Machine Learning Approaches for RNA Editing Prediction

Andrew J. Jung[1,2], Leo J. Lee[1,2], Alice J. Gao[1,2], Brendan J. Frey[1,2]
[1]University of Toronto, [2]Vector Institute
{andrewjung,ljlee}@psi.toronto.edu

Abstract

RNA editing is a major post-transcriptional modification that contributes significantly to transcriptomic diversity and regulation of cellular processes. Exactly how cis-regulatory elements control RNA editing appears to be highly complex and remains largely unknown. However, with the improvement of computational methods for detecting and quantifying RNA editing from large-scale RNA-seq data, it becomes possible to build computational models of RNA editing. Here we report our attempt to develop machine learning models for predicting human A-to-I editing by training on large number of highly confident RNA editing sites supported by observational RNA-seq data. Our models achieve good performance on held-out test evaluations. Furthermore, our deep convolutional model also generalizes well to dataset from a controlled study.

1 Introduction

RNA editing is a post-transcriptional process where the nucleotide sequence of RNA transcripts is modified from its template DNA sequence by substitution, insertion, or deletion[1]. In animals, the most prevalent type of RNA editing is the conversion of adenosine (A) into inosine (I), also known as A-to-I editing[2]. This is catalyzed by adenosine deaminase acting on RNA (ADAR) family of proteins, which binds co-transcriptionally to the double-stranded regions of RNA (dsRNA)[2]. ADAR proteins stochastically edit adenosines and the editing frequency is determined by cis-acting elements (RNA sequence and structures) and trans-acting cellular states[3]. A-to-I editing is involved in the regulation of many other cellular processes, such as RNA splicing[4], mRNA stability[5], and innate immune response against dsRNAs[6]. Additionally, A-to-I editing increases transcriptomic diversity, through recoding of the resulting amino acids[2] and altering microRNA processing[7]. Aberrant editing can result in cancer[8] and diseases in neurological[9], cardiovascular[5], and autoimmune systems[6].

In spite of its prevalence in humans, our understanding of the exact mechanism and regulation of A-to-I editing is still very limited[10]. The editing occurs in dsRNA region and the two bases immediately next to the 'A' influence the activity of ADAR protiens[11]. Although it is proposed that A-to-I editing is regulated by both the sequence and structural context, there is no simple 'motif' that can explain A-to-I editing[3, 10]. Therefore, there has yet been any 'editing code' that can predict A-to-I editing sites given an RNA sequence[10], which is in contrast with other RNA processes where computational codes have been in active development, such as splicing[12–15], RNA protein binding[16–19], and alternative polyadenylation[20, 21]. Recently, there has been an increased interest in A-to-I 'editing code' because of its potential to help with our understanding of RNA editing and applications in diagnostics[10] and RNA therapeutics[22, 23].

Most of the previous computational methods to study RNA editing focused on the accurate detection of editing events from experimental data. This problem is challenging due to technical artifacts and false positives from single-nucleotide polymorphisms (SNP)[24]. Conventional detection methods use series of filters to identify RNA editing sites from RNA-seq[25–31]. Recently, there have been several attempts to use machine learning approaches for more accurate detection of RNA editing[32–34].

Meanwhile, few groups have tackled the prediction problem, using carefully designed experiments to study the editing profiles under different sequence contexts[11, 35]. Most recently, Liu et al. [35] designed experiments to study A-to-I editing by introducing point mutations in the vicinity of three known A-to-I editing sites via CRISPR/Cas-9. The resulting mutagenesis data was used to train editing site-specific gradient boosted trees to predict RNA editing frequency from sequence and secondary structure computed by RNAfold[36]. The site-specific models showed good performance

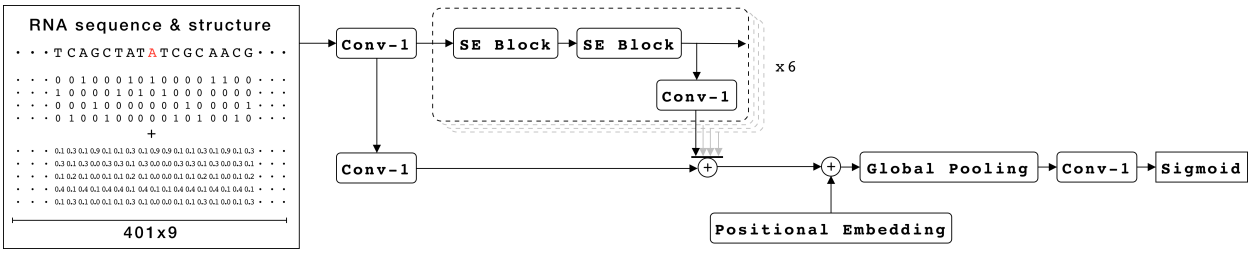

Figure 1: An overview of the deep convolutional model.

over the variants around the same editing site, but did not generalize across different editing sites, potentially due to the small dataset size with only three sites. Nevertheless, this study demonstrated the potential of machine learning approach to learn regulatory features of A-to-I editing.

Here, we take a different approach of modelling A-to-I editing by training on a larger dataset consisting of highly confident editing sites that are observed across many RNA-seq data. Due to the challenges in accurate quantification of RNA editing frequencies, we focus on binary classification models for predicting whether a given 'A' is likely edited or not based on its surrounding RNA sequence and structural context. To test if meaningful features can be learned from the observational dataset without controlled experiments, we evaluate whether our models generalize to the mutagenesis dataset from Liu et al. [35].

2 Methods

2.1 Dataset Construction

In order to train predictive models for classifying whether a given 'A' is edited or not, a balanced dataset consisting of edited and non-edited 'A's is created. High quality editing sites are curated from REDIPortal[37], which contains 4,668,125 A-to-I editing sites previously reported in the literature. Due to the challenges associated with RNA editing detection[24], REDIPortal contains many false positives. Therefore, Genotype-Tissue Expression (GTEx)[38] is used to select highly confident editing sites that are supported by: 1) at least 1,000 RNA-seq experiments (out of 2,642 with the corresponding whole genome sequencing), and 2) at least 50 body sites (out of 55 in GTEx). With these two filtering criteria, there are 25,212 edited 'A's that are well supported across many tissues and experiments.

For each of the curated editing sites, a negative example is sampled uniformly from the non-edited 'A's in the vicinity of 400bp. The size of the sampling window is chosen to ensure that the negative set includes a good mixture of: 1) examples close to the editing sites, which could be challenging for models to distinguish as they share most of genomic context, and 2) examples at a distance to help the models generalize to regions outside of the editing sites. Each of the sampled 'A's are checked to make sure they have not been reported to be editing sites in the literature. To mitigate the risk of learning spurious features, additional analysis is performed to check if the distribution of non-editing sites are similar to the editing sites across different 1) genomic regions (e.g. introns, UTRs, and coding regions), and 2) repeated elements (e.g. Alu, non-Alu, and non-repeat). The final dataset consists of 50,424 examples.

2.2 Computing RNA secondary structure

Although the features controlling A-to-I editing is mostly unknown, it is proposed that dsRNA regions and structural features, such as bulges or loops, influence the ADAR activity[10]. To compute structural features, we use RNAplfold[36] with a wrapper, created by the authors of RNAcontext[39]. The wrapper produces per-base probabilities of it being in 5 different types of structures: paired, hairpin loop, inner loop, multi-loop, or external region. '-W 240 -L 160 -u 1' is used for running RNAplfold.

2.3 Models

We consider four different types of machine learning models to classify whether a given 'A' is edited or not: logistic regression (LR), support vector machines (SVM), gradient boosting (GB), and deep convolutional neural network (CNN).

For LR, SVM, and GB models, input features are derived from 42bp region around a given 'A'. We empirically observe that context larger than 42bp does not improve the performance of the models. The inputs consist of three sets of

Table 1: Performance of the models on the held-out chromosomes

Models	Accuracy	auROC	F1-score[2]
LR	0.853	0.924	0.852
SVM	0.880	0.941	0.878
GB	0.885	0.946	0.882
CNN	0.914	0.965	0.913

features: 1) the two bases neighbouring 'A', 2) identities of 14 non-overlapping 3-mers within the 42bp region, and 3) computed structure represented by per-base probabilities. Features 1) and 2) are one-hot encoded.

For the deep convolutional network, an overview of the model architecture is shown in Figure 1. The model employs a stack of 1D convolutional blocks with residual and skip connections[40, 41], which have shown success on tasks in language modelling[42], speech recognition[43], and computational biology[15]. The model uses sequence context and the computed structures to predict how likely a given 'A' will be edited. The input to the model is a concatenation of one-hot encoded RNA sequence and per-base probabilities, computed as described in section 2.2. The input is convolved by 12 residual blocks with skip connections every 2 blocks. Each residual block includes a Squeeze-Excitation (SE) operation[44] which explicitly models the interdependence between the channels to help the CNN model focus on the most relevant cis-regulatory features. After the last residual block, a sinusoial positional embedding[45] is added to the feature map which is then summarized via global pooling over sequence positions. The combination of the positional embedding with global pooling allows the model to be used with variable input size. We empirically find that, unlike the other three models, larger context up to 400bp improves the performance of the CNN model. All inputs during training are 400bp in length.

One major advantage of the CNN model is data driven feature discovery, removing the need for manual feature engineering from our limited knowledge of RNA editing. Furthermore, the CNN architectures have shown to be more efficient in modelling long range interactions compared to the competing architectures. Lastly, the CNN model with global pooling and the positional embedding can handle flexible input size without much change in the number of parameters, while the number of parameters for the other three models scale linearly with the input size. This could be important for the future development of the model, as A-to-I editing can be influenced by long range interactions up to 10,000bp in length[46].

3 Results

3.1 Evaluation on held-out dataset

The models are evaluated on 20 percent held-out data from chromosomes 2, 14, 15, and 17, for reasons explained in section 3.3. The results are summarized in Table 1[1]. All of the models achieve good performance with the CNN model outperforming the other three.

3.2 Evaluation on dataset without the weak motif

Although there is no well established sequence motif for ADAR proteins, it is hypothesized that they prefer editing sites without 5'G and with 3'G. Even though this is a weak motif, as 3 out of every 16 non-editing sites can have this features by chance, this can be important information the models rely on, which is evidenced by the large weights on these features in LR model. Therefore, the examples without these features could serve as more challenging test subsets.

The held-out testset is split into three groups: 1) sites without 5' G, 2) sites with 3' G, and 3) sites without 5' G and with 3' G. The results are summarized in Table 2. All the models achieve comparable performance as when they are evaluated on the full held-out set from section 3.1.

3.3 Evaluation on the mutagenesis dataset

To evaluate if the models generalize beyond the observational dataset used in training, the mutagenesis dataset created by Liu et al. [35] is used. The study generated 200-300 single and double point mutations for each of the three ADAR editing sites (NEIL1, TTYH2, and AJUBA) from chromsomes 14, 15, and 17, which the models are not trained on. For each mutated sequence, the dataset reports the measured RNA-editing frequency. We further process this dataset

[1] We calculated accuracy and f1-score using the optimal threshold which maximizes accuracy.

Table 2: Performance of the models on the three subsets of the held-out chromosomes

	Without 5'G			With 3'G			Without 5'G and With 3'G		
Models	Accuracy	auROC	F1-score	Accuracy	auROC	F1-score	Accuracy	auROC	F1-score
LR	0.851	0.923	0.850	0.852	0.924	0.852	0.851	0.926	0.851
SVM	0.877	0.941	0.875	0.883	0.943	0.881	0.879	0.944	0.878
GB	0.882	0.944	0.878	0.891	0.946	0.888	0.891	0.947	0.889
CNN	0.912	0.964	0.911	0.915	0.967	0.914	0.916	0.967	0.915

Table 3: Performance of the CNN model on the mutagenesis dataset

Editing substrate	Accuracy	auROC	F1-score
NEIL1	0.841 ($\pm$0.028)	0.881 ($\pm$0.030)	0.844 ($\pm$0.030)
TTYH2	0.876 ($\pm$0.029)	0.917 ($\pm$0.028)	0.875 ($\pm$0.031)
AJUBA	0.752 ($\pm$0.047)	0.742 ($\pm$0.060)	0.788 ($\pm$0.048)

by only selecting samples with high or low editing frequencies and assigning positive or negative labels, respectively. The frequency cutoffs are chosen such that the resulting test set is roughly balanced and contains sufficient number of samples for meaningful evaluations. The processed set contains 161, 122, and 83 samples from NEIL1, TTYH2, and AJUBA, respectively. Even though our models are not trained on classifying editing sites by frequencies, we hypothesize that because editing sites with higher editing frequencies would have 'stronger' editing features, our models could generalize to this new task of distinguishing between the two extreme editing frequencies. Table 3 summarizes the results[3] from only the CNN model, as the other three models performed only slightly better than random. This demonstrates that while the other three models fail to generalize, the CNN model learns features that can be used across different datasets and tasks. This could be attributed to using large dataset to train an expressive CNN model and discover features in a data-driven way.

Performance of the CNN model is slightly worse than on the held-out dataset. This is expected because the model is trained on tissue samples from GTEx whereas HEK293T cell line was used to generate the mutagenesis dataset. Additionally, classifying sites by editing frequencies is different from the task the model is trained on. For AJUBA, the editing frequency gap between the cutoffs for the positives and negatives is small (0.003, in contrast to 0.45 and 0.302 for NEIL1 and TTYH2), so it is more challenging for the CNN model to distinguish samples with subtle editing frequency differences.

4 Discussion

We presented a new approach for predicting human A-to-I editing from RNA sequence and structure. We trained our models on a much larger dataset comprised of editing sites highly supported by GTEx RNA-seq, augmented by sampled non-editing sites. Our CNN model not only achieved good performance on the held-out evaluation, but also generalized to a more challenging task on the mutagenesis dataset. This shows that our model can learn general cis-regulatory features of A-to-I editing, just from the observational dataset.

One major limitation of our models is that we focused only on binary 'editing code' with some ability to distinguish the highly edited sites from the other end of the spectrum. A more complete 'editing code' in the future should be able to predict the editing frequency, and constructing high quality training dataset for this task will require accurate quantification of editing frequency from noisy RNA-seq. Additionally, our models do not take into account the differences in tissue types, so modelling both cis- and trans-regulatory features could be an interesting future work. Another direction is to focus on understanding the features the models learned which can help with our understanding of the editing mechanism and regulation. With a more complete 'editing code', there are many interesting applications that could be explored, such as identifying disease causing variants and RNA therapy using site-directed RNA editing.

[3]Due to the small number of samples, we use boostrapping with 1000 samples to estimate means and uncertainties of the reported metrics.

References

[1] Kazuko Nishikura. A-to-i editing of coding and non-coding rnas by adars. *Nature reviews Molecular cell biology*, 17(2):83–96, 2016.

[2] Kazuko Nishikura. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry*, 79:321–349, 2010.

[3] Meng How Tan, Qin Li, Raghuvaran Shanmugam, Robert Piskol, Jennefer Kohler, Amy N Young, Kaiwen Ivy Liu, Rui Zhang, Gokul Ramaswami, Kentaro Ariyoshi, et al. Dynamic landscape and regulation of rna editing in mammals. *Nature*, 550(7675):249–254, 2017.

[4] Sze Jing Tang, Haoqing Shen, Omer An, HuiQi Hong, Jia Li, Yangyang Song, Jian Han, Daryl Jin Tai Tay, Vanessa Hui En Ng, Fernando Bellido Molias, et al. Cis-and trans-regulations of pre-mrna splicing by rna editing enzymes influence cancer development. *Nature communications*, 11(1):1–17, 2020.

[5] Konstantinos Stellos, Aikaterini Gatsiou, Kimon Stamatelopoulos, Ljubica Perisic Matic, David John, Federica Francesca Lunella, Nicolas Jaé, Oliver Rossbach, Carolin Amrhein, Frangiska Sigala, et al. Adenosine-to-inosine rna editing controls cathepsin s expression in atherosclerosis by enabling hur-mediated post-transcriptional regulation. *Nature medicine*, 22(10):1140–1150, 2016.

[6] Taisuke Nakahama and Yukio Kawahara. Adenosine-to-inosine rna editing in the immune system: friend or foe? *Cellular and Molecular Life Sciences*, pages 1–18, 2020.

[7] Yumeng Wang, Xiaoyan Xu, Shuangxing Yu, Kang Jin Jeong, Zhicheng Zhou, Leng Han, Yiu Huen Tsang, Jun Li, Hu Chen, Lingegowda S Mangala, et al. Systematic characterization of a-to-i rna editing hotspots in micrornas across human cancers. *Genome research*, 27(7):1112–1125, 2017.

[8] Leng Han, Lixia Diao, Shuangxing Yu, Xiaoyan Xu, Jie Li, Rui Zhang, Yang Yang, Henrica MJ Werner, A Karina Eterovic, Yuan Yuan, et al. The genomic landscape and clinical relevance of a-to-i rna editing in human cancers. *Cancer cell*, 28(4):515–528, 2015.

[9] Yukio Kawahara, Kyoko Ito, Hui Sun, Hitoshi Aizawa, Ichiro Kanazawa, and Shin Kwak. Rna editing and death of motor neurons. *Nature*, 427(6977):801–801, 2004.

[10] Eli Eisenberg and Erez Y Levanon. A-to-i rna editing—immune protector and transcriptome diversifier. *Nature Reviews Genetics*, 19(8):473–490, 2018.

[11] Julie M Eggington, Tom Greene, and Brenda L Bass. Predicting sites of adar editing in double-stranded rna. *Nature communications*, 2(1):1–9, 2011.

[12] Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *Journal of computational biology*, 11(2-3):377–394, 2004.

[13] Hui Y Xiong, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan KC Yuen, Yimin Hua, Serge Gueroussov, Hamed S Najafabadi, Timothy R Hughes, et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 2015.

[14] Hannes Bretschneider, Shreshth Gandhi, Amit G Deshwar, Khalid Zuberi, and Brendan J Frey. Cossmo: predicting competitive alternative splice site selection using deep learning. *Bioinformatics*, 34(13):i429–i437, 2018.

[15] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

[16] 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.

[17] Yaron Orenstein, Yuhao Wang, and Bonnie Berger. Rck: accurate and efficient inference of sequence-and structure-based protein–rna binding models from rnacompete data. *Bioinformatics*, 32(12):i351–i359, 2016.

[18] Ilan Ben-Bassat, Benny Chor, and Yaron Orenstein. A deep learning approach for learning intrinsic protein-rna binding preferences. *bioRxiv*, page 328633, 2018.

[19] Shreshth Gandhi, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan J Frey. cdeepbind: A context sensitive deep learning model of rna-protein binding. *bioRxiv*, page 345140, 2018.

[20] Michael KK Leung, Andrew Delong, and Brendan J Frey. Inference of the human polyadenylation code. *Bioinformatics*, 34(17):2889–2898, 2018.

[21] Nicholas Bogard, Johannes Linder, Alexander B Rosenberg, and Georg Seelig. A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, 178(1):91–106, 2019.

[22] Md TA Azad, S Bhakta, and T Tsukahara. Site-directed rna editing by adenosine deaminase acting on rna for correction of the genetic code in gene therapy. *Gene therapy*, 24(12):779–786, 2017.

[23] Liang Qu, Zongyi Yi, Shiyou Zhu, Chunhui Wang, Zhongzheng Cao, Zhuo Zhou, Pengfei Yuan, Ying Yu, Feng Tian, Zhiheng Liu, et al. Programmable rna editing by recruiting endogenous adar using engineered rnas. *Nature biotechnology*, 37(9):1059–1069, 2019.

[24] Gokul Ramaswami and Jin Billy Li. Identification of human rna editing sites: A historical perspective. *Methods*, 107:42–47, 2016.

[25] Gokul Ramaswami, Wei Lin, Robert Piskol, Meng How Tan, Carrie Davis, and Jin Billy Li. Accurate identification of human alu and non-alu rna editing sites. *Nature methods*, 9(6):579–581, 2012.

[26] Ernesto Picardi and Graziano Pesole. Reditools: high-throughput rna editing detection made easy. *Bioinformatics*, 29(14):1813–1814, 2013.

[27] Gokul Ramaswami, Rui Zhang, Robert Piskol, Liam P Keegan, Patricia Deng, Mary A O'connell, and Jin Billy Li. Identifying rna editing sites using rna sequencing data alone. *Nature methods*, 10(2):128–132, 2013.

[28] Qing Zhang and Xinshu Xiao. Genome sequence–independent identification of rna editing sites. *Nature methods*, 12(4):347–350, 2015.

[29] Zongji Wang, Jinmin Lian, Qiye Li, Pei Zhang, Yang Zhou, Xiaoyu Zhan, and Guojie Zhang. Res-scanner: a software package for genome-wide identification of rna-editing sites. *GigaScience*, 5(1):s13742–016, 2016.

[30] Michael Piechotta, Emanuel Wyler, Uwe Ohler, Markus Landthaler, and Christoph Dieterich. Jacusa: site-specific identification of rna editing events from replicate sequencing data. *BMC bioinformatics*, 18(1):7, 2017.

[31] David John, Tyler Weirick, Stefanie Dimmeler, and Shizuka Uchida. Rnaeditor: easy detection of rna editing events and the introduction of editing islands. *Briefings in bioinformatics*, 18(6):993–1001, 2017.

[32] Min-su Kim, Benjamin Hur, and Sun Kim. Rddpred: a condition-specific rna-editing prediction model from rna-seq data. In *BMC genomics*, volume 17, page 5. Springer, 2016.

[33] Heng Xiong, Dongbing Liu, Qiye Li, Mengyue Lei, Liqin Xu, Liang Wu, Zongji Wang, Shancheng Ren, Wangsheng Li, Min Xia, et al. Red-ml: a novel, effective rna editing detection method based on machine learning. *Gigascience*, 6(5):gix012, 2017.

[34] Zhangyi Ouyang, Feng Liu, Chenghui Zhao, Chao Ren, Gaole An, Chuan Mei, Xiaochen Bo, and Wenjie Shu. Accurate identification of rna editing sites from primitive sequence with deep neural networks. *Scientific reports*, 8(1):1–12, 2018.

[35] Xin Liu, Tao Sun, Anna Shcherbina, Qin Li, Kalli Kappel, Inga Jarmoskaite, Gokul Ramaswami, Rhiju Das, Anshul Kundaje, and Jin Billy Li. Learning cis-regulatory principles of adar-based rna editing from crispr-mediated mutagenesis. *bioRxiv*, page 840884, 2019.

[36] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

[37] Ernesto Picardi, Anna Maria D'Erchia, Claudio Lo Giudice, and Graziano Pesole. Rediportal: a comprehensive database of a-to-i rna editing events in humans. *Nucleic acids research*, 45(D1):D750–D757, 2017.

[38] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45 (6):580–585, 2013.

[39] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. Rnacontext: a new method for learning the sequence and structure binding preferences of rna-binding proteins. *PLoS Comput Biol*, 6(7): e1000832, 2010.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[42] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941, 2017.

[43] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[44] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[46] Yulong Song, Wenbing Yang, Qiang Fu, Liang Wu, Xueni Zhao, Yusen Zhang, and Rui Zhang. irclash reveals rna substrates recognized by human adars. *Nature Structural & Molecular Biology*, 27(4):351–362, 2020.