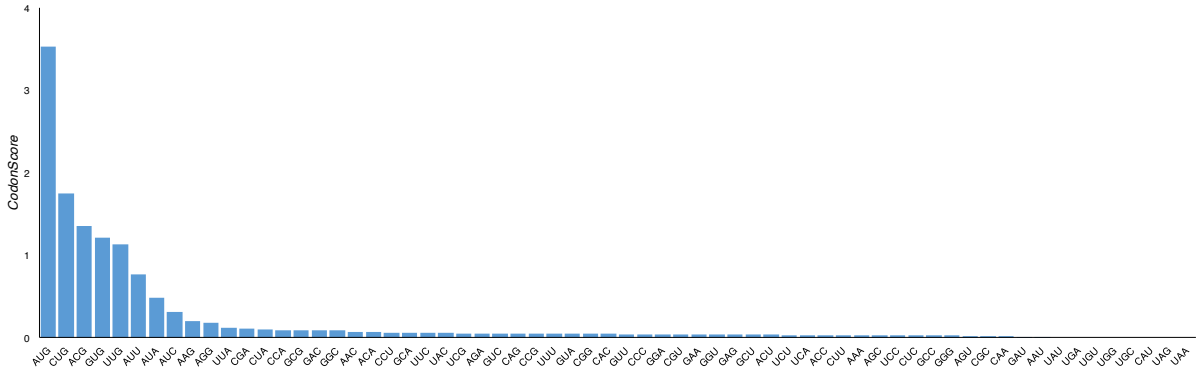
Supplementary Materials for

TITER: predicting translation initiation sites by deep learning

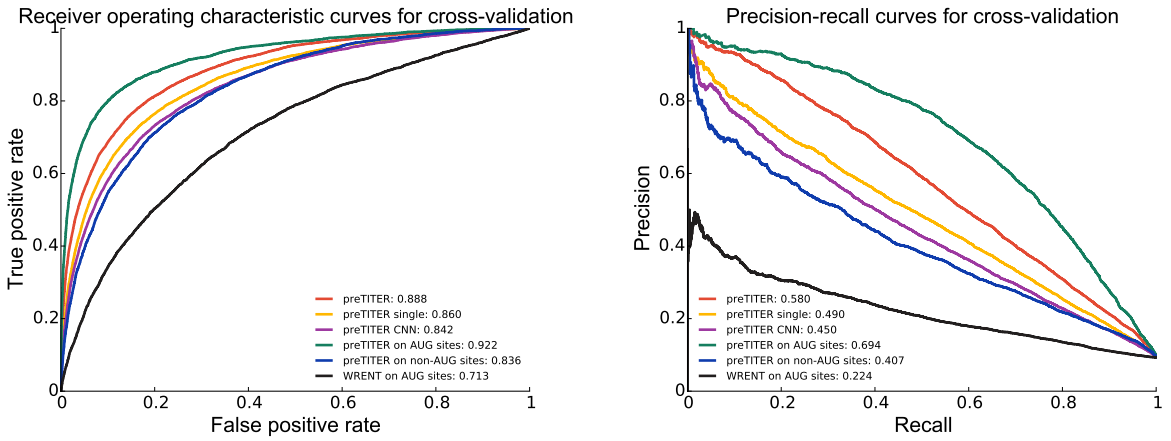
Sai Zhang[1,†], Hailin Hu[2,†], Tao Jiang[3,4,5], Lei Zhang[2,*] and Jianyang Zeng[1,*]

1 Supplementary Figures



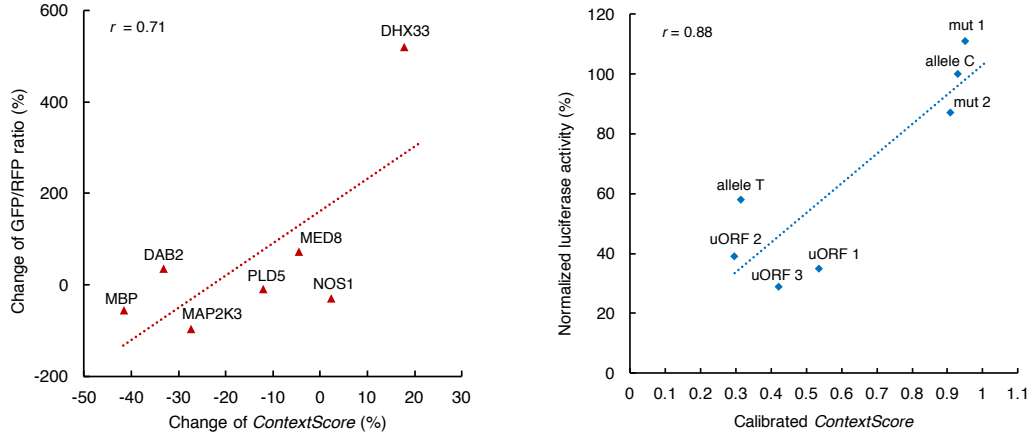
Supplementary Figure 1: *CodonScores* of different codons for the Gao15 dataset. See the main text for more details.

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.
[2] School of Medicine, Tsinghua University, Beijing, China.
[3] Department of Computer Science and Engineering, University of California, Riverside, CA, USA.
[4] MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST/Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[5] Institute of Integrative Genome Biology, University of California, Riverside, CA, USA.
[†] These authors contributed equally to this work.
[*] To whom correspondence should be addressed. Email: `lzhang20@mail.tsinghua.edu.cn` and `zengjy321@tsinghua.edu.cn`.



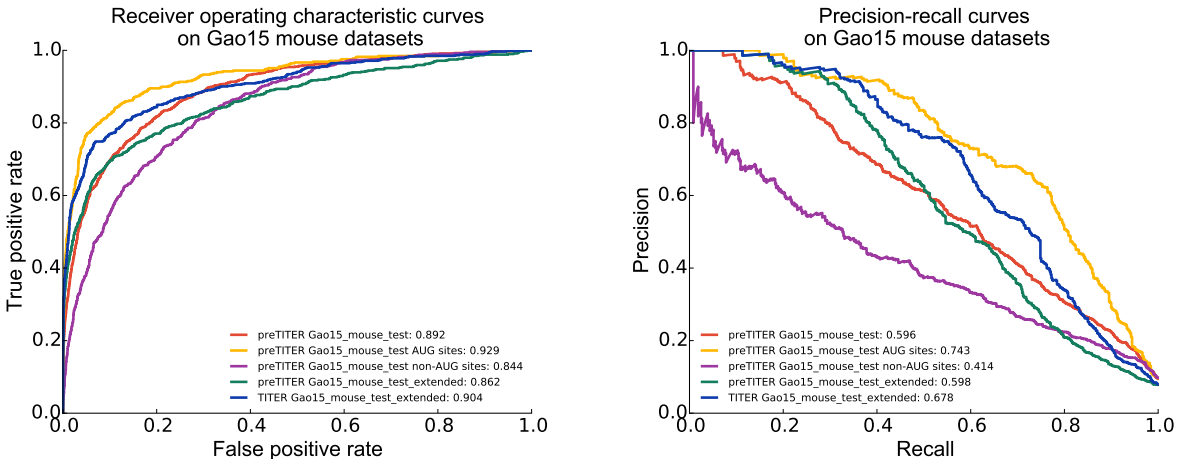
(a) (b)

Supplementary Figure 2: Prediction performance in the five-fold cross-validation tests evaluated by (a) ROC and (b) PR curves as well as the corresponding AUROC and AUPR scores. "preTITER" denotes a preliminary version of our deep learning framework consisting of an ensemble of 32 classifiers that only considered the context features of TISs, i.e., without incorporating the preference of codon composition of TISs. "preTITER single" denotes the preTITER framework consisting of only one classifier. "preTITER CNN" denotes the preTITER framework consisting of only convolutional neural networks, i.e., without using the recurrent layer.

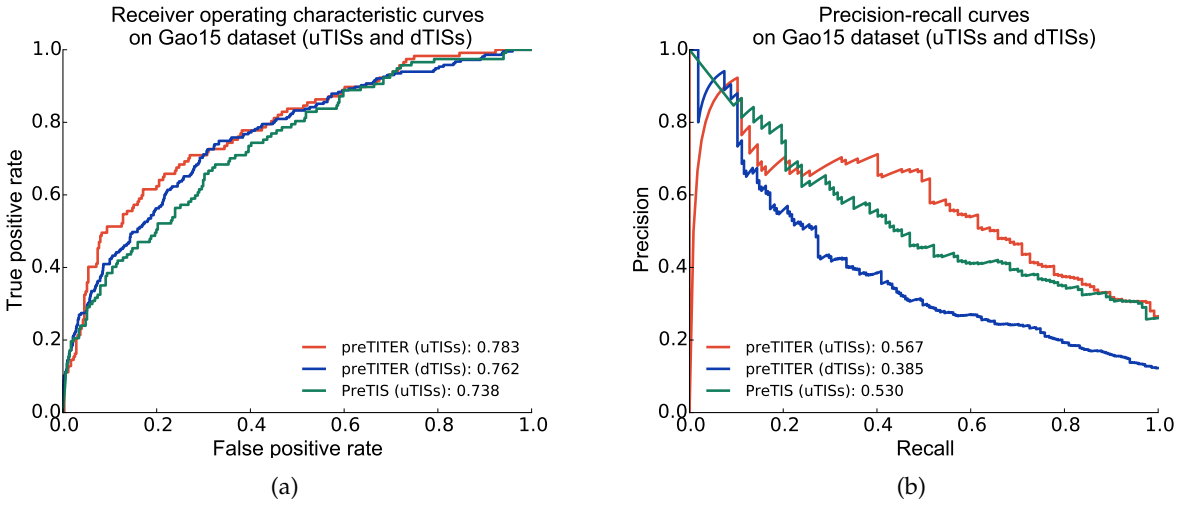


(a) (b)

Supplementary Figure 3: Additional tests on the correlations between the prediction scores of TITER and the experimentally-quantified mutational effects of the sequence variants in the TIS contexts. (a) and (b) correspond to the test results on the mutations derived from the studies in [1] and [2], respectively, in which the corresponding sequences of the real transcripts instead of the plasmid sequences were input into the TITER framework to compute the *ContextScores*.



Supplementary Figure 4: Prediction performance on the mouse test dataset evaluated by (a) ROC and (b) PR curves, respectively. "preTITER" denotes a preliminary version of our deep learning framework that only considered the context features of TISs.



Supplementary Figure 5: Prediction performance on different regions of the Gao15 dataset evaluated by (a) ROC and (b) PR curves, respectively. In particular, we constructed the uTIS dataset based on the same procedure as in the construction of the Gao15_test dataset, except that we only considered TISs in the 5′ UTRs and excluded all the samples that did not satisfy the input conditions of PreTIS. Note that since PreTIS did not provide any specification on its training and test data, the test data used for comparison here may include the training data of PreTIS (e.g., the overlapped uTISs measured by QTI-seq and GTI-seq). Therefore, the comparison in this part is only an under-estimation of the superiority of our method over PreTIS.

2 Supplementary Tables

Supplementary Table 1: The calibrated hyperparameter values for the hybrid deep neural network used in TITER.

Hyperparameter	value
Kernel number	128
Kernel size	3
Max norm of weights in the convolutional layer	3
Pooling length	3
Dropout rate after the pooling layer	0.214
Output dimension of the LSTM layer	256
Dropout rate after the LSTM layer	0.724
Optimizer	Nadam [3]

The hyperparameters were optimized based on the TPE approach [4] (see the main text for more details).

References

[1] W. L. Noderer, R. J. Flockhart, A. Bhaduri, A. J. Diaz de Arce, J. Zhang, P. A. Khavari, and C. L. Wang, "Quantitative analysis of mammalian translation initiation sites by FACS-seq," *Mol Syst Biol*, vol. 10, no. 8, pp. 748–, 2014.

[2] S. E. Calvo, D. J. Pagliarini, and V. K. Mootha, "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans," *Proceedings of the National Academy of Sciences*, vol. 106, no. 18, pp. 7507–7512, 2009.

[3] T. Dozat, "Incorporating Nesterov momentum into Adam," pp. –.

[4] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds.), pp. 2546–2554, Curran Associates, Inc., 2011.