frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 10 October 2023
DOI 10.3389/fbinf.2023.1275787

OPEN ACCESS

EDITED BY
Shicheng Guo,
Johnson & Johnson, United States

REVIEWED BY
Yu Wang,
University of Virginia, United States
Kefan Yang,
Northwestern University, United States
Minghan Yang,
New York University, United States
Tianyi Zhang,
Emory University, United States
Chongbo Yang,
Cornell University, United States

*CORRESPONDENCE
Tsukasa Fukunaga,
fukunaga@aoni.waseda.jp
Michiaki Hamada,
mhamada@waseda.jp

RECEIVED 10 August 2023
ACCEPTED 29 September 2023
PUBLISHED 10 October 2023

CITATION
Hara K, Iwano N, Fukunaga T and Hamada M (2023), DeepRaccess: high-speed RNA accessibility prediction using deep learning.
Front. Bioinform. 3:1275787.
doi: 10.3389/fbinf.2023.1275787

DeepRaccess: high-speed RNA accessibility prediction using deep learning

Kaisei Hara[1,2], Natsuki Iwano[1], Tsukasa Fukunaga[3]* and Michiaki Hamada[1,2,4]*

[1]Department of Electrical Engineering and Bioscience, Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan, [2]Computational Bio Big-Data Open Innovation Laboratory, AIST-Waseda University, Tokyo, Japan, [3]Waseda Institute for Advanced Study, Waseda University, Tokyo, Japan, [4]Graduate School of Medicine, Nippon Medical School, Tokyo, Japan

RNA accessibility is a useful RNA secondary structural feature for predicting RNA-RNA interactions and translation efficiency in prokaryotes. However, conventional accessibility calculation tools, such as Raccess, are computationally expensive and require considerable computational time to perform transcriptome-scale analysis. In this study, we developed DeepRaccess, which predicts RNA accessibility based on deep learning methods. DeepRaccess was trained to take artificial RNA sequences as input and to predict the accessibility of these sequences as calculated by Raccess. Simulation and empirical dataset analyses showed that the accessibility predicted by DeepRaccess was highly correlated with the accessibility calculated by Raccess. In addition, we confirmed that DeepRaccess could predict protein abundance in *E.coli* with moderate accuracy from the sequences around the start codon. We also demonstrated that DeepRaccess achieved tens to hundreds of times software speed-up in a GPU environment. The source codes and the trained models of DeepRaccess are freely available at https://github.com/hmdlab/DeepRaccess.

KEYWORDS

RNA secondary structure, RNA accessibility, machine learning, acceleration, translation efficiency prediction

1 Introduction

RNA molecules play crucial roles in the regulation of diverse cellular processes, and their regulatory functions are closely linked to their structures (Mortimer et al., 2014). For example, tRNAs have to form cloverleaf secondary structures and L-shaped tertiary structures in order to function properly during translation. As another example, in prokaryotic translation, the RNA region upstream of the start codon has a function to regulate protein abundance, and the level of abundance decreases when the region takes a stem structure (de Smit and van Duin, 1990). Accordingly, many experimental and computational studies have been carried out to analyze RNA structures in order to elucidate the relationships between the structures and functions (Bonilla et al., 2022; Wayment-Steele et al., 2022). In particular, computational analyses of RNA secondary structures are frequently performed because of their low cost, moderate accuracy, and high speed (Reuter and Mathews, 2010; Lorenz et al., 2011; Huang et al., 2019; Sato et al., 2021; Fukunaga and Hamada, 2022; Sato and Hamada, 2023).

RNA accessibility is one of the secondary structural features and is defined as the energy required for an RNA region not to form a stem structure. The accessibility is used to predict

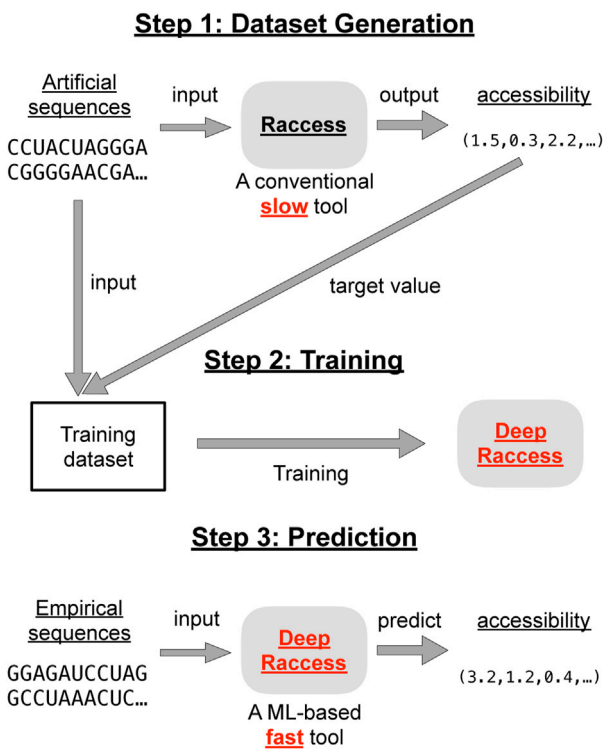

FIGURE 1
The schematic illustration of the DeepRaccess approach. DeepRaccess first trains a model to predict accessibility. The training dataset was composed of artificial RNA sequences and their corresponding accessibility calculated by Raccess. DeepRaccess then rapidly predicts the accessibility of empirical RNA sequences based on the trained predictive model.

RNA-RNA interactions (Agarwal et al., 2015; Fukunaga and Hamada, 2017; Mann et al., 2017) and translation efficiency in prokaryotes (Terai and Asai, 2020) because these molecular processes are more likely to occur when the RNA region of interest is single-stranded. Therefore, several software programs have been developed to calculate the RNA accessibility (Bernhart et al., 2006; Lu and Mathews, 2008; Bernhart et al., 2011; Kiryu et al., 2011; Lange et al., 2012). Some of these programs used a local folding approach, which reduces computational time by ignoring long-distance base pairs (Bernhart et al., 2006; Kiryu et al., 2011; Lange et al., 2012). However, current methods are still too computationally expensive for transcriptome-scale analysis, and thus the development of faster methods for calculating accessibility is an essential research topic. In general, one of the powerful approaches to speed up the calculation is parallel computing, and several parallel algorithms have now been proposed for RNA secondary structure analysis (Fekete et al., 2000; Kawaguchi and Kiryu, 2016). However, the parallel algorithms for RNA secondary structure analysis have not been fully explored, especially in parallel computations using GPUs (Rizk and Lavenier, 2009). This is because most algorithms for RNA secondary structure analysis are based on dynamic programming, which is difficult to parallelize.

In recent years, machine learning-based software acceleration has attracted attention in computer simulation (Um et al., 2020; Kochkov et al., 2021; Sun et al., 2023). Some of these methods used running results of a slow but accurate simulator as training data, and constructs a predictive model that reproduces the simulation results. Since the run of the predictive model is generally much faster than that of the simulator, the accurate predictive model can be seen as a fast alternative to the simulator. In particular, deep learning-based methods have the advantage of using GPUs efficiently based on the deep learning libraries without the need to build specialized algorithms. Machine learning-based acceleration is beginning to be used in bioinformatics, such as phylogenetic tree construction (Azouri et al., 2021) and sequence alignment score calculation (Zheng et al., 2019; Corso et al., 2021; Girgis et al., 2021; Chen et al., 2022). However, there is no research on the application to RNA secondary structure analysis.

In this study, we developed DeepRaccess, a fast accessibility prediction tool based on deep learning-based software acceleration. We confirmed that DeepRaccess could reproduce the results of an existing RNA accessibility calculation method with high accuracy on both simulation and empirical datasets. We also demonstrated that the accessibility calculated by DeepRaccess was moderately correlated with protein abundance in *E. coli*. Finally, we verified that DeepRaccess was significantly faster than an existing method on various datasets in a GPU environment.

2 Materials and methods

2.1 Overview of the DeepRaccess software

DeepRaccess is a machine learning predictor whose input is an RNA sequence and whose output is the accessibility in subregions of the sequence. Figure 1 shows an overview of the DeepRaccess approach. The subregion length l_a is fixed in the training step, and the accessibility of all subregions with the length l_a are the output. When users require the accessibility with a different length l_a, they have to redo the training of the prediction model. In this study, we used 35 as the default value for l_a. Note that this value has been used to predict prokaryotic translation efficiency in a previous study (Terai and Asai, 2020).

The training datasets consisted of RNA sequences as the input and the accessibility as the target values. The RNA sequences were artificially generated (the details are described in Section 2.3), and the accessibility was calculated from the input RNA sequences using Raccess (Kiryu et al., 2011). Raccess adopts a local folding approach that speeds up the computation by ignoring base pairs spanning more than W bases, and can compute the accessibility of all subregions based on a secondary structure score model for a fixed l_a length. The computation is based on dynamic programming, and the time complexity is $O(NW^2)$ where N is the sequence length. In this study, we used the CONTRAfold model as the score model because of its high accuracy (Do et al., 2006), and used 100 as the default value of W. Note that Raccess is the only software that can compute the accessibility for long sequences in a reasonable time under the numerically stable computation. The source code for Raccess is not available from the link given in the original Raccess paper, but can be downloaded from the following link: https://github.com/gterai/raccess.

We set the maximum sequence length in the training dataset to 440, which is the value used in RNABERT (Akiyama and Sakakibara, 2022). We therefore could not predict the accessibility of sequences longer than 440 bases by merely applying DeepRaccess without any modifications. Therefore, for such a long sequence, we predicted the accessibility based on the following procedures. First, DeepRaccess divides the sequence into 440-base subsequences by shifting the window by 330 bases. This means that neighboring subsequences overlap by 110 bases. DeepRaccess then predicted the accessibility of these subsequences and integrated them with the accessibility of the full-length RNA. Specifically, because the accuracy of RNA accessibility declines in regions near the sequence end points (Kiryu et al., 2011), we ignored the accessibility of the 55-base region from the end of each subsequence in the overlapped region.

2.2 Neural network architecture

We used deep neural networks as the predictor. We implemented four representative network architectures using PyTorch and compared their prediction accuracy: 1) Fully Convolutional Network (FCN) (Long et al., 2015), 2) U-Net (Ronneberger et al., 2015), 3) Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and 4) RNABERT (Akiyama and Sakakibara, 2022). Supplementary Tables S1–S4 show the details of each network architecture, respectively. We set the epoch and batch sizes to ten and 256, respectively. We also used AdamW as the optimizer and the values used by RNABERT as the hyperparameters of the optimizer.

The input RNA sequences are embedded into numerical vectors and fed into the neural networks. The FCN and U-Net models used token embedding. This embedding first randomly generates six 120-dimensional numerical vectors corresponding to each of the six states: four RNA bases (A, C, G, U), one undetermined nucleotide (N) and padding. The resulting vectors are then assigned to each state in the input sequences. The BERT and RNABERT models used positional embedding in addition to token embedding. In this embedding, 120-dimensional numerical vectors corresponding to each position in the sequences are randomly generated. Finally, the values of the two embedding results are summed for each base.

We briefly review each network architecture. FCN is a type of CNN architecture that is widely used for image segmentation. FCN does not use fully connected layers and is composed only of convolutional layers. We used a network of 40 convolutional layers with constant channel and unit sizes as the FCN model. U-Net is a variant of FCN, and consists of three parts, bottom-up path, bottleneck, and top-down path. The data is downsampled in the bottom-up path, the computation is performed in convolutional layers with the smallest unit sizes in the bottleneck, and the data is upsampled in the top-down path. The essential feature of U-Net is that the layers on the bottom-up and top-down paths have skip connections. For the U-Net model, we used a network consisting of 3, 35, and 3 layers on the bottom-up path, bottleneck, and top-down path, respectively. BERT was originally developed for natural language processing and is a model in which transformer layers are stacked several times. In this study, we stacked six transformer layers. Transformer can incorporate positional information of elements into the model by using the attention mechanism. RNABERT is a BERT model pre-trained on 76,237 human small ncRNAs in the RNAcentral database (Petrov et al., 2017). Both RNA sequence and structural information were embedded in the learned representation of RNABERT. We fine-tuned the pre-trained RNABERT model in the same way as other models were trained.

2.3 Training datasets

All sequence data used for training were artificially generated. We generated the sequences using two methods: 1) uniform base sampling to generate RNAs that lack strong stem structures and 2) sampling to generate RNAs with strong stem structures similar to small ncRNAs. In this paper, we refer to these methods as the uniform and the structured RNA sampling methods, respectively.

In the uniform sampling method, we first determined the sequence length N by sampling from the uniform distribution $unif(100, 440)$. The bases in the sequences were sampled from the categorical distribution $Cat(x|\pi)$ for the category (A, C, G, U, N), and π was sampled from the Dirichlet distribution $Dir(\pi|\alpha = [1, 1, 1, 1, 0.1])$. π was sampled once per sequence.

In the structured RNA sampling method, after generating a sequence based on the uniform sampling method, we determined the stem length l by sampling from $unif(8, 48)$. We next selected the length d of the region flanked between two stem regions from $unif(3, N - 2l)$. We also selected the start position of the first stem region from $unif(0, N - 2l - d)$. We then substituted the bases in the second stem regions so that the bases were complementary to the bases of the first region. When the base was G or U, whether the base formed a Watson-Crick base pair or a wobble base pair was determined by the Bernoulli distribution $Bern(x|\mu)$. μ was sampled from the Beta distribution $Beta(\mu|\alpha = 4, \beta = 1)$. After that, we substituted the bases in the stem region to create internal loops, and whether a base was substituted or not was determined by $Bern(x|\mu)$. Here, μ was sampled from $Beta(\mu|\alpha = 1, \beta = 15)$, and the base after the substitution was sampled from $Cat(x|\pi)$ using the uniform sampling method. We also substituted the next base after the substituted base according to $Bern(x|\mu)$, and μ was sampled from $Beta(\mu|\alpha = 2, \beta = 1)$.

Using these two methods, we created two training datasets that were a uniform RNA dataset and a structured RNA dataset. In the former, all sequences were generated by the uniform sampling method, while the latter contained half of each of the sequences generated by the two sampling methods. We performed the training on each of the two training datasets and created two predictive models for each architecture. We used 10 million as the default number of sequences per the training.

2.4 Test datasets and evaluation measure

We evaluated the prediction accuracy of DeepRaccess using simulation test datasets and three empirical datasets: Rfam, Gencode, and *E.coli* synthetic mRNA datasets. As the test simulation dataset, we used a dataset generated in the same way as the training data used for the trained model. We used 100 thousand as the number of sequences per the test dataset. The Rfam dataset consisted of 3,105,149 sequences from the Rfam 14.9 database (Kalvari et al., 2021), and most of which are highly structured. The Gencode dataset contains 142,379 transcripts in

TABLE 1 Comparison of the prediction accuracy among the neural network architectures.

Architecture	Uniform		Structured	
	NMSE	Spearman's ρ	NMSE	Spearman's ρ
FCN	**0.0754**	**0.9943**	**0.1148**	**0.9876**
U-Net	0.0984	0.9913	0.2400	0.9788
BERT	0.0912	0.9919	0.2472	0.9734
RNABERT	0.1076	0.9901	0.2442	0.9736

Normalised mean square error (NMSE) is the MSE divided by the target value. "Uniform" and "Structured" mean the evaluation result for the uniform and structured datasets, respectively. The bold values are the highest scores among the neural network architectures.

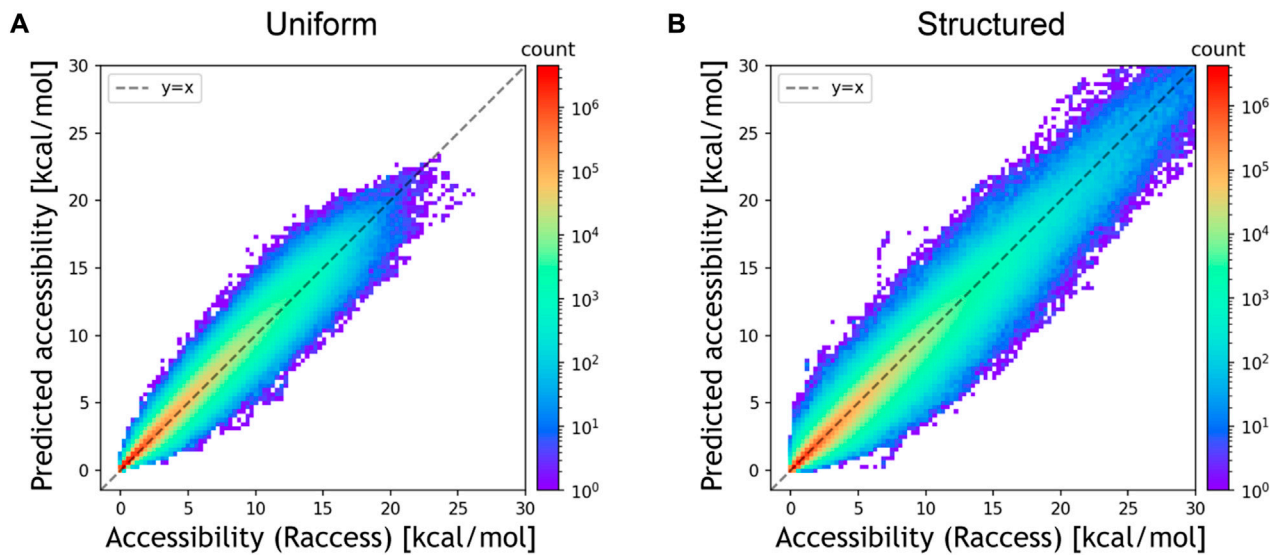

FIGURE 2
Prediction accuracy of the FCN architecture for the simulation datasets [**(A)** the uniform RNA dataset and **(B)** the structured RNA dataset]. The *x* and *y*-axes represent the accessibility calculated by Raccess and predicted accessibility, respectively. The color bar representing the counts is displayed using a log scale.

Gencode version M29 (Frankish et al., 2021). We have removed sequences of less than 35 bases from this analysis. The *E.coli* synthetic mRNA datasets consisted of 244,000 sequences with 120 bases around the start codon of synthetic mRNAs (Cambray et al., 2018; Terai and Asai, 2020). We applied Raccess and DeepRaccess to these datasets, and compared the accessibility calculated by the methods. For the evaluation measure, we used Spearman's rank correlation coefficient (ρ) and normalized mean square error (NMSE), which is the MSE divided by the target value.

To validate the usefulness of DeepRacess, we also evaluated its predictive performance for prokaryotic translation efficiency. We used the *E. coli* synthetic mRNA dataset for this analysis, and calculated Spearman's ρ between protein abundance and the accessibility based on the DeepRaccess. For the comparison, we used the accessibility calculated by Raccess, minimum free energy (MFE) calculated by CONTRAfold (Do et al., 2006), and scores of RBSDesigner (Na et al., 2010) and RBSCalculator (Salis, 2011).

We investigated the computational speed of DeepRaccess and compared it to Raccess. Raccess and DeepRaccess were run in a CPU-only environment (CPU: Intel(R) Xeon(R) Gold 6,148 2.1 GHz, memory: 8 GB). In addition, DeepRaccess was also run in an environment where both CPU and GPU were available (CPU: Intel(R) Xeon(R) CPU E5-2,698 v4 2.2 GHz, GPU: Tesla V100 DGXS 32GB×4, memory: 257GiB).

3 Results

3.1 Accuracy evaluation on simulation datasets

We first evaluated the prediction accuracy of DeepRaccess using simulation test datasets and compared the performances of different neural network architectures. Table 1; Figure 2; Supplementary Figure S1 show the prediction performances of DeepRaccess. We found that the NMSEs were less than 0.25 and the Spearman's ρs were greater than 0.97 in all cases, suggesting that deep learning is effective in predicting RNA accessibility. In addition, the scores based on the structured RNA dataset were worse than those based on the uniform RNA dataset in each architecture. The reason for the difficulty in prediction may be that the structured RNA dataset has a large variance in the RNA accessibility. We also verified that the FCN was the best-

10.3389/fbinf.2023.1275787

TABLE 2 Prediction performances for the empirical datasets.

	Uniform		Structured	
Dataset	NMSE	Spearman's ρ	NMSE	Spearman's ρ
Gencode	0.1352	0.9215	0.1422	0.9159
Rfam	0.5493	0.8721	0.2244	0.9044
E.coli	0.1186	0.8821	0.1520	0.8548

performing architecture in both datasets and therefore used the FCN in the following analyses.

We next investigated the effect of the training data size on the prediction performances using the structured RNA dataset (Supplementary Table S5; Supplementary Figure S2). We confirmed that the prediction accuracy improved as the data size increased, and the accuracy had not yet converged even when the data size was increased to 10 million. Therefore, we should achieve higher prediction accuracy when more sequences are used for the training dataset and more time is spent on training. We also evaluated the effect of the parameters l_a and W on the performances (Supplementary Table S6; Supplementary Figure S3). We found that DeepRaccess had higher performance when l_a was large. In addition, small W resulted in accurate prediction. This may be because the RNA accessibility has small variances and ranges when W is small.

3.2 Accuracy evaluation on the empirical datasets

We then assessed whether DeepRaccess could predict the accessibility of empirical RNA sequences using three datasets (Table 2; Figure 3; Supplementary Figure S4). We verified that the best predictor for each dataset had an NMSE of less than 0.23, and the Spearman's ρ was greater than 0.88 for all datasets. While the predictive model trained on the uniform RNA dataset

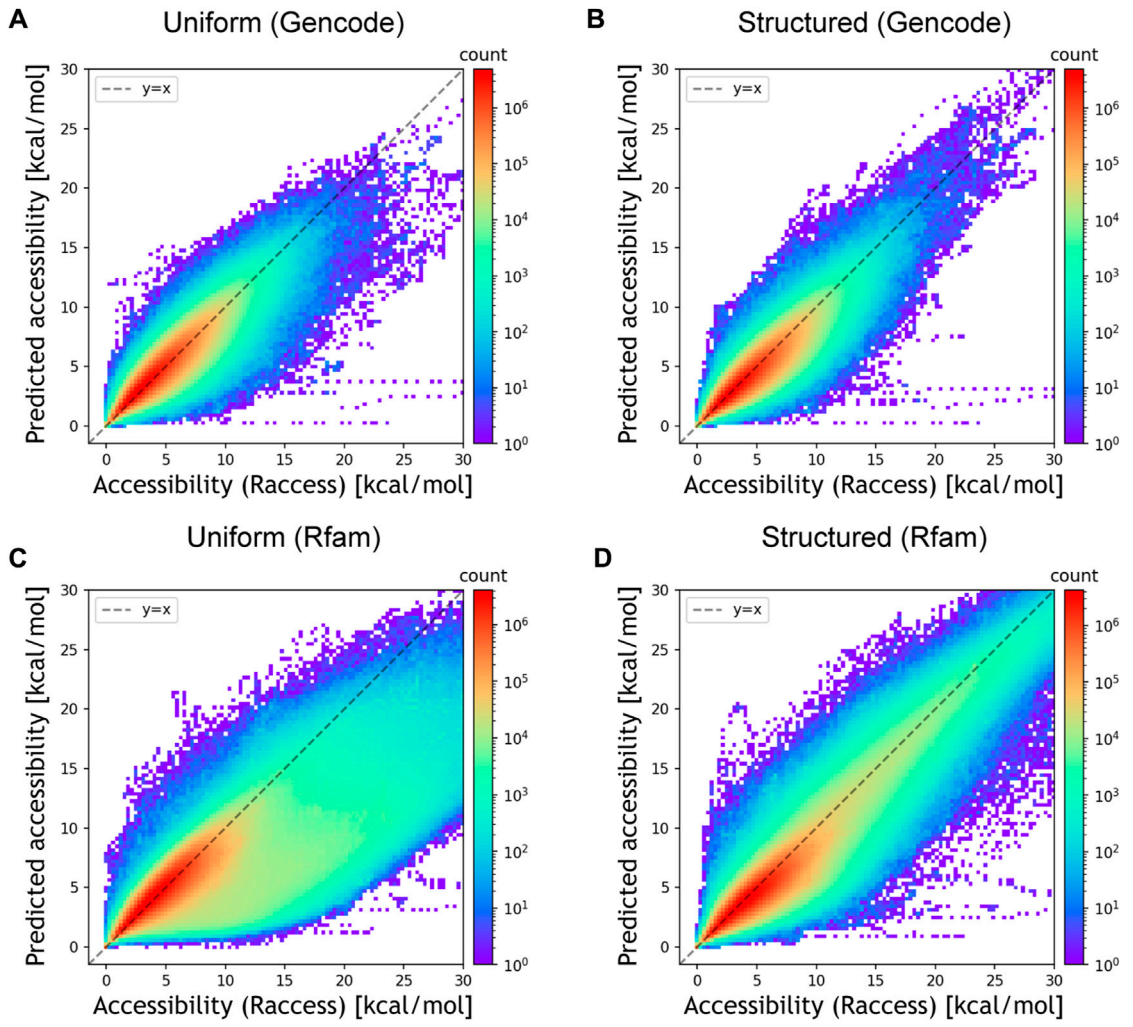

FIGURE 3
Prediction accuracy for the empirical datasets. **(A)** Prediction performances for the Gencode dataset of the predictive model trained on the uniform RNA dataset and **(B)** the structured RNA dataset. **(C)** Prediction performances for the Rfam dataset of the predictive model trained on the uniform RNA dataset and **(D)** the structured RNA dataset. The *x* and *y*-axes represent the accessibility calculated by Raccess and predicted accessibility, respectively. The color bar representing the counts is displayed using a log scale.

TABLE 3 Prediction performances for the protein abundance in the *E.coli* synthetic mRNA dataset.

Measure	DeepRaccess (Uniform)	DeepRaccess (Structured)	Raccess	MFE	RBSDesigner	RBSCalculator
Spearman's ρ	0.585	0.493	0.709	0.605	0.440	0.540

The values of RBSDesigner and RBSCalculator were cited from (Terai and Asai, 2020).

outperformed that trained on the structured RNA dataset for the Gencode and *E.coli* synthetic mRNA datasets, the opposite trend was found for the Rfam dataset. In addition, the prediction results on the Rfam dataset had the highest NMSE of the three datasets. These results are probably due to the fact that the Rfam dataset contains many structured RNAs. Furthermore, for the Rfam and Gencode datasets, we found some data has very low predicted accessibility although the accessibility calculated by Raccess was large. This result means that these regions were predicted to form few stems, even though they actually have strong stem structures. In conclusion, DeepRaccess was also able to predict RNA accessibility with high accuracy for empirical RNA sequences, but its accuracy was insufficient for highly structured RNAs.

We also evaluated the correlation between the protein abundance in *E.coli* and the accessibility calculated by DeepRaccess (Table 3; Figure 4; Supplementary Figure S5). We found that the Spearman's ρ by DeepRaccess trained with the uniform and the structured dataset were 0.585 and 0.493, respectively, indicating that DeepRaccess could predict the protein abundance with moderate accuracy. The prediction accuracy of DeepRaccess trained with the uniform dataset was lower than that based on the accessibility calculated by Raccess, but comparable to that of the MFE score and higher than those of RBSDesigner and RBSCalculator.

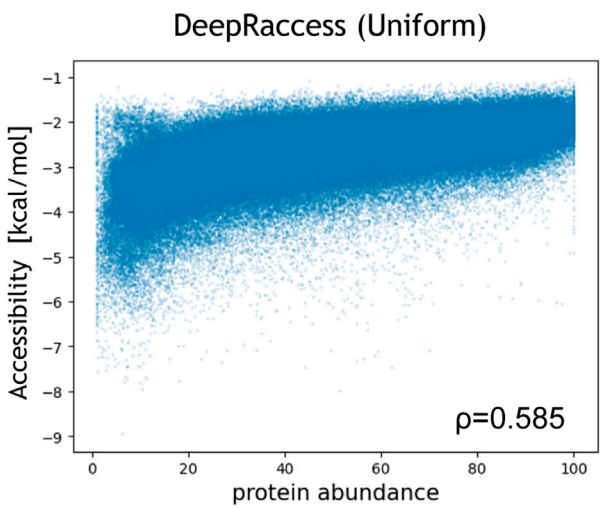

FIGURE 4
Correlation between the protein abundance and the accessibility calculated by DeepRaccess trained with the uniform RNA dataset. The protein abundance was measured by fluorescence-activated cell sorting and was normalized so that the minimum value was 1 and the maximum value was 100 (Cambray et al., 2018). The x- and y-axis represent the protein abundance and the accessibility, respectively.

3.3 Evaluation of the computational speed

Finally, we evaluated the runtime of DeepRaccess. First, we looked at the time taken to train the predictive model and found it to be 2 days and 20 h in the GPU environment. This is not short but the prediction model only needs to be trained once, and thus the long time is not a practical bottleneck. Note that this training step is not necessary when users are using trained models of DeepRaccess. We next assessed the time taken to predict the RNA accessibility (Table 4). In the CPU-only environment, DeepRaccess was not necessarily faster than Raccess, and the superiority depended on the datasets. On the other hand, DeepRaccess was tens to hundreds of times faster in the GPU environment than Raccess in the CPU environment. Although it should be noted that the environment in which the computations were performed was different, we have shown that DeepRaccess was extremely fast compared to Raccess.

4 Discussion

In the current study, we proposed DeepRaccess, a rapid RNA accessibility prediction method based on the deep learning. We evaluated the prediction accuracy and the computational speed of DeepRaccess using the simulation and three empirical datasets. We generated two training datasets, the uniform RNA datasets and the structured RNA datasets. We validated that DeepRaccess had a high level prediction accuracy while exhibiting significantly faster performance on the GPU environment. When calculating the accessibility of RNAs such as mRNAs and long ncRNAs, DeepRaccess trained with the uniform RNA datasets was more effective. On the other hand, when calculating the accessibility of structured RNAs such as short ncRNAs, DeepRaccess trained with the structured RNA datasets was preferable. In addition, we demonstrated that the accessibility of regions around start codons of *E.coli* mRNA calculated by DeepRaccess can predict the protein abundance.

Although DeepRaccess had high prediction accuracy, further improvement in prediction performance is an essential issue. The simplest approach is to increase the number of training data. In this study, we used 10 million RNA sequences as our training data, but we expect to improve the accuracy by using several billion RNA sequences. While increasing data size is difficult in machine learning for bioinformatics in general, our method allows unlimited data growth by generating data through simulation. In addition, there is scope for improvement in training data generation methods. In this study, we employed two sampling methods: the uniform and structured RNA sampling methods. We investigated how the distribution of accessibility in our training dataset differs from those in the empirical dataset (Supplementary Figure S6). As a result, we found that our training datasets, even the structured RNA dataset, tend to have lower accessibility values than empirical datasets. We also investigated the relationship between accessibility and NMSE, and found that there was a correlation

TABLE 4 The run time evaluation on simulation and empirical datasets.

Program	Simulation	Gencode	Rfam	*E.coli*
Raccess	11h52m (628.3)	5d22h (82.0)	6d01h (183.5)	8h53m (201.1)
DeepRaccess: < CPU >	3h02m (160.3)	9d03h (126.5)	4d23h (149.9)	8h24m (190.1)
DeepRaccess: < GPU >	1m08s (1.0)	1h44m (1.0)	47m33s (1.0)	2m39s (1.0)

The rows indicate the run times and the run time ratio of each program to DeepRaccess < GPU > .

between higher accessibility and higher NMSE (Supplementary Figure S7). This may be due to the lack of the data of strong stem region in the dataset. Therefore, utilizing enhanced sequence data generation methods that produce data more akin to empirical RNA sequences should improve the prediction accuracy. Deep generative models, such as generative adversarial networks (Zrimec et al., 2022), should hold promise as potential methods.

Furthermore, the development of neural network architectures is also a promising approach for improving accuracy. In this study, we could not fully optimize the architecture due to time and computational resource constraints. The current study is limited to an initial investigation into identifying the most suitable model from among various architectures, including FCN and BERT. Further optimization of the architecture is an important research topic. Given that we had not yet achieved convergence in prediction accuracy when the data size was increased, the current architecture may be overly complex. The development of lightweight architectures with comparable accuracy to the current study should lead to the faster computation of accessibility. As another example, Corso *et al.* proposed that embedding in the hyperbolic space improves the accuracy of predicting the edit distance between sequences (Corso et al., 2021), and thus applying the non-Euclidean space may also be useful in predicting RNA accessibility (Nickel and Kiela, 2017).

The computational speedup provided by the deep learning method can be applied to the other secondary structural features such as base pairing probabilities (McCaskill, 1990), structural profiles (Fukunaga et al., 2014), and structural entropy (Garcia-Martin and Clote, 2015). Each feature has been used to improve the accuracy of RNA secondary structure prediction (Hamada et al., 2009), to predict RNA-protein binding (Ishida et al., 2020), and to evaluate the effect of base mutations on the structure (Kiryu and Asai, 2012). In particular, the algorithm used to compute these structural features taking into account pseudoknots is extremely slow (Dirks and Pierce, 2003), and thus speeding up the method through deep learning should be an important topic of future research.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

KH: Data curation, Methodology, Software, Writing–review and editing, Formal Analysis, Investigation. NI: Data curation, Methodology, Software, Writing–review and editing. TF: Conceptualization, Funding acquisition, Project administration, Writing–original draft. MH: Conceptualization, Funding acquisition, Project administration, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by JSPS KAKENHI Grant numbers: JP22H04891 and JP23K16997 to TF; JP23H00509, JP22H04925, and JP20H00624 to MH. This research was also supported by AMED under Grant Numbers JP22ama121055, JP21ae0121049, and JP21gm0010008 (to MH).

Acknowledgments

Computations were performed on the NIG supercomputer at ROIS National Institute of Genetics.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1275787/full#supplementary-material

References

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4, e05005. doi:10.7554/elife.05005

Akiyama, M., and Sakakibara, Y. (2022). Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *Nar. Genom Bioinform* 4, lqac012. doi:10.1093/nargab/lqac012

Azouri, D., Abadi, S., Mansour, Y., Mayrose, I., and Pupko, T. (2021). Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat. Commun.* 12, 1983. doi:10.1038/s41467-021-22073-8

Bernhart, S. H., ckstein, U., and Hofacker, I. L. (2011). RNA accessibility in cubic time. *Algorithms Mol. Biol.* 6, 3. doi:10.1186/1748-7188-6-3

Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics* 22, 614–615. doi:10.1093/bioinformatics/btk014

Bonilla, S. L., Vicens, Q., and Kieft, J. S. (2022). Cryo-EM reveals an entangled kinetic trap in the folding of a catalytic RNA. *Sci. Adv.* 8, eabq4144. doi:10.1126/sciadv.abq4144

Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat. Biotechnol.* 36, 1005–1015. doi:10.1038/nbt.4238

Chen, J., Yang, L., Li, L., Goodison, S., and Sun, Y. (2022). Alignment-free comparison of metagenomics sequences via approximate string matching. *Bioinforma. Adv.* 2, vbac077. doi:10.1093/bioadv/vbac077

Corso, G., Ying, Z., Pándy, M., Veličković, P., Leskovec, J., and Liò, P. (2021). Neural distance embeddings for biological sequences. *NeurIPS* 34, 18539–18551.

de Smit, M. H., and van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis. *Proc. Natl. Acad. Sci. U. S. A.* 87, 7668–7672. doi:10.1073/pnas.87.19.7668

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805*.

Dirks, R. M., and Pierce, N. A. (2003). A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24, 1664–1677. doi:10.1002/jcc.10296

Do, C. B., Woods, D. A., and Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22, e90–e98. doi:10.1093/bioinformatics/btl246

Fekete, M., Hofacker, I. L., and Stadler, P. F. (2000). Prediction of RNA base pairing probabilities on massively parallel computers. *J. Comput. Biol.* 7, 171–182. doi:10.1089/10665270050081441

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., et al. (2021). GENCODE 2021. *Nucleic Acids Res.* 49, D916–D923. doi:10.1093/nar/gkaa1087

Fukunaga, T., and Hamada, M. (2022). LinAlifold and CentroidLinAliFold: fast RNA consensus secondary structure prediction for aligned sequences using beam search methods. *Bioinforma. Adv.* 2, vbac078. doi:10.1093/bioadv/vbac078

Fukunaga, T., and Hamada, M. (2017). RIblast: an ultrafast RNA-RNA interaction prediction system based on a seed-and-extension approach. *Bioinformatics* 33, 2666–2674. doi:10.1093/bioinformatics/btx287

Fukunaga, T., Ozaki, H., Terai, G., Asai, K., Iwasaki, W., and Kiryu, H. (2014). CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biol.* 15, R16. doi:10.1186/gb-2014-15-1-r16

Garcia-Martin, J. A., and Clote, P. (2015). RNA thermodynamic structural entropy. *PLoS One* 10, e0137859. doi:10.1371/journal.pone.0137859

Girgis, H. Z., James, B. T., and Luczak, B. B. (2021). Identity: rapid alignment-free prediction of sequence alignment identity scores using self-supervised general linear models. *Nar. Genom Bioinform* 3, lqab001. doi:10.1093/nargab/lqab001

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. (2009). Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* 25, 465–473. doi:10.1093/bioinformatics/btn601

Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., et al. (2019). LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics* 35, i295–i304. doi:10.1093/bioinformatics/btz375

Ishida, R., Adachi, T., Yokota, A., Yoshihara, H., Aoki, K., Nakamura, Y., et al. (2020). RaptRanker: *in silico* RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information. *Nucleic Acids Res.* 48, e82. doi:10.1093/nar/gkaa484

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 49, D192–D200. doi:10.1093/nar/gkaa1047

Kawaguchi, R., and Kiryu, H. (2016). Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome. *BMC Bioinforma.* 17, 203. doi:10.1186/s12859-016-1067-9

Kiryu, H., and Asai, K. (2012). Rchange: algorithms for computing energy changes of RNA secondary structures in response to base mutations. *Bioinformatics* 28, 1093–1101. doi:10.1093/bioinformatics/bts097

Kiryu, H., Terai, G., Imamura, O., Yoneyama, H., Suzuki, K., and Asai, K. (2011). A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. *Bioinformatics* 27, 1788–1797. doi:10.1093/bioinformatics/btr276

Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., and Hoyer, S. (2021). Machine learning-accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2101784118. doi:10.1073/pnas.2101784118

Lange, S. J., Maticzka, D., hl, M., Gagnon, J. N., Brown, C. M., and Backofen, R. (2012). Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.* 40, 5215–5226. doi:10.1093/nar/gks181

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *IEEE CVPR*, 3431–3440.

Lorenz, R., Bernhart, S. H., ner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 26. doi:10.1186/1748-7188-6-26

Lu, Z. J., and Mathews, D. H. (2008). Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res.* 36, 640–647. doi:10.1093/nar/gkm920

Mann, M., Wright, P. R., and Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.* 45, W435–W439. doi:10.1093/nar/gkx279

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29, 1105–1119. doi:10.1002/bip.360290621

Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* 15, 469–479. doi:10.1038/nrg3681

Na, D., Lee, S., and Lee, D. (2010). Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst. Biol.* 4, 71. doi:10.1186/1752-0509-4-71

Nickel, M., and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. *NeurIPS* 30, 7668–7672.

Petrov, A. I., Kay, S. J. E., Kalvari, I., Howe, K. L., Gray, K. A., Bruford, E. A., et al. (2017). RNAcentral: A comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* 45, D128–D134. doi:10.1093/nar/gkw1008

Reuter, J. S., and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinforma.* 11, 129. doi:10.1186/1471-2105-11-129

Rizk, G., and Lavenier, D. (2009). GPU accelerated RNA folding algorithm. *ICCS* 9, 1004–1013.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. *MICCAI* 18, 234–241.

Salis, H. M. (2011). The ribosome binding site calculator. *Methods Enzymol.* 498, 19–42. doi:10.1016/B978-0-12-385120-8.00002-4

Sato, K., Akiyama, M., and Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* 12, 941. doi:10.1038/s41467-021-21194-4

Sato, K., and Hamada, M. (2023). Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. *Brief. Bioinform* 24, bbad186. doi:10.1093/bib/bbad186

Sun, Y., Goll, D. S., Huang, Y., Ciais, P., Wang, Y.-p., Bastrikov, V., et al. (2023). Machine learning for accelerating process-based computation of land biogeochemical cycles. *Glob. Chang. Biol.* 29, 3221–3234. doi:10.1111/gcb.16623

Terai, G., and Asai, K. (2020). Improving the prediction accuracy of protein abundance in *Escherichia coli* using mRNA accessibility. *Nucleic Acids Res.* 48, e81. doi:10.1093/nar/gkaa481

Um, K., Brand, R., Fei, Y. R., Holl, P., and Thuerey, N. (2020). Solver-in-the-loop: learning from differentiable physics to interact with iterative PDE-solvers. *NeurIPS* 33, 6111–6122.

Wayment-Steele, H. K., Kladwang, W., Strom, A. I., Lee, J., Treuille, A., Becka, A., et al. (2022). RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat. Methods* 19, 1234–1242. doi:10.1038/s41592-022-01605-0

Zheng, W., Yang, L., Genco, R. J., Wactawski-Wende, J., Buck, M., and Sun, Y. (2019). SENSE: siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics* 35, 1820–1828. doi:10.1093/bioinformatics/bty887

Zrimec, J., Fu, X., Muhammad, A. S., Skrekas, C., Jauniskis, V., Speicher, N. K., et al. (2022). Controlling gene expression with deep generative design of regulatory DNA. *Nat. Commun.* 13, 5099. doi:10.1038/s41467-022-32818-8