Utilizing Mutations to Evaluate Interpretability of Neural Networks on Genomic Data

Utku Ozbulak[*]
Ghent University

Solha Kang[†]
University of Edinburgh

Jasper Zuallaert
Ghent University

Stephen Depuydt
Ghent University

Joris Vankerschaver
Ghent University

Abstract

Even though deep neural networks (DNNs) achieve state-of-the-art results for a number of problems involving genomic data, getting DNNs to explain their decision-making process has been a major challenge due to their black-box nature. One way to get DNNs to explain their reasoning for prediction is via attribution methods which are assumed to highlight the parts of the input that contribute to the prediction the most. Given the existence of numerous attribution methods and a lack of quantitative results on the fidelity of those methods, selection of an attribution method for sequence-based tasks has been mostly done qualitatively. In this work, we take a step towards identifying the most faithful attribution method by proposing a computational approach that utilizes point mutations. Providing quantitative results on seven popular attribution methods, we find Layerwise Relevance Propagation (LRP) to be the most appropriate one for translation initiation, with LRP identifying two important biological features for translation: the integrity of Kozak sequence as well as the detrimental effects of premature stop codons.

1 Introduction

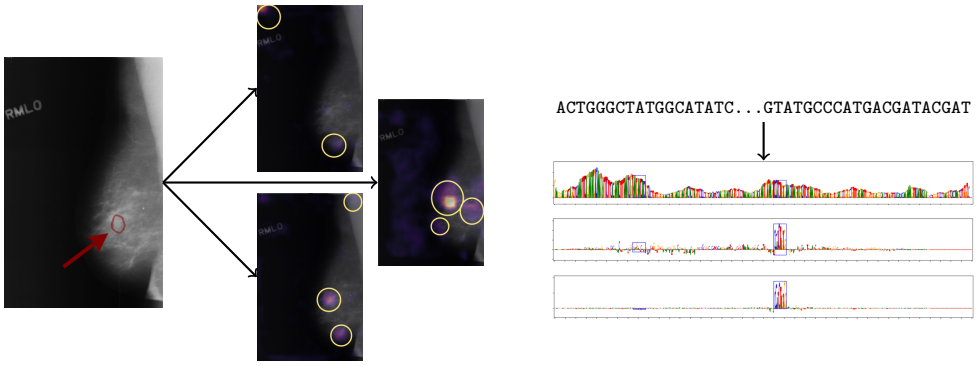
Advancements in machine learning, and deep learning in particular, have contributed to many discoveries in computational biology in recent years, perhaps the most famous research effort being AlphaFold for protein structure prediction [1]. Nevertheless, using DNNs to uncover highly sought-after biologically relevant features remains a non-trivial task [2]. To accomplish this task and to get DNNs to explain their complex reasoning, numerous attribution (also called explainability and interpretability) methods have been employed, with most of these attribution methods originating from the field of computer vision [3]. Unfortunately, even on image data these methods often disagree with each other (see Figure 1). Furthermore, recent research efforts show that the usage of attribution methods may lead to misleading results due to attribution methods not generalizing well or making unjustified causal interpretations [4, 5, 6].

Given the evidence on the instability of attribution methods as well as the availability of a number of unique attribution methods, how can one go about picking the most appropriate attribution method for interpreting DNNs trained with genomic data? A popular method to evaluate the fidelity of attributions for genomic data has been via the usage of synthetic data [7, 8, 9]. Synthetic data in the context of genomics refers to the creation of sequences that contain certain (known) features, such as motifs. Then, the evaluation of attribution methods is performed via observing whether they detect these implanted features or not [9]. In most cases, synthetic data is preferred over genuine data because it can be modified at will, so that specific, quantitative evaluations can be made. On the other

[*]Correspondence: utku.ozbulak@ugent.be

[†]Work done during an internship at Ghent University.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Images (a) Genomic data

Figure 1: An illustration of conflicting attribution maps for (left) breast cancer detection using images [20] and (right) motif detection using genomic data [9].

hand, measuring the generalization of observations from synthetic to real data is not straightforward and, in many cases, an open question. Unlike previous research efforts that employ synthetic data, we will demonstrate in this work the usefulness of point mutations in genuine data to evaluate the interpretability of DNN methods.

Point mutations, in which a single nucleotide is mutated into another one, have a long history in molecular biology in identifying the genetic basis of inherited disases. Although in this context such mutations have been explored mostly in vitro in the past, CRISPR-Cas9 based genome editing enabled an easier and swifter approach [10]. Likewise, these mutations are also employed for computational methods to decipher gene function and regulation [11, 12]. Taking inspiration from these efforts, we employ point mutations to understand the predictions of DNNs.

Employing the relatively well-documented genomics problem of translation initiation site (TIS) detection, we attempt to make a quantitative evaluation of seven popular attribution methods (`DeepLift` [13, 14], `Integrated Gradients` [15], `LRP` [16], `Gradient Shap` [17], `Guided Backpropagation` [18], `Kernel Shap` [17], and `Deconvolution` [19]). With large-scale experiments, we not only reveal large discrepancies between attribution outcomes, we also show how certain attributions can correctly identify well-known biological features.

2 Experimental Settings

Translation initiation site detection – A cornerstone of the central dogma of molecular biology is translation, which denotes the synthesis of appropriate proteins from mRNA molecules [21, 22]. Protein synthesis is principally regulated at the initiation stage, with initiation referring to the assembly of the 80S ribosome in eukaryotes at the start codon with the help of a number of eukaryote initiation factors [23, 24]. As such, precise identification of the translation initiation site on mRNAs is crucial for accurate protein synthesis [22]. TIS detection is a complex problem which found a number of well-recognized solutions that leverage DNNs in recent years [25, 26, 27]. Moreover, the availability of public datasets also makes it an attractive problem for model benchmarking. And finally, the biological processes regulating this process on genome-level are relatively well understood compared to a number of other sequencing problems [23]. Thanks to the properties above, we select TIS-detection to evaluate the faithfulness of attribution methods.

Data – We make use of the `Human-TIS`, `CCDS`, and `Chromosome-21` TIS detection datasets published by [28, 29] where the sequences in these datasets are gathered from human DNA sequences (see the supplementary material for details). TIS-detection is a 2-class classification problem with inputs composed of various permutations of four base pairs (bp) (`A`,`C`, `T`, `G`). The datasets contain sequences consisting of 203 bp with 60 bp in the 5' untranslated region (UTR) and 140 bp downstream of the canonical translation initiation site (`ATG`). A unique property of these data is the codon starting at position 61, which denotes the start of translation. We assign an index of $+1$ to `A` in `ATG` at position 61 and we label upstream (UTR) and downstream (coding region) nucleotides with decreasing and increasing numbers, respectively.

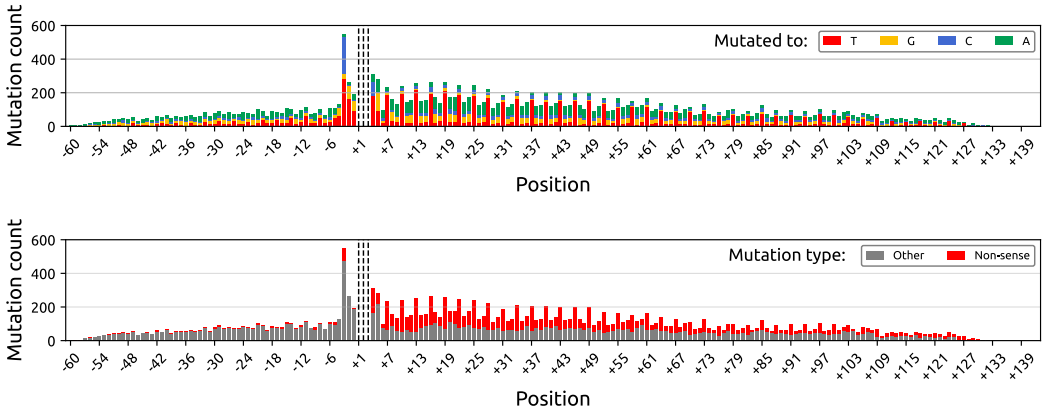


Figure 2: (top) Histogram of mutations per position that, when performed, change the prediction made by the model from TIS-positive to TIS-negative and (bottom) distribution of mutations when separated into non-sense and others.

Model – We use the TISRover architecture described in the work of Zuallaert et al. [27] which achieves state-of-the-art results on this data. Detailed results on the model performance as well as the training routine are provided in the supplementary material.

Attribution methods – As mentioned in Section 1, we evaluate seven commonly used attribution methods given in Section 1. For the implementation of these methods we use the PyTorch Captum library [30]. More details for these attribution methods can be found in their respective paper as well as the comprehensive works of [3, 31].

3 Experimental Results

3.1 Mutations

Given a TIS-positive sequence consisting of 203 bp, we create 600 newly mutated sequences by transforming bp at each position, one at a time, with the other three bases for 200 positions (not mutating the ATG at position 61). Our goal in performing this task is to analyze TIS-positive sequences that have their predictions changed to TIS-negative with the introduction of a single point mutation. Biologically, this means that, due to the introduced mutation, translation either fails or does not begin and the protein is not properly synthesized.

With the above process, we create $717,000$ mutated sequences originating from $1,195$ initially correctly classified TIS-positive sequences. Of those mutated sequences, we observe that $18,682$ have their predictions changed from TIS-positive to TIS-negative, with mutations causing this change.

In Figure 2, we provide the histogram of mutations where the mutation of base pairs to other bases in the graph cause a TIS-positive sequence to be classified as a TIS-negative. Based on results presented in Figure 2, we can align certain patterns with well-known biological features.

Importance of Kozak sequence – For humans as well as other higher eukaryotes, the sequence nearby the start codon is determined to play a significant role in translation initiation, with this (Kozak) sequence defined as 5'- GCCRCCATGG -3' with R representing a purine (i.e., A or G) and ATG representing the start codon [32]. It was discovered that the most influential base pairs in this sequence are the ones in position -3 (R) and +4 (G) relative to the start codon. Our experiments indicate that mutations made in these positions are also the ones that impact the prediction the most, with mutations on position -3 and +4 (often to T, which is the least occurring base pair in Kozak sequence) having the largest influence over prediction.

Premature stop codons – The final stage of translation is termination, which signifies the end of the elongation of the peptide-chain, with this process requiring an in-frame stop codon (TAA, TGA, TAG). Premature termination of translation due to the introduction of a non-sense mutation often leads to specialized processes which prevent the creation of faulty proteins [33]. Our experiments indicate that most of the mutations that lead to TIS-negative prediction are the ones that create a stop codon and thus lead to a non-sense mutation. We provide the proportion of non-sense mutations to others in

Match	Region	DeepLift	Integrated Gradients	LRP	Gradient Shap	Guided Backprop	Deconvolution	Kernel Shap
Exact	Anywhere	41.9%	44.3%	**45.1**%	36.4%	42.0%	27.8%	7.0%
	· Kozak	32.8%	46.0%	**46.7**%	43.1%	37.4%	27.2%	13.3%
	· Upstream	20.1%	28.0%	**29.6**%	25.0%	22.7%	18.7%	7.3%
	· Downstream	49.1%	49.6%	**50.2**%	40.2%	48.3%	30.8%	6.9%
	· · Stop codons	96.7%	96.5%	**97.9**%	78.1%	96.6%	60.2%	11.4%
	· · · `TAA`	96.5%	96.4%	**97.9**%	77.1%	96.5%	63.6%	10.3%
	· · · `TAG`	97.4%	97.2%	**98.3**%	79.5%	96.9%	59.4%	12.1%
	· · · `TGA`	95.3%	95.2%	**97.0**%	76.2%	96.0%	58.4%	10.8%
	· · Other	2.3%	**3.6**%	3.3%	3.0%	1.0%	1.9%	2.5%

Table 1: Percentage of attribution maps created with methods in the first row that identify mutations which occur in various regions of the sequence using the methodology described in Section 3.2.

Figure 2 where it is clear that these mutations mostly cause strong TIS-negative signals downstream of the start codon, as opposed to upstream.

3.2 Attributions

The purpose of attribution methods is to identify important regions of the input data under consideration, where the information contained in these regions leads the examined DNN to make the type of prediction it makes [18]. Outputs of attribution methods are often called attribution maps, and contain a measure of relative importance for each input location (e.g., pixel positions for images or nucleotide positions in genomic data).

By employing the seven attribution methods given in Section 1 and by utilizing the $18,682$ mutated sequences that result in a TIS-negative prediction, we create $130,774$ attribution maps. Based on these attribution maps we identify the most important bp position as the location of largest attribution.

Mutation identification by attribution methods – Given the nature of the experimental routine discussed in Section 3.1, we were able to identify the position that contributes the most to a TIS-negative prediction made by the model. Now, we investigate whether or not attribution methods can identify the positions of those mutations. Note that, although mutations occur at the DNA level, their effects may become apparent at the amino-acid level, which is the reason for stop codons having particular effects as discussed above. As a result of this, we evaluate the match for attribution on the level of codons, meaning that, if the attribution map highlights any of the three base pair of the mutated codon, we consider it a correct match. Furthermore, for mutations occurring in the Kozak sequence, we also consider an attribution to be a correct one if it highlights any region of the Kozak sequence.

Based on the evaluation routine described above, we provide Table 1 which contains the percentage of attribution maps that are able to identify mutated codons for different regions of the sequence. Surprisingly, we observe that attribution methods have large discrepancies among each other when identifying appropriate regions, with `LRP` being the most accurate attribution method in identifying mutated codons. Furthermore, we also observe that stop codons that cause a premature termination of translation can be identified with a surprising accuracy ($+97\%$) even though these mutations are spread all over the downstream area (see Figure 2). Yet, for other mutations that occur in the downstream region, the accuracy of attribution shows a dramatic drop.

4 Conclusions

Utilizing the point mutations which see common use both in vivo and in vitro experiments, in this work, we performed a quantitative evaluation of attribution methods for DNN interpretability. We identified `LRP` to be the most accurate attribution method, with this method identifying biologically relevant features with surprising accuracy, yet still failing for half of the mutation scenarios, which raises the following question: why can attribution methods identify non-sense mutations in downstream regions extremely accurately, yet fail to identify other types of mutations? Supporting the observations of [7] we also believe the answer lies in the evaluation of multiple models with varying degrees of performance in order to measure the correlation between the performance of the model and its ability to identify mutations. Nevertheless, we demonstrated that attribution methods, while being imperfect, can highlight regions relevant to the prediction for translation initiation and urge future research efforts to employ biologically significant experimentations for other datasets as well as for other models in order to discover faithfulness of proposed methods.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[2] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

[3] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[4] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 39–68. Springer, 2022.

[5] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium*, 2020.

[6] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[7] Peter K Koo, Sharon Qian, Gal Kaplun, Verena Volf, and Dimitris Kalimeris. Robust neural networks are more interpretable for genomics. *bioRxiv*, page 657437, 2019.

[8] Peter K Koo and Matt Ploenzke. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence*, 3(3): 258–266, 2021.

[9] Eva I Prakash, Avanti Shrikumar, and Anshul Kundaje. Towards more realistic simulated datasets for benchmarking deep learning models in regulatory genomics. In *Machine Learning in Computational Biology*, pages 58–77. PMLR, 2022.

[10] Masafumi Inui, Mami Miyado, Maki Igarashi, Moe Tamano, Atsushi Kubo, Satoshi Yamashita, Hiroshi Asahara, Maki Fukami, and Shuji Takada. Rapid generation of mouse models with defined point mutations by the CRISPR/Cas9 system. *Scientific Reports*, 4(1):1–8, 2014.

[11] Ivan De Falco, Antonio Della Cioppa, and Ernesto Tarantino. Mutation-based genetic algorithm: performance evaluation. *Applied Soft Computing*, 1(4):285–299, 2002.

[12] Yicheng Zhu, Cheng Soon Ong, and Gavin A Huttley. Machine learning techniques for classifying the mutagenic origins of point mutations. *Genetics*, 215(1):25–40, 2020.

[13] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[14] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[16] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10(7):e0130140, 2015.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[18] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[19] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[20] Jimmy Wu, Bolei Zhou, Diondra Peck, Scott Hsieh, Vandana Dialani, Lester Mackey, and Genevieve Patterson. Deepminer: Discovering interpretable representations for mammogram classification and explanation. 2021.

[21] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, 2009.

[22] Marilyn Kozak. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell*, 15(4):1109–1123, 1978.

[23] Richard J Jackson, Christopher UT Hellen, and Tatyana V Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113–127, 2010.

[24] Vanja Haberle and Alexander Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637, 2018.

[25] Sai Zhang, Hailin Hu, Tao Jiang, Lei Zhang, and Jianyang Zeng. TITER: predicting translation initiation sites by deep learning. *Bioinformatics*, 33(14):i234–i242, 2017.

[26] Chao Wei, Junying Zhang, Xiguo Yuan, Zongzhen He, Guojun Liu, and Jinhui Wu. Neurotis: Enhancing the prediction of translation initiation sites in mrna sequences via a hybrid dependency network and deep learning framework. *Knowledge-Based Systems*, 212:106459, 2021.

[27] Jasper Zuallaert, Mijung Kim, Arne Soete, Yvan Saeys, and Wesley De Neve. TISRover: ConvNets learn biologically relevant features for effective translation initiation site prediction. *International Journal of Data Mining and Bioinformatics*, 20(3):267–284, 2018.

[28] Yvan Saeys, Thomas Abeel, Sven Degroeve, and Yves Van de Peer. Translation initiation site prediction on a genomic scale: beauty in simplicity. *Bioinformatics*, 23(13):i418–i423, 2007.

[29] Wei Chen, Peng-Mian Feng, En-Ze Deng, Hao Lin, and Kuo-Chen Chou. iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical biochemistry*, 462:76–83, 2014.

[30] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, 2020.

[31] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[32] Angelita Simonetti, Ewelina Guca, Anthony Bochler, Lauriane Kuhn, and Yaser Hashem. Structural insights into the mammalian late-stage initiation complexes. *Cell Reports*, 31(1): 107497, 2020.

[33] Louise J Johnson, James A Cotton, Conrad P Lichtenstein, Greg S Elgar, Richard A Nichols, Steven C Le Comber, et al. Stops making sense: translational trade-offs and stop codon reassignment. *BMC Evolutionary Biology*, 11(1):1–8, 2011.

[34] Jasper Zuallaert, Mijung Kim, Yvan Saeys, and Wesley De Neve. Interpretable convolutional neural networks for effective translation initiation site prediction. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1233–1237. IEEE, 2017.

Appendix

Data – An overview of the datasets used in this study is given in Table 2 which details the positive to negative ratio of the labels as well as total number of sequences that fall into each category.

Pre-processing – We use the standard one-hot-encoding routine for genomic data as described in the work of [27, 34] which converts the genomic representation from strings to vectors as follows:

$$\boldsymbol{x} = [\underbrace{\mathtt{N}_{-60}\,\mathtt{N}_{-59}\,\ldots\,\mathtt{N}_{-2}\,\mathtt{N}_{-1}}_{\text{Upstream}}\ \ \mathtt{ATG}\ \ \underbrace{\mathtt{N}_{+4}\,\mathtt{N}_{+5}\,\ldots\,\mathtt{N}_{+142}\,\mathtt{N}_{+143}}_{\text{Downstream}}]\ . \tag{1}$$

with $\mathtt{N}_k \in \{[1,0,0,0],[0,1,0,0],[0,0,1,0],[0,0,0,1]\}$ for `A`, `C`, `T`, and `G`, respectively. We also employ with masking as an augmentation technique in order to improve the accuracy of the model where we replace one-hot representations of bp with empty representations. This operation corresponds to replacing the above vector depictions with a zero vector (i.e., $[0,0,0,0]$).

Model performance – [28] argues that commonly used metrics such as accuracy, true positive rate, and false positive rate can be misleading for datasets with an extreme skew in labels such as `Chromosome-21` and proposed the usage of false-positive rate at fixed sensitivity of 0.8 (FPR.80) for benchmarking in TIS-detection. For easy comparability to the previous research efforts, we also use FPR.80 as an error measurement and provide the results of the best-performing models in Table 3.

Dataset	Total	TIS-positive sequences	TIS-negative sequences	Positive/Negatie ratio	Source
`CCDS`	364,495	13,917	350,578	0.0396	[28]
`Chromosome-21`	1,267,701	258	1,267,443	0.0002	[28]
`Human-TIS`	2,718	1,359	1,359	1	[29]

Table 2: Details of TIS-detection datasets used in this study.

Model	FPR.80
Saeys et al. [28]	0.125
TIS-Rover (reported in [34])	0.031
TIS-Rover (reproduced)	0.032
TIS-Rover (trained with SGD)	0.030
TIS-Rover (SGD + bp masking)	0.029

Table 3: FPR.80 performance of models on `Chromosome-21` dataset.

Distance between attributions and mutations – In the main text, in order to measure the correctness of attribution maps, we measured the whether or not the largest attribution location lies within one of the three bp of the codon that is mutated. In Table 4, we provide experimental results for showing mean and median codon distance between the attributions and mutations for the same regions. With the results displayed in Table 4, we show that, for mutations that do not occur in the Kozak sequence or the ones that do not create a stop codon in downstream region, attribution methods highlight regions that are, on average, far away from the mutations.

Match	Region	DeepLift	Integrated Gradients	LRP	Gradient Shap	Guided Backprop	Deconvolution	Kernel Shap
Distance to mutation	Anywhere	9.3 (3)	8.7 (2)	8.9 (2)	9.9 (5)	10.8 (3)	17.3 (13)	17.2 (15)
	· Kozak	10.3 (7)	6.9 (0)	8.4 (4)	6.8 (0)	10.6 (7)	13.0 (16)	14.3 (11)
	· Upstream	14.2 (11)	10.5 (8)	11.1 (7)	10.4 (8)	13.1 (9)	10.2 (9)	16.6 (12)
	· Downstream	7.6 (1)	8.1 (1)	8.1 (0)	9.8 (3)	10.0 (1)	19.6 (19)	17.4 (16)
	· · Stop codons	1.0 (0)	1.2 (0)	0.8 (0)	3.9 (0)	1.5 (0)	12.1 (0)	16.1 (14)
	· · · `TAA`	1.2 (0)	1.3 (0)	0.9 (0)	4.0 (0)	1.6 (0)	11.9 (0)	16.5 (14)
	· · · `TAG`	0.7 (0)	0.9 (0)	0.7 (0)	3.6 (0)	1.4 (0)	11.4 (0)	15.9 (14)
	· · · `TGA`	1.4 (0)	1.8 (0)	1.2 (0)	4.6 (0)	1.7 (0)	13.5 (0)	16.1 (14)
	· · Other	14.7 (13)	15.4 (13)	15.9 (14)	16.0 (14)	19.0 (17)	27.6 (28)	18.7 (17)

Table 4: Mean (median) distance between the codon highlighted by attribution maps created with methods in the first row and codons that are mutated. The results are provided separately for a number of regions in the second column.