bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Sentences, Words, Attention: A "Transforming" Aphorism of miRNA Discovery

Sagar Gupta[1,3], Vishal Saini[2,3] , Rajiv Kumar[2,3], and Ravi Shankar*[1,3]

[1] Studio of Computational Biology & Bioinformatics,

The Himalayan Centre for High-throughput Computational Biology,

(HiCHiCoB, A BIC supported by DBT, India),

CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),

Palampur (HP), 176061, India.

[2]Biotechnology Division,

CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),

Palampur (HP), 176061, India.

[3]Academy of Scientific and Innovative Research (AcSIR),

Ghaziabad, Uttar Pradesh- 201002

*Corresponding Author: ravish@ihbt.res.in

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

miRNAs are major post-transcriptional regulators. Discovering pre-miRNAs is the core of locating miRNAs and their genomic annotations. Using traditional sequence/structural features many tools have been published to discover miRNAs. However, in practical applications like genomic annotation, their actual performance has been far away from acceptable. This becomes more grave in plants where unlike animals pre-miRNAs are much more complex and difficult to identify. This is reflected by the huge gap between the available software for animal and plant miRNA discovery. Here, we present miWords, an attention based genomic language processing transformer and context scoring deep-learning approach to accurately identify pre-miRNAs in plants which can be extended to other eukaryotes also. During a comprehensive bench-marking the transformer part of miWords alone significantly outperformed the compared published tools with consistent performance while maintaining an accuracy of ~94% across a large number of experimentally validated data. Performance of miWords was also evaluated with *Arabidopsis* genome annotation where also miWords outperformed even those software which essentially use sRNA-seq reads to identify miRNAs. miWords was run across the Tea genome, reporting 821 pre-miRNAs, all validated by RNA-seq data. 10 such randomly selected novel pre-miRNAs were also experimentally validated through qRT-PCR.

Keywords: microRNA, Transformers, Gradient Boosting, Deep learning, Genomics

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

miRNAs are prime regulatory small RNAs (sRNAs) having ~21 bases length which post-transcriptionally regulate most of the genes and stand critical for most of the processes of eukaryotic cells including their development and specialization (1). miRNAs can be found in the intronic as well as intergenic regions (2,3). Mature miRNAs are derived from longer precursor miRNA molecules (pre-miRNAs) which are double-stranded RNAs (dsRNAs) with terminal hairpin loop. Discovering these pre-miRNAs is the central to the problem of finding miRNAs and genomic annotations for them. However, finding these pre-miRNAs remains a challenge, and more so in plants. Unlike animals, in plants mature miRNA formation from the precursors is a single step process. Also in terms of complexity, secondary structures of plant pre-miRNAs are way more complex and longer than those of animal pre-miRNAs (4), making them difficult to detect accurately. The traditionally considered sequence and structural properties and features to identify miRNAs also hold responsibility of difficulty in identifying them as they display lots of variability across the plant genomes in terms of sequence composition properties, structural, and thermodynamic properties.

A comparison of these traditional properties properties like AU%, GC%, length of sequence, number of bulges, terminal loop length, minimum free energy (MFE), maximum bulge size, mismatches and stem length for pre-miRNAs in plants and animals display a good amount of variability and overlap with other genomic elements with several of these properties. Also these values differ a lot between plants and animals, while within plants a good overlap is found between miRNAs and non-miRNAs regions for the same properties (Figure 1 and Supplementary Table 1 Sheet 1). This all suggest that how they are prone to wrongly identify the miRNA regions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Though experimental techniques like direct cloning and quantitative real-time PCR (qRT-PCR) are used identify miRNAs with high expression levels, they are very costly and cumbersome for genomic level identifications and remain mainly limited for experimental validation purpose or studying a handful of miRNAs. Off-late, experiments like RNA-seq and arrays studies have been identified as the main way to identify and profile miRNAs/pre-miRNAs across the genome, but they too essentially require support from computational approaches while using the sequencing read data. The difficulties in the identification of plant miRNAs has been so much that it could be fathomed from the fact that it has been plagued of huge volume of false identifications and reporting and in year 2018 miRBase had to scrap a large number of the reported plant miRNAs data (5). An urgent call was made to critically look into the process of plant miRNA identification and annotation process, following which in the year 2018 some critical guidelines were suggested (6). These studies highlighted the need of support by sRNA-seq for reporting novel miRNAs and seriously questioned the capabilities of existing software pool to identify plant pre/miRNAs which were flooded with false predictions and false positive reportings. The same studies, therefore, also recommended that credibility of any such software must be judged by using it across some well established and annotated genomes like *Arabidopsis*, and evaluate their false positive rates. Since then, most of the approaches to identify plant miRNAs and their precursors have been focused on using sRNA-seq data to report the precursors and their miRNAs while basically founded on the above mentioned traditional properties as descriminators.

In the identification of miRNA precursors, identification of secondary structure patterns, hairpin loops, and their thermodynamic stability have been the most followed approaches. Additionally, homology and conservation patterns were also used to locate similar kind of precursors in other genomes. Tools like MirFinder (7), MicroHARVESTOR (8), RNAmicro (9), MIRcheck (10) show case that. Based on near-perfect complementary properties between plant miRNAs and their

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

targets , other methods like FindMiRNA (11) and MiMatcher (12) were also developed to predict the plant pre-miRNAs.

Lately, with the rise of next generation sequencing techniques, it has become possible to sequence entire expressing miRNAome. This led to the evolution of the tools which utilize the sRNA-seq reads as guide to support their pre-miRNA models. Tools like miR-PREFeR (13), ShortStack (14), PalGrade (15) or miRDeep (16), belong to this category which has also become the default choice of present days miRNA discovery projects. Compared to totally computational approaches, these methods take additional support from the sRNA sequencing data as guide. But they have their own shortcoming and they are not immune to false identifications. First, their dependence on sRNA-seq data makes them not approachable by all and a costlier stuff as one will have to first manage the sequencing experiments and costs related to that. Secondly, these all methods also utilize the same old features and rules which we described above how they work as poor discriminator. Third, they can capture only those pre-miRNAs which express in any given condition, and miRNA expression is highly specific. Many of their rules are not suitable and may even capture non-miRNAs as well as discard genuine miRNAs.

Though sRNA-seq based approaches have become the default choice to detect miRNAs, advent in new machine learning techniques have improved the core miRNA discovery protocol which are defining new developments. Compared with the conservation-based methods, machine-learning-based methods are mainly anchored on sequence and structure based features of pre-miRNAs with more mature automated statistical learning process, sans the rules based approaches. They have specially emphasized upon the RNA secondary structure and hairpin loops identification. Although, groups like Bentwich *et al* suggested that there are about 11 million hairpins in human genome, making it a daunting task to correctly identify miRNA precursor candidates from hairpins (17).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Based on support vector machines (SVMs), Xue et al. developed triplet-SVM with 32 local structure-sequence features (triplet elements) from known human pre-miRNAs which captured structural as well as sequence information (18). This study marked the machine learning revolution in the area of miRNA biology. This was soon followed by SVM based tools like miPred (19), microPred (20). Methods like Probabilistic co-learning models, naïve Bayes, random forest, and kernel density estimation also came into the scenario to identify pre-miRNAs form pseudo hairpins. Advanced bagging based ensemble method was also introduced with miR-BAG (21). However, these developed methods were specifically designed to predict animal pre-miRNAs but rarely for plant pre-miRNAs. Only Triplet-SVM (18) had been tested on the pre-miRNAs from *A. thaliana and O. sativa*. As the plant pre-miRNAs differ greatly from the animal pre-miRNAs and are more complicated, plant miRNA discovery kept lagging while lots of software were developed for animals. Yet, plant miRNA biology took forward the Machine-learning revolution caused by Triplet-SVM with machine-learning tools like PlantMiRNAPred (SVM based) (4), MiPlantPreMat (SVM based) (22), HuntMi (Random Forest based) (23), and plantMirP (Random forest) (24). It rarely witnessed any major entry of deep-learning based plant pre-miRNA discovery tools until very recently. This has also to be noted that majority of these tools don’t comply with the 2018 criteria for plant microRNA annotation (6).

Lately, deep learning (DL) techniques have been implemented successfully to tasks such as speech and image recognition, largely eliminating the manual construction of features and their engineering. They have been very effectively digging out better but hidden features for model building which are otherwise difficult to detect manually (25). In an excellent benchmarking of machine learning based pre-miRNA discovery tools, it was found that the deep learning methods were constintly performing better than the other machine learning approaches, more so when the data was imbalanced (26), seting the pointer for further developments towards DL based

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

approaches. miRNA biology has very recently witnessed some of them for pre-miRNA discovery, which though not developed exclusively for plants, but can be trained on the plant specific data also. This includes appraoches involving Boltzman machines based deep learning DP-miRNA/deepBN (27), deep learning based self organizing maps (SOM) (28), convolution neural nets (CNN) based miRNA classifier like deepMiR (29), convolutional deep residual networks (30), long-short term memory (LSTM) based pre-miRNA classifier like deepMiRGene (31).

DL approaches like convolutional neural networks (CNN) and recurrent neural networks (RNN), the two dominant types of DNN architectures which have shown huge success in image recognition and natural language processing (NLP) (32, 33, 34). CNN takes input in the form of pixeled matrices where it effectively compresses the features to detect the spatial patterns. Natural language processing approaches like recurrent neural nets (RNN) were developed for processing sequential data (34). RNNs evolved further into long short-term memory (LSTM) approach where a bit distanced associations within a sequence could be detected and memorized.

However, despite the forays into DNN based software to detect pre-miRNAs, there remains lots of voids to be filled. First is the inconsistent performance where huge gaps were found when benchmarked across different datasets. Secondly, most of them are still based on direct reading of the input sequences and work with 4-states nucleic acid sequence inputs and 3 states secondary structure inputs. As Deep RBM based software above showed, increasing features in input boosts performance. Third, barring CNN, LSTM, and RBM deep learning approaches require lots of compute power, time, and resources. LSTM like approach can’t be parallelized and become very high on memory consumption and time to train the network. Further to this, they fail to detect the associations effectively if the sequence length is increased or long distanced associations are

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

present. As already mentioned above, the plant pre-miRNAs differ significantly from animal miRNAs from sequence to structural properties, as can be seen from the Figure 1. Plant pre-miRNAs have much larger sequences than animals, which may complicate LSTM based learning. And above all, the existing software at present still fail miserably in their practical application of genome annotation for miRNAs, as noted by some recent studies (6). Recently, a new revolutionary DL architecture, Transformers, has been introduced in the area of artificial intelligence, which has emerged as a highly efficient architecture for language processing tasks (35). It uses self attention mechanism on the input which can be process parallely while more effectively capturing the long distanced associations and contexts within any sequential data. It has outperformed all DL approaches in the machine learning bench-marking and exhibits high promises for much smarter system development .

Inspired by these milestones in Deep Learning, present work proposes a novel hybrid deep learning based approach, where the transformer defines the first phase whose sequence output from its encoder becomes the input to a shallow learning approach, XGBoost, which takes the final classification decision. By unifying the two ML approaches, further performance improvement and balance was obtained when tested across a large amount of validation datasets. The important aspect is that unlike most of the existing deep-learning approaches, miWords sees a genome sequence as a set of sentences composed of words made from monomers, di-mers, and pentamers capturing sequence based information and communication among themselves. They also capture properties like base stacking structural properties and nucleic acids shape properties. Besides this, genome is also seen as a sentence pool made of words from structural triplets from the RNA secondary structures. Context wise their association including long ranged ones are successfully detected which is otherwise missed by the existing software pool for miRNA discovery. A comprehensive benchmarking study was performed whose results showed that miWords

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

outperformed all the compared software with highly significant margin, just based on this transformer part alone.

The real application of such software in genomic sequence annotation of pre-miRNA discovery has been a challenging part where most of the existing software for pre-miRNA discovery generate lots of wrong classification besides being lethargically slow to scan genomic sequences. The existing software pool hardly considers the long distanced relative standing of the genomic regions to characterize pre-miRNA regions, which is one big reason why they end up producing lots of false positives. We noticed that the regions having miRNAs generate a specific transformer scoring (T-score) pattern when compared to the non-miRNA regions, and this could work as a strong discriminator which could further boost pragmatic miRNA discovery with genomic context information. The T-score generated by the transformer across the genomic region forms a scoring plot which captures the relative standing of the scores for various regions on which a CNN module was trained. This CNN part can scan the regions and their T-scores in a relative and contextual manner for the genome wide T-scoring profile. It remarkably lowers the false identifications which is otherwise hardly seen with any existing approaches. In addition to this CNN module, one more optional CNN module has been provided where the user can supply the sRNA-seq data to further enhance the accuracy. Thus, a user can run miWords with and without sRNA-seq support data, and in both the ways can get highly accurate results.

Considering the 2018 guidelines for miRNA discovery and annotations which asks to prove the performance of such software across well annotated genome like *Arabidopsis*, to measure the degree of false positives in the real application, we also ran miWords across the whole genome of *Arabidopsis* and carried out the genomic annotation performance measure while comparing with other existing tools. miWords outperformed all of them with just 10 false positives. Even without

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

considering sRNA-seq read data, miWords outperformed all those well established software which essentially require sRNA sequencing read data, making it much affordable besides being a better performer. Finally, miWords was applied across the Tea genome whose miRNAs annotation is still not well established. It identified a total of 821 pre-miRNAs across the Tea genome. We selected and experimentally validated 10 of these novel pre-miRNAs identified across the Tea genome. miWords has been made freely available as a source code as well as web-server. We expect miWords to drastically improve the scenario of plant genomes annotations and plant miRNA biology.

Materials and Methods

Datasets Sources

In this study data was retrieved for 27 different plant species. For the 27 plants species, the positive set included experimentally validated pre-miRNAs and negative datasets contained mRNAs, rRNAs, snoRNAs, snRNAs, tRNAs, and other non-coding sequences while following the same protocol which we had followed in animal pre-miRNA discovery tool miRBAG (21). The pre-miRNAs sequences and their co-ordinates were fetched from miRBase version 22 for all available plant species including the model plant species (36). Species which had no genome information were discarded, bringing the total number of species to 27. From Ensembl Plants (v51) and NCBI negative instance sequences were downloaded for same selected 27 plant species. The repeated sequences were filtered out. After that, the same number of sequences were extracted randomly for each species with respect to their species pre-miRNAs number. A combined positive and negative datasets were created for all these 27 plants species to build a model that could act as a universal classifier for plants pre-miRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dataset Generation

Uniform length sequences with flanking regions were obtained for every individual positive instance which varied pre-miRNA-wise. For creating the positive dataset, the central base of the terminal loop was treated as the reference point, a standard protocol (21). This placed the reference point appropriately at constant position, providing uniformity across all possible dataset instances, irrespective of variation in length and loop size. Considering the central base of the terminal loop as the midpoint, genomic sequences up to 200 bp were extracted from the flanking regions.

For the negative dataset creation, the terminal loops were identified in the hairpin structures of the negative instances. As followed above, the central base of the terminal loop was mapped as the reference central position for the creation of every negative instance, as in case of the precursors for reference position identification. The negative dataset consisted of different types of RNA sequences including ribosomal RNA, small nucleolar RNA, small nuclear RNA, transfer RNA, mRNA and long non-coding sequences, all taking pseudo-hairpin shapes. For all the instances, the genomic co-ordinate of central reference position was considered for taking 200 bp sequence with equal flanks. With this approach, a consistent sequence length for positive and negative instances was maintained. These sequences were used to create the training and testing datasets for all 27 species. In the entire study this dataset is known as Dataset “A”.

We additionally constructed an independent dataset, known as Dataset “B”, for another layer of unbiased benchamarking. This dataset is based mainly on the instances considered by other tools considered for their benchmarking purpose while covering 75 plant species. Like Dataset “A”, the positive instances were from the pre-miRNAs available at miRBase. But unlike Dataset “A” which has positive instances from only 27 species whose coordinates are available as discussed earlier,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dataset “B” had also those pre-miRNAs from miRBase for the species whose genomic coordinates/information are presently not available. None of the data came from the miRBase (v22) and also redundant sequence were discarded. Negative instances of this dataset were also built from the dataset for the same for the selected tools. Redundant sequences were dropped. For the negative instances, a total of 92,000 instances were collected for this dataset. Dataset “B” was purely used for objective comparative bench-marking on which all comapred tools were tained and tested on common training and testing datasets.

Encoding of sequence data

The Encoder-Decoder architecture with Transformer has emerged as one of the most effective approaches for the neural machine translation, sequence-to-sequence, and binary classification. The prime importance of the method is the potential to train a single end-to-end model directly on source and target sentences having the capability to handle variable length input and output sequences. However, deep networks like transformers, LSTM, and RNN work by performing computation on integers, passing in a group of words won't work. So these input sequences were tokenized for further computation. Provided a character or word sequence and a defined vocabulary set, tokenization is the procedure of cutting it up into unique numerical units, called tokens. These tokenizers along with processes words from the sentence as input and output a unique numerical representation for the tokenized word which becomes input for the embedding layer of the model. Tokenization followed by embedding layer allowed to vectorize the words into a fixed sized (28 elements each) vector of numeric values. The process of tokenization was implemented using TensorFlow (Keras) Tokenizer class for end-to-end tokenization of the positive and negative datasets. We created the Tokenizer object, providing the maximum number of words as our vocabulary size, which we had in the training data. Tokenizing the data while maping the words to

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unique numeric representation, the vocabulary and words within the genomic sequences were encoded.

To encode sequences, every single instance was converted into possible words i.e. monomeric, dinucleotide, and pentameric sequences in an overlapping window. Dinucleotides and pentamers provide structural stacking and shape information, respectively (37, 38). The secondary structural information of these RNA sequences were also used. This information of structure for each sequence was obtained by RNAfold of ViennaRNA Package v2.4.18 (39). RNAfold predicts secondary structure of RNA and gives output in dot-bracket form (“(“,”.“,”)”). To tokenize the information obtained from RNAfold, notations were transformed using the following: (“(“−>”M”, “.”−>”O”, “N” −>”)”. For a sequence length of 200 bases, a maximum length of the input vector was of 793 elements. The encoded monomers, dinucleotides, pentameric sequences and secondary structural words were independent of each another. In this way, the network would determine the correlations between the sequence derived inputs and the structure triplets based inputs in its own way. The encoded instances were then fed as input the transformer encoder to train and evaluate the models.

Implementation of the hybrid Transformers and XGBoost classification system

With the tokenized sequence, encoded vectors were used to build models to classify and distinguish pre-miRNAs from other genomic elements using a deep-shallow learning approach: The multi-headed attention system of Transformer’s encoder which derives the most confident contextual associations and places them into a hidden space vectors, which becomes the input for the XGBoost part to classify and generate the classification score (T-score). Both were implemented using python scikit-learn, XGBoost, Keras, and Tensorflow libraries. In addition, as per the standard practice, the dataset was broken into 70% and 30% as train and test sets, and using the 70% training part the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

model was built and tested upon the 30% totally untouched unseen testing part. Further, in order to assess the consistency of the approach and its observed performance, 10-fold randomized trials were performed where 10 times the entire data-set was randomly split into 70:30 training and testing data, and new model was built and tested from the scratch every time. Also, this was ensured that absolutely no instance overlaped between the train and test set in any fold. This ensured fair training and testing process without any scope of memorization of instances.

The input layer consists of embedding and position encoding layers which operate on matrices representing a batch of sequence samples. Embedding encodes each word ID (unique token number) into a word vector whose length is the embedding size, resulting in a (samples, sequence length, embedding size) shaped output matrix. For any given genomic sequence of length “*l*” (sentence size) , there are “*n*” words in it. Embedding of these arranged words can be represented as:

Let the sentence S = {w_1, w_2,……, w_n}

where, $w_{\mathbf{n}}$ = n^{th} word in the sentences

Every such word in a sentence is converted into a vector of *d-dimension* whose elements (I_d) carry the optimized numeric weights:

W_n = [I_1, I_2, …., I_d] ;

Therefore, the sentence can be represented in the form of embedded words matrix X:

$$X = \begin{bmatrix} w1 \\ w2 \\ \cdots \\ wn \end{bmatrix} \begin{bmatrix} I_{11}, I_{12}, \ldots, I_{1d} \\ I_{21}, I_{22}, \ldots, I_{2d} \\ \ldots\ldots\ldots \\ I_{n1}, I_{n2}, \ldots, I_{nd} \end{bmatrix};$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where, each row corresponds to the word in "S". This matrix has, thus a dimension of *n x d* (number of words in the sentence x the dimension used to represent each word in embedding vector).

Each word of the matrix "X" is also combined with its corresponding positional embedding "P". The position embedding has same dimension "d" as is for the word embedding vector W_n :

X' = X + P

where, P = [p_1, p_2,….., p_d] and the values of P are derived using the following equations:

$P_d = sine\left(index | 10000^{index/no.of\ dimensions} \right)$ for all the even positions in the vector;

$P_d = \cos\left(index | 10000^{index/no.of\ dimensions} \right)$ for all the odd positions in the vector.

Position encoding produces a similarly shaped matrix that can be added to the embedding matrix. Shape of the matrix (samples, sequence length, embedding size) produced by the embedding and position encoding layers is maintained throughout the Transformer, which is finally reshaped by the final output layers. The input embedding layer sends its outputs into the next layer. Similarly, the output embedding layer feeds into the next layer (encoder layer).

The encoder passes input into a multi-head attention layer. The attention module consists of one or more attention heads. The attention module splits its query, key, and value parameters N-ways and feds each split independently through a different head and then merged together to generate a final attention score. This entire process of attention score generation has five major steps:

Step 1: From the above mentioned input matrix X' derived from the embedded words create the Query matrix (Q), Key matrix (K), and Value matrix (V):

Q = X'.W_O , where W_Q is the optimizable weight matrix for the query matrix generation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

K = X'.W_K , where W_K is the optimizable weight matrix for the key matrix generation.

V = X'.W_V , where W_V is the optimizable weight matrix for the value matrix generation.

All these three weight matrices are randomly initialized.

Step 2: Inner product between Query (Q) and Key (K) transpose matrices: Q.K^T

This step establishes the weights for association between the words within a sentence and captures their dependence.

Step 3: Scale the Query-Key inner product to stabilize the gradients:

$Q.K^T/\sqrt{dimension\ of\ the\ key\ vector}$

Step 4: Normalization through softmax function, which ensures that values are in the range of 0-1:

Softmax($Q.K^T/\sqrt{dimension\ of\ the\ key\ vector}$)

Step 5: Compute the attention matrix "A" to get the attention score for each word in the sentence:

This is achieved by taking inner product of the above mentioned softmax normalized query and key inner product with the Value matrix (V):

Softmax ($Q.K^T/\sqrt{dimension\ of\ the\ key\ vector}$) . V

This is a single column vector which holds the attention score for each positional word and their relative closeness in the given sentence.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For multi-headed attention, the same steps were repeated according to the number of heads and their individual attention scores vectors were finally concatenated and forwarded to the block of feed-forward network of the transformer encoder for further processing. Figure 2 provides a snapshot of how this entire system is working. The output from multi-head attention layer passed into dropout layer which helps to reduce over-fitting, which is followed by a layer of normalization. Afterwards the output from this layer passes to a feed-forward layer, which then sends its output to the next dropout layer in the stack. Another layer of feed-forward layer was implemented gathering its input from the previous layer, followed by a layer of GlobalAveragePooling1D which then passes its output to the third Dropout layer of the model. The Dropout layer passes output into the fully connected hidden layers.

The performance of the Transformer was evaluated for a numbers of hidden layers where finally total two hidden layers were found performing the best and the connections between the nodes were made dense. For the model, the number of nodes across the two hidden layers were tuned. All the component layers were optimized for their suitable numbers and component nodes number by iterative additions. Different activation functions were examined for the layers from a pool of available activation function. The fourth Dropout layer takes into from the last hidden layer and passes its output into an LeakyReLu activation function based single node output classification layer was used with binary cross entropy loss function to calculate the loss. “Adadelta” optimizer was used at this point to adjust the weights and learning rates. Adadelta adapts learning rates based on a moving window of gradient updates, instead of accumulating all past gradients even when many updates have been done. The learning rate was set to 0.583 for the optimizer and the model build was trained using 20 epochs and batch sizes of 40 instances. The transformer part derived the hidden features and their relationships which got structured also and on which classification could be done in much superior manner. Since the present problem in this study was not translation but

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

classification, decoders were not needed and instead the encoder output were taken as input for next step of extreme gradient boosting. For this purpose out output of the transformer was passed to the XGBoost classification part.

Optimizaiton of Tansformer-XGBoost system

A gradient boosting framework, XGBoost, is a decision-tree-based ensemble Machine Learning algorithm which has been consistently rated at the top in shallow learning approaches at Kaggle bench-markings. In the classification phase with XGBoost, grid search was applied for parameter optimization using scikit-learn function RandomizedSearchCV. Following hyper-parameters were optimized with the grid search: "eta/learning rate", "max_depth", "objective", "silent", "base_score", "gamma", "subsample", "eta", "colsample_bytree", "n_estimators", "min_child_weight", eval_metric", "tree_method", "reg_alpha", "reg_lambda". Gradient boosted decision trees learn very quickly and may overfit. To overcome this shrinkage was used which slows down the learning rate of gradient boosting models. Size of the decision tree were run on different combinations of max-depth. Values changed until stability was gained as the logloss got stabilized and did not change thereafter. The final max_depth value was 6.

The final model obtained was saved in hierarchical data format 5 (HDF5). Since the entire system is implemented here using TensorFlow and scikit-learn, the HDF5 format provided the graph definition and weights of the model to the TensorFlow structure and saved the model for classification purpose. Each and every hyper-parameter values involved to finalize this hybrid model were fixed using an in-house developed script which tested various combinations of values of the hyper-parameters to pick the best ones. This entire optimization process was done using two

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

different approaches: Random search optimization and Bayesian optimizations. Figure 2 shows the detailed workflow of the implemented architecture.

Performance Evaluation

The performance of the built model was evaluated. Four classes of the confusion matrix namely true positives (TP) , false negatives (FN), false positives (FP), and true negatives (TN) were evaluated. The performance of the raised transformer based model was assessed the performance metrics like sensitivity, specificity, accuracy, F1-Score, and Mathew correlation coefficient (MCC). Sensitivity/True Positive Rate (TPR) defines the proportion of positives which were correctly identified as positives. The specificity value informs about the proportion of negative instances correctly identified. Precision defines the proportion of positives with respect to total true and false positives. F1-score measures the balance between precision and recall. Besides these metrics, Mathew's Correlation Coefficient (MCC) was also considered. MCC is considered among the best metrics to understand the performance where score equally influenced by all the four confusion matrix classes (true positives, false negatives, true negatives, and false positives) (40). A good MCC score is an indicator of robust and balanced model with high degree of performance consistency. AUC/ROC and mean absolute error were also measured for the build model.

Besides this all, the consistency of performance on the developed approach was evaluated through 10-fold random trials of training and testing. Every time the dataset was randomly split into 70:30 ratio with first used to train and second part used to test, respectively. Every single time data was shuffled and random data was selected for building new model from the scratch. Accuracy and other performance measure were calculated for each such model. In order to avoid any sort of imbalance,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

memory, and bias, it was ensured that no overlap of instances existed ever between the train and test sets.

Performance measures were done using the following equations:

$$\text{Acc}=\frac{\text{TN+TP}}{(\text{TN+TP+FN+FP})}$$

$$\text{Specificity}\,(\text{Sp})=\frac{\text{TN}}{(\text{TN+FP})}$$

$$\text{Precision}=\frac{\text{TP}}{(\text{TP+FP})}$$

$$\text{Sensitivity}\,(\text{Sn})=\frac{\text{TP}}{(\text{TP+FN})}$$

$$F\,1-\text{Score}=2\times\left(\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}\right)$$

$$\text{AUC}=\int_0^1 \text{Pr}\,[\text{TP}]\,(v)\,\text{dv}$$

$$\text{MCC}=\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}$$

Where:

TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, Acc = Accuracy, AUC = Area Under Curve.

CNNs based implementation of genomic scanning capabilities using Transformer-XGBoost scoring profiles

Two different CNN modules were constructed for the identification of pre-miRNAs most potential regions across the genomes. First one is genome wide T-scoring based CNN model and the second

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

one is the optional one which uses Read Per Million (RPM) based CNN further improve the identification of pre-miRNAs across the genomes while using sRNA-seq reads data as guide.

The first module works with the Transformer modules scoring output for every genomic position. Thus, the scanned genomic sequence is transformed into its corresponding transformer scores sequence for each scanned position until the last window frame. It appears like a plot image where the T-scoring around the pre-miRNA regions appear different than the regions not having them. This scores sequence is converted into a one hot encoding acting as an input to a convolution layer. The scoring profile input has a dimension of 280X10. If the length of some instance was found shorter it was padded for the empty columns with value of zero. Size of 280 covers the base positions in a window, for each of which corresponding T-score exists. 10 dimensions come from 10 different categories of T-scoring ranging from 0 to 1. To evaluate the performance of the scoring based CNNs, various number of hidden, convulational, maxpooling, and Batchnormalization layers were tested and finally two convolutional, one maxpooling, four batch normalization, and four hidden layers were applied in fully connected manner. The number of the nodes across both the dense hidden layers were tuned based on the number of filters used in the convolution layer. All the component layers were optimized for their best numbers and component nodes number by iterative additions. Additionally, the kernel size and strides were optimized by trying different values in incremental order. A sigmoid activation function based single node classification layer was used with binary crossentropy loss function to calculate the loss. “Adam” optimizer was used at this point to adjust the weights and learning rates. The batch size was set to 4 and the number of epochs was set to 25.

The second CNN based module is optional as it requires availability of short reads sequence mapping information. It takes RPM value for each base while the input genomic sequence is

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

represented in the form of RPM value sequence. For this RPM based second module RPM profiles were created and transformed later into length of 280 vector, where each element holds the scaled normalized sRNA read depth value for the position of nucleotide in the sequence window. Here also, padding was done for the shorter input instances. Similar to the T-scoring based CNN, the performance evaluation of the RPM based CNNs, various number of hidden, convolutional, and maxpooling layers were tested and finally one convolutional, one maxpooling, two batch normalization, and two fully connected hidden layers were selected. The number of the nodes across both the dense hidden layers, the number of filters used in the convolution layer, the kernel size and strides were optimized by trying different values were tuned to find optimal values. As similar to previous CNN, a sigmoid activation function based single node classification layer was used with binary cross entropy loss function to calculate the loss. “Adam” optimizer was used at this point to adjust the weights and learning rates. The batch size was set to 64 and the number of epochs was set to 25.

To train and test the T-scoring based CNN module, another dataset was created from the known pri-miRNA regions of *Oryza sativa.* The 500 bases from their 5’ and 3’ ends were extracted from the genome, which acted as the negative instances which could help recognize the boundaries and shift towards the corresponding miRNA region. All these sequences were represented as corresponding T-score for each base position. Same was done for the pre-miRNA regions also. As discuused earlier, the sequence which is transformed into its corresponding probability scores appears like a plot image where the scorings around the pre-miRNA regions appear different than the flanking regions not having them. This entire data from *Oryza* was considered as training set, where the total number of positvie instances (pre-miRNA regions) were 604, and the total number of non-pre-miRNA regions (the 500 bases flanking sequences around the pre-miRNAs) were 604. Later,

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

scores were converted into a one hot encoding acting as a input to a convolution layer and had the dimension of 280X10.

Similarily with the RPM based CNN module, the extracted sequences from the genome of *Oryza sativa* were mapped back to the genome to calculate it's Read per million (RPM) value for every single base. A total of 131 different samples covering a total of 42 experimental condition, and a total of four billion sRNA-seq reads (161.GB) were considered for raising the RPM CNN module. The fully annotated *Arabidopsis thaliana* genome version (GCA 000001735.1 TAIR10) was used as the test set to benchmark the performance for genomic annotation.

Optimization of CNNs

In the T-scoring based CNNs model, the scores is converted into a one hot encoding acting as a input to a convolution layer and has a length of 280X10, where the various scoring bins defined the first dimension and the base positions in the length of 280 window defined the second dimension. The input layer was followed by a convolution layer containing 64 channels at each position with 2X2 kernel size. The input sequence was padded if the length was shorter than 280 in order to ensure a constant size of the input matrix. The output resulted into 279X9X64 dimension representation after convolution which passes into a MaxPooling layer. This layer included 32 nodes having kernel size 2×2. Max-pooling helped in reducing the dimensions of convoluted sequence into a dimension of 139×4×64 which is then flattened by the flatten layer. The output from flatten layer passes into first dense layer which is followed by Batch Normalization. This layer was used to overcome the over-fitting problem during training process. Likewise, the output from previous layer passes into second dense layer and then into second Batch Normalization layer. Similarly, the input passes through two more combinations of dense and Batch Normalization. It

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

goes finally into the output layer with 50 dimension The output layers had a node with sigmoid activation function. The model was compiled by binary cross entropy loss function to calculate the loss which was optimized with “Adam” optimizer, for a batch size of four and 25 epochs. The model produced probability score for every instance passed.

In the optional RPM based CNNs model, the scaled normalized RPM values of each base became the input to a 1D convolution layer with dimension of 280X1. The convolution layer had 32 channels at each position with kernel dimension of of 5X1 . The output from the convolutional layer passed into a Max-pooling layer whose output was flattened into a flatten layer. The output from the flattened layer passed into two fully connected dense layers. Later, It was followed by a final output layer with sigmoid activation function. The model was compiled by binary cross entropy loss function to calculate the loss which was optimized by using “Adam” optimizer with batch size 64 and epochs of 25. The model produces probability score for every instance passed.

Performance benchmarking for genomic annotations and application demonstration

To identify the pre-miRNAs on *Arabdiopsis thaliana* and *Camellia sinesis*, we downloaded both the genomes from NCBI (‘GCA 000001735.1 TAIR10’ and ‘GCA 004153795.2 AHAU CSS 2’) for performance benchmarking for genome wide pre-miRNAs annotation and as the application demonstration, respectively. The genomes were scanned by the transformer module through an overlapping sliding window of 200 bases up to n-200th position. The generated position-wise scores sequence was scanned through an overlapping sliding window of 280 elements, where every window becomes the input to the CNN modules as described above.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Arabidopsis genome annotation for miRNAs were obtained from miRBase (v22). Seven different published tool's annotations for *Arabidopsis* (miR-PREFeR, ShortStack, mirDeep-P, miRanalyzer, mirDeep2, mirDeep*, MIReNA) were also considered for the corresponding performance measure for the benchmarking process (5, 6). In addition to this, another reference dataset for *Arabidopsis* was retrieved from the work of Bugnon, et al 2021 where they had benchmarked various tools for miRNA identification recently (41). They focused on the performance of the tools on the real situation data like genomes where class imbalance is pronounced with much higher instances of non-miRNA instances.

Validation of identified pre-miRNAs candidates using sRNA-seq reads

For the validation of the identified pre-miRNAs, the sRNA reads were considered by mapping them to the genome. These sRNA-seq fastq data were collected from GEO and SRA databases which had 88 and 104 different read files for *Arabdiopsis thaliana* and *Camellia sinesis,* respectively (Supplementary Table 1 Sheet 2-3). Genomic sequences, annotations and reference RNA sequences were downloaded from NCBI. Trimmomatic v0.39 (42) and in house developed reads processing tool, filteR (43), were used to filter out poor quality reads, read trimming, and for adapter removal. Filtered reads were mapped back to the genome using Hisat2 (44). Complete list of various conditions and sources is available in Supplementary Table 1 Sheet 2-3 . To remove any bias and noise due to some random elements, two different criteria were applied: (i) Reads which appeared more than five times in any given experiment were only considered, and (ii) the mapping region got support from at least for two different experimental conditions, as suggested by the recent guidelines (45). All these reads were subjected to validation for identified pre-miRNAs across the genomes. The co-ordinates of the short reads were obtained from the mapped results and were intersected it with results obtained from miWords utilizing bedtools (46). Since many miRNAs are also homologous, the identified miRNAs by miWords were searched for homology support using

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

blastn with already reported plant miRNAs in miRBase. Related information about sRNA reads file is provided in Supplementary Table 1 Sheet 2-3.

Server and Standalone Implementation

The entire server was implemented in Apache-Linux platform using PHP. Majority of the codes were developed in python and shell. Statistical processing and calculations were implemented through methods were also executed using modules developed with python. The standalone version was developed in python and shell. The entire work was carried out in Open Source OS environment of Ubuntu linux platforms.

Experimental validation: RNA isolation and quantitative real-time analysis

Total RNA was isolated from tea leaves (Camellia sinensis) from the CSIR-Institute of Himalayan Bioresource technology (32° 05’ 59’’N; 76° 34’ 04’’ E; 1305 m a. s. l) experimental farm (CSIR-IHBT-269) in (47). Approximately 10-12 plants more than eight years old were randomly selected from three sub-populations from the same farm, thus representing three biological replicates for leaf sampling. Samples were harvested in liquid nitrogen and stored at -80 °C for RNA isolation. Total RNA was isolated from leaf tissue (100mg) using Trizol (Invitrogen, USA) according to (48). Total RNA was treated with RNase-free DNase I (Invitrogen, USA) as per as manufacturer’s protocol. The cDNA synthesis was performed using random hexamer and SuperSrcipt® III Reverse Transcriptase (Invitrogen, USA), as per as manufacturer’s protocol. Primers of 10 pre-miRNAs were designed using primer 3 software v.0.4.0; Applied Biosystem (Table 1). Quantitative real time-PCR was performed using the standard protocol on Applied Biosystem, USA. In brief, 2.5 µg of the 1/100 dilution of cDNA with water was added to 5.5µl of SYBR green (Thermo scientific, USA), 2.5nM each primer, and water to 10 µl reaction mixture. Amplification was performed with

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

an initial denaturing at 95 °C for 7 min, followed by 40 cycles of 95 °C for 10s, 53 °C for 30 s, and 72 °C for the 30s. Relative expression of each pre-miRNA was calculated using the equation $2^{-\Delta C_T}$ where $\Delta C_T = (C_{TPre\text{-}miRNA} – C_{T18SrRNA})$ (49). 18S rRNA was taken as an internal control to normalize the variance in cDNA. To simplify the relative presentation expression of each pre-miRNA was multiplied by 10^6. All reactions of qRT-PCR were performed using three biological and three technical replicates.

Results and Discussion

The datasets instances

Presently, there are 8,615 known plant pre-miRNAs in the miRNA database miRBase v22 (http://www.mirbase.org/). After the retrieval of data, only those species sequences were kept which had their respective genome information and those having no information were eliminated from the pool. 5,685 pre-miRNAs belonged to 27 plant species which formed the positive sample dataset. Supplementary Table 1 Sheet 4 shows the pre-miRNAs distribution across the 27 plant species.

To construct the negative dataset, same number of instances were collected from the species selected for their corresponding positive dataset. For only 14 out of 27 species, non-coding RNAs were available at Ensembl plants (v51) from which a total of 5,685 RNAs of different classes were retrieved while eliminating the redundant sequences.

Besides raising the trained model and testing it, 10 times random train-test trails have also been done to evaluate the consistency of the transformer approach. For every such trial, a total of 3,978 plant pre-miRNAs and 3,978 negative instances, totaling 7,956 instances in overall, formed the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

training dataset. A total of 3,412 instances formed the testing dataset. It was ensured that no instances overlapped between train and test sets in order to avoid any chance of bias and memory.

Besides the above mentioned dataset, another Dataset "B" was built for absolutely testing and objective comparative bench-marking purpose. This dataset was built from the datasets used by the different tools like miPlantPreMat (22), PlantMiRNAPred (4), HuntMi (23), and plantMirP (24). A total of 16,404 sequences were retreived, covering 75 plant species. After removing similar sequence, only 9,214 (positive instances) plant pre-miRNAs were obtained. Similarily, 92,000 RNAs of different classes (negative instances) were retrieved after removal of the redundant sequences.

In addition to this, to fathom the performance on the real situation data like genomes where class imbalance is pronounced with much higher instances of non-miRNA instances, dataset provided in the study by Bugnon et al (2021) was also considered (41).

Sentences, Words, and Attention! Seeing genome as a pool of sentences through transformers delivers high accuracy

Most of the existing pre-miRNA discovery tools depend upon some traditionally identified feature sets highly focused on the hair-pin loop structures and sequence composition. They build around properties like Minimum free energy (MFE), stem length, AU/GC content, pairing in stem, terminal loop size etc, most of which are inherited from miPred (20). However, these properties exhibit significant differences between animal and plant system, and within plants themselves, these properties exhibit lots of variations. Figure 1 shows the distribution plots for some of such features to build the pre-miRNA models which exhibit a lot of difference from animals as well as variation among themselves and overlap with other types of RNAs. Also, in most of the existing tools, there

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

is absolutely no effort made to record their relative standing and context which largely limit their practical application when used to annotate genomic sequences (6, 26) .

One of these studies clearly showed how poorly most for the existing software to detect pre-miRNAs perform when they face real situation application of performing genomic annotation. They recommended that compared to the traditional machine-learning approaches, it is the need of the time to focus upon the development of the methods based on DL approaches which may perform better than the other machine learning methods. Considering these seminal works and existing limitations of existing machine learning approaches, the current study proposes a revolutionary transformers deep-learning based approach where context and relative standing of the properties have been emphasized upon to come up with a highly accurate and practical pre-miRNA discovery system, miWords.

For the building of a universal model for plant pre-miRNAs for its classification form other types of RNAs, we used 13 different combinations for various sequence input encodings: 1) Monomers, 2) Dimers, 3) Trimers, 4) Pentamers, 5) Structure triplets, 6) Monomers+Dimers, 7) Monomers+Trimers, 8) Monomers+Dimers+Trimers, 9) Monomers+Dimers+Pentamers, 10) Monomers+Dimers+Pentamers+Triplets, 11) Monomers+Dimers+Trimers+Triplets, and 12) Monomers+Dimers+Trimers+Pentamers+Triplets were evaluated for performance of through the raised transformer encoder based model. An assessment was made for each encoding considered where the Dataset “A” was split into 70:30 ratio to form the train and test dataset. Then, the model was trained with the train set and all the above mentioned properties encodings were done accordingly. This protocol came in action as an ablation analysis to evaluate how each of these individual encodings of the sequence was contributing towards the building of model for accurate

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

classification. The sequences were taken in a uniform length of 200 base while taking the reference from the midpoint of the terminal loop, as described in the methods section.

At the first, the Transformer was trained and tested without the XGBoost gradient boosting to evaluate its performance. First, the test of performance was done with monomeric sequence representation and its corresponding encodings fed into the input layer of the Transformer. The observed accuracy for monomeric encodings was just 72.36% covering a total of 200 words per sentence. This was followed by feeding of dimeric, trimeric, and pentameric sequence representations and their encoded sequences into the input layer of the Transformer. This returned an accuracy of 73.21%, 75.36%, and 79.01%, respectively, while covering a total of 199, 198, and 196 words, respectively.

The reasoning for considering monomeric representation was that they capture sequence composition. While the dinucleotide representation has been proven very useful to reflect the base stacking and secondary structure properties (39, 50). Pentamaric sequences are reflective of the nucleic acids shape which determine protein-nucleic acids interactions (38, 51). All these are critical to characterize a pre-miRNAs. Besides the above mentioned sequence based properties, the secondary structure stem-loop based features were also used for the representation and encoding, as miRNAs exist in the stem-loop hairpin form. RNAfold derived stable secondary structure of the considered region was used for the structural representation. For the sequence encoding the extracted dot-bracket secondary structure of these sequences were converted as following: (“(“−>”M”, “.”−>”O”, “N” −>”)”. These encoding were fed into input layer in the form of triplet words covering a total of 198 words per sentence. This fetched an accuracy of 77.09%. As can be seen here now, individually all these properties did not score much and needed information sharing

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with each other. To obtain encoding of a particular type, the sequence was broken down into sub-sequences in an overlapping fashion to gather all the existing possible combinations which were later converted into encodings.

Above evaluation results showed varying influences on pre-miRNA identification were observed for the different encodings. The next step was observing the influence of combining these sequence and structure derived words encodings and learn on them. Combining of the encodings was done in a gradual manner in order to see the additive effect of them on the classification performance. These combination of encoding yielded a better result than using any single encoding. Combining monomers with dinucleotides (399 words) yielded an accuracy of 81.23% while the combination of monomers+trimers (398 words) yielded an accuracy of 84.06%. In addition, the combination of monomers+dimers+trimers (597 words) and monmers+dimers+pentamers (595 words) achieved the accuracies of 87.63% and 89.32%, respectively. As we know, secondary structure holds critical role in miRNA biogenesis, combining these encodings with structure triplets based encodings led to further superior result. Monomers+dinucleotides+trinucleotides+structure triplets (795 words), Monomers+dinucleotides+trinucleotides+pentanucleotides+structure triplets (991 words), and monomers+dinucleotides+pentamers+structure triplets (793 words) combinations yielded the accuracy of 93.67%, 93.84%, and 93.96%, respectively, with the latter one having the better balance between the sensitivity and specificity values. Thus, combinations of different representations and encoding for the genomic sequence markedly improved the performance through the natural language processing approach of transformers. Figure 3 presents the plots for the accuracy, sensitivity and specificity values distribution observed for the various combinations of the sequence encodings for the classification of pre-miRNAs.

After getting an accuracy of 93.96% from the Transformers built from the combination of monomer+dimer+pentamer+structure triplet encodings delivered a good accuracy of 93.96%. There

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was a gap of 0.9% between sensitivity and specificity, though not a big gap, yet we tried to reduce it further. In doing so, the output layer of the transformer having the LeakyReLu activation function was replaced by XGBoost for the classification purpose. XGBoost was the choice as it has come consistently at the top along with deep learning approaches in Kaggle benchmarkings, and performs exceptionally good on structured data and manually extracted features input sets. In our case, the Transformer's encoder became the feature feeder to XGBoost. This also strengthened the performance further while leveraging from the two best and different approaches of machine learning. This hybrid shallow-deep model reduced the performance gap between the sensitivity and specificity to just 0.46% while also increased the accuracy slightly to 94.08% (Supplementary Table 1 Sheet 5). This became the first part of the transformer based pre-miRNA identification system, which can even work independently and can be used directly for pre-miRNA identification. It was further enhanced for pragmatic genomic scanning and annotation purpose which is discussed in the upcoming sections.

Optimization of the Transformer-XGBoost system

Optimization of the hyper-parameters is an important step to derive the best possible model. The transformer part had encoder role which learned across the hidden space of features and presented it to the classification part done by XGBoost part. The transformer encoders had multi-headed attention layer where a total of 14 self attention heads were found best performing. Multihead attentions help a transformer to avoid misunderstanding the relationships between the words by multiple vetting by the different transformer heads for the derived attention scores. The input sequence was padded if the length was shorter than 200 bases in order to ensure a constant size of the input matrix. This output was passed into a dropout layer with dropout fraction 0.1. By passing - dropout fraction, 10% of the hidden units were randomly dropped during the training process of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

model. This layer helped to reduce over-fitting. Later, this output was normalized by the second layer called normalization layer. This layer was followed by a third layer called Feed Forward layer with 14 nodes and followed by second Dropout layer with dropout fraction of 0.1 and second Feed Forward layer with 14 nodes. The output from Feed forward layer passed into GlobalAveragePooling1D layer, followed by the third dropout layer with dropout fraction of 0.16. Next to this layer, pooled feature maps were passed to two fully connected layer. The hidden layers in the present study had two dense layers with both having 38 and 12 hidden nodes with RELU and SELU activation function, respectively. The hidden layer output was passed into the fourth dropout layer in the stack with dropout fraction of 0.17 which passed its output into the last layer. Finally, the output of the dense layer with 12 dimension was passed to the last and final output layer, a node with LeakyReLu activation function. The model was compiled by binary cross entropy loss function to calculate the loss which was optimized by using the “Adadelta” optimizer with learning rate of 0.583. An accuracy of 94.08% was observed for the test set (Dataset A).

In the classification part of the hybrid Transformer-XGBoost, XGBoost takes input from the second fully connected layer of the Transformers stack. Grid search was applied for hyperparameter optimization using scikit-learn function RandomizedSearchCV. Following hyperparameters were finalized after the grid search: params = {"eta/learning rate": 0.22, "max_depth": 6, "objective": "binary:logistic", "silent": 1, "base_score": np.mean(yt), "gamma": 6.4, "subsample": 0.6, "eta": 0.4, "colsample_bytree": 0.83, "n_estimators": 1400, "min_child_weight": 4.76, "eval_metric": "logloss", "reg_alpha": 149.468151996443, "reg_lambda": 0.02399001301159498, "tree_method": 'approx'}. To overcome over-fitting shrinkage was used which slows down the learning rate of gradient boosting models. At the value of 6, stability was gained as the logloss got stabilized and did not change thereafter. The output from the XGBoost returned the probability score (T-Score) for each input sequence. The probability score indicated the confidence of each instance as non-pre-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

miRNA or pre-miRNA. If the T-Score >0.50, the corresponding input sequence was identified as pre-miRNA else a non-pre-miRNA. The final hyperparameters set for the output layer of the implemented model was: {“Activation function”: LeakyReLu, “Loss function”: binary crossentropy, “Optimizer”: Adadelta}. The related information about optimization towards the final model is listed in Supplementary Table 1 Sheet 6-7 and illustrated in Supplementary Figure S1.

Consistent performance across different validated datasets reinforces miWords as a universal classifier for plant pre-miRNAs

As mentioned in the Methods section, for performance testing two different datasets, “A” and “B”, were created. Dataset “A” had 5,684 positive instances and 5,684 negative instances, totalling 11,368 instances. 70% of this dataset was used for training purpose and 30% was kept aside as a totally unseen test set instances in mutually exclusive manner to ensure unbiased performance testing with no scope for memory from data instances. Besides the Dataset “A”, another Dataset “B” was used for completely testing and objective comparative benchmarking purpose. Dataset “B” covered 75 plant species with 9,214 pre-miRNA and 92,000 non-pre-miRNA sequences from miPlantPreMat (22), PlantMiRNAPred (4), HuntMi (23), and plantMirP (24) tools. Related information is summarized in Supplementary Table 1 Sheet 4. All possible redundancy in the datasets was curtailed and a proportionate representation of negative instances from various classes was maintained to ensure a balanced representation.

When the above mentioned Transformer-XGBoost based model was tested over the experimentally validated instances in the test set (Dataset A) it attained an accuracy of 94.08%. The classifier was able to identify a total of 1,601 true negatives out of a total of 1,706 negative instances, approaching a specificity value of 93.85%. The observed sensitivity was at 94.31%, with a total of 1,609 true

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

positives correctly identified while 97 instances were identified wrongly as false negative instances. Similarly, the MCC values for the classifier also exhibited higher score with a value of 0.8816 (Figure 4A).

In order to fathom the consistency of the observed performance on variable train and test set pairs, 10 folds cross random trials was performed where the train and tests instances from Dataset "A" were selected randomly and mutually exclusive non-overlapping manner. Every time the model was built from the scratch using the training data and was tested upon the corresponding test set. This 10-fold validation trials concurred with the above observed performance level and scored in the same range consistently. Mean absolute error (MAE) was calculated between the train and test set utilizing scikit-learn library. The difference between train and test MAE across 10-fold random validation trials was in the range of 0.009 to 0.0132 which indicates the model was trained well with no significant overfitting. All of them achieved good quality ROC curves with high AUC values in the range of 0.9294 to 0.9436 and maintained reasonable balance between specificity and sensitivity. (Supplementary Table 1 Sheet 8). As emerges from the performance metrics evaluation for the build model and their AUC/ROC plots (Supplementary Figure S2), the developed transformer based pre-miRNA classification approach scored high on performance with consistent and reliable performance.

Integrating Transformer as a trainable feature extractor works better with higher dimensions and instances to learn from. In the input layer combined encodings for sequence and structure derived words were used on which the Transformer block gave remarkable results. The 14 multi-head attention system ensured proper attention to each word while mitigating any chance of wrong weighting and association mapping between the words while taking care of right context in the sentence.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

miWords consistently outperforms all the compared tools for pre-miRNA discovery

This study has performed a series of different comparative benchmarkings. The first two are covered in this section. In this comparative benchmarking, the performance of eight compared software was studied across the Datasets “A” and “B”. The compared tools covered the classical machine learning approaches for pre-miRNA discovery as well as recently developed Deep Learning tools: miPlantPreMat (SVM based), HuntMi (Ensemble method of Random Forest), PlantmirP-Rice (Ensembl method of Random Forest) (52), microPred (SVM based), plantMiRP (SVM based), mirDNN (convolutional deep residual networks), deepMir (CNN based) and deepSOM (deep learning based SOM). Besides this, the benchmarking has also considered two different datasets to carry out a fully unbiased assessment of performance of these tools across different datasets. The first dataset considered was the testing dataset part of Dataset “A”. Besides measuring the performance of miWords of this neutral and totally unseen testing part of Dataset “A”, performance of the other eight tools was also benchmarked. The performance measure on the test set of Dataset “A” gave an idea how the compared algorithms in their existing form perform.

The second dataset “B” was used to carry out objective comparative benchmarking, where each of the compared software was trained as well tested across a common dataset in order to fathom exactly how their learning algorithms differed in their comparative performance.

All these eight software were tested across both the datasets which covered more than 27 plant species considered where miWords outperformed all of them across both the datasets, for all the performance metrics considered (Figure 4A). As already reported above for the Dataset “A” test set, miWords scored the accuracy of 94.08% and the MCC value of 0.88816, while displaying a very good balance between sensitivity and specificity with difference of just 0.4% on Dataset “A”. On

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the same Dataset “A”, the second best performing tool was plantmiRP-Rice which scored and accuracy of 87.89% and MCC of 0.75 only, far behind the values observed for miWords. It displayed a gap of just 0.76 between sensitivity and specificity, a gap which is slightly higher than miWords, but yet a good balance between sensitivity and specificity. A Chi-square test confirmed that miWords significantly outperformed the second best performing tool on Dataset “A” comparative benchmarking (p-value<<0.01).

On the Dataset “B”, all these tools were trained on the same common training dataset and tested across the common testing dataset in order to achieve the objective comparative benchmarking of the algorithms. However, two tools, microPred and miPlantPreMat could not be included in this part of bench-marking as both these tools don’t give provision to train on another dataset and rebuild models. Thus, in this part of benchmarking, the remaining six tools and miWords were trained and tested on the common dataset. In this benchmarking also miWords outperformed all the compared tools with significant margin with the similar level of performance (Figure 4B). miWords clocked an accuracy of 93.6% and MCC of 0.87, while displaying a good balance between sensitivity and specificity where gap of only 0.94% was observed. The second best performing tool was HuntMi which attained an accuracy of 90.5% and MCC value of 0.81 but displayed much higher gap of ~7% between sensitivity and specificity scores, exhibiting significant performance imbalance. A Chi-square test done here also confirmed that miWords significantly outperformed the second best performing tool, HuntMi (p-value<<0.01).

Also this needs to be noted in both the bench-marking tests, miWords scored much higher MCC values which suggests consistent and robust performance. MCC gives high score only when a software scores high on all the four performance parameters (true positive, false positive, true negative, false negative). As it is visible from the score distribution for all the metrics (Figure 4A

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and B), miWords also exhibited least dispersion among all. miWords's performance points out that more appropriate features may be learned through training on syntax of words, and their subsequent efficient encoding with multi-headed attention using Transformer as an encoder and feature extractor combined with gradient boosting for classification on that. The full details and data for the benchmarking studies are given in Supplementary Table 1 Sheet 9-10.

Genomic context learning on transformer scores delivers extremely good results on genome wide annotaitons of pre-miRNAs

Performance over standard testing datasets may be claimed good, as has been done by most of the published software in the past. But in actual application of genome annotation, huge performance gaps exist and far below the acceptable limits. Some recent reports have highlighted that how much poor performing most of the existing pri-miRNA discovery tools become in the real situation applications like genomic annotations where most of them end up reporting very high proportion of false positives (5, 41). This has also led to one of the rare event of mass withdrawal of entries of plant miRNAs from databases like miRBase, recently. Taking note of such extreme events in plant miRNA biology, a very insightful commentary was made by Axtell and Meyers (6). There they recommended some protocols to identify genuine pre/miRNAs candidates and suggested a necessary run against some well studied established genome like *Arabidopsis* to compare how much false positive identifications were made by any plant miRNA discovery tool. It has become the standard protocol to assess the success of such tools in real application of locating pre-miRNA regions across a genome. Most of the existing tools perform highly unreliably in genome annotation. The best performing tools were found to be dependent on sRNA sequencing read data to identify pre/miRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

One important problematic factor about all these existing tools is that they hardly acknowledge the role of relative information from the flanking non-miRNA regions in accurate identification of miRNA regions during genome scanning. The relative scoring patterns between miRNA regions and neighborhood non-miRNA regions can become a highly valuable feature for more accurate discrimination. We hardly found any study which performed genomic scanning and tried to further learn on the scoring patterns for pre-miRNA regions and non pre-miRNA regions. A high scoring pre-miRNA region is expected to display higher scoring distribution across its bases along with gradual decline when compared to its non-miRNA flanking regions where scoring is also expected to exhibit random and sharper trend. Doing so would also help in detecting boundaries of the pre-miRNA regions sharply. A t-test between the flanking regions T-score distribution and pre-miRNA region supported this view (p-value < 0.05). Thus, it became another important aspect of miRNA regions and to refine their discovery in genomic context. Therefore, for the first time we conducted such study and trained another DL CNN based system on the obtained T-scoring profiles in actual run across the genomes.

In this process, the transformer part was run across the fully annotated genome of *Oryza Sativa* where the transformer's scoring patterns was recorded for the annotated pre-miRNAs regions and the non-pre-miRNA regions. This was done with a sliding window of size 200 bases, which generated vectors of transformer scores per base for the pre-miRNA regions and non-miRNA regions. The obtained scoring profiles for pre-miRNA regions constituted the positive datasets while the T-score distribution per base for flanking 500 bases both sides of non-pre-miRNA regions constituted the negative dataset. In the methods section above, full details of implementation for this module is already given. The position specific scoring profiles were converted into a matrix of 280X10 dimension, where the rows contained the scoring values in the range of 0-1 in a discrete manner, while the columns captured the base position for the given window size. One-hot encoding

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was done in this matrix where for any given base position, the corresponding value was assigned value from the range of 0-1. This mimicked a pixeled image which now could be passed through Convolution neural nets (CNN). CNN has been brilliant in recognizing spatial patterns, and we expected them to capture the assumption that miRNAs display a completely different scoring pattern for the bases in its region than those belonging to non-miRNA regions. This way a total of 604 experimentally validated *Oryza sativa* pre-miRNA instances belonged to the positive instances dataset, and a total of 604 non-miRNA regions (comprised of randomly selected 5' and 3' flanking non-miRNA neighbors) belonged to the negative dataset, on which the CNN was trained. This also needs to be mentioned here that in order to ensure a fully unbiased treatment, the *Arabidopsis* training instances were removed from the Transformer's training part and the Transformer-XGBoost model was rebuilt without introducing it to the *Arabidopsis* pre-miRNAs during its training.

Now the real test was to check the raised scoring profile based model's performance was on its ability to correctly identify the miRNAs and non-miRNA regions in some very well annotated and studied genome. For this, the *Arabidospsis* genome was taken with its full annotations. *Arabidopsis* has genome size of 119.763 MB and a total of 326 pre-miRNAs are reported for *Arabdiopsis thaliana* in miRBase version 22 (36). In the first phase, for the entire genome for each base the transformer score (T-score) was generated which became the input to the scoring based CNN as a totally unseen test set.

A total of 323 out of the annotated 326 pre-miRNAs of *Arabidopsis* were detected successfully. The next important question was that how much novel miRNAs were identified across this genome, which could be most probably the false positive cases? A total of 771 pre-miRNAs regions were suggested by the transformers, which meant that a total of 448 false positives were called pre-miRNA regions by the transformers. Though this number is very much lesser than what the

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

currently existing pre-miRNA discovery tools and approaches report (including some NGS sRNA-seq data dependent tools) (26), but even this number may be considered substantially high. However, we got an exceedingly good and surprising result when the transformers scores were passed through the above mentioned scoring profile based CNN. Just 29 novel candidates, the potentially false positive cases, were obtained for the entire *Arabidopsis* genome. This was an exceptionally good result, especially when considering the fact that it was not given any sRNA-seq data guidance. And the existing tools which were run across *Arabidsopsis* genome with sRNA-seq read data supports predicted at least 11 false positive pre-miRNA candidates and went predicting up to 12,306 miRNAs despite having sRNA-seq reads support. Also, in general they grossly missed to identify a big number of actual miRNAs with their sensitivity value ranging from 3% to utmost 86%. Figure 5 provides one such comparative benchmarking map between some of these NGS read data guided software performance on *Arabidopsis* genome and miWords's relative standing. Even without using sRNA sequencing read data unlike these category of software, miWords was found outperforming them by big margin.

Besides this all, one more interesting comparative benchmarking analysis was done on an imbalanced dataset recently provided by Bungan et al, 2021 (41). In their benchmarking study, they knowingly created an imbalanced data with much higher negative instances to mimic the actual genome condition where class imbalance is pronounced. There they strongly attracted the attention on the fact that how most of the existing pre-miRNA discovery software performed very poorely. miWords's performance was compared fot this dataset also, and here too it scored the highest for all the performance metrics with huge lead margin than the rest of the six software (Supplementary Table 1 Sheet 11). Thus, the implications of our developed software is going to be high: It can identify pre-miRNA regions across the genomes highly accurately even without getting any help from sRNA sequencing experiments, directly cutting the cost of experiments, time, and efforts.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

And then naturally comes the next question: How good it would perform if some one provides sRNA sequencing data too? miWords has implemented another CNN module which takes input in just 1-D vector manner. The provided sRNA-seq reads data are mapped across the genome and every genomic position can be expressed in the form of RPM-value. The 1-D vector holds the RPM-value. Details of this module is already provided in the methods section above and Supplementary Figure S3. On passing through this optional and second CNN based on RPM values of the bases derived from the sRNA-seq data, the false positive identification further decreased to just 10 cases, a number much lesser than the number of 28 false positive cases reported by best performing software, miR-PREFeR after taking support from sRNA-seq data. Figure 5 provides the comparative benchmarking of miWords with seven best performing tools which use sRNA-seq data, clearly suggesting the top notch performance by miWords.

This also may be noted that in a previous bench-marking study on these compared software, it was found that they are sensitive towards the size of sRNA-seq data and number of studies included. As this data volume and number of studies increases, the number of potential false positives by these software was reported to increase also (13). The reported total number of total novel 28 miRNAs by the best performing tool, miR-PREFeR, was when only two experimental conditions sRNA-seq data were considered. As soon as they considered higher number of samples (6), their reported novel miRNAs number shot up to 49. The same trend was observed for almost all of the compared tools with much higher offshooting. In the present study we had considered comparatively much bigger sRNA-seq data for *Arabidopsis*, a total of 88 samples, and yet did not see such overshooting effect and reported only 10 novel pre-miRNAs. Even without sRNA-seq data support, just based on the Transformer scoring profile based CNN, it outperformed all of the compared software which needed sRNA-seq data.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Thus, miWords emerged not only as the best performing software, but has exceptionally exceeding performance with much stability and reliability. It has emerged as the most suitable software to annotate plant genomes for pre-miRNA/miRNA regions.

Application of miWords on *C. sinensis* genome and experimental validation of the identified pre-miRNAs

To exhibit the applicability of miWords in real scenario of genome scanning for pre-miRNA discovery, miWords was run across the *C. sinensis* genome whose size is 3.06 GB. Tea is the most consumed beverage, a highly important commercial crop with medicinal values also, and which is also highly sensitive to climate change. Though its genome has been revealed, to this date there is no entries for Tea miRNAs in miRBase. Thus, reporting miRNAs of Tea here would not only exhibit the application demonstration of miWords in genome annotations for miRNAs, but also benefit the research groups working on Tea to understand its molecular systems.

The first run of miWords, which was the transformer part, identified 17,044 pre-miRNA regions in the tea genome. The scoring for all these regions were passed to the T-scoring profile based CNN module, which screened them and reported a total of 3,194 pre-miRNA candidates. Finally, the sRNA-seq data supported RPM-CNN module was run on it, which reported a total of 821 pre-miRNA candidates in Tea (Figure 6). For this part of run, we had collected sRNA-seq reads from 104 samples while covering 34 different conditions. This is not necessary that all other discarded pre-miRNA candidates reported by the previous step were false positive, as sRNA-seq data are highly condition specific, and many conditions may not have been captured by the existing sRNA-seq experiments done on Tea. Yet, these 821 pre-miRNAs from Tea genome may be considered as the most confident cases. All of these 821 potential pre-miRNA candidates exhibited sRNA-reads

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

data mapping in multiple samples, with at least five reads mapping in each condition. All this gave strong support evidence to the identified pre-miRNAs as per the 2018 guidelines also.

The final leg of this study belonged to the experimental validation of some of these identified pre-miRNA candidates using quantitative qRT-PCR experiments on 10 such candidates. The cDNA of tea tissue was amplified by qRT-PCR using primers for these 10 pre-miRNAs. Moderate to strong signals were generated for most of the pre-miRNAs (Figure 7). Based on qRT-PCR data, csi-MIR 099 had very strong expression, csi-MIR018, csi-MIR454, csi-MIR582, and csi-MIR646 displayed moderate expression, whereas csi-MIR386, csi-MIR615, and csi-MIR569 had the lowest expression. Interestingly, two sequences namely, csi-MIR329 and csi-MIR696 showed almost negligible expression. This is possible that these two pre-miRNA candidates express themselves more in some other conditions. Thus, this experimental validation exercise supported the findings made by miWords across the Tea genome. Complete Tea miRNAome details are given in the Supplementary Table 1 Sheet 12. The structure and reads information of the experimentally validated Tea pre-miRNAs is also given there in the sheet.

Webserver and standalone implementation

miWords has been made freely available at https://scbb.ihbt.res.in/miWords/ as a very simple to use webserver. The server has been implemented using D3JS visualization library, Python, Javascript, PHP, and HTML5. The user needs to paste the RNA sequences in FASTA format into the text box or upload the RNA sequences in FASTA format and then click the submit button for the identification. After a while the result page appears from where results can be downloaded in a tabular format and sequence wise distribution of probability score in plots is also displayed on the result page.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

However, in actual situation like whole genome sequence scanning, web servers a practically not suitable and one needs standalone version mainly. For heavy duty real scenario application like genome annotation, a standalone version of miWords has also be provided via a link tab on the same server page or can be downloaded from github. The provision to see the results in plots is also given there.

Conclusion

miRNAs define one of the largest regulatory systems of eukaryotes. The mature miRNAs are processed out of their precursors (pre-miRNAs). Though lots of software have been developed to identify pre-miRNAs as their discovery is the core of miRNA biology, in actual practical application of discovery of pre-miRNAs across genomes these software remain far away from the acceptable limits. This is much more grave when dealing with plant genomes where the generally considered properties and features to distinguish a pre-miRNA regions don't work as a strong discriminator. The present work has approached pre-miRNAs as a sentence hidden within genome where the relative arrangements of the words in the form of dinucleotides, trimers, pentamers, and structural triplets define a sentence. Once the syntax is there, one also needs they are read by an intelligent reader which could decode the relationship within the words. This was achieved by a revolutionary deep-learning algorithm of transformers which assigned contextual attention scores to the words withing the sentence using 14 multi-headed transformers which finally passed their learning to an XGBoost classifier to generate classification scores. The next part of this system, miWords, applied a convolution networks system which learned from the transformer-scoring pattern across the genome and partitioned it into the pre-miRNA and non-pre-miRNA vicinity regions to successfully define the boundaries and make it possible to run such software for its practical utilities for genomic annotation. Total four direct and different benchmarking studies were

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

carried out in this study involving more than 10 different published software to identify miRNAs, and in all of them miWords significantly outperformed the compared tools. This also included the class of software which are prime choice for genomic annotation for miRNAs, the Next-gen sequencing reads data guided software. Without using NGS reads guidance, miWords outperformed even that class of software, making it the most suitable and accessible software for genome annotation for miRNAs which can work with much higher accuracy than others even without cost and time on running NGS experiments. Additionally, miWords, also provides an optional module to use sRNA sequencing reads data to further refine the results. As an application demonstration, miWords was run across the Tea genome no identify pre-miRNAs across its genome. A total of 821 pre-miRNAs were identified in the Tea genome, and for all of them sRNA-seq reads gave evidence support. A randomly selected sample of 10 such pre-miRNA candidates were taken for experimental validation, eight of which were significantly expressed and while remaining two gave very low expression values. This experiment validated the approach of miWords and its capabilities to annotate genomes for miRNAs. miWords appears to be the most capable tool which can solve the long pending quest for a software which could be reliably used for genomic annotation for miRNAs.

Declarations

Availability of data and materials

All the secondary data used in present study were publicly available and their due references and sources have been provided in the Supplementary Table 1. Supplementary data files provided contain all data and information generated/used, methodology related details etc have been made

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

available. The software has also been made available at Github at https://github.com/SCBB-LAB/miWords as well as at https://scbb.ihbt.res.in/miWords/.

Funding

RS is thankful to Council of Scientific and Industrial Research and Department of Biotechnology, Govt. of India, for supporting this study through grants: iPRESS [Grant number:34/1/TD-AgriNutriBiotech/NCP-FBR 2020-RPPBDD-TMD-SeMI(MLP-0156)] and HiCHiCoB [Grant number:BT/PR40122/BTIS/137/30/2021], to RS.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SG carried out the computational part and benchmarking of the study. RS conceptualized, designed, analyzed and supervised the entire study. VS and RK carried out the wet lab experiment. RK supervised the wet lab experiment. SG and RS wrote the MS.

Acknowledgments

The work was carried out under the aegis of The Himalayan Centre for High-throughput Computational Biology (HiCHiCoB), a BIC supported by DBT, Govt. Of India. We are thankful to

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CSIR for the funding support they gave for this project. SG is thankful to CSIR and DBT, India for financial support as project associateship. SG and VS are thankful to Academy of Scientific and Innovative Research (AcSIR). This MS has CSIR-IHBT MSID #NA.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1 Fabian,M.R., Sonenberg,N. and Filipowicz,W. (2010) Regulation of mRNA Translation and Stability by microRNAs. Annual Review of Biochemistry, 79, 351–379.

2 Winter,J., Jung,S., Keller,S., Gregory,R.I. and Diederichs,S. (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol, 11, 228–234.

3 Smalheiser,N.R. (2003) EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. Genome Biol, 4, 403.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4 Xuan,P., Guo,M., Liu,X., Huang,Y., Li,W. and Huang,Y. (2011) PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics, 27, 1368–1376.

5 Taylor,R.S., Tarver,J.E., Foroozani,A. and Donoghue,P.C.J. (2017) MicroRNA annotation of plant genomes − Do it right or not at all. BioEssays, 39, 1600113.

6 Axtell,M.J. and Meyers,B.C. (2018) Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. The Plant Cell, 30, 272–284.

7 Bonnet,E., Wuyts,J., Rouzé,P. and Peer,Y.V. de (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. PNAS, 101, 11511–11516.

8 Dezulian,T., Remmert,M., Palatnik,J.F., Weigel,D. and Huson,D.H. (2006) Identification of plant microRNA homologs. Bioinformatics, 22, 359–360.

9 Hertel,J. and Stadler,P.F. (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics, 22, e197–e202.

10 Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA. Molecular Cell, 14, 787–799.

11 Adai,A., Johnson,C., Mlotshwa,S., Archer-Evans,S., Manocha,V., Vance,V. and Sundaresan,V. (2005) Computational prediction of miRNAs in Arabidopsis thaliana. Genome Res., 15, 78–91.

12 Lindow,M. and Krogh,A. (2005) Computational evidence for hundreds of non-conserved plant microRNAs. BMC Genomics, 6, 119.

13 Lei,J. and Sun,Y. (2014) miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. Bioinformatics, 30, 2837–2839.

14 Axtell,M.J. (2013) ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA, 19, 740–751.

15 Doran,J. and Strauss,W.M. (2007) Bio-Informatic Trends for the Determination of miRNA–Target Interactions in Mammals. DNA and Cell Biology, 26, 353–360.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16 Friedländer,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol, 26, 407–415.

17 Bentwich,I., Avniel,A., Karov,Y., Aharonov,R., Gilad,S., Barad,O., Barzilai,A., Einat,P., Einav,U., Meiri,E., et al. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet, 37, 766–770.

18 Xue,C., Li,F., He,T., Liu,G.-P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics, 6, 310.

19 Ng,K.L.S. and Mishra,S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics, 23, 1321–1330.

20 Batuwita,R. and Palade,V. (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics, 25, 989–995.

21 Jha,A., Chauhan,R., Mehra,M., Singh,H.R. and Shankar,R. (2012) miR-BAG: Bagging Based Identification of MicroRNA Precursors. *PLOS ONE*, **7**, e45782.

22 Meng,J., Liu,D., Sun,C. and Luan,Y. (2014) Prediction of plant pre-microRNAs and their microRNAs in genome-scale sequences using structure-sequence features and support vector machine. BMC Bioinformatics, 15, 423.

23 Gudyś,A., Szcześniak,M.W., Sikora,M. and Makałowska,I. (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinformatics, 14, 83.

24 Yao,Y., Ma,C., Deng,H., Liu,Q., Zhang,J. and Yi,M. (2016) plantMirP: an efficient computational program for the prediction of plant pre-miRNA by incorporating knowledge-based energy features. Mol. BioSyst., 12, 3124–3131.

25 LeCun,Y., Bengio,Y., and Hinton,G. (2015) Deep learning. Nature, 521, 436-444.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26 Stegmayer,G., Di Persia,L.E., Rubiolo,M., Gerard,M., Pividori,M., Yones,C., Bugnon,L.A., Rodriguez,T., Raad,J. and Milone,D.H. (2019) Predicting novel microRNA: a comprehensive comparison of machine learning approaches. Briefings in Bioinformatics, 20, 1607–1620.

27 Thomas,J., Thomas,S. and Sael,L. (2017) DP-miRNA: An improved prediction of precursor microRNA using deep learning model. In 2017 IEEE International Conference on Big Data and Smart Computing (BigComp).pp. 96–99.

28 Stegmayer,G., Yones,C., Kamenetzky,L. and Milone,D.H. (2017) High Class-Imbalance in pre-miRNA Prediction: A Novel Approach Based on deepSOM. IEEE/ACM Trans Comput Biol Bioinform, 14, 1316–1326.

29 Tang,X. and Sun,Y. (2019) Fast and accurate microRNA search using CNN. BMC Bioinformatics, 20, 646.

30 Yones,C., Raad,J., Bugnon,L.A., Milone,D.H. and Stegmayer,G. (2021) High precision in microRNA prediction: A novel genome-wide approach with convolutional deep residual networks. Comput Biol Med, 134, 104448.

31 Park,S., Min,S., Choi,H.-S. and Yoon,S. (2017) Deep Recurrent Neural Network-Based Identification of Precursor microRNAs. In Advances in Neural Information Processing Systems. Curran Associates, Inc., Vol. 30.

32 Krizhevsky,A., Sutskever,I. and Hinton,G.E. (2017) ImageNet classification with deep convolutional neural networks. Commun. ACM, 60, 84–90.

33 Vieira,J.P.A. and Moura,R.S. (2017) An analysis of convolutional neural networks for sentence classification. *2017 XLIII Latin American Computer Conference (CLEI)*, 10.1109/CLEI.2017.8226381.

34 Mandic,D. and Chambers,J. (2001) Recurrent neural networks for prediction: learning algorithms, architectures and stability Wiley, Chichester.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

35 Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,Ł. and Polosukhin,I. (2017) Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vol. 30.

36 Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. Nucleic Acids Research, 47, D155–D162.

37 Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res, 41, W56–W62.

38 Heikham,R. and Shankar,R. (2010) Flanking region sequence information to refine microRNA target predictions. J Biosci, 35, 105–118.

39 Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**, 3429–3431.

40 Chicco,D. and Jurman,G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21**, 6.

41 Bugnon,L.A., Yones,C., Milone,D.H. and Stegmayer,G. (2021) Genome-wide discovery of pre-miRNAs: comparison of recent approaches based on machine learning. Briefings in Bioinformatics, 22, bbaa184.

42 Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

43 Gahlan,P., Singh,H.R., Shankar,R., Sharma,N., Kumari,A., Chawla,V., Ahuja,P.S. and Kumar,S. (2012) De novo sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major transcriptome adjustments. *BMC Genomics*, **13**, 126.

44 Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, **12**, 357–360.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

45 Jha,A., Panzade,G., Pandey,R. and Shankar,R. (2015) A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Research*, **43**, 8713–8724.

46 Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

47 Kumari,M., Thakur,S., Kumar,A., Joshi,R., Kumar,P., Shankar,R. and Kumar,R. (2019) Regulation of color transition in purple tea (Camellia sinensis). Planta, 251, 35.

48 Rosas-Cárdenas,F. de F., Durán-Figueroa,N., Vielle-Calzada,J.-P., Cruz-Hernández,A., Marsch-Martínez,N. and de Folter,S. (2011) A simple and efficient method for isolating small RNAs from different plant species. Plant Methods, 7, 4.

49 Schmittgen,T.D., Jiang,J., Liu,Q. and Yang,L. (2004) A high-throughput method to monitor the expression of microRNA precursors. Nucleic Acids Res, 32, e43.

50 Černý,J., Božíková,P., Svoboda,J. and Schneider,B. (2020) A unified dinucleotide alphabet describing both RNA and DNA structures. Nucleic Acids Research, 48, 6367–6381.

51 Sharma,N.K., Gupta,S., Kumar,A., Kumar,P., Pradhan,U.K. and Shankar,R. (2021) RBPSpot: Learning on appropriate contextual information for RBP binding sites discovery. IScience, 24, 103381.

52 Zhang,H., Wang,H., Yao,Y. and Yi,M. (2020) PlantMirP-Rice: An Efficient Program for Rice Pre-miRNA Prediction. Genes (Basel), 11, 662.

Tables

Table 1: Primer list for 10 selectedpPre-miRNAs from Tea genome taken for validation through qPCR.

Pre-mi RNA sequences	**Primers (5’-3’)**
csi-MIR018 FP	CGATGTGGCTGCAAATATGA
csi-MIR018 RP	TTCCGACTCCGATTCCACTA
csi-MIR099 FP	CAAAATGTAAGGGTGCAAAGTG
csi-MIR099 RP	ACAACCGCTAATGCCCTAAC

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

csi-MIR386 FP	CAGCAGAACCGGAGGAGTAA
csi-MIR386 RP	ATCACCCAAATCGGATCTCA
csi-MIR454 FP	TCATGTGAGTATGCTTCCGGA
csi-MIR454 RP	AACCGGTCTCTCCTCATACG
csi-MIR569 FP	CAAACTTTGAGACAGTGTAAGCAA
csi-MIR569 RP	GAAAAATTCGGAAGAAGAAGACA
csi-MIR582 FP	CTTCCGAACCCTCTTCTGTG
csi-MIR582 RP	TAAAAGCCGGAGCAATAGGA
csi-MIR615 FP	TCAAACAAGGCCTAAGTGTTC
csi-MIR615 RP	CCCTTCCTAGGTTAGACTTTTT
csi-MIR329 FP	CAAATGGTGCCACGCAAAT
csi-MIR329 RP	TGTGTGGTGGTAGTAGCATACAAT
csi-MIR646 FP	AACTCACCGCAAACATAGGC
csi-MIR646 RP	GACCCTAAATCCTCTGAAGGTG
csi-MIR696 FP	GAGTTGTTGGCCAGGTTCTG
csi-MIR696 RP	TGGCCTACACTGATACTTTCTCT
18sRNAFP	ACACCCTGGGAATTGGTTT
18sRNARP	GTATGCGCCAATAAGACCAC

Figure legends

Figure 1: Distribution pattern of traditionally considered properties for miRNA characterization. A) Pattern of distribution comparison between animals and plants pre-miRNAs. Values differ a lot between animals and plants as unlike animal pre-miRNAs, plants pre-miRNAs display much more complexity and variability. **B)** Pattern of distribution comparison between pre-miRNAs v/s other RNAs in plants. As can be seen clearly that most of these properties are actually not good descriminators as lots of overlap of their values occur between pre-miRNAs and other RNAs.

Figure 2: Detailed pipeline of the workflow. The image provides the brief outline of the entire computation protocol implemented to develop the Transformer-XGBoost based model to identify pre-miRNAs. This illustrates how a genomic sequence can be seen as a sentence composed of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

words and their related arrangements which can be efficiently learned through multi-headed transformers. The various nuclotides k-mers and RNA secondary structure triplets define the words for any given regions (the sentence). The words and their attention scores are evaluated through query, key, and value matrices which are then passed to different layers of deep-learning protocol to present its learning for classification job through XGBoost.

Figure 3: Ablation analysis for five main properties in discriminating between the negative and positive instances. Impact of combination of the monomer, dimer, trimer, pentamers, and structure triplet properties based sequence encodings. These encodings appeared highly additive and complementary to each other as the performance in accurately identifying pre-miRNAs increased substantially as they combined together.

Figure 4: Comparative bench-marking results for miWords for two different datasets. (A) Bechmarking result on Dataset A. Here all the comapred tools were tested on the testing part of the Dataset A which was totally unseen and untouched for all the compared tools including miWords. This gives a view of how the compared software would behave in their existing form and models. **(B)** Objective Comparative Benchmarking on Dataset B. Here, all the compared tools were first trained on a common dataset for training and then tested on a common mutually exclusive dataset for their performance. This gave a clear view on the performance of each of the compared algorithms. From the plots it is clearly visible that for all these datasets, miWords consistently and significantly outperformed the compared tools for all the compared metrics.

Figure 5: Comparative benchmarking for genomic annotation capability and performance. The 2018 miRNAs discovery guidelines (6) noted that most of the existing software fail to perform even reasonably for their actual application for genomic annotations and report a huge number of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

false positives, while on standard bench-marking datasets they claim high accuracy. In order to assess how much prone a software is towards making false positive claims, they should be benchmarked against well annotated genomes like *Arabidopsis* which is now not expected to have any newer miRNAs. Any reporting of novel miRNAs on such genome should be considered as a false positive case and accordingly the performance of a software may be rated. In this performance bench-marking, miWords was compared to the tools which are most preferred ones for genomic annotation at the present, as they use sRNA sequencing reads data as help guide to reduce their false positive predictions. As can be seen from this bench-marking plot, miWords with sRNA-seq reads support (miWords-R) as well as without it (miWords-T) outperforms all the compared tools for all these performance metrics. In *Arabidopsis*, miWords identified all of its pre-miRNAs correctly except three of them, and reported only 10 false positives, the lowest of all.

Figure 6: Details of the workflow carried out to annotate the Tea genome for pre-miRNAs. A total of 821 pre-miRNAs were identified in *C. sinensis*. This workflow also illustrate how the transformer-scoring (T-score) can be utilized by the next CNN modules to refine the results while learning over the genomic context of the scoring pattern for miRNA and neighboring non-miRNA regions. It also shows how using sRNA-seq reads based CNN module can help further refine the result.

Figure 7: Experimental validation of the identified pre-miRNAs in *C. sinensis*. miRNA expression analysis of selected 10 pre-miRNAs by quantitative real time-PCR in tea leaves (*Camellia sinensis*). Pre-miRNAs expression is presented relative to 18S rRNA. Mean ± SD of triplicate quantitative real-time PCR from a single cDNA sample.

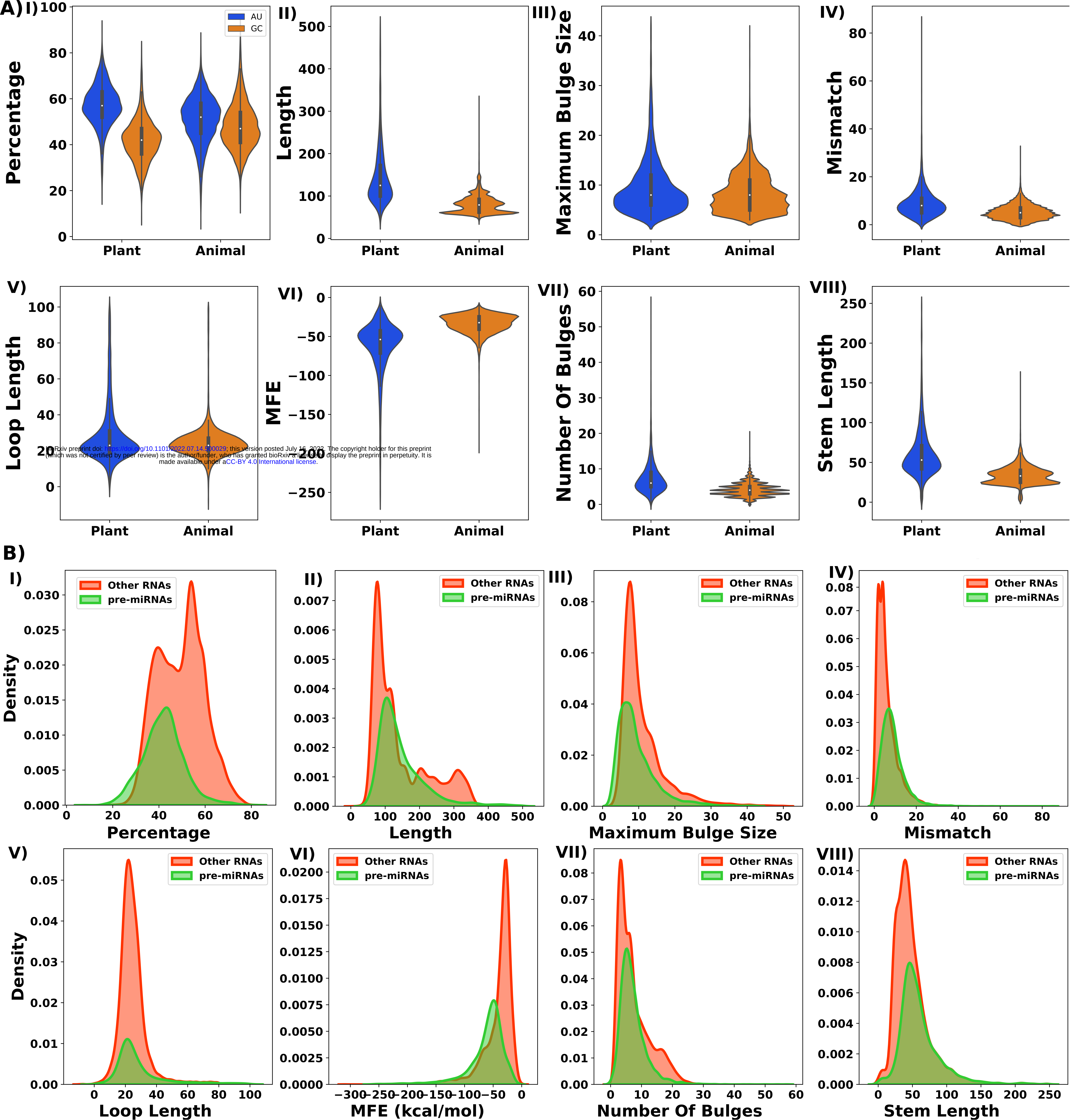
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary information

Supplementary Figure S1: Optimization results for hyperparameters for transformers part of the hybrid Transformer-XGBoost model. A) Batch size optimization, **B)** Dropout rate optimization, **C)** Second Dropout rate optimization, **D)** Learning rate, **E)** Number of units per dense layer 1, **F)** Epoch size optimization **G)** Embedding size, **H)** Number of units per dense layer inside Transformer, I) Number of dense layers 2, and J) Number of Attention heads.

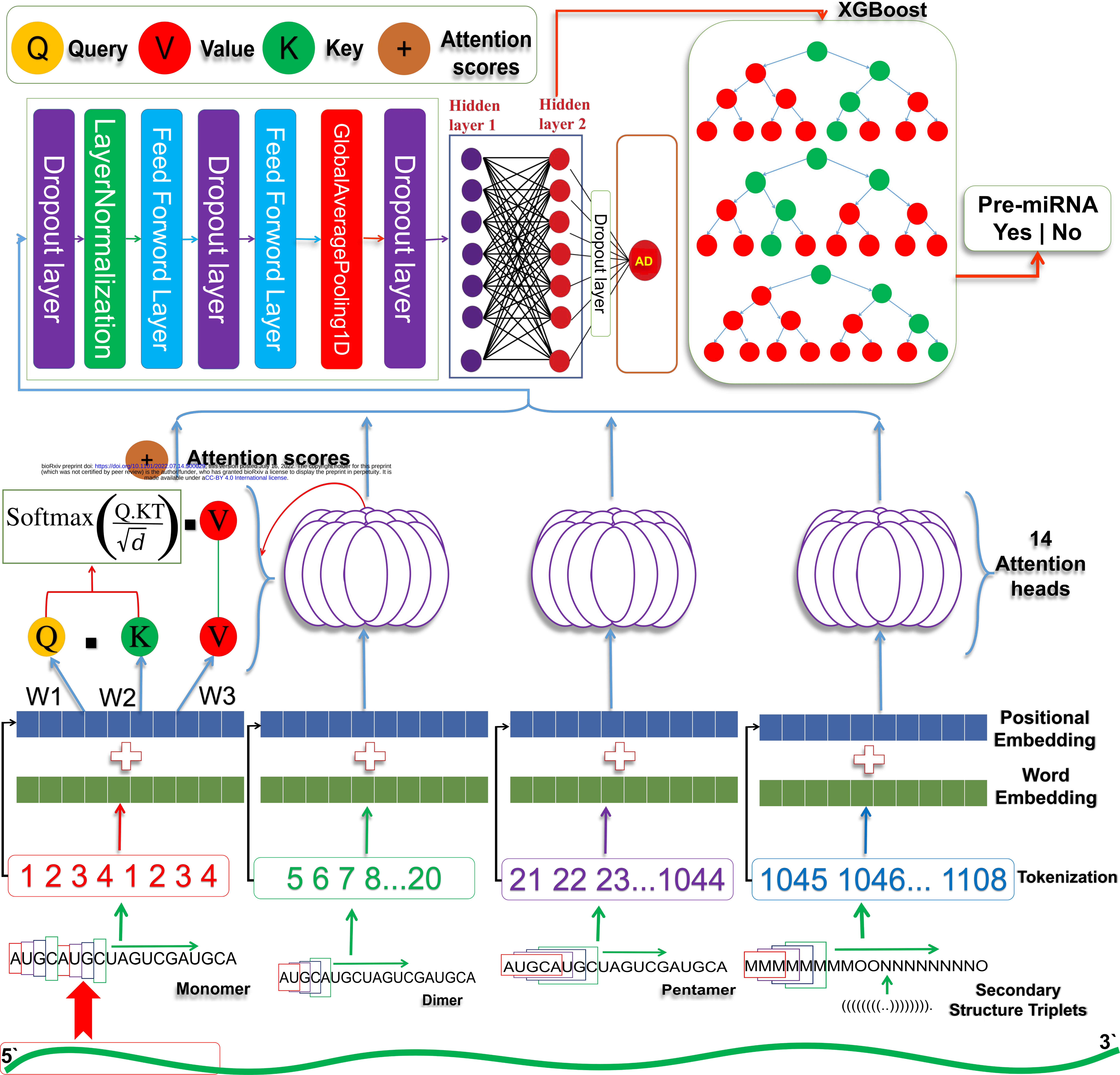
Supplementary Figure S2: AUC/ROC plot for Ten fold cross validation. The AUC/ROC plots forthe hybrid models for the testset clearly showcase the robustness and highly reliable performance of the implemented hybrid Transformer-XGBoost model.

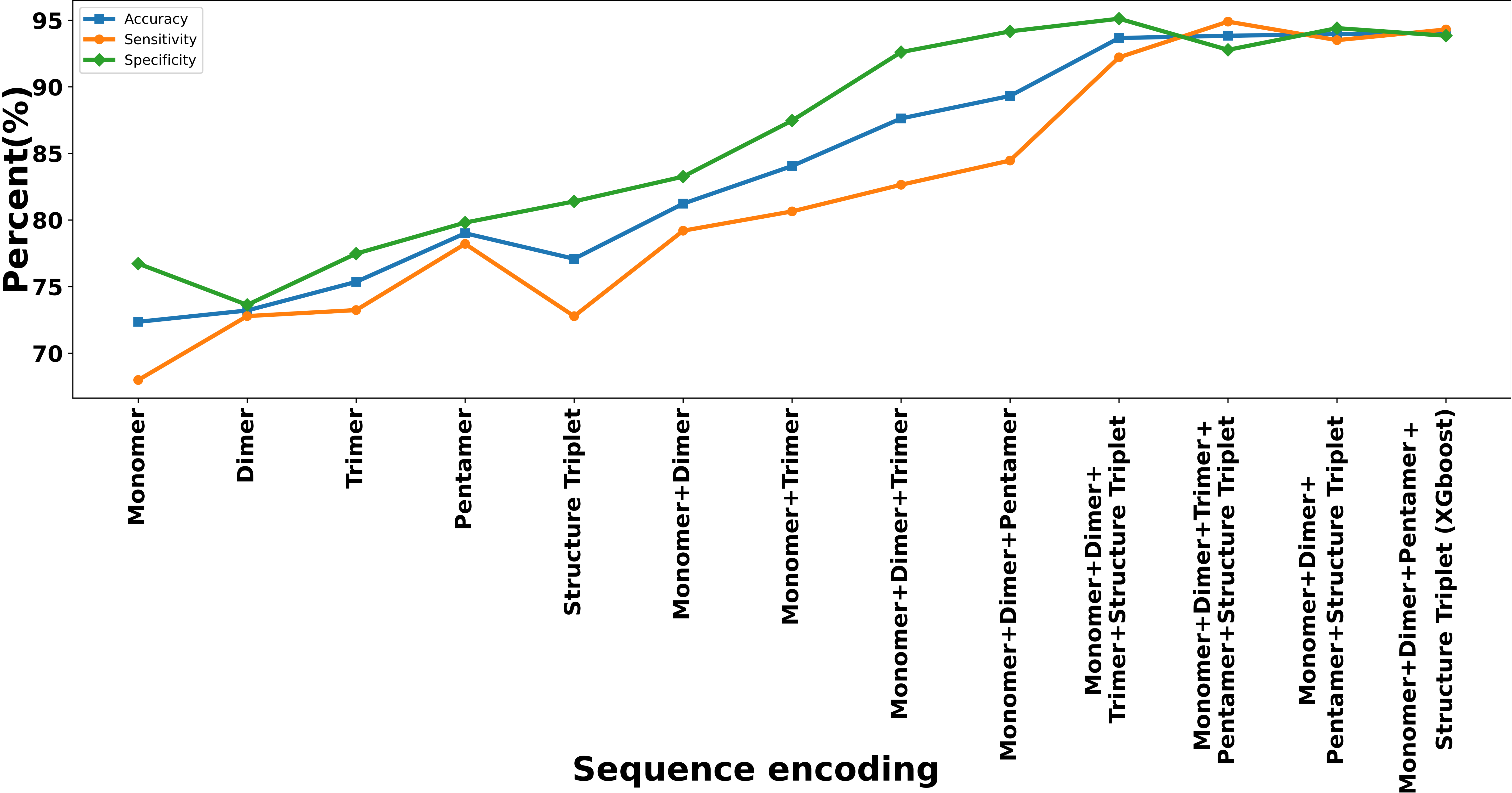
Supplementary Figure S3: Detailed pipeline of the scoring and RPM profile. A) The image provides the outline of CNN architecture implemented for scoring based profiles for better classification of pre-miRNAs. **B)** Architecture of RPM based CNN model implemented second level classification of pre-miRNAs.

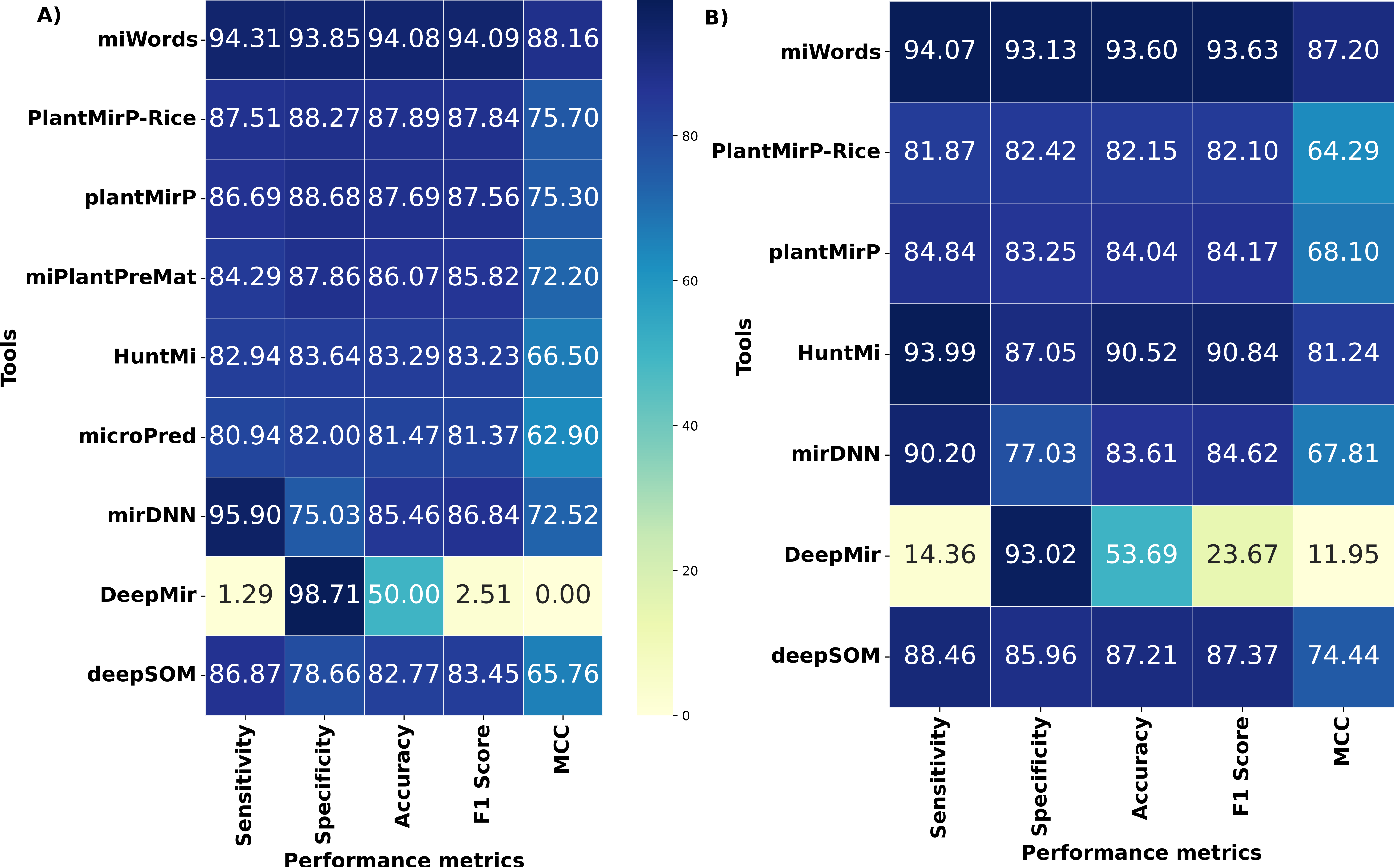


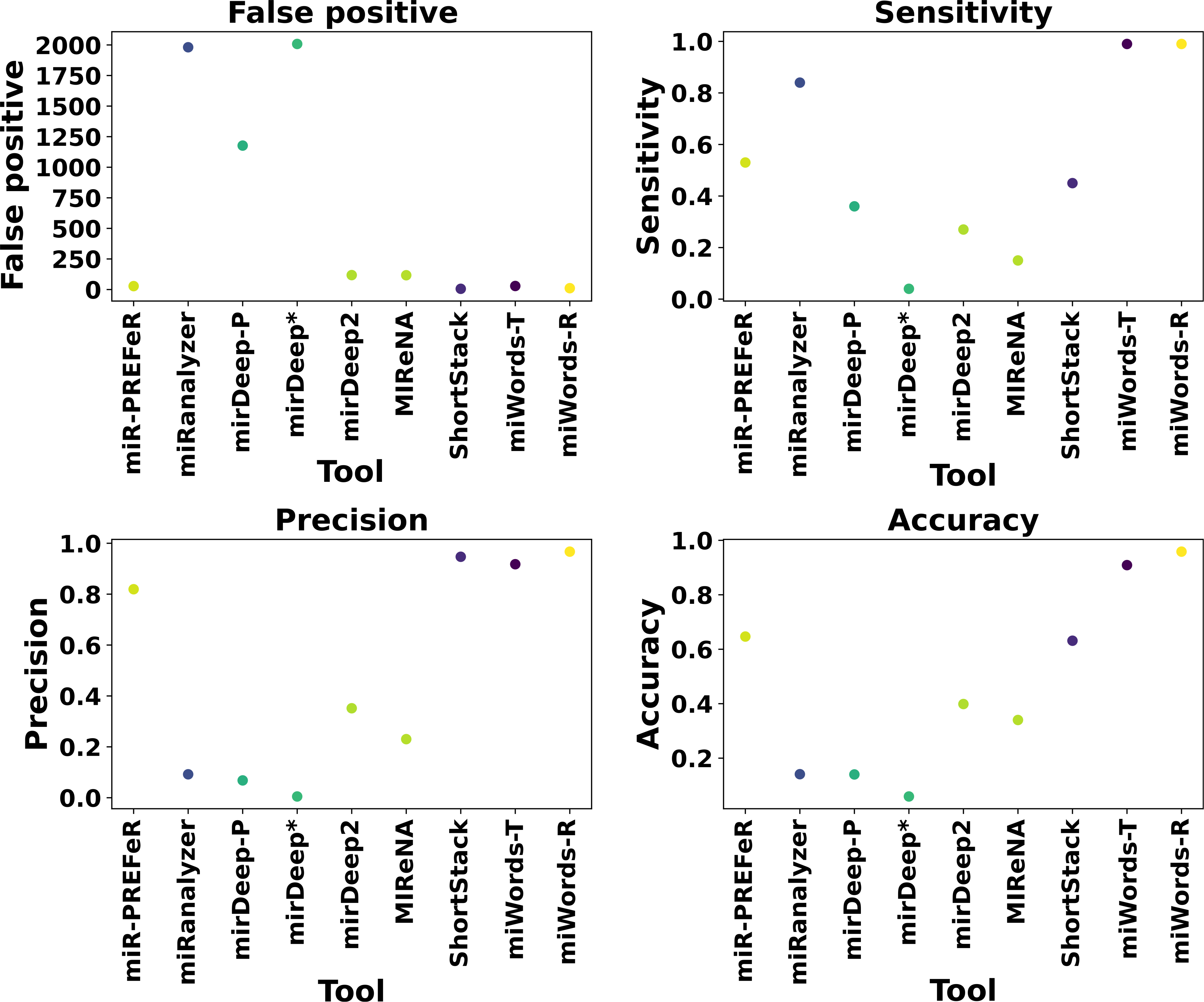
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

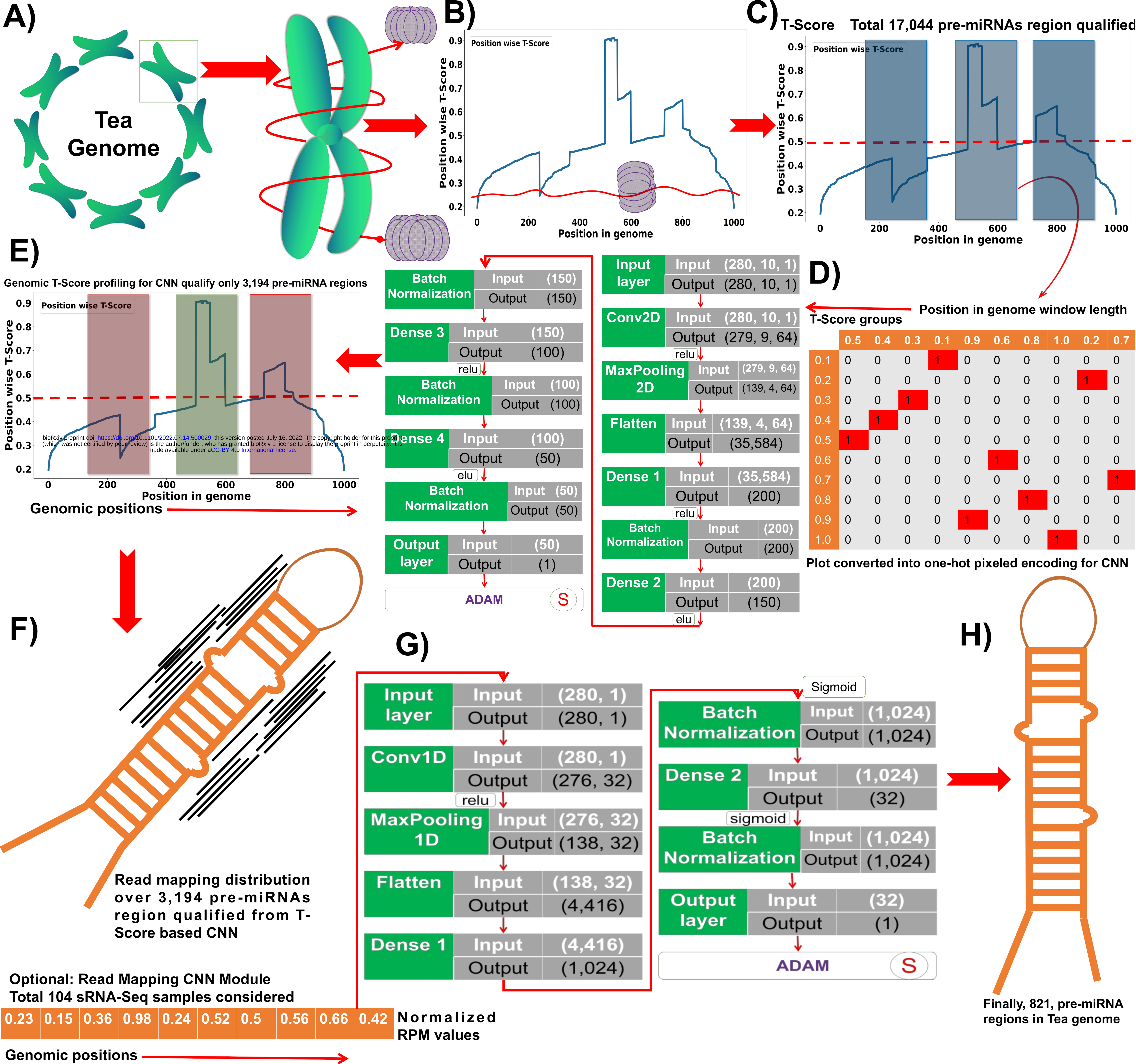
bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




Accuracy
Sensitivity
Specificity
Percent(%)
70
75
80
85
90
95
Monomer
Dimer
Trimer
Pentamer
Structure Triplet
Monomer+Dimer
Monomer+Trimer
Monomer+Dimer+Trimer
Monomer+Dimer+Pentamer
Monomer+Dimer+
Trimer+Structure Triplet
Monomer+Dimer+Trimer+
Pentamer+Structure Triplet
Monomer+Dimer+
Pentamer+Structure Triplet
Monomer+Dimer+Pentamer+
Structure Triplet (XGboost)
Sequence encoding

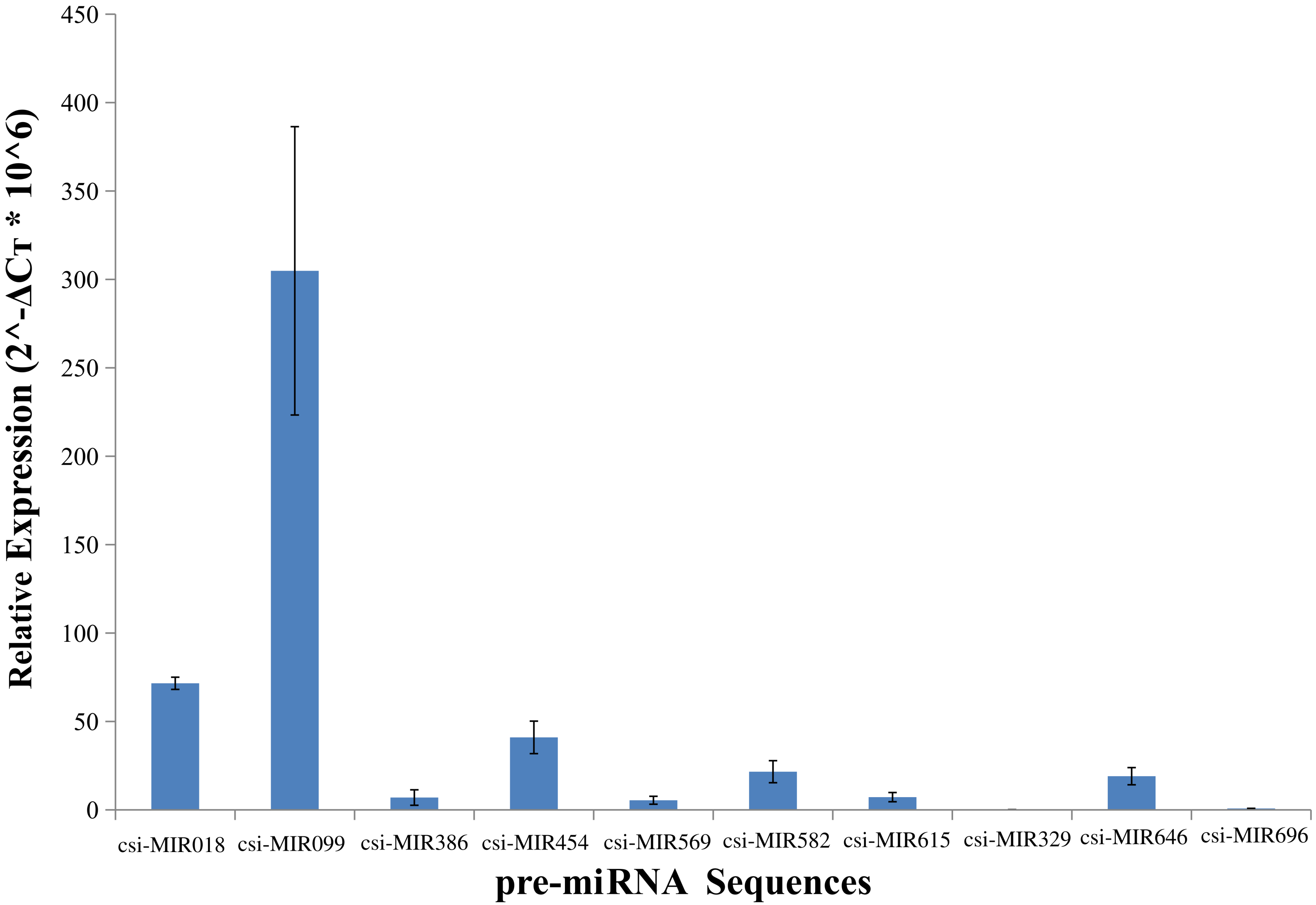

A)
Tools
miWords 94.31 93.85 94.08 94.09 88.16
PlantMirP-Rice 87.51 88.27 87.89 87.84 75.70
plantMirP 86.69 88.68 87.69 87.56 75.30
miPlantPreMat 84.29 87.86 86.07 85.82 72.20
HuntMi 82.94 83.64 83.29 83.23 66.50
microPred 80.94 82.00 81.47 81.37 62.90
mirDNN 95.90 75.03 85.46 86.84 72.52
DeepMir 1.29 98.71 50.00 2.51 0.00
deepSOM 86.87 78.66 82.77 83.45 65.76
Sensitivity
Specificity
Accuracy
F1 Score
MCC
Performance metrics
80
60
40
20
0
B)
Tools
miWords 94.07 93.13 93.60 93.63 87.20
PlantMirP-Rice 81.87 82.42 82.15 82.10 64.29
plantMirP 84.84 83.25 84.04 84.17 68.10
HuntMi 93.99 87.05 90.52 90.84 81.24
mirDNN 90.20 77.03 83.61 84.62 67.81
DeepMir 14.36 93.02 53.69 23.67 11.95
deepSOM 88.46 85.96 87.21 87.37 74.44
Sensitivity
Specificity
Accuracy
F1 Score
MCC
Performance metrics


False positive
False positive
2000
1750
1500
1250
1000
750
500
250
0
Sensitivity
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Precision
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Accuracy
1.0
0.8
0.6
0.4
0.2
miR-PREFeR
miRanalyzer
mirDeep-P
mirDeep*
mirDeep2
MIReNA
ShortStack
miWords-T
miWords-R
Tool


A)
Tea Genome
B)
Position wise T-Score
Position wise T-Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
200
400
600
800
1000
Position in genome
C)
T-Score
Total 17,044 pre-miRNAs region qualified
Position wise T-Score
Position wise T-Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
200
400
600
800
1000
Position in genome
E)
Genomic T-Score profiling for CNN qualify only 3,194 pre-miRNA regions
Position wise T-Score
Position wise T-Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
200
400
600
800
1000
Position in genome
Genomic positions
Batch Normalization | Input | (150)
Output | (150)
Dense 3 | Input | (150)
Output | (100)
relu
Batch Normalization | Input | (100)
Output | (100)
Dense 4 | Input | (100)
Output | (50)
elu
Batch Normalization | Input | (50)
Output | (50)
Output layer | Input | (50)
Output | (1)
ADAM
S
Input layer | Input | (280, 10, 1)
Output | (280, 10, 1)
Conv2D | Input | (280, 10, 1)
Output | (279, 9, 64)
relu
MaxPooling 2D | Input | (279, 9, 64)
Output | (139, 4, 64)
Flatten | Input | (139, 4, 64)
Output | (35,584)
Dense 1 | Input | (35,584)
Output | (200)
relu
Batch Normalization | Input | (200)
Output | (200)
Dense 2 | Input | (200)
Output | (150)
elu
D)
Position in genome window length
T-Score groups
Plot converted into one-hot pixeled encoding for CNN
F)
Read mapping distribution over 3,194 pre-miRNAs region qualified from T-Score based CNN
Optional: Read Mapping CNN Module
Total 104 sRNA-Seq samples considered
0.23 | 0.15 | 0.36 | 0.98 | 0.24 | 0.52 | 0.5 | 0.56 | 0.66 | 0.42
Normalized RPM values
Genomic positions
G)
Input layer | Input | (280, 1)
Output | (280, 1)
Conv1D | Input | (280, 1)
Output | (276, 32)
relu
MaxPooling 1D | Input | (276, 32)
Output | (138, 32)
Flatten | Input | (138, 32)
Output | (4,416)
Dense 1 | Input | (4,416)
Output | (1,024)
Sigmoid
Batch Normalization | Input | (1,024)
Output | (1,024)
Dense 2 | Input | (1,024)
Output | (32)
sigmoid
Batch Normalization | Input | (1,024)
Output | (1,024)
Output layer | Input | (32)
Output | (1)
ADAM
S
H)
Finally, 821, pre-miRNA regions in Tea genome

	0.5	0.4	0.3	0.1	0.9	0.6	0.8	1.0	0.2	0.7
0.1	0	0	0	1	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	1	0
0.3	0	0	1	0	0	0	0	0	0	0
0.4	0	1	0	0	0	0	0	0	0	0
0.5	1	0	0	0	0	0	0	0	0	0
0.6	0	0	0	0	0	1	0	0	0	0
0.7	0	0	0	0	0	0	0	0	0	1
0.8	0	0	0	0	0	0	1	0	0	0
0.9	0	0	0	0	1	0	0	0	0	0
1.0	0	0	0	0	0	0	0	1	0	0

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.14.500029; this version posted July 16, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


Relative Expression (2^-ΔCт * 10^6)
450
400
350
300
250
200
150
100
50
0
csi-MIR018
csi-MIR099
csi-MIR386
csi-MIR454
csi-MIR569
csi-MIR582
csi-MIR615
csi-MIR329
csi-MIR646
csi-MIR696
pre-miRNA Sequences