Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TransformerCas9: an interpretable multi-scale Transformer with hierarchical attention for CRISPR/Cas9 sgRNA activity prediction

Qingling Song
Chongqing Medical University
Jiansheng Li
Chongqing Medical University
Haisong Liu
Chongqing Medical University
Liuzhi Chen
Chongqing Medical University
Ziyu Zhai
Chongqing Medical University
Jiaju Qi
Chongqing Medical University
Jie Cao
Bishan Hospital Affiliated to Chongqing Medical University
Baoru Han
Chongqing Medical University
Feng Liu
fengliu@cqmu.edu.cn

Chongqing Medical University

Research Article

Keywords: CRISPR/Cas9, sgRNA activity prediction, Transformer

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11574502/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TransformerCas9: an interpretable multi-scale Transformer with hierarchical attention for CRISPR/Cas9 sgRNA activity prediction

Qingling Song[1, 2, 3], Jiansheng Li[2, 3], Haisong Liu[2], Liuzhi Chen[2], Ziyu Zhai[2], Jiaju Qi[2], Jie Cao[1, *], Baoru Han[2, *], Feng Liu[1, 2, *]

[1] Bishan Hospital Affiliated to Chongqing Medical University, Chongqing 402760, China.

[2] College of Artificial Intelligence Medicine, Chongqing Medical University, Chongqing 400016, China.

[3] These authors contributed equally: Qingling Song, Jiansheng Li.

[*] To whom correspondence should be addressed.

Abstract

Background: Accurate prediction of single-guide RNA (sgRNA) activity is critical for effective CRISPR/Cas9 genome editing. Despite substantial advances in computational modeling, existing approaches remain limited by suboptimal predictive performance, restricted generalizability across diverse Cas9 variants, and insufficient interpretability.

Results: We present TransformerCas9, an interpretable Transformer-based framework that integrates multi-scale K-mer encoding with a hierarchical attention mechanism to simultaneously capture local sequence features and long-range contextual dependencies,

offering a new framework for accurate sgRNA activity prediction. Systematic analyses of the contributions of key architectural components and biological features further demonstrate the effectiveness of the proposed framework. On the HT_Cas9_Test benchmark dataset, TransformerCas9 achieves a Spearman correlation coefficient of 0.788 and a Pearson correlation coefficient of 0.790, surpassing all evaluated baseline models. Furthermore, across seven datasets encompassing wild-type, high-fidelity, and PAM-flexible SpCas9 variants, TransformerCas9 consistently outperforms six state-of-the-art approaches, achieving the highest performance on five datasets and the best overall ranking among all evaluated methods. Interpretability analyses leveraging intrinsic attention weights, Integrated Gradients attribution paired with in silico perturbation, and complementary explanation approaches consistently highlight biologically relevant sequence determinants associated with sgRNA activity, supporting the biological plausibility of the learned representations.

Conclusions: **O**ur work introduces TransformerCas9 as an accurate, robust, and interpretable framework for sgRNA activity prediction, which adopts multi-scale K-mer encoding, multi-head self-attention mechanism and a tailored MLA module to learn comprehensive sequence representations, providing a practical tool for CRISPR guide RNA design.

Keywords: CRISPR/Cas9, sgRNA activity prediction, Transformer

Contact: 840331cj@163.com; baoruhan@cqmu.edu.cn; fengliu@cqmu.edu.cn

Background

The CRISPR/Cas9 system has emerged as a revolutionary tool for genome editing [1] due to its exceptional targeting precision [2, 3], demonstrating significant potential across

various applications including disease therapeutics [4–7], agricultural breeding [8–11], and other domains [12, 13]. The success of CRISPR/Cas9 largely hinges on the design efficacy of sgRNAs [14], which determine both the specificity and activity of direct DNA recognition and cleavage. Consequently, the accurate prediction of sgRNA activity has become a critical challenge in the field.

Despite substantial efforts in leveraging machine learning and deep learning to predict sgRNA activity, existing approaches are still constrained by limited feature representation, insufficient architectural innovation, and "black-box" nature, resulting in suboptimal prediction accuracy, poor generalizability, and a lack of internal model interpretability. Existing approaches rely on integrating hand-crafted biophysical properties [15, 16] to augment sequence representation, a process that introduces external preprocessing pipelines and computational redundancy, leading to unnecessary resource overhead. Architecturally, traditional models such as convolutional neural networks (CNNs) [17] and recurrent neural networks (RNNs) [18] may be less direct in modeling long-range dependencies than self-attention based frameworks within sgRNA sequences [19, 20]. Moreover, the "black-box" nature of deep learning models hampers their applicability in uncovering biological mechanisms [21–23].

To address these limitations, we proposed TransformerCas9, a novel multi-scale deep learning model based on the Transformer [24] framework. Our approach is characterized by three key features: (i) a pure sequence-driven paradigm, where TransformerCas9 operates directly on raw nucleotide sequences without requiring pre-defined biophysical or structural features, as empirical evaluations demonstrate that incorporating external physical features yields only marginal performance gains; (ii) a hybrid K-mer encoding

with multi-level attention, which leverages multi-scale nucleotide tokenization [25] and hierarchical attention to simultaneously capture local base patterns and global sequence dependencies; and (iii) intrinsic interpretability, achieved by leveraging its multi-level attention mechanism to extract global attention weights, thereby providing a transparent and intrinsically interpretable perspective on sequence-activity relationships.

We demonstrated the performance of TransformerCas9 in predicting sgRNA activities across benchmark evaluations on the HT_Cas9_Test [26] dataset. TransformerCas9 outperformed ten machine learning models and four conventional deep learning models with a Spearman correlation coefficient (R) of 0.788 and a Pearson correlation coefficient (r) of 0.790. We found that sgRNAs with a higher TransformerCas9 editing score resulted in substantially higher activities (3.52-fold) compared with those with lower scores. These results highlight the superior capacity of the TransformerCas9 framework in learning the complex relationships between sgRNA sequence variations and editing activities. Furthermore, we validated the broad applicability and superior performance of TransformerCas9 across seven datasets encompassing three distinct categories: wild-type *Streptococcus pyogenes* Cas9 (SpCas9), engineered high-fidelity variants, and non-canonical protospacer adjacent motifs (PAMs). Notably, when compared against six state-of-the-art (SOTA) models, TransformerCas9 achieved the highest R in five out of the seven datasets and ranked first in average R among all evaluated models, demonstrating its broad applicability capability across diverse SpCas9 variants. Furthermore, interpretability analyses reveal that activity is influenced by purine-rich regions and seed region bases, consistent with known biological insights. Our work provides an effective computational tool for sgRNA design and offers a transparent framework for characterizing the

sequence-activity relationships of CRISPR/Cas9 systems. These findings contribute to the advancement of interpretable artificial intelligence into the field of genome engineering.

Methods

TransformerCas9 Framework Design

To predict the activity of sgRNA sequences, we developed TransformerCas9, a nucleotide-level language model. The overall framework of TransformerCas9 is organized into a progressive computational pipeline. The process begins with the K-mer (K = 1, 2, 3) embedding and positional encoding, which preprocesses raw nucleotide sequences into structured, multi-scale representations across different K-mer scales. Subsequently, these representations are fed into a stack of weight-shared Transformer Encoder Modules to capture long-range dependencies across multiple representational subspaces via multi-head self-attention. To further distill and integrate these features, a tailored Multi-Level Attention (MLA) Module is employed to hierarchically aggregate local functional motifs and global inter-scale patterns. Finally, the Fully Connected Network (FCN) Module maps the comprehensive feature representation to a continuous activity score, quantifying the expected activity of sgRNA. The detailed architecture and operational mechanisms of each module are described in the following.

Embedding and Positional Encoding

Input sequences are first fed into the K-mer Encoding (KE) Module, where sequence data are transformed into embedding matrices. Specifically, the module employs a customized multi-scale K-mer encoding strategy with K = 1, 2, and 3. For an input sequence of length M, it is segmented into overlapping K-mers of length k, resulting in subsequences of length

$M - k + 1$.

Each K-mer is then mapped into a d-dimensional vector space through a learnable embedding matrix. To encode positional information within the sequence, we adopt sinusoidal positional encoding based on the following formula:

$$PE_{(pos,\,2i)} = \sin\left(pos \,/\, 10000^{2i\,/\,d}\right)$$
$$PE_{(pos,\,2i+1)} = \cos\left(pos \,/\, 10000^{2i\,/\,d}\right)$$

where pos denotes the position of a nucleotide, i is the embedding dimension index, and d is the embedding size. This encoding scheme enables the model to capture absolute position information, enhancing its sequence-awareness.

Encoder Module

The encoded K-mer representations are passed through stacked Transformer encoder blocks for feature extraction. Each encoder block includes a multi-head self-attention mechanism, computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent the query, key, and value matrices. Each attention layer is followed by a residual connection and layer normalization:

$$\text{Output} = \text{LayerNorm}(X + \text{Attention}(Q, K, V))$$

This design enhances gradient flow and training stability. Each encoder also contains a position-wise feedforward network (FFN):

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

which models nonlinear interactions and restores the original dimensionality.

MLA Module

To simultaneously capture local dependencies within sequences and global correlations across different encoding scales, we designed a hierarchical MLA Module. The module integrates multi-scale features through K-mer level attention and multi-scale attention [27] to dynamically filter and fuse representations from the three independent K-mer branches (K = 1, 2, 3).

First, for each individual K-mer branch, the sequence feature matrix—after being processed by the respective Transformer encoder layers—is fed into a K-mer level attention pooling layer. This stage computes position-specific importance weights along the sequence dimension to emphasize critical functional motifs. Given the encoded feature matrix $X \in \mathbb{R}^{B \times L \times D}$ from a specific K-mer branch, the attention weights α along the sequence dimension and the resulting feature representation F_k (k = 1, 2, 3) are computed as:

$$\alpha = \text{softmax}(XW_\alpha)$$

$$F_k = \sum_i \alpha_i x_i$$

where X represents the sequence feature matrix, W_α is a learnable projection matrix, and x_i denotes its feature vector at position i.

Next, the module performs cross-scale attention fusion to capture inter-scale patterns. These pooled representations are stacked into a unified multi-scale matrix F_{stack} and processed by a multi-scale attention mechanism:

$$F_{\text{stack}} = \text{stack}(F_1, F_2, F_3)$$

$$\beta = \text{softmax}(F_{\text{stack}} W_\beta)$$

$$F_{output} = \sum_{k=1}^{3} \beta_k F_k$$

where W_β is a trainable fusion weight matrix, β denotes the adaptive attention weights across different K-mer scales, and F_{output} is the fused multi-scale representation. This hierarchical design allows the model to adaptively balance fine-grained mononucleotide features with broader multi-nucleotide contextual interactions.

FCN Moudle

The fused features from MLA Module are passed to a fully connected neural network (FCN) for regression. The FCN comprises several linear layers with ReLU activations and Dropout (rate = 0.1) to mitigate overfitting. The final prediction is computed as:

$$y = W_n(\ldots \mathrm{ReLU}(W_2 \mathrm{ReLU}(W_1 x + b_1) + b_2)\ldots) + b_n$$

where x is the MLA Module output and W_i, b_i (for i in $\{1, ..., n\}$) are trainable parameters.

Datasets and Preprocessing

We curated and integrated eight publicly available high-throughput CRISPR/Cas9 datasets to comprehensively train and evaluate TransformerCas9. The benchmark dataset was derived from the study by Kim et al. 2019 [26], containing SpCas9 activity data generated in HEK293T cells, comprising: HT_Cas9_Train (n = 12,832), HT_Cas9_Test (n = 542).

To evaluate the generalization performance of our model across various SpCas9 variants, we employed seven datasets. These datasets were categorized based on the engineered properties of the nucleases: (i) the wild-type SpCas9 (WT-SpCas9, n = 55,604) [28]; (ii) engineered high-fidelity variants targeting standard NGG PAMs, including Sniper-Cas9 (n = 37,794) [29, 30], SpCas9-HF1 (n = 56,888; 23 nt) [28], eSpCas9 (n =

58,617) [28] and HypaCas9 (n = 37,853) [30]; and (iii) PAM-flexible variants targeting non-canonical PAMs, including SpCas9-NG (n = 30,585) [31, 30], xCas9 (n = 37,738) [32, 30]. Target site sgRNA sequence overlap between HT_Cas9_Train and all test datasets was evaluated using Hamming distance[33]. Sequences with >90% sequence similarity to any training sample were considered as training-test overlap (Additional file: Table S1). All datasets represent sgRNAs paired with corresponding indel frequencies. To ensure consistency across different scales, activity values were normalized to the [0, 1] range using min–max normalization[34] as follows:

$$Y_{\mathrm{norm}_i} = \frac{Y_i - Y_{\min}}{Y_{\max} - Y_{\min}}$$

To evaluate the importance of biological features, melting temperature (Tm), secondary structure, minimum free energy (MFE), and GC content [28, 35] were calculated using the ViennaRNA package and concatenated with sequence representations prior to model input. However, these features were introduced exclusively for performance comparison and are omitted from the final TransformerCas9 framework, thereby preserving the computational efficiency and simplicity of a purely sequence-based inference process.

Evaluation Metrics

To quantitatively assess the performance of TransformerCas9, we employed five widely used evaluation metrics: R, r, the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean squared error (RMSE) [36, 37]. The correlation-based metrics (R and r) measure the monotonic and linear relationships, respectively, between the predicted and experimentally measured activities. R^2 represents the proportion of variance

explained by the model, and the error-based metrics (MAE and RMSE) quantify the absolute prediction errors. All five metrics were calculated under a strictly partitioned data regime to avoid information leakage. As representative metrics, the formulas for R and r were defined as follows:

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where d_i is the difference in ranks between the predicted and actual values for the *i-th* sgRNA, and n is the number of samples; x_i and y_i represent the predicted and actual activities, $\bar{x}$ and $\bar{y}$ are their respective means. These five metrics were selected to comprehensively reflect the rank-order agreement (R), linear correlation (r), prediction variance fit (R^2), and absolute value proximity (MAE and RMSE) between model predictions and ground truth.

Comparison of TransformerCas9 with baseline models and SOTA models

To facilitate comprehensive comparisons, we implemented ten machine learning algorithms and four conventional deep learning models as baselines, and evaluated their performance against TransformerCas9. The machine learning baselines comprised Support Vector Machine (SVM) [38], Multi-layer Perceptron (MLP) [39], K-Nearest Neighbors (KNN) [40], Decision Tree (DT) [41], Random Forest (RF) [42], Gradient Boosting Decision Tree (GBDT) [43], AdaBoost [44], XGBoost [45], LightGBM [46], and CatBoost [47]. The conventional deep learning baselines included CNN, RNN, Long Short-Term Memory (LSTM) [48], and a standard Transformer. Each of these fourteen baseline

methods was trained on HT_Cas9_Train and evaluated on the corresponding HT_Cas9_Test. All training pipelines, hyperparameter optimization protocols, and evaluation metrics were kept strictly identical across all methods to isolate the performance impact of the models themselves.

Furthermore, we compared TransformerCas9 with six SOTA models. Among them, TransCrispr [35] and DeepHF [28] were fully evaluated in their original studies. In contrast, CRISPR-ONT [49], AttnToCrispr [50], and C-RNNCrispr [20] lacked WT-SpCas9 evaluations, and DeepSpCas9 [26] lacked both WT-SpCas9 and HypaCas9 evaluations. To complete these evaluations, we used their official pre-trained models and evaluation scripts to test them on the respective test sets.

Implementation Details

Our model was implemented using PyTorch (v1.12.1) under Python (v3.10.0), with computational tasks executed on four NVIDIA RTX A6000 GPUs (48 GB VRAM per GPU). To ensure fairness and reproducibility, the baseline framework was uniformly fixed across all experiments, including a fixed network backbone (32-dimensional embeddings and 2,048-dimensional feed-forward networks), the MSE loss function, the Adam optimizer, a weight decay of 0.1, with the random seed fixed at 1. Under this framework, hyperparameters were systematically optimized via a randomized grid search on the primary benchmark dataset (HT_Cas9_Train). The complete candidate search spaces, alternative values, and the final optimal configuration utilized for evaluating the HT_Cas9_Test dataset—including a batch size of 2,048 and a learning rate of 0.007—are comprehensively documented in Additional file: Table S2.

Statistical Analysis

To ensure the robustness of our results, the data presented in the figures as mean ± standard deviation (s.d.) were generated from five independent experimental runs, each initialized with a distinct random seed. Two-tailed paired t-tests were utilized to evaluate statistical differences across these runs. For comparisons between two groups, statistical significance was evaluated using the Mann–Whitney U test. For all analyses, statistical significance was defined as $p < 0.05$.

RESULTS

Architectural Design of TransformerCas9

To enhance the accuracy and interpretability of sgRNA activity prediction, we developed TransformerCas9, a deep learning-based computational model inspired by natural language processing techniques. The model comprises three key components: input, feature learning, and regression prediction (Fig. 1). The K-mer Encoding (KE) Module serves as the input component, which preprocesses input sgRNA sequences by performing nucleotide embedding at various K-mer scales (K = 1, 2, 3) and incorporating positional encoding, transforming raw sequences into structured representations optimized for downstream learning. The core of the framework lies in the feature learning component, which employs multiple weight-shared Transformer Encoder Modules combined with an MLA Module to automatically learn comprehensive and interpretable representations of sgRNA sequences. Specifically, the Encoder Module leverages a multi-head self-attention mechanism to capture long-range dependencies between nucleotides within the input sequence and enables the model to attend to the input sequence from multiple representational subspaces in parallel, thereby enhancing the model's capacity to detect complex interaction patterns among nucleotides. To deepen sequence representation and

improve interpretability, we designed a tailored MLA Module wherein a K-mer Level Attention (KLA) block yields distinct feature weight vectors for each K-mer scale, which are then stacked and blended with a global sequence weight vector via a Multi-Scale Attention (MSA) block to generate a comprehensive sgRNA representation, encoding both fine-grained and global sequence information. The FCN Module performs the final regression step, utilizing multiple fully connected feedforward layers to map the learned sequence representation to an activity score.

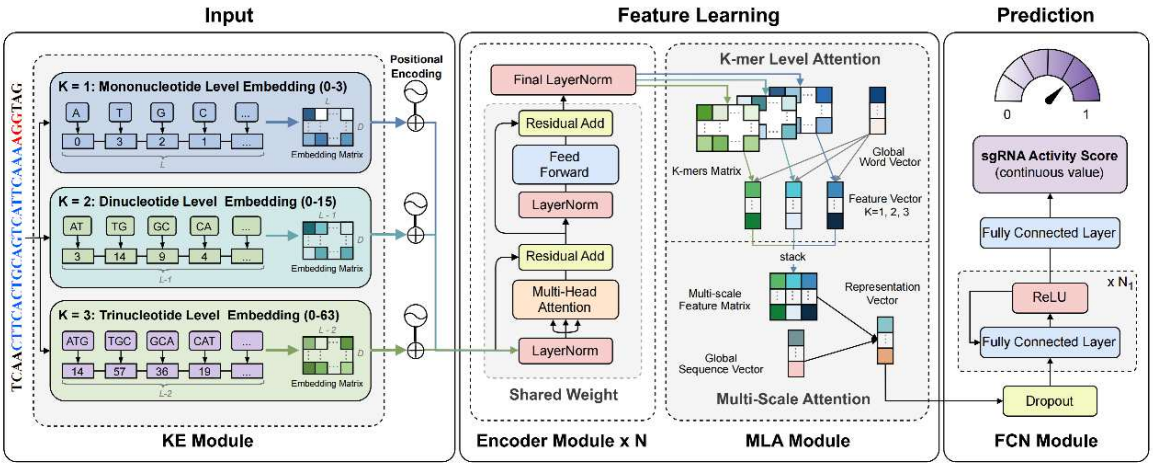


Fig. 1. | Overview of the model framework. The framework consists of three components: Input, Feature Learning, and Prediction. **Input Component**: sgRNA sequences are processed by the KE Module through multi-scale K-mer embeddings: mononucleotide ($K = 1$), dinucleotide ($K = 2$), and trinucleotide ($K = 3$). Each scale yields a feature matrix of dimension L_i x D (L_i, sequence length at scale i; D, embedding dimension), integrated with positional encodings. The 30-nucleotide sgRNA sequence comprises the 4 nt upstream of the sgRNA target site (black), the 20 nt protospacer (blue), the 3 nt PAM (red), and the 3 nt downstream of the sgRNA target site (black); **Feature Learning Component**: A stack of N Encoder Modules with shared weights; the MLA Module comprises K-mer Level Attention and Multi-Scale Attention; **Prediction Component**: The FCN Module contains N_1 hidden fully connected layers that map the representation vector to a continuous activity score (range 0 – 1).

Ultimately, this score serves as the quantitative prediction of the indel frequency associated with each sgRNA–Cas9 complex.

Individual Contributions of Key Modules and Bio-Features

To systematically evaluate the individual contributions of the KE and MLA Modules, we assessed the effects of the multi-scale K-mer encoding strategy and multi-level attention mechanism. Focusing on the KE Module, the single-nucleotide baseline ($K = 1$) achieved a mean R of 0.736 on the HT_Cas9_Test dataset (Fig. 2a). Incorporating dinucleotide or trinucleotide features yielded a 2.4% increase ($K = 1, 2$; mean $R = 0.754$) and 2.2% increase ($K = 1, 3$; mean $R = 0.752$) over the baseline, respectively. Remarkably, the full K-mer features ($K = 1, 2, 3$) achieved mean R of 0.785, representing a statistically significant 6.7% increase over the baseline. We further validated the contribution of the KE Module across two independent datasets, WT-SpCas9 and SpCas9-HF1. Importantly, compared to the respective $K = 1$ baselines, we observed that the full K-mer features strategy achieved mean R gains of 18.3% (WT-SpCas9) and 8.3% (SpCas9-HF1) (Fig. 2b). These results confirm that hierarchical integration of multi-scale patterns is essential for capturing the complex contextual dependencies governing sgRNA activity.

When assessing the contribution of each attention level, we compared four architectural configurations: a baseline Transformer, its counterparts augmented with KLA or MSA alone, and the full MLA mechanism (Fig. 2c). The baseline Transformer achieved $R = 0.736$. Adding KLA alone improved performance by 3.1% ($R = 0.759$). In comparison, incorporating MSA alone yielded a 2.3% improvement over the baseline ($R = 0.753$). Notably, the model's performance was significantly improved when both attention mechanisms were jointly integrated, yielding a 6.7% increase over the baseline ($R = 0.785$). This result shows that KLA and MSA are mutually reinforcing. The latter captures cross-K-mer dependencies across broader sequence contexts, while the former refines local feature selection. Similarly, the full MLA-equipped Transformer outperformed the

baseline, improving *R* by 18.3% for WT-SpCas9 and 15.5% for SpCas9-HF1, respectively (Fig. 2d). These results demonstrate that the performance gains are not specific to a single experimental context but reflect the model's robust capacity to represent sequence-activity relationships.

To investigate the impact of explicit biological features, we extended the model to incorporate biological priors by introducing a bio-feature extraction module and evaluated its performance against the base TransformerCas9 model across three

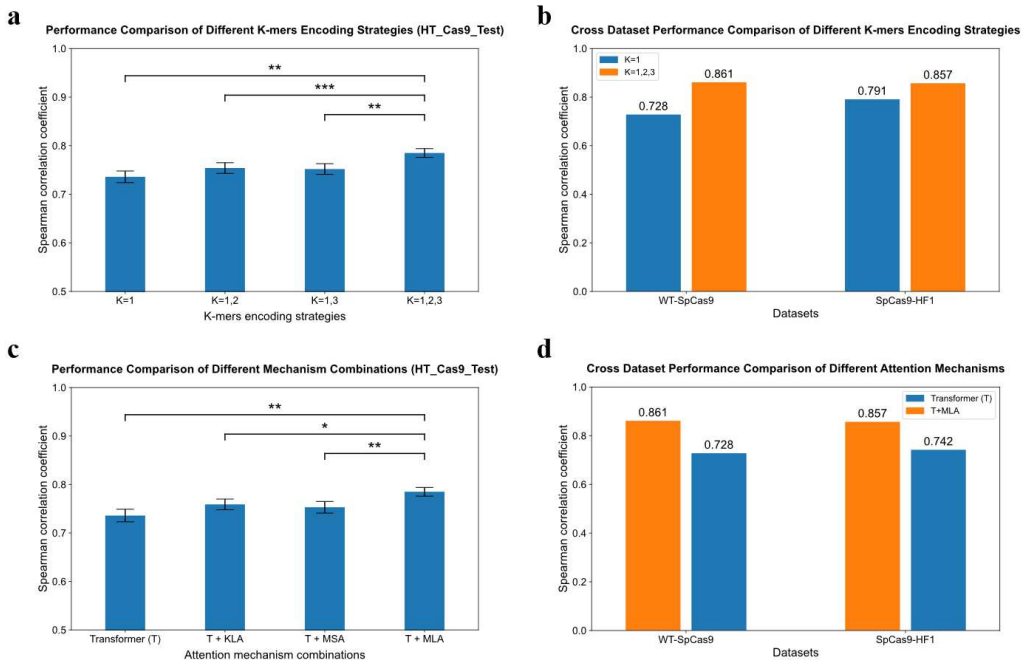


Fig. 2. Validation of the multi-scale encoding and hierarchical attention within the TransformerCas9 framework. a-b, Performance comparison of different K-mer encoding strategies evaluated on (**a**) the HT_Cas9_Test dataset (n = 542) and (**b**) across WT-SpCas9 (n = 55,604) and SpCas9-HF1 (n = 56,888) datasets (K = 1 vs. K = 1, 2, 3). **c-d**, Performance comparison of different attention mechanisms evaluated on (**c**) the HT_Cas9_Test dataset and (**d**) across WT-SpCas9 and SpCas9-HF1 datasets (Transformer [T] vs. T+MLA). For **a** and **c**, data represent mean ± s.d. of n = 5 independent replicates; *p < 0.05, **p < 0.01, ***p < 0.001 (two-tailed paired t-tests with Bonferroni correction).

independent datasets. The bio-feature extraction module enables the model to go beyond pure sequence-based inputs by integrating widely adopted, biologically meaningful features, such as Tm, secondary structure, MFE, and GC content. However, the integration

of these features yielded only marginal improvements in predictive accuracy (Additional file: Fig. S1). The median *R* increased by a mere 0.23% (0.861 vs. 0.863) for WT-SpCas9, while it increased by 0.58% (0.857 vs. 0.862) and 0.82% (0.851 vs. 0.858) for the high-fidelity SpCas9-HF1 and eSpCas9, respectively. These negligible gains underscore that TransformerCas9 autonomously captures essential latent biological representations from raw sequences, rendering manual feature engineering largely redundant.

Evaluation of TransformerCas9 Predictive Capacity

To systematically evaluate the predictive capability of TransformerCas9 on sgRNA activity, we first assessed its performance on the HT_Cas9_Test dataset. The model exhibited high concordance between predicted and experimentally measured activities, achieving *R* of 0.788 and *r* of 0.790 (Fig. 3a). The robust predictive performance was further evidenced by R^2 of 0.512, alongside MAE of 11.333% and RMSE of 14.300%. These results indicate that TransformerCas9 can reliably capture the variance in CRISPR/Cas9 activity within complex genomic contexts.

Next, we assessed the model's potential as a design tool by examining its capacity to distinguish high-activity sgRNAs [36, 51]. Using a predicted activity threshold of 15%, we observed that sgRNAs with higher predicted editing scores (≥ 15%) consistently exhibited significantly higher experimental activities compared to those with lower predicted scores (< 15%). Specifically, the high-score group displayed a 3.52-fold increase in activity relative to the low-score group, as confirmed by the Mann–Whitney U test ($p < 0.001$) (Fig. 3b), demonstrating the practical utility of TransformerCas9 in sgRNA design.

To further evaluate its performance through a comparative analysis, we compared TransformerCas9 against 14 established baseline models across both classical machine

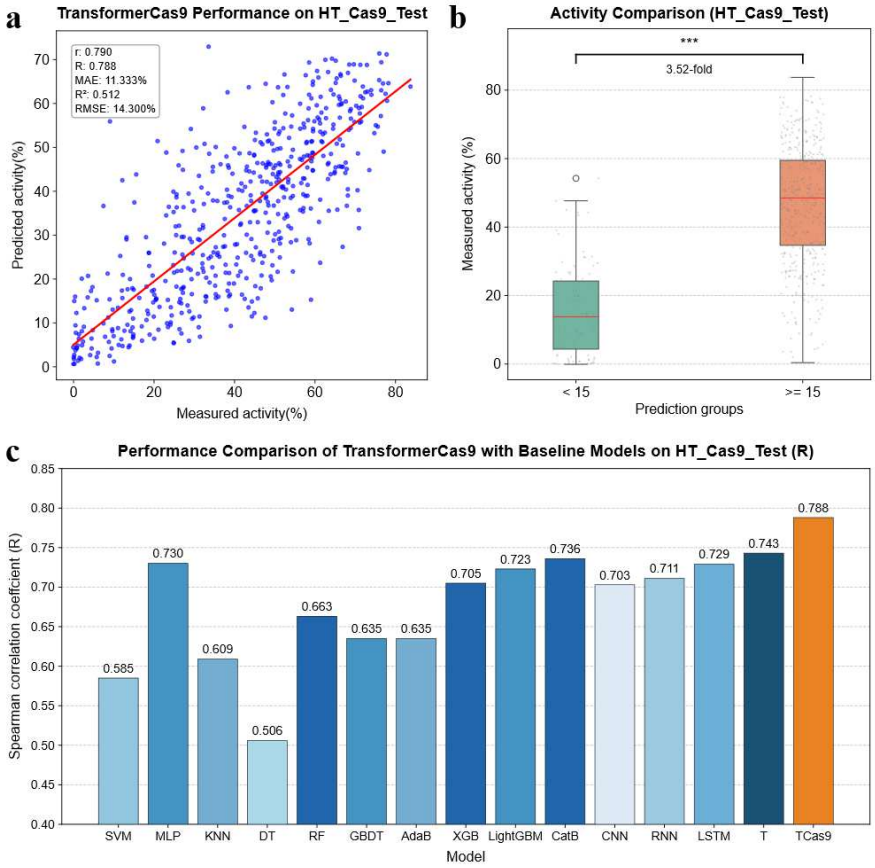


Fig. 3. Benchmarking TransformerCas9 on sgRNA activity prediction. a, Predictive accuracy of TransformerCas9 on the HT_Cas9_Test dataset (n = 542) assessed by multiple evaluation metrics (r, R, R^2, MAE, RMSE). The red solid line in plot represents the linear regression line. **b**, Comparison of measured activities between sgRNAs predicted to have high (≥ 15%) and low (< 15%) editing scores. (***$p < 0.001$, Mann–Whitney U test). The red horizontal line within each box indicates the median; the box spans the interquartile range (IQR, 25th–75th percentile); whiskers extend to 1.5× IQR; and circles beyond the whiskers represent outliers. Overlaid gray dots represent individual sgRNA measurements. **c**, Performance comparison of TransformerCas9 with various traditional machine learning and deep learning models on the same dataset. (AdaB: AdaBoost; XGB: XGBoost; CatB: CatBoost; T: Transformer; TCas9: TransformerCas9)

learning and deep learning frameworks. TransformerCas9 consistently exhibited superior predictive power, achieving the highest scores in both correlation coefficient metrics. Notably, R represented a relative improvement of 6.0%-55.7% over all baseline models (Fig. 3c), while r exceeded all comparators by 5.3%-52.2% (Additional file: Fig. S2). These advancements are attributed to TransformerCas9's hybrid architectural design,

which effectively integrates the temporal modeling capacity of LSTM (R = 0.729), and the global context awareness of Transformer's self-attention mechanism (R = 0.743). This synergy enables accurate modeling of the nonlinear biological dependencies underlying sgRNA-Cas9 activity (Methods).

Broad applicability of TransformerCas9 across SpCas9 variants

To systematically evaluate the broad applicability of TransformerCas9, we assessed its performance across seven SpCas9 datasets, including wild-type SpCas9 (WT-SpCas9), four engineered high-fidelity variants (Sniper-Cas9, SpCas9-HF1, eSpCas9, and HypaCas9) and two PAM-flexible variants (SpCas9-NG and xCas9) (Methods). The TransformerCas9 framework achieved strong predictive performance across all datasets, with R values ranging from 0.846 to 0.941, yielding a high overall mean of 0.890 (Fig. 4a). Additionally, the model's high predictive accuracy was further confirmed across other metrics, with average values of 0.918 for r, 0.833 for R^2, 6.005% for MAE, and 8.537% for RMSE (Additional file: Fig. S3a-d). These results indicate that the model performed robustly on both standard NGG-targeting and non-canonical PAM-targeting variants, demonstrating broad applicability across diverse engineered nucleases. The individual alignment of predictions and experimental values for each variant is further visualized in the detailed scatter plots (Additional file: Fig. S4). Furthermore, applying the 15% predicted activity threshold across these seven datasets reliably stratified sgRNAs into high- and low-scoring groups. Importantly, the high-score groups exhibited consistently and significantly elevated experimental activities, with fold changes ranging from 3.59 to 7.10 ($p < 0.001$) (Additional file: Fig. S5). These findings suggest that TransformerCas9 successfully captures distinct biophysical determinants governing different SpCas9

variants, enabling its broad application across diverse engineered nucleases with varying fidelity profiles and PAM specificities.

We further performed a comparative analysis against six SOTA models (TransCrispr, CRISPR-ONT, DeepHF, AttnToCrispr, C-RNNCrispr, and DeepSpCas9) across the same

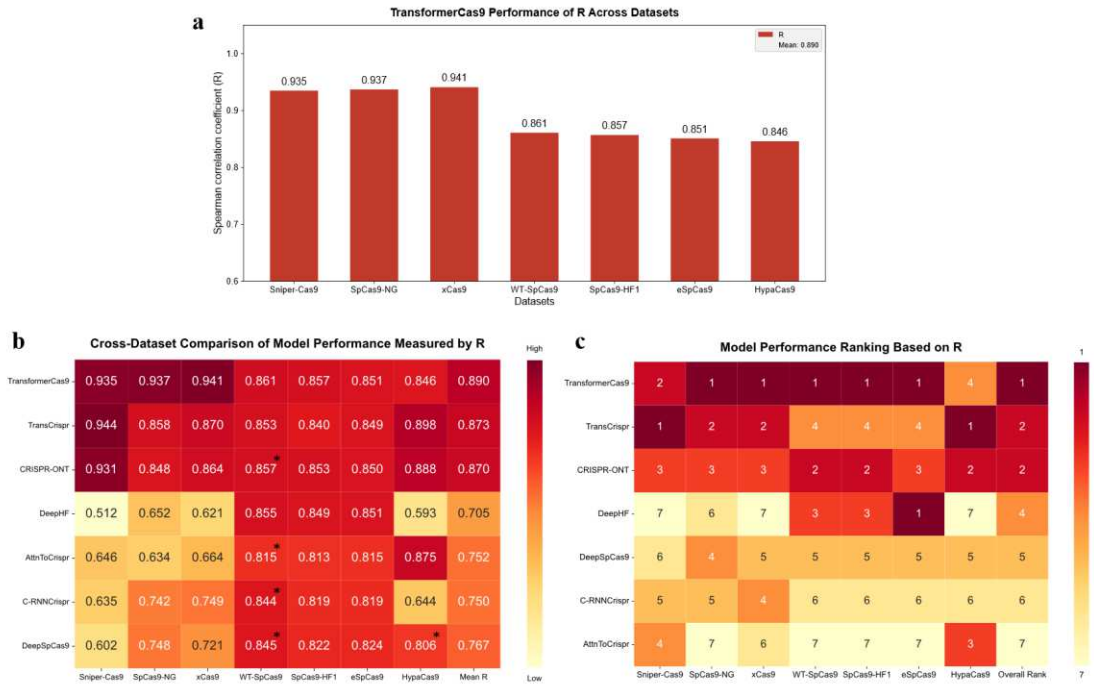


Fig. 4. Evaluation of the applicability of TransformerCas9 against SOTA deep learning models across SpCas9 variants. a, Spearman correlation coefficient (*R*) achieved by TransformerCas9 across seven datasets. **b**, Comparison of *R* between TransformerCas9 and six representative deep learning models across seven datasets. The rightmost column shows the mean *R* across datasets of each model. Asterisks (*) indicate reproduced results in this study. (Methods) **c**, Model rankings derived from *R* on each dataset. Rank 1 corresponds to the best-performing model. The overall rank was calculated as the average rank across all datasets.

seven datasets (Fig. 4b). TransformerCas9 consistently achieved strong performance across most datasets, achieving the highest *R* on five datasets (SpCas9-NG, xCas9, WT-SpCas9, SpCas9-HF1 and eSpCas9) and competitive scores on Sniper-Cas9. Specifically, TransformerCas9 outperformed the runner-up (TransCrispr) on SpCas9-NG (relative improvement of 9.2%; $R = 0.937$) and xCas9 (relative improvement of 8.2%; $R =$

0.941), with even more pronounced improvements of approximately 25% over DeepSpCas9 and C-RNNCrispr. Overall, TransformerCas9 achieved the highest mean *R* across all seven datasets, representing a 26.2% improvement over DeepHF and an 18.4% improvement over AttnToCrispr (Additional file: Fig. S3e). This systematic superiority is further supported by the overall model ranking (Fig. 4c), where TransformerCas9 achieved the lowest mean rank of 1.57 (standard deviation = 1.13), whereas TransCrispr and CRISPR-ONT followed as joint-second, and DeepHF and AttnToCrispr exhibited limited generalizability. Collectively, these results indicate the robust representation capacity of our proposed Transformer-based framework, demonstrating its generalizability and potential utility in both CRISPR guide RNA design and protein engineering contexts across diverse SpCas9 variants.

Intrinsic interpretability of TransformerCas9

To uncover the sequence features learned by TransformerCas9, we first extracted K-mer level attention weights from the multi-level attention mechanism embedded in the model. Notably, the learned attention patterns were strongly aligned with established biological mechanisms of the CRISPR/Cas9 system. The model assigned significantly higher attention weights to the seed region (positions 16–20), identifying these PAM-proximal positions as critical determinants of on-target activity (Fig. 5a, Additional file: Table S3). The most impactful nucleotide resides at position 20, where the guanine (G) is strongly preferred and cytosine (C) is disfavored, a pattern that reflects an overall purine bias at this site (Fig. 5b). Additional positional constraints were captured by the model, specifically an adenine (A) bias at position 16 and a preference for C at position 18. This observation is consistent with previously reported sgRNA activity rules and structural

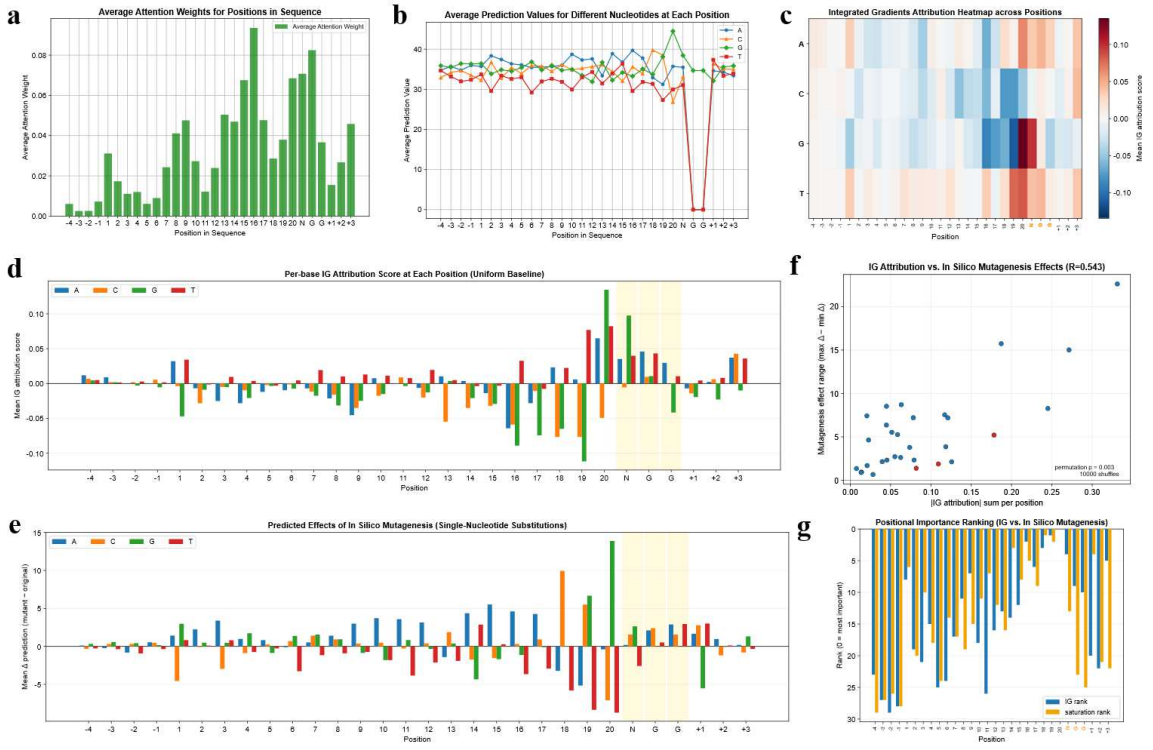


Fig.5. Interpretability analysis results of TransformerCas9. The 30-nucleotide sgRNA sequence comprises the 4 nt upstream of the sgRNA target site (position: -4 to -1), the 20 nt protospacer (1 to 20), the PAM (N, G, G), and the 3 nt downstream of the sgRNA target site (+1 to +3). **a**, Average attention weights of the Multi-Level Attention (MLA) Module across all sequence positions. **b**, Average predicted activity (%) for each nucleotide type across all positions, with color-coded lines representing A (blue, circle), C (orange, triangle), C (green, diamond), and T (red, square). **c-d**, Integrated Gradients (IG) attribution analysis ($K = 1$ branch, uniform baseline). **c**, Heatmap of IG attribution scores across sequence positions and nucleotide types. **d**, Per-base IG attribution scores at each position, with positive and negative values denoting feature promotion and inhibition, respectively. (left to right on the horizontal line of each site are A, C, G, T). **e**, Predicted effects of in silico single-nucleotide substitutions. **f-g**, Comparison of positional importance between IG attribution and in silico mutagenesis (ISM). **f,** Scatter plot showing the correlation between IG-derived importance and ISM effect range. **g**, Positional importance ranking comparison between IG attribution and in silico mutagenesis (where rank 0 denotes the most important position).

constraints of the Cas9-R-loop complex [52–55].

Next, we integrated Integrated Gradients (IG) [56, 57] attribution with DNA sequence perturbation via in silico saturation mutagenesis (ISM) [58] to further dissect positional sensitivity at the nucleotide level. Both methods consistently demonstrated that the core seed region exhibits high sensitivity to single-nucleotide variations (Fig. 5c-d),

highlighting robust substitution responses at key functional sites (e.g., G preference at position 20) (Fig. 5e). Finally, cross-validation between IG attribution and ISM demonstrated a statistically significant positive correlation in positional importance (Fig. 5f, R = 0.543, $p < 0.05$), alongside largely consistent positional rankings (Fig. 5g). Together, these results demonstrate that TransformerCas9 can identify functionally critical positions governing editing activity, with attention patterns closely aligned with established biological mechanisms of the CRISPR/Cas9 system.

Finally, we trained a CatBoost model and interpreted its predictions using both SHapley Additive exPlanations (SHAP) [59] and Local Interpretable Model-agnostic Explanations (LIME) [21, 22] to complement the attention-based analysis from an external perspective. The SHAP summary revealed that position 20 had the highest mean absolute SHAP value, indicating its dominant contribution to model output. This finding corroborates the attention-based analysis from TransformerCas9, underscoring the functional significance of position 20 (Additional file: Fig. S6a-b). Consistently, using the LIME analysis on the same model reinforced these insights by assigning the highest positive contribution value of +4.059 to position 20, significantly exceeding that of other nucleotide positions (Additional file: Fig. S6c). Together, these results strongly emphasize the biological and computational importance of this site [60].

Collectively, the convergence of TransformerCas9's intrinsic interpretability, sequence perturbation, and external model-agnostic validations consistently identified and validated biologically meaningful positions in the sgRNA sequence [52–55, 60]. These results underscore that TransformerCas9 provides both predictive accuracy and mechanistic interpretability, which bridges the gap between black-box deep learning and

the mechanistic underpinnings of CRISPR/Cas9 sequence-function relationships.

Discussion

Accurate prediction of CRISPR/Cas9 sgRNA targeting activity remains a central objective in genome engineering research [61–63]. In this study, we developed TransformerCas9, a Transformer-based deep learning model designed for the accurate prediction of CRISPR/Cas9 activity. By leveraging multi-head self-attention mechanisms, the model captures global contextual features within sgRNA sequences, transcending the local receptive field constraints inherent in conventional convolutional frameworks. TransformerCas9 further integrates a multi-level attention mechanism with multi-scale representations, which jointly encode fine-grained K-mer patterns and global multi-scale dependencies within a unified framework. This design allows the model to learn biologically relevant hierarchical features without relying on predefined handcrafted features. Comprehensive evaluations across multiple public datasets demonstrate the superiority of this architectural design in both predictive performance and biological interpretability.

To address the "black-box" nature of deep learning models, we adopted a multi-perspective interpretability framework. First, the analysis of K-mer level attention weight matrices revealed key nucleotide positions with elevated attention values corresponding to the PAM-proximal region, consistent with their known biological significance [28, 64]. Second, profiling the mean predicted values of each nucleotide at individual positions confirmed the presence of distinct, position-specific (positions 16, 18, and 20) nucleotide preferences captured by TransformerCas9. These findings align with established biological knowledge [55] and support the biological plausibility of our

model's predictions.

Despite these advancements, several limitations highlight key directions for future study. Data heterogeneity arising from variations in cell types, experimental platforms, and measurement techniques introduces systematic biases that hinder model generalization [65, 66]. Although evaluations across multiple datasets demonstrate that the model is robust to different SpCas9 variants, its predictive performance remains constrained by the scale and diversity of available training data. Future work should expand experimental datasets, establish unified data standards and processing pipelines, and implement rigorous quality control to reduce heterogeneity.

Beyond data-level limitations, the model's applicability to more complex editing scenarios needs to be further improved. While the current study accommodates both canonical and PAM-flexible variant, practical therapeutic applications often rely on chemically modified sgRNAs to enhance stability [67] and reduce immunogenicity [68]. Incorporating dynamic positional encoding and chemical descriptor embeddings may enable broader compatibility with novel sgRNA formats. Furthermore, in clinical gene editing, off-target cleavage at unintended genomic loci remains a critical safety concern. Although our evaluation included several high-fidelity variants engineered to minimize off-target risk, TransformerCas9 currently focuses exclusively on predicting on-target cleavage efficiency. To guarantee both high efficiency and genomic safety, incorporating off-target prediction by modeling the mismatch tolerance between sgRNAs and non-target genomic DNA remains an essential next step. Addressing these limitations will be essential for advancing TransformerCas9 toward broader and more clinically relevant applications.

Conclusions

This study presents TransformerCas9, an intrinsically interpretable multi-scale deep learning model that accurately predicts sgRNA activity from raw nucleotide sequences without relying on pre-defined biophysical features. By utilizing hybrid K-mer encoding and multi-level attention, TransformerCas9 consistently outperforms traditional machine learning and state-of-the-art deep learning models across diverse SpCas9 variants. Crucially, its transparent attention mechanism captures biologically relevant patterns—such as seed-region base preferences—providing a highly effective computational tool for optimized sgRNA design.

Abbreviations

A	adenine base
C	cytosine base
CNNs	convolutional neural networks
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cas9	CRISPR-associated protein 9
DT	Decision Tree
FCN	Fully Connected Network
FFN	feedforward network
G	guanine base
GBDT	Gradient Boosting Decision Tree
IG	Integrated Gradients
IQR	interquartile range
ISM	in silico mutagenesis

539	KE	K-mer Encoding
540	KLA	K-mer Level Attention
541	KNN	K-Nearest Neighbors
542	LIME	Local Interpretable Model-agnostic Explanations
543	LSTM	Long Short-Term Memory
544	MAE	mean absolute error
545	MFE	minimum free energy
546	MLA	Multi-Level Attention
547	MLP	Multi-layer Perceptron
548	MSA	Multi-Scale Attention
549	PAMs	protospacer adjacent motifs
550	RF	Random Forest
551	RMSE	root mean squared error
552	RNNs	recurrent neural networks
553	R^2	coefficient of determination
554	sgRNA	single-guide RNA
555	SHAP	SHapley Additive exPlanations
556	SOTA	state-of-the-art
557	SpCas9	*Streptococcus pyogenes* Cas9
558	R	Spearman correlation coefficient
559	r	Pearson correlation coefficient
560	SVM	Support Vector Machine
561	Tm	melting temperature

Supplementary material

File name: Additional file.pdf

Description: This file contains supplementary figures and supplementary tables supporting the analyses in this study.

Acknowledgements

The authors acknowledge the High-Performance Computing Platform of Chongqing Medical University for providing computational resources.

Author contributions

Q.S., J.L., and F.L. conceived and designed the study. Q.S. and J.L. performed the experiments and analyzed the data. J.L. and F.L. designed the TransformerCas9 framework and completed the model training and evaluation on benchmark datasets. Q.S. and J.L. conducted the cross-variant data training and associated analysis. H.L. and L.C. contributed to the framework evaluation and baseline model assessment. H.L., Z.Z., and J.Q. contributed to the visualization for the explainability analysis. Q.S. and J.L. wrote the manuscript. B.H. and F.L. supervised the study. F.L. and J.C. mentored the writing of the thesis. All authors reviewed, revised, and approved the final manuscript.

Funding

This work was supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202500461).

Data availability

We have made the source code for TransformerCas9 and the custom Python scripts used

for the activity prediction available on GitHub at https://github.com/BioAIGroup/TransformerCas9. All datasets with activity labels used in this study were derived from previously published papers. Specifically, the HT_Cas9_Train and HT_Cas9_Test datasets are provided as Tables S1 and S2 in the original study [26]; datasets for WT-SpCas9, SpCas9-HF1, and eSpCas9 are provided as Supplementary Data 2 [28]; and the Sniper-Cas9, SpCas9-NG, xCas9, and HypaCas9 datasets are available in Supplementary Table 8 of the corresponding paper [30].

Code availability

The source code can be found at https://github.com/BioAIGroup/TransformerCas9.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Barrangou R, May AP. Unraveling the potential of CRISPR-Cas9 for gene therapy. Expert Opin Biol Ther. 2015;15:311–4. https://doi.org/10.1517/14712598.2015.994501.

2. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity.

Science. 2012;337:816–21. https://doi.org/10.1126/science.1225829.

3. Pennisi E. The CRISPR Craze. Science. 2013;341:833–6. https://doi.org/10.1126/science.341.6148.833.

4. Osborn MJ, Gabriel R, Webber BR, DeFeo AP, McElroy AN, Jarjour J, et al. Fanconi Anemia Gene Editing by the CRISPR/Cas9 System. Hum Gene Ther. 2015;26:114–26. https://doi.org/10.1089/hum.2014.111.

5. Wu YX, Liang D, Wang YH, Bai MZ, Tang W, Bao SM, et al. Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9. Cell Stem Cell. 2013;13:659–62. https://doi.org/10.1016/j.stem.2013.10.016.

6. Hendriks D, Clevers H, Artegiani B. CRISPR-Cas Tools and Their Application in Genetic Engineering of Human Stem Cells and Organoids. Cell Stem Cell. 2020;27:705–31. https://doi.org/10.1016/j.stem.2020.10.014.

7. Long C, McAnally JR, Shelton JM, Mireault AA, Bassel-Duby R, Olson EN. Prevention of muscular dystrophy in mice by CRISPR/Cas9–mediated editing of germline DNA. Science. 2014;345:1184–8. https://doi.org/10.1126/science.1254445.

8. Akanmu AO, Asemoloye MD, Marchisio MA, Babalola OO. Adoption of CRISPR-Cas for crop production: present status and future prospects. PeerJ. 2024;12:e17402. https://doi.org/10.7717/peerj.17402.

9. Zhu HC, Li C, Gao CX. Applications of CRISPR–Cas in agriculture and plant biotechnology. Nat Rev Mol Cell Biol. 2020;21:661–77. https://doi.org/10.1038/s41580-020-00288-9.

10. Mohamed HI, Khan A, Basit A. CRISPR-Cas9 System Mediated Genome Editing Technology: An Ultimate Tool to Enhance Abiotic Stress in Crop Plants. J Soil Sci Plant

Nutr. 2024;24:1799–822. https://doi.org/10.1007/s42729-024-01778-x.

11. Chen F, Chen L, Yan Z, Xu J, Feng L, He N, et al. Recent advances of CRISPR-based genome editing for enhancing staple crops. Front Plant Sci. 2024;15:1478398. https://doi.org/10.3389/fpls.2024.1478398.

12. Verdezoto-Prado J, Chicaiza-Ortiz C, Mejía-Pérez AB, Freire-Torres C, Viteri-Yánez M, Deng LR, et al. Advances in environmental biotechnology with CRISPR/Cas9: bibliometric review and cutting-edge applications. Discov Appl Sci. 2025;7:167. https://doi.org/10.1007/s42452-025-06609-x.

13. Ferreira R, David F, Nielsen J. Advancing biotechnology with CRISPR/Cas9: recent applications and patent landscape. J Ind Microbiol Biotechnol. 2018;45:467–80. https://doi.org/10.1007/s10295-017-2000-6.

14. Dang Y, Jia GX, Choi J, Ma HM, Anaya E, Ye C, et al. Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biol. 2015;16:280. https://doi.org/10.1186/s13059-015-0846-3.

15. Noshay JM, Walker T, Alexander WG, Klingeman DM, Romero J, Walker AM, et al. Quantum biological insights into CRISPR-Cas9 sgRNA efficiency from explainable-AI driven feature engineering. Nucleic Acids Res. 2023;51:10147–61. https://doi.org/10.1093/nar/gkad736.

16. Ren FR, Ren C, Zhang Z, Duan W, Lecourieux D, Li SH, et al. Efficiency Optimization of CRISPR/Cas9-Mediated Targeted Mutagenesis in Grape. Front Plant Sci. 2019;10:612. https://doi.org/10.3389/fpls.2019.00612.

17. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278–324. https://doi.org/10.1109/5.726791.

18. Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. 2015. https://doi.org/10.48550/arXiv.1506.00019.

19. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5:157–66. https://doi.org/10.1109/72.279181.

20. Zhang GS, Dai ZM, Dai XH. C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. Comput Struct Biotechnol J. 2020;18:344–54. https://doi.org/10.1016/j.csbj.2020.01.013.

21. Wellawatte GP, Gandhi HA, Seshadri A, White AD. A Perspective on Explanations of Molecular Prediction Models. J Chem Theory Comput. 2023;19:2149–60. https://doi.org/10.1021/acs.jctc.2c01235.

22. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016. https://doi.org/10.48550/arXiv.1602.04938.

23. Xiao LM, Wan YQ, Jiang ZR. AttCRISPR: a spacetime interpretable model for prediction of sgRNA on-target activity. BMC Bioinformatics. 2021;22:589. https://doi.org/10.1186/s12859-021-04509-6.

24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.

25. Zhang GS, Xie HZ, Dai XH. DeepIndel: An Interpretable Deep Learning Approach for Predicting CRISPR/Cas9-Mediated Editing Outcomes. Int J Mol Sci. 2024;25:10928. https://doi.org/10.3390/ijms252010928.

26. Kim HK, Kim Y, Lee S, Min S, Bae JY, Choi JW, et al. SpCas9 activity prediction

673 by DeepSpCas9, a deep learning–based model with high generalization performance. Sci
674 Adv. 2019;5:eaax9249. https://doi.org/10.1126/sciadv.aax9249.
675 27. Zhou MY, Ma J, Yang HQ, Jiang LX, Mo Y. Sequential Attention Module for
676 Natural Language Processing. 2021. https://doi.org/10.48550/arXiv.2109.03009.
677 28. Wang DQ, Zhang CD, Wang B, Li B, Wang Q, Liu D, et al. Optimized CRISPR
678 guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat Commun.
679 2019;10:4284. https://doi.org/10.1038/s41467-019-12281-8.
680 29. Lee JK, Jeong E, Lee J, Jung M, Shin E, Kim Y, et al. Directed evolution of
681 CRISPR-Cas9 to increase its specificity. Nat Commun. 2018;9:3048.
682 https://doi.org/10.1038/s41467-018-05477-x.
683 30. Kim N, Kim HK, Lee S, Seo JH, Choi JW, Park J, et al. Prediction of the
684 sequence-specific cleavage activity of Cas9 variants. Nat Biotechnol. 2020;38:1328–36.
685 https://doi.org/10.1038/s41587-020-0537-9.
686 31. Nishimasu H, Shi X, Ishiguro S, Gao L, Hirano S, Okazaki S, et al. Engineered
687 CRISPR-Cas9 nuclease with expanded targeting space. Science. 2018;361:1259–62.
688 https://doi.org/10.1126/science.aas9129.
689 32. Hu JH, Miller SM, Geurts MH, Tang WX, Chen LW, Sun N, et al. Evolved Cas9
690 variants with broad PAM compatibility and high DNA specificity. Nature. 2018;556:57–63.
691 https://doi.org/10.1038/nature26155.
692 33. Browne TS, Edgell DR, Gloor GB. Better data for better predictions: Data
693 curation improves deep learning for sgRNA/Cas9 prediction. PeerJ. 2026;14:e20706.
694 https://doi.org/10.7717/peerj.20706.
695 34. Cao XH, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve

696 classification accuracies in biomedical data. BMC Bioinformatics. 2016;17:359.
697 https://doi.org/10.1186/s12859-016-1236-x.

698 35. Wan Y, Jiang Z. TransCrispr: Transformer based hybrid model for predicting
699 CRISPR/Cas9 single guide RNA cleavage efficiency. IEEE/ACM Trans Comput Biol
700 Bioinform. 2023;20:1518–28. https://doi.org/10.1109/TCBB.2022.3201631.

701 36. Liu F, Huang S, Hu J, Chen X, Song Z, Dong J, et al. Design of prime-editing
702 guide RNAs with deep transfer learning. Nat Mach Intell. 2023;5:1261–74.
703 https://doi.org/10.1038/s42256-023-00739-w.

704 37. Kumari S, Nie J, Chen H-S, Ma H, Stewart R, Li X, et al. Evaluation of gene
705 association methods for coexpression network construction and biological knowledge
706 discovery. PloS One. 2012;7:e50411. https://doi.org/10.1371/journal.pone.0050411.

707 38. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
708 https://doi.org/10.1007/BF00994018.

709 39. Rumelhart DE, Hintont GE, Williams RJ. Learning representations by
710 back-propagating errors. 1986.

711 40. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory.
712 1967;13:21–7. https://doi.org/10.1109/TIT.1967.1053964.

713 41. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
714 https://doi.org/10.1007/BF00116251.

715 42. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
716 https://doi.org/10.1023/A:1010933404324.

717 43. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann
718 Stat. 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.

44. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55:119–39. https://doi.org/10.1006/jcss.1997.1504.

45. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

46. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.

47. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2018.

48. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

49. Zhang GS, Zeng T, Dai ZM, Dai XH. Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. Comput Struct Biotechnol J. 2021;19:1445–57. https://doi.org/10.1016/j.csbj.2021.03.001.

50. Liu Q, He D, Xie L. Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas System using attention boosted deep learning and network-based gene feature. PLOS Comput Biol. 2019;15:e1007480. https://doi.org/10.1371/journal.pcbi.1007480.

742 51. Yu G, Kim HK, Park J, Kwak H, Cheong Y, Kim D, et al. Prediction of efficiencies
743 for diverse prime editing systems in multiple cell types. Cell. 2023;186:2256-2272.e23.
744 https://doi.org/10.1016/j.cell.2023.03.034.
745 52. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using
746 the CRISPR/Cas9 system. Science. 2014;343:80–4.
747 https://doi.org/10.1126/science.1246981.
748 53. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al.
749 Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation.
750 Nat Biotechnol. 2014;32:1262–7. https://doi.org/10.1038/nbt.3026.
751 54. Xu H, Xiao T, Chen C-H, Li W, Meyer CA, Wu Q, et al. Sequence determinants of
752 improved CRISPR sgRNA design. Genome Res. 2015;25:1147–57.
753 https://doi.org/10.1101/gr.191452.115.
754 55. Zhu WT, Xie HZ, Chen YW, Zhang GS. CrnnCrispr: An Interpretable Deep
755 Learning Method for CRISPR/Cas9 sgRNA On-Target Activity Prediction. Int J Mol Sci.
756 2024;25:4429. https://doi.org/10.3390/ijms25084429.
757 56. Cheng X, Li Z, Shan R, Li Z, Wang S, Zhao W, et al. Modeling CRISPR-Cas13D
758 on-target and off-target effects using machine learning approaches. Nat Commun.
759 2023;14:752. https://doi.org/10.1038/s41467-023-36316-3.
760 57. Wang X, Cheng X, Li Z, Ma S, Zhang H, Chen Z, et al. A comprehensive
761 benchmark for multiple highly efficient base editors with broad targeting scope. 2024.
762 https://doi.org/10.1101/2024.12.17.628899.
763 58. Li VR, Zhang Z, Troyanskaya OG. CROTON: An automated and variant-aware
764 deep learning framework for predicting CRISPR/Cas9 editing outcomes. Bioinformatics.

765 2021;37 Supplement_1:i342–8. https://doi.org/10.1093/bioinformatics/btab268.
766 59. Mak JK, Störtz F, Minary P. Comprehensive computational analysis of epigenetic
767 descriptors affecting CRISPR-Cas9 off-target activity. BMC Genomics. 2022;23:805.
768 https://doi.org/10.1186/s12864-022-09012-7.
769 60. Konstantakos V, Nentidis A, Krithara A, Paliouras G. CRISPR–Cas9 gRNA
770 efficiency prediction: An overview of predictive tools and the role of deep learning.
771 Nucleic Acids Res. 2022;50:3616–37. https://doi.org/10.1093/nar/gkac192.
772 61. Dong C, Gou YW, Lian JZ. SgRNA engineering for improved genome editing and
773 expanded functional assays. Curr Opin Biotechnol. 2022;75:102697.
774 https://doi.org/10.1016/j.copbio.2022.102697.
775 62. Moon SB, Kim DY, Ko J-H, Kim J-S, Kim Y-S. Improving CRISPR Genome
776 Editing by Engineering Guide RNAs. Trends Biotechnol. 2019;37:870–81.
777 https://doi.org/10.1016/j.tibtech.2019.01.009.
778 63. Xiang X, Corsi GI, Anthon C, Qu KL, Pan XG, Liang X, et al. Enhancing
779 CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. Nat
780 Commun. 2021;12:3238. https://doi.org/10.1038/s41467-021-23576-0.
781 64. Xue L, Tang B, Chen W, Luo J. Prediction of CRISPR sgRNA Activity Using a
782 Deep Convolutional Neural Network. J Chem Inf Model. 2019;59:615–24.
783 https://doi.org/10.1021/acs.jcim.8b00368.
784 65. Trivedi V, Mohseni A, Lonardi S, Wheeldon I. Balanced Training Sets Improve
785 Deep Learning-Based Prediction of CRISPR sgRNA Activity. ACS Synth Biol.
786 2024;13:3774–81. https://doi.org/10.1021/acssynbio.4c00542.
787 66. Zhang GS, Luo Y, Dai XH, Dai ZM. Benchmarking deep learning methods for

predicting CRISPR/Cas9 sgRNA on- and off-target activities. Brief Bioinform. 2023;24:bbad333. https://doi.org/10.1093/bib/bbad333.

67. Hendel A, Bak RO, Clark JT, Kennedy AB, Ryan DE, Roy S, et al. Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. Nat Biotechnol. 2015;33:985–9. https://doi.org/10.1038/nbt.3290.

68. Kim S, Koo T, Jee H-G, Cho H-Y, Lee G, Lim D-G, et al. CRISPR RNAs trigger innate immune responses in human cells. Genome Res. 2018;28:367–73. https://doi.org/10.1101/gr.231936.117.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile.pdf