Deciphering the rules of mRNA structure differentiation *in vivo* and *in vitro* with deep neural networks in *Saccharomyces cerevisiae*

Haopeng Yu[1,2], Wenjing Meng[1,2],Yuanhui Mao[1,2],Yi Zhang[1,2],Qing Sun[1,2],Shiheng Tao[1,2,*]

[1]College of Life Sciences and State Key Laboratory of Crop Stress Biology in Arid Areas, Northwest A&F University, Yangling, Shaanxi, China

[2]Bioinformatics Center, Northwest A&F University, Yangling, Shaanxi, China

[*]Corresponding Author: shihengt@nwsuaf.edu.cn

Abstract

In vivo mRNA structure is influenced by versatile factors in translation process, resulting in a significant differentiation of mRNA structure from that *in vitro*. Due to multiple factors caused the differentiation of *in vivo* and *in vitro* mRNA structures, it was tough to perform a further accuracy analysis of mRNA structures in previous studies. In this study, we elaborated a deep neural networks (DNN) model to predict structural stability of mRNA *in vivo*, by fitting six quantifiable features that may affect mRNA folding: ribosome density, minimum folding free energy, GC content, mRNA abundance, ribosomal initial density and position of mRNA structure. Simulation mutations of mRNA structure was designed and then fed into trained DNN model to compute their structural stability state. We found the unique effects of these six features on mRNA structural stability in vivo. Strikingly, a double-sided effect of ribosomal density on mRNA structural stability was identified. It could be speculated that higher stable structure leaded to an accumulation of ribosomes in the region of low ribosomal density, whereas higher ribosomal density resulted in unstable mRNA structure. Additionally, the trend and extent of these six features affecting mRNA structural stability *in vivo* could also be accurately predicted by DNN model. Recruitment of the deep neural networks provides a new paradigm to decipher the differentiation of mRNA structure *in vivo* and *in vitro*. This knowledge on mechanisms of influencing factors on mRNA structural stability will facilitate the design and functional analysis of mRNA structure *in vivo*.

Introduction

mRNA is a key component of the translation system. It not only carries protein coding information, but also an important vector for translation regulation information by folding into mRNA structures[1]. Notably, mRNA structures at 5'UTR and RBS region have a significant effect on translation efficiency[2,3]. Recent studies have shown that in-cell RNA structures are quite different characteristics *in vivo* and *in vitro* (DMS-probing[4,5], DMS-MaPseq[6], icSHAPE[7] and SHAPE-MaP[2]), some general conclusions have been drawn: the number of *in vivo* mRNA structures was far less than that *in vitro*[4], and mRNA structures in highly expressed genes, especially long-range base pairing structures, appeared translation destabilizes[2]. However, it challenges to carry out this difference further and accurately to the specific analysis of each mRNA structure, owing to the dynamic and uniqueness of the mRNA structure itself, that is, each mRNA structure has a 'personal' distinctive structure.

In order for the ribosome to proceed translation smoothly, it is necessary to unwind the mRNA structure firstly, and a complex interaction between translating ribosome and mRNA structure occurs during translation. Meanwhile, translating ribosome is one of the most important factors that cause structural differences *in vitro* and *in vivo*. Ribosome profiling is a method for determining the exact position of ribosomes in transcriptome with single-nucleotide resolution during translation by deeply sequencing ribosomal protected mRNA fragments[8,9]. This approach can resolve the interaction between translating ribosome and mRNA structure during translation effectively, for example, the mRNA structure in CDS region influences protein co-translational

folding by affecting the efficiency of ribosome translation or is more extreme, with causing ribosomal pauses[10-12]. Besides, factors that affect the mRNA structural stability *in vivo* could been primarily classified below several parts: gene features, for instance, mRNA abundance and translation initiation rate; structure features, including mRNA structural strength like minimum free energy of structural subsequences; and other factors such as mRNA structure location in gene. Due to the multifaceted influence of these factors, precise prediction mRNA structural stability became a great challenge.

Deep learning, or artificial neural networks, is a type of machine learning that solves complex problems by learning "big data". Deep neural network (DNN), a type of deep learning, has multiple hidden layers and units, which can model complex non-linear relationships, was widely used in natural language processing[13], speech recognition[14] or the sensational alpha Go[15]. At present, deep learning has been applied in biological 'big data', such as detection of breast cancer based on histological images[16], analysis of ribosomal stalling sites based on ribosome profiling[17] and prediction of DNA- and RNA-binding proteins[18]. We proposed a novel deep learning DNN model to fit mRNA structure features, subsequently, the structural state *in vivo* was predicted, revealing the impact range of these features on the *in vivo* mRNA structural stability.

As the simplest eukaryote, *S. cerevisiae* was first and most commonly applied to the research of mRNA structure *in vivo* and ribosome occupancy[4,5,8]. Therefore, in this study, *S. cerevisiae* was selected as the research species to carry out a series of work on the analysis of mRNA structure *in vivo*. Among *in vivo* experiments of detecting mRNA structure, DMS-probing determination of

mRNA structure in *S. cerevisiae*, by setting a threshold value of Gini coefficient and Pearson correlation, was demonstrated to be effective[4]. Thus, we proposed marking the *in vitro* and *in vivo* mRNA structures by this method and divide them into two classes (0 and 1, structures with stable and unstable trend *in vivo*), then DNN model was trained with these data. After having a precise model to predict structural state *in vivo*, we can mutate the different features of these mRNA structures and decipher the particular impact of these features on mRNA structural stability (Figure 1a).

Results

Constructing a Deep Learning Model with 6 mRNA Structure Features.

We proposed a deep neural network (DNN) model, which fit mRNA structures differentiation *in vitro* and *in vivo* with the quantifiable features that may influence structural stability, thereby realizing the ability to predict mRNA structural stability *in vivo* (Figure 1a). Unlike previous mRNA structure prediction models aiming at constructing RNA structures[19], this deep learning model was mainly used to predict whether mRNA structures *in vivo* performed stable or unstable state. DMS profiling[4] is a chemical probing method, methylated adenine and cytosine bases in single-stranded RNA, especially suggests a global feature of *in vivo* and *in vitro* as well as denatured mRNA structures by measuring Gini index[20] and Pearson correlation coefficient. Higher Gini index revealed stronger structural region, and vice versa weaker structural region. Normally, due to a variety of factors such as translating ribosomes, the *in vitro* structure was

usually stronger than that *in vivo*[4]. Therefore, 'disappear trend' (DIS) was designated to measure the differences of mRNA structure *in vivo* and *in vitro*, which was equal to Gini index of mRNA structures *in vitro* minus that *in vivo*. In this study, the whole mRNA structures of *Saccharomyces cerevisiae in vitro* were classified into two classes by DIS. Class 1 was structures beyond mean plus standard deviation (SD), which represented structures with unstable state. Class 0 was structures below the mean minus SD, which yielded the structure profiles of a stable state (Figure 1a). After multiple thresholds of classification were tested, one times SD was determined for ensuring the amount of training sets and deep learning model accuracy.

Stable and Unstable mRNA structures could appear on the same gene, meanwhile several structural features differences might be interpreted as the existence of these two states (Figure 2a).Previous studies found that a versatility of factors can affect mRNA structure *in vivo*, such as ribosome density (RD), the gene expression level, tRNA levels and mRNA structure stability itself[2,11,12,21]. Six quantitative features of them were chose for modeling: gene features such as reads per kilo base per million mapped reads (RPKM) of RNA-Seq implied mRNA abundance and initial translation ribosome density (INI) calculated by ribosome profiling data; structure subsequence features such as minimum free energy (MFE) and GC contents of subsequence(GC) represent mRNA structure stability; other features such as structural relative position in gene (POS) and ribosome density of structure region (RD). It's worth mentioning that RD and INI value in this project were averaged from five different studies of ribosome profiling data in wild-type yeast[22-26]. It was revealed that mRNA structure DIS value was remarkably positive

relevant to RD and RPKM, implying that high ribosomal density and high mRNA abundance were not conducive to the formation of the stable mRNA structure (Figure 1b), the results were consisted with observations from prior studies[2,27]. Six features showed significant differences (p value<2.2e-16) of two mRNA structure classes. However, only the visual difference was revealed on the RD, RPKM, and GC distributions (Figure 1C). It could be initially concluded that these six features do not affect the mRNA structural stability alone, and it was tough to resolve their contribution to mRNA structural stability from the correlation and distribution. Therefore, we adopted DNN modeling and deciphering the relationship between these complex regulatory factors.

We established the DNN sequential model by Tensorflow deep learning framework[28]. By adjusting the hyperparameters of the model, the final model has 11 hidden layers (10×64 and 1×1), the activation function adopted ReLU[29] and Sigmoid, the optimization function adopted Adam[30] to accelerate the training process, and batch normalization[31] was added in order to prevent overfitting of the model training (Figure 5, Table S1). The average accuracy of the trained DNN model of 10-fold cross-validation reached 99.92%. The DNN model that can accurately predict *in vivo* mRNA structure state through *in vitro* structural features was established.

Ribosome density revealed a double-sided effect on mRNA structure.

Translating ribosome is a critical factor that affects mRNA structural stability. Previous studies have demonstrated that a negative correlation between ribosomal translation efficiency and *in vivo* mRNA structure stability, that is, high translation efficiency corresponds to a low structural region[2,5,27,32]. Two main hypotheses of this correlation in the preceding studies were proposed. One is that the mRNA structure forms an energy barrier that reduces the translation efficiency of ribosomes. The other is the case that the translation of ribosomes causes the *in vivo* structure to fail to fold into a structure, especially long-paired structures. In order to solve this causal relationship between ribosome and RNA structures, we performed single-factor gradient mutations on the ribosome density (RD) of randomly selected 5000 stable structure regions and 5000 unstable structural regions, based on the assumptions that no significant changes in other parameters after RD mutation (Figure 2d). After the RD of 64% unstable structures were weakened to the minimum value, they were transformed into a stable structure. It's also revealed that double-side effect of RD impact mRNA structural stability: Low RD was positively correlated with structural stability and this tendency was more obvious in the original stable structure (Figure 2d, red line). Contrarily, RD in a higher range was negative correlated with structural stability. This double-side effect of ribosome density interpreted that stronger structures hindered ribosomes translation especially high translation rate ribosome (equal to low ribosome density), which would slow translation rate and leaded to increase ribosome density. However, it was arduous to re-increase its density of high ribosome density region by mRNA structural barrier, but instead unwound the structure and leading to *in vivo* mRNA structural instability.

We proposed the concept of “feature based structure transition region” of mRNA structure *in vivo*, which performed by determining the inflection point of DNN model’s prediction of various features’ gradient mutation. Two “RD based structure transition regions” were revealed from the single-factor mutation results. It was shown that the positive correlation region was below -2.3656 and the negative correlation region were from -2.36568 to 0.149073 of unstable structures and -2.36568 to 2.26147 of stable structure, representing mRNA structure tended to be stable or unstable as RD increasing (Figure 2d, Table S3). Pronounced effect of higher ribosomal density on the stable mRNA structure was showed compared to the unstable structural transformation region. It was worthwhile to mention that the boundary values of these transition regions were decided basing on the inflection points of structural transformed curve based on five studies’ ribosome profiling data with data process method (Corresponding raw data was showed in Table S4).

Further analysis divided the area into four groups (Figure 2d). The originally stable structures, G1 and G2 groups were closer to the gene 5’ end and preferred to locate in lower expression genes (low RPKM) comparing to G3 and G4 groups (Figure 2e). Gene upstream preference and low RPKM was also reflected in the group of remaining stable structures as same as group G1 and G2, additionally, the lower initial efficiency translation was obviously observed. The original unstable four-group structure has mild difference, but the group of remaining unstable structures was more focused on highly expressed genes (Figure 2e). Single-factor gradient mutations were difficult to achieve in the “real” experiment, thus the use of high-precision models to predict the general

correlation between *in vivo* structures and RDs through simulated-mutations was exactly the effective way to decipher this rule.

mRNA structure prone to be stable at 5'-end gene.

Our trained model provided accurately prediction of yeast mRNA structural state *in vivo* through six features of the structure. Here, we sought to establish the relationship between mRNA structure and its relative position in the gene CDS region. Previous studies analyzed the *in vitro* structure of various species to derive a tendency that stable RNA structure lean to form at downstream of the translation initiation codon (AUG), which might favor ribosome formation in ribosome 'Ramp'[33-35]. This feature could also be seen from the statistics of the 2 classes (Stable and unstable) mRNA structure in the yeast, with a higher proportion of stable mRNA structure at the 5' and 3' ends of the gene (Figure 3b). We changed the relative position (POS) of the mRNA structure on gene by gradient mutations and found that the state of 86.38% originally unstable structures transformed to stable after transferring structures to the downstream of the AUG (Figure 3a). As the mRNA structure was being transferred to gene upstream, the proportion of stable structures increased. The mRNA structure of the middle part of the gene remained basically unchanged, and the stable state of the structure at the 3' end (beyond 0.8) revealed the slight trend of stable-transformed especially to unstable structures. It was indicated that stable mRNA structure tended to form at the 5' and 3' ends of the gene, which was consistent with previous studies[33,35], and stable structure itself was easily formed *in vivo* at upstream of the gene. Combined with the previous RD results, it can be concluded that the mRNA structure and RD at

gene 5' end prone to have a "cooperate effect", in other terms, the structure maintained the ribosome density to a certain level for more efficient translation, and the RD also remained in a certain range to ensure mRNA structure stability. Therefore, we concluded that region below approximately 0.3 was regarded as "POS based positive structural transition region", signifying that mRNA structure performed more stable in the gene upstream (approximately <30%). Considering that it's arduous to communicate all the changes of the whole features, mutations in this fraction also carried out using single-factor mutation, involving a scheme that mutation only occurred with the feature of relative position (POS). Furtherly, the proposal of co-mutation of the position and RD was designed by changing the POS and RD features to the first structure of the 5' end of gene, similar result was obtained as well (Figure 3c). Small fraction of unstable structures in position mutations revealed unchanged state and obvious difference of RPKM may account for this appearance (P value < 0.0001, Figure 3d).

mRNA abundance and initiation RD effect on mRNA structure stability.

Two primary characteristics of the gene affecting the structural stability: mRNA abundance and translation initiation ribosome density (could be translated into initiation rate through data processing), which were obtained by RNA-Seq data and ribosome profiling data respectively. mRNA abundance (RPKM value in this project) represented mRNA copy number, which has a certain correlation but not equaling to the protein synthesis efficiency collected by ribosome profiling[36-38]. From the ratios of different mRNA abundance genes of two classes (class 0 and class 1)structures, it was presented that high mRNA abundance was more likely to cause

structural instability while there was mild difference in low mRNA abundance (Figure 3f).The mRNA structure served as a barrier for ribosome translation and it is naturally tough to stay stable for maintaining the translation efficiency of highly expressed genes. As there was no strong "translation task" for low-expressed genes, structural instability was not strongly needed, hence, the structure tended to be preferably stable. According to single-factor mutation, the effect of mRNA abundance on mRNA structural stability can be divided into positive and negative correlation regions (Figure 3e). The statistical result of fig.3f was probably a good confirmation of the negative correlation region, reflecting that few stable structures appeared on high expression genes of authentic yeast. However, the positive correlation region couldn't be well explained by fig.3f, assumed that low abundance genes might be affected by other features. mRNA structure and translation efficiency were involved by particular translation initial rate[21]. In our previous correlation analysis, we also found a strong correlation between the ribosomal density in the structural region and the ribosomal translation initiation density of the gene (Figure 1b).mRNA structure was more likely to be stable when the initial ribosome density was either in the lower range (below -3.71493) or in the especially higher extent (from 2.057627 to 3.608730), however, it was apt to be unstable when a subsequent particularly high initial ribosome density reached (beyond 3.608730) (Figure 3g).The lower initial density represented a higher initial rate and higher density was always accompanied with a lower initial rate. It was assumed that lower initial rate was prone to keep the mRNA structural stability while extremely low initial rate was prone to cause structural instability, which could be interpreted as the mRNA structure on low initial rate gene may effectively highlight the effects of its translational barrier function. In this

part, structural instability was caused by the extremely low initial rate, which was consistent with the conclusion in ribosome density mentioned above that excessively high ribosome density may also lead to structural instability. We have also tried to mutate two features involving with mRNA abundance and translation initiation rate at the same time, but the structural transformation trend fluctuated greatly and no obvious regularity could be concluded.

The Effect of the mRNA Structure Subsequence composition on its stability.

The minimum free energy (MFE) and GC content was used to evaluate the strength of mRNA structure base on its sequence composition[21,33,39,40]. MFE were significantly negative correlated with GC and the Pearson correlation reached -0.45 (Figure 1b). Compared with mutations on other features observed previously, the result of single-factor mutations presented very special both MFE and GC (Figure 4a,4b). The MFE of unstable structure below -4.14 (corresponding to -47.83 kJ·mol-1) became stable, and the trend of stable structure tendency showed no remarkable change. With a mean value 0 (corresponding to -19.18 kJ·mol-1) as the inflection point, the transition of stable and unstable structures had a "V" trend. When the changes of normalized MFE value headed to both sides of the mean 0, the stable and unstable structure would shift toward the opposite state to itself. In general, the reduction of MFE means the enhancement of the mRNA structure *in vitro*, but it couldn't fully explain the results of this study. When the MFE was less than the mean value 0, the originally unstable structure tended to be stable as the MFE decreased, which was in line with conventional wisdom. Unexpectedly, the instability of the original stable structure was uniquely observed in this condition, which was precisely a

confirmation of the conclusion that there existed a double-side effect between the mRNA structure and ribosome density referred to above fig. 2d, supposing that when the original stable mRNA structure continuously to be strengthened its stability by reducing MFE, an increasing of the ribosome density was appeared until reaching a threshold and the excessive ribosome occupancy would in turn unwind the mRNA structure causing unstable state. Further, in terms of single-factor mutation on GC contents, a novel finding was discovered that unstable structures at low GC content were more likely to become stable in particularly. Unlike other influencing factors mentioned above, changes of mRNA structure sequence itself leading to a direct influence on its stability, while the mean value (0 of MFE corresponding to -19.18 kJ·mol-1 and 40.25% of GC contents) was the optimum condition that endowed mRNA structure of keeping its original state.

We proposed that the mRNA structure tolerance could be tested by raising and lowering MFE of the mRNA structure sequence. Randomly selected unstable structures and stable structures were used to carry out gradient-decrease MFE mutations and gradient-rising MFE mutations respectively. As is shown in fig.4c and fig.4d, with the sequential introduction of four gradient mutations, altered state of mRNA structure was divided into four classes of T1-T4, and which were step-up representatives indicating enhancement of the mRNA structure tolerance gradually (Figure 4c and 4d). In conclusion, high gene expression levels and high ribosome densities could lead to more tolerant unstable structures, maintaining their own state of instability, thus make it more difficult to change unstable structures through mutations; High gene expression levels and high ribosome densities could make the stability of the stable structure less tolerant, weakening

the state of its own stability, resulting in a state that was more likely to change the stable structure through mutation; Above all, it was conclusively proved gene expression levels and ribosomal density contributed to structural tolerance.

Phylogenetic conserved mRNA structures prone maintained their *in vivo* state.

Phylogenetic conserved mRNA structures were supported by DMS-seq and inspected by phylogenetic conservation analysis[4]. Positive and negative mutations in the three features, RD, RPKM and INI, of mRNA structure revealed that the conserved mRNA structure was more able to maintain its own stability state compared to random structures when subjected to the same degree of mutation (Figure 4e). Consequently, conserved structure was more tolerant to fluctuations in the intracellular environment than the random structure. It was worthwhile mentioning that the conserved structure not only had a stable state *in vivo*, but also had an unstable structure measured by DIS and both of them met the tolerant results. It raised the possibility that unstable structure *in vivo* was also involved a certain function during translation.

Discussion

Considering the invention of large-scale, high-throughput methods for *in vivo* mRNA structure detection, previous studies have statistically derived the general conclusions of *in vivo* and *in vitro* mRNA structural differences, such as, the number of *in vitro* mRNA structures was more than that *in vivo*, and the *in vivo* and *in vitro* mRNA structural differences were mainly caused by ribosome translation[2,4-7]. Due to multiple factors caused the differences of *in vivo* and *in vitro* structures, it was tough to perform a further accuracy or "personalized analysis" of mRNA structures in previous studies.

In this study, we used a novel deep learning approach to accurately model with labeled mRNA structures in yeast and six quantitative structural features (ribosome density of structure region, minimum free energy and GC content of subsequence, gene mRNA abundance and ribosomal initial density, structural relative position). Laterly, trained deep neural network (DNN) model could accurately predict the stable or unstable state of each mRNA structure *in vivo* (accuracy reached to 99.92%). Then the rules explaining the effect of these six features on mRNA structural stability were resolved, involving with simulated-mutations of these six features and subsequently predicting by DNN model. We found double-sided effect between ribosomal density (RD) and mRNA structural stability, i.e., mRNA structural stability was positively correlated with RD in the low ribosome density region and negatively correlated with RD in the high ribosome density region. It might be explained that the mRNA structure was more capable of exerting an energy barrier in the low ribosome density region, consequently, the higher

ribosome density was caused by more stable mRNA structure; and the mRNA structure *in vivo* was arduous to show stability in the high ribosome density region, owing to the high ribosome occupancy on mRNA structure region. In prior studies, the relationship of RD and mRNA structural stability was not uniform[2,4,27]. In this study, the presumed theory“double-sided effect”on RD and mRNA structural stability was emerged and evidenced by ribosomal density analysis, which was origin from a sparkling thought inspired by the inflection point of RD value at the ‘V’ shape (Figure 2d). Meanwhile, we found that the mRNA structure at the front of the gene presented more stable state, which was consisted with observations from prior studies[33,35]. What’s more, mRNA abundance and initial ribosome density, as members of genetic characteristics, accounted for its various effects on the stability of the mRNA structures to varying degrees. In addition, mutation of mRNA structural subsequence itself could also directly cause changes in the structural state, showing that the mean values of MFE and GC were the optimum condition allowing mRNA structure of keeping its original state. Also, mRNA structural tolerance of MFE adjustment depended more on other factors such as RD or mRNA abundance. Eventually, phylogenetic conserved mRNA structures were prone to be more tolerant in the structural features fluctuations and more capable of keeping their own state when compared with the random mRNA structures. It was worth mentioning that the effective range of these features on the mRNA structural stability could have been provided through the DNN model prediction.

This is the first study to use deep learning approach to decipher the rules of the mRNA structure differentiation the *in vitro* and *in vivo*. It had been proven that deep neural networks could be

used to solve complex regulatory issues in the translation process in this study. The DNN model accurately predicted the relationship between these six mRNA structural features and its structural state, meanwhile, it also resolved the influence range of these features on structural stability. However, some limitations are worth noting. Previous studies have shown that codon bias and tRNA abundance could affect the ribosome translation rhythm and afterwards influence the mRNA structural stability *in vivo*[12,21,41]. Codon bias and tRNA abundance were not considered separately in our model, on causality that the ribosome translation information of these two factors was already carried through ribosome density data. Our results are encouraging and should be validated in a larger amount of mutant yeast species. Future works should therefore focus on the following points: codon biases and tRNA abundance substituted for ribosome density, as two respective features, were involved in impact analysis on mRNA structural stability; And more species, like arabidopsis and mouse, as well as more detection methods, like SHAPE-MaP and icSHAPE, should be considered in state prediction of mRNA structure *in vivo*.

Materials and Methods

Raw Data Preprocessing. *Saccharomyces cerevisiae* RNA structure data was obtained from DMS-profiling experiment (containing *in vivo*, *in vitro* and denatured data sets), which was published in the NCBI GEO database, accession number GSE458034. We settled DMS signals of specified genome position in each gene, trimmed G/T sites in gene sequence, then used slide windows of 50nt (full-length slide windows, i.e. after add G and T sited, were about 100nt) to divide trimmed-genes into regions, meanwhile filtered low DMS signal regions (<1 per nt). Gene sequences were retrieved from SGD database assembly R62. Multiple ribosome profiling data (RP) was combined to closer to the true *S. cerevisiae* ribosome occupancy. RP data of wild-type yeast was based on the following studies: Zinshteyn and Gilbert et al. 2013[22], Lareau et al. 2014[23], Albert et al. 2014[24] (BY and RM samples), Young et al. 2016[25] and Nissley et al. 2016[26](Rep1 and Rep2 samples). RNA-Seq data was denoted by Albert et al. 2014 (BY and RM samples)[24]. Then RP and RNA-Seq data were trimmed and mapped to the yeast genome (assembly R62) by Bowtie and normalized by RPKM of RNA-Seq data[42]. For the yeast fragments, the off-set 15 nucleotides from 5' end represented the P-site of ribosome's position[36]. Therefore, ribosome occupancy at nucleotides resolution was given by located the genome position of 15th nucleotide of ribosomal fragments. For all these RP data by different studies, ribosome footprint of corresponding position was summed.

Classified RNA Structures from DMS-Probing Data. We defined RNA structure by two metrics: Pearson's correlation coefficient ($r<0.55$) and the Gini coefficient ($Gini<0.14$) as

described method[4]. After removing the G/T sites in the sequence, the 50 nt sliding window was used to divide the CDS region of all genes. Firstly, we determined the structure *in vitro* by the threshold of r and Gini. In order to more reasonably analyzed the disappearance trend of RNA structure, we proposed a novel measure for mRNA structural region. DIS, which was equivalent to subtracting the *in vitro* and *in vivo* Gini coefficients of the structural region (G). For mRNA structural region of length *n*, the DMS signal per sites isS_i, and its DIS and Gini coefficient is:

$$\mathrm{DIS} = G_{vitro} - G_{vivo}$$

$$\mathrm{G} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|S_i - S_j\right|}{2n\sum_{i=1}^{n} S_i}$$

Calculation and Standardization Six Features of mRNA Structures. We chose six structural features as a measure of mRNA structure. The feature TE (translation efficiency) was the protein synthesis rate of structure region[43]; MFE was the minimum free energy of a local subsequence, calculated by by Vienna (v.2.1.9) package RNAfold[44] function and visualized by RNAstructure[45]; INI was the translation efficiency at gene initial region (+15nt); RPKM represented the relative number of transcripts of a gene and was calculated by RNA-Seq data; GC represented the GC content of the structural region sequence; POS was the relative position of the structural region on the gene, which was obtained by dividing the sequence length by the central position of the structural region. The distribution of these six features, of which three, TE, INI and RPKM were exponentially distributed, so a logarithmic process was performed. Prior to deep learning training, parameters (except Pos and GC) were normalized ($(x-\mu)/\sigma^2$) to make the model more accurate.

Established Deep Neural Networks Model. We established a deep neural networks model to

predict whether RNA structures *in vitro* would have stable or unstable tendency *in vivo*. By labeling known-state structures and predicting a sequence using these six features, a prediction model can be generated. DIS was used to evaluate the degree of RNA structural destabilization *in vivo* and its distribution was similar to a normal distribution. To ensure a sufficient number of training set and development set (Train/Dev set), we decided to use ($\mu \pm \sigma$) as the threshold value. Class 1 was defined as DIS greater than $\mu+\sigma$, meaning that such structures had an unstable tendency *in vivo*; class 0 was tagged as DIS greater than $\mu-\sigma$, representing that such structures had a stable tendency *in vivo*. Therefore, it became a binary classification problem. In this project, we used the Tensorflow[28], a deep learning framework created by Google Brain team, and Keras API[46] to construct a DNN model, solving the complex problem of RNA structural state prediction *in vivo*. This framework can greatly simplify the process of building the DNN sequential model by using the built-in functions. By adjusting the hyperparameters of the model, the final model has 11 hidden layers (10×64 and 1×1), the activation function adopted ReLU[29] and Sigmoid, the optimization function adopted Adam[30] to accelerate the training process, and batch normalization[31] was added in order to prevent overfitting of the model training (Figure 5). Final parameters of the model were determined using ROC curve (Figure S2, Table S1) and 10-fold cross-validation was used to test the accuracy of the model.

Single factor mutation and prediction. We performed single-factor mutation experiments on the six features of mRNA structure to test the effect of these features on the structural stability. From the development set, 5,000 unstable structures and 5000 stable structures were randomly

selected. Each of the other parameters was subject to 100 equal parts gradient mutation ranging from its minimum value to its maximum. After that, the model of the mutated structure was predicted again using the DNN model. Transformed mRNA structure was defined as the change of mRNA structure state (i.e., be classified to another group) after the mutation. The ratio of transformed mRNA structure was equal to the number of transformed structures divided by the total structures. In the results of gradient mutation, with the increase of each one of the six features, the increase of mRNA structural stability was named as “positive correlated transition region”, and the decline of mRNA structural stability was regarded as “negative correlated transition region”.

Ribosome density feature analysis. mRNA structures were divided into 5 groups, G1 to G4 and “stable” or “unstable” structures. G1 to G4 groups were according to the midpoints of “ribosome density based positive and negative correlated transition region”. mRNA structure remained stable during the mutation of the ribosome density gradient was named “stable”, and the structure that remained unstable was called “unstable”.

mRNA structural tolerance analysis of MFE. Randomly selected 5000 high MFE mRNA structures and 5000 low MFE mRNA structures (MFE threshold: $\mu \pm \sigma$) were performed four times gradient mutations (decrease 1 SD of MFE per mutation for high MFE structures and increase 1 SD of MFE for low MFE structures). The state of mRNA structure after each mutation was recorded as the basis for measuring structural tolerance.

The original mRNA structure state was changed at the first mutation was called “T1”, that is, the

least tolerate mRNA structure and the state transformed until the 4th mutation was named “T4”, that is, the most tolerant mRNA structure. mRNA structures with constant state during the mutation process were not considered and minimal back-mutation structures was trimmed either.

Tolerance analysis of conserved structures. Conservative structural data from Rouskin’s study[4]. Random structure was randomly extracted from the total mRNA structure and excluded the list of conserved structures. Three parameters, RD, RPKM and INI, that may fluctuate within the cell, an one-fold standard deviation mutation was performed (Up was +SD and down was -SD).

ACKNOWLEDGMENTS. This work was supported by the National Natural Science Foundation of China (Grants 3177080732). The authors would like to thank Professor Shiheng Tao for valuable guidance of the writing of this thesis and critically reviewing the manuscript.

References:

1 Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet***15**, 469-479, doi:10.1038/nrg3681 (2014).

2 Mustoe, A. M. *et al*. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell***173**, 181-195 e118, doi:10.1016/j.cell.2018.02.034 (2018).

3 Espah Borujeni, A. *et al*. Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Res***45**, 5437-5448, doi:10.1093/nar/gkx061 (2017).

4 Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature***505**, 701-705, doi:10.1038/nature12894 (2014).

5 Ding, Y. *et al*. *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature***505**, 696-700, doi:10.1038/nature12756 (2014).

6 Zubradt, M. *et al*. DMS-MaPseq for genome-wide or targeted RNA structure probing *in vivo*. *Nat Methods***14**, 75-82, doi:10.1038/nmeth.4057 (2017).

7 Spitale, R. C. *et al*. Structural imprints *in vivo* decode RNA regulatory mechanisms. *Nature***519**, 486-490, doi:10.1038/nature14263 (2015).

8 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science***324**, 218-223, doi:10.1126/science.1168978 (2009).

9 Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet***15**, 205-213, doi:10.1038/nrg3645 (2014).

10 Sabi, R. & Tuller, T. A comparative genomics study on the effect of individual amino acids on ribosome stalling. *Bmc Genomics***16 Suppl 10**, S5, doi:10.1186/1471-2164-16-S10-S5 (2015).

11 Zur, H. & Tuller, T. Strong association between mRNA folding strength and protein abundance in S. cerevisiae. *EMBO Rep***13**, 272-277, doi:10.1038/embor.2011.262 (2012).

12 Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A***107**, 3645-3650, doi:10.1073/pnas.0909910107 (2010).

13 Doan, S., Conway, M., Phuong, T. M. & Ohno-Machado, L. Natural language processing in biomedicine: a unified system architecture overview. *Methods Mol Biol***1168**, 275-294, doi:10.1007/978-1-4939-0847-9_16 (2014).

14 Hinton, G. *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition. *Ieee Signal Proc Mag***29**, 82-97, doi:10.1109/Msp.2012.2205597 (2012).

15 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature***529**, 484-+, doi:10.1038/nature16961 (2016).

16 Ciresan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. *Med Image Comput Comput Assist Interv***16**, 411-418 (2013).

17 Zhang, S. *et al.* Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning. *Cell Syst***5**, 212-220 e216, doi:10.1016/j.cels.2017.08.004 (2017).

18 Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol***33**, 831-838, doi:10.1038/nbt.3300 (2015).

19 Ouyang, Z., Snyder, M. P. & Chang, H. Y. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome research***23**, 377-387, doi:10.1101/gr.138545.112 (2013).

20 Wittebolle, L. *et al.* Initial community evenness favours functionality under selective stress. *Nature***458**, 623-626, doi:10.1038/nature07840 (2009).

21 Mao, Y., Liu, H., Liu, Y. & Tao, S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae. *Nucleic Acids Res***42**, 4813-4822, doi:10.1093/nar/gku159 (2014).

22 Zinshteyn, B. & Gilbert, W. V. Loss of a conserved tRNA anticodon modification perturbs cellular signaling. *PLoS genetics***9**, e1003675, doi:10.1371/journal.pgen.1003675 (2013).

23 Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife***3**, e01257, doi:10.7554/eLife.01257 (2014).

24 Albert, F. W., Muzzey, D., Weissman, J. S. & Kruglyak, L. Genetic influences on translation in yeast. *PLoS genetics***10**, e1004692, doi:10.1371/journal.pgen.1004692 (2014).

25 Young, S. K., Palam, L. R., Wu, C., Sachs, M. S. & Wek, R. C. Ribosome Elongation Stall Directs Gene-specific Translation in the Integrated Stress Response. *J Biol Chem***291**, 6546-6558, doi:10.1074/jbc.M115.705640 (2016).

26 Nissley, D. A. *et al.* Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to co-translational folding. *Nat Commun***7**, 10341, doi:10.1038/ncomms10341 (2016).

27 Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol***10**, 770, doi:10.15252/msb.20145524 (2014).

28 Abadi, M. *et al.* TensorFlow: a system for large-scale machine learning. *Software available from tensorflow.org* (2016).

29 Nair, V. & Hinton, G. E. in *International Conference on International Conference on Machine Learning.* 807-814.

30 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

31 Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal

Covariate Shift. 448-456 (2015).

32 Gorochowski, T. E., Ignatova, Z., Bovenberg, R. A. & Roubos, J. A. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res***43**, 3022-3032, doi:10.1093/nar/gkv199 (2015).

33 Mao, Y., Li, Q., Wang, W., Liang, P. & Tao, S. Number variation of high stability regions is correlated with gene functions. *Genome Biol Evol***5**, 484-493, doi:10.1093/gbe/evt020 (2013).

34 Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell***141**, 344-354, doi:10.1016/j.cell.2010.03.031 (2010).

35 Tuller, T. *et al.* Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome biology***12**, R110, doi:10.1186/gb-2011-12-11-r110 (2011).

36 Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell***147**, 789-802, doi:10.1016/j.cell.2011.10.002 (2011).

37 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature***473**, 337-342, doi:10.1038/nature10098 (2011).

38 Schwanhausser, B. *et al.* Corrigendum: Global quantification of mammalian gene expression control. *Nature***495**, 126-127, doi:10.1038/nature11848 (2013).

39 Mao, Y. *et al.* Genome-wide analysis of selective constraints on high stability regions of mRNA reveals multiple compensatory mutations in Escherichia coli. *Plos One***8**, e73299, doi:10.1371/journal.pone.0073299 (2013).

40 Zhang, J., Kuo, C. C. J. & Chen, L. A. GC content around splice sites affects splicing through pre-mRNA secondary structures. *Bmc Genomics***12**, doi:Artn 90 10.1186/1471-2164-12-90 (2011).

41 Supek, F. The Code of Silence: Widespread Associations Between Synonymous Codon Biases and Gene Function. *J Mol Evol***82**, 65-73, doi:10.1007/s00239-015-9714-8 (2016).

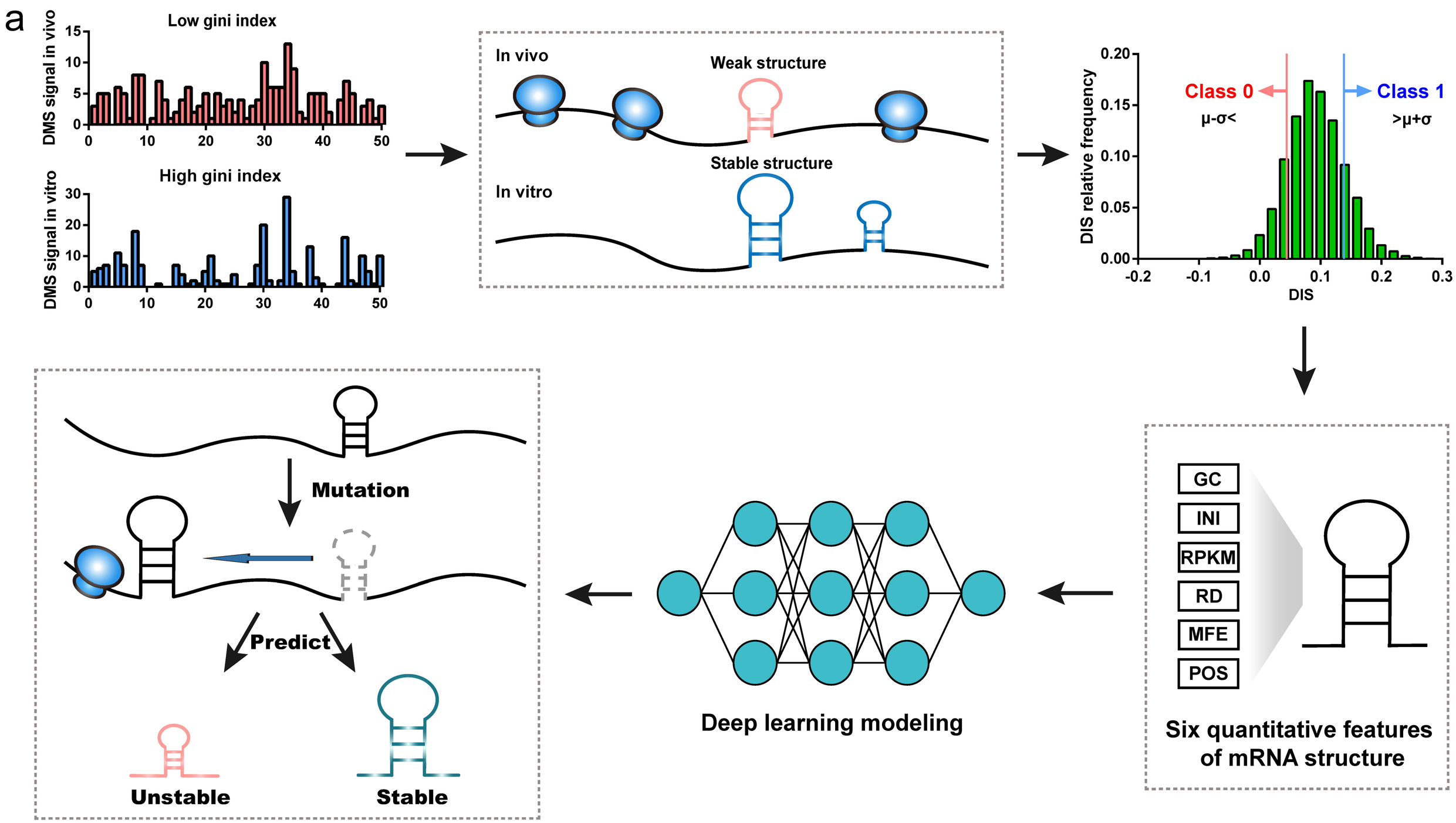
42 Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc***7**, 1534-1550, doi:10.1038/nprot.2012.086 (2012).

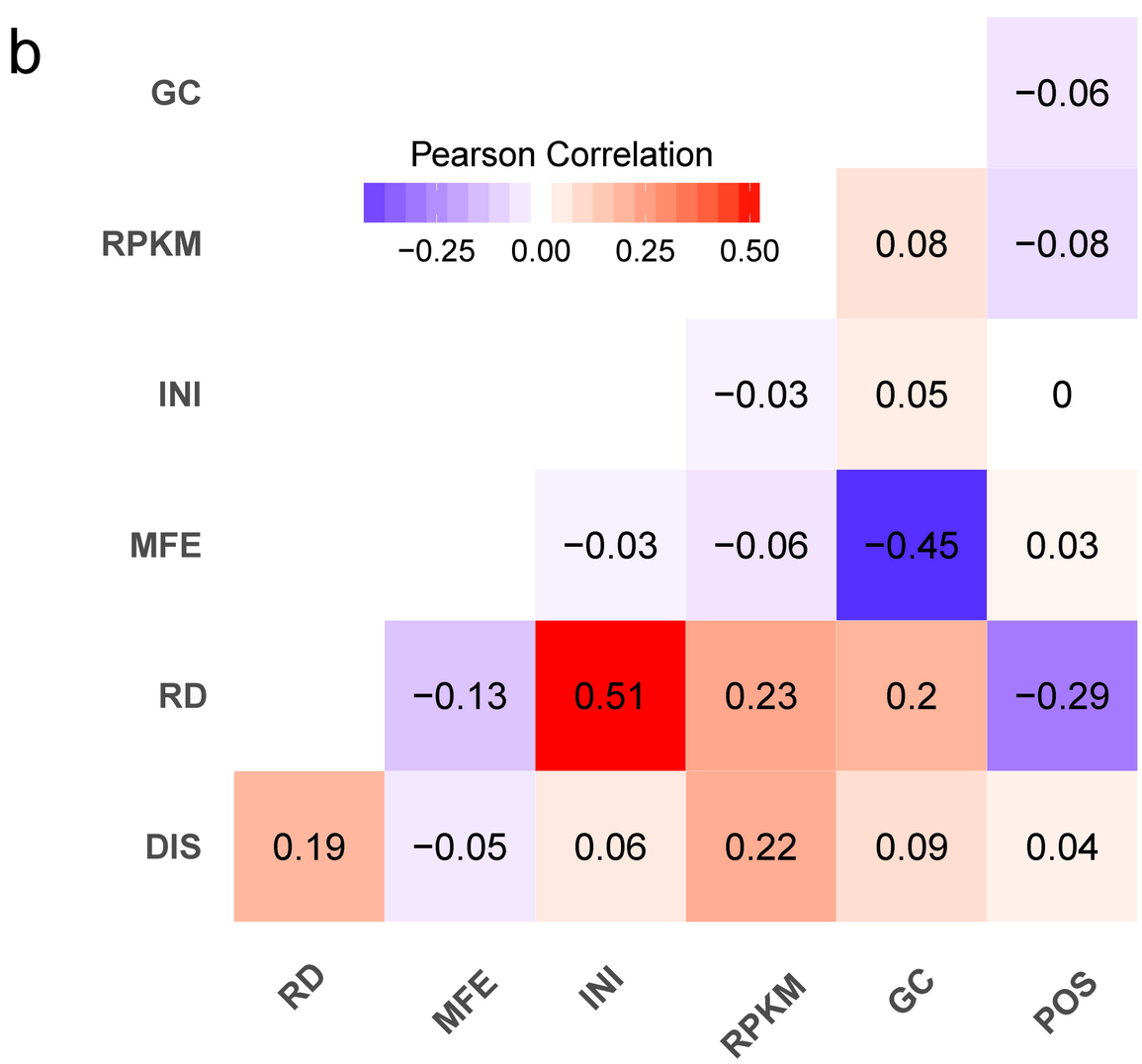
43 Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell***157**, 624-635, doi:10.1016/j.cell.2014.02.033 (2014).

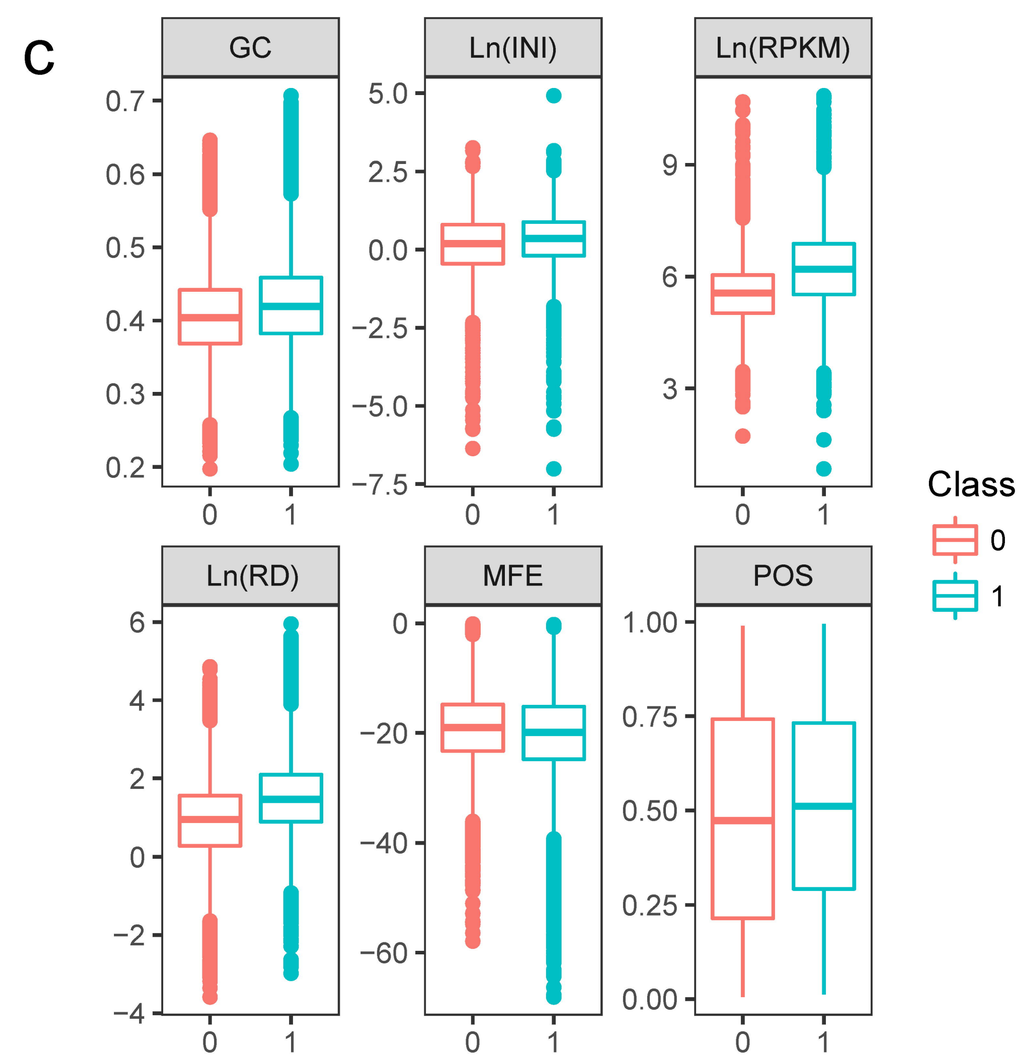
44 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol***6**, 26, doi:10.1186/1748-7188-6-26 (2011).

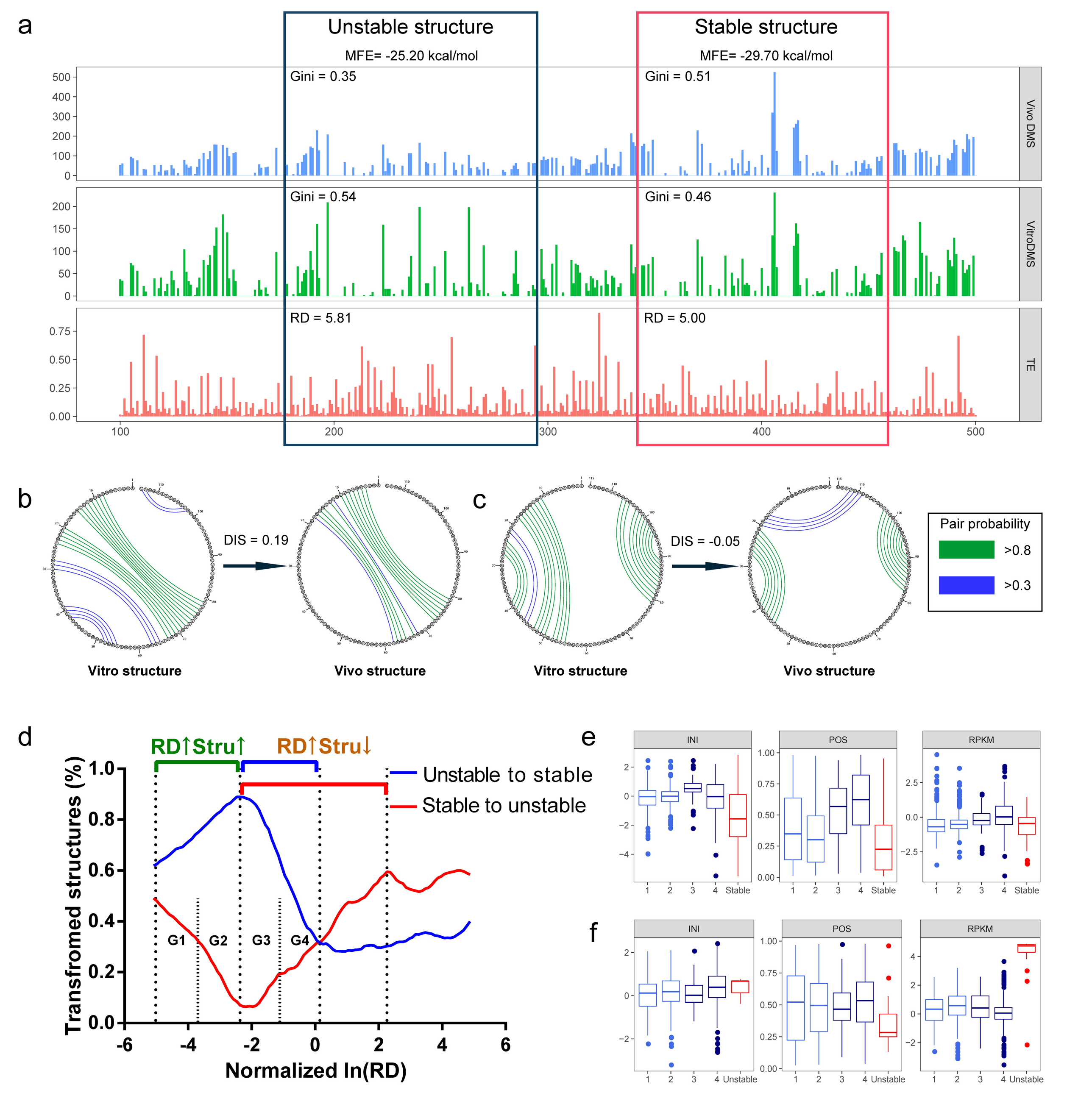
45 Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics***11**, 129, doi:10.1186/1471-2105-11-129 (2010).

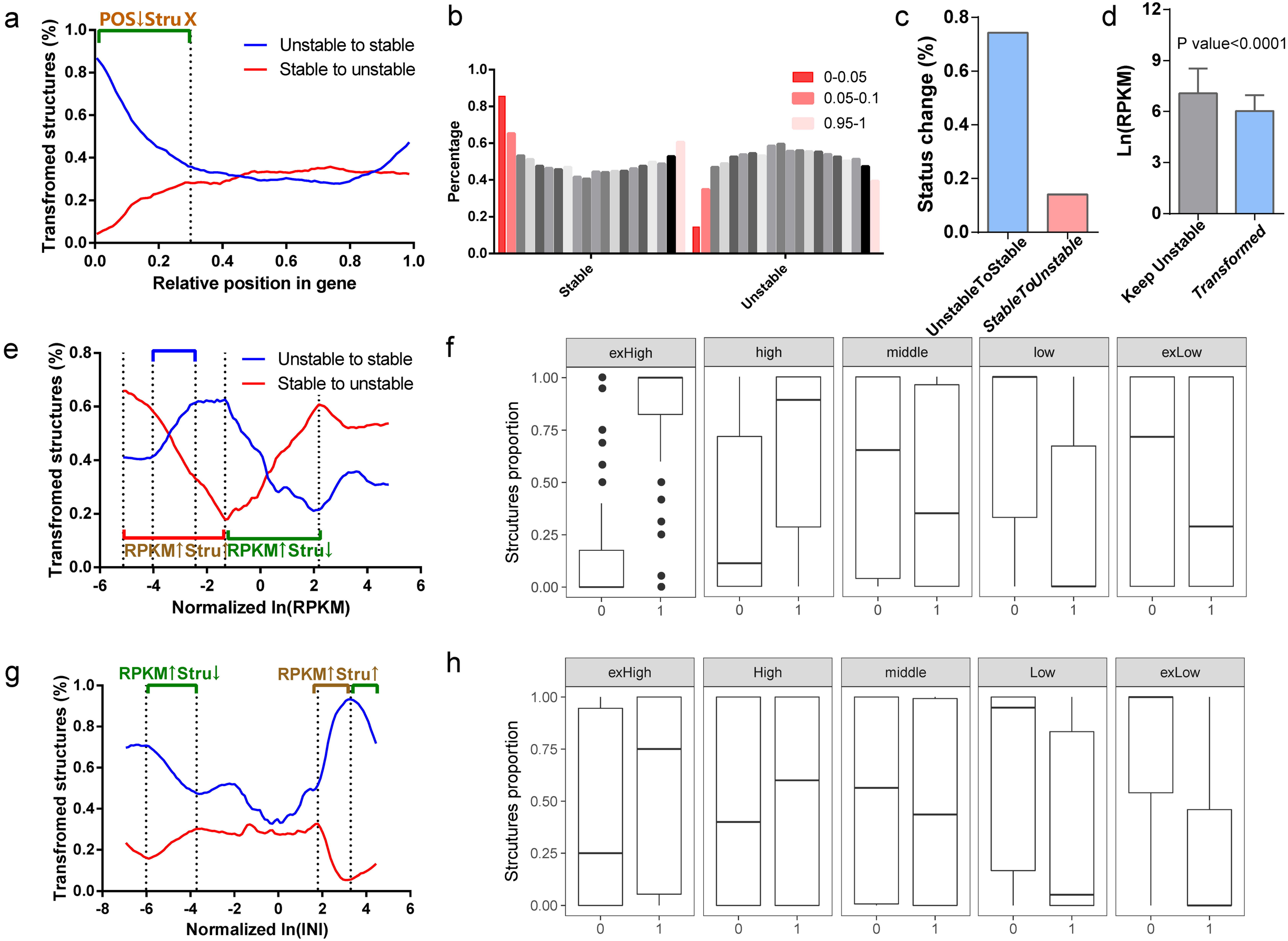
46 Chollet, F. Keras. *GitHub repository, https://github.com/fchollet/keras* (2015).

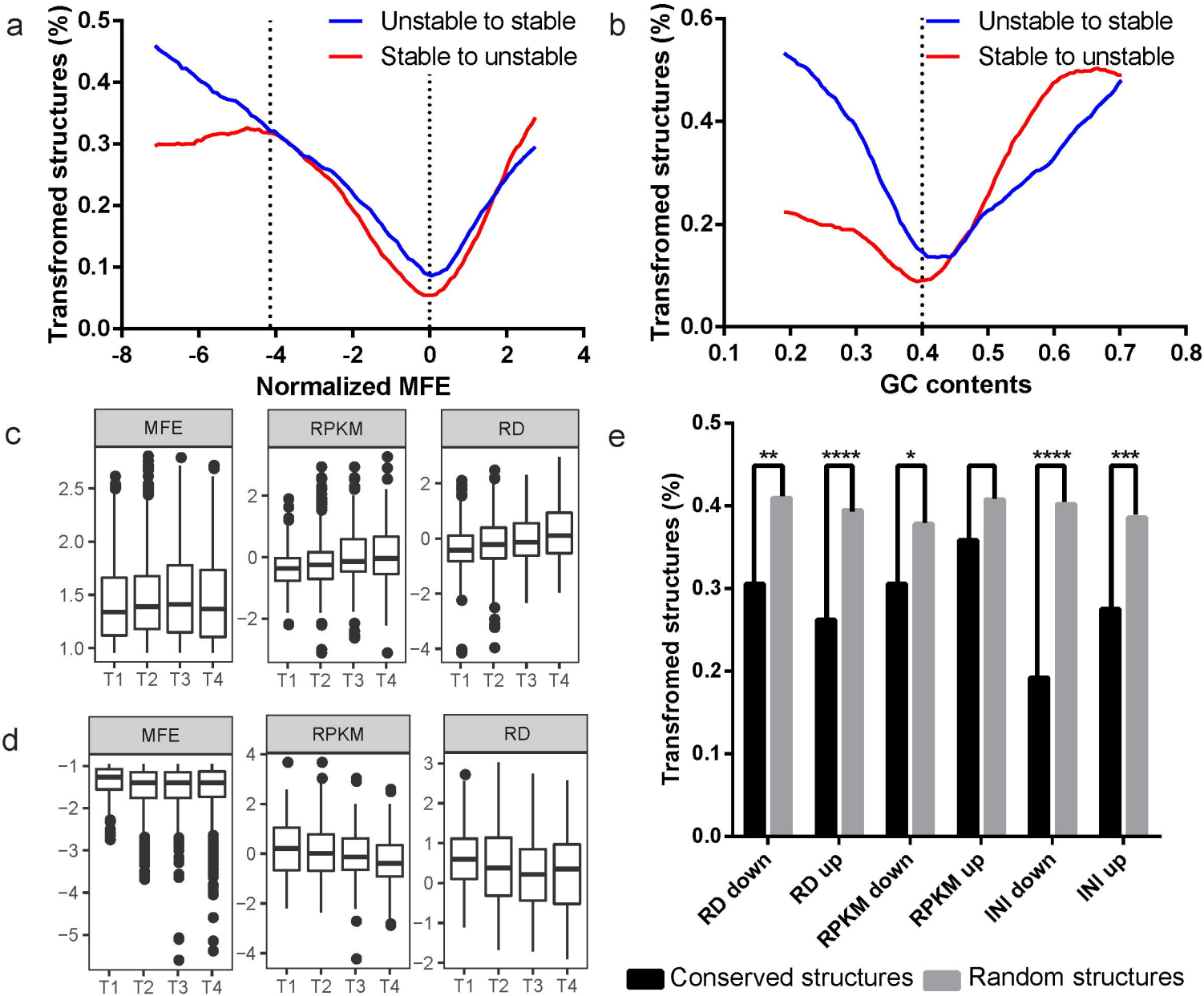

a
Low gini index
DMS signal in vivo
High gini index
DMS signal in vitro
In vivo
Weak structure
Stable structure
In vitro
DIS relative frequency
Class 0
μ-σ<
Class 1
>μ+σ
DIS
Mutation
Predict
Unstable
Stable
Deep learning modeling
GC
INI
RPKM
RD
MFE
POS
Six quantitative features of mRNA structure

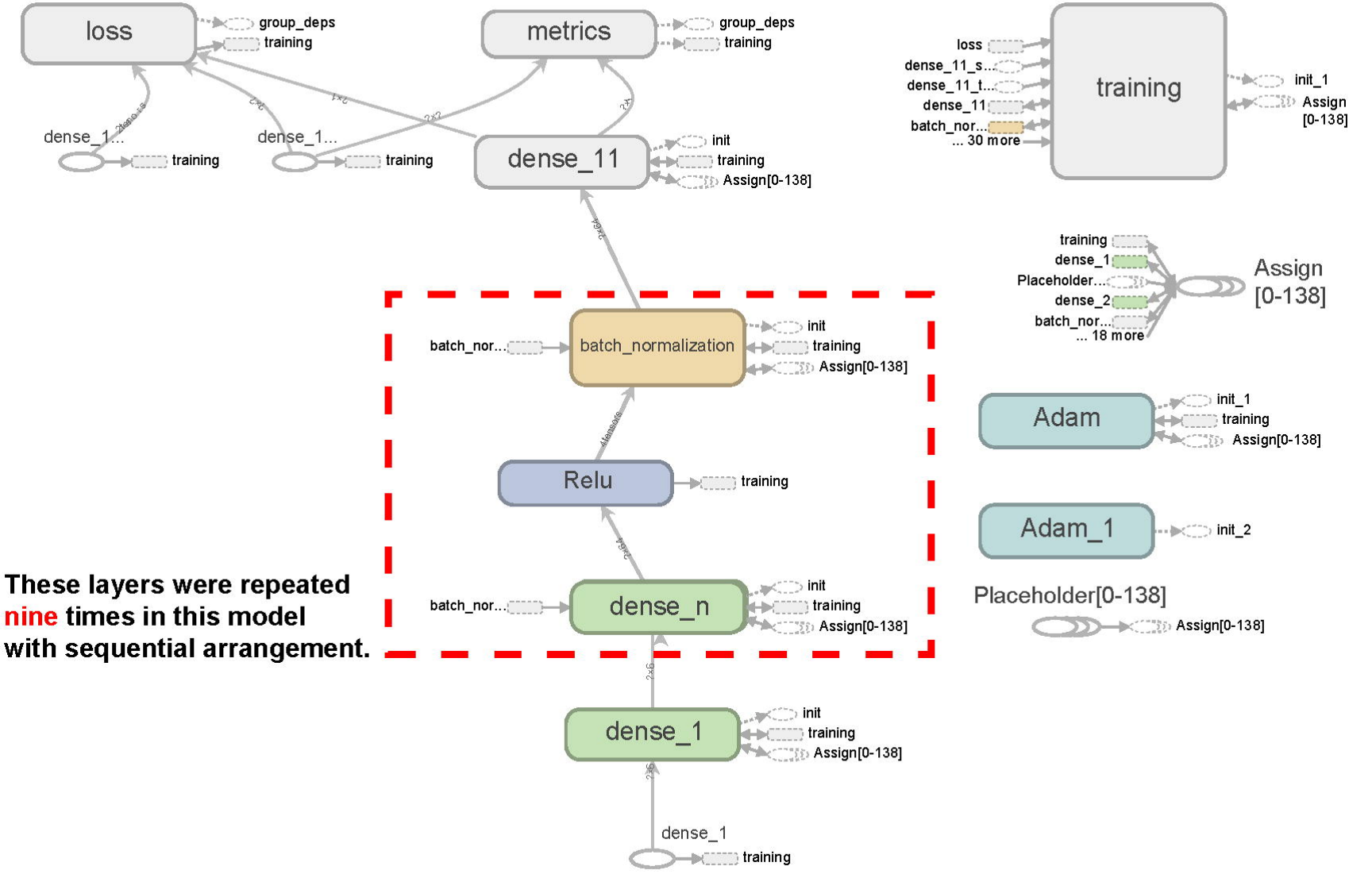

b
Pearson Correlation
−0.25 0.00 0.25 0.50
GC −0.06
RPKM 0.08 −0.08
INI −0.03 0.05 0
MFE −0.03 −0.06 −0.45 0.03
RD −0.13 0.51 0.23 0.2 −0.29
DIS 0.19 −0.05 0.06 0.22 0.09 0.04
RD MFE INI RPKM GC POS


c
GC
Ln(INI)
Ln(RPKM)
Ln(RD)
MFE
POS
Class
0
1


a
Unstable structure
MFE= -25.20 kcal/mol
Stable structure
MFE= -29.70 kcal/mol
Gini = 0.35
Gini = 0.51
Vivo DMS
Gini = 0.54
Gini = 0.46
VitroDMS
RD = 5.81
RD = 5.00
TE
b
DIS = 0.19
Vitro structure
Vivo structure
c
DIS = -0.05
Vitro structure
Vivo structure
Pair probability
>0.8
>0.3
d
RD↑Stru↑
RD↑Stru↓
Unstable to stable
Stable to unstable
Transfromed structures (%)
G1
G2
G3
G4
Normalized ln(RD)
e
INI
POS
RPKM
Stable
f
Unstable


a
POS↓Stru X
Unstable to stable
Stable to unstable
Transfromed structures (%)
Relative position in gene
b
0-0.05
0.05-0.1
0.95-1
Percentage
Stable
Unstable
c
Status change (%)
UnstableToStable
StableToUnstable
d
P value<0.0001
Ln(RPKM)
Keep Unstable
Transformed
e
Unstable to stable
Stable to unstable
RPKM↑Stru↑
RPKM↑Stru↓
Transfromed structures (%)
Normalized ln(RPKM)
f
exHigh
high
middle
low
exLow
Strcutures proportion
g
RPKM↑Stru↓
RPKM↑Stru↑
Transfromed structures (%)
Normalized ln(INI)
h
exHigh
High
middle
Low
exLow
Strcutures proportion


a
Transfromed structures (%)
Unstable to stable
Stable to unstable
Normalized MFE
b
Transfromed structures (%)
Unstable to stable
Stable to unstable
GC contents
c
MFE
RPKM
RD
T1
T2
T3
T4
d
MFE
RPKM
RD
T1
T2
T3
T4
e
Transfromed structures (%)
RD down
RD up
RPKM down
RPKM up
INI down
INI up
Conserved structures
Random structures


loss
group_deps
training
metrics
group_deps
training
dense_1...
training
dense_1...
training
dense_11
init
training
Assign[0-138]
training
loss
dense_11_s...
dense_11_t...
dense_11
batch_nor...
... 30 more
init_1
Assign
[0-138]
training
dense_1
Placeholder...
dense_2
batch_nor...
... 18 more
Assign
[0-138]
batch_normalization
batch_nor...
init
training
Assign[0-138]
Adam
init_1
training
Assign[0-138]
Relu
training
Adam_1
init_2
Placeholder[0-138]
Assign[0-138]
dense_n
batch_nor...
init
training
Assign[0-138]
These layers were repeated
nine times in this model
with sequential arrangement.
dense_1
init
training
Assign[0-138]
dense_1
training