Allocator is a graph neural network-based framework for mRNA subcellular localization prediction*

Fuyi Li[1,2*], Yue Bi[3], Xudong Guo[1], Xiaolan Tan[4], Cong Wang[1], and Shirui Pan[4,5]

[1] College of Information Engineering, Northwest A&F University, Yangling, 712100, China
[2] South Australian immunoGENomics Cancer Institute, The University of Adelaide, SA 5005, Australia.
[3] Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia
[4] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
[5] School of Information and Communication Technology, Griffith University, QLD 4222, Australia

Abstract. The asymmetrical distribution of expressed mRNAs tightly controls the precise synthesis of proteins within human cells. This non-uniform distribution, a cornerstone of developmental biology, plays a pivotal role in numerous cellular processes. To advance our comprehension of gene regulatory networks, it is essential to develop computational tools for accurately identifying the subcellular localizations of mRNAs. However, considering multi-localization phenomena remains limited in existing approaches, with none considering the influence of RNA's secondary structure. In this study, we introduce 'Allocator', a deep learning-based model that seamlessly integrates both sequence-level and structure-level information, significantly enhancing the prediction of mRNA multi-localization. The Allocator models equip four efficient feature extractors, each designed to handle different inputs. Two are tailored for sequence-based inputs, incorporating multilayer perceptron and multi-head self-attention mechanisms. The other two are specialized in processing structure-based inputs, employing graph neural networks. Benchmarking results underscore Allocator's superiority over state-of-the-art methods, showcasing its strength in revealing intricate localization associations. Furthermore, we have developed a user-friendly web server for Allocator, enhancing its accessibility. You can explore Allocator through our web server, available at http://Allocator.unimelb-biotools.cloud.edu.au.

Keywords: mRNA · subcellular localisation · multi-class classification · multi-label prediction · graph neural network.

1 Introduction

The asymmetric distribution of mRNA has been discovered to be closely associated with various cellular processes, laying the groundwork for local protein translation and organism development [1,2]. This significance of asymmetric distribution has been shown to be universal across different species [3,4]. In particular, Bouvrette *et al.* found that around 80% of cellular RNA species exhibit non-uniform distribution, with these patterns being broadly conserved throughout evolution [5,6]. Recent advancements in high-resolution imaging technologies have further revealed that mRNA asymmetrical distribution, known as mRNA subcellular localization, plays a crucial role in temporal and spatial control of gene expression [7]. For example, in the D. *melanogaster* embryo, cytoplasmic *Nos* mRNA binds to RBP Smaug, inhibiting its own translation and triggering decay. However, in later development, *Nos* mRNA is localized to the posterior pole and replaces Smaug with RBP Oskar, thereby protecting transcripts from degradation. However, the extensive time and substantial financial resources required for wet lab experiments pose significant limitations on conducting thorough investigations, thus impeding our comprehensive understanding of the RNA localization mechanism.

In recent years, there have been various computational approaches for predicting RNA subcellular localizations. They offer effective and time-saving alternatives to traditional laboratory experiments. These approaches

* This work is supported by the National Natural Scientific Foundation of China (grant 62202388). Fundamental Research Funds for the Central Universities (Z1090222021). The first two authors contributed equally to this study.
*To whom correspondence should be addressed: Fuyi Li, College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China 712100. Email: fuyi.li@nwafu.edu.cn.

can be categorized into two types. The first type regards localization prediction as a single-label multi-class task, where the goal is to identify a single localization for a given transcript. Models such as RNATracker [8], iLoc-mRNA [9], mRNALoc [10], and SubLocEP [11] adopted this strategy to build their predictive models. However, not all transcripts serve a single function; they may be localized in different compartments to fulfil different roles, as seen with *Nos* mRNA. Therefore, developing a single-label model falls short of providing comprehensive predictions for multiple localizations and fails to reveal the complex relationships between labels. It's essential to consider the broad spectrum of multi-localization in model development. To address this issue, DM3Loc [12] and clarion [13] were introduced by treating it as a multi-label task. DM3Loc leveraged a multi-head self-attention mechanism for predictions, while clarion employed XGBoost based on a problem transformation strategy for classification. Both of them offer predictions for multiple labels simultaneously, significantly advancing the state of the art in mRNA subcellular localization prediction.

Nevertheless, DM3Loc and clarion were developed solely based on sequence information. In addition to the primary mRNA sequence, the secondary structure also plays a pivotal role in the localization process. Indeed, mRNAs are often targeted and bound by RNA-binding proteins (RBPs) via the *cis*-localization element, which can act as signals for delivery to a specific cellular localization [14]. For instance, in the 3'UTR of fruit fly *bicoid* mRNA, several 50-nt sequences, known as bicoid localization elements, have been identified. These elements form stem-loop structures that facilitate intermolecular interactions [15]. This insight has motivated us to incorporate mRNA secondary structure into the model construction. The similarity between RNA structure and graph representation enables the utilization of graph neural networks (GNNs) in this task. GNNs have proven effective in capturing intricate patterns and correlations, showcasing remarkable performance across diverse bioinformatics tasks [16,17]. Recent years have witnessed the emergence of several influential GNN architectures, each offering a distinct set of strengths: MPNN (message-passing neural network) passes information between nodes to update node representations. GCN (graph convolutional network) uses graph convolution operations for local information propagation. GIN (graph isomorphism network) aims to identify structural similarities and solve isomorphism challenges. GAT (graph attention network) applies attention to determine the weights of information exchange between different nodes. In this paper, we proposed a novel approach, termed Allocator, that leverages the power of GIN to analyze structural features, thereby providing a substantial boost to the model's performance.

Allocator is a deep learning-based method designed to identify multiple subcellular localizations of mRNA. It incorporates various network architectures, including multilayer perceptron (MLP), self-attention, and GIN, to ensure precise and reliable predictions. Two input types are generated from the mRNA sequences: features obtained from the k-mer and CKSNAP encoding schemes and a graph representation of the mRNA secondary structure. These inputs undergo feature learning through two numerical extractors and two graph extractors. The learned features are then combined and fed into fully connected layers to generate predictions for six subcellular localization categories. Allocator has been demonstrated to be outstanding in revealing label correlations according to six multi-label evaluation metrics. To our knowledge, Allocator is the first approach that utilizes graph neural networks to capture mRNA secondary structure information to predict mRNA subcellular localization.

2 Material and Methods

2.1 Benchmark dataset

In this study, we employed a benchmark dataset comprising over 26,000 entries for mRNA subcellular localization in *Homo sapiens*. These entries were originally sourced from RNALocate [18] and curated by DM3Loc [12]. The dataset includes a total of 17,298 unique mRNA sequences in FASTA format. All sequences have undergone sequence similarity reduction using CD-HIT-EST [19] with an 80% similarity threshold. Their lengths range from 186-nt to 30,609-nt, with an average length of 3689-nt. We then divided this dataset into the training set, validation set, and testing set in an 8:1:1 ratio. **Table 1** displays the specific distribution for each subcellular localization within the three subsets. The 'Sequences' column represents the actual number of FASTA sequences, while other columns, such as 'Nucleus', count only the number of labels (localizations). It is clear that the label relationship between mRNA and localization labels is not limited to one-to-one but also

one-to-many, which can be addressed as a multi-label problem. Furthermore, it is important to highlight our decision to utilize the original sequences rather than truncating sequences in order to preserve the complete genomic information.

Table 1: The distribution for six subcellular localizations in training, validation, and testing set

Data set	Sequences	Nucleus	Exosome	Cytosol	Ribosome	Membrane	ER
Training	13838	9479	13733	7710	4159	2555	1577
Validation	1729	1233	1705	990	539	339	206
Testing	1731	1204	1710	945	509	336	193
Total	17298	11916	17148	9645	5207	3230	1976

2.2 Allocator's framework

Allocator is a multi-view parallel deep learning framework that is designed for mRNA multi-localization prediction. It is composed of three main modules: feature engineering, feature extractor, and prediction module, as depicted in **Figure 1**. The feature engineering module involves preparing model inputs by leveraging information from both the primary sequence and secondary structure. This step results in two sets of numerical features and two sets of graph-based features, forming the foundation for the subsequent model training. Then, these features are processed through the feature extractor module to capture intricate and complex patterns. There are four parallel extractors, with two specifically for numerical features and the other two for graph-based features. Lastly, the outcomes of these extractors are concatenated and fed into the output module, consisting of two fully connected layers, to generate the predictive probabilities. More detailed explanations of each module are presented in the following sections.

Feature engineering Regarding the numerical features, we employed two encodings, k-mer and CKSNAP (k-spaced nucleic acid pairs), for extracting primary sequence characteristics [20,21]. These encodings have demonstrated their versatility in various bioinformatics tasks [13,22,32]. Specifically, the k-mer encoding calculates the frequency of k adjacent nucleic acids. In our study, we considered the frequencies of 1-mer, 2-mer, 3-mer, 4-mer and 5-mer, resulting in a feature vector with the dimensionality of 1364. On the other hand, CKSNAP represents the frequency of nucleic acid pairs by any k nucleic acid. We chose k values of 0, 1, 2, 3, 4 and 5, leading to a 96-dimensional feature vector. The feature extraction process described above was conducted using the iLearn toolkit, where additional details can be found [20].

Incorporating the RNA's secondary structure into the model can offer valuable insights into its folding patterns, including structural elements like stem-loops and pseudoknots. However, experimentally determined RNA secondary structures are minimal and insufficient for our dataset. Therefore, we turned to predictive secondary structures to generate the graph-based features. For each RNA molecule, we represent it as a graph with individual bases as nodes. The graph is constructed using two types of edges: one connecting adjacent bases and the other representing base-pairing interactions derived from a predictive tool. We have prepared two distinct graphs predicted by two popular popular prediction tools, i.e., RNAFold [23] and LinearFold [24]. Both of these graphs include the formation of hydrogen bonds and phosphodiester bonds. In addition, each node is denoted by a 10-dimensional feature vector that integrates four different encodings: one-hot, NCP (nucleotide chemical property), EIIP (electronion interaction pseudopotentials, and ANF (accumulated nucleotide frequency) [25]. (Further details about the node features are provided in **Table S1**).

Numerial extractor To facilitate the feature extractor's capabilities at both local and global levels, we introduced multiple types of neural network architectures into the design of our numerical extractors. These two numerical extractors share a standard set of components, which primarily include an MLP layer and a

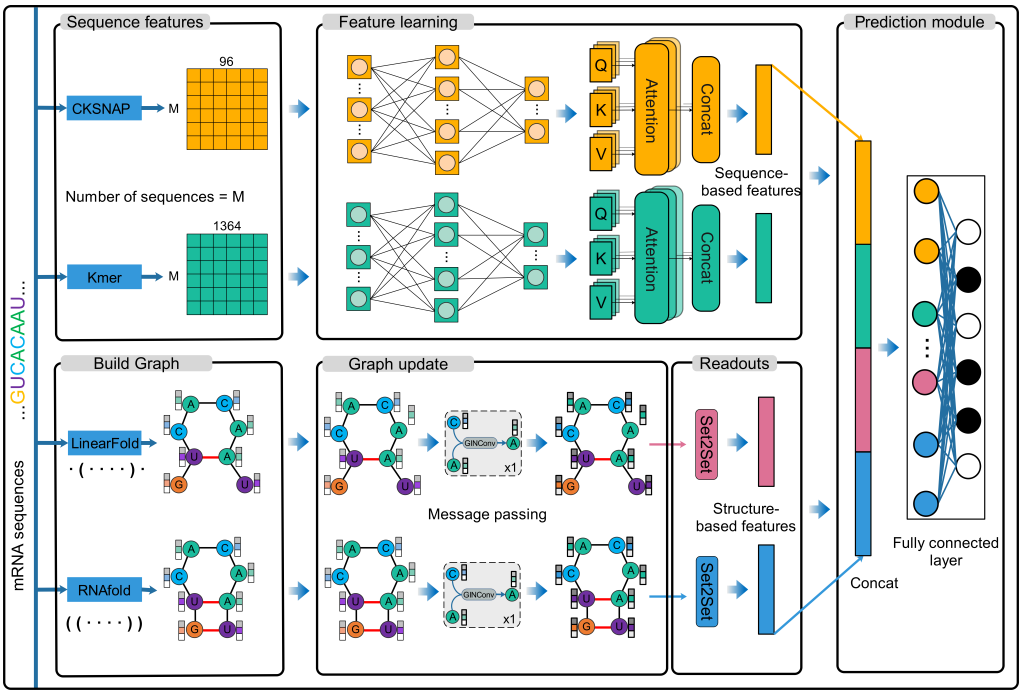

Fig. 1. The framework of Allocator.

multi-head self-attention layer. As a core extractor component, the multi-head self-attention can weigh the importance of individual words within the input sequences. This mechanism is achieved by combining multiple attention heads, each relying on the scaled dot products to calculate the attention scores. The specific process can be defined as follows:

$$MultiHeadAttention\,(Q,\ K, V) = Concat\,(head_1,\ \ldots, head_n)\,W^0 \tag{1}$$

$$where\ head_i = Attention\,(Q_i,\ K_i, V_i) = softmax\left(\frac{Q_i \times K_i^T}{\sqrt{d_{k_i}}}\right) V_i \tag{2}$$

$$Q_i = X \times W_i^Q \tag{3}$$

$$K_i = X \times W_i^K \tag{4}$$

$$V_i = X \times W_i^V \tag{5}$$

where X represents the input of attention module, W_i^Q, W_i^K, W_i^V are learnable parameter matrices, and d_{k_i} is the dimension of matrix K_i.

Graph extractor We utilized graph isomorphic networks to establish Allocator's graph extractors, which is a specialized type of graph network used for various graph-related tasks. Graph Isomorphic Network (GIN) [26] is designed to aggregate information from neighbouring nodes, as formulated below:

$$h_v^{(l+1)} = MLP^{(l+1)}\left((1+\epsilon) \bullet h_v^{(l)} + \sum_{u \in \mathcal{N}(v)} h_u^{(l)}\right) \tag{6}$$

where $h_v^{(l)}$ is the representation of node v at layer l, ϵ is a learnable parameter, and $MLP^{(l+1)}$ is a multilayer perceptron used to update the node's representation. Following the GIN layer, we employed a set2set layer, acting as a global graph pooling mechanism, to capture and summarize information from the entire set. This layer outputs a fixed-size representation for subsequent combinations with other extractors' outputs. To simplify the code writing, we used the PyTorch Geometric package [27] to implement the construction of a graph neural network. For more detailed model parameters for the Allocator, please refer to **Supplementary Table S2**.

Prediction module Our prediction module is a feedforward network with two layers. After concatenating the outcomes of the four extractors, the prediction module performs the linear transformation and then applies a sigmoid function in the second layer to produce the final output, which is formulated as follows:

$$output = Sigmoid\left(W_2 \times \left(W_1 \times X_{concat} + b_1\right) + b_2\right) \tag{7}$$

where W_1, W_2, b_1, and b_2 are learnable parameters.

2.3 Model training

During the training of Allocator, we employed the Adam [28,29,30] optimization algorithm to enhance the training process and optimize the model's parameters. To measure the dissimilarity between the true labels and the predicted probabilities, we used the binary cross-entropy function that is defined as follows:

$$Loss_i = \sum_{1}^{6} y_{ij} \cdot \log p_{ij} + \left(1 - y_{ij}\right) \cdot log\left(1 - p_{ij}\right) \tag{8}$$

where y_{ij} takes the value 0 or 1, and p_{ij} takes the value in $[0, 1]$. When the y_{ij} value is 0, it means that the mRNA sample x_i does not belong to the corresponding mRNA subcellular localization compartment j (1, 2, 3, 4, 5, 6); on the contrary, if the y_{ij}value is 1, it means that the mRNA sample x_i is located in the corresponding compartment j. Our training objective is to make the predicted probability value as close to 1 or 0 as possible, with the model providing separate predicted probability values for each location compartment.

2.4 Model evaluation

To evaluate Allocator's overall performance, we utilized six evaluation metrics designed explicitly for multi-label problems, including example-based accuracy (Acc_{exam}), average precision, coverage, one-error, ranking loss, and hamming loss. These metrics offer a comprehensive assessment of Allocator's capabilities across all six labels/locations. The descriptions of these metrics are provided below:

$$Acc_{exam} = \frac{1}{t}\sum_{i=1}^{t} \frac{|P_i \cap Y_i|}{|P_i \cup Y_i|} \tag{9}$$

$$Average\ Precision = \frac{1}{t}\sum_{i=1}^{t} \frac{1}{|Y_i|}\sum_{y \in Y_i} \frac{|\{y' \,|Rank_f\left(x_i, y'\right) \leq Rank_f\left(x_i, y\right), y' \in Y_i\}|}{Rank_f\left(x_i, y\right)} \tag{10}$$

$$Coverage = \frac{1}{t}\sum_{i=1}^{t} max_{y' \in Y_i} Rank\left[f\left(x_i, y'\right)\right] - 1 \tag{11}$$

$$One - error = \frac{1}{t}\sum_{i=1}^{t} I\left(\arg max_{y' \in Y_i} f\left(x_i, y'\right) \notin Y_i\right) \tag{12}$$

$$Ranking\ Loss = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{|Y_i|\left|\bar{Y}_i\right|}I\left(f\left(x_i,y'\right)\leq f\left(x_i,y''\right),y'\in Y_i,\ y''\in\bar{Y}_i\right) \tag{13}$$

$$Hamming\ Loss = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{q}\left|P_i\Delta Y_i\right| \tag{14}$$

where $f(\cdot)$ represents the classifier; Y_i and $\overline{Y}_i$ represent the set of real labels and the complement of this set, respectively; P_i represents the set of predicted labels; (x_i, y_i) $\{1 \leq i \leq t\}$ represents a multi-label instance; $Rank_f(x, y)$ indicates the descending ranking of y in Y; $|\cdot|$ indicates the cardinality of the set, $q = |Y_i|$; Δ is the symmetric difference between two sets; and $I(\cdot)$ indicates the count that satisfies the condition.

3 Results and discussion

3.1 Allocator's multi-label performance

Prior to this, DM3Loc used one-hot representations to achieve impressive results; however, it might overlook capturing relationships between nucleotide sequences and structural information. The Allocator's key innovation lies in the incorporation of the RNA secondary structure. We sought to investigate whether this structural information can provide additional insights. As part of our exploration to explore the effect of secondary structure, we retrained a new model only utilizing numerical extractors with sequence-based features of k-mer and CKSNAP, that is, Allocator (seq). This model differs from the original Allocator (seq + structure). Both models were trained with a learning rate of 0.0003 and a dropout rate of 0.1, lasting for 200 epochs. We conducted a comparison among Allocator (seq), Allocator (seq + structure), and DM3Loc as illustrated in **Figure 2**. Firstly, it is evident that Allocator (seq + structure) leads to a significant improvement in terms of (Acc_{exam}). Additionally, it reduces the coverage, ranking loss, and hamming loss when compared to Allocator (seq). This demonstrates that integrating the secondary structure definitely enhances performance, highlighting the importance of multi-source features.

In addition to the multi-label evaluation, we also conducted a single-label assessment of the Allocator's performance. We used accuracy and Matthews correlation coefficient (MCC) to measure the efficacy of each label/localization (**See supplementary Table S3**). Allocator's accuracies are higher than DM3Loc in all six localization predictions. Notably, for the exosome's prediction, the Allocator model achieves an accuracy of 98.80%, surpassing DM3Loc's 73.71% by approximately 25 percentage points. However, Allocator's MCC performance falls short of that achieved by DM3Loc. We found that the Allocator's performance is not satisfactory for imbalanced labels in the case of exosome, membrane and ER localization. This can be attributed to the model's training direction, which emphasizes label relationships and tends to favour the majority classes. In contrast, DM3Loc defined a range of cutoffs for each localization, leading to an increase in MCC but a decrease in accuracy. Indeed, addressing this inherent problem requires the inclusion of more training samples in the future. Overall, while Allocator may not excel in terms of MCC, it still offers a valuable perspective for uncovering label relationships.

3.2 Architectural analysis of Allocator

The feature extractor module serves as the essential core of the Allocator models and is responsible for extracting intricate patterns from RNA sequences and structural features. To understand the contributions of each component in Allocator, we conducted comprehensive experiment analyses from the views of numerical extractors, graph extractors, and the entire architecture.

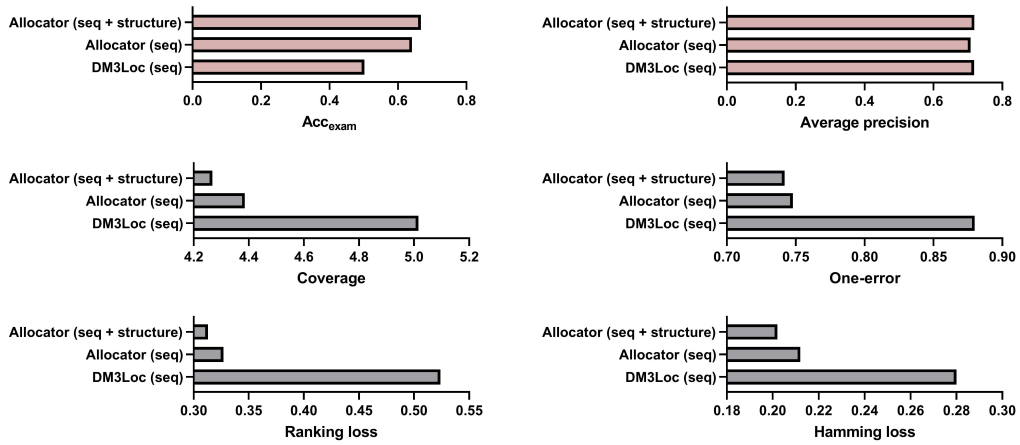

Fig. 2. illustrates the performance of Allocator across six multi-label evaluation metrics.

Numerical extractor study We explored four network combinations to optimize the numerical extractor, which includes using only MLP, MLP+attention, CNN+LSTM [31], and CNN+LSTM+attention [30]. The MLP consists of three hidden layers, while both CNN and LSTM have a single layer. As illustrated in **Table 2**, the combination of MLP+attention achieves the best results in terms of Accexam and average precision, along with yielding the lowest coverage, one-error, and hamming loss. The combination of CNN and LSTM (CNN+LSTM) did not demonstrate superior performance, possibly due to the input features not being context-independent. Furthermore, we observed a significant improvement when the attention mechanism was added compared to the model without attention. Therefore, we adopted the combination of MLP and attention (MLP+attention) for the numerical extractors, aiming to extract in-depth attributes from sequence-based features.

Table 2: Performance comparison on the test set using different neural network types

Type	Accexam	Average precision	Coverage	One-errorl	Ranking loss	Hamming loss
N: MLP +attention & G: GIN (Allocator)	0.667	0.719	4.268	0.742	0.313	0.202
N: MLP	0.647	0.709	4.344	0.770	0.337	0.218
N: CNN + LSTM	0.638	0.716	4.421	0.763	0.355	0.221
N: CNN + LSTM + attention	0.650	0.708	4.388	0.762	0.312	0.205
G: GAT	0.650	0.710	4.347	0.748	0.323	0.209
G: GCN	0.635	0.716	4.439	0.744	0.352	0.216
G: MPNN	0.640	0.707	4.355	0.760	0.346	0.221
Without N-kmer	0.656	0.714	4.360	0.753	0.309	0.201
Without N-CKSNAP	0.646	0.712	4.352	0.763	0.347	0.219
Without G-LinearFold-SS	0.652	0.716	4.334	0.758	0.333	0.212
Witheout G-RNAFold-SS	0.641	0.708	4.369	0.776	0.347	0.222

* Abberiations: numerical extractor (N); graph extractor (G).

Graph extractor study As mentioned in the introduction, GIN, GAT, GCN, and MPNN are different types of graph neural networks, each possessing unique characteristics and approaches to handling graph-structural data. To guide the development of our graph extractors, we explored these four network types to evaluate their impact on model performance. The results suggest that the GIN-based model consistently attains the highest levels of Acc_{exam} and average precision, thereby establishing GIN as the optimal choice for Allocator's graph extractors.

Ablation study After completing the entire design of Allocator, we made further efforts to reduce complexity and streamline the architecture. To achieve this, we conducted ablation experiments, systematically removing individual extractors to assess their effect. These ablations included the removal of the numerical extractor for k-mer (i.e., N-kmer), the removal of the numerical extractor for CKSNAP (N-CKSNAP), the exclusion of the graph extractor for secondary structure from Linearfold (G-linearFold-SS), and the exclusion of the graph extractor for secondary structure from RNAfold (G-RNAFold-SS). Also, as listed in **Table 2**, when we removed individual extractors, the Acc_{exam} and average precision metrics display decreased performance in comparison to the complete Allocator. Notably, in contrast to N-kmer and G-LinearFold-SS, N-CKSNAP and G-RNAfold-SS exhibit a more pronounced influence on Allocator. As a result, we decided to retain all four extractors to maintain a more robust model capacity.

3.3 Analysis of single-localization

In this section, we analyzed to explore the impact of the Allocator's predictions on single localizations. Firstly, we performed a statistical analysis using the test set, comparing the actual and predicted samples of six different localizations. As depicted in **Figure 3A**, we observed that the distributions for exosome and ribosome closely matched between the actual and predicted samples. In contrast, for the nucleus and cytosol, the predicted samples exceeded the actual samples, while for membrane and ER, the opposite was observed. One potential explanation is the impact of imbalanced samples across different localizations, as these imbalances appear to have a cross-influence on one another. Nevertheless, it's crucial to highlight that, on the whole, the actual labels and those predicted remain highly similar, providing evidence of the Allocator's outstanding predictive ability.

Afterwards, we selected correctly predicted samples for each localization separately and observed the layer representations for each of the four extractor combinations. The reason for not being able to choose the same samples is that their feature outputs are the same since this is a multi-label problem, and thus, no differences can be observed. Furthermore, we excluded the exosome category from our analysis due to its extreme imbalance, and we could not obtain a sufficient number of negatives for this category. **Figure 3B** displays the average representation values for positives and negatives across the remaining five locations. Notably, we observe significant differences in the categories of nucleus, cytosol, membrane and ER. This highlights the effectiveness of combining the outputs from the numerical and graph extractors, thereby reinforcing the superiority of the Allocator structure.

3.4 Webserver and software development

To facilitate easy access to Allocator, we have developed a user-friendly web server, which is freely available at http://allocator.unimelb-biotools.cloud.edu.au. The web server is hosted on a robust Linux server, equipped with four cores, 16 GB of memory, and a spacious 200 GB hard disk. This powerful configuration enables the server to handle various tasks and deliver a smooth user experience. With Allocators's web server, users are required to upload or paste the mRNA sequences in FASTA format to submit an online prediction task. Once the prediction is completed, the results will be presented in a tabular format. Users can download the results in different file formats, such as EXCEL, TXT, and CSV. We also offer a detailed guide on the 'Help' page of our website, explaining the specific steps for using Allocator. Please note that the online prediction can simultaneously accommodate up to 5 mRNA sequences. For prediction involving a larger range of sequences, please download the standalone software of Allocator for fast and efficient local predictions.

4 Discussion and conclusion

The widespread identification of mRNA subcellular localizations is crucial in understanding intricate gene regulatory mechanisms. To complement labour-intensive wet lab experiments, several computational approaches have been proposed for predicting mRNA subcellular localizations. However, most existing methods overlook the multi-localization aspect of mRNA and treat it as a simple single-label multi-class task. While a multi-label and multi-class model, DM3Loc, has been developed based on the self-attention neural network, there is

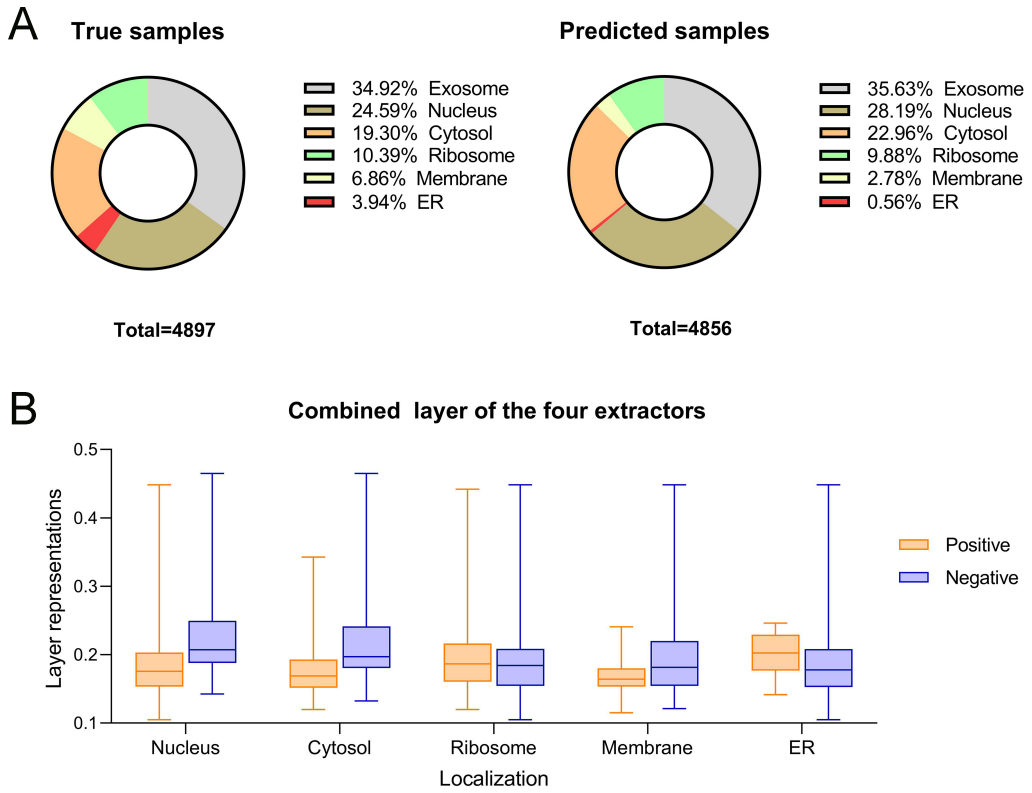

Fig. 3. (A) shows the distribution of actual and Allocator-predicted samples for all six localization categories; (B) displays the layer output representations for both positives and negatives in five localizations, excluding exosome.

a gap in addressing the complexities of mRNA subcellular localizations due to the dynamic cellular processes. The secondary structure of mRNA is closely related to its functionality and may contain crucial information regarding transport, such as *cis*-localization element. Building on this insight, we introduced a deep learning model named Allocator, designed to identify multiple mRNA localizations simultaneously. Allocator not only considers mRNA primary sequence to generate numerical features but also raises the secondary structure to create the graph representations. The main components of Allocator models consist of two numerical extractors specialized for sequence features and two graph extractors for graph representations. This heterogeneous network architecture empowers the Allocator models to exhibit robust performance. In summary, this paper showcases the effectiveness of the Allocator framework, which holds promise for application in various biological analysis tasks in future research.

Despite achieving satisfactory results in revealing label relationships, Allocator still faces challenges performing well on imbalanced labels. Additionally, it is essential to emphasize that the mRNA secondary structure used in this work is derived from prediction tools and may not accurately represent the actual secondary structure of mRNA, potentially impacting the model's performance. We believe that with the ongoing advancement and expansion of experimental data and databases, these challenges can be effectively mitigated and resolved, leading to an enhancement in the model's capabilities.

References

1. Ephrussi A, Dickinson LK, Lehmann R. Oskar organizes the germ plasm and directs localization of the posterior determinant nanos, Cell 1991;**66**(1):37-50.

2. Weatheritt RJ, Gibson TJ, Babu MM. Asymmetric mRNA localization contributes to fidelity and sensitivity of spatially localized systems, Nature structural & molecular biology 2014;**21**(9):833-839.
3. Long RM, Singer RH, Meng X et al. Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA, Science 1997;**277**(5324):383-387.
4. Kloc M, Zearfoss NR, Etkin LD. Mechanisms of subcellular mRNA localization, Cell 2002;**108**(4):533-544.
5. Bouvrette LPB, Cody NA, Bergalet J et al. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells, Rna 2018;24(1):98-113.
6. Li J, Zhang L, He S et al. SubLocEP: a novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning, Briefings in Bioinformatics 2021;**22**(5):bbaa401.
7. Martin KC, Ephrussi A. mRNA localization: gene expression in the spatial dimension, Cell 2009;**136**(4):719-730.
8. Yan Z, Lécuyer E, Blanchette M. Prediction of mRNA subcellular localization using deep recurrent neural networks, Bioinformatics 2019;**35**(14):i333-i342
9. Zhang Z-Y, Yang Y-H, Ding H et al. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens 2021;**22**(1):526-535.
10. Garg A, Singhal N, Kumar R et al. mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization 2020;**48**(W1):W239-W243.
11. Li J, Zhang L, He S et al. SubLocEP: a novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning 2021;**22**(5):bbaa401.
12. Wang D, Zhang Z, Jiang Y et al. DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism, Nucleic acids research 2021;**49**(8):e46-e46.
13. Bi Y, Li F, Guo X et al. Clarion is a multi-label problem transformation method for identifying mRNA subcellular localizations, Briefings in Bioinformatics 2022;**23**(6):bbac467.
14. Das S, Vera M, Gandin V et al. Intracellular mRNA transport and localized translation, Nature Reviews Molecular Cell Biology 2021;**22**(7):483-504.
15. Buxbaum AR, Haimovich G, Singer RH. In the right place at the right time: visualizing and understanding mRNA localization, Nature Reviews Molecular Cell Biology 2015;**16**(2):95-109.
16. Réau M, Renaud N, Xue LC et al. DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces, Bioinformatics 2023;**39**(1):btac759.
17. Kang C, Zhang H, Liu Z et al. LR-GNN: A graph neural network based on link representation for predicting molecular associations, Briefings in Bioinformatics 2022;**23**(1):bbab513.
18. Zhang T, Tan P, Wang L et al. RNALocate: a resource for RNA subcellular localizations 2017;**45**(D1):D135-D138.
19. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation sequencing data 2012;**28**(23):3150-3152.
20. Chen Z, Zhao P, Li F et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data 2020;**21**(3):1047-1057.
21. Lee D, Karchin R, Beer MAJGr. Discriminative prediction of mammalian enhancers from DNA sequence 2011;**21**(12):2167-2180.
22. Chen R, Li F, Guo X et al. ATTIC is an integrated approach for predicting A-to-I RNA editing sites in three species 2023:bbad170.
23. Lorenz R, Bernhart SH, Höner zu Siederdissen C et al. ViennaRNA Package 2.0 2011;6:1-14.
24. Huang L, Zhang H, Deng D et al. LinearFold: linear-time approximate RNA folding by 5'-to-3'dynamic programming and beam search 2019;**35**(14):i295-i304.
25. Liu Q, Chen J, Wang Y et al. DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites, Briefings in Bioinformatics 2021;**22**(3):bbaa124.
26. Xu K, Hu W, Leskovec J et al. How powerful are graph neural networks? 2018.
27. Fey M, Lenssen JEJapa. Fast graph representation learning with PyTorch Geometric 2019.
28. Kingma DP, Ba J. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 2014.
29. Li F, Chen J, Leier A et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites, Bioinformatics 2020;**36**(4):1057-1065.
30. Li F, Guo X, Bi Y et al. Digerati–A multipath parallel hybrid deep learning framework for the identification of mycobacterial PE/PPE proteins, Computers in Biology and Medicine 2023:107155.
31. Liu Q, Fang H, Wang X et al. DeepGenGrep: a general deep learning-based predictor for multiple genomic signals and regions, Bioinformatics 2022;**38**(17):4053-4061.
32. Li F, Guo X, Jin P et al. Porpoise: a new approach for accurate prediction of RNA pseudouridine sites, Briefings in Bioinformatics 2021;**22**(6):bbab245.