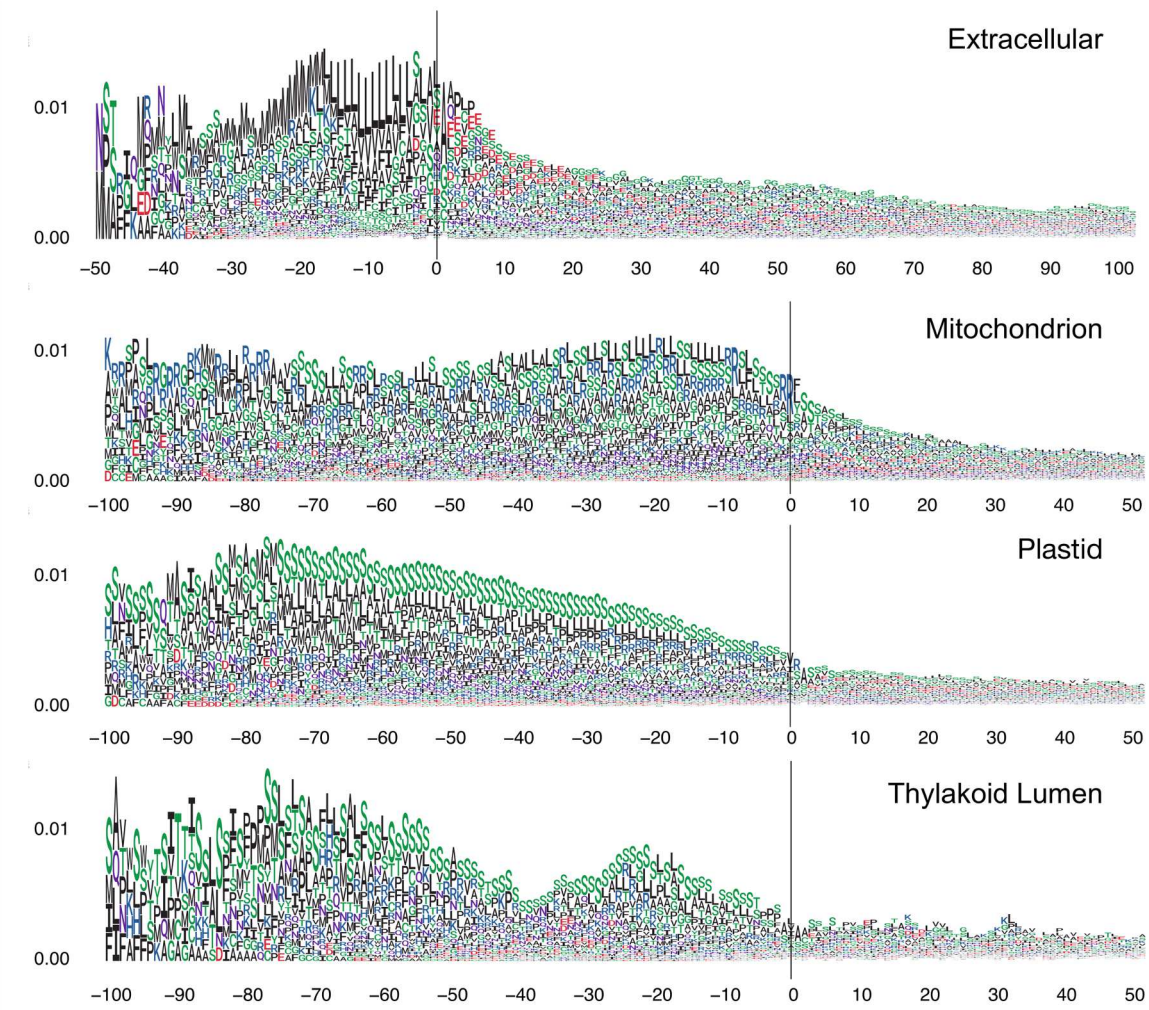


Figure 3. The attention weight visualization by aligned N-terminal sequences at the cleavage site for proteins localized at extracellular, mitochondrion, plastid and thylakoid lumen. The vertical lines indicate the cleavage sites. For extracellular proteins, the range cover 50 AAs before cleavage cite and 100 AAs after the cleavage cite. For the other three classes of proteins, the range covers 100 AAs before the cleavage cite and 50 AAs after the cleavage cite.

the proteins from all four subcellular localizations. This indicates that the high attention weights near the N terminus are mostly contributed by signal peptides and transit peptides.

Looking at the attention weights at the termini of proteins in all ten subcellular localizations (Figure S2-S4), it is apparent that the attention weight increases towards the termini in all cases although more so at the C-terminus. We, therefore, wondered if there is a terminus attention bias introduced by the MULocDeep method. We did a control experiment to test for such a terminus bias on the proteins. For each localization, we randomly shuffled the order of amino acids for each protein sequence. Then we plotted the attention weights aligned at termini for these four localizations and did find some terminus bias (Figure S5). We can use this to distinguish true and false-positive signals, which all give high attention weights near the termini. As shown in Figure S5, false-positive signals are characterized by a gradual

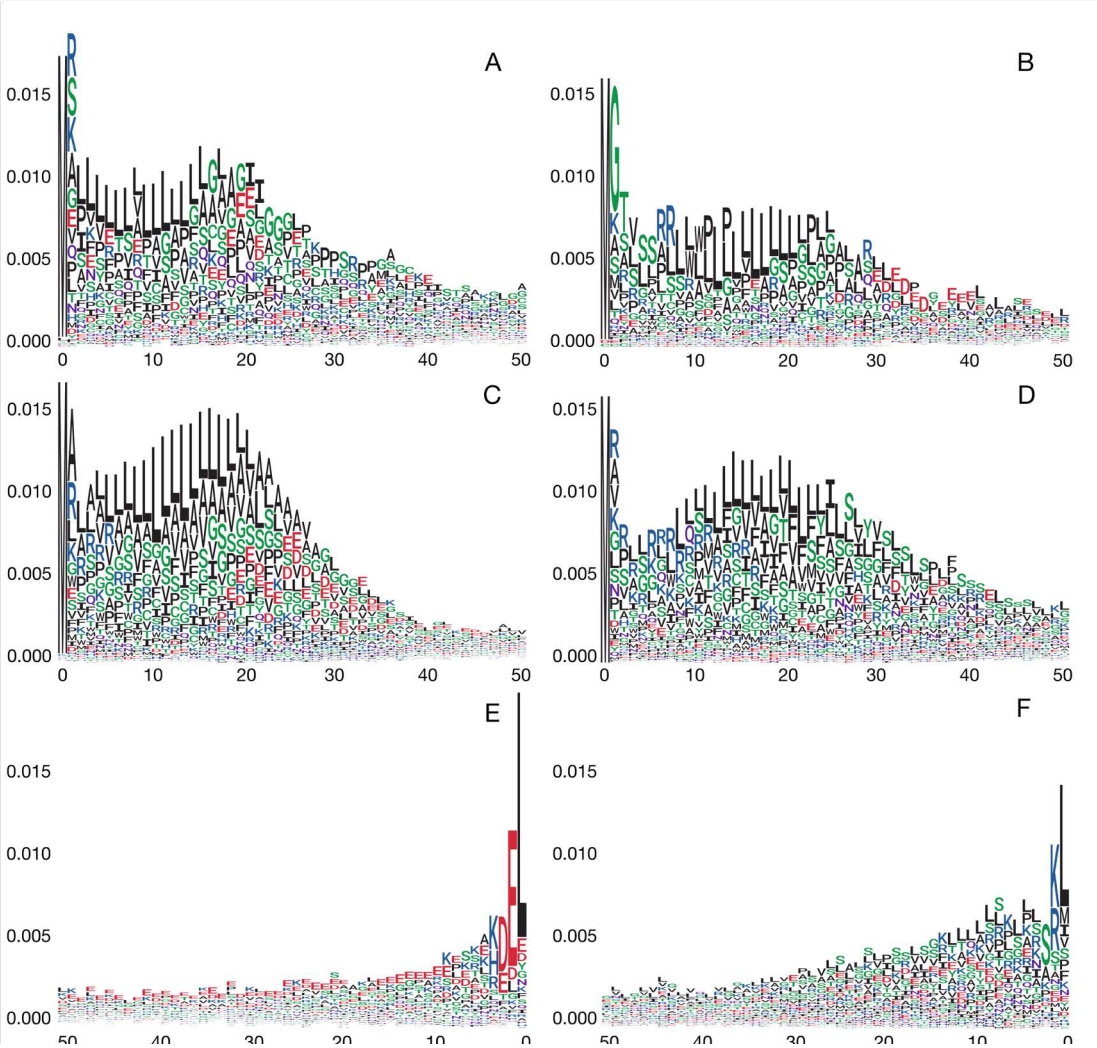


Figure 4. Attention weights of 50 amino acids near the termini at the sub-organellar level, which suggests potential localization signals. (A) N-terminus of cytoplasmic granule. (B) N-terminus of cell surface. (C) N-terminus of endoplasmic reticulum lumen. (D) N-terminus of Golgi apparatus, Golgi stack membrane. (E) C-terminus of endoplasmic reticulum lumen. (F) C-terminus of peroxisome proteins.

decrease from the terminus and the absence of dominant amino acids at each residue position. In contrast, the features of true signals are: amino acids in each position are more conserved, the high attention signal lasts relatively longer and sometimes the highest attention weight appears as a 'bump' away from the termini (in all four localizations shown in Figure S2). Hence, although the terminus attention bias exists, the attention weights in MULocDeep may add values to illustrate biologically significant signals.

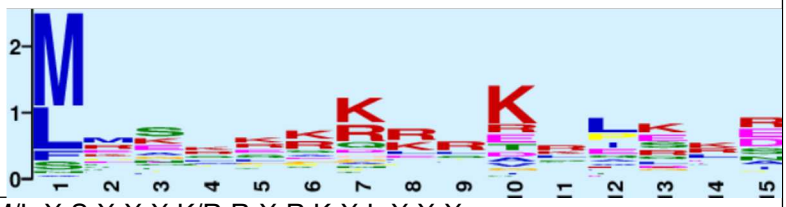
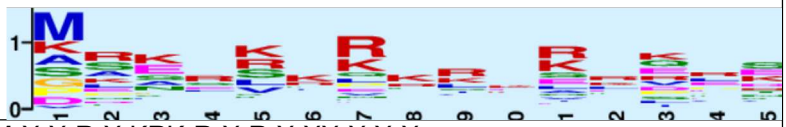
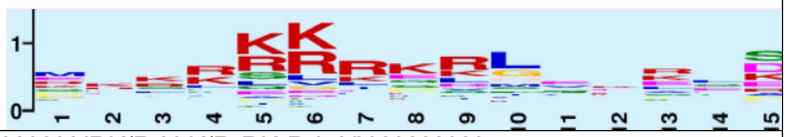
Localization	Rank	Motif	
Nucleus	1 (4061)	Logo	
		RE.	M/L-X-S-X-X-X-K/R-R-X-R-K-X-L-X-X-X
	2 (3680)	Logo	
		RE.	M-X-X-R-X-KRK-R-X-R-X-XX-X-X-X
	3 (2001)	Logo	
		RE.	X-X-X-XRK/R-X-K/R-RK-R-L-XX-X-X-X-X

Figure 5. The top three GLAM2 results for segments from nucleus proteins. For each result, its rank, score, sequence logo and the regular expression (RE) of the motif are given.

Since the attention weights of a subcellular localization reflect the average of attention weights of its suborganellar localizations, there still could be suborganellar localizations that have strong N-terminal or C-terminal sorting signals even though the subcellular sorting signal is not obvious. We, therefore, show the attention weights of proteins at the suborganellar localizations in nucleus, cytoplasm, cell membrane, endoplasmic reticulum, Golgi apparatus, lysosome, and peroxisome (Figures S6-S12). We found several suborganelle localizations with strong signals near the termini. These signals combining one significant signal at the subcellular level are shown in Figure 4, including N-terminal signal peptides in proteins from cytoplasmic granule (Figure 4A), cell surface (Figure 4B) and endoplasmic reticulum lumen (Figure 4C), and perhaps also in Golgi apparatus membrane and Golgi stack membrane (Figure 4D), which all resemble the endoplasmic reticulum signal peptide observed for extracellular proteins in Figures S2 and S4. A C-terminal KDEL/HDEL signal in the endoplasmic reticulum lumen proteins (Figure 4E), a C-terminal SRL/SKL/SRM signal for the peroxisome (Figure 4F), and a less clear sequence for the peroxisome membrane (Figure S12)[35,36] are observed.

Besides the signals near the termini, we also analyzed the attention weights in the middle. We aligned sequences in the same way as we did at the termini, but no signal was found at all. A likely reason is that the signals in the middle do not appear in the same position for different proteins. Thus, the sorting signal in the middle of protein sequences was analyzed and visualized with the help of the GLAM2 tool[37] in the MEME Suite, which can discover over-represented, position-independent motifs in protein sequences (see Method for details). A well-known internal signal is the nuclear localization signal (NLS). The visualization in Figure 5 using nuclear proteins was obtained by setting the "initial columns" (initial number of aligned columns in the motif) equals to 15, the "maximum columns" (upper bound on the number of aligned columns in the motif) equal to 30 while other parameters remain the default in the GLAM2 online tool. The classical NLS pattern includes stretches of 4-5 Lys or Arg (e.g. KKKK)[32], which was readily recognized by GLAM2 (Figure 5).

We carried out similar GLAM2 analyses on proteins from six more subcellular localizations (Figures S13-S18) and found a number of well-defined internal signals where the signal found in the cytoplasm deserves special mention (Figure S13, motif rank 2) – A L/I/VxxxxxxL/V/I/F motif, which is known to be a nuclear export signal[38].

Application of MULocDeep in human proteome

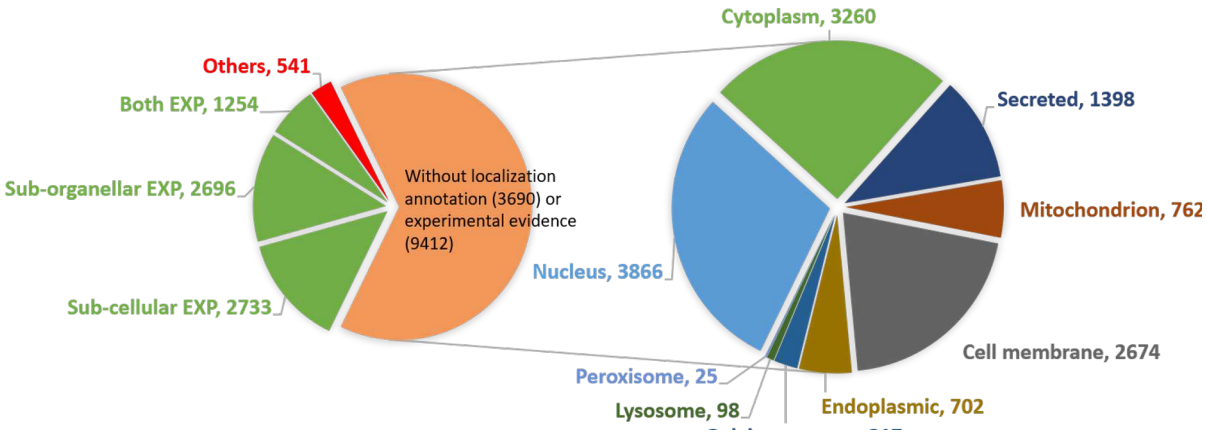


Figure 6. The human proteome localization pie chart. The left pie chart is based on the protein localization annotation, including with or without annotation, and with or without experimental evidence (ECO: 0000269). "Others" include those proteins contain annotations not in any of the 10 major sub-cellular or 45 sub-organellar localization annotations in MULocDeep, or have different localizations for different protein isoforms. The right pie chart is based on the distribution of prediction for the proteins in the orange part in the left pie chart.

We applied MULocDeep in human proteome and drew pie charts of statistics in Figure 6. The human proteome data is collected from the UniProt database release 2020_2[27]. According to the left pie chart in Figure 6, many human proteins have localization annotations. The green part in the left pie chart are proteins with experimentally verified localization annotations (ECO: 0000269). We further divided this part based on if the experimental evidence is marked on subcellular level, suborganellar level, or both. Since

subcellular localization can be inferred from suborganellar annotation, one can consider all the proteins in the green part have subcellular level experimental evidence. For the proteins without experimental evidence or without localization annotation at all (proteins in the orange part in the left pie chart), we combined them together and applied MULocDeep. The right pie chart is drawn based on the prediction results. Particularly, more than half of the predicted proteins are localized at nucleus and cytoplasm. This result is consistent with the conclusion in Thul *et al*[39], which was obtained by mapping 12,003 human proteins at a single-cell level using immunofluorescence microscopy. Furthermore, according to Figure 6, the number of human mitochondrial proteins is 1388 (combining the proteins with experimental evidence and the proteins predicted by MULocDeep). This number matches the conclusion in Calvo *et al*[40], which estimated 1100-1400 distinct proteins in the human mitochondrial proteome. Above all, the prediction results in Figure 6 reflects the true localization distribution in human and demonstrates the ability of MULocDeep for proteome-wise annotation.

The MULocDeep web server

We developed a user-friendly website (http://www.mu-loc.org/) to make the application of the MULocDeep method more accessible. Every page comes with a “help” button, which explains how the specific page works. Every user has his or her personal workspace in our database in which they can manage their jobs conveniently. Many web servers offer protein localization prediction. MULocDeep is unique in that this is the only web server that can predict all 45 suborganellar localizations within 10 main subcellular compartments. And to the best of our knowledge, it is the only web server that provides an amino acid level interpretation and visualization. Users can use this web server as a protein localization prediction tool or a hypothesis generator regarding protein sorting signal motifs.

Discussion

In this paper, we present “MULocDeep”, a deep learning model for protein localization prediction (shown in Figure 1). The core of the model is the bidirectional LSTM to handle protein sequence information and the multi-head self-attention to assign weights to each amino acid of a sequence for interpretation. Some methods[17] added CNN layers before LSTM layers and reported high performance. Here, we discard CNN to pursue a residue-level interpretation resolution for more biological insight.

To maximize prediction accuracy, we used a Bayesian optimization method to determine the hyperparameters. However, experiments showed that the performance was insensitive to the selection of hyperparameters (Table S2). This indicates that our model is robust. The selected configuration of hyperparameters remained the same when training a variant model to compare with others using their own training data, instead of optimizing a new set of hyperparameters with the specific data, which could have yielded slightly better results. The comparison with the DeepMito method on the *DeepMito* dataset again shows that MuLocDeep is insensitive to hyperparameters. Although no hyperparameters were retrained, MuLocDeep still outperformed the DeepMito method significantly (Table 2). When comparing MULocDeep with DeepLoc, the results indicate that they are about equally good at localization prediction

(each method has 5 categories with higher MCC than the other method in Table S1). However, from the perspective of interpretability and the prediction performance as a tool, MULocDeep is superior to DeepLoc (according to the results in Table 2).

We published a dataset of the experimentally extracted mitochondrial proteome from three species (*Mito3* dataset) and used it to test the performance of subcellular localization prediction by different methods (Table 2). We also collected a comprehensive dataset for 45 suborganelle protein localization from the UniProt database (*UniLoc* dataset). The MULocDeep model was trained based on the UniLoc-*train* dataset combining the training samples in the *DeepLoc* dataset. Tables S3 and S4 show the cross-validation results at the suborganelle level with different thresholds. We noticed that there are classes with very low MCC and some of these classes have very small samples. Thus, further dividing these samples into eight folds would result in folds with tens of positive samples and thousands of negative samples for a specific class. The calculation of MCC for such cases is not meaningful.

The *UniProt-test* dataset and the *Mito3* dataset were used for a comprehensive evaluation of localization classifiers at both subcellular and suborganellar levels (Table 2). The MULocDeep method generally outperforms other methods at both levels. But we also noticed that the inner membrane in the mitochondrion is better predicted by DeepMito than by our method probably due to different training data, as we have already demonstrated the MULocDeep's algorithmic superiority over DeepMito using the same *DeepMito* datasets as the training data (Table 2).

The attention weight mechanism in the MULocDeep model has the ability to detect sorting signals at the protein termini and in the middle of protein sequences. The interpretation was validated by matching the known transit peptides for proteins located in the nucleus, extracellular, mitochondria, plastid, and thylakoid lumen (Figure S2). Particularly, plastid and thylakoid lumen proteins have a high attention peak at position 2 from the N terminus, which is enriched in alanine. A region enriched in serine follows. The signal peptide in extracellular proteins apparently consists of a positively charged amino acid followed by a number of hydrophobic residues, a well-known endoplasmic reticulum localization signal (Figure S4)[41]. The mitochondrion transit peptides are enriched in arginine and leucine. For other subcellular (Figures S3 and S4) and suborganellar (Figures S6-S12) classes of proteins, we also found high attention regions near both N- and C-termini. Although most of them appear to be caused by terminus bias, several known signals were recognized; in particular, a well-known N-terminal endoplasmic reticulum signal peptide recognized in extracellular proteins (Figure S2) was also seen in proteins from cytoplasmic granule (Figure 4A) and cell surface (Figure 4B), both of which pass through the endoplasmic reticulum on the way to those localizations[41]. This N-terminal signal was particularly strong in proteins belonging to the endoplasmic reticulum lumen (Figure 4C), where it was seen together with a C-terminal KDEL/HDEL signal (Figure 4E), which is an endoplasmic reticulum retention signal[42-44]. We also found a clear C-terminal SRL/SKL/SRM signal for the peroxisome (Figure 4F). It turns out to be a targeting signal for peroxisome proteins[35,36].

To analyze the attention in the middle of proteins, we used the GLAM2 tool in the MEME suite, and we demonstrate the effectiveness of this approach by visualizing the attention of proteins located in the nucleus (Figure 5). We also found a very clear L/I/VxxxxxxL/V/I/F motif in proteins belonging to the cytoplasm (Figure S13), which is known to be a nuclear export signal[38]. Note that the analysis by GLAM2 was influenced by the number of top amino acids and the window size we chose. The parameters of the GLAM2 itself could also affect the results. Changing the parameter configuration would result in different visualization results, but the general motif information remains the same. For other localizations, where the mechanisms are unclear, we provide a list of attention visualizations in Figures S13-S18. They could be used as candidate motifs for these localizations.

There are limitations to the MULocDeep method, which may lead to several future works. First of all, more advanced machine-learning methods, such as graph-based neural networks, could be applied for feature representation and localization prediction. Secondly, a more rigorous confidence assessment of each predicted localization could be provided instead of the current prediction probabilities, and each predicted motif can be given a confidence assessment as well. Furthermore, the prediction performance could be further improved by building animal/plant specific models. Finally, future applications could be extended to localization-related disease studies, e.g. to predict the impact of a mutation in localization. With that said, we are still a long way from understanding protein localization fully. Given the complexities of decoding signal peptides or extracting features that can distinguish detailed levels of localizations, the interpretability of MUlocDeep makes it possible to generate hypotheses regarding protein sorting mechanisms. Some of them can be verified by current knowledge, making the rest worth exploring by combining computational prediction methods and experimental verification. By making the datasets, the software code, and the MULocDeep website publicly accessible, our work constitutes a major step toward understanding the protein localization mechanism.

Methods

Datasets

The three-species mitochondrial proteome dataset (Mito3). This dataset contains the mitochondrial proteome extracted from three plant species and was generated as follows (details can be found in Supplementary Note 2). First, mitochondria were isolated from an *Arabidopsis thaliana* cell culture, *Solanum tuberosum* tubers, and *Vicia faba* roots. Then, the mitochondrial proteins of the three species were identified and quantified using shotgun mass spectrometry according to Thal *et al*.[45]. The generated raw files were analysed with the Proteome Discoverer software (Thermo Fisher Scientific, Dreieich, Germany) using the Mascot (Matrix Science, London, UK) search engine against in-house protein sequence databases, depending on the analyzed species: *Arabidopsis thaliana*, *Solanum tuberosum* and *Vicia faba* spectra were queried against the TAIR 10 database, the *Solanum tuberosum* protein database and the *Medicago truncatula* protein database 4.0, respectively. A total of 8002 proteins from these three species were identified in this step. Finally, the proteins of Vicia and Solanum were assigned to their Arabidopsis orthologues by pairwise BLAST searches of the underlying sequence datasets. The identified

proteins from the three species with the respective Arabidopsis orthologues are shown in the Supplementary Data 1-3. In total, 4778 unique mitochondrial proteins were identified. Since this dataset is used for localization classifier evaluation at the subcellular level (for evaluation of predicting if a protein is mitochondrial or not), we balanced it by collecting 8002 plant proteins that were labelled as non-mitochondrial in the UniProt database (release 2020_2). The proteins in this dataset were not used in training the model.

Suborganelle dataset (UniLoc). The protein sequence and localization annotations were downloaded from the UniProt database[27], release 2018_11 and 2020_2. For both release versions, we collected all metazoan and Viridiplantae proteins with the Uniport evidence level <= 2, which means these proteins' localizations have experimental evidence at transcript or protein level. Only proteins with suborganelle localization annotations under ten main subcellular localizations were kept. One protein can have more than one localization annotation, and we chose suborganelle localizations with more than 50 proteins and ignored others in this study. Finally, 45 suborganelle localizations remained. This is the most comprehensive multi-label dataset for protein localization annotations down to the suborganelle level to date. In the release 2018_11, a total of 38,747 proteins were obtained. This part of the dataset was for training (*UniLoc-train*) and was divided into 8 folds. We used CD-hit[46] to ensure that proteins in different folds have identity less than 40% sequence identify. In the release 2020_2, we removed the protein sequences already found in release 2018_11 or used in training the model, after which 1211 proteins remained. This part of the dataset (basically release 2018_11 to release 2020_2) was used for protein localization classifier evaluation (*UniLoc-test*). The statistics for the dataset are shown in Figure S20. It is possible that there are still proteins in *UniLoc-test* homologous to proteins in *UniLoc-train,* but it is expected that *UniLoc-test* has homology to other method's training data as well, so the performance bias probably applies to all methods in a similar way. Such a comparison without removing homology also reflects real applications from actual users.

DeepLoc dataset. The DeepLoc method predicts 10 main protein localizations at the subcellular level. One of the contributions of this work is that it provided a benchmark dataset[17]. It was extracted from the UniProt database, release 2016_04. The following criteria were applied when they collected the data: Eukaryotic, not fragments, encoded in the nuclear genomes, longer than 40 amino acids and experimentally annotated (ECO: 0000269). They merged the similar localizations or subclasses of the same localization into 10 main localizations (as shown in Table S1). Unlike our suborganelle dataset, the DeepLoc dataset is a multi-class but not multi-label dataset, i.e., proteins with more than one subcellular localization were filtered out. A total of 13,858 proteins were obtained after the filtering process (See Table S5 for category details). The training and test data were already separated with less than 30% identity. Like we did to the suborganelle dataset, we divided the training data into 8 folds, each fold has similar distributions of localizations with other folds. Proteins in each fold share less than 40% sequence identity with proteins in any other fold.

DeepMito datasets. The DeepMito method focuses on the prediction of sub-mitochondrial protein localization. Two datasets were used in DeepMito method. The *SM424-18* dataset was derived from the UniProt database (release 2018_02). They filtered the proteins by selecting all non-fragment protein sequences with evidence at the protein level for experimentally determined subcellular localization in one of the four sub-mitochondrial compartments: outer membrane, intermembrane space, inner membrane and matrix. They further reduced the redundancy using the CD-HIT program so that 424 mitochondrial proteins remained sharing at most 40% sequence identity. The other dataset was the *SubMitoPred* dataset, which was derived from the UniProt database release 2014_10. The protein selection criteria were: Full-length proteins >50 residues, single experimental sub-mitochondrial localization, and internal redundancy reduced at 40% sequence identity using CD-HIT. The dataset comprises 570 mitochondrial proteins distributed in the same four compartments as in the SM424-18 dataset. Both datasets were split into folds, 10 for the SM424-18 dataset and 5 for the SubMitoPred dataset. The category details can be found in Table S6.

Protein sequence representation

An encoded amino acid contains two parts. The first 5 digits come from the first five eigenvectors of a comprehensive list of 237 physical-chemical properties for each amino acid[47]. As we did in the domain boundary prediction study[48], these 5-number descriptors can represent each amino acid for computational efficiency while maintaining almost all the information. The last 20 numbers come from the position-specific scoring matrix (PSSM) profile of a protein. A protein's PSSM profile is usually generated through a multiple sequence alignment against a large database. Some methods try to accelerate the process by searching a relatively small database first, and if no hit is found then use a large database instead[17,49]. We further shortened this process by two steps, first by scanning the Swissprot[50] using PSI-blast[46]. The Swissprot database is a much smaller database than UniProt, yet most of the proteins that we studied find hits (147 of 38,747 were missed in the suborganelle dataset). If no hits were retrieved, in the next step, the BLOSUM62 encoding[51] was applied directly, which did not take any search time. In this way, we saved much computational time without significant performance decrease (see Results). Since the length of proteins varies, we fixed the protein-encoding length at 1000 AA. If a protein exceeded this length, the first 500 amino acids from N-terminus and the last 500 amino acids from C-terminus were preserved and combined. If a protein had a shorter sequence, we padded it to 1000 AA at the end and masked the padding part for the following calculation.

Parameter tuning and neural network training

The hyperparameters in the models are determined through a Bayesian optimization process. These hyperparameters include the hidden dimensions, the number of heads in attention, regularizers, dropout rates, *etc*. The MULocDeep model is an ensemble of eight "sub-models" derived from an 8-fold cross validation. Each of these eight sub-models was optimized individually and the hyperparameters of the sub-model that achieved the highest accuracy were set as the global optimum. Table S7 lists the

hyperparameter optimization results for all eight sub-models in the cross validation. The hyperparameter configuration in sub-model 1 was selected as the global optimum. Therefore, in the MULocDeep model, all the eight sub-models used the same global optimized hyperparameters. We also tested the performance of an ensemble of sub-models where each sub-model 1) used its own optimized hyperparameters, 2) used the same, but not the global optimized, hyperparameters (select two from seven non-global optima), or 3) used the same random hyperparameters. The results are shown in Table S2. It turns out that the difference in performance among various ensemble models was insignificant, except for a notably bad performance by using the randomly selected hyperparameters. We also conducted an experiment to test the performance of individual models using the globally optimized hyperparameters. The results are listed in Table S8. Comparing the results in Table S2, the ensemble models generally have a better performance than individual models.

The *UniLoc-train* dataset and the training samples in the *DeepLoc* dataset were already divided into 8 folds as described above. Eight models were trained where each of them used 7 folds as training and 1 fold was left for evaluation. The final MULocDeep model is the ensemble of eight models, and the prediction of a protein is the average of predictions from these eight models. All these models used the same global optimized hyperparameters. To train each of these models, the *UniLoc-train* dataset and the *DeepLoc* dataset were utilized iteratively. We trained them using the samples with only subcellular localization labels for 50 epochs, and then trained another 50 epochs using the samples with suborganelle localization labels. When a training sample had both suborganelle and subcellular (inferred from suborganelle) annotation, each element in the matrix (Figure 1) yielded a loss using a binary cross entropy loss function after a sigmoid activation function (Lost 1). The maximum prediction score under each organelle (each column in the matrix in Figure 1) was extracted and went through another binary cross entropy (Lost 2). If a training sample only had the subcellular localization information, the Lost 1 was not used, only the Lost 2 was calculated after the Max-Pooling operation.

The training process was written using the Keras package (version 2.3.0) and run using a NVIDIA GeForce RTX 2080 Ti GPU. The training time for the MULocDeep model was roughly 2 min for one epoch.

Bayesian optimization

We formulate the accuracy “ACC” as the objective function and it is a function of all the hyperparameters. A Gaussian process is used as the surrogate model to approximate the objective function. We use the expected improvement (EI) as the acquisition function, which directs sampling to areas where an improvement over the current best observation is likely. The acquisition jitter, which trades off exploitation (high objective) and exploration (high uncertainty) was set as 0.05.

Since the optimization process would take a long time, we used samples with subcellular localization labels only (training samples in the *DeepLoc* dataset). We had already divided these samples into 8 folds using CD-hit[46] and the sequence identity between proteins in different folds was below 40%. Then, an 8-fold cross validation was performed. Each hyperparameter has a searching space. During the cross

validation, 7 folds were used for training a sub-model under one specific hyperparameter configuration for 40 epochs. The remaining fold evaluated the accuracy in each epoch. The highest accuracy on the validation fold during the 40 epochs was recorded as the accuracy for this hyperparameter configuration. In total, 150 hyperparameter combinations were tested. Figure S21 shows the accuracy along the testing process. The hyperparameter configuration which achieved the highest accuracy among the 150 combinations was used as the optimized configuration for this sub-model. Thus, the optimization process would run 40 (test one specific configuration) * 150 (150 configurations to test in total) * 8 (8-fold cross-validation), which is 48,000 epochs in total. Finally, each of the eight sub-models had its own optimized hyperparameter configuration and the corresponding accuracy achieved. The configuration with the highest accuracy was selected as the global optimized configuration that was used for the MULocDeep model and its variant models.

Multi-head self-attention

The multi-head self-attention[24] uses the overall semantics of the whole sentence formed by multiple components in a sentence. So, multiple hops of attention are needed to focus on different parts of the sentence. Our method borrows this idea and sets the number of heads equals to 41 (derived from the hyperparameter tuning). The final weight of each amino acid is the average of the 41 weights. Then we could analysis if any "important parts" of a protein sequence are responsible for the protein localization. The attention matrix A is calculated as Eq. (1)

$$A = softmax(W_{s2}\tanh(W_{s1}H^T)) \quad \text{Eq. (1)}$$

where H is the 1000-by-180 embedding sequence output from bidirectional LSTM. W_{s1} is a weight matrix with a shape of 369-by-180. W_{s2} is a matrix of parameters with shape 41-by-369. The attention matrix A is returned separately for interpretation. The sequence embedding M, calculated as the weighted sum by multiplying *A* and *H* (Eq. 2), is also returned for further prediction.

$$M = AH \quad \text{Eq. (2)}$$

When training the models, we applied the penalization term P below[24]

$$P = \|(AA^T - I)\|_F^2 \quad \text{Eq. (3)}$$

where A is the attention matrix, *I* is an identity matrix, $\|\cdot\|_F$ stands for the Frobenius norm of a matrix. We multiple the penalization term with an attention regularizer and added the product to the model's loss. The loss gets high if two attention vectors are identical. Thus, by using this penalization term, we encourage the attention vectors to concentrate on different parts of a protein sequence.

Attention Visualization

To visualize the attention for one protein, the average of the protein's attention matrix was calculated along the dimension of "heads", which is 41 in our model. A 1000-long vector is left and ready for visualization.

To visualize the attention of a group of proteins that belong to the same category, different methods are applied based on the following two scenarios:

(1) For analyzing attention at termini, we simply align the first 50 or the last 50 amino acids one by one from left to right or right to left depending on if it is N-terminus or C-terminus. For each position, the frequency of amino acids and the weight of attention (average along with the number of "heads" and proteins) are obtained. Then we used the R package "ggseqlogo"[52] to visualize the attention.

(2) For analyzing attention in the middle of proteins, we kept the entire protein sequence and ranked the amino acid in it based on their attention weights. We selected the top 5 amino acids, each of them combining the surrounding 20 amino acids (10 window size at each side) to form a segment. A final segment was obtained by concatenating all segments by a string of "X" with the same length of window size. All segments belonging to the same class were analysed by the GLAM2[37], which is a tool in the MEME Suite 5.1.0 that can discover variable-length, gapped motifs.

Evaluation criteria

We used accuracy (ACC), Matthew's correlation coefficient (MCC)[53], recall, precision, area under receiver operating characteristic curve (ROC_auc), and area under precision & recall curve (P&R_auc) to evaluate our method and compared with others. For unbalanced datasets, measurements such as ACC, recall and precision would introduce bias and overestimate a method's performance. MCC considers true and false positives and negatives, and is generally regarded as a balanced measure even if the classes are of very different sizes[54]. The definitions of ACC, MCC, recall and precision are listed in Eqs. 4-7:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Eq. (4)}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{Eq. (5)}$$

$$recall = \frac{TP}{TP+FN} \quad \text{Eq. (6)}$$

$$precision = \frac{TP}{TP+FP} \quad \text{Eq. (7)}$$

where *TP, FP, TN, FN* are true positive, false positive, true negative and false negative prediction, respectively. Measurements such as MCC, recall, precision and ACC can be used in binary prediction cases.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[55] partner repository with the dataset identifier PXD019987. Datasets used to train and test the MULocDeep model can be found in the GitHub repository (https://github.com/yuexujiang/MULocDeep).

Code availability

All the codes for training and testing the MULocDeep model can be found in the GitHub repository (https://github.com/yuexujiang/MULocDeep).

References

1 Davis, J. R., Kakar, M. & Lim, C. S. Controlling protein compartmentalization to overcome disease. *Pharmaceutical research* **24**, 17-27, doi:10.1007/s11095-006-9133-z (2007).

2 Hung, M. C. & Link, W. Protein localization in disease and therapy. *Journal of cell science* **124**, 3381-3392, doi:10.1242/jcs.089110 (2011).
3 Hagmann, M. Protein zip codes make Nobel journey. *Science* **286**, doi:10.1126/science.286.5440.666 (1999).
4 Goldenberg, N. M. & Steinberg, B. E. Surface charge: a key determinant of protein localization and function. *Cancer Res* **70**, 1277-1280, doi:10.1158/0008-5472.CAN-09-2905 (2010).
5 Walther, T. C. & Mann, M. Mass spectrometry–based proteomics in cell biology. *Journal of Cell Biology* **190**, 491-500 (2010).
6 Schubert, W. *et al.* Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature biotechnology* **24**, 1270-1278 (2006).
7 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
8 Goldberg, T. *et al.* LocTree3 prediction of localization. *Nucleic acids research* **42**, W350-355, doi:10.1093/nar/gku396 (2014).
9 Briesemeister, S. *et al.* SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *Journal of proteome research* **8**, 5363-5366 (2009).
10 Blum, T., Briesemeister, S. & Kohlbacher, O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC bioinformatics* **10**, 274 (2009).
11 Briesemeister, S., Rahnenfuhrer, J. & Kohlbacher, O. YLoc--an interpretable web server for predicting subcellular localization. *Nucleic acids research* **38**, W497-502, doi:10.1093/nar/gkq477 (2010).
12 Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic acids research* **35**, W585-587, doi:10.1093/nar/gkm259 (2007).
13 Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics* **31**, 3269-3275, doi:10.1093/bioinformatics/btv367 (2015).
14 Small, I., Peeters, N., Legeai, F. & Lurin, C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**, 1581-1590, doi:10.1002/pmic.200300776 (2004).
15 Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology* **300**, 1005-1016 (2000).
16 Emanuelsson, O., Brunak, S., Von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols* **2**, 953 (2007).
17 Almagro Armenteros, J. J., Sonderby, C. K., Sonderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387-3395, doi:10.1093/bioinformatics/btx431 (2017).
18 Zhang, N. *et al.* MU-LOC: A Machine-Learning Method for Predicting Mitochondrially Localized Proteins in Plants. *Front Plant Sci* **9**, 634, doi:10.3389/fpls.2018.00634 (2018).
19 Almagro Armenteros, J. J. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance* **2**, doi:10.26508/lsa.201900429 (2019).
20 Savojardo, C., Bruciaferri, N., Tartari, G., Martelli, P. L. & Casadio, R. DeepMito: accurate prediction of protein submitochondrial localization using convolutional neural networks. *Bioinformatics* **36**, 56-64 (2019).
21 Wang, X., Zhang, W., Zhang, Q. & Li, G. Z. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics* **31**, 2639-2645, doi:10.1093/bioinformatics/btv212 (2015).

22 Wan, S., Mak, M.-W. & Kung, S. Transductive learning for multi-label protein subchloroplast localization prediction. *IEEE/ACM transactions on computational biology and bioinformatics* **14**, 212-224 (2016).
23 Javed, F. & Hayat, M. Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC. *Genomics* **111**, 1325-1332 (2019).
24 Lin, Z. *et al.* A structured self-attentive sentence embedding. *arXiv preprint* (2017).
25 Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint* (2014).
26 Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735-1780 (1997).
27 UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506-D515, doi:10.1093/nar/gky1049 (2019).
28 Fukasawa, Y. *et al.* MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol Cell Proteomics* **14**, 1113-1126, doi:10.1074/mcp.M114.043083 (2015).
29 Sperschneider, J. *et al.* LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific reports* **7**, 1-14 (2017).
30 Ding, H., Liu, L., Guo, F.-B., Huang, J. & Lin, H. Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. *Protein peptide letters* **18**, 58-63 (2011).
31 Lin, H., Chen, W., Yuan, L., Li, Z. & Ding, H. Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta biotheoretica* **61**, 259-268 (2013).
32 Dingwall, C. & Laskey, R. A. Nuclear targeting sequences—a consensus? *Trends in biochemical sciences* **16**, 478-481 (1991).
33 Costantino, A., Balzola, F. & Bounous, G. Changes in biliary secretory immunoglobulins A in mice fed whey proteins. *Minerva dietologica e gastroenterologica* **35**, 241-245 (1989).
34 Lomax, M. I., Welch, M. D., Darras, B. T., Francke, U. & Grossman, L. I. Novel use of a chimpanzee pseudogene for chromosomal mapping of human cytochrome c oxidase subunitIV. *Gene* **86**, 209-216 (1990).
35 Ast, J., Stiebler, A. C., Freitag, J. & Bolker, M. Dual targeting of peroxisomal proteins. *Front Physiol* **4**, 297, doi:10.3389/fphys.2013.00297 (2013).
36 Reumann, S. Specification of the peroxisome targeting signals type 1 and type 2 of plant peroxisomes by bioinformatics analyses. *Plant Physiol* **135**, 783-800, doi:10.1104/pp.103.035584 (2004).
37 Frith, M. C., Saunders, N. F., Kobe, B. & Bailey, T. L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS computational biology* **4**, e1000071 (2008).
38 Kosugi, S., Hasebe, M., Tomita, M. & Yanagawa, H. Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic* **9**, 2053-2062, doi:10.1111/j.1600-0854.2008.00825.x (2008).
39 Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, doi:10.1126/science.aal3321 (2017).
40 Calvo, S. E. & Mootha, V. K. The mitochondrial proteome and human disease. *Annual review of genomics human genetics* **11**, 25-44 (2010).
41 Lodish H, B. A., Zipursky SL, et al. *Molecular Cell Biology*. (W. H. Freeman; 4th edition, 2000).
42 Denecke, J., De Rycke, R. & Botterman, J. Plant and mammalian sorting signals for protein retention in the endoplasmic reticulum contain a conserved epitope. *EMBO J* **11**, 2345-2355 (1992).

43 Gomord, V. *et al.* The C-terminal HDEL sequence is sufficient for retention of secretory proteins in the endoplasmic reticulum (ER) but promotes vacuolar targeting of proteins that escape the ER. *Plant J* **11**, 313-325, doi:10.1046/j.1365-313x.1997.11020313.x (1997).
44 Stornaiuolo, M. *et al.* KDEL and KKXX retrieval signals appended to the same reporter protein determine different trafficking between endoplasmic reticulum, intermediate compartment, and Golgi complex. *Mol Biol Cell* **14**, 889-902, doi:10.1091/mbc.e02-08-0468 (2003).
45 Thal, B., Braun, H. P. & Eubel, H. Proteomic analysis dissects the impact of nodulation and biological nitrogen fixation on Vicia faba root nodule physiology. *Plant molecular biology* **97**, 233-251 (2018).
46 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
47 Venkatarajan, M. S. & Braun, W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *Molecular modeling annual* **7**, 445-453 (2001).
48 Jiang, Y., Wang, D. & Xu, D. DeepDom: Predicting protein domain boundary from sequence alone using stacked bidirectional LSTM. *Pac Symp Biocomput* **24**, 66-75 (2019).
49 Tsirigos, K. D., Peters, C., Shu, N., Kall, L. & Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic acids research* **43**, W401-407, doi:10.1093/nar/gkv485 (2015).
50 Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45-48 (2000).
51 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915-10919 (1992).
52 Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645-3647 (2017).
53 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta -Protein Structure* **405**, 442-451 (1975).
54 Boughorbel, S., Jarray, F. & El Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one* **12** (2017).
55 Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research* **47**, D442-D450 (2019).

Acknowledgements

This work was supported by the US National Institutes of Health grants R21-LM012790 and R35-GM126985. We would like to thank Dr. Hao Lin for providing suggestions in defining subcellular and suborganellar categories. We would like to thank Dr. Ning Zhang for providing the evaluation results by the MU-LOC method.

Author information

Affiliations

Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, Columbia, Missouri, USA
Yuexu Jiang, Duolin Wang, Yifu Yao & Dong Xu

Institute of Plant Genetics, Leibniz University Hannover, Hannover, Germany
Holger Eubel & Patrick Künzler

Department of Molecular Biology and Genetics, Aarhus University, Forsøgsvej 1, DK-4200 Slagelse, Denmark
Ian Max Møller

Contributions

D.X. conceived and supervised the project. Y.J. designed the neural network, developed the code, wrote most of the manuscript. D.W. helped with the neural network design, prepared the UniProt dataset. H.E., P.K., I.M.M. prepared the mitochondrial proteome dataset, provided biological insights, and helped write the manuscript. D.W. and Y.Y. helped with the result evaluation and comparison. Y.Y. developed the MULocDeep website.

Corresponding author

Correspondence to Dong Xu.

Ethics declarations

Competing interests

None declared.

Additional information

Supplementary information

Supplementary information

Supplementary Data 1. Identified Vicia Proteins (as Medicago IDs) with Medicago Sequences and Arabidopsis Match

Supplementary Data 2. Identified Solanum Proteins with Sequences and Arabidopsis Match

Supplementary Data 3. Identified Arabidopsis Proteins with Sequences

Figures

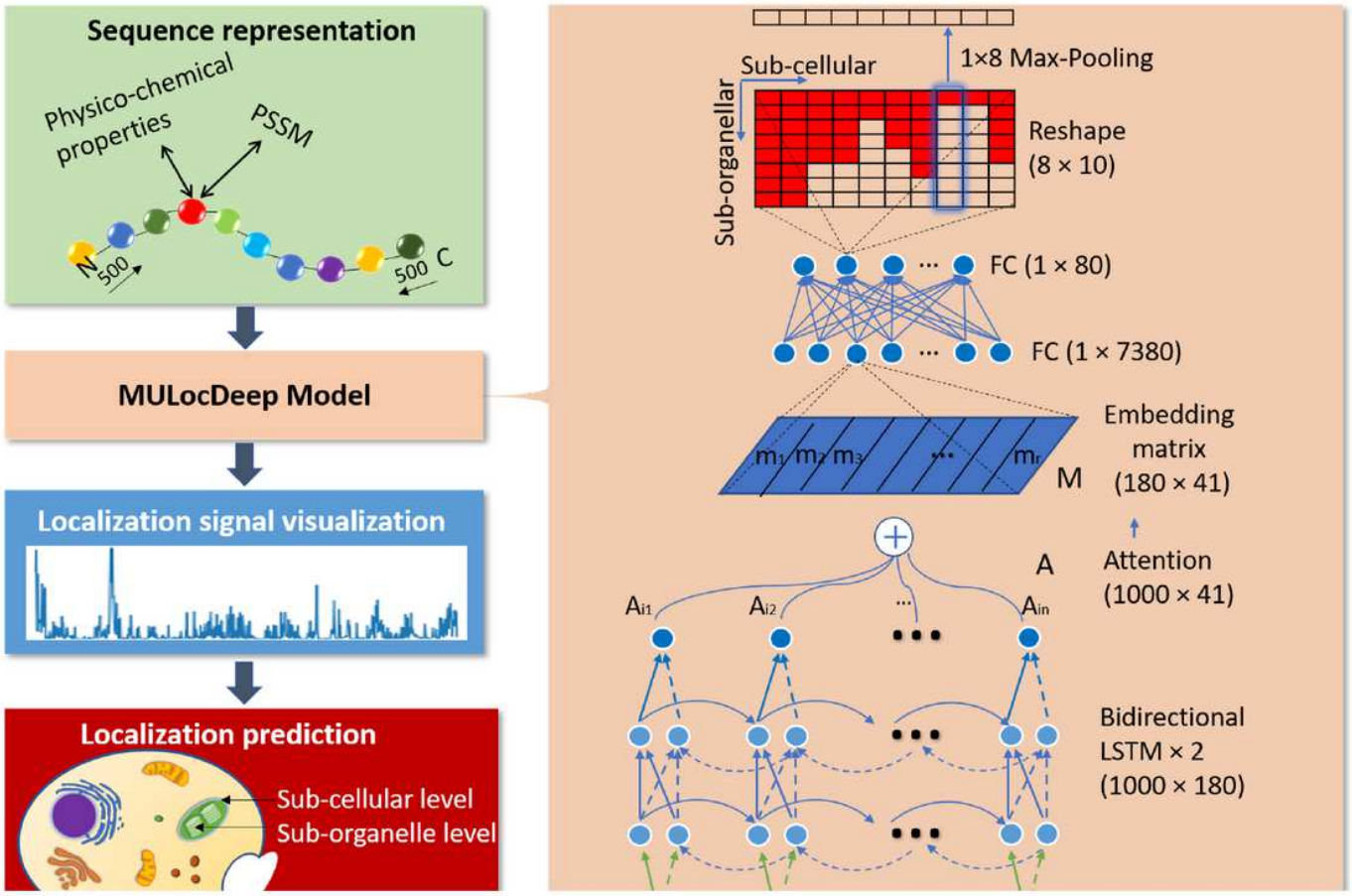


Figure 1

MULocDeep workflow and neural network architecture. The workflow is composed of four steps: (1) Protein sequence representation, (2) training the MuLocDeep model, (3) localization signal visualization, and finally (4) localization prediction. The details of the neural network architecture are displayed in the right panel.

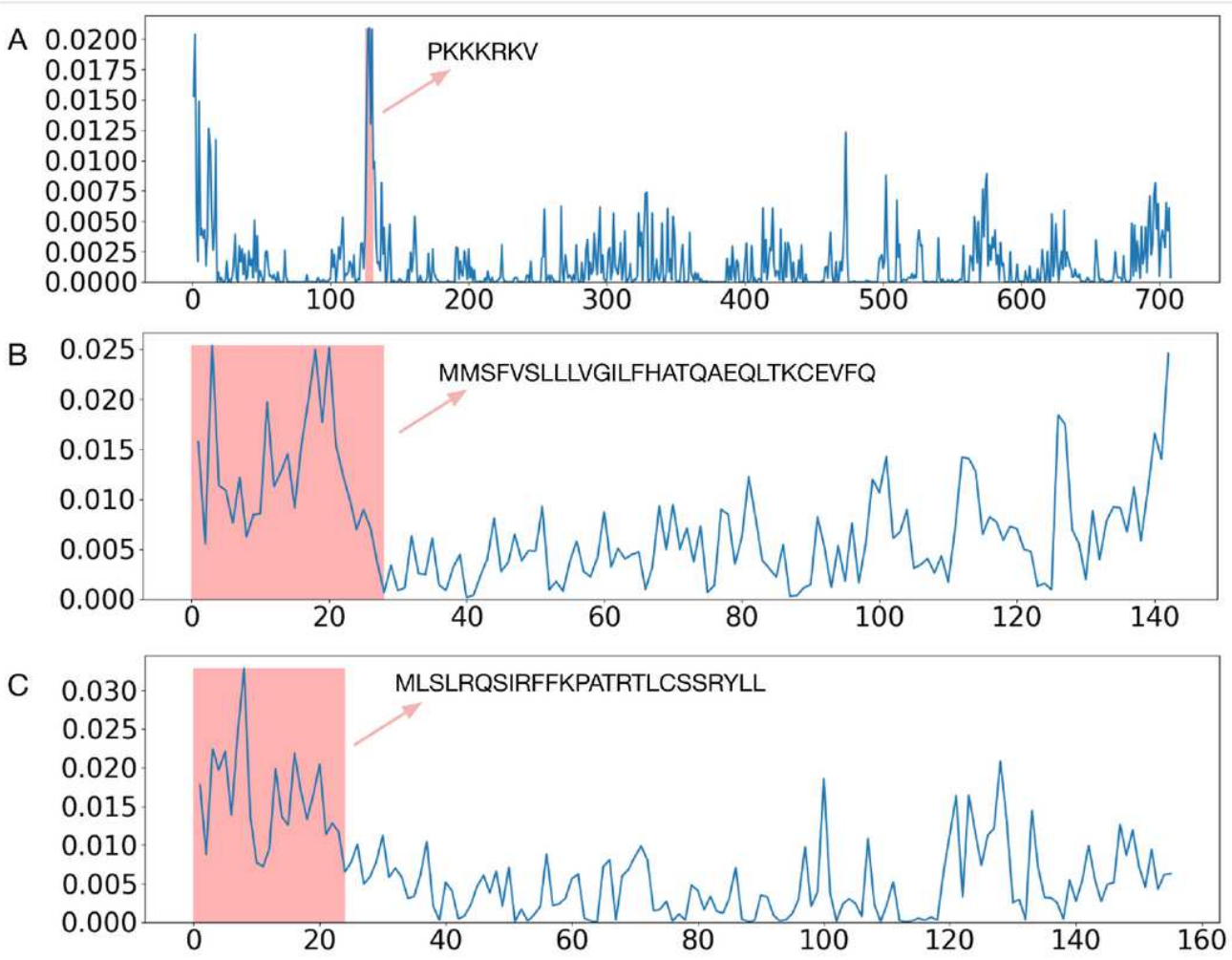


Figure 2

The Visualization of attention weights for (A) SV40 large T antigen (P03070), (B) lactalbumin (P09462), and (C) cytochrome oxidase subunit 4 (P04037). The region of the known sorting motif is highlighted in peach and labelled with the sequence of known localization signals.

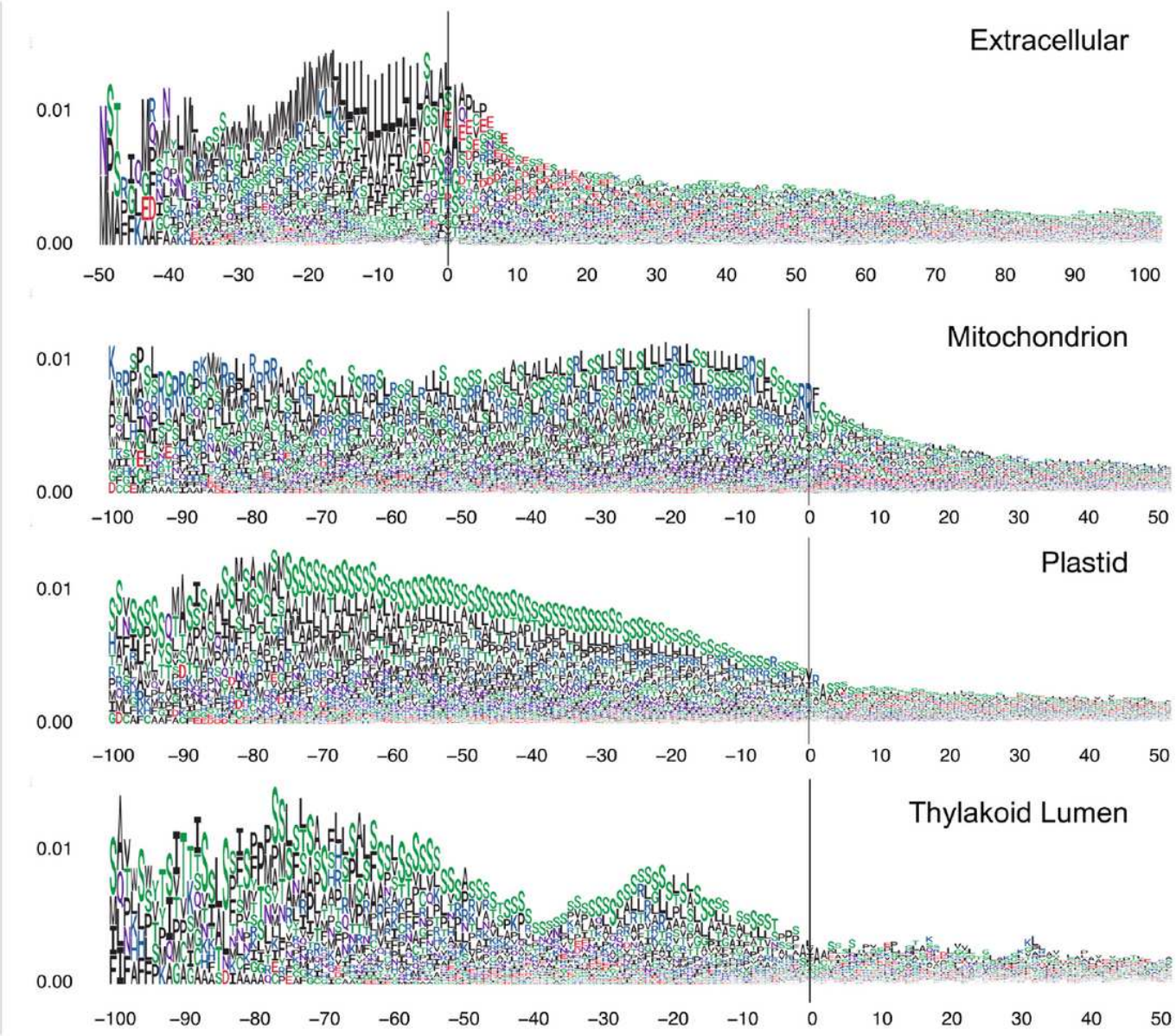


Figure 3

The attention weight visualization by aligned N-terminal sequences at the cleavage site for proteins localized at extracellular, mitochondrion, plastid and thylakoid lumen. The vertical lines indicate the cleavage sites. For extracellular proteins, the range cover 50 AAs before cleavage cite and 100 AAs after the cleavage cite. For the other three classes of proteins, the range covers 100 AAs before the cleavage cite and 50 AAs after the cleavage cite.

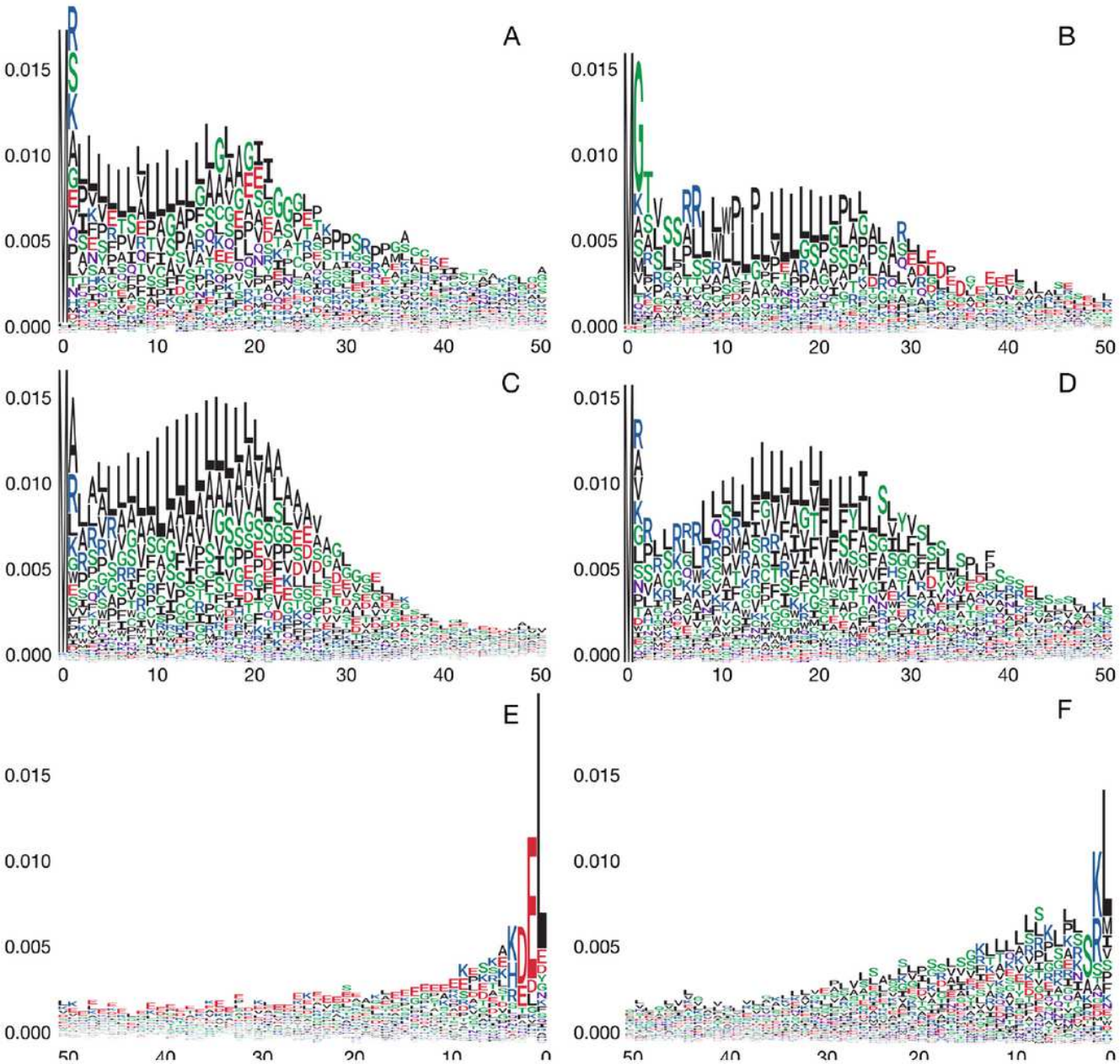


Figure 4

Attention weights of 50 amino acids near the termini at the sub-organellar level, which suggests potential localization signals. (A) N-terminus of cytoplasmic granule. (B) N-terminus of cell surface. (C) N-terminus of endoplasmic reticulum lumen. (D) N-terminus of Golgi apparatus, Golgi stack membrane. (E) C-terminus of endoplasmic reticulum lumen. (F) C-terminus of peroxisome proteins.

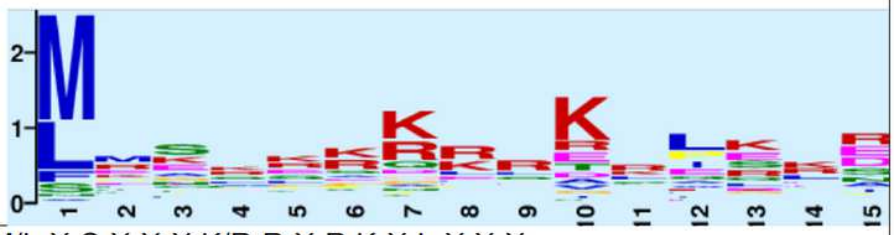
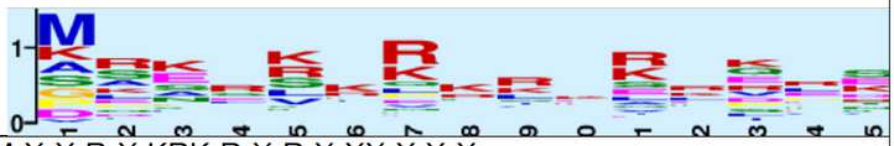
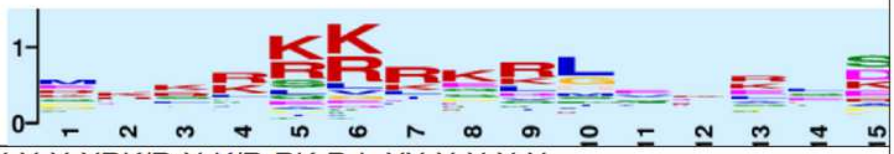
Localization	Rank	Motif	
Nucleus	1 (4061)	Logo	
		RE.	M/L-X-S-X-X-X-K/R-R-X-R-K-X-L-X-X-X
	2 (3680)	Logo	
		RE.	M-X-X-R-X-KRK-R-X-R-X-XX-X-X-X
	3 (2001)	Logo	
		RE.	X-X-X-XRK/R-X-K/R-RK-R-L-XX-X-X-X-X

Figure 5

The top three GLAM2 results for segments from nucleus proteins. For each result, its rank, score, sequence logo and the regular expression (RE) of the motif are given.

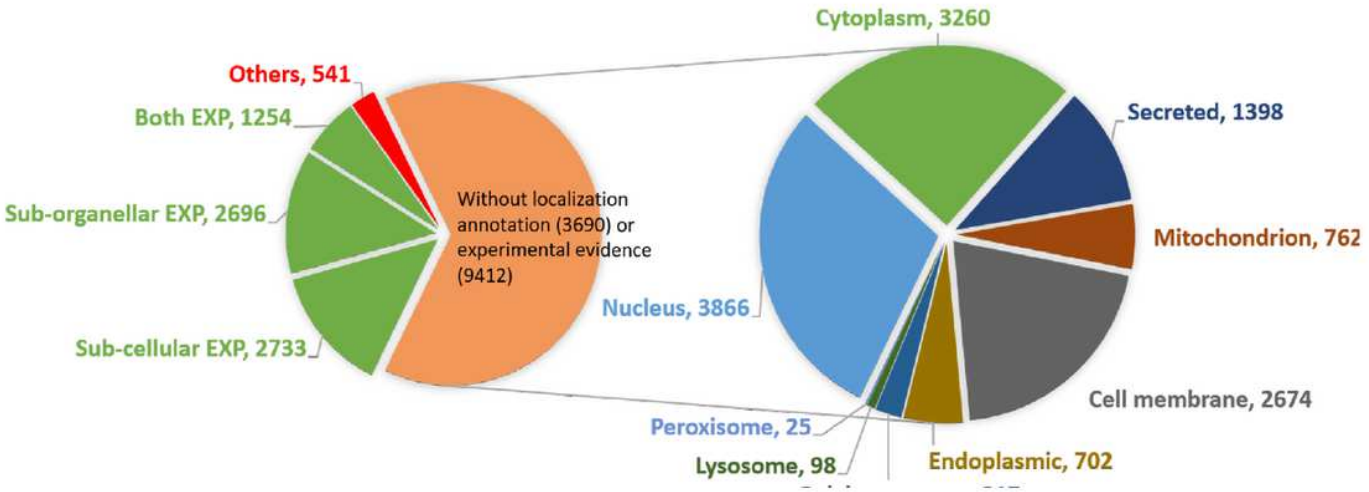


Figure 6

The human proteome localization pie chart. The left pie chart is based on the protein localization annotation, including with or without annotation, and with or without experimental evidence (ECO: 0000269). “Others” include those proteins contain annotations not in any of the 10 major sub-cellular or 45 sub-organellar localization annotations in MULocDeep, or have different localizations for different

protein isoforms. The right pie chart is based on the distribution of prediction for the proteins in the orange part in the left pie chart.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Dongsupplemental782020submit.pdf