Academic Journal of Medicine & Health Sciences

ISSN 2616-5791 Vol. 6, Issue 1: 15-22, DOI: 10.25236/AJMHS.2025.060103

Application and evaluation of deep neural network fusion architecture in predicting COVID-19 mRNA vaccine degradation

Nana Guo[1,a], Junxi Li[2,b], Xin Guo[1,c,*]

[1]Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation
[2]Herzen University, St. Petersburg, Russian Federation
[a]NnG.cn@outlook.com, [b]lijunxi111@yandex.com, [c]go7.s@outlook.com
[*]Corresponding author

Abstract: *The COVID-19 outbreak highlighted the importance of mRNA vaccines; however, the thermal instability of mRNAs poses many challenges for vaccine production, storage, and transport, and accurately predicting their degradation is critical to safeguarding vaccine quality and efficacy. Traditional prediction methods always have the disadvantages of long experimental periods, excessive errors and unstable biological environments. Although machine learning and deep learning approaches can compensate for the shortcomings of traditional methods, using only one of these models to predict COVID-19 mRNA vaccine degradation is not effective. So, we propose a fusion model GGTC of GRU, GNN, Transformer and CNN. The results show that our fusion of GRU, CNN, Transformer and GNN models not only improves the accuracy of model prediction, but also improves the generalisation ability of the model.*

Keywords: *mRNA Vaccine; Artificial Neural Networks; Attention Mechanism; Model Fusion*

1. Introduction

Messenger ribonucleic acid vaccines (mRNA vaccines) are a novel vaccine technology that play an important role in disease prevention, for example in the fight against infectious diseases [1]. The core component of an mRNA vaccine is a piece of messenger ribonucleic acid (mRNA) that encodes an antigenic protein. When the vaccine is injected into the body, the mRNA enters the cell and uses the cell's own ribosomal translation mechanism to translate the genetic information carried by the mRNA into antigenic proteins. These antigenic proteins are processed inside the cell and displayed on the cell surface, activating the body's immune system. The immune system recognizes these foreign antigens and produces a specific immune response involving the production of antibodies and the activation of immune cells such as T-cells [2]. The immune system is able to quickly recognise and eliminate a real pathogen when it invades the body, preventing disease. In the COVID - 19 mRNA vaccine, the mRNA encodes the spiny protein (S protein) of the new coronavirus [3].

The development cycle for mRNA vaccines can be significantly shortened compared to traditional vaccines [4]. This is because they do not require the cultivation of large numbers of pathogens or the complex protein purification process of conventional vaccines. mRNA itself is a non-replicating nucleic acid with no ability to infect [5]. It is also degraded in the body by normal cellular metabolic processes and does not integrate into the human genome. This is a clear safety advantage over some traditional vaccines, such as live attenuated vaccines, which may have the risk of responding to mutations that cause disease. mRNA is a relatively unstable molecule, which places high demands on the storage and transport conditions of mRNA vaccines [6]. Many mRNA vaccines need to be stored at ultra-low temperatures to ensure mRNA integrity and vaccine efficacy. It is therefore important for us to study the degradation of the COVID-19 mRNA vaccine.

There are three main methods for predicting COVID-19 mRNA vaccine degradation; methods based on physicochemical properties, methods based on nucleic acid sequence characterisation and methods based on enzymatic reactions [7-10]. According to the Arrhenius equation, the degradation of mRNA was observed by incubating the vaccine at different high temperatures and then extrapolating the degradation rate at normal storage temperature, but the experimental period is long and the extrapolation results may have errors. Bioinformatics software was used to predict the secondary structure of mRNAs and analyse their structural features such as stem-loop and hairpin. mRNAs with more double-stranded regions are relatively more stable, but the secondary structure of mRNAs changes dynamically and does not reflect their true stability in complex biological environments [11]. Of course, we can also calculate the GC content of the mRNA, and a higher GC content may make the mRNA more thermally stable, but the base composition is only one of the influencing factors, and we cannot accurately predict the degradation based on this alone [12]. Different types of nucleases were added to mRNA vaccine samples to simulate nuclease degradation of mRNA to understand the sensitivity of mRNA to different nuclease, but in vitro tests cannot fully simulate the complex enzymatic environment in vivo [13]. The above three traditional methods have problems in predicting COVID-19 mRNA vaccine degradation, leading to excessive errors between the final prediction and the real results.

In mRNA vaccine degradation prediction, the support vector machine (SVM) algorithm attempts to find an optimal hyperplane such that the sum of the distances of all sample points to that hyperplane is minimised. Features related to vaccine degradation are used as input vectors to train the SVM model to predict the extent of vaccine degradation [14]. However, SVM can suffer from high computational complexity on a large scale and has limited ability to model complex non-linear relationships. Recurrent Neural Networks (RNN) are able to take into account the sequence information of the bases in a sequence when processing mRNA sequences [15]. At each time step, the RNN receives one base information and computes the next hidden state based on the current input and hidden state. For COVID - 19 mRNA vaccine degradation prediction, the RNN can learn the pattern of the effect of bases at different positions in the sequence on degradation. However, traditional RNNs suffer from the problem of gradient vanishing or gradient explosion, and it can be difficult to effectively learn long-term dependencies when dealing with long sequences. Long Short-Term Memory Networks (LSTM) is an improvement of RNNs by introducing gating units to solve the gradient problem of traditional RNNs [16]. In predicting mRNA vaccine degradation, LSTM is better able to remember key information in long sequences. LSTM remembers important degradation-related features that appear at the beginning of mRNA sequences and consistently uses this information in subsequent calculations to predict degradation rates [17]. LSTM is able to capture the complex temporal dependencies in sequences more accurately, which is beneficial for analysing the effects of interactions between distant bases on the degradation of mRNA sequences. However, the LSTM structure is more complex and less computationally efficient, so another variant of the RNN, the Gated Control Loop Unit (GRU) neural network, emerged to solve this problem.

In the COVID-19 mRNA vaccine degradation prediction task, the GRU controls the delivery and updating of information through update and reset gates, and the GRU is able to rapidly learn the feature representations of the mRNA sequences and predict degradation based on these features [18]. To predict mRNA vaccine degradation, the mRNA molecule can be represented as a graph structure, with bases as nodes and interactions between bases as edges. By propagating information through the graph, the graph neural network (GNN) is able to learn the feature representations of the nodes and edges and predict the degradation of the whole mRNA molecule accordingly [19-20]. Transformer's self-attention mechanism is able to compute the degree of association between any two positions in a sequence directly, without having to process the sequence step by step like an RNN [21]. When predicting mRNA vaccine degradation, it can comprehensively capture the interactions between individual bases in a sequence, and both close and distant relationships can be effectively modelled. For mRNA vaccine degradation prediction, CNNs can treat mRNA sequences as one-dimensional signals and perform feature extraction by sliding the convolution kernel over the sequence. Attention networks are better at capturing global features, while traditional convolutional neural networks (CNNs) are better at capturing local features [22]. For mRNA vaccine degradation prediction, CNNs can treat mRNA sequences as one-dimensional signals and perform feature extraction by sliding the convolution kernel over the sequence.

We select four different types of models - GRU, CNN, Transformer and GNN - for fusion. The GRU model works well with time-series data such as mRNA sequences, capturing information about the order of bases in the sequence as well as patterns of base combinations. CNNs are good at extracting local features and can identify specific structures in mRNA sequences. Transformer excels at capturing global sequence information and complex relationships. GNN uses information about the graph structure of mRNA molecules to learn the association of base interactions with degradation. Model fusion can integrate the advantages of different models and compensate for the shortcomings of a single model, thus improving the accuracy and stability of the forecast.

2. Method

2.1 Dataset

The COVID-19 mRNA vaccine dataset was derived from the RMDM database [23]. There are a total of 6,034 samples in this data set. The training set has a total of 2,400 samples with a length of 68 sequences. The test set has 3,634 samples with sequences of length 91. The details of this dataset are presented in Table 1.

Table 1: COVID-19 mRNA vaccine dataset

Parameters	Explanation
id_seqpos	An arbitrary identifier for each sample.
reactivity	Reaction values for the first 68 bases in the sequence are used to determine the likely secondary structure of the RNA sample.
deg_Mg_pH10	Reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating with magnesium at high pH (pH 10).
deg_pH10	Reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating without magnesium at high pH (pH 10).
deg_Mg_50C	Reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating with magnesium at high temperature (50 degrees Celsius).
deg_50C	Reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating without magnesium at high temperature (50 degrees Celsius).

Table 1 contains the meanings indicated by the six columns in the dataset. The experimental data were all obtained for the first 68 bases under the five conditions in Table 1. Minimum value across all 5 conditions must be greater than -0.5. Mean signal/noise across all 5 conditions must be greater than 1.0. Signal/noise is defined as mean (measurement value over 68 nts) / mean (statistical error in measurement value over 68 nts). To help ensure sequence diversity, the resulting sequences were clustered into clusters with less than 50% sequence similarity, and the 629 test set sequences were chosen from clusters with 3 or fewer members. That is, any sequence in the test set should be sequence similar to at most 2 other sequences.

In mRNA vaccine degradation studies, a matrix of class correlation coefficients for mRNA genes can help identify which genes are associated with mRNA vaccine stability [24-25]. The matrix of class correlation coefficients of mRNA genes can be used to extract very useful features. Highly correlated genomes may have redundant information, and representative genes from them can be selected as features to reduce the dimensionality of the data while preserving important information [26]. We selected the class correlation coefficient matrix of the mRNA gene with the serial number id_0a2bbe37e and visualised the results as shown in 1.

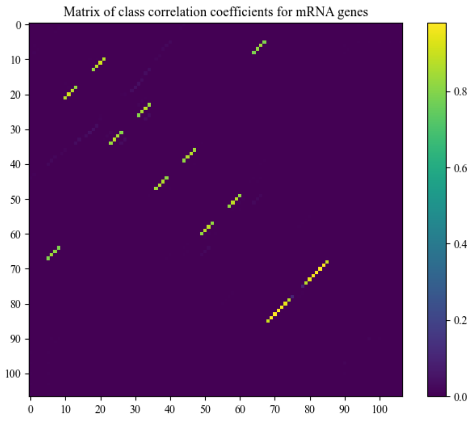

Figure 1: Matrix of class correlation coefficients for mRNA genes

We can see that the colours in Figure 1 gradually change from dark blue in the lower left corner to yellow in the upper right corner. Dark blue indicates lower correlation coefficients, close to 0 or negative, and yellow indicates higher correlation coefficients, close to 1. Dark blue areas indicate low or negative correlation coefficients between genes. Yellow areas indicate a high positive correlation between genes.

We selected the mRNA sequence with the sequence number id_0a2bbe37e and estimated the structural sequence of the mRNA using the gamma parameter, as shown in 2. The secondary structure of the mRNA is visualised as shown in Figure 3.

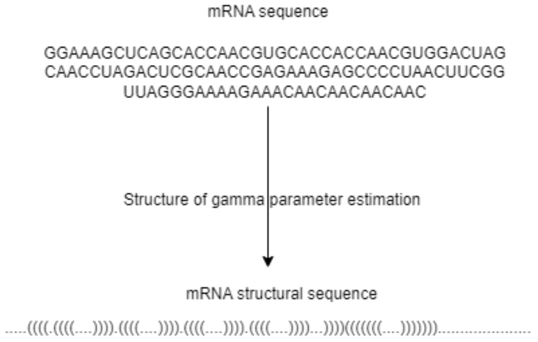
mRNA sequence
GGAAAGCUCAGCACCAACGUGCACCACCAACGUGGACUAG
CAACCUAGACUCGCAACCGAGAAAGAGCCCCUAACUUCGG
UUAGGGAAAAGAAACAACAACAACAAC
Structure of gamma parameter estimation
mRNA structural sequence
.....((((.((((....)))).((((....)))).((((....)))).((((....))))...))))(((((((....))))))).....................

Figure 2: Estimation of structural sequence of mRNA by gamma parameter

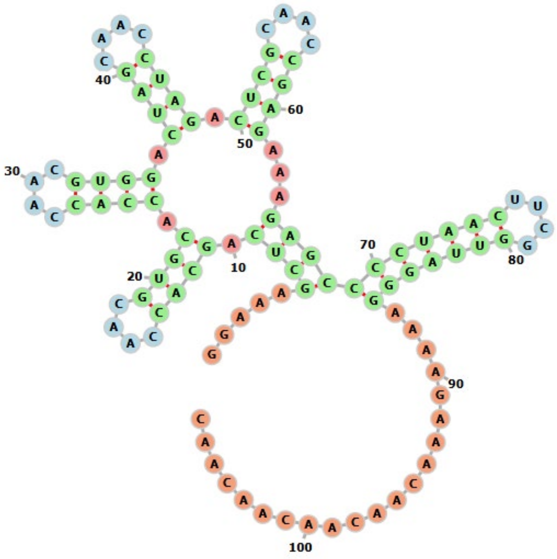

Figure 3: mRNA secondary structure visualisation

Figures 2 and 3 illustrate the mRNA secondary structure. The secondary structure of mRNA plays an important role in predicting mRNA degradation in neural networks such as graph neural networks and CNNs, which can provide a structural building block for graph neural networks and help capture long-range interactions. Provides CNNs and transformers with local feature patterns and increased feature dimensions. It also helps GRUs capture information about sequence and structure dynamics and account for long-term dependencies, helping each neural network to better mine and degrade relevant features to improve prediction accuracy.

2.2 Feature Engineering

Feature engineering plays a multifaceted role in predicting mRNA degradation using neural networks. It is capable of representing and transforming data, encoding biological information such as the sequence and structure of mRNAs into numerical forms that can be processed by neural networks, and integrating features from different sources and dimensions to construct comprehensive feature matrices. By using clustering analysis to extract complex features based on graph structure and mining potential feature relationships, neural networks can capture long-range interactions, spatial relationships and common differences within mRNA molecules [27]. The rich features provide diverse inputs to the model, increasing the model's generalisation ability and training effect, and improving prediction accuracy on new data.

Feature engineering in the prediction of mRNA vaccine degradation mainly consists of the following steps:

1) Feature extraction from .npy files.

2) Addition of features to the dataset.

3) Data pre-processing and feature integration.

4) Clustering and data segmentation.

5) Data storage.

Specific secondary structure patterns, some combinations of stem-loop structures, can affect the binding site and affinity of the degrading enzyme to the mRNA and thus the rate of degradation, and extraction of these structural features can provide neural networks with critical information to predict degradation.

We chose to use the KMeans clustering algorithm for the clustering operation, which groups the mRNA molecules based on the similarity of the data, adding the cluster_id column to the data. This helps to reveal natural grouping structures in the data, and the neural network can use this clustering information to learn common features of molecules within different groups, and different features between groups, to better capture potential patterns associated with degradation.

Different types of features can reflect different aspects of mRNA molecules, which helps to improve the generalisation ability of the model to adapt to the task of predicting degradation of different types of mRNA molecules and reduce the risk of overfitting. In addition to sequence features, structure-related features can help the neural network to distinguish the effect of changes in the secondary structure of mRNA on degradation under different environmental conditions, such as different pH and temperature, and thus predict degradation more accurately.

The data segmentation combines a variety of factors, such as the reactivity column and the cluster_id column, generating fold5 and fold10 columns for labelling data in different folds. This allows a cross-validation strategy to be used during model training, making rational use of the data for multiple training and evaluation. Data partitioning in the feature engineering phase ensures that the training and validation sets are representative in terms of data distribution and feature combinations, providing a reliable data base for neural network training.

3. Experiments

We first train the model using Gated Recurrent Unit (GRU). The GRU gated recurrent unit is a recurrent neural network that slightly improves on the LSTM by combining the forget gate and the input gate into a single “update gate”, as well as combining the cell state and the hidden state [28].

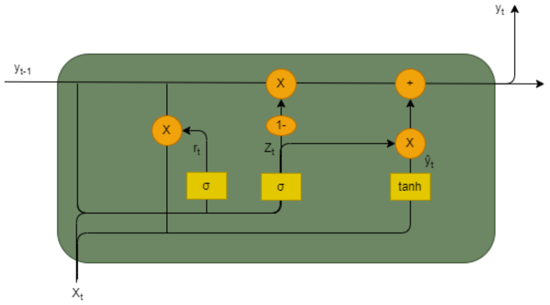

Figure 4: GRU Framework

As shown in Figure 4, we design a GRU network with multiple bi-directional GRU layers stacked together, and each bi-directional GRU layer consists of bi-directional GRU units inside, which can capture both forward and reverse sequence information. Each bi-directional GRU layer receives the data output from the previous layer, and the final output is information about the entire sequence, with the shape keeping the sequence length dimension constant.

We chose to use four neural network structures, GRU, GNN, Transformer and CNN, for fusion to obtain the GGTC model.

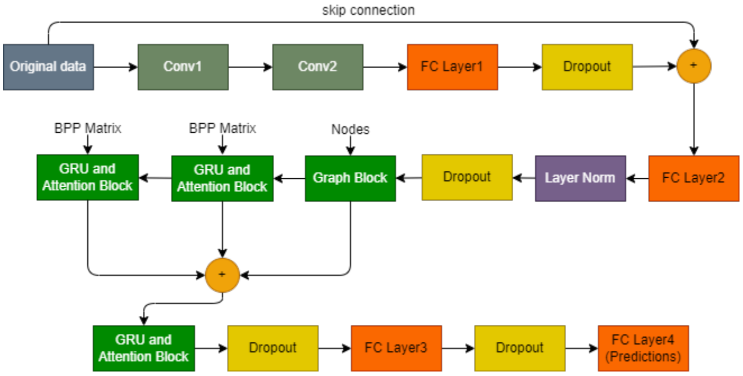

Figure 5: GGTC Architecture

As shown in Figure 5, in Conv1, the convolution kernel slides over the raw data in certain steps to extract preliminary local features. The data passed from Conv1 to Conv2 is the feature map after the first layer of convolutional processing. Conv2 builds on this by using different convolutional kernels to further extract more complex and abstract features.

The feature map output from Conv2 is a multi-dimensional array representing features extracted at different locations and on different channels. Before passing to FC Layer1, these feature maps need to be spread into one-dimensional vectors. This process actually converts the spatially distributed feature information into a long vector so that the fully connected layer can process it. Each neuron in FC Layer1 is connected to all the elements in this one-dimensional vector, and the local features extracted by the convolutional layer are combined into a higher-level feature representation through the operation of weighted summation and activation function. This step is an important conversion from local to global features.

There are paths from Conv1 and Conv2 to the BPP Matrix and Nodes, respectively. The BPP Matrix represents the probability of base pair formation in an mRNA sequence, and is an important tool for describing the secondary structure of mRNAs. The Nodes represent the key feature points in the structure of mRNAs. GRU and Attention Blocks are used to address long-range dependencies in mRNA sequences and focus on sequence regions that are important for degradation rates. Graph Block is used to process the secondary structure of mRNA. It converts the structural information of mRNA into a graph representation and processes it through a graph neural network (GNN) to extract structure-related features.

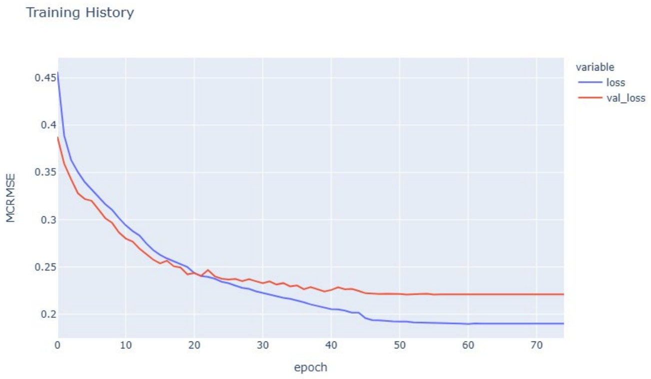

Figure 6: GGTC Training History

As shown in Figure 6, the degradation process of mRNA vaccines involves many factors, including sequence information, secondary structure, etc. GRU processes sequence order information, GNN processes structural information, Transformer captures global associations, and CNN extracts local patterns. This fusion can make full use of the multi-modal information of mRNA data, more comprehensively understand the characteristics of mRNA vaccines, and thus more accurately predict the degradation rate.

4. Discussion and Conclusion

The GGTC model proposed in this study combines four different types of model architectures: GRU, GNN, Transformer and CNN. GRU is able to effectively deal with long-term dependencies in sequence data and capture temporal features of mRNA sequences, which is a powerful support for understanding the dynamics of the degradation process, while GNN focuses on modelling the complex network of relationships, such as the molecular structure, by analysing the interactions within and between mRNA molecules. GNN focuses on the modelling of complex networks such as molecular structures, and explores the potential structural factors affecting degradation through the analysis of intra- and inter-molecular interactions of mRNAs. Transformer, with its powerful parallel computing capability and ability to capture global information, is able to comprehensively integrate multi-dimensional information related to mRNAs and enhance the model's capability of recognizing complex patterns. CNN, with its convolutional operation, is good at extracting local features, which is very helpful in analysing the local structures and the global information of mRNA sequences. The convolution operation of CNN is good at extracting local features and has unique advantages in analysing the local structure and patterns of mRNA sequences. By fusing these four models, the GGTC model can fully utilize the strengths of each model and make up for the deficiencies of a single model.

However, there is still room for further improvement of the GGTC model. In terms of data acquisition, we used a relatively comprehensive public dataset of mRNA vaccines, but due to the complexity and limitations of biological experiments, the scale and diversity of the data still need to be improved. There is also model interpretability, which requires more research to develop effective interpretation methods to improve the understanding of the decision-making process of GGTC models.

References

[1] Kramps T, Elbers K. Introduction to RNA vaccines. RNA Vaccines 2016; 1499:1–11.

[2] Barbier A.J., Jiang A.Y., Zhang P., Wooster R., Anderson D.G. The clinical progress of mRNA vaccines and immunotherapies. Nat. Biotechnol. 2022; 40:840–854. doi: 10.1038/s41587-022-01294-2.

[3] Zhang, N. N. et al. A thermostable mRNA vaccine against COVID-19. Cell 182, 1271–1283.e1216 (2020).

[4] Waickman AT, Victor K, Newell K, et al. MRNA-1273 vaccination protects against SARS-COV-elicited lung inflammation in nonhuman primates. JCI Insight 2022;7(13): e160039.

[5] Baden LR, El Sahly HM, Essink B, et al. Efficacy and safety of the mRNA-1273 sars-cov-2 vaccine. N Engl J Med 2021;384(5):403–16.

[6] Leppek K, Byeon GW, Kladwang W, et al. Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. Nat Commun 2022;13(1):1536.

[7] Pardi, N., Hogan, M., Porter, F. et al. mRNA vaccines — a new era in vaccinology. Nat Rev Drug Discov 17, 261–279 (2018). https://doi.org/10.1038/nrd.2017.243

[8] Crommelin, D. J., Sindelar, R. D., & Meibohm, B. (2021). From bench to bedside: mRNA as a drug. Journal of Pharmaceutical Sciences, 110 (4), 1075-1091.

[9] Jin, YH., Cai, L., Cheng, ZS. et al. A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). Military Med Res 7, 4 (2020). https://doi.org/10.1186/s40779-020-0233-6

[10] Lazarus, J.V., Ratzan, S.C., Palayew, A. et al. A global survey of potential acceptance of a COVID-19 vaccine. Nat Med 27, 225–228 (2021). https://doi.org/10.1038/s41591-020-1124-9

[11] Do CB, Woods DA, Batzoglou S. Contrafold: RNA secondary structure prediction without physics-based models. Bioinformatics 2006;22(14):90–98.

[12] Andronescu M. Rnasoft: a suite of RNA secondary structure prediction and design software tools. Nucleic Acids Res 2003;31(13):3416–22.

[13] Zhang NN, Li XF, Deng YQ, et al. A thermostable mRNA vaccine against covid-19. Cell 2020; 182:1271–83.

[14] Ahuja A.S., Reddy V.P., Marques O. Artificial intelligence and COVID-19: A multidisciplinary approach. Integr. Med. Res. 2020;9(3) - PMC - PubMed

[15] Muneer, Amgad et al. "iVaccine-Deep: Prediction of COVID-19 mRNA vaccine degradation using deep learning." Journal of King Saud University. Computer and information sciences vol. 34,9 (2022): 7419-7432. doi: 10.1016/j.jksuci.2021.10.001

[16] S. Asif Imran, M. Tariqul Islam, C. Shahnaz, M. Tafhimul Islam, O. Tawhid Imam and M. Haque, "COVID-19 mRNA Vaccine Degradation Prediction using Regularized LSTM Model," 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering

(WIECON-ECE), Bhubaneswar, India, 2020, pp. 328-331, doi: 10.1109/WIECON-ECE52138.2020.9398044.

[17] Qaid, Talal S et al. "Deep sequence modelling for predicting COVID-19 mRNA vaccine degradation." PeerJ. Computer science vol. 7 e597. 22 Jun. 2021, doi:10.7717/peerj-cs.597

[18] Wayment-Steele HK, Kladwang W, Watkins AM, et al. Deep learning models for predicting RNA degradation via dual crowdsourcing. Nat Mach Intell. 2022;4(12):1174-1184. doi:10.1038/s42256-022-00571-8

[19] Zichen Wang, Vassilis N. Ioannidis, Huzefa Rangwala, Tatsuya Arai, Ryan Brand, Mufei Li, and Yohei Nakayama. 2022. Graph Neural Networks in Life Sciences: Opportunities and Solutions. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 4834–4835. https://doi.org/10. 1145/ 3534678. 3542628

[20] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graphconvolutional neural network model for the prediction of chemical reactivity. Chemical science 10, 2 (2019), 370--377.

[21] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos- Ruiz, NinaMDonghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. 2020. A deep learning approach to antibiotic discovery. Cell 180, 4 (2020), 688--702.

[22] Ing, S.H., Abdullah, A.A., Mashor, M.Y., Mohamed-Hussein, Z., Mohamed, Z., & Ang, W.C. (2022). Exploration of hybrid deep learning algorithms for covid-19 mrna vaccine degradation prediction system. International Journal of Advances in Intelligent Informatics.

[23] Cordero P., et al. (2012). An RNA Mapping DataBase for Curating RNA Structure Mapping Experiments. Bioinformatics 28(22): 3006-3008.

[24] van Dam, Sipko et al. "Gene co-expression analysis for functional classification and gene-disease predictions." Briefings in bioinformatics vol. 19,4 (2018): 575-592. doi:10.1093/bib/bbw139

[25] Zhao Y, Li H, Fang S, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res 2016;44: D203–8.

[26] Zeisel A, Munoz-Manchado AB, Codeluppi S, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015; 347:1138–42.

[27] Alex Reneau, Jerry Yao-Chieh Hu, Ammar Gilani, and Han Liu. 2023. Feature programming for multivariate time series prediction. In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 1204, 29009–29029.

[28] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)