

RESEARCH

Open Access



Unified mRNA Subcellular Localization Predictor based on machine learning techniques

Saleh Musleh¹, Muhammad Arif¹, Nehad M. Alajez^{2,3} and Tanvir Alam^{1*}

Abstract

Background The mRNA subcellular localization bears substantial impact in the regulation of gene expression, cellular migration, and adaptation. However, the methods employed for experimental determination of this localization are arduous, time-intensive, and come with a high cost.

Methods In this research article, we tackle the essential challenge of predicting the subcellular location of messenger RNAs (mRNAs) through Unified mRNA Subcellular Localization Predictor (UMSLP), a machine learning (ML) based approach. We embrace an *in silico* strategy that incorporate four distinct feature sets: kmer, pseudo k-tuple nucleotide composition, nucleotide physicochemical attributes, and the 3D sequence depiction achieved via Z-curve transformation for predicting subcellular localization in benchmark dataset across five distinct subcellular locales, encompassing nucleus, cytoplasm, extracellular region (ExR), mitochondria, and endoplasmic reticulum (ER).

Results The proposed ML model UMSLP attains cutting-edge outcomes in predicting mRNA subcellular localization. On independent testing dataset, UMSLP achieved over 87% precision, 94% specificity, and 94% accuracy. Compared to other existing tools, UMSLP outperformed mRNALocator, mRNAloc, and SubLocEP by 11%, 21%, and 32%, respectively on average prediction accuracy for all five locales. SHapley Additive exPlanations analysis highlights the dominance of k-mer features in predicting cytoplasm, nucleus, ER, and ExR localizations, while Z-curve based features play pivotal roles in mitochondria subcellular localization detection.

Availability We have shared datasets, code, Docker API for users in GitHub at: <https://github.com/smusleh/UMSLP>.

Keywords Multiclass classification, mRNA, Subcellular Localization, Machine learning

Introduction

Messenger RNA (mRNA) denotes an RNA molecule characterized by a singular strand, complementary to a corresponding DNA strand within a genome. Throughout transcription, these RNA molecules undergo a series of modifications, encompassing splicing, capping, and polyadenylation. These modifications serve to facilitate their intranuclear mobility and eventual exportation to the cytoplasm, followed by secretion into extracellular domains, as outlined by [1]. A significant milestone in the exploration of mRNA subcellular localization was achieved through the findings of

*Correspondence:

Tanvir Alam

talam@hbku.edu.qa

¹ College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

² Translational Cancer and Immunity Center (TCIC), Qatar Biomedical Research Institute (QBRI), Hamad Bin Khalifa University, Doha, Qatar

³ College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar



Jeffery et al., who revealed asymmetric distribution of mRNA in ascidian embryos and eggs, thus contributing to the advancement of this field [2]. Subsequent inquiries into this phenomenon unveiled the non-random arrangement of mRNAs associated with cytoskeletal proteins within the cytoplasm, offering insights into a plausible mechanism for quantifying their concentration, as investigated by [3]. Over the course of time, researchers have revealed a correlation between mRNA localization and an array of cellular functions, alongside their pivotal regulatory roles within cellular environments, as highlighted by [4]. The spatial allocation of mRNAs also exerts a critical influence on the temporal and spatial control of gene expression, contributing significantly to diverse cellular processes. These encompass cell migration, adaptive cellular responses, maintenance of cellular polarity, orchestration of synaptic plasticity associated with enduring memory, assembly of protein complexes, and the modulation of selective translation, as elucidated in works by [5–8]. Additionally, gaining insight into the factors influencing mRNA localization and the resulting functional outcomes may pave the way for innovative therapeutic interventions aimed at modifying cellular functions through the manipulation of mRNA localization. Figure 1 illustrates

a diagrammatic portrayal depicting the subcellular-level localization of mRNA.

Advancements in experimental techniques have allowed the detection of subcellular localization for numerous RNAs [9]. RNA fluorescent in situ hybridization (RNA-FISH) stands as a dependable experimental method for discerning the localization of mRNA. State-of-the-art technologies like smFISH [10], MERFISH [11] along with its commercially available variant, seqFISH+ [12], as well as GeoMx DSP [13] provide high-resolution images of individual transcripts while delivering both quantitative (RNA copy) and qualitative (subcellular localization) data. However, this method is time-consuming, labor-intensive, and limited to specific tissues [14, 15]. In recent times, advanced high-throughput methodologies like APEX-RIP and CeFra-seq have been introduced to ascertain the subcellular positioning of RNA. However, it's worth noting that the data derived from APEX-RIP, as demonstrated by [16] or CeFra-seq as discussed by [17] might exhibit inherent noise and a potential deficiency in achieving high precision, as indicated by [1]. All currently available experimental methodologies employed to ascertain mRNA localization are characterized by their considerable cost and time requirements. As a result, a burgeoning interest has

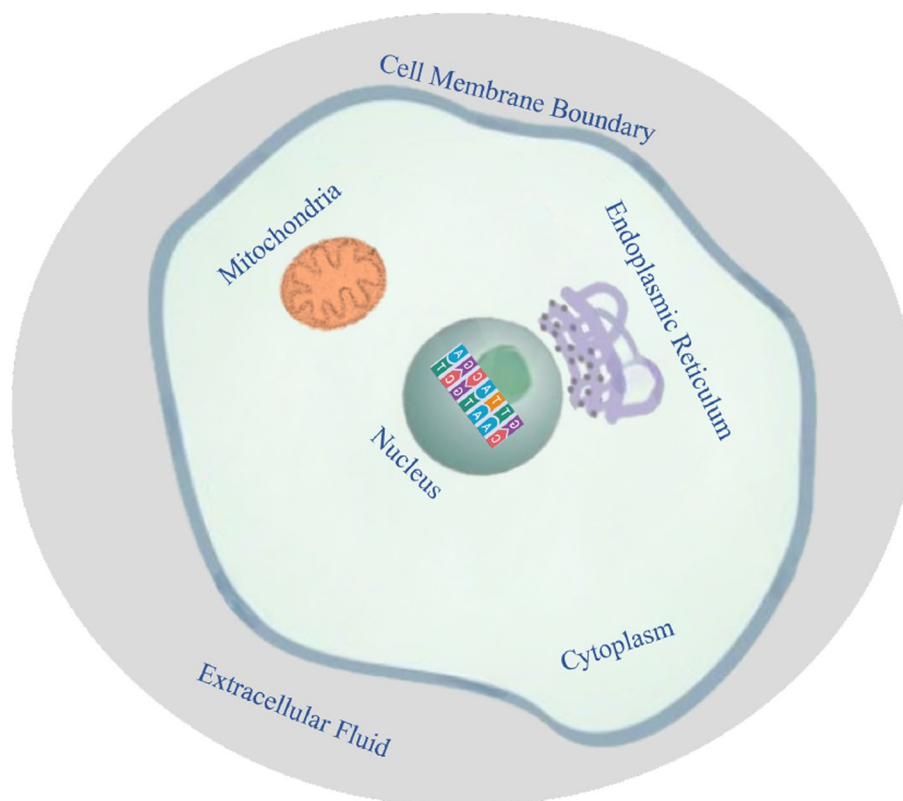


Fig. 1 An animal cell model with five subcellular localization: cytoplasm, nucleus, ER, ExR and mitochondria

emerged within the RNA research community to devise *in silico* approaches grounded in machine learning (ML) models to tackle multiple challenges effectively [18–20] including the efforts aimed at tackling the aforementioned challenges [21].

RNATracker was the very first *in silico* method for predicting mRNA subcellular localization [1]. The authors considered both the mRNA sequence and its corresponding secondary structure as input features for prediction model. The mRNA sequence was represented using a 4-bit one-hot encoding, while the projected secondary structure adopts a 6-bit one-hot encoding scheme. To manage variations in sequence lengths, sequences exceeding 4000 nucleotides were subjected to truncation at the 5' terminal, while sequences falling short were padded with zero values. Subsequently, this embedding was fed into a hybrid architecture comprising a convolutional neural network (CNN) and bidirectional long short-term memory with attention mechanism. The authors evaluated the model on benchmark datasets sourced from APEX-RIP and CeFra-Seq as well though data derived from CeFra-Seq and APEX-RIP might be inherently noisy and potentially lack a high degree of accuracy [1]. In iLoc-mRNA, Zhang et al. used k-mer approach (with k=9) to derive distinctive features from the mRNA sequence to predict its localization [22]. Then, the authors leveraged the ANOVA technique coupled with the binomial distribution to discern and select a refined subset of features from the initial k-mer set. A support vector machine (SVM) employing a radial basis function (RBF) was deployed to undertake the prediction of mRNA subcellular localization. In a recent study, Garg et al. introduced the mRNAloc for predicting mRNA subcellular localization across five distinct locales: nucleus, cytoplasm, extracellular region (ExR), endoplasmic reticulum (ER), and mitochondria [23]. Beginning with input mRNA transcript, the authors meticulously crafted pseudo k-tuple nucleotide composition (PseKNC) features, varying k values across the range of 2 to 5. These generated features were subsequently fed into a SVM model for predicting mRNA subcellular localization. All these methods considered one vs. rest (OvR) approach for the prediction of mRNA localization. But, one of the major drawbacks of OvR approach is to build multiple models which will take more time to find optimized models. Moreover, in terms of model deployment it will take more space and runtime.

Considering the limitation of OvR approach researchers have focused on developing unified multi-class classification model, where only one model is developed for predicting all classes. Based on our literature survey, we found only two recent articles that considered unified approach of multi-class classification of mRNA

subcellular localization prediction. The first work was SubLocEP where Li et al. considered sequence and physicochemical properties of nucleotide to generate features and feed into LightGBM model to predict mRNA subcellular localization [22]. In a five-fold cross-validation (CV) experiment, the authors reported an average accuracy of 65%. And in the independent datasets, the accuracy outcomes spanned a range from 48.68% to 60.10%. In the second research work, Tang et al. introduced the mRNAlocater as a unified multiclass classification model for the same purpose [24]. The authors used PseKNC (with k values ranging from 2 to 6) and PseEIP features. Then the authors considered a fusion of CatBoost, XGBoost and LightGBM model to achieve a better result for mRNA subcellular localization prediction. Recently in some literatures, authors used deep learning based models for the subcellular localization of mRNAs. In DM3Loc [25], authors used CNN with attention mechanism to predict the subcellular localization of mRNA. In RNALight [26], authors used CNN and RNN based networks, but could not outperform k-mer based LightGBM based model. A brief summary of these methods are highlighted in Table 1.

Building upon the prior discussion, in this article we focused on building a Unified mRNA Subcellular Localization Predictor (UMSLP) model with higher accuracy for the most common subcellular localization of mRNAs within cells. The contribution of the present article can be summarized as follows:

1. We proposed a novel combination of feature, encompassing kmer analysis, PseKNC, physicochemical properties of nucleotide and the utilization of 3D sequence representation through Z-curve transformation.
2. The proposed model UMSLP with the selected set of features outperformed the existing methods for the same purpose in almost all evaluation metrics.
3. We have provided a Docker container and an associated API, enabling users to employ our model for the localization prediction of their input sequences. The source code along with the Docker container has been provided to the community in GitHub as well.

Material and methods

Dataset collection

In our experiments, we utilized the identical datasets formulated by Garg et al. [23] as the foundation for training our model in mRNA subcellular localization prediction that was collected from RNALocate v2 database. Only mRNAs which are present in one subcellular localization is considered. mRNAs present in multiple locations and they were discarded

Table 1 A brief overview of the previous works using ML based approach for mRNA subcellular localization prediction

Reference	Year	Subcellular localizaiton	#location	Model	Approach and Features
RNATracker [1]	2019	CeFra-Seq (Cytosol,Nuclear, Membrane, Insoluble); APEX-RIP (Cytosol, Nuclear, ER, Mitochondria)	4	Unified; CNN, BLSTM, Attention mechanism	One hot encoding of sequence
iLoc-mRNA [22]	2021	Four customized locations by authors : C1, C2, C3, C4 covering Cytosol, Cytoplasm, Ribosome, ER, Nucleus, Exosome, Mitochondria, Dendrite	4	SVM	OvR; k-mer
mRNALoc [23]	2020	Cytoplasm, Nucleus, ER, ExR, Mitochondria	5	SVM	OvR; Pse-KNC
SubLocEP [27]	2021	Cytoplasm, Nucleus, ER, ExR, Mitochondria	5	LGBM	Unified; k-mer, PseKNC, physicochemical properties (PseEIIIP)
mRNALoacter [24]	2021	Cytoplasm, Nucleus, ER, ExR, Mitochondria	5	LGBM, XGBoost, CatBoost	Unified; PseKNC, physicochemical properties (PseEIIIP)
MSLP [28]	2022	Cytoplasm, Nucleus, ER, ExR, Mitochondria, Cytosol, Pseudopodium, Posterior, Ribosome, Exosome	10	CatBoost	OvR; k-mer, PseKNC, physicochemical properties PseEIIIP, DPCP, TPCP, Z-curve
DM3Loc [25]	2021	Cytosol, Nucleus, ER, Exosome, Ribosome, Membrane	6	CNN with multi-head self-attention	Sequence only
RNALight [26]	2023	Cytoplasm, Nucleus	2	LightGBM	k-mer

from the analysis as suggested in prior works [24, 28]. To mitigate potential homology bias, the sequences were subjected to clustering using the BLASTClust program from the standalone BLAST package, utilizing the parameters S40 and -L 0.7, as detailed in [23]. As mentioned in [23], 1972 mRNAs having more than one subcellular localizations were dropped from our analysis. The resultant benchmark dataset comprised a total of 14,909 sequences, with distribution as follows: 6,376 mRNAs localized in the cytoplasm, 1,426 in the endoplasmic reticulum (ER), 885 in the extracellular region (ExR), 421 in mitochondria, and 5,831 in

the nucleus. Following the approach outlined in [23], our training set for ML model comprised five-sixths of the sequences allocated for each location, while the remaining one-sixth was reserved for independent testing to assess the model’s performance. For an overview of the datasets, we refer to Fig. 2 and Table 2.

Feature extraction

Kmer related features

We analyzed each mRNA by the its corresponding DNA sequence calculating the occurrences of consecutive nucleotides of different lengths (k values of 1,

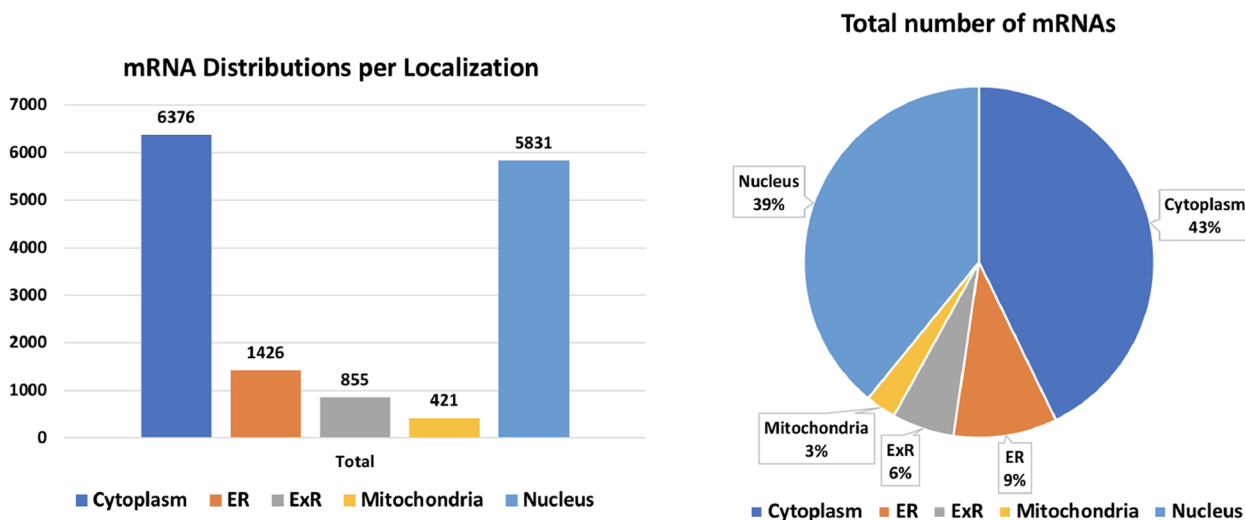


Fig. 2 Counts and distributions of mRNAs in different subcellular localizations. We can observe that cytoplasm and nucleus covers more than 82% of the mRNAs in the dataset

Table 2 Number of sequences per location in training and testing dataset

Location	Training Dataset $\frac{5}{6}$	Testing Dataset $\frac{1}{6}$	Total
Cytoplasm	5,310	1,066	6,376
ER	1,185	241	1,426
ExR	710	145	855
Mitochondria	350	71	421
Nucleus	4,855	976	5,831

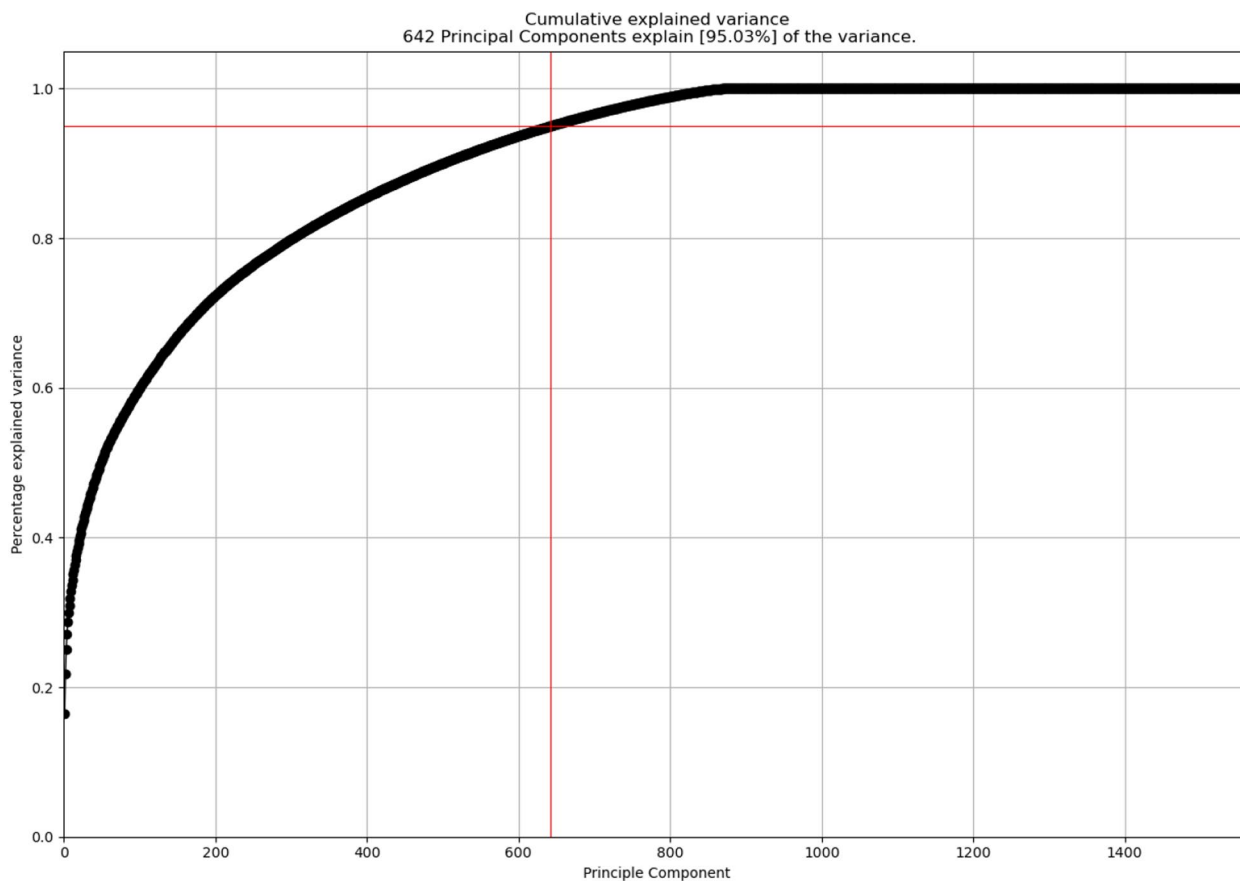
2, 3, 4, and 5) throughout the entire transcript. These counts were then normalized based on the length of the sequence and adjusted for the possible combinations of nucleotides of a given length. As a result, we obtained a 1360-dimensional feature vector (with dimensions 16, 64, 256, and 1024 originating from di-nucleotides, tri-nucleotides, quad-nucleotides, and penta-nucleotides, respectively) to represent each input sequence. This feature vector was subsequently utilized as input for our ML models.

$$Kmer_i = \frac{N_i}{L}, k = 2, 3, 4, \text{ and } 5 \quad (1)$$

From every sequence, a feature vector was generated, possessing a size of 1360, with N_i denoting the tally of k -mers within the transcript and L signifying the length of the mRNA transcript.

Pseudo k -tuple nucleotide composition (PseKNC)

The pseudo k -tuple nucleotide composition (PseKNC) is designed to capture the impact of nucleotide sequence arrangement within DNAs, reflecting their implications, as discussed by [29] and expounded upon by [30]. This preservation of sequence order is achieved through the utilization of physicochemical properties inherent to the constituent oligonucleotides. The dimension of the resultant feature vector is denoted as $(4^k + \lambda)$, where the positive integer k signifies the k -mer's highest correlation rank observed along a DNA sequence. In our specific scenario, k values of 2, 3, 4, and 5 were implemented, accompanied by a value of

**Fig. 3** PCA Analysis of the features. 642 principal components cover over 95% variance of the dataset

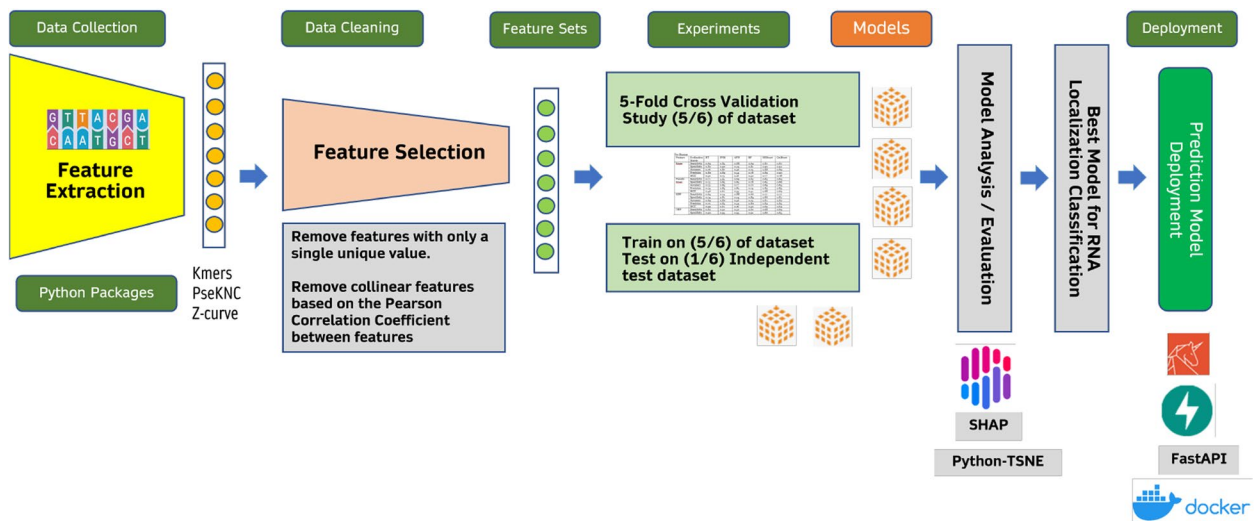


Fig. 4 UMSLP Computational Pipeline. First, benchmark dataset was collected. Then data cleaning and pre-processing was done to remove noisy data points. Feature engineering was utilized to generate relevant features and building ML model. SHAP based analysis was involved to explain the model. Finally, docker based solution is provided for the community

Table 3 Results from Ablation Study using CatBoost model for CV

Feature	CatBoost	Precision	Recall	Acc	Specificity	F1 Score
k-mer	Cytoplasm	0.74	0.88	0.82	0.77	0.80
	ER	0.94	0.38	0.94	1.00	0.54
	ExR	0.84	0.20	0.95	1.00	0.32
	Mitochondria	0.96	0.85	0.99	1.00	0.90
	Nucleus	0.81	0.87	0.87	0.87	0.84
	avg	0.86	0.63	0.91	0.93	0.68
Pse-KNC	Cytoplasm	0.64	0.77	0.72	0.68	0.70
	ER	0.72	0.13	0.91	0.99	0.22
	ExR	0.53	0.07	0.94	1.00	0.13
	Mitochondria	0.83	0.63	0.99	1.00	0.71
	Nucleus	0.67	0.75	0.76	0.76	0.71
	avg	0.68	0.47	0.86	0.89	0.49
Z-curve 144	Cytoplasm	0.68	0.83	0.76	0.70	0.74
	ER	0.87	0.21	0.92	1.00	0.34
	ExR	0.77	0.05	0.94	1.00	0.09
	Mitochondria	0.96	0.92	1.00	1.00	0.94
	Nucleus	0.75	0.81	0.82	0.83	0.78
	avg	0.81	0.57	0.89	0.91	0.58
Z-curve 44	Cytoplasm	0.72	0.84	0.79	0.75	0.78
	ER	0.81	0.33	0.93	0.99	0.47
	ExR	0.80	0.11	0.95	1.00	0.19
	Mitochondria	0.94	0.87	0.99	1.00	0.90
	Nucleus	0.77	0.84	0.84	0.84	0.81
	avg	0.81	0.60	0.90	0.92	0.63

($\lambda = 10$). This configuration yielded the creation of 16, 64, 256, 1024, and 10 features, respectively. These features were amalgamated to forge a feature vector of 1370 dimensions for each mRNA DNA sequence.

Z curve representation for trinucleotide frequencies

The Z-curve theory is a geometric approach utilized for visualizing genome sequences within three-dimensional space, as introduced by [31] and further elaborated by [32]. By applying the Z-transform technique, the frequencies of nucleotides A, C, G, and T, as well as their various combinations (kmer), present within the sequence or open reading frame, undergo a transformation into a three-dimensional space, as originally proposed by [33]. In our feature engineering process, we embraced the subsequent representation of the Z-curve:

1. Phase-independent frequency: This encoding of Z-curve represented by a 48-bit descriptor as follows:

$$\begin{cases} x_{LM} = [p(LMA) + p(LMG)] - [p(LMC) + p(LMT)] \\ y_{LM} = [p(LMA) + p(LMC)] - [p(LMG) + p(LMT)] \\ z_{LM} = [p(LMA) + p(LMT)] - [p(LMC) + p(LMG)] \end{cases} \quad (2)$$

where the normalized frequency of trinucleotides JLM , LMC , LMG , LMT are represented by $p(LMA)$, $p(LMC)$, $p(LMG)$, $p(LMT)$ respectively. The dimension of the feature matrix is 48.

2. Phase-specific frequency: This is succinctly represented utilizing Z-curve parameters, encapsulated within a 144-bit descriptor as outlined below:

$$\begin{cases} x_{LM}^k = [p^k(LMA) + p^k(LMG)] - [p^k(LMC) + p^k(LMT)] \\ y_{LM}^k = [p^k(LMA) + p^k(LMC)] - [p^k(LMG) + p^k(LMT)] \\ z_{LM}^k = [p^k(LMA) + p^k(LMT)] - [p^k(LMC) + p^k(LMG)] \end{cases} \quad (3)$$

where k represents the three position of nucleotides at potential codons. The normalized frequency of trinucleotides LMA , LMC , LMG , LMT at different positions were represented by $p^k(LMA)$, $p^k(LMC)$, $p^k(LMG)$, $p^k(LMT)$ respectively generating a feature vector of dimension 144.

Physicochemical properties of mRNA genes

To capture the physicochemical characteristics of nucleotides, three types of features were employed: (1) Pse-EIIP: Pseudo Electron-ion interaction pseudopotentials of trinucleotide, (2) DPCP: Dinucleotide physicochemical properties, and (3) TPCP: Trinucleotide physicochemical properties. To derive these features from the mRNA DNA sequence, the iLearnPlus [34] tool was utilized. These generated a feature vector of size 3200. The details can be found in [28].

Data cleansing and normalization

Initially, our dataset comprised 6122 features, including 1360 from Kmer, 1370 from PseKNC, 64 from PseEIIP, 2368 from DPCP, 768 from TPCP, 48 from Z-Curve 48-bit, and 144 from Z-Curve 144-bit. To reduce dimensionality, we considered two different approaches. In the first approach, we applied principal component analysis (PCA) on the dataset and considered only Principal components covering 95% variability of the dataset (Fig. 3).

In the second approach, we eliminated feature columns with a single static value and it dropped 70

Table 4 ML model results based on PCA

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
Cytoplasm	0.73	0.87	0.81	0.76	0.79
ER	0.94	0.36	0.94	1.00	0.52
ExR	0.78	0.16	0.95	1.00	0.27
Mitochondria	0.98	0.86	1.00	1.00	0.91
Nucleus	0.80	0.85	0.86	0.86	0.82
Avg	0.84	0.62	0.91	0.92	0.66
XGBoost					
Cytoplasm	0.73	0.87	0.81	0.76	0.80
Er	0.91	0.38	0.94	1.00	0.54
ExR	0.78	0.16	0.95	1.00	0.26
Mitochondria	0.95	0.85	0.99	1.00	0.90
Nucleus	0.81	0.85	0.86	0.87	0.83
Avg	0.83	0.62	0.91	0.92	0.66
Decision Tree					
Cytoplasm	0.66	0.65	0.71	0.75	0.65
ER	0.31	0.35	0.86	0.92	0.33
ExR	0.16	0.16	0.91	0.95	0.16
Mitochondria	0.75	0.64	0.98	0.99	0.69
Nucleus	0.68	0.69	0.75	0.80	0.68
Avg	0.51	0.50	0.84	0.88	0.50
GNB					
Cytoplasm	0.63	0.20	0.61	0.91	0.30
Er	0.18	0.55	0.71	0.73	0.27
ExR	0.16	0.34	0.86	0.90	0.22
Mitochondria	0.42	0.95	0.96	0.96	0.58
Nucleus	0.59	0.58	0.68	0.74	0.59
Avg	0.40	0.52	0.76	0.85	0.39
MLP					
Cytoplasm	0.76	0.79	0.80	0.81	0.77
ER	0.62	0.54	0.92	0.97	0.58
ExR	0.35	0.18	0.93	0.98	0.24
Mitochondria	0.93	0.81	0.99	1.00	0.87
Nucleus	0.78	0.83	0.84	0.85	0.81
Avg	0.69	0.63	0.90	0.92	0.65

features. Then we drop features with collinearity above 98% and it dropped 4487 features. Consequently, this preprocessing phase yielded a dataset with 1565 features (Supplementary File S1). Then we normalized these features using z-score normalization technique. Normalization step is essential in many ML algorithms, guaranteeing a consistent scale across features. It mitigates the risk of certain features overshadowing others during a model's training. The second approach provided us the better results compared to the PCA based dimension reduction and therefore, we considered the second approach as part of the proposed final workflow (see the Results section).

Development of classification models

Within the realm of ML, classification entails the task of determining the category to which a new instance belongs from a predefined set of classes. In a multi-class classification problem, a sample might be classified into more than two classes. To conduct our investigation, we integrated a range of ML training algorithms, specifically including CatBoost, XGBoost, Decision Tree, Gaussian Naive Bayes (GNB) and Multi-layer Perceptron (MLP) classifier. Moreover, we used

StackingClassifier to fuse the output of two models : CatBoost, XGBoost and use a classifier to compute the final prediction of model. We fine-tuned the hyperparameters of these models based on GridSearchCV hyperparameter method using Scikit-learn. Then we selected the most optimal parameters among the designated hyperparameter choices.

Performance evaluation of the models

Assessing the effectiveness of a ML model is a crucial aspect in its development process. For the assessment of our model's performance, we utilized a method known as k-fold cross-validation (CV). This technique involves dividing the dataset into "k" equal-sized subsets. In our study, we adopted Five-fold cross-validation (CV), which means the data is partitioned into five distinct segments or folds, each containing comparable data. During each iteration, four folds were utilized for training purposes, while one fold was held back for testing. Cross-validation offers valuable insights into the model's ability to perform on new and unseen data, allowing for an assessment of its generalization capabilities. Moreover, it facilitates the generation of more precise estimation of the model's predictive performance, leading to a

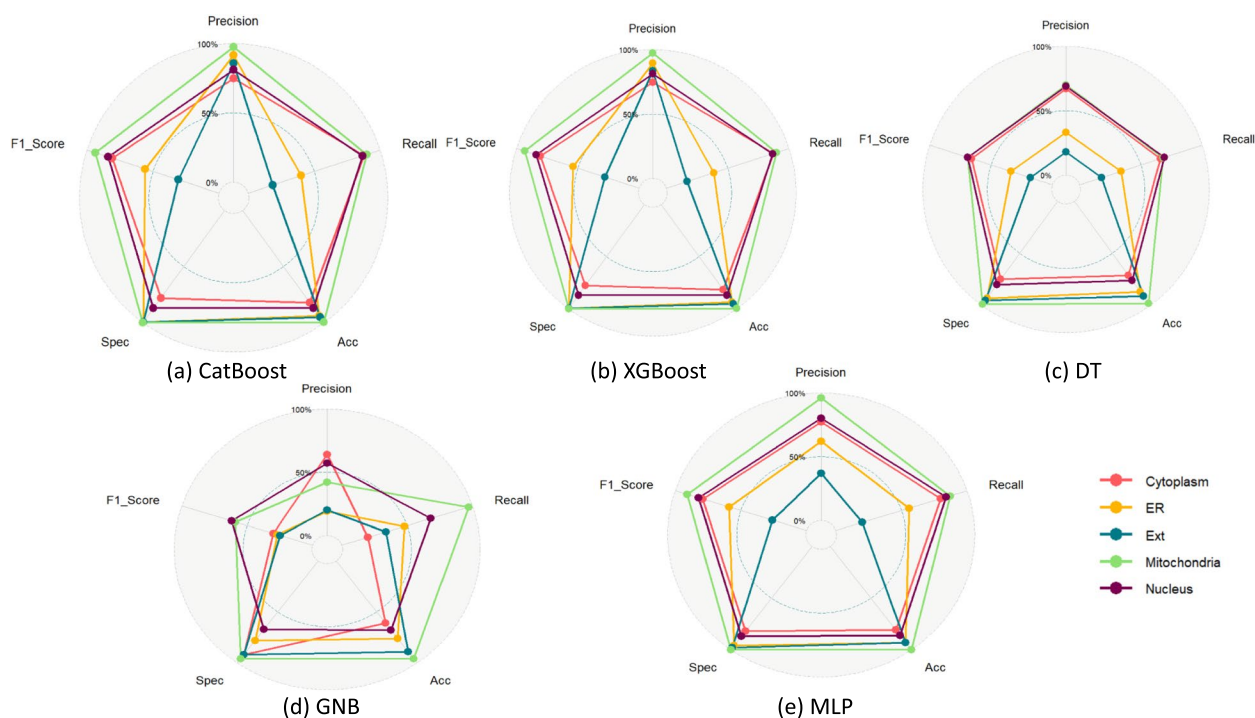


Fig. 5 UMSLP Models Performance per localization for different ML models using five fold cross validation. Each plot shows different evaluation metrics for a model highlighting per class result as well as average across all classes

heightened accuracy when evaluating its effectiveness. In this approach, 80% of the data was dedicated to training the model, while the remaining 20% was reserved for validating its performance. We have generated the confusion matrix structure to enable various evaluation metrics to gauge the quality and effectiveness of the model, and provide insights into how well it performed with the given dataset. The following metrics were utilized for this evaluation:

$$Accuracy(Acc) = \frac{TP + TN}{TP + FN + FP + TN} \quad (4)$$

$$Specificity(Sp) = \frac{TN}{FP + TN} \quad (5)$$

$$Sensitivity(Sn) = \frac{TP}{TP + FN} \quad (6)$$

$$Precision(Pr) = \frac{TP}{TP + FP} \quad (7)$$

$$F1 - Score(F1) = \frac{2 * Precision * Recall}{Precision + recall} \quad (8)$$

In the context of our analysis, a true positive (*TP*) signifies a correct prediction that aligns with the actual truth. Similarly, a true negative (*TN*) indicates a correct rejection of a prediction that corresponds with the true absence. On the other hand, a false positive (*FP*) occurs when predictions are deemed true but do not align with reality. Lastly, a false negative (*FN*) pertains to predictions that are considered false despite being true in actuality.

Explanation of algorithms

The interpretability and explainability of ML models has gained significant attention in recent times. Users not only value the performance of the models but also seek insights into the underlying classification process. While certain models, such as Decision Trees, inherently offer transparency, others, like Neural Networks, often operate as black-box models. We employed the SHapley Additive exPlanations (SHAP) technique to interpret the models developed in this study. This method was introduced by Lundberg et al. in 2020 [35] that enables the interpretation of the output of ML models. By leveraging the conventional Shapley values derived from game theory and their associated extensions, SHAP establishes a connection between optimal credit distribution and localized explanations.

Model deployment

The overall computational workflow for the data processing, model development and model deployment is depicted in Fig. 4. Once the model is developed we used

Table 5 UMSLP 5 Fold CV Results Per Model - All Locations

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
Cytoplasm	0.75	0.88	0.82	0.78	0.81
ER	0.92	0.40	0.94	1.00	0.56
ExR	0.86	0.19	0.95	1.00	0.31
Mitochondria	0.98	0.90	1.00	1.00	0.93
Nucleus	0.81	0.87	0.87	0.87	0.84
Avg	0.86	0.65	0.92	0.93	0.69
XGBoost					
Cytoplasm	0.74	0.88	0.82	0.78	0.81
ER	0.89	0.39	0.94	1.00	0.54
ExR	0.83	0.17	0.95	1.00	0.28
Mitochondria	0.97	0.90	1.00	1.00	0.93
Nucleus	0.81	0.87	0.87	0.87	0.84
Avg	0.85	0.64	0.91	0.93	0.68
Decision Tree					
Cytoplasm	0.67	0.66	0.71	0.75	0.66
ER	0.33	0.34	0.87	0.93	0.34
ExR	0.18	0.18	0.91	0.95	0.18
Mitochondria	0.70	0.68	0.98	0.99	0.69
Nucleus	0.69	0.69	0.76	0.80	0.69
Avg	0.51	0.51	0.85	0.88	0.51
GNB					
Cytoplasm	0.64	0.20	0.61	0.92	0.30
ER	0.19	0.48	0.76	0.78	0.27
ExR	0.20	0.34	0.89	0.92	0.25
Mitochondria	0.42	0.97	0.96	0.96	0.59
Nucleus	0.57	0.68	0.68	0.67	0.62
Avg	0.41	0.53	0.78	0.85	0.41
MLP					
Cytoplasm	0.77	0.80	0.81	0.82	0.79
ER	0.62	0.56	0.93	0.96	0.59
ExR	0.37	0.20	0.93	0.98	0.26
Mitochondria	0.96	0.87	1.00	1.00	0.91
Nucleus	0.80	0.84	0.86	0.87	0.82
Avg	0.71	0.65	0.91	0.93	0.67
StackEnsemble					
Cytoplasm	0.79	0.86	0.84	0.83	0.82
ER	0.75	0.61	0.94	0.98	0.67
ExR	0.56	0.25	0.95	0.99	0.35
Mitochondria	0.92	0.94	1.00	1.00	0.93
Nucleus	0.83	0.85	0.87	0.89	0.84
Avg	0.77	0.70	0.92	0.94	0.72

Docker environment to setup the model and deploy it in Docker environment. Users can download the Docker from GitHub and use the proposed model as a tool for mRNA subcellular localization prediction. Details of the Docker is provided in the GitHub: <https://github.com/smusleh/UMSLP>.

Motif analysis for mRNA sequence from different subcellular localizations

For the motif analysis of each mRNA subcellular localization, we used STREME [36] to discover ungapped small motifs that are enriched in a location compared to the background sequence. We used the default setup in

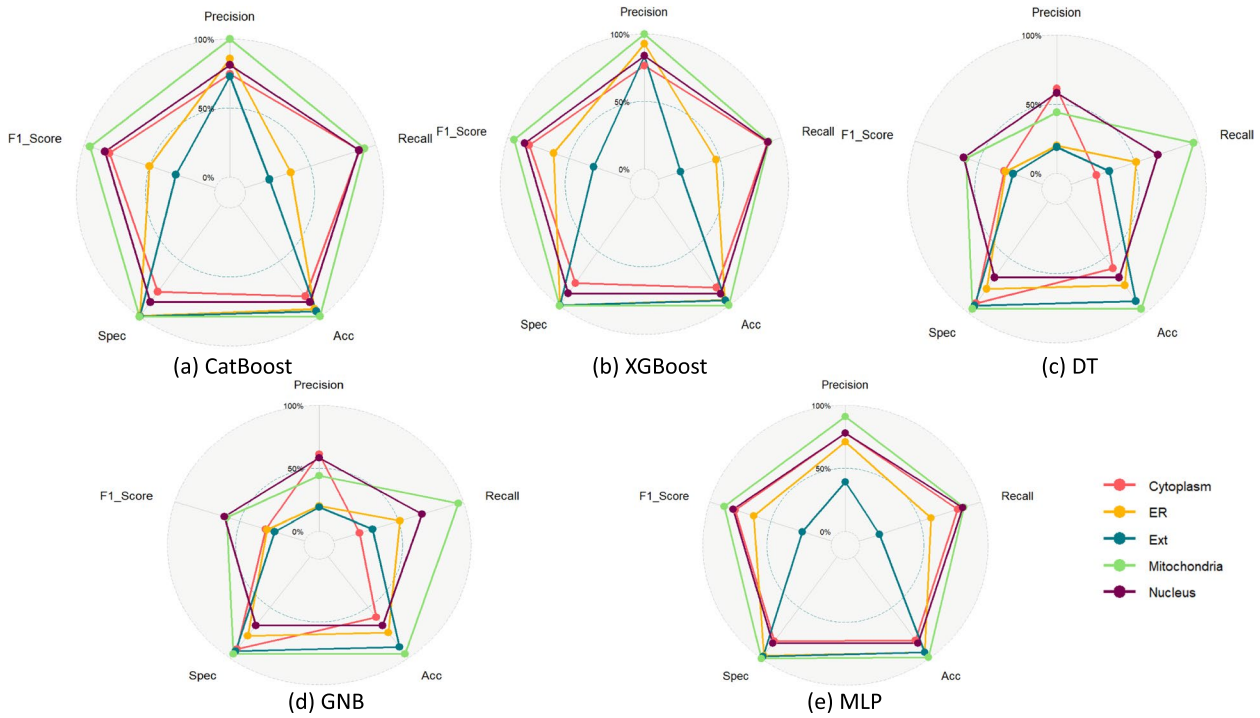


Fig. 6 UMSLP models performance for localizations on independent test set. Each plot shows different evaluation metrics for a model highlighting per class result as well as average across all classes

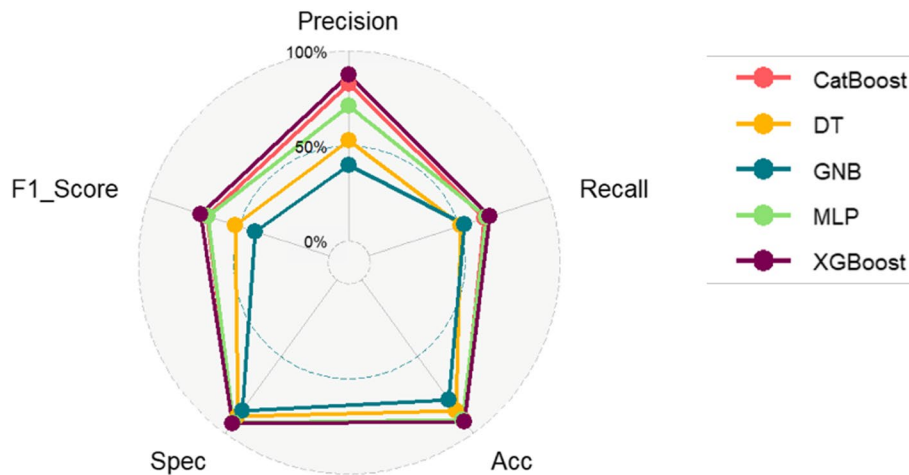


Fig. 7 Model performance across all classes (locations) based on independent set results

STREME with motif length 8 to 15 with a p-value threshold 0.05.

Results

Post data cleaning and selection step, we conducted two separate experiments: we employed a five-fold cross-validation (CV) technique. Then we followed a conventional train-and-test approach where we use $\frac{5}{6}$ of the dataset for training and the remaining $\frac{1}{6}$ of the dataset as independent test set.

Ablation study on the selected features

We first analyzed individual types of features as input to the models for evaluating their effectiveness. The ablation study on different types of features would help us to understand the relative contribution of features in making the final model. The results of ablation study with CatBoost model are highlighted in Table 3. From Table 3, we can observe that, k-mer based features had the most distinguishing power compared to the other types of features. Both Z-curve with 44 bits and 144bits achieved similar performance at the average accuracy of 91% and 92%, respectively. Among all the features, PseKNC was shown to be the least performing with an average accuracy of 89%. Supplementary File 02 highlights the results of ablation study using other models.

Performance of the models using PCA based dimension reduction technique

We first applied PCA to reduced the dimensionality of the input features. Then we selected PC covering 95% variability of the dataset. Then we feed those PC into models to check the performance of the models (Table 4). Based on the ML model performance, we can observe that both XGBoost and CatBoost model performed at similar level and they can be considered as the best performing models. Other models i.e., DT, MLP, GNB were not close them in terms of evaluation metrics.

Performance of the models in cross validation

We used CatBoost, XGBoost and Decision Tree models, MLP and GNB for the prediction task. For dimension reduction, we finally considered the collinearity based dimension reduction as mentioned in the [Material and methods](#) section. A radar diagram in Fig. 5 show and visualize performance metrics of the three models.

The CV results for each model are shown in Table 5. We have also calculated the mean values of all performance indicators including precision, recall, accuracy, specificity and F1 score. As shown in the Table 5, the

average score for all performance metrics, the CV results is using the CatBoost and XGBoost models are very close with CatBoost being the best one. We implemented the StackEnsemble technique using boosting algorithms to

Table 6 UMSLP Results on Independent Dataset - All Locations

Model/Class	Pr	Sn	Acc	Sp	F1
CatBoost					
Cytoplasm	0.75	0.87	0.82	0.78	0.80
ER	0.86	0.35	0.93	0.99	0.50
ExR	0.73	0.19	0.95	1.00	0.30
Mitochondria	1.00	0.91	1.00	1.00	0.96
Nucleus	0.81	0.87	0.87	0.87	0.84
Avg	0.83	0.64	0.91	0.93	0.68
XGBoost					
Cytoplasm	0.76	0.89	0.84	0.80	0.82
ER	0.93	0.47	0.95	1.00	0.62
ExR	0.84	0.18	0.95	1.00	0.30
Mitochondria	1.00	0.90	1.00	1.00	0.95
Nucleus	0.84	0.89	0.89	0.89	0.86
Avg	0.87	0.67	0.92	0.94	0.71
DT					
Cytoplasm	0.69	0.68	0.73	0.77	0.68
ER	0.38	0.43	0.88	0.92	0.40
ExR	0.15	0.12	0.91	0.96	0.13
Mitochondria	0.75	0.63	0.98	0.99	0.68
Nucleus	0.70	0.72	0.77	0.80	0.71
Avg	0.53	0.51	0.85	0.89	0.52
GNB					
Cytoplasm	0.61	0.20	0.60	0.91	0.30
ER	0.20	0.51	0.75	0.78	0.29
ExR	0.19	0.30	0.89	0.93	0.23
Mitochondria	0.44	0.96	0.96	0.96	0.60
Nucleus	0.58	0.68	0.68	0.68	0.62
Avg	0.40	0.53	0.78	0.85	0.41
MLP					
Cytoplasm	0.78	0.81	0.82	0.83	0.79
ER	0.71	0.59	0.94	0.97	0.64
ExR	0.39	0.17	0.94	0.98	0.24
Mitochondria	0.91	0.86	0.99	1.00	0.88
Nucleus	0.78	0.85	0.85	0.85	0.81
Avg	0.71	0.65	0.91	0.93	0.67
StackEnsemble					
Cytoplasm	0.78	0.84	0.83	0.82	0.81
ER	0.75	0.63	0.94	0.98	0.68
ExR	0.63	0.26	0.95	0.99	0.37
Mitochondria	0.92	0.94	1.00	1.00	0.93
Nucleus	0.82	0.85	0.87	0.88	0.84
Avg	0.78	0.70	0.91	0.93	0.72

assess whether the fusion would enhance the outcomes. However, the combination of XGBoost and CatBoost within the ensemble failed to surpass the performance of the individual models (Table 5).

Based on the results on independent set, we selected XGBoost as the final model for UMSLP.

Tables 6 show the in detail report of the models covering all evaluation metrics on independent dataset.

Performance of the models in Independent Dataset

As shown in the figure the average score for all performance indicators, the independent test results is using the CatBoost and XGBoost models are very close with XGBoost being the best one (Fig. 6).

Figure 7 highlights the average score of all evaluation metrics for the models we tested on independent set.

Motifs discovered from mRNA sequence from different subcellular localization

Figure 8 highlights the top two motifs identified from mRNA sequences coming from different subcellular localizations. We can observe that the top two motifs from nucleus and ExR are very similar- mainly having long sequences of As (or Ts in reverse complement) and

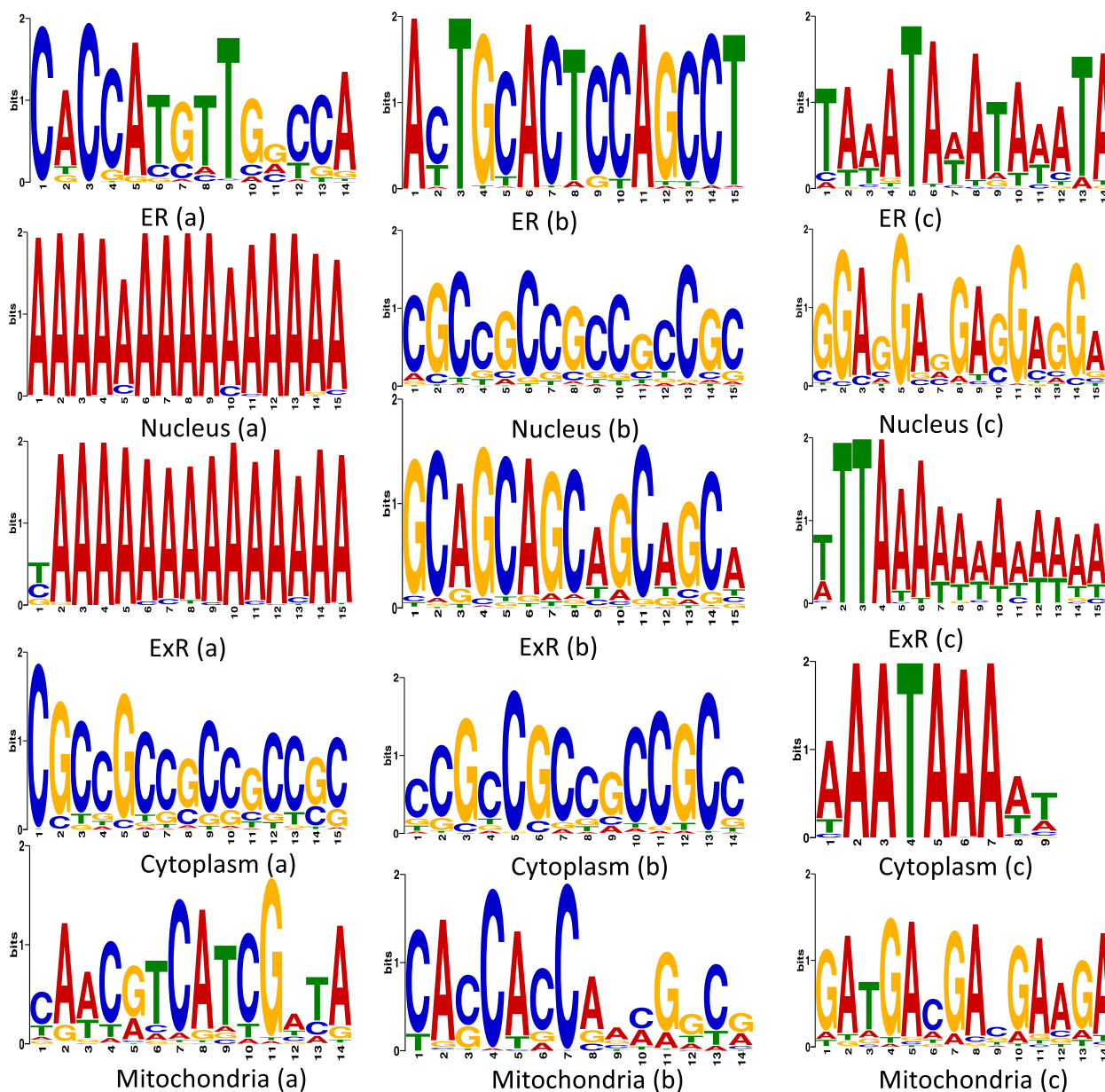


Fig. 8 Top ranked motifs identified from mRNA sequences coming from different subcellular localizations

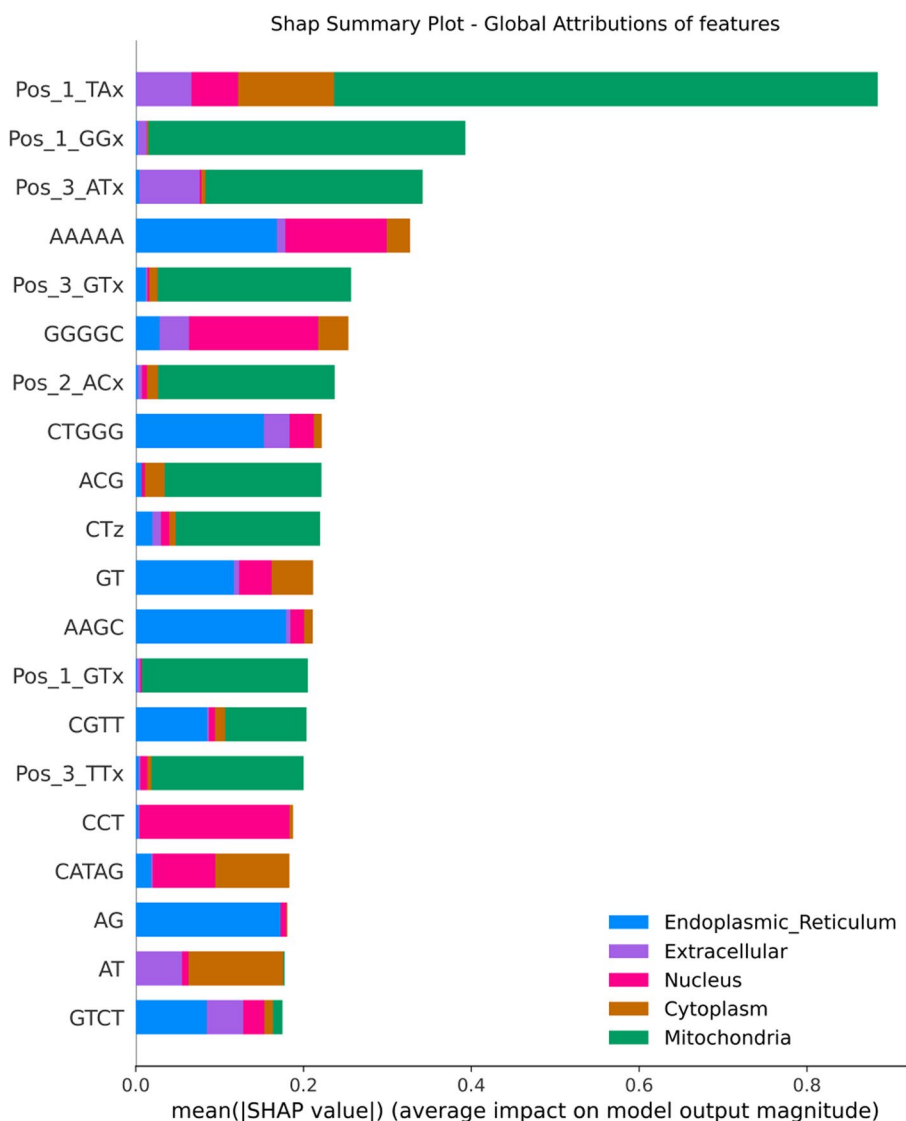


Fig. 9 SHAP Summary Plot - Global Feature Importance

CG dominant regions. This also hints a plausible explanation of low recall value of models for the prediction of ExR (Table 5, Table 6). As the underlying sequence pattern in nucleus and ExR are very similar, and the nucleus class has many more mRNA sequence compared to ExR (6376 vs 855) then the model tries to predict them into nucleus over ExR. The distribution of the motif location over the sequences are prevalent in 3' UTR as well as in 5' UTR regions and the coding region (Supplementary File 03) and this aligns perfectly to the findings mentioned in [22]. Supplementary File 03 summarizes top motifs from all locations with their statistical details and positional distribution.

Towards explainability of the proposed model and features SHAP values provide a quantified measure of both the extent and direction (whether positive or negative) of a feature's influence on a prediction. For visualizing the global attributions of features across all localizations we used SHAP Waterfall plot (Fig. 9). For the impact of top ranked features in each subcellular localization we used Beeswarm plots (Figs. 10, 11, and 12).

As illustrated in Fig. 9, Z-curve features of both 144-bit (Pos_*), 48-bit (CTz), and Kmer (K=3), nine features that have the most predictive power for the Mitochondria localization (green dominant area) followed by the other localization. Looking into the details of these

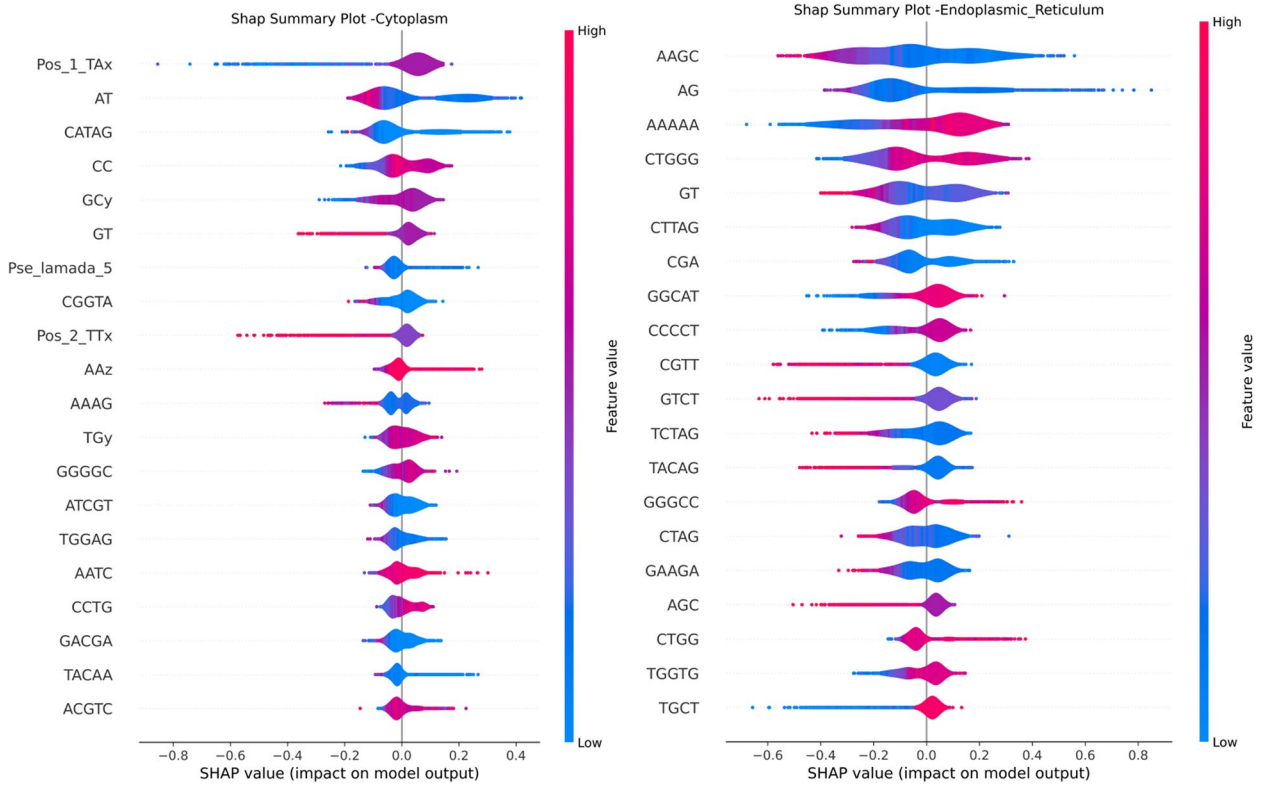


Fig. 10 SHAP Summary Plot - Cytoplasm - ER - Feature Importance

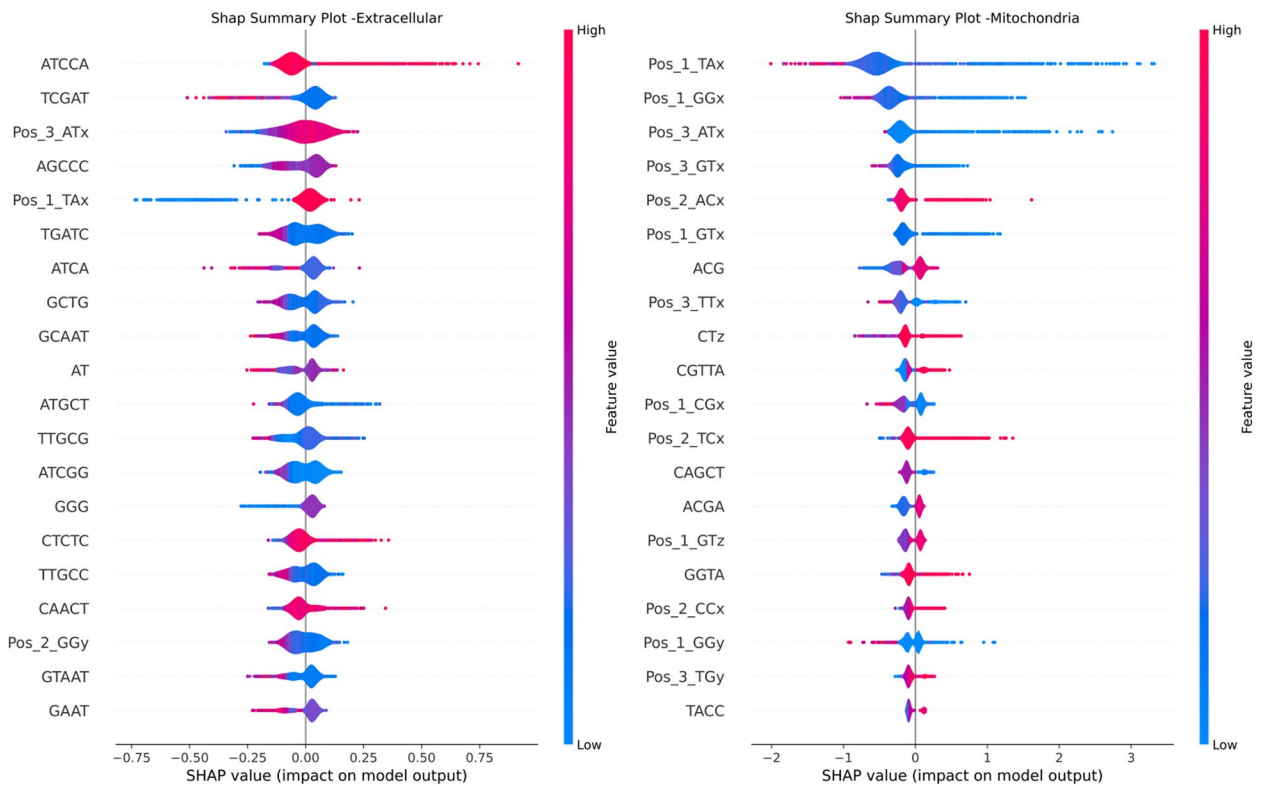


Fig. 11 SHAP Summary Plot - ExR - Mitochondria Feature Importance

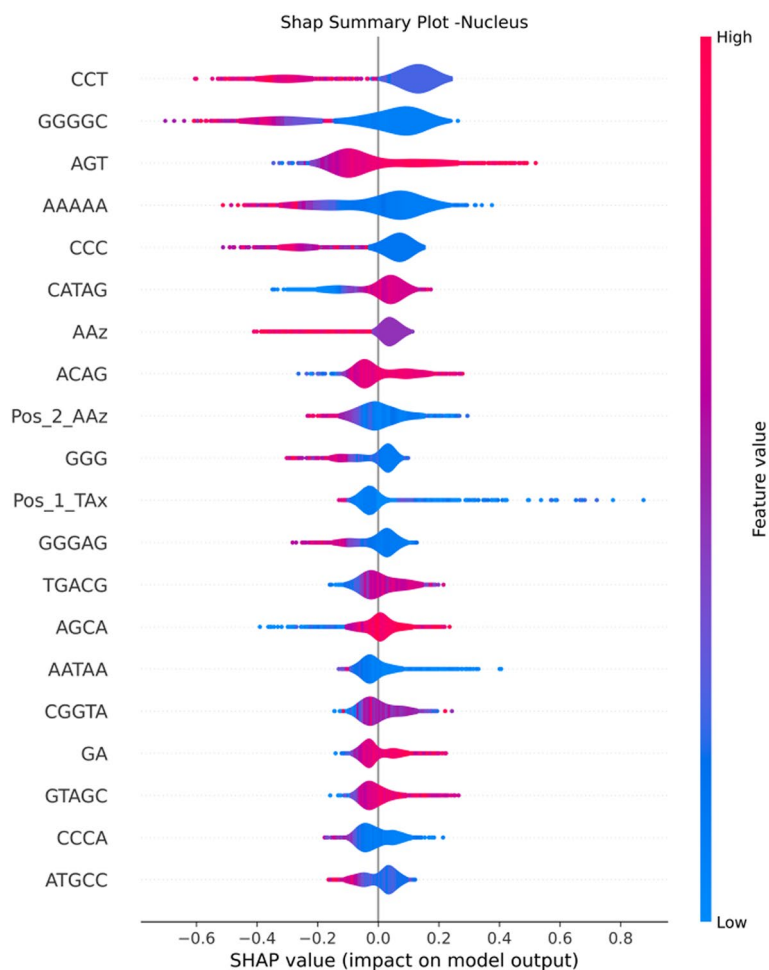


Fig. 12 SHAP Summary Plot - Nucleus Feature Importance

features in the SHAP plots of each localization, we can observe that these Z-curve features are the top 4 features impacting the prediction are Pos_1_Tx, Pos1_GGx, Pos_3_ATx, and Pos_3_GTx, and their negative values of have positive impact (positive Shap values) driving for Mitochondria prediction Fig. 11. Figure 9 shows also that Pos_1_TAx also contributes to the prediction of Extracellular, Nucleus and Cytoplasm but to a much lesser extent than Mitochondria prediction.

We can observe that k-mer based features have higher contribution in the prediction of nucleus and ER. This is evident from Figs. 9 and 12 where negative SHAP values of GGGGC and CCT highly contribute to the prediction of Nucleus localization. For the ER we can observe that k-mers CTGGG, AAGC and AG globally impact the model towards ER prediction (see Fig. 10). The AAAAA Kmer is contributing to both ER and Nucleus. Positive SHAP values of AAAAA drive the model towards the ER prediction and negative SHAP values drive the model towards the Nucleus prediction. The two top k-mer

features contributing to the prediction of Cytoplasm are CATAG and AT. Negative SHAP values of these two features impact the model prediction for cytoplasm. Finally, the top feature that contributes to the ExR localization prediction is the positive values of the ATCCA Kmer (Fig. 11).

In summary, features of type Z-Curve from both 48 bit and 144 bit along with Kmer ones are the dominant feature contributing to the prediction. A few features from type PseKNC where contributing to the predictions. Majority of the Physiochemical and PseKNC features were removed during the features preprocessing phase due to high correlations.

Discussion

In this study, we introduced an unified multi-class classification ML model UMSLP for mRNA subcellular localization problem. Utilizing the optimal set of proposed features, UMSLP outperforms existing models

Table 7 UMSP Comparisons against other tools. “-”: Not reported

Location (total sequence)	Method	Precision %	Recall %	Acc %	F1Score %
Cytoplasm (1066)	UMSLP	76.50	88.90	83.50	82.20
	mRNALocator	55.20	79.64	63.75	65.21
	mRNALoc	-	73.26	64.55	-
Nucleus (976)	UMSLP	83.70	89.00	88.90	88.80
	mRNALocator	91.40	26.13	70.19	40.64
	mRNALoc	-	50.20	69.35	-
ER (241)	UMSLP	92.60	47.10	94.60	62.40
	mRNALocator	100.00	9.13	91.24	16.73
	mRNALoc	-	75.10	69.23	-
ExtR(145)	UMSLP	83.90	18.30	95.10	30.10
	mRNALocator	26.38	95.86	84.23	41.37
	mRNALoc	-	81.38	58.10	-
Mitochondria (71)	UMSLP	1.00	90.00	99.70	94.70
	mRNALocator	44.36	83.10	96.56	57.84
	mRNALoc	-	87.32	96.88	-
Avg(all locals)	UMSLP	67.54	66.66	92.36	71.64
	mRNALocator	63.47	58.77	81.19	44.36
	mRNALoc	-	73.45	71.62	-
	SubLocEP	61.70	60.10	60.10	57.80

for the same problem. During feature preprocessing step, we noticed a strong correlation between two types of features: PseKNC and Kmer, specifically when their K values were equal (K=2,3,4, and 5). Moreover, we observed a significant negative correlation between the TPCP Physicochemical properties and Kmer, PseKNC, and EIIP features. As a result, we retained only 1360, 13, 192 features from Kmer, PseKNC, Z-curve, respectively leading to a dataset with dimensions of 14,909 sequences by 1,565 features, a significant reduction in the dataset's dimensionality. All features from Physicochemical properties were dropped finally. Through extensive analysis, we found that Kmer features with K=2, 3, 4, and 5 play a crucial role in predicting Cytoplasm, ExR, ER, and Nucleus localizations. Moreover, the Z-curve 144-bit features proved to be dominant in predicting mitochondria localization.

Compared to the existing methods such as mRNALocator, mRNALoc, SubLocEP, the proposed XGBoost based UMSP model achieved the best results using this optimal set of features and outperformed them in almost all performance evaluation metrics (Table 7). It is noticeable that that, UMSP outperformed the existing tools for the same purpose on multiple evaluation metrics. Unified multiclass classification based models SubLocEP and mRNALocator highlighted the average results on five subcellular localizations and UMSP outperformed both of them in all

evaluation metrics (Table 7). Compared to mRNALocator, UMSP achieved slightly lower precision in predicting nucleus (UMSP:mRNALocator=83:91) and ER(UMSP:mRNALocator=92:100), but the sensitivity of mRNALocator was too low compared to UMSP (UMSP:mRNALocator=89:26 for nucleus, and UMSP:mRNALocator=47:09 for ER), indicating that the proposed mRNALocator model for nucleus and ER was biased towards positive class ignoring the prediction of other classes. We also compared UMSP against mRNALoc which was a model based on OvR approach of multiclass classification problem. For all the five subcellular localization prediction task, UMSP outperformed mRNALoc in all evaluation metrics indicating the superiority of unified multiclass prediction approach over OvR approach.

Conclusion

In summary, our study underscores the crucial significance of distinct subcellular localization of mRNAs in eukaryotic cells. The proposed ML based which is highly accurate would support in identifying the locales of mRNAs that necessitate extensive manual work, significant expenses, and lengthy wet-lab protocols. Moreover the shared code, data and Docker would support the community to use and improve the proposed solution in near future. In the future, our emphasis will be on studying the multi-label subcellular localizations of mRNAs

due to their presence in multiple locations. We plan to employ CNN, LSTM, and attention-based mechanisms to enhance prediction accuracy. However, past studies indicate that deep learning models have not achieved the performance levels of traditional feature-based machine learning models in this regard [26]. Additionally, there's a need to enhance the quality of datasets pertaining to mRNA localization by encompassing a broader range of locations and conducting more experiments. Our future efforts will involve expanding mRNA localization databases by incorporating scientific evidence.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10077-9>.

Additional file 1. List of selected features for UMSLP.

Additional File 2. Results of ablation study using different ML models.

Additional File 3. Motifs identified based on mRNA sequence from different subcellular localizations.

Acknowledgements

We would like to thank the College of Science and Engineering, Hamad Bin Khalifa University for providing support for this research. We would like to thank Rory Johnson for his comments on the manuscript.

Authors' contributions

Conceived and Design: TA. Experiments: SM. Initial Draft: TA, SM. Analysis and Writing: All authors.

Funding

Open Access funding provided by the Qatar National Library. Open access publication of this article was supported by the Qatar National Library (QNL), Qatar.

Availability of data and materials

We have shared datasets, code, Docker API for users in GitHub at: <https://github.com/smusleh/UMSLP>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 21 September 2023 Accepted: 1 February 2024

Published online: 07 February 2024

References

- Yan Z, Lécuyer E, Blanchette M. Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*. 2019;35(14):i333–42.
- Jeffery WR, Tomlinson CR, Brodeur RD. Localization of actin messenger RNA during early ascidian development. *Dev Biol*. 1983;99(2):408–17.
- Lawrence JB, Singer RH. Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell*. 1986;45(3):407–15.
- Kuriyan J, Eisenberg D. The origin of protein interactions and allostery in colocalization. *Nature*. 2007;450(7172):983–90.
- Martin KC, Ephrussi A. mRNA localization: gene expression in the spatial dimension. *Cell*. 2009;136(4):719–30.
- Jung H, Gkogkas CG, Sonenberg N, Holt CE. Remote control of gene function by local translation. *Cell*. 2014;157(1):26–40.
- Medioni C, Mowry K, Besse F. Principles and roles of mRNA localization in animal development. *Development*. 2012;139(18):3263–76.
- Di Liegro CM, Schiera G, Di Liegro I. Regulation of mRNA transport, localization and translation in the nervous system of mammals (Review). *Int J Mol Med*. 2014;33(4):747–62.
- Cui T, Dou Y, Tan P, Ni Z, Liu T, Wang D, et al. RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. *Nucleic Acids Res*. 2022;50(D1):D333–9.
- Raj A, Van Den Bogaard P, Rifkin SA, Van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. 2008;5(10):877–9.
- Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. 2015;348(6233):aaa6090.
- Eng CHL, Lawson M, Zhu Q, Dries R, Kouloua N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*. 2019;568(7751):235–9.
- Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol*. 2020;38(5):586–99.
- Chen J, McSwiggen D, Ünal E. Single Molecule Fluorescence In Situ Hybridization (smFISH) Analysis in Budding Yeast Vegetative Growth and Meiosis. *J Vis Exp*. 2018;(135):e57774.
- Meyer C, Garzia A, Tuschl T. Simultaneous detection of the subcellular localization of RNAs and proteins in cultured cells by combined multi-color RNA-FISH and IF. *Methods*. 2017;118–119:101–10.
- Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, Ting AY. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *Elife*. 2017;6:e29224.
- Lefebvre FA, Cody NAL, Bouvrette LPB, Bergalet J, Wang X, Lécuyer E. CeFra-seq: Systematic mapping of RNA subcellular distribution properties through cell fractionation coupled to deep-sequencing. *Methods*. 2017;126:138–48.
- Musleh S, Islam MT, Alam T. LNCRI: Long Non-Coding RNA Identifier in Multiple Species. *IEEE Access*. 2021;9:167219–28.
- Salhi A, Essack M, Alam T, Bajic VP, Ma L, Radovanovic A, et al. DES-ncRNA: A knowledgebase for exploring information about human micro and long noncoding RNAs based on literature-mining. *RNA Biol*. 2017;14(7):963–71.
- Basit SA, Qureshi R, Musleh S, Guler R, Rahman MS, Biswas KH, et al. COVID-19Base v3: Update of the knowledgebase for drugs and biomedical entities linked to COVID-19. *Front Public Health*. 2023;11:1125917.
- Alam T, Al-Absi HRH, Schmeier S. Deep Learning in LncRNAome: Contribution, Challenges, and Perspectives. *Noncoding RNA*. 2020;6(4):47.
- Zhang ZY, Yang YH, Ding H, Wang D, Chen W, Lin H. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief Bioinforma*. 2021;22(1):526–35.
- Garg A, Singhal N, Kumar R, Kumar M. mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. *Nucleic Acids Res*. 2020;48(W1):W239–43.
- Tang Q, Nie F, Kang J, Chen W. mRNALocater: Enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. *Mol Ther J Am Soc Gene Ther*. 2021;29(8):2617–23. <https://doi.org/10.1016/j.ymthe.2021.04.004>.
- Wang D, Zhang Z, Jiang Y, Mao Z, Wang D, Lin H, et al. DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Res*. 2021;49(8):e46–e46.
- Yuan GH, Wang Y, Wang GZ, Yang L. RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. *Brief Bioinforma*. 2023;24(1):bbc509.

27. Li J, Zhang L, He S, Guo F, Zou Q. SubLocEP: a novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. *Brief Bioinforma.* 2021;22(5):bbaa401. <https://doi.org/10.1093/bib/bbaa401>.
28. Musleh S, Islam MT, Qureshi R, Alajez N, Alam T. MSLP: mRNA subcellular localization predictor based on machine learning techniques. *BMC Bioinformatics.* 2023;24(1):1–23.
29. Liu B, Liu F, Fang L, Wang X, Chou KC. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics (Oxford, England).* 2015;31(8):1307–9. <https://doi.org/10.1093/bioinformatics/btu820>.
30. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinforma.* 2020;21(3):1047–57. <https://doi.org/10.1093/bib/bbz041>.
31. Gao F, Zhang CT. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics (Oxford, England).* 2004;20(5):673–81. <https://doi.org/10.1093/bioinformatics/btg467>.
32. Zhang R, Zhang CT. A Brief Review: The Z-curve Theory and its Application in Genome Analysis. *Curr Genomics.* 2014;15(2):78–94. <https://doi.org/10.2174/1389202915999140328162433>.
33. Zhang CT, Zhang R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.* 1991;19(22):6313–7.
34. Chen Z, Zhao P, Li C, Li F, Xiang D, Chen Y-Z, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.* 2021;49(10):e60. <https://doi.org/10.1093/nar/gkab122>.
35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* 2017;30:4768–77.
36. Bailey TL. STREME: accurate and versatile sequence motif discovery. *Bioinformatics.* 2021;37(18):2834–40.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.