RESEARCH

Multi-task learning uncovers robust translation cis-regulatory features

Weizhong Zheng[1], John H.C. Fong[1], Yuk Kei Wan[1], Athena H.Y. Chu[1], Yuanhua Huang[1,4], Alan S.L. Wong[1,3] and Joshua W.K. Ho[1,2*]

Abstract

Background: Many studies have found that sequence in the 5' untranslated regions (UTRs) impacts the translation rate of an mRNA, but the regulatory grammar that underpins this translation regulation remains elusive. Deep learning methods deployed to analyse massive sequencing datasets offer new solutions to motif discovery. However, existing works focused on extracting sequence motifs in individual datasets, which may not be generalisable to other datasets from the same cell type. We hypothesise that motifs that are genuinely involved in controlling translation rate are the ones that can be extracted from diverse datasets generated by different experimental techniques. In order to reveal more generalised cis-regulatory motifs for RNA translation, we develop a multi-task translation rate predictor, *MTtrans*, to integrate information from multiple datasets.

Results: Compared to single-task models, *MTtrans* reaches a higher prediction accuracy in all the benchmarked datasets generated by various experimental techniques. We show that features learnt in human samples are directly transferable to another dataset in yeast systems, demonstrating its robustness in identifying evolutionarily conserved sequence motifs. Furthermore, our newly generated experimental data corroborated the effect of most of the identified motifs based on *MTtrans* trained using multiple public datasets, further demonstrating the utility of *MTtrans* for discovering generalisable motifs.

Conclusions: *MTtrans* effectively integrates biological insights from diverse experiments and allows robust extraction of translation-associated sequence motifs in 5'UTR.

Keywords: artificial intelligence; multi-task learning; motif discovery; eukaryotic translation; explainable AI

*Correspondence: jwkho@hku.hk
[1]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China
[2]Laboratory of Data Discovery for Health ($D^2$4H) Limited, Hong Kong Science Park, Hong Kong SAR, China
Full list of author information is available at the end of the article

Introduction

For eukaryotic translation systems, the primary point of regulation of translation occurs at the initiation stage [1, 2, 3]. The 5' untranslated regions (5'UTR) encode many important sequence features, collectively shaping the initiation efficiency [4, 5]. While the molecular mechanism of translation regulation has been studied for decades, it remains challenging to predict how well the 5'UTR of a transcript impacts its translation. Discovery of predictive sequence motifs in the 5'UTR, and how the interplay of these elements exert their regulatory effects, remain an important research area in molecular biology.

Several sequencing-based techniques have been developed to assess the translation rate for the individual transcript in a high-throughput fashion. These techniques include Ribosome Profiling [6], Fluorescence-Activated Cell Sorting (FACS), and Massively Parallel Assays (MPA) [7, 8]. Ribosome profiling is broadly used to probe the ribosome protected fragments and derive the translation efficiency (TE) of the natural transcripts in the cells in which the TE of a transcript is shaped by the interplay of the regulators from several regions more than just 5' UTR [9]. By contrast, FACS and MPA generally work by selecting from a pool of cells delivered with

different synthetic 5'UTRs inserted into the target protein which can be directly used to measure translation rate. We can therefore ascribe the changes in translation rate directly to their 5'UTR region. FACS-based sequencing methods usually link to translation rate by inserting the designed 5'UTR into the coding sequence of a fluorescence protein. MPA methods are more flexible in pairing 5'UTR with the translation rate, including using yeast growth rate as the readout and directly using polysome profiling. While the number of transcripts is limited by the genes expressed in the cell, artificial sequences can be massively synthesised, and their sequences are not constrained by natural selection, enabling a much larger sequence space to explore. Pairing the 5'UTR sequence with the translation rate, each technique confers unique information for building predictive models which are finally used to yield and test the regulatory elements [4, 8].

Methods to identify the regulatory code of translation control from these high throughput data evolve quickly. Previous methods required domain knowledge to firstly list out potential elements and tested them with the sequencing data afterward [4, 9, 10, 11]. Recently, it has become a new paradigm in the area of regulatory genomics to use deep learning methods to learn the sequence motifs by the model itself [7, 8, 12, 13, 14, 15]. However, deep neural networks can capture unintended rule that relies on dataset-specific covariates, also called the 'short-cut' [16, 17]. Even though seemingly successful in one dataset, short-cut features may fail to generalise to slightly different circumstances. Existing studies often only extracted and evaluated the elements in a single library or with datasets generated by the same technique. It is therefore challenging to distinguish real translation-related motifs from the false positives caused by dataset-specific artefacts [16]. A key premise of our work is that *robust sequence features* in 5'UTR that predict translation rate across multiple datasets generated by different experimental techniques are more likely to be true regulatory elements for mRNA translation.

In this work, we propose a multi-task learning framework to learn from various datasets and predict the translation rate of a given 5'UTR sequence. Multi-task learning optimises several objectives at the same time and can be regarded as a regularisation technique, for example, the hard-parameter-sharing, to prevent the model from over-committing to a single task, enforcing the generalisability of the learnt features [18, 19]. We here assume that the genuine regulatory elements can be found in multiple datasets, even if they were generated by different experimental techniques. Our hard-parameter-sharing design enables the model to craft features beneficial for all tasks while retaining task specificity in the higher layers. Our model significantly outperforms existing methods in all the tested datasets. Further, our model identifies sequence features that are generalisable to an independent dataset collected from a different species. We performed an analysis to determine what combinations of tasks would lead to the most robust predictive models. As further validation, we conducted an independent FACS screening experiment to validate the identified sequence motifs.

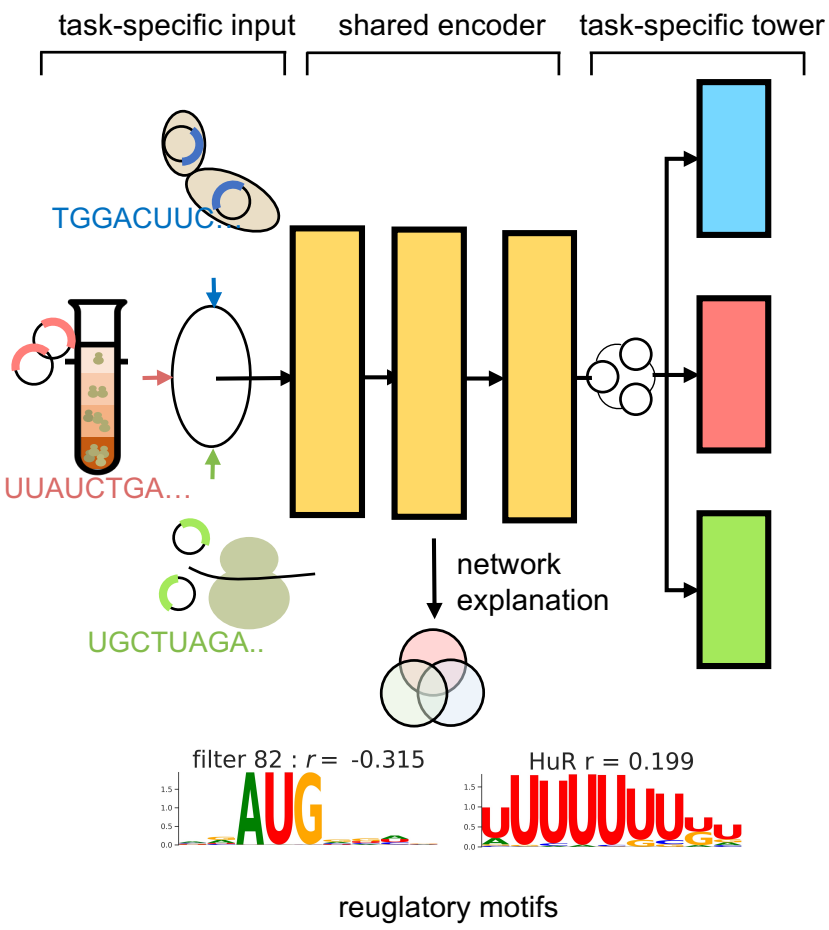

Figure 1 The model architecture of *MTtrans*. The leftmost are datasets from different systems: blue for the designed 5'UTR with yeast massive parallel assay (MPA); red for designed 5'UTR MPA in with polysome profiling; green for ribosome profiling for natural 5'UTR. The same coloring is applied to task-specific tower.

Result

Multi-task model for translation prediction learns the shared patterns from multiple experimental systems

Our multi-task learning model, *MTtrans*, was built to predict the translation rate conferred by the 5'UTR sequences of the transcripts for a wide range of experimental systems. We adapted the canonical hard parameter sharing architecture which consists of a shared encoder (bottom), a task switch and several task-specific upper layers (towers) (Figure 1). To train the model, we collected datasets from different translation systems and of diverse techniques, including three artificially synthesized libraries constructed with MPA combined with polysome profiling in human HEK293T cell line, one synthetic library from MPA combined with yeast growth and three ribosome profiling studies measuring the translation efficiency of the nature transcripts from HEK293T cell line, PC-3 cell line and human muscle tissue respectively.

In general, the shared bottom was designed to extract and assemble local sequence patterns and the towers will learn to reorient the extracted pattern for each task (Figure 1). Basically, the bottom module was mainly made up of four convolution layers. Analogous to the scanning mechanism of the translation initiation, the one-dimensional convolution filters were used to scan through the input sequences from the 5' to 3' end to detect the patterns of interest. One important merit of using convolutional neural networks is the emergence of many network explanation approaches in recent years [15, 20, 21], allowing for the visualisation of the regulatory signal encoded in the higher layers. The tower module is built up with a two-layer Gated Recurrent Unit network (GRU) to deal with the various length of inputs and a dense layer to project the GRU memory to the output value, which is the predicted translation rate.

MTtrans better coordinates MPA tasks and improves translation rate prediction

We started with three MPA-derived libraries to evaluate the effectiveness of our *MTtrans* model. The *single-task* counterparts were considered as the baseline to compare how well the models capture task specificity. Another baseline called *mixing* was trained with a dataset merging the three libraries. The involvement of *mixing* was to account for the benefit of the larger sample size brought by integrating datasets. There is no technical bias between the tasks as they were measured by the same protocol. Instead, other covariates exist like batch effect, length variance and shift in sequence composition between randomly synthetic sequences and natural sequences[22, 23, 24].

Our multi-task method can predict the translation rate more accurately than the alternatives in all three tasks. Although the current methods for the same task MPA-U also have excellent performance ($r^2 = 0.934$ by Sample et al., $r^2 = 0.940$ by our single-task implement)[8], our model can consistently exceed the state-of-art result in all tasks and push translation rate prediction to the highest of 0.943 in r^2 (Figure 2a&b). For model *mixing*, despite the gain over the *single-task* in task MPA-V, it cannot compensate for the performance loss in the other two tasks, implying a task conflict among them. The results also suggest that it can be harmful to mix the datasets without considering their discordance and it is necessary to model the task specificity. Thus, with the task-specific towers, our method alleviates the conflict and improves the prediction in all tasks, turning the mixture of tasks into mutually beneficial.

MTtrans learns more transferable sequence features

We aim to learn genuine biological motifs. A characteristic of a genuine biological motif is that it is more likely to be conserved across species. In this analysis, we test whether features extracted from MTtrans based on the source of human datasets can be transferred to a target dataset from *Saccharomyces cerevisiae*. A common way to assess how transferable the features are is to copy the first few layers of the source model as a feature extractor for the target task [25]. The performance of the target model with the fixed encoder can imply how general are these sequence features. As the model species to study the eukaryotic translation system, *Saccharomyces cerevisiae* shares a number of conserved elements with human beings. To reduce

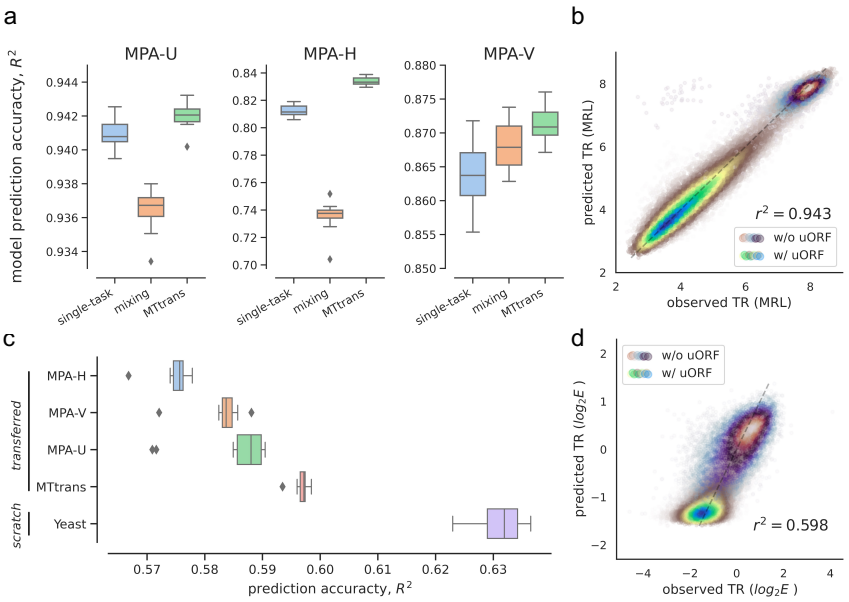

Figure 2 Combining multiple MPA as tasks achieved the highest prediction accuracy and learn more generalisable representation. **a** Performance for predicting mean ribosome loading (MRL) which is the translation rate for each task: MPA-U, MPA-H and MPA-V. The y axis shows the r^2 of regression. The considered baseline is the single-task counterpart shown in blue. The same single-task model trained with the mixed dataset, termed "mixing", is also compared. **b** the scatter plot showing the performance for task MPA-U. A high fidelity of r^2 at 0.943 is reached. Two palettes (yellow-green and blue-red) are used to represent the density of the sequences with uORF and without uORF respectively. **c** Transferring to task MPA-yeast with encoder trained from human MPA datasets. The first three models start with ST are those trained from single-task. Model $3M$ takes the bottom module from multi-MPA model while for $Scratch$, the entire model is trained *de novo* from MPA-yeast. **d** The scatter plot of *MTtrans* prediction to observed translation rate in different conditions of *uORF*. The colour shows the same uORF condition as in b. The readout of translation rate in the yeast system is the growth enrichment of yeast cells, short for E.

the variation brought by experimental techniques, we chose a dataset constructed with a similar MPA system, but the translation rate was measured in yeast cells.

MTtrans has learnt more transferable motifs than the single-task models significantly. As shown in Figure 2c, single-task models have displayed different predictive power, with the task *MPA-H* being the lowest while *MPA-U* being the highest. Instead of learning the average of them, *MTtrans* turned out to adapt to the yeast system better than the three single-task models, suggesting that *MTtrans* has captured more generalisable regulatory patterns. However, *MTtrans* was not on par with the model trained *from scratch* (purple box in Figure 2c), which showed the evolutionary divergence between the two translation systems despite a considerable convergence they shares. For example, in the human 293T cell line, the translation rate of sequences with *uORF* can still range widely, while that of yeasts is restricted in a very narrow range (see Figure 2b and 2d). The species specificity shaped by the evolution process could not be easily overcome by putting more knowledge from another species.

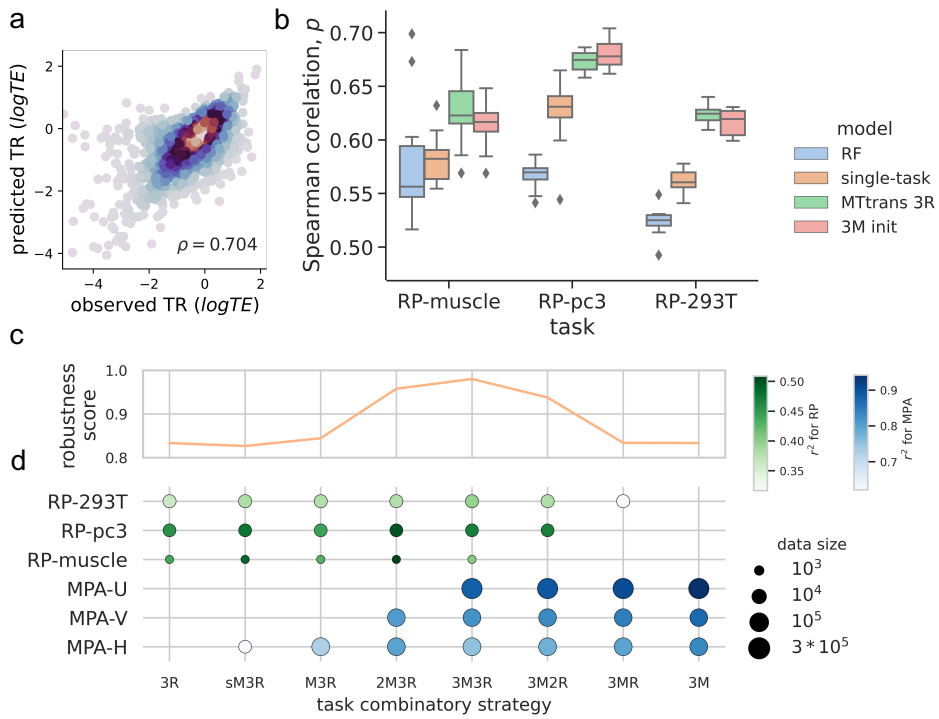

Figure 3 MTtrans supports a variety of task groupings to make effective information sharing and promote the prediction for the noisy ribosome profiling data. **a** Comparison between predicted logarithmic translation Efficiency (TE) and true measurement for cell line PC-3. **b** Model performance in three ribosome profiling datasets. The *MTtrans* here which is actually strategy *3R* in **d**, was trained by the three RP datasets. The red boxes show a*3R* model initialised with *bottom* module from *3M*. **c**-**d** Exploiting various task grouping to optimize *MTtrans* performance and robustness. **c** The upper line plot shows the defined robustness score of the strategy. **d** The dot plot below shows what tasks are combined and the accuracy of each selected task. The x-axis is shared by the two subplots. The strategy defined by each combination is named by the dataset used (see Method). The grouping strategy is displayed by an increasing proportion of MPA tasks over RP from left to right.

MTtrans better predicts the translation rate of endogenous transcripts in human cell lines

We next turned to the genomic 5'UTRs in their actual genomic context, whose translation rate is usually measured with ribosome profiling. It is challenging to quantify the contribution of 5'UTR to the final translation level due to the confounders, including other regulatory regions and post-transcriptional modifications. Even considering all the known *cis*-regulatory elements, Li et al. only explained less than 50% of the variance in human kidney ribosomal profile [9]. The insufficient modelling might be caused by the high technical noise of the data or the incomplete understanding of the translation control. Herewith MTtrans, we attempt to fill the missing features by integrating the noisy ribosome profiling datasets.

To train the *MTtrans* on ribosome profiling data, we selected three sequencing libraries from human HEK293T cell line [26], PC-3 cell line [27] and muscle tissue [28]. With the same datasets, Cao et al. built a random forest (RF) regressor to predict the translation efficiency (TE) of each transcript [11]. They manually crafted features to describe the known regulators in 5'UTR and thousands of kmer frequen-

cies to cover other possible elements. Their RF model and the single-task model were taken as the baseline. In addition to the normal *MTtrans* initialised randomly (*3R*), we design another *MTtrans* initialised with the bottom module of model *3M* to transfer the features learnt from MPA tasks. Highly similar transcripts of the same gene were filtered for all the datasets to avoid information leakage from the training set to the evaluation set (see Methods).

MTtrans greatly surpasses other methods in the noisy RP data (Figure 3a&b). Deep learning-based algorithms seem to be more suitable in modelling RP when we compare *single-task* with *RF*, which may be explained by the learnt regulatory features not included in the RF model. Notably, a consistent performance gain was conferred by *MTtrans* in all cell lines. The improvement was not only observed in small source tasks like *muscle* but also held true for tasks of larger size (*i.e. PC-3*, Figure 3a). However, the model initialised by *3M*, has failed to offer extra advantages over random initiation. Moreover, the models inheriting the bottom module of *3M* will collapse (Supplementary Figure 1), which evidences the discrepancy between MPA tasks and RP tasks in terms of the features they require.

Optimised task grouping enables *MTtrans* to learn robust features

So far *MTtrans* has shown its modelling advantage in integrating data of the same techniques. While features derived from that might enjoy less batch effect error, patterns spanning technical barriers are particularly more interesting and can only be learnt by modelling data of different sequencing techniques. To learn a set of general features, various task grouping strategies were designed with a different mixture of MPA tasks and RP tasks. To order the combination, we defined a gradient with *3R* and *3M* at the two ends, between which are mixture at different levels (Figure 3d). Notably, a robustness score, with the most robust model scoring 1, is defined here by considering the accuracy relative to single-technique strategy (*3R* and *3M*) and the number of tasks supported (see Methods).

Generally, task performance positively correlates with the proportion of itself or highly related tasks in the training data, but an appropriate task selection can overcome the data imbalance. For the encoder with limited model capacity, the hidden features were updated in a competitive way that only those universal features can be kept during the training. With more MPA tasks joined, MPA tasks enjoyed a stable increase in the prediction accuracy, which climbed to the peak at *3M*. Interestingly, RP tasks could also benefit from the integrative training with MPA tasks by referring to more refined sequence features identified from the large sequence space (see RP tasks in model *M3R* and *2M3R*, supplementary figure 2). The best performance for task RP-293T was achieved by strategy *3M3R*. Both *2M3R* and *3M2R* scored highly in robustness score (0.95 and 0.93, respectively), suggesting the benefit of integrating the tasks of different sources.

Thus, training the multi-task model required a careful balance of their covariates, such as the sequencing techniques and cell type. Taken together, model *3M3R* stood out from all other tested combinations (Figure 3c) because it retained a comparable performance on diverse tasks regarding the technique and cell type, maximising the chance to learn more general motifs.

Deeper features are more predictive

An emerging approach in explainable deep learning models is to interrogate the convolution filters from a discriminative model, which has been applied to various biological questions, for example, the sequence motifs of the TF-binding and CpG methylation sites [7]. Many studies only interpret the parameters of the first convolution layer [12, 13], but one could argue that features extracted from higher layers confer more abstraction and are more likely to represent the regulatory grammar [7, 15]. Before delving into each convolution filter, we need to figure out which layer in the encoder compresses the most necessary complexity. Taking the output from each layer as an individual feature set, four random forest models were built to predict the translation rate directly (Figure 4a). The random forest models were applied here to evaluate the feature set because it does not generate any hidden feature.

Except for task *RP-muscle*, features coming from a higher layer can better describe the regulation in UTR sequences. The turning point for MPA tasks happened at layer 2 but layer 4 for RP tasks (Figure 4c). Despite the performance gain by using deeper features, the increment can not be simply attributed to the depth because the receptive field of the convolution neurons also increases by layers. By restructuring the encoder to only one layer but with different kernel sizes, we obtained shallow features with aligned receptive fields as the features from each layer (Figure 4c&d). A larger kernel size did not provide extra merit for shallow features, making it cryptic. Comparing the trend of deep and shallow features, we could see that depth is critical in shaping the sequence pattern regulating the translation.

Random forest is an explainable algorithm that can quickly evaluate the feature importance. Here the top 20 important features from layer 2 and layer 3 are shown (Figure 4b). The receptive field of the filter controls the length of motifs, so features are converted to 5-mer and 9-mer by the max activation patch (see Method). The scaled importance score for each task will sum up to rank the features. Features detecting AUG are emphasized in both layers and all the tasks. Especially, multiple highlighted features caught the position and flanking nucleotide around the AUG, which will refine the sequence context of the upstream start codon. There are rich biological insights delivered by the layer 2 features already, which also significantly influence the translation rate (Supplementary Figure 3). Among the top important layer 2 motifs, 'UUUCC' and 'GCGGC' are found to impact protein-RNA interaction[29], and 'UGGGU' impacts the secondary structure of the RNA [30, 31], a known regulatory factor of the translation process.

Discovery of 5'UTR sequence motifs that predict translation rate

We next dissected the *MTtrans* model to see what features it has learned. Trained as a regressor, the encoder of *MTtrans* will naturally adopt the discriminative features to quantify the translation rate accurately. We can visualise these features into sequence motifs by explaining the convolutional filters. Since model *3M3R* is compatible with the most translation measurement tasks, the motifs explained from it may be more versatile.

To explore the sequence motifs, we use the maximum activation patch method to visualise individual convolutional filters at layer 3. As discussed earlier, the features

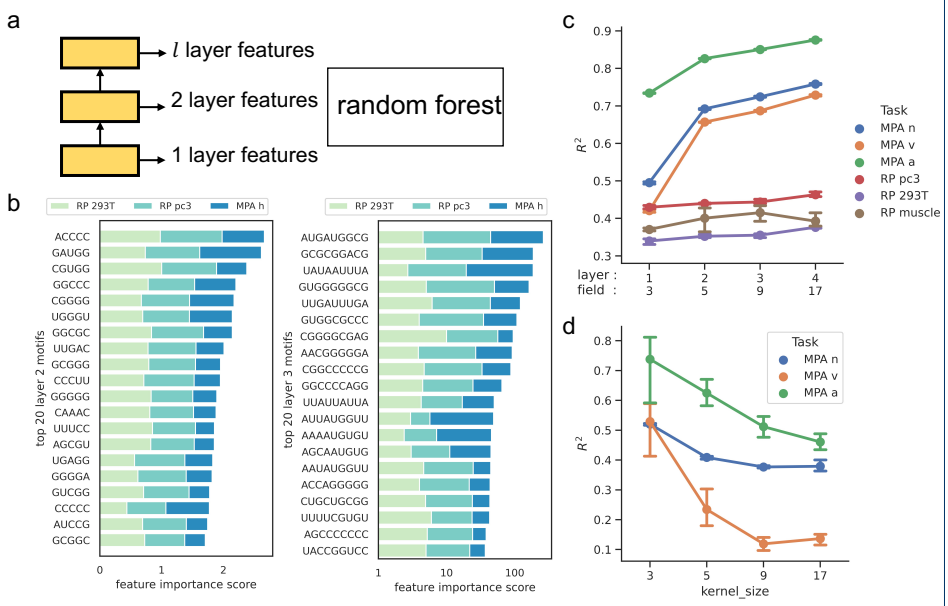

Figure 4 Interrogation of features from each layer.**a** The feature map encoded by each convolution layer of *MTtrans* was redirected as the input for random forest models, whose prediction accuracy indicates how informative each feature sets are. **b** The top 20 most important features for layer 2 and layer 3 features encoded by *3M3R*. The feature importance scores for the RF model trained for task RP-293T, RP-PC3 and MPA-H were sorted by their sum. **c** The performance (in R^2) of random forest model using *MTtrans* features. The x-axis displayed which layer the features came from and the receptive field. The colour of the line denotes the dataset used. **d** The performance of 1-layer-only convolutional networks. The kernel size of the filter pairs to the receptive field of each layer in the original model architecture.

encoded from at least layer 2 are worth explaining (Figure 4c), and we choose layer 3 here for a longer receptive field to cover the $5\tilde{7}$bp long annotated binding sites of RNA-binding Proteins (RBP). The sequence windows that can activate the neuron the most is clipped out from the sequence and grouped to generate the position weight matrix. With this method, each convolutional filter will be interpreted as a sequence motif with a receptive field of nine nucleotides. The correlation between neural activities and the translation rate can indicate the effect of the motif it derives.

Refined by six datasets, the sequence motifs from model *3M3R* are supposed to cover two cell lines as well as two sequencing techniques. 28 motifs are significantly similar to the RBP binding motifs annotated by RNAcompete [32], indicating that *MTtrans* captured the biological meaningful elements to some degree(Figure 5). For example, HuR have been proved to bind to 5'UTR; PABR4 has high poly-U affinity[33]; RMB4 is known to suppress the cap-dependent initiation process [34].

In Figure 5, by construction, motifs having similar position weight matrices are closer in the UMAP space. The motifs are roughly separated into two groups according to their predicted effect on translation rate: the positive ones at the upper region of the plot, and negative ones at the lower half. The upper right region (called the positive end in the rest of the article) is condensed with U-enriched motifs and annotated motifs. However, the symmetric A-enriched cluster is not observed, indicating that *MTtrans* has captured patterns that are more sophisticated than the

AU ratio. It holds true for the GC-enriched motifs which are mostly encoded in the lower-left region (called the negative end). Highly GC enriched motifs like filter 236 and the binding site of RBM4 repress translation, but filter 197 with A and U shows a stronger repressive effect. More directly, the GC ratio has been included as the feature by Cao et al [11]. The performance gap between their RF model and our RF with *MTtrans* features also implies the need to refine the sequence signature.

The upstream start codon (uAUG) is a well-characterised regulatory signal. The flanking sequences around the start codon formulate the Kozak consensus sequence which strongly influences the translation initiation[1, 2]. There are a substantial number of the discovered motifs contain AUG. In particular, filter 105 reconstructs the preference of purines (adenine or guanine) 3 nt ahead of the uAUG. As previously demonstrated, uAUG in 5'UTR is generally repressive, matching our finding that 56 out of 67 uAUG motifs negatively correlate to the translation rate. Most of them have guanine attached to the uAUG, which is highly conserved among species [35]. Comparatively, the positive uAUG motifs are more interesting as they define the sequence context to cancel the upstream initiation. An apparent pattern is a Uracil or Adenine flanking around the uAUG, which is opposite to the preference of the Kozak consensus (Supplementary figure 4). Lastly, the upstream stop codon in 5'UTRs can close the uORF and thus rescue the translation of the main ORF. The filter 165 highlighted the stop codon UGA and was assigned with a positive effect, matching our known biological mechanism.

Thus, by interpreting the learnt shared encoder, we can rediscover many genuine regulatory signals, suggesting the advantage of *MTtrans* as a motif discovery tool.

The discovered regulatory motifs can be experimentally validated

As we hypothesized that the motifs learnt by *3M3R* are less technique-specific, we applied a different sequencing technique, FACS, to test the consistency of the motif under a new condition. To validate the effect of the discovered motifs, mostly *de novo*, we constructed a FACS library with around 8,000 sequences consisting of designed 5'UTRs and a fixed CDS region coding the GFP. The engineered plasmids were delivered into the HE293T cell line which is coherent with two of the dataset which *MTtrans* was trained from. The cells were sorted to select cells with high GFP expression for further sequencing, so that the read count can indicate the translation altered by the 5'UTR sequence (Figure 6a). In the end, there are a total of 5515 sequences binned and reaching the desired read depth.

MTtrans then encoded the UTR sequences of the FACS library to identify which filters they could maximally activate. For sequences that activate a motif uniquely, their read count distribution will suggest the effect of the motif. Among the 123 inferred positive motifs, 72 have a higher reads count than the average in the FACS library. For negative motifs, 83 out of the 141 motifs show lower reads count averagely (Figure 6b). These generalisable motifs have taken up around 60% of that discovered by *MTtrans* , demonstrating the validity of *MTtrans* as a motif discovery tool.

We explore the generalisable motif in the same reduced space. The dots of the same regulatory effect tend to locate closely in the space. Especially, the positive group and the negative group are largely positive and negative respectively (Figure 6c),

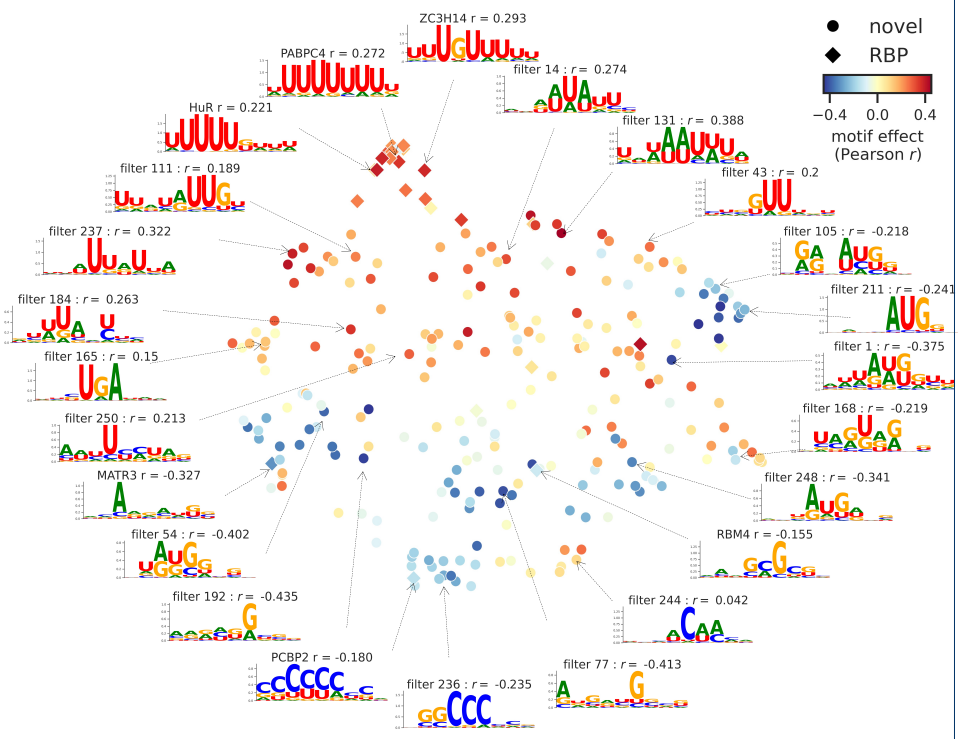

Figure 5 Sequence motifs discovered by MTtrans. A total of 256 motifs are projected into UMAP according to their position weight matrix. Colour denotes the regulatory effect estimated in dataset MPA-H, which is the Pearson correlation between motifs activity and translation rate. *Diamonds* indicate motifs that are significantly similar to known RBP in [32].

showing a high coherence in sequence preference and regulatory effect. Therefore, the U-enriched and GC-enriched motifs are widely applicable signals in controlling the translation process.

In Figure 6d, we showed the most positive motifs that have been validated as well as the negative ones. The read count distribution of the uniquely-activating sequences shows a high variance, probably because of the different sequence contexts the motifs locate in. The median of the read count can represent the overall effect of the motifs by comparing it to the average of the whole library. The same positive motifs could also be yielded by the motif enrichment analysis seeing their ratio in the top 25% sequences over the rest (Supplementary Figure 5). In summary, the highlighted positive motifs are worth studying and are likely the genuine *cis*-regulating elements in terms of their consistency across analyzing methods and experimental systems.

Thus, in the independent FACS library, the sequence motifs discovered by model *3M3R* can largely maintain the same regulatory direction as estimated. The consistency of the learned features regardless of the sequencing techniques manifested the advantage of integrating multiple datasets to learn the regulatory sequences motifs.

Discussion

In this study, we demonstrate that a multi-task learning framework, which we named *MTtrans*, is effective to predict mRNA translation rate using 5'UTR sequences

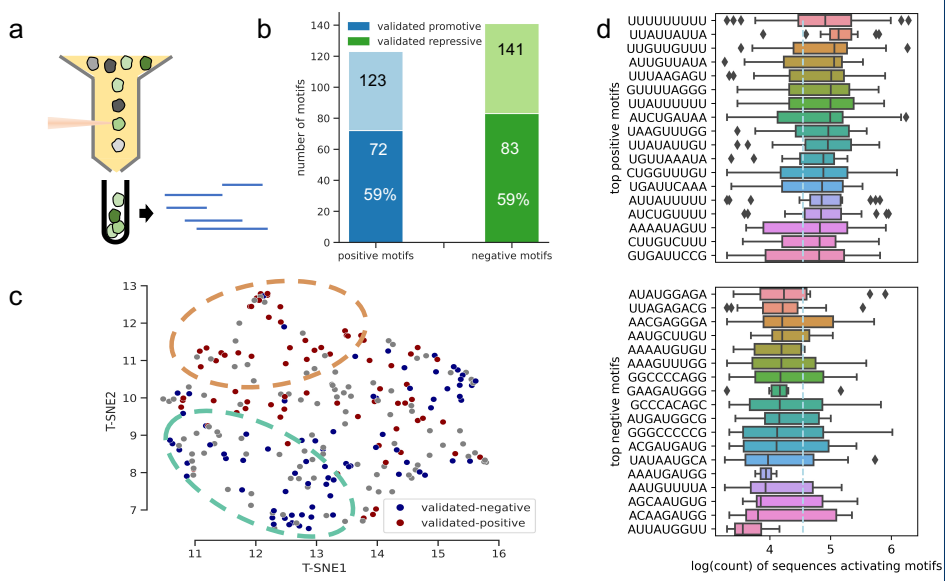

Figure 6 FACS experiment to validate the discovered motifs. **a** cells carrying designed UTR are first sorted by the expressed GFP fluorescence and then sequenced. **b** Out of 123 positive motifs (light blue), 72 of them (darker blue) are identified by our independent FACS experiment (59%). Out of 141 negative motifs (light green), 83 of them (darker green) are also identified (59%). **c** The distribution of validated motif in the same UMAP space as Fig5. *Reds* are motifs showing positive effect in the FACS library while the *blues* are negative. Dots lines highlight 2 clusters of the motifs. **d** Displays are the top positive and negative motifs validated in the FACS library. The UTRs sequences that activate individual motifs are grouped to show their effect on translation.

using diverse experimental data. *MTtrans* works by treating each dataset as an individual task so that useful information across tasks can be underlined repeatedly. Importantly, we found that this framework enables the extraction of highly robust sequence motifs that predict the increase or decrease of translation rate. This is achieved by the hard-sharing architecture in our multi-task model.

Previous studies on transcription factor motif discovery usually can extract predictive features from the first convolutional layer already [12, 13]. Koo et al. even propose an ambiguous pooling to enforce the motifs detection finished in the first layer [14]. Interestingly, in the context of translation rate prediction based on 5'UTR, it is necessary to look beyond the first layer to extract features. Furthermore, we found that the improvement does not result from the longer receptive field (Figure 4 c d). As the deeper features are assembled from the lower layer patterns for CNNs, translation rate prediction may thereby require better arranging the basal motifs to sense the sequence context. This might partially explain how the CNN model surpasses the k-mer-linear model [8] because k-mer models lose the sequential relationship of the features despite a broader k-mer coverage. Future work should explore how to best extract features from different layers of neural networks for the prediction of transcription factor binding and translation rate.

A core premise of our work is that more robust and transferable features can be extracted from a model learned from diverse data sets generated from different experimental techniques and cell types. To evaluate this hypothesis, we tested the use of different combinations of data as the tasks in *MTtrans* and found, indeed,

the most cross-platform model is the one with a balanced task source (*3M3R* in Figure 3d). Models trained with the data generated by the same technique, such as MPA, obtained the specificity for the same setup (Figure 2a), but fail to predict results generated by another experimental technique (Supplementary Figure 1). This suggests that such models are learning the short-cut to some degree, more explicitly, the experimental-specific artefact that may be not biologically relevant to translation. Therefore, our *MTtrans* framework is indeed important to allow multiple data types to be integrated to identify robust features that can work well across different conditions.

Although we have yielded robust regulatory motifs with maximally activated patches (Figure 5), there is still rich information in the model that is not fully explained, such as the motif interaction. Recent studies analyse the motif interaction in an instance-based way. Ziga et al. identified the cooperative TF binding interaction by permuting the distance of two candidate motifs [36]. For *MTtrans*, the GRU layers in the task-specific tower take in the spatial array of the motif activation from the last convolutional filters and model their long-term interaction. Visualisation of the GRU layer, however, usually requires an extra architectural plug-in like attention [37]. Although the task-specific syntax explanation of the GRU layers is not the focus of this work, future work can reveal how the essential elements are coordinated to shape the translation rate.

Overall, *MTtrans* provides a solution to extract the robust translation regulatory elements in the 5'UTR from a collection of related yet noisy systems. We expect this multi-task framework can extend to learn the technique-invariant determinant for other sequence modelling questions to distil the key signal.

Methods

MTtrans model

Our multi-task learning model *MTtrans* applies the hard parameter sharing to encode the information from several task-specific inputs. The model consists of a shared encoder f_e to embed sequences into a shared space and the task-specific towers f_w^t to capture the variance of the translation rate for the t-th task. Suppose a total of T tasks which are sequence libraries from any techniques, are integrated to train the model and the data for task t is denoted as $\{x_i^t, y_i^t\}$ with paired translation rate $y_i^t \in \mathbb{R}, t = \{1, ..., T\}, i = \{1, ..., N^t\}$.

The shared encoder

We stacked four 1d-convolution filters layers as the main building block for the shared encoder. Raw sequences will first be one-hot encoded, ending up a 4 dimensional input x_i^t for *conv1*. Noted we don't include any pooling operation in the encoder following the convolution layer for a better sense of the location. Instead, Batch Normalisation (BN) and drop-out operation is performed after each convolution layer to stabilise the activation and have a more robust model performance. The hidden unit at layer l is computed by $h_i^{t,l} = \mathrm{ReLU}(W^l h x_i^{t,l-1} + b^l)$ and the batch-normalized activation by $a_i^{t,l} = \mathrm{BN}(h_i^{t,l})$.

Table 1 hyper parameters of the shared encoder

	out channel	padding	stride	kernel size
conv1	128	1	1	3
conv2	256	1	2	3
conv3	256	1	2	3
conv4	256	1	2	3

The task specific tower

The number of towers corresponds to the number of tasks selected. The t-th tower f_w^t consists of 2 layer Gated Recurrent Units and an output dense layer. The hidden size of the GRU is set to 80 and the hidden state of the last time point is taken to connect to the output layer. Taken together, the translation rate is predicted by $\hat{y}_i^t = f_w^t(a_i^{t,L}) = f_w^t(f_e(x_i^t))$ wherein $a_i^{t,L}$ is the activation from the last convolution layer.

Model training

One of the biggest differences in training our multi-task model is that only the t-th towers will be updated at one time. Technically, there is a task switch activating the corresponding tower for each mini-batch sampled from $\mathcal{X}^t$, while the shared encoder receives gradient for all the tasks. Each instance in $\mathcal{X}^t$ is a pair of sequences x_i^t and the translation rate y_i^t, which is defined by the mean ribosome loading for MPA tasks [8] and the translation efficiency for RP tasks [26, 27, 28]. The cost is quantified by the naive mean square error weighted by λ^t for $t \in \{1, .., T\}$:

$$\mathcal{L}^t = \lambda^t \sum_{i \in B} [y_i^t - f_w^t(f_e(x_i^t))]^2$$

A learning rate scheduler is applied to wrap the *Adam* optimizer so f_w^t and f_e will be updated in a rate that is boosted in the beginning and then goes through a dramatic decay. The small learning rate in the late training phase restricts the parameters in a narrow range. Therefore, each f_w^t docks to a similar θ_{f_e} and stabilise the models. The learning rate ϵ is mediated by the largest dimension of the model d, two constants τ_1, τ_2 and a step-renewing variable δ, following by $\delta = \delta + \min(\tau_2, 2\delta)$.

$$\theta_{f_w^t}^* = \theta_{f_w^t} - \epsilon \nabla_{\theta_{f_w^t}} \mathcal{L}^t$$

$$\theta_{f_e}^* = \theta_{f_e} - \epsilon \sum_{t \in T} \nabla_{\theta_{f_e}} \mathcal{L}^t$$

$$\epsilon = d^{-\frac{1}{2}} \times \min(\delta^{-\frac{1}{2}}, \delta * (\tau_1^{-\frac{3}{2}}))$$

Model evaluation

All the datasets used in our study were split into train, validation and test set. For all the MPA tasks, the same train-test splitting was kept as in Sample et al [8] so that the model performance was directly comparable. The remaining dataset was further sectioned to train and validation set in a 10-fold cross-validation manner. The training process was terminated when the validation loss converged. In the

test set, the prediction accuracy R^2 of each fold was computed from the predicted values $\hat{y^t}$ and the true translation rate y^t. The final R^2 was the average of all 10 fold. For the RP tasks, 10-fold cross-validation splitting was firstly performed to get the train-val set and the test set, the former of which was further split to train and validation set in a ratio of 9:1. Similarly, the prediction accuracy was calculated by averaging the 10 fold but in a measurement of Spearman correlation coefficient.

Data pre-processing

We collected a total of 7 public translation profiling datasets to evaluate MTtrans, including 3 massively parallel assays by polysome profiling, 1 massively parallel assay by yeast growth and 3 ribosome profiling for 3 cell lines.

The massively parallel assays used are GSM3130435, GSM3130443 and GSM4084997 from the Gene Expression Omnibus (GEO). They are termed MPA-U, MPA-H and MPA-V in the text for brief. Their pre-processing are described in [8]. The Mean Ribosome Loading (MRL) is defined as the translation rate for these three libraries.

A similar technique performed on the yeast uses the growth assay to probe the translation level. Sample GSM2793752 under the accession GSE104252 is used in this study with the pre-processing described in [7] and is termed MPA-yeast. The growth rate is taken to represent the translation rate.

For ribosome profiling data, three datasets collected from HKE293T cell line [26], PC-3 cell line [27] and muscle tissue [28] were selected because they are the benchmark datasets on which the random forest model with the handcraft features [11] was built. The ribosome profiling data were first pre-processed as in [11]. Later, to prevent data leakage, we filter the isoforms with the same sequence 100nt upstream of the start codon and keep the most abundant isoforms. The translation rate is defined as the logarithmic translation efficiency (TE), which is the division of ribosome protected fragments (RPF) reads count from the ribosome profiling to the reads count of transcripts from a paired RNA-seq sample.

For FACS library, the total plasmid pool was used to constructed the reference. The consensus sequences were taken as the reads to map back to the reference file to quantify their count using *Bowtie2* [38].

Maximally Activating Patches

The hidden features formulated within the convolution neuron represent the local sequence combinations that are useful for predicting the translation rate. By explaining the model, we can therefore reveal the sequence motif those convolution blocks are detecting. Here we use Maximum Activation Patch to prioritize a number of sequence regions activating the neuron. We first selected a set of sequence windows P_c, that show the highest activation values for convolutional filter c.

$$P_c = \underset{\{x^t_{i,j} \in \mathcal{X}^t\}}{\arg\min}\ a_i^{t,l}$$

Piling up the short sequences (patches) in P_c, we can calculate the probability of nucleotide at each position, which is the Position Weight Matrices(PWMs). From these, the convolutional filters can be converted to a sequence logo with python package *logomaker version 0.8*.

Task grouping search

The performance of the multi-task model is influenced by model capacity, dataset size and the relationship among tasks [19, 39]. We thereby set the channel size of the encoder to be the same for each strategy regardless of the number of tasks grouped. The dataset size of MPA related tasks is generally higher than the RP tasks, so each MPA task is down-sampled to represent a comparative data size with the RP tasks. Model *3R* and *3M* are two extreme strategies tasked with datasets from the same sequencing techniques. To search for an optimal task grouping strategy with diverse tasks, we empirically tested the strategy in the following order:

1. On top of *3R*, we first introduced a sub-sampled MPA-H dataset to have a light touch on the massively parallel assay, making a new combination *sM3R*
2. Based on the previous result, we replaced the sub-sampled MPA-H with the full-sized MPA-H to make a new strategy *M3R*, increasing the proportion of MPA-tasks.
3. The strategy *2M3R* consisted of two MPA tasks, with MPA-V added in, and all three RP tasks. The involvement of MPA-V here alleviates the technical imbalance on RP side.
4. Adding the MPA-U dataset, model *3M3R* integrated all the datasets so the number of towers for the RP side and MPA side is now equal. The direction of updating for the encoder is no more biased to the RP side.
5. We then removed the smallest task RP-muscle, to make up the new strategy *3M2R*. The task infinity of RP-muscle is also the smallest of the other five.
6. By removing the task RP-PC3, the strategy *3MR* totally consists of datasets performed in the HEK-293T cell line.

Evaluate the combination strategies

The models trained by different task grouping strategies is also evaluated in the test set with respect to the selected tasks. The same prediction accuracy r^2 is used. Considered the difference of range in r^2 between RP tasks and MPA tasks. Two colorbars were used to indicated their performance (green for RP tasks and blue for MPA tasks), both in Figure 3d and Supplementary Figure 2.

We also defined the robustness score to give a numerical summary of the strategy. Here we denote σ^t as the performance in r^2 for tasks t. The score sums up the mean relative performance of RP task over model *3R*, the average relative performance of MPA tasks over model *3M* and the number of tasks supported $n_{supported}$ over the total number tasks n_{total}, which is 6 in this scenario.

$$\text{robustness score} = \mathbb{E}_{t \in RP} \frac{\sigma^t}{\sigma^t_{3R}} + \mathbb{E}_{t \in MPA} \frac{\sigma^t}{\sigma^t_{3M}} + \frac{n_{supported}}{n_{\text{total}}}$$

Motif similarity comparison

Important translation regulatory elements can be captured independently by models trained with different datasets. In order to identify which motifs are significantly similar to annotated RBP binding motifs, motifs similarity analysis was performed using *tomtom* from the tool-kit *meme-suite version 5.4.1*. The query motifs are explained from the models in the way we described and target motifs are labelled by RNAcompete[32] which is a built-in database in *meme-suite version 5.4.1*. Only

motif pairs satisfying $FDR \leq 0.05$ and $E - value \leq 10$ at the same time will be noted as significantly similar.

Validate discovered motifs with the FACS library

Our core assumption is that collectively training the model with datasets from diverse techniques can lead to a more biologically relevant feature set. The FACS sequencing method was used here as a complement of the RP and MPA to test the robustness of the sequences motifs derived from *MTtrans*. When sorting the cells carrying 5'UTRs with high translation efficiency, their reads count y in the sorted bin becomes a proxy of the translation efficiency.

We firstly encoded the 5'UTRs by *MTtrans* (the exact model was *3M3R*) to extract the feature map of layer 3. Then we searched the sequences $\mathcal{S}_c$ that can uniquely activate the motifs. For each convolutional filter, the 5'UTR sequences were then ranked by the activated values in the feature map. The top-ranked sequences were regarded as the activated sequence for the corresponding filter. Each sequence can only be assigned to one filter. When one sequence ranks top for multiple filters, the filters with the highest activation values will be assigned to it, and it will be added to the uniquely activating set $\mathcal{S}_c$, where c denotes the convolution filter (channel). The effect of a motif converted from filter c is therefore defined as $\varrho_c = \mathbb{E}_{\{i \in \mathcal{S}_c\}} y_i - \mu$, where μ is the average read count of the library $\mu = \mathbb{E}_{\{i \in \mathcal{X}^{\text{FACS}}\}} y_i$. In other words, if the reads count of the activated set $\mathcal{S}_c$ is higher than average, we have a positive ϱ_c, indicating c is a positive motif and *vice versa*. Lastly, we compared the estimated effect with the effect they displayed in the FACS library. If their direction is the same, the motif effect is consistent in the new library.

Design and construction of 5' UTR library

Ribo-Seq and RNA-Seq data on HEK293T and hESC cells were extracted from RPFdb v2.0 [40] and Gene Expression Omnibus [41] respectively. Reads Per Kilobase of transcript per Million mapped reads (RPKM) was used as the primary input. Rare genes with RPKM less than 1 were eliminated. The translation efficiency (TE) score was defined as the ratio of RPKM of the ribosome protected fragments (RPFs) by ribosome profiling over the average RPKM of the transcripts by RNA-seq. Average RNA-seq RPKM was calculated from 3 HEK293T and 4 hESC samples respectively. Natural 5' UTR sequences were defined from -100 to -1 with respect to the start codon of the coding sequence on mRNA transcripts documented in the Reference sequence (RefSeq) database [42]. The -100 to -9 sequences were extracted and attached CTAGCCACC containing a NheI overhang and Kozak sequence for cloning purposes. Synthetic sequences were generated by dividing the -100 to -9 sequences into 7mer or 31mer windows and scrambling them, followed by CTAGCCACC attachment. The natural sequences were ranked according to TE scores. Top and bottom 5% of the population in each cell line were selected and 4,800 of them were placed in the library pool, with 1,032 HEK293T specific, 1,532 hESC specific and 2,236 double specific 5' UTRs. 3,200 synthetic sequences were also selected for the library with 1,400 generated from HEK293T specific, 1,400 generated from hESC specific and 400 generated from double specific 5' UTRs. Together, an 8,000 library was designed. The library was then generated by the GenScript Precise Synthetic Oligo Pools service.

To insert the 5' UTR library into an expression plasmid, the 5' UTR sequences were firstly amplified from the Oligo Pool by Phusion DNA polymerase (New England Biolabs). The resulting fragments were digested by SbfI and BbsI (New England Biolabs) and ligated into a SbfI and NheI (New England Biolabs) digested lentiviral GFP expression plasmid using T4 DNA ligase (New England Biolabs) overnight at 16℃. The ligation products were transformed into chemical competent E. coli strain DH5α and selected by 50g ml1 carbenicillin. The resulting library was extracted and purified using Plasmid Midi (Qiagen) kits.

Cell culture and cell sorting

HEK293T cells were obtained from the American Type Culture Collection (ATCC) and cultured in DMEM supplemented with 10% FBS and 1× antibiotic–antimycotic (Life Technologies) at 37 °C with 5% CO_2.

Lentivirus production of the 5' UTR library was similar as previously described [43]. The resulting lentivirus was titrated to a multiplicity of infection of 0.15 to achieve low copy number delivery of the 5 'UTR library into HEK293T cells while achieving 100 fold representation.

On day 7 post-transduction, cell sorting was performed on a BD Influx cell sorter (BD Biosciences). Live single cells were gated and sorted by GFP signals. A two-round sorting approach was used by enriching the total GFP population and further sorting them into three bins according to their GFP intensity.

Deep sequencing

Genomic DNA of collected samples were extracted by DNeasy Blood and Tissue kit (Qiagen). The inserted 5 'UTR region was amplified by PCR using Kapa HiFi Hotstart Ready-mix (Kapa Biosystems). The fragments were gel purified by QIAquick Gel Extraction Kit (Qiagen) and sent to HKU LKS Faculty of Medicine Centre of PanorOmic Science for SMRTbell Library construction and Single Molecule, Real-Time (SMRT) sequencing (PacBio).

Appendix

There is one additional file containing Supplementary Figures 1-5.

Acknowledgements
We thank Dr. Sung Chul Kwon for his helpful suggestions about filtering the transcripts and Dr. Chen Qiao for help with model training. We also thank our colleague Xinyi Lin and Yiming Chao for their feedback on the data visualisation and testing the pipeline.

Funding
This work was supported in part by AIRInnoHK administered by Innovation and Technology Commission and the National Natural Science Foundation of China Excellent Young Scientists Fund (32022089).

Abbreviations
RP : Ribosome profiling; MPA : Massively Parallel reporter Assay; FACS : Fluorescence-Activated Cell Sorting; uAUG : upstream start codon; uORF : upstream Open Reading Frame; CNN : Convolutional Neural Network; GRU : Gated Recurrent Unit;

Availability of data and materials
The code to re-implement *MTtrans* can be access from https://github.com/holab-hku/MTtrans and the FACS library is also available are available from Gene Expression Omnibus (GEO) under the accession of GSE201766.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
All the authors have approved the final version of the manuscript.

Authors' contributions
A.S.L.W. and J.W.K.H. conceived the ideas of the study. W.Z. developed the method, implemented the software, and performed all the experiments under the supervision of Y.H., A.S.L.W. and J.W.K.H.. Y.K.W. performed some initial analyses, and contributed to the initial design of the sequence library of the FACS experiment. J.H.C.F. performed the FACS experiment. A.H.Y.C. contributed to data preprocessing. W.Z. and J.W.K.H. wrote the initial version of the manuscript. All authors read and approved the final manuscript.

Authors' information
Correspondence and requests for materials should be addressed to jwkho@hku.hk.

Author details
[1]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China. [2]Laboratory of Data Discovery for Health ($D^2$4H) Limited, Hong Kong Science Park, Hong Kong SAR, China. [3]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China. [4]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China.

References
1. Jackson, R.J., Hellen, C.U., Pestova, T.V.: The mechanism of eukaryotic translation initiation and principles of its regulation. Nature reviews Molecular cell biology **11**(2), 113–127 (2010)
2. Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., Bartel, D.P.: Improved ribosome-footprint and mrna measurements provide insights into dynamics and regulation of yeast translation. Cell reports **14**(7), 1787–1799 (2016)
3. Riba, A., Di Nanni, N., Mittal, N., Arhné, E., Schmidt, A., Zavolan, M.: Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. Proceedings of the national academy of sciences **116**(30), 15023–15032 (2019)
4. Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L.B., Weinberger, A., Segal, E.: Deciphering the rules by which 5-utr sequences affect protein expression in yeast. Proceedings of the National Academy of Sciences **110**(30), 2792–2801 (2013)
5. Araujo, P.R., Yoon, K., Ko, D., Smith, A.D., Qiao, M., Suresh, U., Burns, S.C., Penalva, L.O.: Before it gets started: regulating translation at the 5 utr. Comparative and functional genomics **2012** (2012)
6. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., Weissman, J.S.: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. science **324**(5924), 218–223 (2009)
7. Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S., Seelig, G.: Deep learning of the regulatory grammar of yeast 5 untranslated regions from 500,000 random sequences. Genome research **27**(12), 2015–2024 (2017)
8. Sample, P.J., Wang, B., Reid, D.W., Presnyak, V., McFadyen, I.J., Morris, D.R., Seelig, G.: Human 5 utr design and variant effect prediction from a massively parallel translation assay. Nature biotechnology **37**(7), 803–809 (2019)
9. Li, J.J., Chew, G.-L., Biggin, M.D.: Quantitative principles of cis-translational control by general mrna sequence features in eukaryotes. Genome biology **20**(1), 1–24 (2019)
10. Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., Wang, C.L.: Quantitative analysis of mammalian translation initiation sites by facs-seq. Molecular systems biology **10**(8), 748 (2014)
11. Cao, J., Novoa, E.M., Zhang, Z., Chen, W.C., Liu, D., Choi, G.C., Wong, A.S., Wehrspaun, C., Kellis, M., Lu, T.K.: High-throughput 5 utr engineering for enhanced protein production in non-viral gene therapies. Nature communications **12**(1), 1–10 (2021)
12. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology **33**(8), 831–838 (2015)
13. Zeng, H., Edwards, M.D., Liu, G., Gifford, D.K.: Convolutional neural network architectures for predicting dna–protein binding. Bioinformatics **32**(12), 121–127 (2016)
14. Koo, P.K., Eddy, S.R.: Representation learning of genomic sequence motifs with convolutional neural networks. PLoS computational biology **15**(12), 1007560 (2019)
15. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., *et al.*: Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics **53**(3), 354–366 (2021)
16. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
17. DeGrave, A.J., Janizek, J.D., Lee, S.-I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
18. Kumar, A., Daume III, H.: Learning task grouping and overlap in multi-task learning. arXiv preprint arXiv:1206.6417 (2012)
19. Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? In: International Conference on Machine Learning, pp. 9120–9132 (2020). PMLR
20. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

21. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp. 3145–3153 (2017). PMLR
22. De Lucrezia, D., Slanzi, D., Poli, I., Polticelli, F., Minervini, G.: Do natural proteins differ from random sequences polypeptides? natural vs. random proteins classification using an evolutionary neural network. PLoS One **7**(5), 36634 (2012)
23. Yu, J.-F., Cao, Z., Yang, Y., Wang, C.-L., Su, Z.-D., Zhao, Y.-W., Wang, J.-H., Zhou, Y.: Natural protein sequences are more intrinsically disordered than random sequences. Cellular and Molecular Life Sciences **73**(15), 2949–2957 (2016)
24. Yona, A.H., Alm, E.J., Gore, J.: Random sequences rapidly evolve into de novo promoters. Nature communications **9**(1), 1–10 (2018)
25. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? Advances in neural information processing systems **27** (2014)
26. Andreev, D.E., O'Connor, P.B., Fahey, C., Kenny, E.M., Terenin, I.M., Dmitriev, S.E., Cormican, P., Morris, D.W., Shatsky, I.N., Baranov, P.V.: Translation of 5 leaders is pervasive in genes resistant to eif2 repression. Elife **4**, 03971 (2015)
27. Hsieh, A.C., Liu, Y., Edlind, M.P., Ingolia, N.T., Janes, M.R., Sher, A., Shi, E.Y., Stumpf, C.R., Christensen, C., Bonham, M.J., *et al.*: The translational landscape of mtor signalling steers cancer initiation and metastasis. Nature **485**(7396), 55–61 (2012)
28. Wein, N., Vulin, A., Falzarano, M.S., Szigyarto, C.A.-K., Maiti, B., Findlay, A., Heller, K.N., Uhlén, M., Bakthavachalu, B., Messina, S., *et al.*: Translation from a dmd exon 5 ires results in a functional dystrophin isoform that attenuates dystrophinopathy in humans and mice. Nature medicine **20**(9), 992–1000 (2014)
29. Kuret, K., Amalietti, A.G., Ule, J.: Positional motif analysis reveals the extent of specificity of protein-rna interactions observed by clip. bioRxiv (2021)
30. Komatsu, K.R., Taya, T., Matsumoto, S., Miyashita, E., Kashida, S., Saito, H.: Rna structure-wide discovery of functional interactions with multiplexed rna motif library. Nature communications **11**(1), 1–14 (2020)
31. Schultz, J., Maisel, S., Gerlach, D., Müller, T., Wolf, M.: A common core of secondary structure of the internal transcribed spacer 2 (its2) throughout the eukaryota. Rna **11**(4), 361–364 (2005)
32. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., *et al.*: A compendium of rna-binding motifs for decoding gene regulation. Nature **499**(7457), 172–177 (2013)
33. Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., *et al.*: The mrna-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell **46**(5), 674–690 (2012)
34. Lin, J.-C., Hsu, M., Tarn, W.-Y.: Cell stress modulates the function of splicing regulatory protein rbm4 in translation control. Proceedings of the National Academy of Sciences **104**(7), 2235–2240 (2007)
35. Kozak, M.: An analysis of 5'-noncoding sequences from 699 vertebrate messenger rnas. Nucleic acids research **15**(20), 8125–8148 (1987)
36. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., *et al.*: Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics **53**(3), 354–366 (2021)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
38. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. Nature methods **9**(4), 357–359 (2012)
39. Kokkinos, I.: Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6129–6138 (2017)
40. Wang, H., Yang, L., Wang, Y., Chen, L., Li, H., Xie, Z.: Rpfdb v2. 0: an updated database for genome-wide information of translated mrna generated from ribosome profiling. Nucleic acids research **47**(D1), 230–234 (2019)
41. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.*: Ncbi geo: archive for functional genomics data sets—update. Nucleic acids research **41**(D1), 991–995 (2012)
42. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.*: Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic acids research **44**(D1), 733–745 (2016)
43. Choi, G.C., Zhou, P., Yuen, C.T., Chan, B.K., Xu, F., Bao, S., Chu, H.Y., Thean, D., Tan, K., Wong, K.H., *et al.*: Combinatorial mutagenesis en masse optimizes the genome editing activities of spcas9. Nature methods **16**(8), 722–730 (2019)