IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Hybrid Neural Networks for Off-Target Prediction in Gene Editing: Integrating Sequence and Biophysical Features

Parsh Verma

Pomperaug High School, 380 Old Waterbury Rd, Southbury, Connecticut, 06488, USA; parshpune@gmail.com

ABSTRACT: The introduction of CRISPR-Cas9 revolutionized genetics but brought challenges, notably off-target effects leading to undesired DNA changes. This research proposes a hybrid neural network integrating sequencing information, biophysical features, and CG content from gRNA sequences to predict off-target effects better. In this research, various neural network architectures, both simple and complex models with attention mechanisms, were developed and analyzed against three cell lines. Emphasis was given to F1 scores to enhance accuracy while reducing false positives. The study demonstrated that while complex models with attention mechanisms excel at identifying off-target effects, simpler architectures can also achieve excellent results for specific cell lines when paired with well-selected features. These findings will contribute to developing more reliable predictive models, improving the safety and efficacy of gene editing technologies.

KEYWORDS: Computational Biology and Bioinformatics, Genomics, Machine Learning, Gene Editing, CRISPR.

Introduction

CRISPR-Cas9 is a groundbreaking gene-editing technology derived from a bacterial immune system that cuts foreign DNA. It uses two key components: guide RNA (gRNA) to direct the Cas9 enzyme to a specific DNA sequence, and the Cas9 protein to cleave the DNA. Although this allows precise modifications, off-target edits can occur when gRNA binds to similar sequences, causing unintended DNA cleavage and potential mutations, especially in regions with high CG content. Some studies suggest that higher GC content in the gRNA can stabilize the DNA-RNA duplex, potentially increasing on-target activity, while others indicate that target sequences with a GC content higher than 70% may increase the likelihood of off-target effects.[1]

While CRISPR-Cas9 has revolutionized gene editing, its utility is limited by off-target effects that can cause harmful genetic alterations. For instance, recent studies have shown that Cas9 tolerates up to three mismatches between the gRNA and genomic DNA, which increases the chances of off-target effects in regions with higher GC content.[2] The need for accurate prediction of these off-target effects is critical for the safety and effectiveness of CRISPR-based therapies. Existing models often overlook complex, non-linear biophysical interactions, which can significantly impact the accuracy of off-target predictions. Recent advancements have sought to address these challenges by incorporating factors such as DNA binding energy and chromatin accessibility, which have led to some improvements in prediction accuracy. However, significant gaps remain, particularly in the integration of multiple feature types and accounting for non-linear dependencies.

This research addresses these gaps by introducing a hybrid neural network that integrates both sequence information and biophysical properties, such as binding energy, to enhance off-target effect predictions. This hybrid neural network refers to a model that combines different types of neural networks—specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs), along with attention mechanisms. CNNs are utilized for capturing spatial features within the sequence, while RNNs, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), help capture dependencies in sequential data. The attention mechanism further enhances the model by focusing on the most relevant parts of the input sequence, allowing for improved prediction accuracy. For instance, CRISPR-Net employs a recurrent convolutional network to score gRNA-target pairs with mismatches and indels.[3]

By considering both the surrounding sequence and its inherent properties, including CG content, this approach offers improved accuracy in identifying potential off-target loci in CRISPR-Cas9 applications. It aligns with the latest methodologies that utilize deep learning to account for complex sequence dependencies and class imbalances in genomic data. The proposed hybrid model combines convolutional and recurrent neural network components with attention mechanisms to capture complex interactions, contributing to the refinement of gene-editing technologies and the development of more advanced predictive models in genomics.

Sequence-Based Prediction Models:

Early efforts in CRISPR-Cas9 off-target prediction primarily relied on sequence homology and position-weight matrices (PWM). Tools like CRISPRmit and CCTop ranked off-target sites based on their similarity to the gRNA target.[4] Despite their utility, these models had limited accuracy, functioning well only in specific genome regions due to unconsidered interacting factors.

Integration of Biophysical Features:

Recent studies have incorporated biophysical features alongside gene sequences to predict off-target effects more ac-

DOI: 10.36838/v7i2.113

curately. For example, a study demonstrated that manipulating Cas9's biophysical attributes, such as minimizing nonspecific RNA-DNA interactions, significantly reduces off-target mutations while maintaining high on-target efficiency.[5] The BoostMEC model further improves prediction by considering binding energy, cleavage efficiency, and other biophysical factors, reducing false positives and enhancing off-target site identification.[6]

Machine Learning Approaches:

Machine learning, particularly deep learning, has revolutionized CRISPR-Cas9 off-target prediction. DeepCRISPR uses a deep convolutional denoising neural network (DCDNN) and CNN to predict both on- and off-target activities of sgRNAs with improved accuracy.[7] Similarly, DeepCRISTL, developed in 2024, applies deep transfer learning to predict CRISPR/Cas9 editing efficiency in specific cellular contexts, fine-tuning predictions with smaller, biologically relevant datasets for superior context-specific accuracy.[8]

Role of CG Content and Emerging Trends:

CG content is increasingly recognized as a critical factor in off-target prediction due to its influence on CRISPR-Cas9 binding specificity. High CG content regions are linked to increased off-target cleavage risks, highlighting the need to consider nucleotide composition in predictive models.[1] Additionally, integrating multi-omics data, such as chromatin openness and epigenetic changes, has shown significant promise. A study demonstrated that combining sequence analyses with multi-omics data markedly improves off-target predictions across various cell lines.[9]

Relevance to the Current Study:

This study aims to enhance off-target prediction by developing a hybrid neural network that integrates sequence data, biophysical features, and CG content. By testing models across different cell lines like HEK293, K562, and U2OS, the study builds on recent advancements in CRISPR-Cas9 research, incorporating novel computational approaches to improve the safety profiles of gene editing technologies.

This study hypothesizes that a hybrid neural network model integrating sequence information and biophysical properties—such as binding energy and CG content—will significantly improve the accuracy of predicting CRISPR-Cas9 off-target effects compared to existing models that only utilize sequence-based features or standard deep learning approaches. It is anticipated that the proposed hybrid model will demonstrate superior performance across multiple metrics (accuracy, precision, and recall) in diverse cell lines such as HEK293, K562, and U2OS, owing to its ability to effectively capture non-linear dependencies and leverage multiple relevant features for off-target identification.

Methods

Data Preparation:

The dataset utilized in this research can be found in the crisprSQL database.[10] It consists of CRISPR-Cas9 gene editing sequences along with corresponding cleavage frequencies, energy features, and other biophysical attributes. The dataset was pre-processed to remove any rows containing negative cleavage frequencies, ensuring that only biologically plausible data is retained for further modeling. Additionally, CG content was calculated for each guide RNA (gRNA) sequence by counting the number of cytosine (C) and guanine (G) nucleotides divided by the length of the sequence.

Using k-mers (k=2) best described as short motifs; the gRNA sequences were transformed. These k-mers help in capturing nearby sequence information that is very crucial for understanding how one character in the string depends upon others. The sequences were then tokenized after which k-mer sequences had their padded equivalents created so that all models could have inputs with the same lengths maintained.

Out of the available options for consideration, five energy parameters (energy_1 to energy_5) and CG content were chosen as input features for training. The dependent variable, which is cleavage frequency, was turned into a binary variable by using a threshold determined from the median cleavage frequency of the training dataset so that anything more than or equal to the median was assigned as 1 while the rest as 0.

Furthermore, the SMOTE-Tomek resampling technique was used to correct any imbalance in the distribution of classes among the training data samples.

Cell Lines Used:

In this dataset, the tests were performed to validate if the models were applicable to the following 3 cell lines.

HEK293: A human embryonic kidney cell line known for high transfection efficiency.

K562: A human myelogenous leukemia cell line, commonly used in hematopoietic research, helping assess CRISPR-Cas9 effects in blood cells.

U2OS: A human osteosarcoma cell line frequently used in cancer research, aiding in evaluating CRISPR-Cas9 specificity in cancer-related contexts.

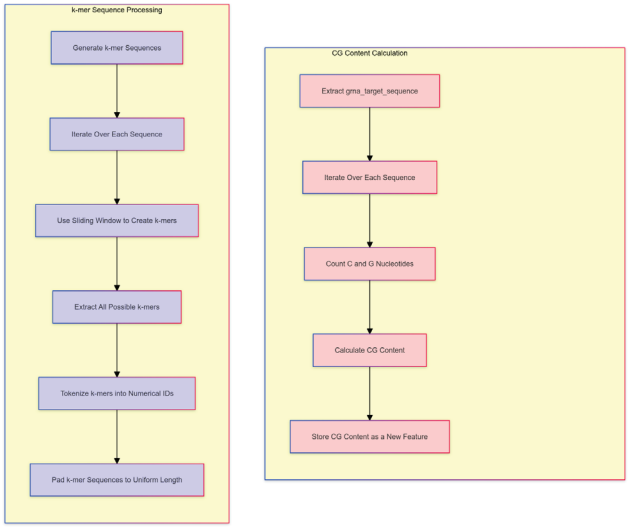

Figure 1: ***Feature Engineering (CG Content & gRNA sequence processing)***
This figure depicts the feature engineering steps for k-mer sequence processing and CG content calculation. The left side shows the creation and tokenization of k-mers, while the right side outlines the calculation of CG content from sequences.

Neural Network Architectures:

Three neural network architectures were implemented to predict off-target effects of CRISPR-Cas9 gene editing, integrating both sequence-based and biophysical features: single-layered models, multi-layered models without attention, and multi-layered models with attention. This study also analyzes how effective and strong they were when applied in predicting off-target effects in various cell lines.

To streamline the discussion of the various models used in this study, each model variant within these architectures is referenced by a unique identifier (M1 through M10) as follows:

- **M1:** BiDirectional LSTM (Architecture 1)
- **M2:** CNN (Architecture 1)
- **M3:** GRU (Architecture 1)
- **M4:** LSTM (Architecture 1)
- **M5:** CNN+BiDirectional LSTM (Architecture 2)
- **M6:** CNN+GRU (Architecture 2)
- **M7:** CNN+LSTM (Architecture 2)
- **M8:** CNN+BiDirectional LSTM+Attention (Architecture 3)
- **M9:** CNN+GRU+Attention (Architecture 3)
- **M10:** CNN+LSTM+Attention (Architecture 3)

These model identifiers will be consistently used throughout the paper for clarity.

Base Convolutional Hybrid:

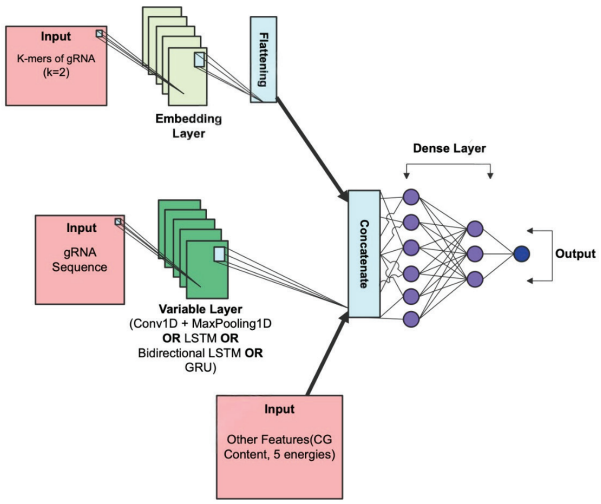

Figure 2: ***Architecture 1 – Base Convolutional Hybrid***
This figure shows a neural network architecture where k-mers of gRNA are processed through an embedding layer, gRNA sequences through a variable layer, and CG content with energy features are directly concatenated. The combined features are passed through a dense layer to predict CRISPR-Cas9 off-target effects.

Architecture 1 implemented multiple models based on different core neural network components: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM, and Gated Recurrent Unit (GRU). This architecture used three primary inputs: gRNA sequences, k-mer representations (encoded with embeddings), and additional biophysical features such as energy values and CG content.

CNN Model: Extracted local sequence features, integrated with k-mer embeddings and biophysical features for classification.

LSTM Model: Replaced the CNN layer with an LSTM layer, which was combined with k-mer embeddings and biophysical features for classification.

Bidirectional LSTM Model: Utilized Bidirectional LSTM layers to capture sequence dependencies in both directions, enhancing context understanding.

GRU Model: Substituted LSTM with a GRU layer, offering computational efficiency while maintaining sequence pattern recognition.

Sequential Convolutional Hybrid:

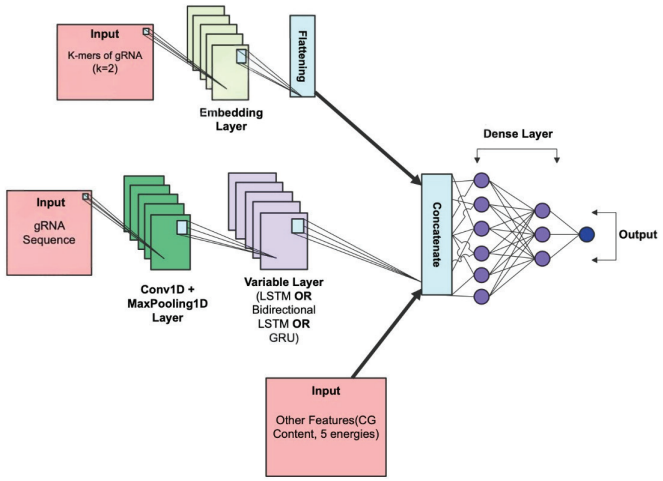

Figure 3: ***Architecture 2 – Sequential Convolutional Hybrid***
This figure shows a neural network where k-mers of gRNA are embedded and flattened, gRNA sequences pass through Conv1D and a variable layer (LSTM, bidirectional LSTM, or GRU), and CG content with energy features are concatenated before entering a dense layer. The architecture outputs predictions for CRISPR-Cas9 off-target effects.

In **Architecture 2**, various models have been enhanced by combining Convolutional Neural Networks (CNN) and recurrent neural network (RNN) layers—particularly LSTM, Bidirectional LSTM, and GRU.

CNN+LSTM Model: The CNN layer is followed by a max-pooling layer in this architecture, and then the sequence data goes through an LSTM layer. After passing through a dense layer, the final LSTM output is concatenated with the k-mer embeddings and other biophysical features.

CNN+Bidirectional LSTM Model: Bidirectional LSTM layers were employed in this model in place of the standard ones to improve the capturing of bidirectional dependencies within the sequence data.

CNN+GRU Model: For this model, the decision was to include a GRU layer instead of LSTM, as once again it was designed to be computationally efficient, while still performing well.

DOI: 10.36838/v7i2.113

Attention-Augmented Sequential Hybrid:

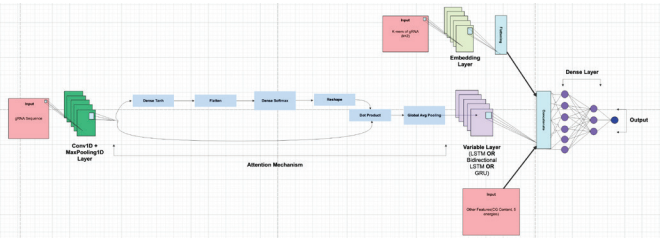

Figure 4: ***Architecture 3 – Attention-Augmented Sequential Hybrid***
This figure illustrates an attention-augmented neural network architecture where gRNA sequences pass through Conv1D and MaxPooling1D layers, followed by an attention mechanism applied to the processed sequence. K-mers of gRNA and other features (CG content and energy scores) are processed separately and concatenated with the attention-modulated output before passing through dense layers for CRISPR-Cas9 off-target prediction.

Architecture 3 introduces attention mechanisms to enhance the prediction capability of the model by focusing on important parts of the input sequence data.

Attention Block: This is an additional mechanism added to the previous architecture, architecture 2. The attention mechanism was employed on the CNN layer output. Following max-pooling, the attention layer performed arithmetic mean on the input sequence. This enabled the model to concentrate on certain areas crucial in predicting cleavage.

CNN+LSTM+Attention Model: The CNN layer outputs are first processed by the attention layer before being passed through an LSTM layer. The final LSTM output, together with the k-mer embeddings and biophysical features, was concatenated and passed through a dense layer for classification.

CNN+Bidirectional LSTM+Attention Model: In this model, the LSTM layer is substituted by the Bidirectional LSTM layer to enable better context representation via attention-improved sequence data.

CNN+GRU+Attention Model: This is an improved version of the LSTM+attention model that capitalizes on attention benefits in a computationally efficient manner by replacing the LSTM sequential dependency handling layer with a GRU one.

Mathematical Optimization in Enhancing Off-Target Prediction Accuracy:

- ***Gradient Descent in the Context of Off-Target Prediction:***

Optimization is crucial for neural networks predicting off-target effects in CRISPR-Cas9 gene editing, as it enables the model to identify subtle patterns in genomic data. The complex relationship between sequence-specific features, biophysical properties, and gene editing outcomes requires a finely tuned model, achieved through careful calculus-based optimization.

The primary goal of this research is to minimize prediction errors while maximizing the identification of off-target effects. This is directly tied to optimizing the loss function $L(\theta)$, which represents the cost associated with the discrepancy between predicted off-target sites and actual data. The loss function can be visualized as a landscape with multiple peaks and valleys, each point representing a specific configuration of model parameters. The objective is to find the global minimum, where predictions are most accurate.

The gradient of the loss function $\nabla_\theta L(\theta)$ indicates how much each parameter needs to be adjusted to move closer to the optimal configuration. For instance, when predicting if a specific gRNA sequence will cause an off-target effect, the gradient guides adjustments to weights associated with features like CG content or binding energy to improve accuracy.

Each component of this gradient vector represents the partial derivative of the loss function with respect to a specific parameter. The chain rule, a fundamental calculus concept, is applied to compute these derivatives, particularly in deep neural networks where layers of transformations are stacked.

Backpropagation and Adjusting Model Weights:

This study utilized backpropagation to compute gradients, particularly useful for complex characteristics like the biophysical properties of DNA interactions. For example, if the model incorrectly predicts an off-target site, backpropagation identifies necessary adjustments by analyzing features such as binding or unzipping energy.

The update rule for each parameter 0; during gradient descent is given by:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \cdot \frac{\partial L}{\partial \theta_i}$$

The η (learning rate) is a key hyperparameter that influences the speed at which the model converges at a solution. In the prediction of CRISPR-Cas9 off-targets, this procedure spans numerous epochs during which the model is trained to tell apart true from false off-target sites.

Importance of Learning Rate and Convergence:

Selecting the appropriate learning rate is crucial. A high learning rate can cause the model to overshoot the optimal solution, while a low rate risks becoming stuck in local minima. This study tested various learning rates, finding that 0.001 allowed the model to efficiently learn complex genomic relationships, reducing the loss function over time and improving prediction sensitivity. This rate was optimal for navigating the intricate landscape without becoming trapped in suboptimal configurations.

Evaluation Metrics:

In this research, the evaluation metrics used help to determine how well the model can predict off-target effects. It is possible to identify the strong and weak points of every architectural approach by considering various indicators of model performance.

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

Accuracy (ACC): Measures the proportion of correct predictions out of all predictions made.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (P): Focuses on the model's ability to correctly identify true positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (R): Measure that is very sensitive since failure to identify an off-target effect could have serious consequences.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: Combination of recall and precision that gives an overall idea of the model's overall effectiveness in balancing between the two metrics.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Measure in determining if a model can tell the difference between two classes.

Balanced Accuracy (BA): Considers sensitivity and specificity, hence it has become a useful tool for assessing the performance of models in the presence of class imbalance.

Although all the metrics mentioned have been evaluated across all architectures, the F1 score has been used to determine the optimal threshold for predictions.

Results

Metrics - Multiple Cell Lines:

The performance of the model variants across three different neural network architectures—Architecture 1 (Base Convolutional Hybrid), Architecture 2 (Sequential Convolutional Hybrid), and Architecture 3 (Attention-Augmented Sequential Hybrid)—was evaluated using metrics such as accuracy, precision, recall, F1-score, AUC, and balanced accuracy across multiple cell lines.

Table 1: ***Metrics – Multiple Cell Lines Together***
Performance metrics for different models and architectures across multiple cell lines, with highlighted cells indicating the best results.

Architecture	Model	Accuracy	Precision	Recall	F1_score	AUC	Balanced_accuracy
Architecture 1	BiDirectional LSTM	0.7370	0.7271	0.7589	0.7423	0.8287	0.7370
Architecture 1	CNN	0.7368	0.7715	0.6743	0.7181	0.8369	0.7368
Architecture 1	GRU	0.7387	0.7129	0.8021	0.7541	0.8316	0.7390
Architecture 1	LSTM	0.7347	0.7328	0.7582	0.7377	0.8324	0.7345
Architecture 2	CNN+BiDirectional LSTM	0.7304	0.7300	0.7520	0.7331	0.8283	0.7312
Architecture 2	CNN+GRU	0.7459	0.7361	0.7677	0.7512	0.8354	0.7460
Architecture 2	CNN+LSTM	0.7369	0.7731	0.6804	0.7178	0.8326	0.7362
Architecture 3	CNN+BiDirectional LSTM+Attention	0.7331	0.7223	0.7584	0.7392	0.8228	0.7327
Architecture 3	CNN+GRU+Attention	0.7307	0.7613	0.6832	0.7155	0.8298	0.7311
Architecture 3	CNN+LSTM+Attention	0.7323	0.7120	0.7897	0.7455	0.8289	0.7326

Architecture 1 (Base Convolutional Hybrid):

This architecture leverages CNN or other RNN layers for feature extraction and classification, capturing spatial patterns within input sequences.

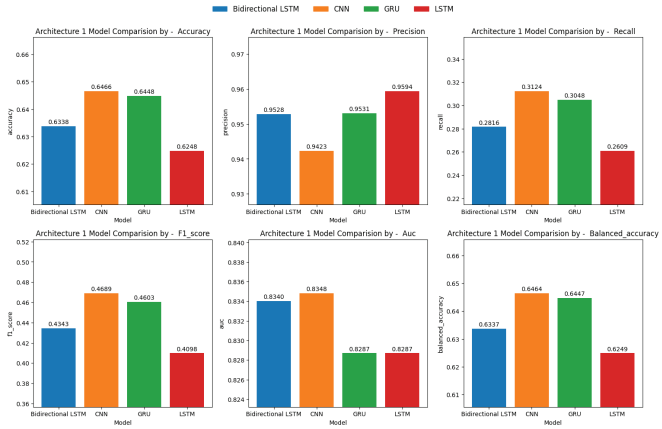

Figure 5: ***Metric Calculations for Architecture 1 (Base Convolutional Hybrid)*** Bar charts comparing accuracy, precision, recall, F1 score, AUC, and balanced accuracy for different models within Architecture 1.

Architecture 1 leveraging Convolutional Neural Networks (CNNs) for feature extraction and classification, demonstrated notable performance across various metrics. The GRU variant achieved the highest accuracy of 73.87%, reflecting its ability to identify complex patterns when trained and validated across all cell lines. The CNN model excelled in precision at 77.15%, indicating its robustness in minimizing false positives. Moreover, the GRU variant led in recall at 80.21% and F1 score at 75.41%, balancing precision and recall effectively. The CNN model's AUC of 83.69% highlighted its strong discriminative power, making the GRU variant in this architecture a strong candidate for off-target prediction, particularly in environments requiring balanced accuracy and precision.

Architecture 2 (Sequential Convolutional Hybrid):

This architecture integrates CNNs with sequential layers (LSTM, Bi-Directional LSTM, or GRU), enhancing the model's ability to capture sequence dependencies.

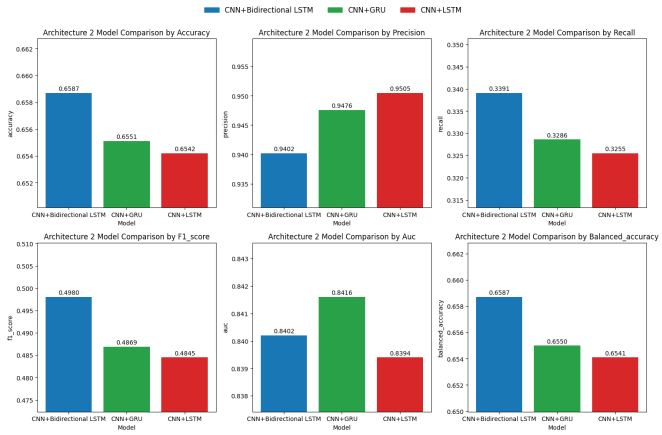

Figure 6: ***Metric Calculations for Architecture 2 (Sequential Convolutional Hybrid)*** Bar charts comparing accuracy, precision, recall, F1 score, AUC, and balanced accuracy for different models within Architecture 2.

In Architecture 2, combining convolutional layers with sequential layers such as GRU and LSTM enhanced performance, particularly in scenarios requiring sensitivity to sequence dependencies. The CNN+GRU variant demonstrat-

DOI: 10.36838/v7i2.113

ed the highest accuracy at 74.59% and balanced accuracy at 74.60%, proving effective in contexts requiring a balance between sensitivity and specificity. The CNN+LSTM variant excelled in precision at 77.31%, crucial for accurate predictions, while the CNN+GRU model led in recall at 76.77% and F1 score at 75.12%. This architecture's effectiveness in off-target prediction suggests its utility in complex genomic environments where sequence interactions are critical.

Architecture 3 (Attention-Augmented Sequential Hybrid):

This architecture introduces attention mechanisms to enhance the model's focus on critical parts of the input sequences.

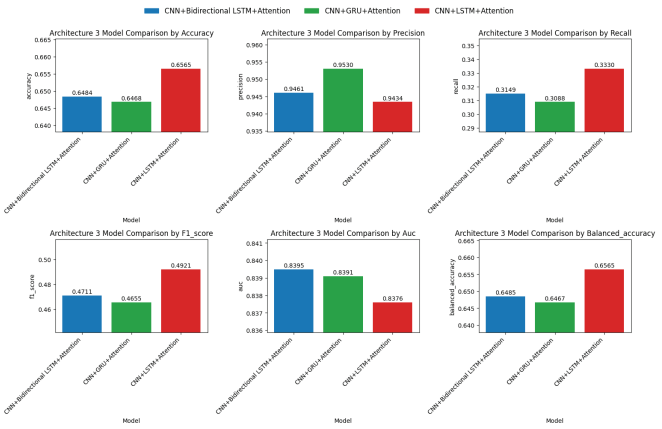

Figure 7: ***Metric Calculations for Architecture 3 (Attention-Augmented Sequential Hybrid)***
Bar charts comparing accuracy, precision, recall, F1 score, AUC, and balanced accuracy for different models within Architecture 3.

Introducing attention mechanisms in this architecture aims to enhance the model's ability to focus on relevant sequence features, particularly beneficial in scenarios requiring the identification of subtle off-target effects. The CNN+BiDirectional LSTM+Attention model achieved the highest accuracy (73.31%) and balanced accuracy (73.27%) CNN+LSTM+Attention model led in recall at 78.97% and F1 score at 74.55%, balancing precision and recall while maintaining precision at 71.20%. Although the addition of attention mechanisms improved performance, it was not very impactful when experimented against all the cell lines together.

Cell Line-Specific Metrics:

The performance of the models was further analyzed on individual cell lines to assess their robustness against these 3 cell lines available in the dataset.

Table 2: ***Metrics – 3 Cell Lines – All Model Variants***
Performance metrics, including accuracy, precision, recall, F1 score, AUC, and balanced accuracy, across U2OS, HEK293, and K562 cell lines for all model variants. Highlighted cells indicate top-performing metrics against the model name.

Model	U2OS						HEK293						K562					
	Accuracy	Precision	Recall	F1_score	AUC	Balanced_accuracy	Accuracy	Precision	Recall	F1_score	AUC	Balanced_accuracy	Accuracy	Precision	Recall	F1_score	AUC	Balanced_accuracy
M1	0.7492	0.7707	0.7131	0.7381	0.8412	0.7482	0.7153	0.7286	0.7081	0.7103	0.8047	0.7169	0.8299	0.8694	0.7793	0.8207	0.8785	0.8287
M2	0.7415	0.7948	0.6518	0.7155	0.8377	0.7415	0.7262	0.7465	0.6854	0.7137	0.8082	0.7259	0.8378	0.8974	0.7624	0.8239	0.8990	0.8378
M3	0.7475	0.7685	0.7103	0.7367	0.8387	0.7471	0.7168	0.7086	0.7392	0.7214	0.8020	0.7157	0.8191	0.8707	0.7529	0.8061	0.8740	0.8175
M4	0.7475	0.7760	0.7054	0.7355	0.8391	0.7472	0.7143	0.7040	0.7572	0.7249	0.8048	0.7147	0.8122	0.8618	0.7456	0.7988	0.8698	0.8113
M5	0.7585	0.7733	0.7471	0.7542	0.8422	0.7586	0.6989	0.6860	0.7571	0.7117	0.7846	0.7007	0.6991	0.7066	0.6993	0.6955	0.7793	0.7013
M6	0.7551	0.7773	0.7221	0.7461	0.8397	0.7553	0.6880	0.7003	0.6674	0.6801	0.7754	0.6886	0.6991	0.6931	0.7209	0.7041	0.7831	0.6981
M7	0.7599	0.7814	0.7304	0.7525	0.8408	0.7604	0.6771	0.6977	0.6608	0.6618	0.7706	0.6786	0.7070	0.7510	0.6464	0.6834	0.7847	0.7094
M8	0.7521	0.7655	0.7580	0.7609	0.8417	0.7614	0.7014	0.6996	0.7126	0.7006	0.7762	0.6998	0.7050	0.7157	0.6938	0.7011	0.7678	0.7050
M9	0.7566	0.7486	0.7777	0.7607	0.8392	0.7556	0.6999	0.6711	0.7890	0.7243	0.7798	0.7003	0.7071	0.7236	0.6752	0.6957	0.7855	0.7077
M10	0.7629	0.7739	0.7487	0.7587	0.8413	0.7623	0.6915	0.6704	0.7693	0.7127	0.7756	0.6928	0.7100	0.7133	0.7043	0.7062	0.7770	0.7098

Figure 8: ***Metric Comparison for Cell Line U2OS***
U2OS: Bar charts comparing accuracy, F1 score, and AUC across different models within the U2OS cell line.
HEK293: Bar charts comparing accuracy, F1 score, and AUC across different models within the HEK293 cell line.
K562: Bar charts comparing accuracy, F1 score, and AUC across different models within the K562 cell line.

Cell Line U2OS:

The U2OS, human osteosarcoma cell line was used to evaluate different model variants' performance under the three architectures, considering its relevance in gene editing studies.

The CNN+LSTM+Attention model (M10) in Architecture 3 exhibited the highest accuracy at 76.29% and balanced accuracy at 76.23%. This model also achieved a recall of 74.87% and an F1 score of 75.87%, making it particularly effective in identifying true off-target effects while maintaining a balance between precision and recall. Precision for this model was 77.39%, further demonstrating its reliability in minimizing false positives in this context. Overall, the attention mechanism significantly contributed to the model's ability to focus on relevant sequence features, enhancing its predictive capability in the U2OS cell line.

Cell Line HEK293:

The HEK293 cell line from the dataset was used to assess all the model variants' performance in this widely studied cellular environment.

For the HEK293 cell line, the CNN model (M2) from Architecture 1 outperformed other models, achieving the highest accuracy at 72.62% and precision at 74.65%. The recall was 68.54%, and the F1 score was 71.37%, indicating a strong balance between true positive and false positive rates. The model also showed a robust AUC of 80.82%, confirming its capability to distinguish between on-target and off-target effects. These metrics suggest that the simpler CNN-based approach is highly effective for this cell line, where adding RNN layers or attention mechanisms provided no significant improvement in predictive power.

Cell Line K562:

The K562, a human immortalized myelogenous leukemia cell line, with its distinct challenges in gene editing, was used to evaluate the performance of different model variations across the three architectures.

In the K562 cell line, the CNN variant (M2) in Architecture 1 consistently delivered the best performance, with the highest accuracy at 83.78%, precision at 89.74%, F1 score at 82.39%, and AUC at 89.90%. The recall was also strong at 76.24%, indicating its effectiveness in identifying true off-targets. The balanced accuracy of 83.78% highlights the model's consistent performance across different datasets. These results emphasize the efficacy of the CNN model in this specific context, where the simplicity of the architecture outperformed more complex models incorporating RNN layers and attention mechanisms.

Cell Line Specific Metrics - Conclusion:

Overall, these results suggest that the CNN variant in Architecture 1 (M2) consistently outperformed other models in the K562 and HEK293 cell lines, demonstrating its strong predictive power and reliability in the context of gene editing specific to this cell line. On the other hand Architecture 3, CNN+LSTM+attention (M10) worked outperformed others for U2OS cell line. This suggests, that while RNN layers and attention mechanisms worked in specific cases even simpler architecture works better in this particular context.

■ Discussion

This study shows the effectiveness of hybrid neural networks in predicting CRISPR-Cas9 off-target effects, emphasizing the importance of selecting the right architecture. Simpler CNN-based models, like Architecture 1's CNN variant (M2), excelled in K562 and HEK293 cell lines, while Architecture 3's CNN+LSTM+Attention model (M10) performed best in complex scenarios like the U2OS cell line.

The results indicate that the choice of model architecture should be based on the complexity and characteristics of the genomic environment. For simpler, well-characterized environments, such as the K562 and HEK293 cell lines, the straightforward CNN model (M2) provided reliable accuracy and precision without the need for more computationally intensive mechanisms like attention layers. In contrast, for more complex cell lines like U2OS, the combination of LSTM and attention mechanisms, as seen in Architecture 3, was essential to capture subtle dependencies and nuances in the data.

Furthermore, the attention mechanism used in Architecture 3 was particularly valuable in focusing on the most important parts of the sequence, leading to improved recall and balanced performance metrics. However, this architecture was less impactful when generalized across all cell lines, suggesting that specialized model architectures are better suited for specific genomic contexts. This points to the need for targeted application of model architectures rather than a one-size-fits-all approach.

Despite these successes, there is room for improvement. The models used a limited set of biophysical features, suggesting that including more detailed data, such as chromatin states or 3D genomic architecture, could enhance accuracy. Further optimization of attention mechanisms, possibly through advanced strategies like multi-head attention, could better capture complex gRNA interactions. Moreover, additional work is required to establish when and where more complex models, such as those with attention mechanisms, should be prioritized over simpler ones, considering both computational efficiency and predictive accuracy.

Future research should focus on real-world validation. While the F1 Score balanced precision and recall in setting thresholds, these thresholds could be refined in lab settings for better real-world applicability. Expanding scalability and generalization across diverse genomic contexts will be key to enhancing the robustness of these models in CRISPR-Cas9 applications.

■ Conclusion

This study demonstrated the effectiveness of hybrid neural networks in predicting off-target effects in CRISPR-Cas9 gene editing. The results showed that with proper feature selection and engineering, even simpler models could outperform more complex architectures. Architecture 1's CNN-based models, particularly the CNN variant (M2), achieved the highest accuracy and precision in the HEK293 and K562 cell lines, proving the value of straightforward approaches. Architecture 2's integration of CNNs with sequential models like GRU and LSTM, exemplified by the CNN+GRU variant (M6), underscored the importance of sequence dependency, achieving the highest accuracy across the dataset. Architecture 3, with attention mechanisms, offered the most advanced prediction capability, particularly in the U2OS cell line with the CNN+LSTM+Attention model (M10), excelling in both accuracy and balanced accuracy. Compared to existing machine learning approaches, the use of attention layers in this study improved recall by allowing the model to focus on critical regions of the sequence, demonstrating a clear advantage over previous methods. This suggests that the hybrid model surpasses existing models in specific and general genomic environments, successfully achieving the aims outlined in the hypothesis.

In conclusion, the choice of architecture should be aligned with the specific application needs. For simpler, well-understood genomic contexts, straightforward CNN models are sufficient, providing robust predictions with minimal computational overhead. Conversely, complex models with attention mechanisms excel in identifying off-target effects in more nuanced and diverse genomic environments. These findings pave the way for developing more refined predictive tools to enhance the safety and effectiveness of CRISPR-Cas9 applications. Future research should prioritize optimizing these models for practical CRISPR-based therapeutic applications, focusing on refining threshold settings in laboratory environments to improve real-world efficacy. Additionally, addressing scalability and generalization across diverse genomic contexts will be key to further improving these models' robustness and ensuring their applicability in varied gene-editing scenarios.

DOI: 10.36838/v7i2.113

Availability and implementation

The code for models is available at: https://github.com/parshverma/CrisprHybridNN.

Acknowledgments

I want to thank Dallin Lowder, a Predoctoral Fellow at Baylor College of Medicine, very much for supporting me all through this research and being my mentor. The knowledge he had on gene editing and bioinformatics greatly contributed to the development of this study while his positive criticism served to improve on it. I also thank him for leading me through the research mazes, for spurring me on, and giving me some very useful ideas.

References

1. Modrzejewski, D., Hartung, F., Lehnert, H., Sprink, T., Kohl, C., Keilwagen, J., & Wilhelm, R. (2020). Which Factors Affect the Occurrence of Off-Target Effects Caused by the Use of CRISPR/Cas: A Systematic Review in Plants. *Frontiers in plant science*, *11*, 574959. https://doi.org/10.3389/fpls.2020.574959
2. Guo, C.; Ma, X.; Gao, F.; Guo, Y. Off-Target Effects in CRISPR/Cas9 Gene Editing. Frontiers in Bioengineering and Biotechnology **2023**, 11 (1143157). https://doi.org/10.3389/fbioe.2023.1143157
3. Lin, J., Zhang, Z., Zhang, S., Chen, J., & Wong, K. C. (2020). CRISPR-net: A recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels. Advanced Science, 7(13), 1903562. https://doi.org/10.1002/advs.201903562
4. Stemmer, M.; Thumberger, T.; del Sol Keyer, M.; Wittbrodt, J.; Mateo, J. L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLOS ONE* **2015**, *10* (4), e0124633. https://doi.org/10.1371/journal.pone.0124633.
5. Fu, Y.; Foden, J. A.; Khayter, C.; Maeder, M. L.; Reyon, D.; Joung, J. K.; Sander, J. D. High-Frequency Off-Target Mutagenesis Induced by CRISPR-Cas Nucleases in Human Cells. *Nature Biotechnology* **2013**, *31* (9), 822–826. https://doi.org/10.1038/nbt.2623.
6. Zarate, O. A.; Yang, Y.; Wang, X.; Wang, J.-P. BoostMEC: Predicting CRISPR-Cas9 Cleavage Efficiency through Boosting Models. *BMC Bioinformatics* **2022**, *23* (1). https://doi.org/10.1186/s12859-022-04998-z.
7. Chuai, G.; Ma, H.; Yan, J.; Chen, M.; Hong, N.; Xue, D.; Zhou, C.; Zhu, C.; Chen, K.; Duan, B.; Gu, F.; Qu, S.; Huang, D.; Wei, J.; Liu, Q. DeepCRISPR: Optimized CRISPR Guide RNA Design by Deep Learning. *Genome Biology* **2018**, *19* (1) https://doi.org/10.1186/s13059-018-1459-4.
8. Elkayam, S.; Ido Tziony; Orenstein, Y. DeepCRISTL: Deep Transfer Learning to Predict CRISPR/Cas9 On-Target Editing Efficiency in Specific Cellular Contexts. *Bioinformatics* **2024**. https://doi.org/10.1093/bioinformatics/btae481.
9. Ruan, Y., Wang, J., Yu, M., Wang, F., Wang, J., Xu, Y., Liu, L., Cheng, Y., Yang, R., Zhang, C., Yang, Y., Wang, J., Wu, W., Huang, Y., Tian, Y., Chen, G., Zhang, J., & Jian, R. (2023). A multi-omics integrative analysis based on CRISPR screens re-defines the pluripotency regulatory network in ESCs. Communications biology, 6(1), 410. https://doi.org/10.1038/s42003-023-04700-w
10. Störtz, F.; Minary, P. CrisprSQL: A Novel Database Platform for CRISPR/Cas Off-Target Cleavage Assays. Nucleic Acids Research 2021, 49 (D1), D855–D861. https://doi.org/10.1093/nar/gkaa885.

Authors

Parsh Verma is a high school student planning to major in Biology and Applied Mathematics. Parsh has a deep interest in exploring the intersections of science, technology, and mathematics, with a focus on solving complex problems through innovative approaches.