Supplementary Information

TransG4: An interpretable deep-learning approach for sequence-based G-quadruplex prediction

1 Details of Dataset

Table S1. Number of training, validation and test sets under K^+ and K^++PDS in G4-seq.

Dataset	Train	Valid	Test
K^+	311,736,885	31,356,341	29,534,138
K^++PDS	315,440,197	31,652,070	29,923,093

The human genome consists of 23 pairs of chromosomes, including 22 pairs of autosomes and one pair of sex chromosomes (XY). To train the model and select optimal hyperparameters, this study employed a chromosome-wise hold-out strategy for dataset partitioning: chromosome 1 was used as the test set, chromosome 2 as the validation set, and the remaining chromosomes for training. During training, all training data were shuffled to prevent any potential bias introduced by a specific data ordering.

The table above summarizes the number of samples in the training, validation, and test sets used in this experiment.

2 Impact of kernel size

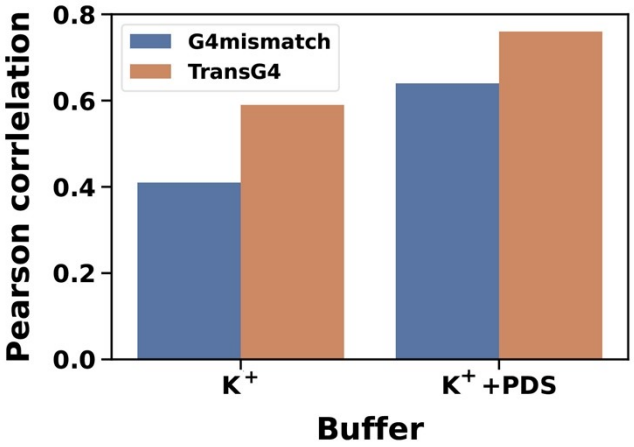
Table S2. Impact of kernel size on mismatch scores prediction under K^+.

Kernel size	MAE	RMSE	R^2	R
3	0.843	1.773	0.689	0.871
9	0.845	1.763	0.695	0.873
15	**0.837**	**1.763**	**0.702**	**0.873**

As shown in the table above, a kernel size of 15 yields better performance compared to kernel sizes of 3 and 9.

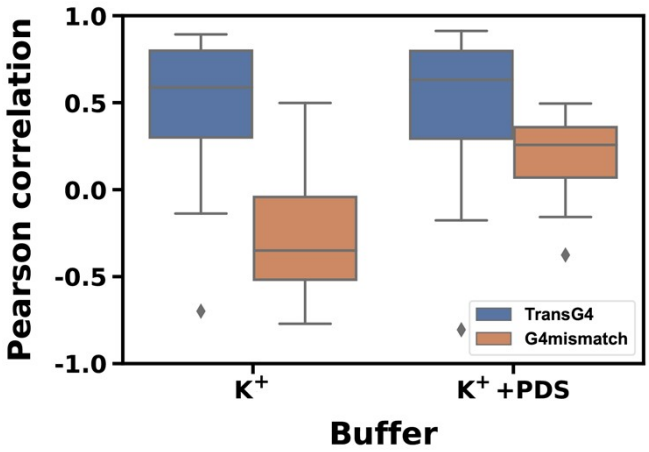
3 Supplementary Figures

Figure S1. The performance evaluation of G4 mismatch score predictions on G4-forming sequences data (the sequences fall under the PQS: $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$). TransG4 significantly outperforms G4mismatch under K^+ and K^++PDS.



The figure shows that under both K^+ and K^++PDS conditions, TransG4 achieves significantly higher Pearson correlation coefficients than G4mismatch, indicating that TransG4 performs better in predicting G4 mismatch scores.

Figure S2. Performance evaluation of TransG4 on an independent microarray dataset. Predictions generated by TransG4 show a strong correlation with PDS-binding intensities measured in the microarray experiment.



The box plot shows that under both K^+ and K^+ +PDS conditions, TransG4 exhibits consistently higher and more stable Pearson correlation coefficients, whereas G4mismatch shows weaker performance with greater variability.

5 Model result display

Table S3. 15 TransG4 inferred motifs under K^++PDS.

Entry	Inferred motif	Position	Mismatch	Predicted mismatch
chr1 10990	CGGGGGGAGGGTGGC	178	58.4	43.2
chr1 11065	CGGGGGGAGGGTGGC	103	47.4	48.7
chr1 12087	GGGAAAGATTGGAGG	184	35.7	20.3
chr1 13152	GCTGGAGAAGGGGAG	39	39.2	34.6
chr1 15697	GCAAGAGCAGGGGGT	94	52.4	49.5
chr1 15772	GCAAGAGCAGGGGGT	19	39.5	35.4
chr1 16028	CCGGGAGGTGGGGAA	26	42.4	39.6
chr1 16658	GTGGGGGCGGTGGGG	166	38.6	42.6
chr1 16733	GTGGGGGCGGTGGGG	91	51.1	51.5
chr1 18248	TGGGCAGCAGGGCAG	28	36.1	37.6
chr1 19318	AAGGGAGGGGGAGGA	193	40.4	22.4
chr1 19393	AAGGGAGGGGGAGGA	118	46.4	50.1
chr1 19468	AAGGGAGGGGGAGGA	43	36.3	41.4
chr1 19963	GGGGCAGTGGGAGGG	89	42.5	52.4
chr1 21013	CCGGGTGGTGGGGAG	110	36.5	42.2

Fifteen representative motifs inferred by TransG4 under K^++PDS conditions. For each motif, the genomic location, motif sequence, relative position, experimentally measured mismatch score, and the corresponding predicted mismatch score are reported.

6 Hardware and reproducibility

The model is trained using an NVIDIA RTX 2080ti graphics processing unit (GPU) (11GB) and an Intel(R) Xeon(R) Gold 6230 CPU.

All source code and preprocessed datasets are publicly available at:

https://github.com/M-nianjj/TransG4-deep-learning-approach-for-sequence-based-G-quadruplex-prediction

This repository includes scripts for environment setup, dataset processing, model fine-tuning, and evaluation.