Deepred-Mt: Deep representation learning for predicting C-to-U RNA editing in plant mitochondria

Alejandro A. Edera[1] Ian Small[2] Diego H. Milone[1]*
M. Virginia Sanchez-Puerta[3,4]†

[1]Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Ciudad Universitaria, Santa Fe, Colectora Ruta Nacional No 168 km. 0, Paraje El Pozo, Santa Fe, 3000, Argentina.

[2]ARC Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, WA 6009, Australia.

[3]IBAM, Universidad Nacional de Cuyo, CONICET, Facultad de Ciencias Agrarias, Almirante Brown 500, Chacras de Coria, M5528AHB, Argentina.

[4]Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Cuyo, Padre Jorge Contreras 1300, M5502JMA, Argentina.

Abstract

In land plant mitochondria, C-to-U RNA editing converts cytidines into uridines at highly specific RNA positions called editing sites. This editing step is essential for the correct functioning of mitochondrial proteins. When using sequence homology information, edited positions can be computationally predicted with high precision. However, predictions based on the sequence contexts of such edited positions often result in lower precision, which is limiting further advances on novel genetic engineering techniques for RNA regulation. Here, a deep convolutional neural network called Deepred-Mt is proposed. It predicts C-to-U editing events based on the 40 nucleotides flanking a given cytidine. Unlike existing methods, Deepred-Mt was optimized by using editing extent information, novel strategies of data augmentation, and a large-scale training dataset, constructed with deep RNA sequencing data of 21 plant mitochondrial genomes. In comparison to predictive methods based on sequence homology, Deepred-Mt attains significantly better predictive performance, in terms of average precision as well as F1 score. In addition, our approach is able to recognize well-known sequence motifs linked to RNA editing, and shows that the local RNA structure surrounding editing sites may be a relevant factor regulating their editing. These results demonstrate that Deepred-Mt is an effective tool for predicting C-to-U RNA editing in plant mitochondria. Source code, datasets, and detailed use cases are freely available at `https://github.com/aedera/deepredmt`.

Keywords: Representation learning, Sequence classification, Convolutional neural networks, Land plants, Mitochondrial genomes, C-to-U RNA editing.

1 Introduction

In land plants, the editosome is a highly sophisticated molecular machine able to bind organellar RNA molecules. It post-transcriptionally converts cytidines to uridines (C-to-U) at highly specific RNA positions called editing sites (esites). Editing restores well conserved amino acids that are essential for protein functioning [Covello and Gray, 1989, Giegé and Brennicke, 1999, Gualberto et al., 1989, Hiesel et al., 1989]. In general, protein products translated from unedited RNAs result in severe or lethal phenotypes [Liu et al., 2013, Toda et al., 2012, Schallenberg-Rüdinger et al., 2013, Li et al., 2014]. Editosome recognition is governed by *cis* elements or molecular signals found in the sequence context of edited cytidines [Mulligan et al., 2007, Giegé and Brennicke,

*Corresponding author: dmilone@sinc.unl.edu.ar

†Corresponding author: mvsanchezpuerta@fca.uncu.edu.ar

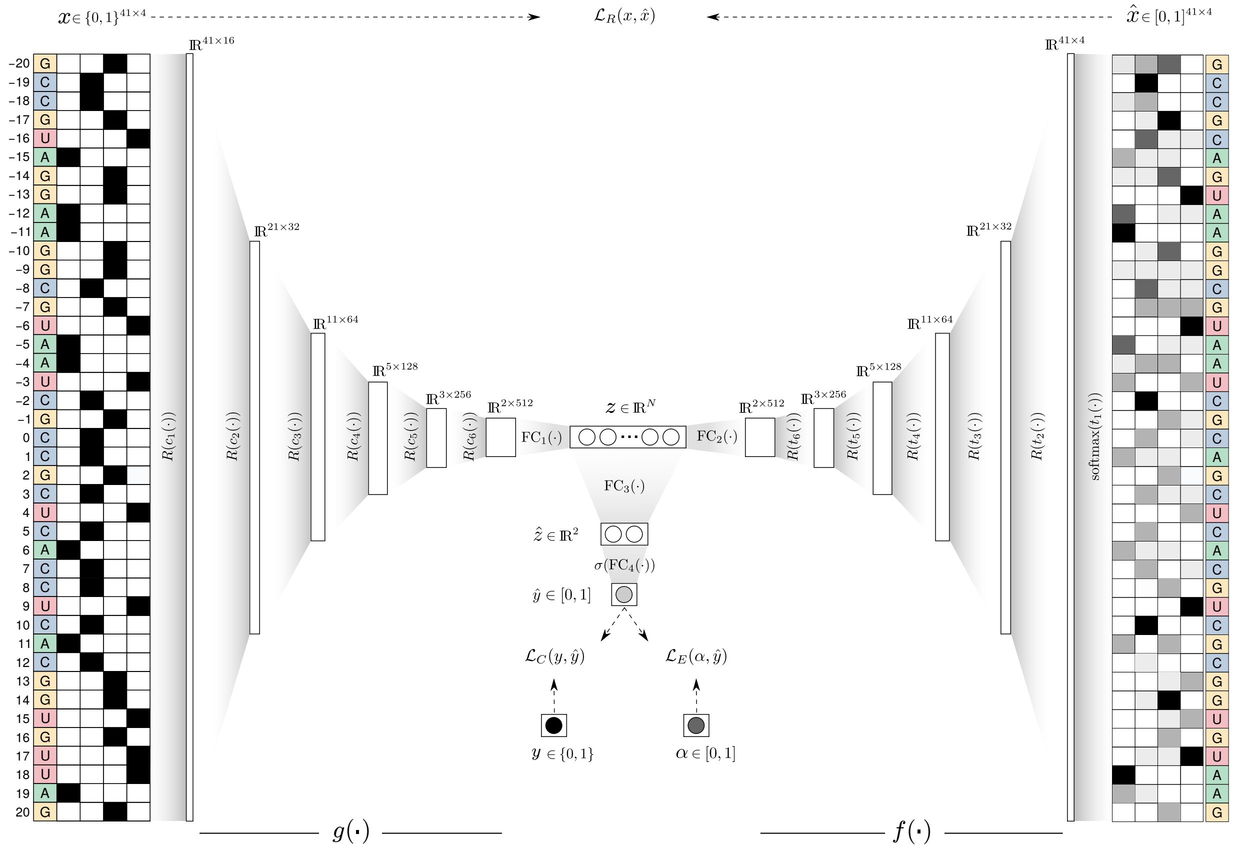

Figure 1: Architecture of Deepred-Mt. Sequence motifs of an input RNA sequence x are detected by a deep convolutional network g and encoded into a dense vector z. Motif detection is assisted by an additional deep convolutional network f, which transforms the vector z back to the input sequence. A score $\hat{y}$ calculated from the vector z is used to predict if the input sequence is edited. The network is optimized to match the score $\hat{y}$ with the real edited state y and editing extent α of the input sequence.

1999, Kubo and Kadowaki, 1997]. These signals are not only found in the 5' sequence region of target cytidines [Barkan et al., 2012, Takenaka et al., 2013, Yagi et al., 2013] but also in the 3' sequence region [Choury et al., 2004, Farré et al., 2001, Neuwirt et al., 2005]. However, despite the specificity exhibited by editosome recognition [Barkan et al., 2012], a further characterization of *cis* elements remains recalcitrant, limiting further advances on editosome binding prediction [Takenaka et al., 2013, Yagi et al., 2013, Kobayashi et al., 2019] as well as in genetic engineering techniques for RNA regulation [Yan et al., 2019].

A significant step towards a better characterization of the *cis*-elements involved in RNA editing consists in designing models capable of predicting editing sites by mimicking how the editosome recognizes its targets. In other words, such models must be able to predict editing events by exclusively using the information encoded by the sequence context of cytidines. Although several *ab initio* models have attempted this in the past [Du et al., 2007, Du and Li, 2008, Cummings and Myers, 2004, Thompson and Gopal, 2006, Yura et al., 2008], they suffer from many false positives [Mower, 2009], degrading precision. Attempts at decreasing the number of false positives have failed even when such models have used additional input information, such as codon position, free energy, amino acids patterns, and extremely large flanking sequences [Thompson and Gopal, 2006, Cummings and Myers, 2004, Du et al., 2007]. In contrast, simpler computational methods based on sequence homology are able to achieve better predictions [Mower, 2009, Lenz et al., 2018]. These methods based on sequence homology exploit the fact that RNA editing tends to restore amino acids at very well conserved positions. However, this sequence homology approach results unsuitable (on their own) to study the *cis* elements involved in RNA editing, because of not using sequence context information.

Here, a novel *ab initio* model called Deepred-Mt is proposed. It predicts C-to-U editing sites by only using the neighboring nucleotides of target cytidines. Neighboring nucleotides are processed by using a deep convolutional neural network whose architecture is inspired in recent advances in sequence classification tasks [Kalchbrenner et al., 2014, Zhang et al., 2015, Alipanahi et al., 2015, Acharya et al., 2017, Romdhane and Pr, 2020]. We optimized Deepred-Mt to especially predict C-to-U editing events on coding regions of plant mitochondrial genomes. To this aim, we created a large-scale dataset containing 138,597 cytidines from 21 plant mitochondrial genomes. These cytidines were carefully labeled, as either edited or unedited, by using 13 billion high-quality paired-end reads. Unlike existing methods, Deepred-Mt optimization makes use of two novel sources of information. First, it includes editing extent information (i.e., in how many RNAs the same site is edited), to be able to capture molecular patterns linked to RNA editing efficiency [Chateigner-Boutin and Hanson, 2003, Mower and Palmer, 2006, Phreaner et al., 1996]. Second, it uses data augmentation techniques, to include more general domain knowledge about mitochondrial RNA sequences. Our experiments show that Deepred-Mt predictions are significantly better than those of PREP-Mt [Mower, 2009] and PREPACT [Lenz et al., 2018], the state-of-the-art methods for predicting RNA editing based on sequence homology. In addition, the more confident predictions of Deepred-Mt are consistent with well-known sequence motifs linked to RNA editing, suggesting that Deepred-Mt is recognizing meaningful sequence information.

2 Deepred-Mt

2.1 Architecture

Given an RNA sequence consisting of a central cytidine flanked by 20 nucleotides on each side, Deepred-Mt computes a prediction score $\hat{y}$ that indicates if this central cytidine is edited. This score is computed from a vector z which compactly encodes motifs automatically found in the input RNA sequence. Two deep convolutional neural networks are responsible for motif detection: $g(\cdot)$ and $f(\cdot)$. The whole architecture of Deepred-Mt is shown in Figure 1. Here, the network g builds the vector z, whereas the network f transforms this vector z back to a sequence $\hat{x}$ that reconstructs the input sequence x. Although this reconstruction step is not strictly needed for computing the score $\hat{y}$, it is a crucial piece for representation learning [Bengio et al., 2013], which plays a central role during parameter optimization [Hinton and Salakhutdinov, 2006]. This is because the reconstruction step helps the network g to construct vectors z that serve as meaningful representations of the input sequences.

To process the input sequence, it is first converted into a matrix $x \in \{0,1\}^{41\times 4}$, where rows represent the nucleotides in each sequence position as a one-hot vector (i.e., a binary vector with a single component set to one). Then this matrix is progressively processed by the network g that has the form

$$g(x) = (R \circ c_6 \circ R \circ c_5 \dots R \circ c_1)(x),$$

where $\circ$ denotes a function composition, and c_i is a 1-D convolution layer whose output is passed through a function R. This function R improves motif detection by carrying out a (batch) normalization, which adjusts the mean and standard deviation of inputs [Ioffe and Szegedy, 2015], followed by a rectification, which clamps all negative values to zero [Nair and Hinton, 2010]. The output of each convolutional layer is a matrix with a number of rows and columns that is approximately half and double, respectively, with respect to those of the input matrix. This dimensionality change enables higher convolutional layers to process longer chunks of sequence, allowing the detection of complex data patterns [LeCun et al., 1990]. The network g outputs a matrix that is then linearly transformed by a fully connected layer FC_1 into the vector $z \in \mathbb{R}^N$. With this vector z, both the reconstructed sequence $\hat{x}$ and the score $\hat{y}$ are computed.

To reconstruct the input sequence, a fully-connected layer FC_2 transforms the vector z back to a matrix used as input for the network f. The architecture of this network f mirrors that of the network g, but its convolutional layers [Wojna et al., 2019], as denoted by $t_i(\cdot)$ in Figure 1, are reversely sorted and also transposed. This mirrored architecture is very often employed in representation learning [Bengio et al., 2013] and is what allows reconstructing the input sequence. The output of the last transposed convolutional layer, $t_1(\cdot)$, is a 41-by-4 matrix. The rows of this matrix are then normalized to one, by using a softmax function [Goodfellow et al., 2016], to obtain the reconstructed sequence $\hat{x}$.

The prediction score $\hat{y}$ is computed as

$$\begin{aligned} \hat{z} &= \mathrm{FC}_3(z), \\ \hat{y} &= \sigma(\mathrm{FC}_4(\hat{z})). \end{aligned}$$

Here, the fully-connected layers FC_3 and FC_4 linearly reduce the dimensionality of the vector z into vectors with dimensionalities $\mathbb{R}^2$ and $\mathbb{R}$, respectively. The score $\hat{y} \in [0, 1]$ is computed by passing the one-dimensional output of FC_4 through a sigmoid function σ.

2.2 Learning

All convolutional and fully-connected layers of Deepred-Mt are functions parameterized by a set of parameters θ. We optimized these parameters by using an $N = 5$ and a training dataset D consisting of 41-bp RNA sequences with central positions containing cytidines labeled with two values: $y \in \{0, 1\}$ and $\alpha \in [0, 1]$. The first categorical value indicates if the central cytidine is edited $(y = 1)$ whereas the second real value indicates its editing extent.

To optimize the parameters, a so-called loss function is often used, which measures how wrong the training dataset D is predicted for a given set of parameters θ. For our problem, we use the following loss function

$$\mathcal{L}(\theta, D) = \sum_{(x,y,\alpha)\in D} \underbrace{\mathcal{L}_R(x, \hat{x})}_{\text{reconstruction loss}} + \underbrace{\mathcal{L}_C(y, \hat{y})}_{\text{classification loss}} + \underbrace{\mathcal{L}_E(\alpha, \hat{y})}_{\text{editing extent loss}} \quad ,$$

where $\mathcal{L}_R$, $\mathcal{L}_C$, and $\mathcal{L}_E$ are loss components called the reconstruction loss, classification loss, and editing extent loss, respectively. This equation shows that the loss function $\mathcal{L}$ is computed independently for each training sequence x based on its reconstruction $\hat{x}$ and prediction score $\hat{y}$.

Internally, the three loss components are defined in terms of widely used functions employed for loss calculation [Goodfellow et al., 2016]. The reconstruction loss $\mathcal{L}_R$ uses the categorical cross-entropy to measure how similar the sequence x and its reconstruction $\hat{x}$ are: $\mathcal{L}_R = -\sum_k \sum_i x_i^k \log(\hat{x}_i^k)$, where k and i index the 41 sequence positions and the four nucleotides, respectively. The classification loss $\mathcal{L}_C$ uses the binary cross-entropy to compare how similar the label y and the prediction score $\hat{y}$ are: $\mathcal{L}_C = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y}))$. The editing extent loss uses the mean absolute error to measure how far the editing extent α is from the score $\hat{y}$: $\mathcal{L}_E = -|\alpha - \hat{y}|$. The editing extent loss $\mathcal{L}_E$ allows the model to capture editing extent patterns, such as the inefficient edition of synonymous cytidines [Mower and Palmer, 2006]. To this aim, the editing extent of a cytidine is used as a probability of editing, to represent how frequent a cytidine is edited. In this way, the optimization of the loss $\mathcal{L}_E$ incorporates editing extent information by calibrating Deepred-Mt predictions. This calibration encourages predictions to be overconfident for sites with higher editing extents but underconfident for those with lower editing extents.

An algorithm often used to optimize loss functions is stochastic gradient descent [Bottou et al., 2018], which calculates the gradient of a loss function over small subsets of the training dataset called mini-batches. Since unedited cytidines outnumber edited ones [Giegé and Brennicke, 1999], our training dataset is unbalanced and can result in a suboptimal optimization of parameters [Krawczyk, 2016]. This was circumvented by building each mini-batch with an equal number of edited and unedited sequences that were randomly drawn from the training dataset. To improve predictive power, 20% of the training dataset was used as a validation set. Optimization was stopped by monitoring the loss function on this validation set [Caruana et al., 2001]. We made the source code of Deepred-Mt publicly available[1].

2.3 Novel learning strategies

To boost predictions of editing events, we design two novel learning strategies: synthetic augmentation and task-related augmentation. Both strategies consist in significantly increasing the diversity and amount of the data available for training, without actually collecting new data. These strategies are inspired by the fact that using large-scale training data can dramatically improve the predictive performance of deep neural networks [Wang et al., 2017, Erhan et al., 2010].

The synthetic augmentation strategy (SAS) consists in artificially creating new training sequences by varying existing ones. During mini-batch building, sequences were varied on the fly by using two techniques. The first technique randomly replaces the esites of a sequence, if any, by uridines. This enables Deepred-Mt to mimic editosome functioning, which still recognizes sites on coding RNAs [Binder et al., 1994, Mareéchal-Drouard et al., 1993] even when neary esites are in diverse configurations of edited and unedited states [Phreaner et al., 1996, Lu et al., 1996, Wilson and Hanson, 1996]. The second technique is rather standard and consists in

[1] https://github.com/aedera/deepredmt.

occluding sequence positions, by assigning zero vectors to their corresponding rows in the one-hot matrix x. This prevents Deepred-Mt from making predictions that overly rely on very specific sequence patterns, since some sequence regions are not fully observable. In this strategy, two types of occlusions were used: the first randomly occludes 50% of the sequence positions, while the other occludes k consecutive sequence positions, where k is an integer drawn from the range between 0 and 10.

To achieve a stronger effect of the synthetic augmentation strategy, we also followed the standard training methodology employed by previous *ab initio* methods [Cummings and Myers, 2004, Du et al., 2007]. In this methodology, the sequence contexts of the same sites shared by the mitochondrial genomes of different species are included as part of the training set. Despite having a high sequence identity, these homologous sequences can have some differences in certain surrounding positions that can lead to a better parameter optimization. This is because such differences contribute to expanding the training set with new variants, which are also magnified by the synthetic augmentation strategy. In addition, conservation patterns among the positions surrounding homologous sites have shown to be crucial for identifying editing motifs [Mulligan et al., 2007]. Moreover, homologous sequences with 100% sequence identity are still useful for capturing editing patterns. They contribute to discriminating better editing events that are functionally important, since such events are well conserved across homologous sites [Brenner et al., 2019].

The task-related augmentation strategy (TAS) consists in slightly changing the predictive task of Deepred-Mt to be able to use more abundant data along with the original training data. To achieve this, we make the original predictive task of Deepred-Mt more general. In this new predictive task, sequences were predicted as edited not only when their central positions have esites but also when they have uridines (U) homologous to esites. The rationale behind this is that these uridines play the same role as their homologous esites when they are translated into amino acids. This is because C-to-U RNA editing restores well-conserved amino acids [Covello and Gray, 1989]. Moreover, since the mutational rate of plant mitochondria is exceptionally lower than other genomes [Wolfe et al., 1987], we can assume that the sequence contexts of such uridines are very similar to their homologous esites. Thus, these uridines are very likely prone to be recognizable by the editosome, for example, when they are mutated into cytidines Choury and Araya [2006]. To be able to use this task-related sequences in our problem, the uridines in their central positions were replaced by esites. Our experiments demonstrate that this simple augmentation strategy results in substantial improvements in predictive performance.

3 Materials and experimental setup

3.1 Large-scale data collection

Mitochondrial genomes of land plants were sampled from the National Center for Biotechnology Information whenever they were fully sequenced and also had abundant paired-end RNAseq data available in the European Nucleotide Archive. This sampling excluded mitochondrial genomes whose RNA is U-to-C edited. This is because U-to-C RNA editing can preclude a reliable identification of C-to-U editing events, and it also makes harder the discovery of C-to-U editing signals, since a same site can be both C-to-U and U-to-C edited [Groth-Malonek et al., 2007, Knie et al., 2016, Schallenberg-Rüdinger and Knoop, 2016]. In addition, the mitochondrial genome of *Amborella trichopoda* was excluded from sampling, since it has an unusual number of gene copies that precludes a reliable esite identification [Rice et al., 2013]. In total, we sampled 21 mitochondrial genomes for which a total of 194 sets of RNAseq were downloaded (Table S1) and processed with fastp v0.20.0 [Chen et al., 2018], for quality-based filtering and adapter trimming. A total of 13,008,426,633 paired-end RNA reads were obtained with qualities above Q20 on average (Table S1). From the sampled genomes we extracted 723 sequences encoding well-known and intact protein-coding genes on which paired-end RNA reads were mapped using Bowtie v2.3.4.3 [Langmead and Salzberg, 2012]. To enhance read-mapping, sequence ends were extended by 102 bp. A total of 83,800,590 read pairs were aligned. To reduce the number of misaligned read pairs as well as those harboring sequencing errors, aligned read pairs were discarded whenever they had more than two mismatches with respect to the non-C reference positions (to exclude potential esites). After this step, we obtained a total of 79,286,452 read pairs.

We constructed deep RNAseqs by pooling the aligned read pairs when they belonged to different RNAseqs but came from the same species. This substantially increases read depths, leading to more reliable identifications of esites. In addition, it overcomes some of the frequent problems arising from using a single RNAseq, such as intra-species and tissue-specific biases as well as sequencing artifacts [Zehrmann et al., 2008, Stone and Storchova, 2015, Wu et al., 2015]. As a result, we obtained at least 20 high-quality reads aligned on 99% of the coding cytidines (Figure S1). The proportion of T bases aligned on a nucleotide was defined as the editing extent. Esites were identified by looking for cytidines whose editing extents were greater than 10%, as is usual in other studies [Wu et al., 2015, Štorchová et al., 2018, Zheng et al., 2020].

3.2 Datasets

RNA sequences were extracted from the 723 protein-coding sequences by centering a window at every position (position 0) and taking 20 upstream (-20 to -1) and 20 downstream (+1 to +20) nucleotides. In total, 652,634 RNA sequences were extracted, with A (172,395), unedited C (126,166), edited C (12,431), G (140,611), and U (201,031) in their central positions. The extracted sequences were used to define the following three datasets.

Training dataset. It included all the sequences with central positions containing unedited and edited Cs. Central positions were labeled with 0s and 1s when containing unedited and edited cytidines, respectively. In addition, these central positions were also labeled with their corresponding editing extents.

Task-related sequences. These sequences were obtained by using the proposed task-related augmentation strategy. We found that a total of 97,397 sequences with central positions containing Us homologous to esites (Fig. S2). The central positions of these task-related sequences were labeled with 1s and also with their corresponding editing extents.

Control dataset. We constructed a control dataset by including a fake RNA editing signal in a random downstream position of training sequences labeled with 1. This fake signal was defined as the motif "GGCG". One of the four nucleotides of this fake signal was randomly changed to another nucleotide according to the editing extent of a sequence. Concretely, each time the fake signal was included in a sequence, one nucleotide was changed with a probability inversely proportional to the editing extent of the sequence. Therefore, fake signals included in sequences with low editing extents were more likely to deviate from the motif "GGCG" in one nucleotide, simulating editing extent patterns linked to sequence motifs.

3.3 Baseline methods

As baseline methods for predicting C-to-U esites, we used PREP-Mt [Mower, 2009] and PREPACT [Lenz et al., 2018]. The predictive performance of these methods is closely tied to their own internal data, composed of full gene sequences. These data, not publicly available, were selected by experts from specific plant species to achieve reliable predictions. As clarified below, both baseline methods make predictions based on sequence homology.

PREP-Mt. Given an input protein-coding RNA sequence, PREP-Mt translates it into a protein sequence that is aligned against a set of reference protein sequences. These reference sequences were previously translated from gene sequences whose esites have been experimentally verified and replaced by uridines. Once aligned, the input sequence is scanned for amino-acid mismatches yielded by codons containing cytidines. If these amino-acid mismatches can be solved by replacing cytidines by uridines in the input sequence, then these replaced cytidines are predicted as esites. A score is finally calculated for each predicted esite based on how well their codons are conserved across the reference sequences [Mower, 2009]. It is worth noting that PREP-Mt is unable to make predictions for synonymous cytidines, since they lead to no amino acid changes. For our experiments, we used the public web service of PREP-Mt[2].

PREPACT. Similar to PREP-Mt, PREPACT aligns a given input protein-coding RNA sequence against a reference set of RNA sequences whose esites have been already identified. Cytidines in the input sequence are predicted as edited whenever they are aligned with reference esites [Lenz et al., 2018]. Unlike PREP-Mt, PREPACT is able to predict esites at the third codon position, in which C-to-U RNA editing yields synonymous amino acid changes. We used the public web service of PREPACT with standard options: BLASTx to detect C-to-U changes, self position labeling, and the 25 mitochondrial genomes available as references[3]. To be able to obtain a predictive score for each input cytidine, we used as a score the relative number of PREPACT reference sequences in which the input cytidine was predicted as edited.

3.4 Evaluation metrics

Precision and recall are widely used to evaluate prediction performance. Precision measures how many sequences predicted as edited are truly edited: $TP/(TP + FP)$, where TP (True Positives) are the number of sequences

[2] http://prep.unl.edu

[3] http://www.prepact.de/prepact-main.php

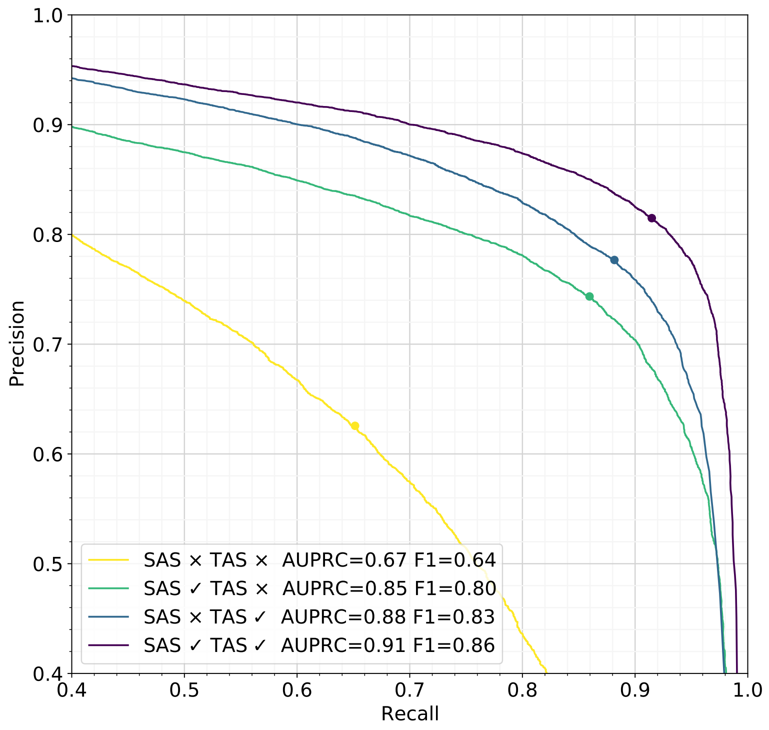

Figure 2: Impact of learning strategies on predictions of Deepred-Mt. Curves plot different precision-recall performances, where points indicate maximum F1 scores. Check and cross marks indicate presence or absence, respectively.

labeled and predicted as edited, while FP (False Positives) are the number of sequences labeled as unedited but predicted as edited. Recall measures how many of the truly edited sequences are correctly predicted: $TP/(TP + FN)$, where FN (False Negatives) are the number of sequences labeled as edited but predicted as unedited.

However, precision and recall depend on a user-defined threshold (or cutoff), used for binarizing method predictions. A widely used measure for evaluating the overall predictive performance is the maximum value of the F1 score. The F1 score is the harmonic mean of precision and recall, and takes value between 0 (worst) and 1 (best). Similarly, the area under the receiver operating characteristic (ROC) curve is widely used for assessing predictive performance over all thresholds. However, the area under the ROC curve can be misleading in scenarios where data are unbalanced [Saito and Rehmsmeier, 2015]. Instead, the area of the precision-recall curve (AUPRC) is often used when data are imbalanced.

4 Results

4.1 Motif recognition

We first conducted a controlled experiment to evaluate the motif recognition performance of Deepred-Mt. We used the control dataset, in which the downstream regions of the edited sequences contain a fake editing signal "GGCG" that was randomly degraded according to the editing extents of the sequences. By using a stratified 5-fold cross validation for evaluating model performance, Deepred-Mt obtained a high predictive performance: an AUPRC of 0.94 and a maximum F1 score of 0.88 (Figure S3). By further inspecting false positive predictions, we found that they were mainly yielded by wild-type occurrences of the fake signal in sequences with unedited cytidines in their central positions, which fooled network predictions. Although the fake signal does not resemble the motifs found in real edited sequences, the result of this experiment is still important as it clearly demonstrates that the architecture of Deepred-Mt is capable of detecting sequence motifs.

Next, we analyzed the vector representations z constructed by Deepred-Mt from the control sequences. The aim of this analysis was to assess if the fake signal was correctly encoded by such vectors. To this analysis, we trained Deepred-Mt on the whole control dataset to then use it to obtain the vectors z of the control sequences labeled as edited. The resulting vectors were visually inspected by reducing their dimensionality to just two dimensions, using UMAP [McInnes et al., 2018]. We found that the vectors formed 18 well separated groups (Figure S4A). When calculating a consensus sequence for each group separately, the fake signal was almost perfectly recovered in all but one group (Figure S4B). Notably, in this group with no consensus sequence, its sequences show the lowest average editing extent. This indicates that such sequences harbor fake signals with changed nucleotides. These results show that Deepred-Mt was able to capture the erratic fake signal, as well as its editing extent patterns. This highlights the advantages of using methods capable for detecting sequence motifs, such as Deepred-Mt, as compared to methods that are exclusively based on sequence homology.

4.2 Impact of learning strategies

In Section 2.3, we proposed the synthetic and task-related augmentation strategies designed to achieve a better optimization of the parameters of Deepred-Mt, aiming to improve its predictive power. Therefore, we analyzed quantitatively how these strategies affected the predictive performance of Deepred-Mt. To determine the specific impact of each learning strategy, we compared the predictive performance achieved by Deepred-Mt when its parameters were optimized using none (SAS × TAS ×), either (SAS ✓ TAS ×, or SAS × TAS ✓), or both learning strategies (SAS ✓ TAS ✓). The predictive performance of these four versions of Deepred-Mt was evaluated by using a stratified, 5-fold cross-validation on the training set.

The results are shown in Figure 2 where the predictive performance of each version of Deepred-Mt is represented as a precision-recall curve which aggregates the results over the 5 folds of a stratified cross-validation. Taking as baseline the performance of Deepred-Mt when using none of the learning strategies (SAS × TAS ×), the use of the synthetic augmentation strategy (SAS ✓ TAS ×) shows an improvement in predictions. This is reflected both in the AUPRC, which increases from 0.67 to 0.85, as well as in the maximum F1 score, which increases from 0.64 to 0.80. Similarly, the use of the task-related augmentation strategy (SAS × TAS ✓) results also in predictions better than those of the baseline version of Deepred-Mt. In this case, the AUPRC is increased from 0.67 to 0.88 and the maximum F1 score is increased from 0.64 to 0.83. However, when both strategies are used together (SAS ✓ TAS ✓), the resulting predictive performance is the highest. The combination of both strategies achieves an AUCPRC of 0.91 and a maximum F1 score of 0.86. This analysis shows that the use of both learning strategies allows Deepred-Mt to achieve a better predictive performance.

4.3 Homologous sequences

Here, we evaluate how the predictive capacity of Deepred-Mt is affected when homologous sequences are excluded from the training set. To evaluate the impact of this change, we trained Deepred-Mt using both learning strategies on the full training dataset (H ✓) and also on the same training dataset but depleted of homologous sequences (H ×). Next, we compared the predictive performance of the two trained models on 41-bp RNA sequences extracted from the mitochondrial genome of the hornwort *Anthoceros agrestis*. This genome was chosen because it harbors many very specific C-to-U esites that were experimentally identified [Gerke et al., 2020], and are not included in the training data used in this study.

Figure 3 shows the precision-recall curves obtained from the predictive performances of the two trained models. When homologous sequences are excluded from the training set (H ×), Deepred-Mt achieves an AUPRC and maximum F1 score equal to 0.31 and 0.36, respectively. Notably, these performance values are significantly improved when Deepred-Mt is trained with homologous sequences (H ✓). In this case, Deepred-Mt achieves an AUPRC and maximum F1 score equal to 0.58 and 0.56, respectively, which are both higher than the corresponding values obtained when excluding homologous sequences. This significant improvement on prediction power when including homologous sequences suggests that such sequences enable Deepred-Mt to capture relevant information about esites, which is leveraged to yield better predictions. Moreover, this result also demonstrates the importance of following the standard methodology used by previous *ab initio* models, in which homologous sequences are also included in the training set.

4.4 Comparing predictions of Deepred-Mt

To assess the predictive power of Deepred-Mt, we compared its predictive performance with those of two state-of-the-art predictive methods, PREP-Mt and PREPACT. For this comparison, we used a leave-one-species out

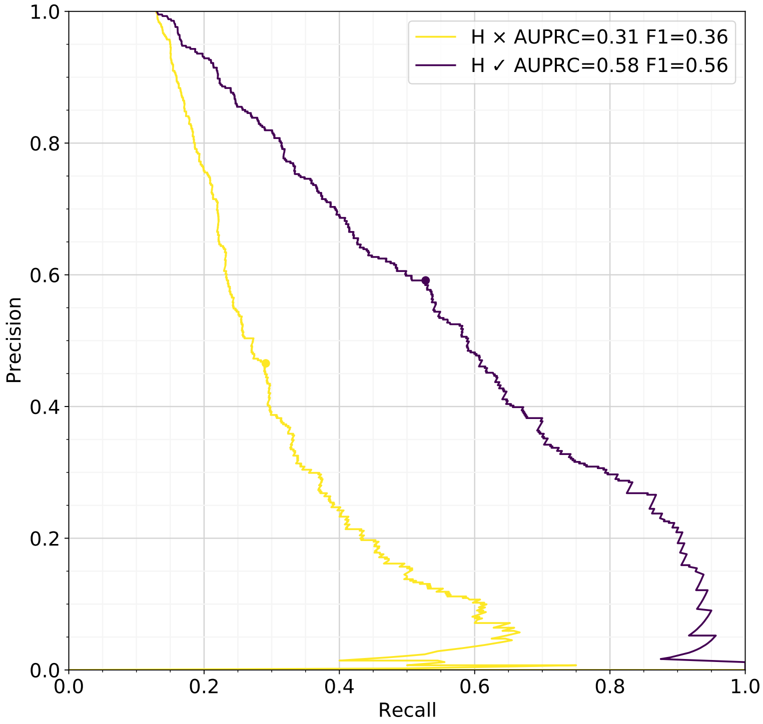

Figure 3: Impact of homologous sequences on predictions of Deepred-Mt. Curves plot different precision-recall performances, where points indicate maximum F1 scores. Check and cross marks indicate presence or absence, respectively.

cross validation on the training dataset. In each of the 21 folds of this cross validation, Deepred-Mt was trained on all training sequences except those coming from one species, which were instead used as testing data to estimate the prediction error. To train Deepred-Mt, we used both learning strategies. Our comparison was carried out on two variations of the training dataset: one excluding sequences whose central cytidines were synonymous, and the other including such sequences. Learning Deepred-Mt on each of these datasets required approximately 88 epochs, which is equivalent to ∼1 hour on an Nvidia Titan RTX GPU, and predicting all the sequences with the learned models spent 12 seconds.

Figure 4A shows the predictive performances of the three methods, represented by precision-recall curves, when synonymous sites are excluded. Summary statistics computed from these curves are shown in Table 1. This table shows that Deepred-Mt has a value of AUPRC (0.96) that is significantly higher than those of both PREPACT (0.91) and PREP-Mt (0.88). Similarly, Deepred-Mt achieves a higher F1 score (0.92) as compared to those achieved by PREPACT (0.89) and PREP-Mt (0.92). When analyzing the baseline methods, we can observe that PREPACT achieves a higher AUPRC (area under the precision-recall curve), while PREP-Mt instead achieves a higher maximal F1 score (point over the precision-recall curve). This means that the average

Table 1: Performances for predicting C-to-U RNA editing events when synonymous sites are either excluded or included. Boldface indicates the best predictive performances.

Method	Excluded		Included	
	AUPRC	F1 score	AUPRC	F1 score
PREPACT	0.91	0.89	0.79	0.82
PREP-Mt	0.88	0.91	0.76	0.84
Deepred-Mt	**0.96**	**0.92**	**0.91**	**0.86**

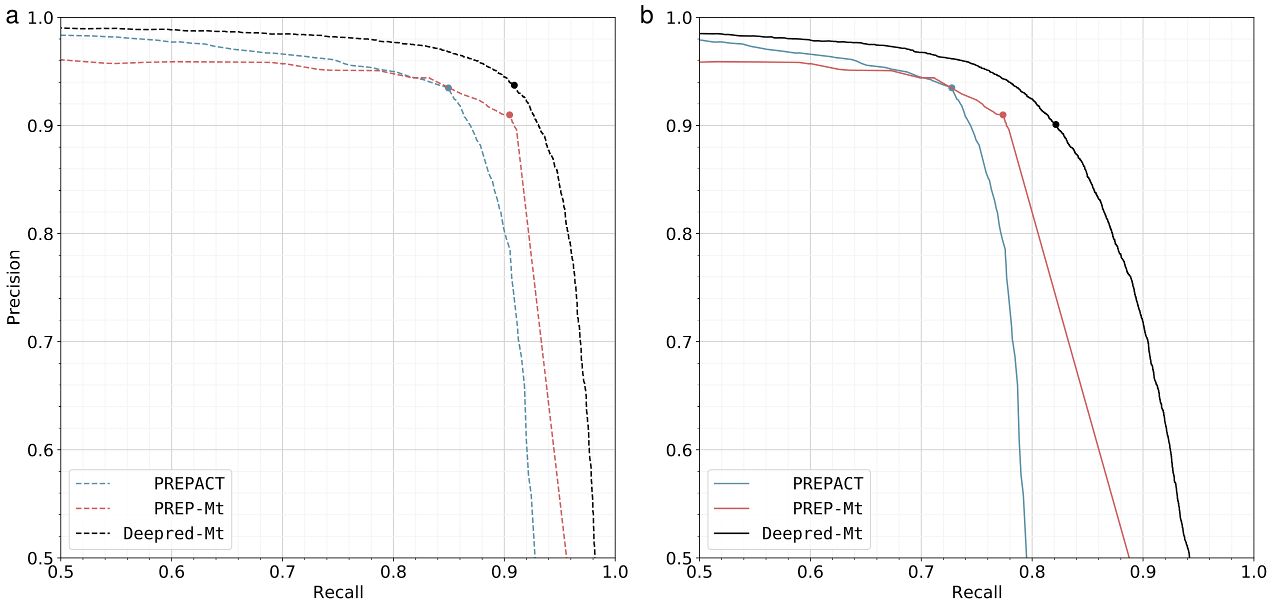

Figure 4: Predictive performances of Deepred-Mt and baselines. Curves plot precision-recall performances when synonymous sites are (a) excluded (dashed) or (b) included (solid). Points over curves indicate where the maximum F1 score is achieved.

precision of PREPACT is higher than that of PREP-Mt, while there is a threshold for PREP-Mt that yields a more balanced combination of precision and recall. However, in comparison to the baselines, Deepred-Mt predictions are better in both characteristics: higher average precision and more balanced precision and recall.

Figure 4B shows the predictive performances when synonymous sites are included. Table 1 shows statistics summarizing these performances. We can see that these results are in line with those obtained when excluding synonymous sites. The best predictive performance is again achieved by Deepred-Mt in both the AUPRC (0.91) and the F1 score (0.86), which are higher than the corresponding values achieved by PREPACT (AUPRC=0.79, F1=0.82) and PREP-Mt (AUPRC=0.76, F1=0.84). This is a relevant result because including synonymous sites makes the predictive problem much harder. This is due to synonymous sites are very specific (i.e., they are poorly conserved across species), and harbor degraded editing signals [Mower and Palmer, 2006, Mower, 2008]. These results demonstrate the advantages of using Deepred-Mt to predict C-to-U esites in mitochondrial genomes.

4.5 Detecting RNA editing motifs

One major drawback of the computational methods that use sequence homology to make predictions is that they cannot be used for exploratory analysis, such as finding novel sequence motifs associated with RNA editing. Here, we demonstrate that Deepred-Mt is well suited for this task. To this aim, we present an analysis showing that confident predictions of Deepred-Mt are linked to well-known sequence motifs associated with RNA editing.

For this analysis, we trained Deepred-Mt using both learning strategies on the fully training dataset. The resulting model was used then to score all the heterologous sequences in the training dataset, where all esites were left as cytidines. Next, we only retained those sequences that were correctly predicted, using a threshold of 0.5. With the scores of these sequences, we estimated their second quartiles separately for unedited and edited sequences. We used these quartiles for defining scores as confident. In the case of the retained unedited sequences, their scores were defined as confident if they were above its second quartile. Similarly, the scores of the retained edited sequences were defined as confident if they were below its second quartile. Finally, we estimated entropy-based consensus sequences [Schneider et al., 1986] from those sequences with confident scores. Since codon position can have an effect on the composition of the sequence contexts of the edited cytidines [Cummings and Myers, 2004, Mower, 2008, Mulligan et al., 2007], consensus sequences were estimated separately according to the codon position of the central positions.

Figure 5 shows the results of this analysis in which we can see in the first row the consensus structures of the unedited sequences in the first (2,115), second (2,150), and third (2,268) codon position, respectively. The

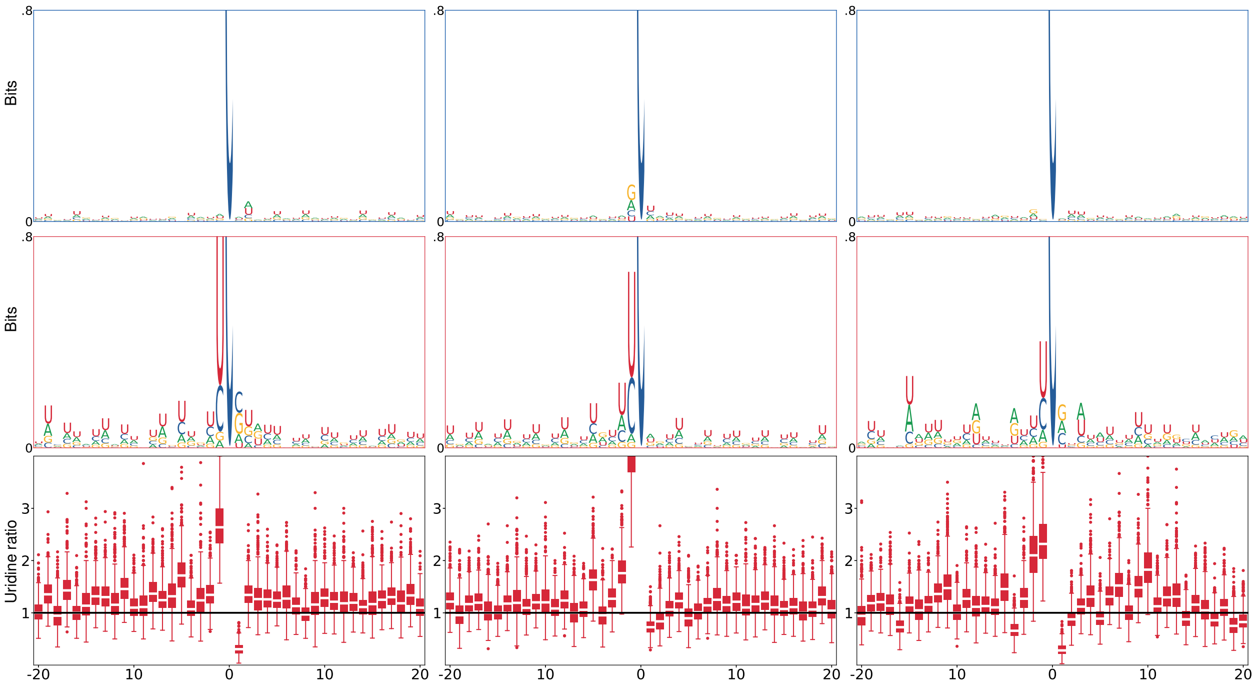

Figure 5: Consensus sequences associated with predictions of Deepred-Mt. The left, middle, and right columns show sequences with center positions in the first, second, and third codon positions, respectively. First and second rows show unedited and edited sequences, respectively. The third row shows boxes plotting ratios indicating the degree of uridine enrichment in edited sequences relative to unedited ones. Enrichment is indicated by ratios above the black horizontal line.

second row shows the consensus structures of the edited sequences in first (327), second (647), and third (53) codon position, respectively. These results show that highly-confident predictions of Deepred-Mt are associated with motifs of editing sites which have been reported in numerous studies. For example, unedited sequences with central positions in the second codon position show a high incidence of guanines in position -1, as previously observed [Choury et al., 2004, Oldenkott et al., 2019]. Instead, edited sequences show a high incidence of uridines in positions -2 and -1 for the three codon positions, and in position -5 for the first and second codon positions. This is in line with previous observations [Giegé and Brennicke, 1999, Mulligan et al., 2007]. In addition, this result also shows that the codon position has some effect on consensus sequences. This is observed in the high variability of the consensus structures of the edited sequences in the third codon position, as previously reported [Cummings and Myers, 2004, Mower, 2008, Mulligan et al., 2007].

Notably, the consensus sequences of the edited sequences show an uridine-enrichment at many positions, which is not observed in their equivalent positions in the unedited sequences. This has been also observed in other studies but not thoroughly analyzed in large-scale data [Cummings and Myers, 2004, Kindgren et al., 2015, Takenaka et al., 2007]. Therefore, we assessed this enrichment quantitatively as follows. For each sequence position, a ratio was calculated between the number of uridines in the edited and unedited sequences. This ratio was estimated 1,000 times by taking random samples of 100 unedited and 100 edited sequences. The last row of Figure 5 shows the distributions of the obtained ratios for each sequence position. This result shows that ratios are mainly concentrated above 1.0. For example, we can see that ratios are more frequently found above the upper quartiles (upper whiskers) than below the lower quartiles (lower whiskers). This indicates that edited sequences are enriched for uridines, and suggests that this enrichment may be a relevant factor for RNA editing.

5 Conclusions

In this study, Deepred-Mt is proposed as a novel model to predict C-to-U RNA editing in plant mitochondrial genomes. To make such predictions, Deepred-Mt recognizes sequence motifs in RNA sequences by using a deep convolutional neural network. Our results show that Deepred-Mt can achieve better predictive performance as compared to state-of-the-art methods. We empirically demonstrate that this superior predictive performance of Deepred-Mt is the result of a conjunction of factors: the use of recent advances in sequence processing, the preparation of large-scale training data, the incorporation of editing extent information as training data, and the design of novel strategies for augmenting the amount of training data.

To make predictions, Deepred-Mt mimics the recognition mode used by the editosome, the molecular apparatus responsible for RNA editing. The capability of Deepred-Mt for motif recognition represents a step forward to achieve a better characterization of the *cis*-elements involved in RNA editing. We found that Deepred-Mt was able to recognize well-known sequence motifs associated with RNA editing, such as the conservation of uridines in the position -1 and RNA regions enriched in uridines.

Funding

This work has been supported by the Agencia Nacional de Promoción Científica y Tecnológica (FONCyT) from Argentina (grant numbers PICT 2016-0555, PICT 2017-0691, and PICT 2018-3384).

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author's contributions

Alejandro A. Edera: Conceptualization of this study, data analysis, architecture design, result analysis, manuscript writing and reviewing. **Ian Small**: Conceptualization of this study, result analysis, manuscript writing and reviewing. **Diego H. Milone**: Conceptualization of this study, architecture design, result analysis, manuscript writing and reviewing. **M. Virginia Sanchez-Puerta**: Conceptualization of this study, data analysis, result analysis, manuscript writing and reviewing.

References

Patrick S Covello and Michael W Gray. RNA editing in plant mitochondria. *Nature*, 341(6243):662–666, 1989. doi: 10.1038/341662a0.

Philippe Giegé and Axel Brennicke. RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. *Proceedings of the National Academy of Sciences*, 96(26):15324–15329, 1999. doi: 10.1073/pnas.96.26.15324.

José M Gualberto, Lorenzo Lamattina, Géraldine Bonnard, Jacques-Henry Weil, and Jean-Michel Grienenberger. RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature*, 341 (6243):660–662, 1989. doi: 10.1038/341660a0.

Rudolf Hiesel, Bernd Wissinger, Wolfgang Schuster, and Axel Brennicke. RNA editing in plant mitochondria. *Science*, 246(4937):1632–1634, 1989. doi: 10.1126/science.2480644.

Yu-Jun Liu, Zhi-Hui Xiu, Robert Meeley, and Bao-Cai Tan. Empty pericarp5 encodes a pentatricopeptide repeat protein that is required for mitochondrial RNA editing and seed development in maize. *The Plant Cell*, 25(3):868–883, 2013. doi: 10.1371/journal.pgen.1008305.

Takushi Toda, Sota Fujii, Ko Noguchi, Tomohiko Kazama, and Kinya Toriyama. Rice MPR25 encodes a pentatricopeptide repeat protein and is essential for RNA editing of *nad5* transcripts in mitochondria. *The Plant Journal*, 72(3):450–460, 2012. doi: 10.1111/j.1365-313x.2012.05091.x.

Mareike Schallenberg-Rüdinger, Peter Kindgren, Anja Zehrmann, Ian Small, and Volker Knoop. A DYW-protein knockout in *Physcomitrella* affects two closely spaced mitochondrial editing sites and causes a severe developmental phenotype. *The Plant Journal*, 76(3):420–432, 2013. doi: 10.1111/tpj.12304.

Xiao-Jie Li, Ya-Feng Zhang, Mingming Hou, Feng Sun, Yun Shen, Zhi-Hui Xiu, Xiaomin Wang, Zong-Liang Chen, Samuel SM Sun, Ian Small, et al. Small kernel 1 encodes a pentatricopeptide repeat protein required for mitochondrial *nad7* transcript editing and seed development in maize (*Zea mays*) and rice (*Oryza sativa*). *The Plant Journal*, 79(5):797–809, 2014. doi: 10.1111/tpj.12584.

R Michael Mulligan, Kenneth LC Chang, and Chia Ching Chou. Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. *Molecular biology and evolution*, 24(9):1971–1981, 2007. doi: 10.1093/molbev/msm125.

Nakao Kubo and Koh-ichi Kadowaki. Involvement of 5' flanking sequence for specifying RNA editing sites in plant mitochondria. *FEBS letters*, 413(1):40–44, 1997. doi: 10.1016/s0014-5793(97)00873-9.

Alice Barkan, Margarita Rojas, Sota Fujii, Aaron Yap, Yee Seng Chong, Charles S Bond, and Ian Small. A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet*, 8(8): e1002910, 2012. doi: 10.1371/journal.pgen.1002910.

Mizuki Takenaka, Anja Zehrmann, Axel Brennicke, and Knut Graichen. Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. *PloS one*, 8(6):e65343, 2013. doi: 10.1371/journal.pone.0065343.

Yusuke Yagi, Shimpei Hayashi, Keiko Kobayashi, Takashi Hirayama, and Takahiro Nakamura. Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PloS one*, 8(3):e57286, 2013. doi: 10.1371/journal.pone.0057286.

David Choury, Jean-Claude Farre, Xavier Jordana, and Alejandro Araya. Different patterns in the recognition of editing sites in plant mitochondria. *Nucleic acids research*, 32(21):6397–6406, 2004. doi: 10.1093/nar/gkh969.

Jean-Claude Farré, Gabriel Leon, Xavier Jordana, and Alejandro Araya. *Cis* recognition elements in plant mitochondrion RNA editing. *Molecular and Cellular Biology*, 21(20):6731–6737, 2001. doi: 10.1128/mcb.21.20.6731-6737.2001.

Julia Neuwirt, Mizuki Takenaka, Johannes A Van Der Merwe, and Axel Brennicke. An in vitro RNA editing system from cauliflower mitochondria: editing site recognition parameters can vary in different plant species. *Rna*, 11(10):1563–1570, 2005. doi: 10.1261/rna.2740905.

Takehito Kobayashi, Yusuke Yagi, and Takahiro Nakamura. Comprehensive prediction of target RNA editing sites for PLS-class PPR proteins in *Arabidopsis thaliana*. *Plant and Cell Physiology*, 60(4):862–874, 2019. doi: 10.1093/pcp/pcy251.

Junjie Yan, Yinying Yao, Sixing Hong, Yan Yang, Cuicui Shen, Qunxia Zhang, Delin Zhang, Tingting Zou, and Ping Yin. Delineation of pentatricopeptide repeat codes for target RNA prediction. *Nucleic acids research*, 47(7):3728–3738, 2019. doi: 10.1093/nar/gkz075.

Pufeng Du, Tao He, and Yanda Li. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. *Biochemical and biophysical research communications*, 358(1):336–341, 2007. doi: 10.1016/j.bbrc.2007.04.130.

Pufeng Du and Yanda Li. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. *Journal of Theoretical Biology*, 253(3):579–586, 2008. doi: 10.1016/j.jtbi.2008.04.006.

Michael P Cummings and Daniel S Myers. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC bioinformatics*, 5(1):1–7, 2004. doi: 10.1186/1471-2105-5-132.

James Thompson and Shuba Gopal. Genetic algorithm learning as a robust approach to RNA editing site prediction. *BMC bioinformatics*, 7(1):145, 2006. doi: 10.1186/1471-2105-7-145.

Kei Yura, Yuki Miyata, Tomotsugu Arikawa, Masanobu Higuchi, and Mamoru Sugita. Characteristics and prediction of RNA editing sites in transcripts of the moss *Takakia lepidozioides* chloroplast. *DNA research*, 15(5):309–321, 2008. doi: 10.1093/dnares/dsn016.

Jeffrey P Mower. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic acids research*, 37(suppl_2):W253–W259, 2009. doi: 10.1093/nar/gkp337.

Henning Lenz, Anke Hein, and Volker Knoop. Plant organelle RNA editing and its specificity factors: enhancements of analyses and new database features in PREPACT 3.0. *BMC bioinformatics*, 19(1):255, 2018. doi: 10.1186/s12859-018-2244-9.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, 2014.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015. doi: 10.1038/nbt.3300.

U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 89:389–396, 2017. doi: 10.1016/j.compbiomed.2017.08.022.

Taissir Fekih Romdhane and Mohamed Atri Pr. Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss. *Computers in Biology and Medicine*, 123:103866, 2020. doi: 10.1016/j.compbiomed.2020.103866.

Anne-Laure Chateigner-Boutin and Maureen R Hanson. Developmental co-variation of RNA editing extent of plastid editing sites exhibiting similar *cis*-elements. *Nucleic acids research*, 31(10):2586–2594, 2003. doi: 10.1093/nar/gkg354.

Jeffrey P Mower and Jeffrey D Palmer. Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*. *Molecular Genetics and Genomics*, 276(3):285–293, 2006. doi: 10.1007/s00438-006-0139-3.

Christina G Phreaner, Mark A Williams, and R Michael Mulligan. Incomplete editing of *rps12* transcripts results in the synthesis of polymorphic polypeptides in plant mitochondria. *The Plant Cell*, 8(1):107–117, 1996. doi: 10.1105/tpc.8.1.107.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814, 2010.

Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

Zbigniew Wojna, Vittorio Ferrari, Sergio Guadarrama, Nathan Silberman, Liang-Chieh Chen, Alireza Fathi, and Jasper Uijlings. The devil is in the decoder: Classification, regression and gans. *International Journal of Computer Vision*, 127(11):1694–1706, 2019. doi: 10.1007/s11263-019-01170-8.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, vol. 1, 2016.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. doi: 10.1007/s13748-016-0094-0.

Rich Caruana, Steve Lawrence, and Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems*, pages 402–408, 2001.

Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11, 2017.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208, 2010.

Stefan Binder, Anita Marchfelder, and Axel Brennicke. RNA editing of tRNA^{Phe} and tRNA^{Cys} in mitochondria of *Oenothera berteriana* is initiated in precursor molecules. *Molecular and General Genetics MGG*, 244(1): 67–74, 1994. doi: 10.1007/BF00280188.

L Mareéchal-Drouard, D Ramamonjisoa, A Cosset, JH Weil, and A Dietrich. Editing corrects mispairing in the acceptor stem of bean and potato mitochondrial phenylalanine transfer RNAs. *Nucleic acids research*, 21 (21):4909–4914, 1993. doi: 10.1093/nar/21.21.4909.

Bingwei Lu, Robin K Wilson, Christina G Phreaner, R Michael Mulligan, and Maureen R Hanson. Protein polymorphism generated by differential RNA editing of a plant mitochondrial *rps12* gene. *Molecular and cellular biology*, 16(4):1543–1549, 1996. doi: 10.1128/mcb.16.4.1543.

Robin K Wilson and Maureen R Hanson. Preferential RNA editing at specific sites within transcripts of two plant mitochondrial genes does not depend on transcriptional context or nuclear genotype. *Current genetics*, 30(6):502–508, 1996. doi: 10.1007/s002940050162.

Wolfram Georg Brenner, Malte Mader, Niels Andreas Müller, Hans Hoenicka, Hilke Schroeder, Ingo Zorn, Matthias Fladung, and Birgit Kersten. High level of conservation of mitochondrial rna editing sites among four populus species. *G3: Genes, Genomes, Genetics*, 9(3):709–717, 2019. doi: 10.1534/g3.118.200763.

Kenneth H Wolfe, Wen-Hsiung Li, and Paul M Sharp. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences*, 84 (24):9054–9058, 1987. doi: 10.1073/pnas.84.24.9054.

David Choury and Alejandro Araya. RNA editing site recognition in heterologous plant mitochondria. *Current genetics*, 50(6):405–416, 2006. doi: 10.1007/s00294-006-0100-3.

Milena Groth-Malonek, Ute Wahrmund, Monika Polsakiewicz, and Volker Knoop. Evolution of a pseudogene: exclusive survival of a functional mitochondrial *nad7* gene supports haplomitrium as the earliest liverwort lineage and proposes a secondary loss of RNA editing in marchantiidae. *Molecular Biology and Evolution*, 24 (4):1068–1074, 2007. doi: 10.1093/molbev/msm026.

Nils Knie, Felix Grewe, Simon Fischer, and Volker Knoop. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns–a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. *BMC evolutionary biology*, 16(1):1–12, 2016. doi: 10.1186/s12862-016-0707-z.

M Schallenberg-Rüdinger and V Knoop. Coevolution of organelle RNA editing and nuclear specificity factors in early land plants. In *Advances in Botanical Research*, volume 78, pages 37–93. Elsevier, 2016. doi: 10.1016/bs.abr.2016.01.002.

Danny W Rice, Andrew J Alverson, Aaron O Richardson, Gregory J Young, M Virginia Sanchez-Puerta, Jérôme Munzinger, Kerrie Barry, Jeffrey L Boore, Yan Zhang, Claude W dePamphilis, et al. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science*, 342(6165):1468–1473, 2013. doi: 10.1126/science.1246275.

Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018. doi: 10.1093/bioinformatics/bty560.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357, 2012. doi: 10.1038/nmeth.1923.

Anja Zehrmann, Johannes A van der Merwe, Daniil Verbitskiy, Axel Brennicke, and Mizuki Takenaka. Seven large variations in the extent of RNA editing in plant mitochondria between three ecotypes of *Arabidopsis thaliana*. *Mitochondrion*, 8(4):319–327, 2008. doi: 10.1016/j.mito.2008.07.003.

James D Stone and Helena Storchova. The application of RNA-seq to the comprehensive analysis of plant mitochondrial transcriptomes. *Molecular Genetics and Genomics*, 290(1):1–9, 2015. doi: 10.1007/s00438-014-0905-6.

Zhiqiang Wu, James D Stone, Helena Štorchová, and Daniel B Sloan. High transcript abundance, RNA editing, and small RNAs in intergenic regions within the massive mitochondrial genome of the angiosperm *Silene noctiflora*. *BMC genomics*, 16(1):938, 2015. doi: 10.1186/s12864-015-2155-3.

Helena Štorchová, James D Stone, Daniel B Sloan, Oushadee AJ Abeyawardana, Karel Müller, Jana Walterová, and Marie Pažoutová. Homologous recombination changes the context of cytochrome b transcription in the mitochondrial genome of *Silene vulgaris* KRA. *BMC genomics*, 19(1):1–17, 2018. doi: 10.1186/s12864-018-5254-0.

Peng Zheng, Dongxin Wang, Yuqing Huang, Hao Chen, Hao Du, and Jumin Tu. Detection and analysis of C-to-U RNA editing in rice mitochondria-encoded ORFs. *Plants*, 9(10):1277, 2020. doi: 10.3390/plants9101277.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Philipp Gerke, Péter Szövényi, Anna Neubauer, Henning Lenz, Bernard Gutmann, Rose McDowell, Ian Small, Mareike Schallenberg-Rüdinger, and Volker Knoop. Towards a plant model for enigmatic u-to-c rna editing: the organelle genomes, transcriptomes, editomes and candidate rna editing factors in the hornwort anthoceros agrestis. *New Phytologist*, 225(5):1974–1992, 2020. doi: 10.1111/nph.16297.

Jeffrey P Mower. Modeling sites of RNA editing as a fifth nucleotide state reveals progressive loss of edited sites from angiosperm mitochondria. *Molecular biology and evolution*, 25(1):52–61, 2008. doi: 10.1093/molbev/msm226.

Thomas D Schneider, Gary D Stormo, Larry Gold, and Andrzej Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431, 1986. doi: 10.1016/0022-2836(86)90165-8.

Bastian Oldenkott, Yingying Yang, Elena Lesch, Volker Knoop, and Mareike Schallenberg-Rüdinger. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in *Escherichia coli*. *Communications biology*, 2(1):1–8, 2019. doi: 10.1038/s42003-019-0328-3.

Peter Kindgren, Aaron Yap, Charles S Bond, and Ian Small. Predictable alteration of sequence recognition by RNA editing factors from *Arabidopsis*. *The Plant Cell*, 27(2):403–416, 2015. doi: 10.1105/tpc.114.134189.

Mizuki Takenaka, Daniil Verbitskiy, Johannes A van der Merwe, Anja Zehrmann, Uwe Plessmann, Henning Urlaub, and Axel Brennicke. In vitro RNA editing in plant mitochondria does not require added energy. *FEBS letters*, 581(14):2743–2747, 2007. doi: 10.1016/j.febslet.2007.05.025.