TIS TRANSFORMER: RE-ANNOTATION OF THE HUMAN PROTEOME USING DEEP LEARNING

A PREPRINT

Jim Clauwaert*
Department of Data Analysis and Mathematical Modelling
Ghent University, Belgium
jim.clauwaert@ugent.be

Zahra McVey
Novo Nordisk Research Centre Oxford
Novo Nordisk Ltd.
Oxford, United Kingdom
zqab@novonordisk.com

Ramneek Gupta
Novo Nordisk Research Centre Oxford
Novo Nordisk Ltd.
Oxford, United Kingdom
rmgp@novonordisk.com

Gerben Menschaert
Department of Data Analysis and Mathematical Modelling
Ghent University, Belgium
gerben.menschaert@ugent.be

November 17, 2021

ABSTRACT

The precise detection of translation initiation sites is essential for proteome delineation. In turn, the accurate mapping of the proteome is fundamental in advancing our understanding of biological systems and cellular mechanisms. We propose TIS Transformer, a deep learning model for the determination of translation start sites, based on information embedded in processed transcript nucleotide sequences. Through the application of deep learning techniques first designed for natural language processing tasks, we have developed an approach that achieves state-of-the-art performances on the prediction of translation initiation sites. TIS Transformer utilizes the FAVOR+ algorithm for attention calculation, enabling processing of full transcript sequences by the model. Analysis of input importance revealed TIS Transformer's ability to detect key features of translation, such as translation stop sites and reading frames. Furthermore, we demonstrate TIS Transformer's ability to detect multiple peptides on a transcript, and peptides encoded by short Open Reading Frames (sORFs), either alongside a canonical coding sequence or in long non-coding RNAs. Using a cross-validation scheme, we apply TIS Transformer to re-annotate the full human transcriptome.

Keywords Deep learning · Translation · Transcriptomics · Self-attention

1 Introduction

Translation is the generation of proteins from messenger RNA (mRNA). The nucleotide sequence of mRNA not only encodes proteins through its codon structure, but also influences other factors, such as the location and efficiency of the translation process [28]. The complexity of the nucleotide sequence is highlighted by the variety of gene products, playing roles in multiple molecular processes. Partly due to this complexity, existing mappings of the genome and transcriptome are still based on the combination of proof-of-existence data with statistical tests. Genome annotation platforms such as Ensembl, NCBI and USCS rely largely on sequence alignment methods [1, 23]. Flaws of such systems are inherent to the knowledge at hand, where biases and errors are known to be propagated with time [8]. Subsequently, changes in existing annotations are made as new data on gene products or molecular mechanisms become available.

*To whom correspondence should be addressed. Tel: +32 926 49922; email: jim.clauwaert@ugent.be

A comprehensive understanding of translation can advance the identification of novel proteins. However, the underlying biological complexity, and the obscurity of the ground truth, have hindered the development of successful approaches. In recent years, black-box machine learning models have gained attention due to their ability to attain high performances when sufficient data is available. These predictive models can be generated without the requirement of a complete understanding of the underlying biological determinants. Prior work has attempted to predict TISs sites based solely on intronless transcript sequences. The GenemarkS-T model, proposed by Tang et al. [22], applies an iterative supervised training approach on individual clusters, separated through unsupervised learning on G+C content of the transcript. Deep learning techniques have also gained success due to their automated feature learning [7], as can be seen through the application of neural networks on sequence and omics data. Zhang et al. [34] use a combination of convolutional and recurrent layers to process a transcript sequence of fixed length around the TIS . Similarly, Zuallaert et al. [35] have applied convolutional neural networks to determine the location of TISs. Although the predictive performance of these models still poses restrictions on their wider applications, insights can be gained from the evaluation of the underlying decision-making process.

Deep learning techniques have the unique strength of deducing relevant features when applied to complex problems and big data. Standard machine learning methods and traditional neural network architectures remain restrictive in mapping the input features to the target label. Relations are, for most architectures (e.g. fully-connected layers, convolutions), explicitly mapped by position. In contrast to these methods, attention allows the model to determine the relevancy between two pieces of information based on their information embedding and (relative) positioning. For many natural language processing tasks, architectures incorporating attention have been proven superior [4, 18]. The previously popular recurrent neural networks are now being largely replaced by networks applying attention, such as the transformer network [24]. Some recent studies have successfully investigated the use of transformer networks on genome data, such as proposed by Zaheer et al. [33], for the prediction of promoter regions and chromatin profiling. Ji et al. [10] introduce DNABERT, a transformer model pre-trained on genome data that can be fine-tuned for a specific tasks such as the annotation of promoters or splice sites.

The computational cost of attention imposes limitations on the maximum input length of sequential data. Specifically, this cost scales quadratically with respect to the length over which attention is calculated, where the length of previous studies is traditionally capped at 512 units [10, 24]. Several recent studies have focused on overcoming this limitation through the use of approximation methods, allowing transformer networks to apply attention over larger sequences [27, 29, 5].

In this work, we propose TIS Transformer, a tool for the identification of TISs based on the transcript sequence, using a proposed scaling solution for computing attention. Models are trained and evaluated on the processed human transcriptome; therefore, input sequence data do not contain introns or intergenic regions. We apply the Performer model, a high performing transformer architecture that scales linearly with the input length, enabling us to apply attention over the full transcript sequences [5]. We achieve state-of-the-art performances in the prediction of TIS and we observe that model error is correlated with the quality score of annotated TISs or transcripts. Evaluation on the decision-making process shows that the trained model implicitly detects both translation stop sites and reading frames. Furthermore, although not included in the annotations used for training, we are able to identify multiple CDSs on single transcripts, concurring with previously reported uORFs [16]. Given the high performances of the tool and the aforementioned advantages, we believe this approach to be a robust new method for the discovery and annotation of novel proteins, applicable in a wide range of study designs. By using a cross-evaluation scheme across the chromosomes, we successfully provided annotations for the full human transcriptome. Finally, as experimental data is not needed, the model can be used to assess or identify proteins expressed on newly discovered or designed transcripts. Given the high performances of TIS Transformer, we believe this approach to be a robust new method for the discovery and annotation of novel proteins, with wide applications ranging from genetic engineering to target discovery.

2 Material and Methods

2.1 Training objective

The goal of the study is to create a predictive model that is able to detect the occurrence of proteins on the transcriptome. Rather than predicting a paired TIS and translation stop site, we decided to achieve this goal through the detection of TISs. Translation stop sites are determined post-hoc through detection of the first in-frame stop codon. As such, the model is optimized to perform binary classification between TISs and non-TISs for each position of the human transcriptome.

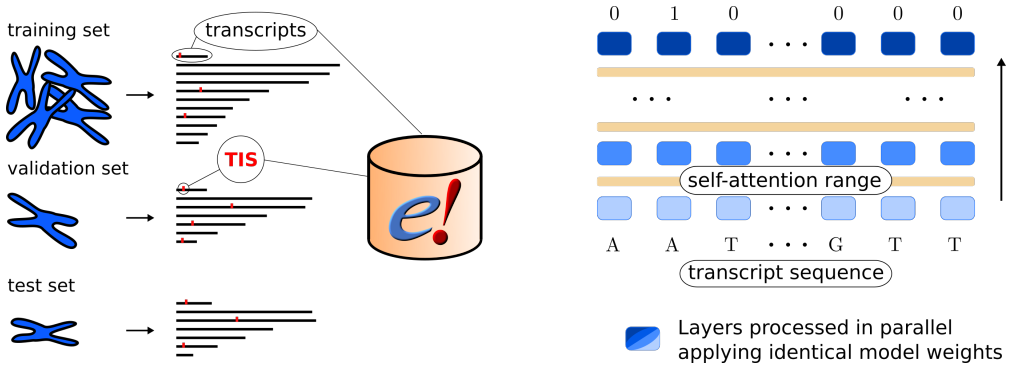

Figure 1: **Schematic of the data and model set-up.** (**left**) The Ensembl annotation (version 102) is used to determine transcript sequences and translation initiation sites (TISs). Transcripts are grouped by chromosome to create a training, validation and test set. (**right**) The performer model allows processing of full transcript sequences, evaluating data through the layers in parallel to obtain model outputs at each position. The model architecture can handle varying input lengths, as identical model weights are applied to transform the data. Through self-attention, sequential information from any site on the transcript can be queried by the model to determine the presence of TISs at any position.

2.2 Model input and output

The model processes the full sequence of the transcript to impute the presence of TISs at each position (Figure ref1). The input of the model is represented by $\boldsymbol{H}^{\text{in}} \in \mathbb{R}^{n \times d}$, where $n < 25,000$ and d the dimension of the hidden states. $\boldsymbol{H}_i^{in}$ denotes the nucleotide embedding (A, T, C ,G , N, [START], [STOP]) at position i of the transcript (thymine is used to proxy uracil throughout the paper). The input embeddings are trained as part of the optimization process. The [START] and [STOP] tokens are used at before the beginning and after the end of the transcript, relatively. Identical in dimensions to the input embeddings, the output $\boldsymbol{H}^{\text{out}} \in \mathbb{R}^{n \times d}$ of the transformer network is processed by a pair of feed-forward layers that result in a model output at each nucleotide position of the transcript. The loss is computed from predictions at inputs $\in \{A, T, C, G, N\}$.

2.3 Transformer architecture

The transformer architecture is a recent methodology that implements (self-)attention for a group of inputs. Through positional encodings, a sequential ordering can be imposed between these inputs. Through the attention mechanism, the network is optimized to determine the flow of information based on the information itself. The transformer network is characterized by several sequential layers with identical design.

The principal element is the use of self-attention, performed by a module called the attention head. Here, a query, key and value matrix are derived from the input X in each layer:

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_q^\top, \boldsymbol{X}\boldsymbol{W}_k^\top, \boldsymbol{X}\boldsymbol{W}_v^\top, \tag{1}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n \times d_k}$. For the first layer, $\boldsymbol{X}$ equals $\boldsymbol{H}^{\text{in}}$. Multiplication and subsequent normalization between $\boldsymbol{Q}$ and $\boldsymbol{K}^\top$ returns an $n \times n$ matrix, whose values determine the flow of information to (e.g. $(\boldsymbol{Q}\boldsymbol{K}^\top)_{i,:}$) and from (e.g. $(\boldsymbol{Q}\boldsymbol{K}^\top)_{:,i}$) each hidden state through multiplication with $\boldsymbol{V}$. For each attention head, a matrix $\boldsymbol{Z}$ is calculated, holding the combined information for each input hidden state.

$$\boldsymbol{Z} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

where $\boldsymbol{Z} \in \mathbb{R}^{n \times d_k}$. Multiple sets of $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ make it possible to route multiple types of information from $\boldsymbol{X}$ in each layer. The outputs $\boldsymbol{Z}$ that are derived from different attention heads are concatenated and transformed to fit the dimension size of the input $\boldsymbol{X}$, with which they are summed (i.e. residual connection). Each layer is built using the same components, but each feature unique weights which determine how the information is routed. The architecture allows processing of transcripts of different lengths, as the same weights are applied to calculate the $\boldsymbol{q}$, $\boldsymbol{k}$ and $\boldsymbol{v}$ vectors for each hidden state, which are calculated in parallel and represented as rows in the matrices $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$.

To allow long-range attention over full transcripts, we used a recent innovation in calculating attention introduced by Choromanski et al. [5]. Equation 2 is exchanged with the Fast Attention Via Positive Orthogonal Random Features (FAVOR+) algorithm, which utilizes random feature maps decompositions for approximation in order to obtain a computational complexity that scales linearly with respect to the length of the sequence. [5].

2.4 Data sets and model optimization

Due to its active updates, the Ensembl assembly of the Human genome (GRCh38.p13; release 102) was selected to construct the input data. The data comprises 89,736 (38.7%) protein-coding and 142 184 (61.3%) non-coding transcripts. These constitute 374,446,434 RNA nucleotide positions, of which 89,736 are positively labeled as a TIS (0.024%). The training, validation, and test set are allocated in accordance to chromosomes, thereby grouping transcript isoforms and proteoforms together. A total of 24 models were trained to sequentially use each chromosome as a test set, where the other chromosomes make the training and validation set (Figure 1, Supplementary Table A1). For the benchmark results, chromosome 11 and 4 were randomly selected to be the validation and test set at the start of the study. Training was stopped when a minimum loss on the validation set is obtained. All reported performance metrics are obtained on the test sets.

Notwithstanding the size of the data set and overall high computational requirements of transformer architectures, model optimization is possible on a single RTX 2080 and converged after ca. 35 hours due to the relative shallowness of the final transformer architecture (Supplementary Tables A3–A5). In an attempt to improve performances and the speed at which the validation loss converges, a base model was first created through a self-supervised learning scheme (i.e. masked language modelling) on the transcript data. Here, the training objective is the imputation of randomly masked inputs to the model, allowing an initial mapping of transcriptome semantics by the model. The optimization task was performed using a similar set-up as first described by Zaheer et al. [33] (see Supplementary Files for additional information). Both the self-supervised and supervised training process apply the cross-entropy loss function. The details of varying model architecture performances are given by Supplementary Table A8 and Supplementary Figure A1–4.

3 Results

3.1 Benchmark

The novel transformer network we developed for the detection of TIS was then benchmarked against recently published studies. TITER is a neural network that applies convolutional and recurrent architectures to process transcript sequences [34]. The study evaluates the model through the use of custom annotations obtained from ribosome profiling data. TISRover is convolutional neural network trained using annotations of the consensus coding sequence data set (CCDS) [35]. Unlike the method discussed in this paper, previous architectures are generally not applied to evaluate each position on the full transcriptome due to the differences in which data is processed. Specifically, both methods evaluate a 200-nucleotide window around the site of interest, where the input sequence required for the imputation of neighboring sites, albeit largely identical, have to be processed separately. Evaluation of the full transcriptome thereby requires the model to process a total sequence length that is 200 times the full transcriptome length. By reducing the data set, it is possible to obtain feasible training times for our benchmark experiment. To achieve this, only ATG sites were applied to train and evaluate both methods, effectively reducing training times a hundredfold. Further details on the benchmark experiment are provided in the Supplementary Files. Benchmark results are given in Table 1.

Comparing the receiver operating characteristic curve (ROC AUC) and area under the precision-recall curve (PR AUC) scores of all methods on all transcripts derived from chromosome 4, the TIS Transformer largely outperforms previous methods (Table 1). Despite the limitations discussed, the discrepancy between performances is corroborated by an improved architectural design and previous studies indicating the superiority of transformer models on processing sequence data.

Table 1: **Performance of the TIS Transformer on annotating TIS sites on the human transcriptome.** To benchmark our method with previous studies, the data set was reduced by only evaluating ATG sites. Performances are obtained on the test set (chromosome 4)

Task	Model	Model weights	ROC AUC	PR AUC
All sites	TIS Transformer	∼900K	99.99	85.63
ATG sites	TIS Transformer	∼900K	99.63	83.59
	TISRover [35]	∼530K	97.20	43.32
	TITER [34]	∼431K	61.99	1.69

3.2 In-depth evaluation of false predictions

The PR AUC score of the benchmark indicates that the TIS Transformer is well capable of handling the large class imbalance between the positive and negative set. The output probabilities of the negative and positive samples can be

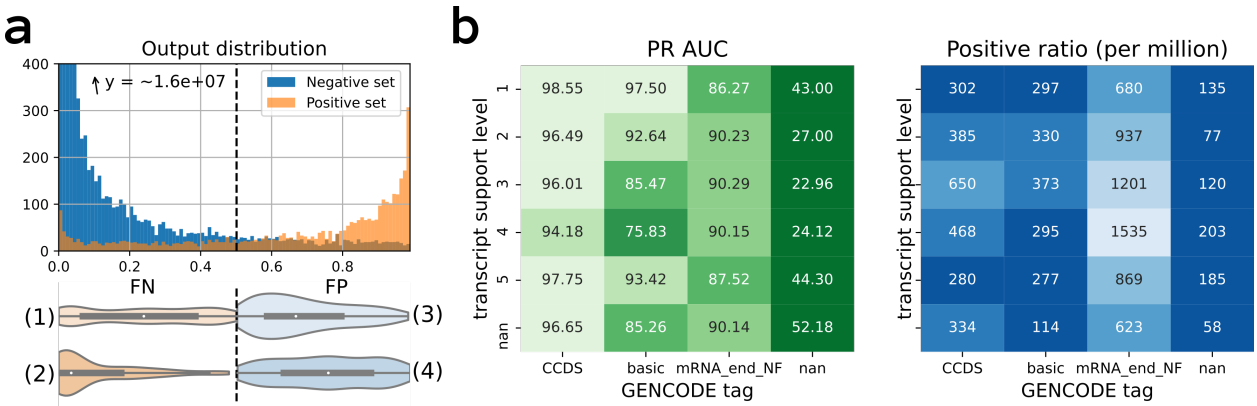

Figure 2: **Model performances on predicting translation initiation sites.** (**a**) A histogram of negative (blue) and positive (orange) labels (top). The dotted line represents a possible threshold separating the positive and negative predictions. (bottom) Violinplots showing the distribution of model outputs grouped by the false negative predictions (2) with and (1) without at least one alternative (i.e. same transcript) predicted translation initiation site present. False positive predictions are grouped by resulting proteins (3) not returning and (4) returning a match using pBLAST or PROTEOFORMER. (**b**) Model performances binned by transcripts tags. PR AUC performances (left) and the ratio of positive samples in each group (right) averaged for all chromosomes are obtained by binning the transcripts according to transcript support level and tags given to the annotated translation initiation site or transcript (if any, otherwise nan). Data includes the prediction of all transcripts, utilizing the 24 different models cross evaluated using each chromosome.

clearly distinguished as they are concentrated at 0 and 1, respectively (Figure 2a). Similarly, this can be observed when evaluating the model outputs at the single transcript level (e.g. Figure 4).

By grouping the transcripts and TISs of the test sets in accordance to characterizing tags, several patterns emerge from the data (Figure 2b). The utilized tags are allocated by GENCODE, where the transcript support level quantifies the support for a transcript to exist, based on available *in vivo* evidence [30]. The most common tags have been included, such as those given to TISs that are part of the consensus coding sequence database (CCDS), the transcripts identified as representative transcripts for each gene (basic), and the transcripts for which the mRNA end could not be confirmed (mRNA_end_NF). Supplementary Figure A5 includes more information for each group including the ROC AUC scores and number of nucleotide positions (samples) in each group.

We observe three patterns of interest. First, the performance of the model is correlated to the transcript support level for the bulk of annotations (CCDS, basic). With the exception of transcript support level 5, which is denoted as 'no single transcript supports the model structure'. We hypothesise the presence of these transcripts within the annotated set to be supported by factors not captured by the transcript support level, thereby explaining their relatively higher performances. Second, the model performs worse for the group of transcripts where no tags are given to the transcript or TISs. Third, the differences in performance between groups is largely uncorrelated to the ratio of positive samples along rows and columns, which could otherwise have been the driving factor for differences in PR AUC scores.

Of further interest are the nucleotide positions for which the model returns uncertain probabilities (i.e. between 0 and 1), constituting only a fraction of total samples in the test set. For evaluation purposes, a set of positive predictions is selected equal to the number of positive annotations provided by Ensembl. For the false negative predictions, a distinction exists between output distributions of false negatives with and without an alternative TIS on the transcript (Figure 2a). The false positive predictions were evaluated against external data sets using two additional methods (Supplementary Table A2). Firstly, using pBLAST, the proteins resulting from the false positive TISs are searched in the UniProt databases (Swiss-Prot/TrEMBL/isoforms). The conditions applied required an identical start, stop, length, and 95% sequence overlap between two proteins. On average across the chromosomes, 35% of the false positive matched a protein in the UniProt database. Secondly, using ribosome profiling data incorporating both cycloheximide and lactimidomycin treated experiments, a high resolution footprint can be obtained around translated TISs [13]. Expressed TISs (GSE58207, GSE74279) have been analyzed for Ensembl version 102 using the PROTEOFORMER tool [25]. Ribosomal profiling hits from both data sets agreed with 17% of the false positive predictions. For comparison, the applied ribosomal data accounted for 54% of the Ensembl annotated TISs. Together, an average of 45% of the false positive predictions are substantiated by either pBLAST or PROTEOFORMER (Figure 2a; Supplementary Table A2).

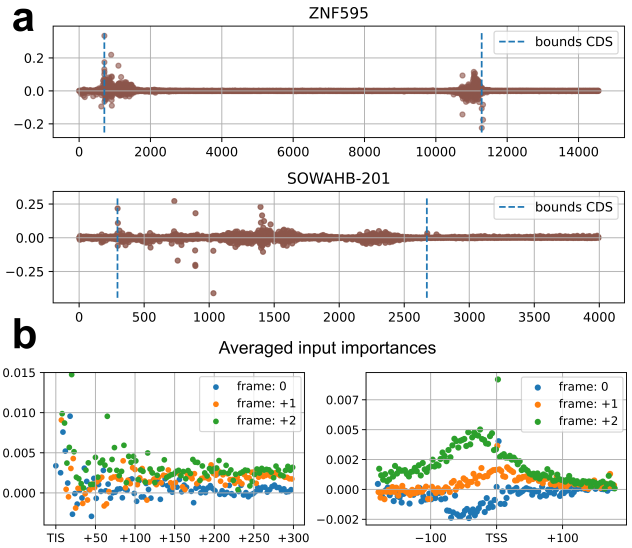

Figure 3: **Attribution scores to input nucleotides obtained through integrated gradients.** (**a**) Scores shown are given by the model for the translation initiation site on position 708 of WDFY3 (ENST00000295888) and at position 296 of SOWAHB (ENST00000334306). (**b**) Attribution scores for the positions surrounding the translation initiation site (TIS) and translation stop site (TSS) averaged for all annotated translation initiation sites and companion translation stop sites in Ensembl.

3.3 Input analysis

Several techniques exist that allow us to gain insights into the decision-making process of the trained model. Here, we apply integrated gradients to evaluate the relative contribution of input nucleotides on the transcript to the predicted output. Integraded gradients, first introduced by [21], utilizes the partial derivatives the inputs with respect to the output to assign attribution scores. We observe several expected patterns, such as the high importance of the candidate TIS itself and surrounding areas (e.g. Kozak sequence context). Interestingly, in addition to this, we determine that both the translation stop sites and reading frame have influence towards determining TISs (Figure 3b). This reveals the ability of the model architecture to learn biological features of translation, without being provided this information explicitly. We hypothesise that the relatively higher importances of the second and third element of the reading frame can be ascribed to the higher importance towards determining the exact identity of the amino acid at each position. Further analysis of these attribution score profiles offer exciting opportunities as these may reveal additional biological factors influencing the translation process(Figure 3a).

3.4 Small proteins

Previously, it was postulated that the minimum length of functional proteins is around 100 amino acids [6]. Today, multiple studies have reported proteins shorter than 100 amino acids [3, 26], fulfilling roles in different types of regulatory mechanisms [11, 31]. Nonetheless, small open reading frame encoded peptides (SEPs) continue to be underrepresented in existing annotations [9]. With sequence alignment algorithms suffering from low statistical power for shorter sequences, more evidence is needed to differentiate the false from the true positives [19].

The introduction of high-performing machine learning models raises the opportunity to use these methods for the detection of novel proteins. In theory, proposed prediction methods should not suffer from bias against SEPs due to low statistical power, as the model ideally detects translation initiation sites based on relevant biological factors. However, the absence of SEPs in the data used to train and validate novel approaches might introduce biases during training and/or can obscure straight-forward evaluation of the results.

Figure 4 shows the output of the TIS Transformer for several high-profile recently validated SEPs, of which some are still missing from the annotated set of Ensembl. Akimoto et al. [2] prove the existence of three upstream open reading frames (uORFs) for the MKKS gene through proteome analysis, serving as a regulatory mechanism (peptoswitch). uMKKS0, uMKKS1 and uMKKS2 are reported to be 43, 63, and 50 amino acids long, respectively (Figure 4a). Matsumoto et al. [15] could validate the expression of a 90 amino acid sORF on the SPAAR gene (Figure 4b). The micropeptide is shown to be an important factor in regulating biological pathways related to muscle regeneration. Two

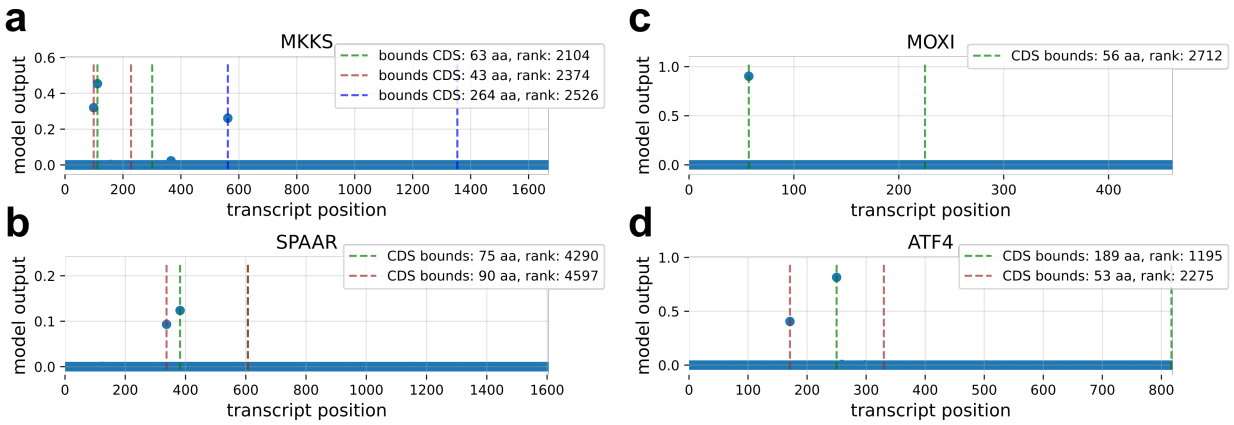

Figure 4: **Example transcripts with predicted micropeptide expression.** Shown are the model outputs (y-axis) for each position of the transcript (x-axis). For high TIS predictions, the bounds of the resulting CDS are given, as well as their length and prediction rank (n-highest prediction) on the chromosome. Given are the predictions for the transcripts (**a**) MKKS (ENST00000652676), (**b**) SPAAR (ENST00000443779), (**c**) MOXI (ENST00000426713), and (**d**) ATF4 (ENST00000676346).

studies reported the existence of a 56 SEP, found to affect mitochondrial respiration [14, 20] (Figure 4c). Young et al. [32] report the existence of an upstream CDS of 53 codons overlapping with the ATF-4 coding region (Figure 4d). Supplementary Table A7 and Figure A7 feature more information and model predictions of recently recovered small ORFs [17].

In contrast to the Ensembl database used for training, the model allows for multiple high ranking TIS per transcript (Figure 4A/C). To illustrate, selecting the top k predictions on each chromosome results in a total of 657 transcripts where multiple TISs are predicted on the human genome. Several additional examples of such transcripts are shown in Supplementary Figure A8.

It was revealed that the transformer model successfully identifies several of the recently reported SEPs. Importantly, a reverse correlation exists between the model output probability scores and the lengths of the resulting proteins (Supplementary Figure A6). As shown in the previous section, the model is capable of outlining the bounds of the resulting CDSs. As such, it is possible that the lack of SEPs in the data results in a negative bias towards the candidate TISs that result in small CDSs.

4 Discussion

Recent advancements of machine learning in processing sequential data, mainly introduced in the field of natural language processing, portend new opportunities towards the creation of predictive models on biological sequence data. In this study, we introduce TIS Transformer, a predictive tool capable of processing information from the full transcript sequence rather than a fixed-length cutout around the position of interest [34, 35]. The use of self-attention makes it possible to determine relevant pieces of information based on the information itself. With respect to performance, attention-based models have shown to be superior when dealing with complex relations in sequential data [24]. In contrast to previous methods requiring a fixed-window approach, the proposed architecture brings improvements in computational efficiency, where the input transcript only needs to be processed once to predict the occurrence of TISs at each position of that transcript. We find TIS Transformer to advance the detection of TISs on processed transcripts by a large margin.

The introduction of the FAVOR+ mechanism, which reduces the complexity of the attention step, is essential in making the computation of attention over full transcript sequences feasible. Evaluation of the decision-making process of the model has corroborated the importance of the full transcript sequence in successfully identifying TISs (Figure 3. Clear patterns were observed that match properties of translation, where both the reading frames and translation stop sites of resulting proteins can have relevance in the decision-making process. The importance of these sites in determining the existence of a TISs is not straightforward. These may as well be related to biases introduced by existing (flawed) data sets, rather than based on relevant biological factors. To illustrate, prior detection of resulting proteins are an important factor when determining TISs. As such, the positive set can present biases against specific TISs, such as those that result in unstable proteins, initiate faulty translation events, or translate into SEPs. It is possible that the importance

of the amino acid composition or the length of the protein is a result of these existing biases. Another example is the absence of non-cognate initiation codons from the training data and predictions, with only 7 (0.006%) instances (all CTG) present in the top k predictions of all chromosomes.

In this study, the positive set incorporates all annotations provided by Ensembl, thereby incorporating a set of molecular systems or transcript types that impede the translation process. Specifically, included are TISs that result in nonsense mediated-decay and transcripts missing validated coding mRNA ends (GENCODE tags CDS_end_NF/mRNA_end_NF). Even though the TISs of these groups are part of the positive set in this study, the inclusion of valid TISs that result in undetectable/invalid protein products is debatable due to the likelihood of missing annotations in each set. Future work might thereby focus on differentiating between types of translation events when curating the positive labels.

While the correlation between transcript support level and model performances is not unexpected (Figure 2), its observation is nonetheless of importance as an indicator of the underlying limitations of current annotations. Biologically, no correlation between transcript support level and the presence of TISs is expected, where differences in performances are expected to be caused by noise in the data, rather than shortcomings of the proposed model. Using additional steps, we substantiate the hypothesis that false positive predictions are at least in part being driven by shortcomings in existing annotations. Ca. 50% of the false positive results were matched with entries of the UniProt database and ribosome profiling data (Supplementary Table A2). Future work may focus on further deconvolution of the causes of false positive and false negative results, with multiple opportunities for further optimisation of both the annotated input data and the predictive model.

Given the success of deep learning in a plethora of fields, its application and adaptation by various platforms has gained considerable traction in recent years. Most notably has been the release of AlphaFold, which has been incorporated in platforms (e.g. UniProt) to serve as the go-to method for protein structure prediction [12]. Similarly, the availability of TIS Transformer can aid the community in various ways, such as offering insights into existing annotations, and driving experimental improvements towards validating candidate proteins. Analysis of the results and decision-making process can further the understanding of the biological drivers of translation. Lastly, the model can be used to impute TISs on novel transcripts, an important step in the design of biological pathways. To facilitate accessibility to our research, we present a re-annotation of the full human transcriptome following a cross validation scheme.

5 Data availability

All discussed models are made available for the community, and can thereby be applied custom transcripts. The input data, unprocessed model outputs, and curated predictions for each chromosome, as used in this study, are available at `https://github.com/jdcla/TIS_transformer`. The listed python repository provides all scripts used to obtain published results, including those to run the benchmarks on listed studies. Provided scripts can be applied to train TIS Transformer on new data and organisms .

6 Funding

The work presented was sponsored by Novo Nordisk Ltd with the work being carried out jointly by Ghent University and Novo Nordisk employees. Novo Nordisk took part in the study design and overall supervision and guidance of the project, with a focus on the evaluation of biological relevance. Ghent University was responsible for the study design, technical development, testing and implementation of the deep learning model.

References

[1] Bronwen L. Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M. J. Searle. The Ensembl gene annotation system. *Database*, 2016(baw093), January 2016.

[2] Chizuru Akimoto, Eiji Sakashita, Katsumi Kasashima, Kenji Kuroiwa, Kaoru Tominaga, Toshiro Hamamoto, and Hitoshi Endo. Translational repression of the McKusick–Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1830(3):2728–2738, March 2013.

[3] Shea J. Andrews and Joseph A. Rothnagel. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, 15(3):193–204, March 2014.

[4] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading. *arXiv:1601.06733 [cs]*, September 2016.

[5] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking Attention with Performers. *arXiv:2009.14794 [cs, stat]*, March 2021.

[6] B. Dujon, D. Alexandraki, B. André, W. Ansorge, V. Baladron, J. P. G. Ballesta, A. Banrevi, P. A. Bolle, M. Bolotin-Fukuhara, P. Bossier, G. Bou, J. Boyer, M. J. Buitrago, G. Cherét, L. Colleaux, B. Dalgnan-Fornier, F. del Rey, C. Dion, H. Domdey, A. Düsterhöft, S. Düsterhus, K.-D. Entian, H. Erfle, P. F. Esteban, H. Feldmann, L. Fernandes, G. M. Fobo, C. Fritz, H. Fukuhara, C. Gabel, L. Gaillon, J. M. Carcia-Cantalejo, J. J. Garcia-Ramirez, M. E. Gent, M. Ghazvini, A. Goffeau, A. Gonzaléz, D. Grothues, P. Guerreiro, J. Hegemann, N. Hewitt, F. Hilger, C. P. Hollenberg, O. Horaitis, K. J. Indge, A. Jacquier, C. M. James, J. C. Jauniaux, A. Jimenez, H. Keuchel, L. Kirchrath, K. Kleine, P. Kötter, P. Legrain, S. Liebl, E. J. Louis, A. Maia e Silva, C. Marck, A.-L. Monnier, D. Möstl, S. Müller, B. Obermaier, S. G. Oliver, C. Pallier, S. Pascolo, F. Pfeiffer, P. Philippsen, R. J. Planta, F. M. Pohl, T. M. Pohl, R. Pöhlmann, D. Portetelle, B. Purnelle, V. Puzos, M. Ramezani Rad, S. W. Rasmussen, M. Remacha, J. L. Revuelta, G.-F. Richard, M. Rieger, C. Rodrigues-Pousada, M. Rose, T. Rupp, M. A. Santos, C. Schwager, C. Sensen, J. Skala, H. Soares, F. Sor, J. Stegemann, H. Tettelin, A. Thierry, M. Tzermia, L. A. Urrestarazu, L. van Dyck, J. C. van Vliet-Reedijk, M. Valens, M. Vandenbo, C. Vilela, S. Vissers, D. von Wettstein, H. Voss, S. Wiemann, G. Xu, J. Zimmermann, M. Haasemann, I. Becker, and H. W. Mewes. Complete DNA sequence of yeast chromosome XI. *Nature*, 369(6479):371–378, June 1994.

[7] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, July 2019.

[8] Alexander P. Fields, Edwin H. Rodriguez, Marko Jovanovic, Noam Stern-Ginossar, Brian J. Haas, Philipp Mertins, Raktima Raychowdhury, Nir Hacohen, Steven A. Carr, Nicholas T. Ingolia, Aviv Regev, and Jonathan S. Weissman. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Molecular Cell*, 60(5):816–827, December 2015.

[9] Martin C. Frith, Alistair R. Forrest, Ehsan Nourbakhsh, Ken C. Pang, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, Timothy L. Bailey, and Sean M. Grimmond. The Abundance of Short Proteins in the Mammalian Proteome. *PLOS Genetics*, 2(4):e52, April 2006.

[10] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*.

[11] Richard A. Jorgensen and Ana Elena Dorantes-Acosta. Conserved Peptide Upstream Open Reading Frames are Associated with Regulatory Genes in Angiosperms. *Frontiers in Plant Science*, 0, 2012.

[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[13] Sooncheol Lee, Botao Liu, Soohyun Lee, Sheng-Xiong Huang, Ben Shen, and Shu-Bing Qian. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 109(37):E2424–E2432, September 2012.

[14] Catherine A. Makarewich, Kedryn K. Baskin, Amir Z. Munir, Svetlana Bezprozvannaya, Gaurav Sharma, Chalermchai Khemtong, Akansha M. Shah, John R. McAnally, Craig R. Malloy, Luke I. Szweda, Rhonda Bassel-Duby, and Eric N. Olson. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. *Cell Reports*, 23(13):3701–3709, June 2018.

[15] Akinobu Matsumoto, Alessandra Pasut, Masaki Matsumoto, Riu Yamashita, Jacqueline Fung, Emanuele Monteleone, Alan Saghatelian, Keiichi I. Nakayama, John G. Clohessy, and Pier Paolo Pandolfi. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, 541(7636):228–232, January 2017.

[16] Hélène Mouilleron, Vivian Delcourt, and Xavier Roucou. Death of a dogma: Eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Research*, 44(1):14–23, January 2016.

[17] Jonathan M. Mudge, Jorge Ruiz-Orera, John R. Prensner, Marie A. Brunet, Jose Manuel Gonzalez, Michele Magrane, Thomas Martinez, Jana Felicitas Schulz, Yucheng T. Yang, M. Mar Albà, Pavel V. Baranov, Ariel

Bazzini, Elspeth Bruford, Maria Jesus Martin, Anne-Ruxandra Carvunis, Jin Chen, Juan Pablo Couso, Paul Flicek, Adam Frankish, Mark Gerstein, Norbert Hubner, Nicholas T. Ingolia, Gerben Menschaert, Uwe Ohler, Xavier Roucou, Alan Saghatelian, Jonathan Weissman, and Sebastiaan van Heesch. A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. *bioRxiv*, page 2021.06.10.447896, June 2021.

[18] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics.

[19] Andrea Pauli, Eivind Valen, and Alexander F. Schier. Identifying (non-)coding RNAs and small peptides: Challenges and opportunities. *BioEssays*, 37(1):103–112, 2015.

[20] Colleen S. Stein, Pooja Jadiya, Xiaoming Zhang, Jared M. McLendon, Gabrielle M. Abouassaly, Nathan H. Witmer, Ethan J. Anderson, John W. Elrod, and Ryan L. Boudreau. Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Reports*, 23(13):3710–3720.e8, June 2018.

[21] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*, June 2017.

[22] Shiyuyun Tang, Alexandre Lomsadze, and Mark Borodovsky. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research*, 43(12):e78–e78, July 2015.

[23] F. Thibaud-Nissen, M. DiCuccio, W. Hlavina, A. Kimchi, P. A. Kitts, T. D. Murphy, K. D. Pruitt, and A. Souvorov. P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *Journal of Animal Science*, 94(suppl_4):184–184, September 2016.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[25] Steven Verbruggen, Elvis Ndah, Wim Van Criekinge, Siegfried Gessulat, Bernhard Kuster, Mathias Wilhelm, Petra Van Damme, and Gerben Menschaert. PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-assisted Proteogenomic Hunt for New Proteoforms*. *Molecular & Cellular Proteomics*, 18(8, Supplement 1):S126–S140, August 2019.

[26] Rui Vitorino, Sofia Guedes, Francisco Amado, Manuel Santos, and Nobuyoshi Akimitsu. The role of micropeptides in biology. *Cellular and Molecular Life Sciences*, 78(7):3285–3298, April 2021.

[27] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-Attention with Linear Complexity. *arXiv:2006.04768 [cs, stat]*, June 2020.

[28] Gavin S. Wilkie, Kirsten S. Dickson, and Nicola K. Gray. Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. *Trends in Biochemical Sciences*, 28(4):182–188, April 2003.

[29] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nystr\"omformer: A Nystr\"om-Based Algorithm for Approximating Self-Attention. *arXiv:2102.03902 [cs]*, March 2021.

[30] Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, January 2016.

[31] Mujie Ye, Jingjing Zhang, Meng Wei, Baihui Liu, and Kuiran Dong. Emerging role of long noncoding RNA-encoded micropeptides in cancer. *Cancer Cell International*, 20(1):506, October 2020.

[32] Sara K. Young and Ronald C. Wek. Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *The Journal of Biological Chemistry*, 291(33):16927–16935, August 2016.

[33] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences.

In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc., 2020.

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

[35] 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, January 2018.