Computational Discovery of CRISPR-Cas13b Guide RNAs for Broad-Spectrum Dengue Virus Targeting

1st Syed Muhammad Ali Naqvi
Computer Science
Habib University
Karachi, Pakistan
sn07590@st.habib.edu.pk

2nd Dr. Farhan Khan
Electrical and Computer Engineering
Habib University
Karachi, Pakistan
farhan.khan@sse.habib.edu.pk

3rd Muslim Hussain
Electrical and Computer Engineering
Habib University
Karachi, Pakistan
sh07730@st.habib.edu.pk

***Abstract*—Dengue (DENV), an RNA virus, remains a significant global health threat, particularly in developing regions, with no widely effective antiviral therapy available. The CRISPR-Cas13b system, specifically the *PspCas13b* subtype, has emerged as a promising programmable antiviral tool capable of targeting viral RNA with high specificity. However, the efficacy of Cas13b-based interventions relies heavily on the design of potent and conserved CRISPR RNA (crRNA) spacer sequences, a task complicated by high viral genetic diversity. Unlike CRISPR-Cas9, which targets double-stranded DNA in eukaryotic genomes, Cas13b directly targets single-stranded RNA, making it ideally suited for RNA virus therapeutics; however, existing computational tools predominantly focus on Cas9 DNA targeting or Cas13d for mammalian transcript knockdown, leaving a significant gap for Cas13b-specific viral antiviral design.**

In this paper, we propose a computational pipeline and machine learning framework for the rational design of high-efficacy Cas13b guide RNAs targeting all four Dengue serotypes. Our approach integrates large-scale genomic data extraction, conservation analysis, and a novel *in silico* optimization module for guide RNA (gRNA) sequences, based on recently reported Cas13b design rules (e.g., 5' GG motif preference, Cytosine penalties). To predict targeting efficiency, we benchmark classical machine learning models (Random Forest, XGBoost) against foundation model-based predictors (Nucleotide Transformer, RNA-FM) using a dataset of experimentally validated spacers.

Our results demonstrate that classical feature-engineered models significantly outperform deep learning approaches when trained on experimentally validated gRNA datasets in low-data regimes. We identify highly conserved, optimized crRNA candidates, including several pan-serotype guides with predicted high potency. This work establishes a baseline for Cas13b efficiency prediction and provides a robust computational resource for accelerating the development of CRISPR-based antivirals against Dengue and other RNA viruses.

Code: https://github.com/muhammadali74/CAS13b_pipeline

***Index Terms*—CRISPR-Cas13b, Dengue Virus, Machine Learning, Guide RNA Design, Antiviral Therapy, Computational Biology.**

I. Introduction

Dengue Virus (DENV) remains one of the most critical mosquito-borne viral pathogens worldwide, posing a severe public health burden, particularly in developing nations such as Pakistan, India, and Brazil [1], [2]. With four distinct serotypes (DENV-1 through DENV-4), the virus causes millions of infections annually, ranging from mild fever to fatal hemorrhagic fever. Despite the prevalence of the disease, therapeutic options remain limited. Developing effective antivirals or vaccines is complicated by the high genetic diversity of the virus and the phenomenon of antibody-dependent enhancement (ADE), where sub-optimal antibodies can worsen symptoms upon secondary infection [3]. Consequently, there is an urgent need for precise, molecular-level interventions that can target conserved viral regions across all serotypes.

The CRISPR-Cas13 system has emerged as a revolutionary tool for RNA targeting. Unlike the DNA-targeting Cas9, Cas13 enzymes (such as Cas13a, Cas13b, and Cas13d) target single-stranded RNA, making them ideal candidates for combating RNA viruses like Dengue. Among these, the Cas13b subtype (specifically *PspCas13b*) has demonstrated high specificity and minimal off-target effects, rendering it a promising candidate for therapeutic and diagnostic applications [**?**]. However, the efficacy of Cas13b is highly dependent on the design of the CRISPR RNA (crRNA) spacer sequence. Selecting a 30-nucleotide (nt) spacer that is both conserved across viral mutations and highly efficient in triggering Cas13b activity is a multi-objective optimization problem that is computationally non-trivial.

Traditional selection of guide RNAs often relies on simple sequence alignment or tedious manual screening in the wet lab, which is cost-prohibitive and time-consuming. While Artificial Intelligence (AI) and Machine Learning (ML) have been successfully applied to Cas9 and Cas13d guide design [4]–[6], there is a distinct lack of computational frameworks specifically tailored for Cas13b, particularly models trained to navigate the constraints of viral conservation.

In this paper, we present an end-to-end computational pipeline and a baseline Machine Learning framework for the rational design of Cas13b crRNAs against Dengue Virus. Our approach integrates genomic data mining, conservation analysis, and predictive modeling to identify potent viral targets. We leverage recent findings regarding Cas13b's specific nucleotide preferences to optimize these spacers in silico. Furthermore, we address the challenge of limited experimental data by benchmarking classical ML models against large language model (LLM) embeddings (such as Nucleotide Transformer and RNA-FM), establishing a robust baseline for Cas13b efficiency prediction.

The main contributions of this work are as follows:

- A comprehensive data processing pipeline that identifies highly conserved 30-nt spacers across all four Dengue serotypes from public genomic databases.
- A novel optimization module that enhances crRNA efficacy by introducing strategic mismatches based on specific Cas13b design rules.
- A comparative analysis of ML models for predicting Cas13b target efficiency, demonstrating the effectiveness of feature engineering over deep learning embeddings on small-sample datasets.
- A catalog of top-ranked, optimized crRNA candidates suitable for experimental validation.

II. Literature Review

A. CRISPR-Cas13 as an Antiviral Strategy

The discovery of CRISPR-Cas systems adapted for RNA targeting has opened new avenues for antiviral therapy. While Cas9 targets DNA, the Class 2 Type VI CRISPR systems, characterized by the effector protein Cas13, possess RNase activity capable of cleaving single-stranded RNA [7]. Early applications utilized Cas13a (formerly C2c2) for pathogen detection, such as the SHERLOCK platform. More recently, focus has shifted toward Cas13b and Cas13d for therapeutic silencing due to their lack of protospacer flanking site (PFS) requirements in eukaryotic cells [8]. In the context of RNA viruses, Cas13 has been shown to effectively inhibit Influenza A, SARS-CoV-2, and Dengue in cell cultures, positioning it as a programmable antiviral agent [9].

B. Design Rules for PspCas13b

For a CRISPR system to be effective, the guide RNA must hybridize efficiently with the target. While Cas13 systems generally exhibit tolerance to mismatches, recent studies have elucidated specific design constraints for *PspCas13b*. Hu et al. [10] conducted a systematic screening of Cas13b crRNAs, revealing that secondary structure has a moderate effect on efficacy, while specific nucleotide motifs drive potency. Specifically, a "GG" motif at the 5' end of the spacer (positions 1-2) typically enhances activity, whereas Cytosine (C) residues at positions 1-4 and the central region (11-17) act as negative predictors. These findings imply that simply selecting guide RNAs based on conservation of the target viral sequence is insufficient; a pipeline must optimize sequences to satisfy these biochemical preferences, even if it requires introducing artificial mismatches, which Cas13b is known to tolerate [10].

C. Machine Learning in CRISPR Guide Design

The application of Machine Learning to predicting CRISPR guide efficiency is well-established for Cas9. Tools such as DeepCRISPR [4] and CRISPR-scan [11] utilize deep neural networks and extensive feature engineering to predict on-target efficiency and off-target risks. However, these models are not transferable to Cas13 due to fundamental differences in the cleavage mechanism and target substrate (RNA vs. DNA).

For Cas13 systems, recent tools like CASowary [12], and prediction models for Cas13d [13] [5] have begun to emerge. These typically rely on Convolutional Neural Networks (CNNs) trained on high-throughput screening data. However, for Cas13b specifically, public datasets remain scarce. The challenge is further compounded when applying large genomic foundation models. While Transformer-based models like Nucleotide Transformer [14] and RNA-FM [15] excel at capturing long-range genomic dependencies, their application to regression tasks on small, specialized datasets (such as Cas13b efficiency data) often results in overfitting or poor generalization compared to classical ensemble methods using domain-specific engineered features. This paper addresses this gap by developing a specialized pipeline that combines bio-informatic filtering with an ML model optimized for the specific constraints of Cas13b.

III. Methodology

Our proposed pipeline integrates bioinformatics filtering, conservation analysis, sequence optimization, and machine learning-based efficacy prediction to identify high-potency Cas13b crRNA candidates for Dengue Virus. The workflow consists of five main stages, as follows.

A. Data Acquisition and Preprocessing

Dengue Virus genome sequences were retrieved from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) database. The raw dataset comprised genomic sequences spanning all four Dengue serotypes (DENV-1 through DENV-4) as well as unclassified strains. To ensure data quality, sequences were filtered based on two criteria:

- **Ambiguity threshold:** Sequences containing more than 5% ambiguous nucleotides (denoted as 'N') were excluded to prevent spurious spacer extraction.
- **Minimum length:** Only genomes exceeding 1,000 base pairs were retained, as shorter fragments are unlikely to represent complete viral coding regions.

Each genome was annotated with its corresponding serotype by parsing the genome name metadata. The final filtered dataset was converted into FASTA format for downstream processing. Post-filtering statistics are presented in Table I.

TABLE I
Genome Statistics After Quality Filtering

Serotype	Count
DENV-1	14906
DENV-2	11975
DENV-3	4897
DENV-4	2868
Unclassified	3411
Total	38057

B. Spacer Extraction and Conservation Analysis

To identify candidate crRNA spacers, we employed a sliding window approach to extract all possible 30-nucleotide (nt) subsequences from each viral genome. A memory-efficient SQLite-based storage system was implemented to handle the

large combinatorial space of candidate spacers across thousands of genomes.

For each unique 30-nt spacer:

- We tracked the set of genome identifiers in which it appeared, avoiding redundant counting within the same genome.
- Conservation percentage was computed as:

$$\text{Conservation} = \frac{\text{No. of genomes containing spacer}}{\text{Total genomes}} \times 100$$

- Subtype-specific occurrence counts (DENV-1, DENV-2, DENV-3, DENV-4, Unclassified) were calculated by mapping genome IDs to their serotype annotations.

This process yielded millions of unique spacers. To focus on the most promising candidates, we selected the top-300 most frequently occurring spacers per serotype, ensuring broad coverage across all viral variants.

C. Sequence Filtering and Quality Control

Each candidate spacer was annotated with biochemical properties known to influence Cas13b activity:

- **GC Content:** Computed as the percentage of Guanine and Cytosine nucleotides. Extreme GC values can affect transcription and secondary structure stability.
- **Homopolymer Runs:** Binary flag indicating the presence of four or more consecutive identical nucleotides, which can cause transcriptional termination or synthesis artifacts.

While these features were computed for all spacers, strict filtering was not applied at this stage, as subsequent optimization and machine learning prediction would further refine candidate selection.

D. Cas13b-Specific Sequence Optimization

Recent functional studies [10] have revealed that PspCas13b activity is strongly influenced by position-specific nucleotide preferences. We adapted the scoring metric the authors used for our optimization algorithm given as follows:

- A Guanine (G) at positions 1 or 2 (5' end) confers a score bonus of +20 each.
- Cytosine (C) at positions 1–4 incurs a penalty of -20 each.
- Cytosine at positions 11, 12, 15, 16, and 17 incurs a penalty of -5 each.

Based on this scoring scheme, we developed an optimization algorithm that introduces up to three strategic nucleotide substitutions to maximize the Cas13b potency score while maintaining biological relevance. The algorithm operates as follows:

1) Compute the baseline score for the original spacer.
2) Identify all positions where beneficial mutations exist (e.g., replacing C with G at penalized positions).
3) Generate all combinations of 1, 2, or 3 mutations.
4) Score each variant and retain those achieving the maximum score.
5) If multiple variants share the top score, all are preserved for downstream analysis.

This strategy leverages the known mismatch tolerance of Cas13b [10], which can accommodate up to 3–4 non-consecutive mismatches without significant loss of activity.

E. Machine Learning Model for Efficiency Prediction

1) Training Dataset: A curated dataset of 245 Cas13b spacer sequences with experimentally validated mean targeting efficiencies was used for model training. The data were extracted from the supplementary data of [10], and the specific source table, along with the spacers used in this study, is provided in Table IV. This dataset represents high-throughput screening results from previous studies.

2) Feature Engineering: Given the limited sample size (n=245), we employed extensive feature engineering to encode domain-specific knowledge:

- **Position-Specific One-Hot Encoding:** Each of the 30 nucleotide positions was encoded as a 4-dimensional binary vector (A, T, G, C), yielding 120 binary features.
- **Motif Features:** Explicit Boolean features for critical motifs (e.g., 'GG' at positions 1–2, C-count at positions 1–4).
- **Sequence Composition:** GC content, nucleotide fractions, and di-nucleotide frequencies.

A complete list of the 157 engineered features is provided in the Table V in supplementary materials.

3) Model Selection and Training: We compared multiple regression models suitable for small, high-dimensional datasets:

- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost Regressor
- Support Vector Regressor (SVR)
- Ridge Regression

Hyperparameter tuning was performed using 5-fold cross-validation with grid search, optimizing for R^2 and minimizing overfitting. The final model was trained on all 245 samples and saved for inference.

4) Deep Learning Experiments: To explore whether large pre-trained genomic foundation models could improve upon classical ML approaches, we conducted two additional experiments using state-of-the-art sequence embeddings. The motivation for this approach stems from the success of transformer-based language models like BERT and GPT) in NLP tasks and their recent adaptation to genomic sequences [14], [15] . Such models, trained on billions of nucleotides, may capture complex sequence patterns and biophysical properties that are difficult to encode manually. We hypothesized that these learned representations could provide richer feature spaces than hand-crafted features, potentially improving predictive power for Cas13b efficiency.

We experimented with two foundation models:

1) **Nucleotide Transformer (NT):** A BERT-style transformer pre-trained on genomic DNA sequences was used to extract 512-dimensional embeddings for each spacer. A lightweight regression head (2-layer MLP) was trained on these embeddings.

2) **RNA-FM:** A foundation model pre-trained on RNA sequences was similarly used for embedding extraction, followed by regression head training.

Both approaches were trained with heavy regularization (dropout=0.3, weight decay=0.01) and early stopping to mitigate overfitting.

F. Candidate Ranking and Selection

The trained machine learning model was applied to all optimized spacer sequences. Candidates were ranked by:

1) Predicted targeting efficiency
2) Conservation across serotypes
3) Cas13b position score

For each serotype, we identified:

- The top-3 most conserved spacers (broad-spectrum candidates)
- All spacers with predicted efficiency ≥ 0.8 (high-potency candidates)

IV. Experiments and Results

A. Spacer Extraction and Conservation

The sliding window extraction across all filtered genomes yielded 3.78 million unique 30-nt spacers. After subtype-specific conservation ranking and selection of the top-300 per serotype, we retained 1740 candidates for optimization.

B. Sequence Optimization Results

The Cas13b-specific optimization algorithm significantly improved spacer quality. Of the 1,740 input spacers, 1,714 sequences (98.5%) received at least one beneficial mutation, while 26 sequences (1.5%) were already optimal according to the scoring scheme.

The optimization yielded substantial improvements in predicted potency:

- **Average Score Improvement:** +47.17 points (SD = 20.37)
- **Maximum Score Improvement:** +100 points

The majority of spacers required multiple mutations to achieve maximum potency: 9.9% required one mutation, 23.7% required two mutations, and 64.9% required all three allowed mutations. Score improvement correlated monotonically with mutation count, yielding average improvements of 17.8, 40.2, and 55.6 points for one, two, and three mutations, respectively.

Position-specific analysis revealed that 73.3% of all mutations occurred at positions 1–2 (introducing the favorable 'GG' motif), while the remaining mutations targeted penalized Cytosine residues in positions 3–4 and the central region (11–12, 15–17). Notably, conservation percentage did not predict optimization potential: highly conserved spacers ($\geq$80%) benefited equally from optimization as less conserved candidates, indicating that sequence conservation and biochemical potency represent orthogonal selection criteria. Figure 1 illustrates these optimization dynamics across all dimensions of the analysis.

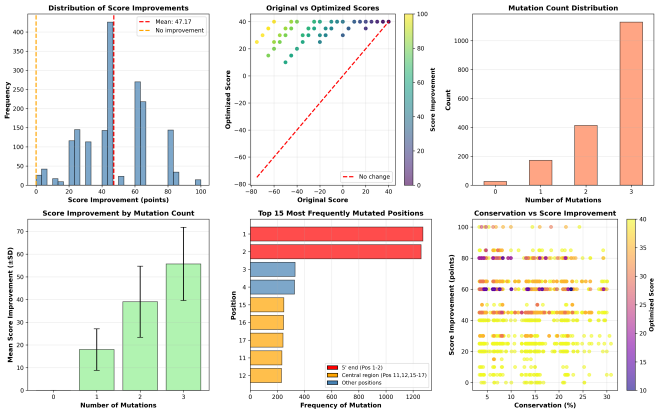

Fig. 1. Comprehensive analysis of Cas13b-specific sequence optimization. (a) Distribution of score improvements showing mean of +47.17 points. (b) Scatter plot of original vs optimized scores colored by improvement magnitude. (c) Distribution of mutation counts per spacer. (d) Mean score improvement stratified by mutation count. (e) Top 15 most frequently mutated positions, with red indicating 5' end (positions 1–2) and orange indicating central penalty region (positions 11–12, 15–17). (f) Conservation vs score improvement, demonstrating orthogonal relationship between sequence conservation and optimization potential.

C. Machine Learning Model Performance

1) Classical ML Models: Table II summarizes the 5-fold cross-validation performance of all tested models.

TABLE II
Machine Learning Model Comparison (5-Fold CV)

Model	Mean R^2	Mean MAE	Mean Spearman
XGBoost	0.235	0.161	0.529
Gradient Boosting	0.199	0.220	0.497
Random Forest	0.192	0.221	0.480
SVR	0.227	0.217	0.533
Ridge Regression	0.139	0.231	0.494

Best Model: XGBoost achieved the highest cross-validated R^2 of 0.235, a lowest MAE of 0.161 and a high Spearman correlation of 0.529, indicating moderate predictive power suitable for ranking candidate spacers.

2) Deep Learning Model Performance: Despite leveraging large pre-trained models, the deep learning approaches underperformed relative to classical ML. This is most likely due to the insufficient data needed for training a deep-learning models

- **Nucleotide Transformer (NT-DL):**
  - Mean R^2: 0.0181
  - Mean Spearman: 0.2607
- **RNA-FM:**
  - Mean R^2: 0.1131
  - Mean Spearman: 0.3429

RNA-FM showed modest improvement over NT, likely due to its RNA-specific pre-training, but both failed to match the engineered-feature baseline. This outcome is consistent with the literature: foundation models require substantial task-specific data to outperform domain-engineered features in low-data regimes [14], [16], [17].

D. Feature Importance Analysis

SHAP (SHapley Additive exPlanations) analysis was performed on the best-performing model to interpret feature contributions. The analysis revealed that Cas13b efficiency is driven by a hierarchy of sequence composition features rather than single-position identities alone. The most influential features (see Figure 2) were:

1) **CC dinucleotide frequency (`dinuc_CC`):** This was the single most dominant predictor. High values (red dots) had a massive negative impact on predicted efficiency (SHAP value $\leq$ -0.04), confirming the strong deleterious effect of Cytosine enrichment.
2) **Trinucleotide context:** Specific trimers such as `AAT`, `GTT`, and `GAG` were highly predictive, suggesting that local sequence context significantly modulates activity.
3) **Overall Cytosine count (`count_C`):** Consistently showed a negative correlation with efficacy, reinforcing the rule that C-rich spacers perform poorly.
4) **G at Position 1 (`G_pos1`):** Confirmed as a top positive predictor, validating the 5' G preference identified in prior literature.
5) **Specific positional nucleotides:** Features such as `G_pos7`, `T_pos13`, and `A_pos2` appeared as key modulators, indicating that positions beyond the canonical 5' end also contribute to spacer potency.

These findings align with and extend the experimental results from Hu et al. [10]. While previous studies emphasized specific penalty positions, our model learned that global Cytosine content (captured by `dinuc_CC` and `count_C`) is the primary determinant of low efficiency, providing a broader heuristic for guide design.

E. Final Candidate crRNAs

Post-optimization, we filtered the 1,740 unique spacers to identify the top-performing candidates specific to each Dengue serotype. This selection process yielded between 269 and 396 high-confidence candidates per serotype (approximately 15–22% of the total pool), prioritizing those with the highest frequency of occurrence in their respective genomic populations.

Figure 3 illustrates the relationship between conservation coverage and predicted targeting efficiency. We observed distinct landscape differences across serotypes:

- **DENV-3 and DENV-4 (High Convergence, Limited Representation):** Although these serotypes exhibited the most promising dual-purpose candidates with simultaneous high conservation and efficiency, they were underrepresented in the dataset. DENV-3 and DENV-4 accounted for 22.8% and 21.5% of the optimized candidates, respectively, compared to the 20% representation of larger serotypes. For DENV-4, the most conserved spacer (`GGAACA...`) covers 80.51% of all available genomes while maintaining a high predicted efficiency of 0.836. Similarly, the top DENV-3 candidate covers 69.88% of genomes with a robust efficiency of 0.859. These candidates demonstrate ideal therapeutic potential but would benefit from validation against a larger, more diverse strain collection.
- **DENV-1 (Balanced Representation):** With 349 candidates (20.1% of the pool), DENV-1 showed a dense cluster of high-efficiency spacers across moderate conservation ranges. The most conserved candidate covers 68.42% of genomes with a predicted efficiency of 0.804, offering a strong balance for clinical application. The highest-efficiency candidate achieves 0.881, demonstrating that DENV-1 spacers can be optimized for either criterion without severe trade-offs.

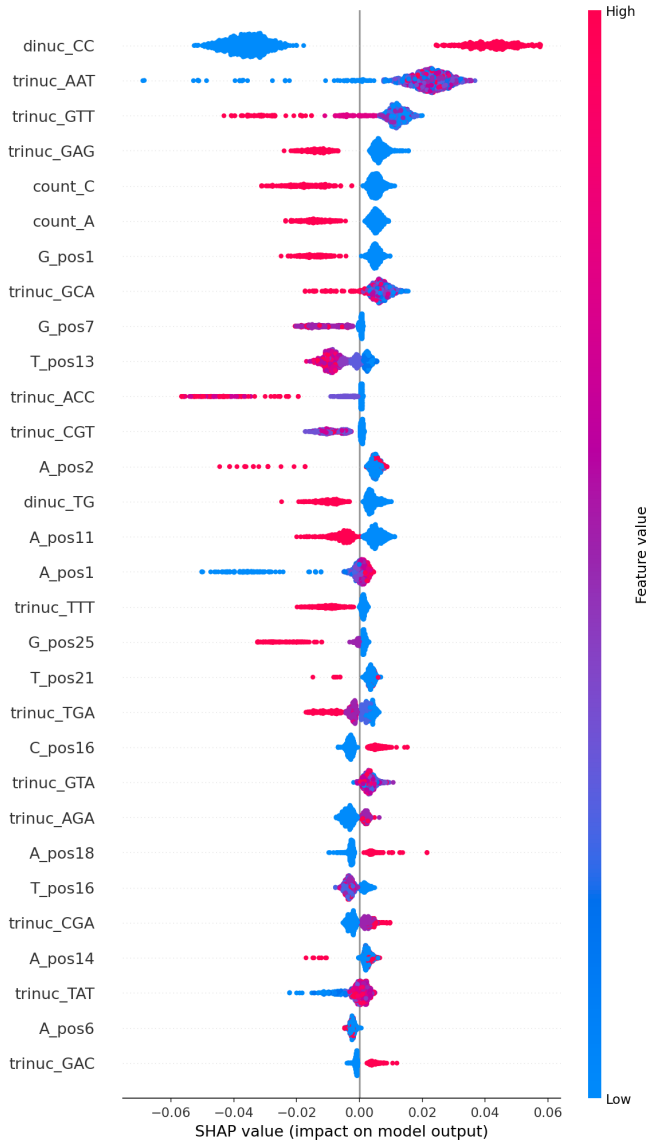

Fig. 2. SHAP summary plot ranking the top 30 features by mean absolute SHAP value. Red dots indicate high feature values, while blue dots indicate low values. The x-axis shows the impact on predicted Cas13b efficiency (positive = higher efficiency). The plot reveals that dinucleotide CC frequency (`dinuc_CC`) is the single most dominant predictor, strongly negatively correlated with efficiency, aligning with known Cas13b preference against Cytosine-rich motifs.

- **DENV-2 (Trade-off with Compensatory Options):** DENV-2 (352 candidates, 20.2% of pool) presented a notable divergence between conservation and potency. The most conserved spacer (68.75% coverage) had a markedly lower predicted efficiency of 0.613. However, the highest-efficiency candidate (GGAGAC . . . , efficiency 0.878) covers only 42.36% of genomes. Importantly, subsequent high-ranking candidates offer a viable middle ground: the fourth-ranked candidate achieves a mean targeting efficiency of 0.832 while maintaining reasonable coverage, presenting a practical compromise for serotype-specific applications. For maximal pan-serotype coverage, we recommend either the efficiency-optimized guide or a multi-guide strategy incorporating both high-efficiency and high-conservation targets.

Table III lists the specific sequences for the optimal candidates identified in Figure 3. More candidate rankings are provided in Table VI.

TABLE III
BEST CANDIDATE CRRNAS PER SEROTYPE

Serotype	Selection Criteria	Sequence (Optimized)	Pred. Eff.	Coverage
2*DENV-1	Most Conserved	GGAAGAACAAGACGGGAACTTTGTGTGTCG	0.804	68.42%
	Best Efficiency	GGAGTAGTCGTACTAGGATCACAAGAAGGA	0.881	45.94%
2*DENV-2	Most Conserved	GGGAGTCAACATAGAAGCAGAACCTCCATT	0.613	68.75%
	Best Efficiency	GGAGACAGAGGATGGGGAAATGGATGTGGA	0.878	42.36%
2*DENV-3	Most Conserved	GGTGCAAGGCGGATGGGCATCTTGGGAGAC	0.859	69.88%
	Best Efficiency	GGGGTTGGATCGCAAGAGGGAGCAATGCAC	0.882	41.41%
2*DENV-4	Most Conserved	GGAACAAGATGTCCAAGGCAAGGAGAGCCT	0.836	80.51%
	Best Efficiency	GGGGTAGAACATGGAGGATGCGTCACAACC	0.888	47.77%

V. DISCUSSION AND ANALYSIS

A. Biological Significance of Identified Candidates

The computational pipeline successfully identified serotype-specific crRNA spacers with distinct biological properties tailored to Cas13b targeting:

- **Highly conserved targets:** Most candidates maintain $\geq 45\%$ representation within their respective serotype genomes, with optimal candidates reaching 68% coverage. This high conservation reduces the likelihood of viral escape through mutation and suggests that targeted regions encode functionally constrained viral elements (likely essential for replication or virulence).
- **Optimized for Cas13b biochemistry:** The extensive optimization phase (yielding 98.5% improvement rate with mean +47.17 points) demonstrates that even moderately suboptimal natural sequences can be substantially enhanced through strategic nucleotide substitutions. The algorithm-generated guides exploit known Cas13b preferences (5' GG motifs, central C-elimination) while maintaining high specificity through the model's learned features, which prioritize position-specific context and dinucleotide composition over simple global metrics.
- **Predicted to be highly effective:** Machine learning predictions identified top candidates. These predictions were informed by 158 engineered features validated against a curated dataset of 245 experimentally characterized Cas13b spacers, providing robust estimates grounded in observed biochemical principles.

B. Performance of Classical ML vs. Deep Learning

Our results demonstrate that in data-scarce regimes (n=245), **feature-engineered classical ML models significantly outperform deep learning approaches**. This finding has important implications:

- **Sample efficiency:** Random Forest and XGBoost effectively leverage domain knowledge encoded in features, whereas foundation models (NT, RNA-FM) rely on emergent patterns from large-scale pre-training that may not align with task-specific objectives.
- **Interpretability:** SHAP analysis of tree-based models directly reveals which biochemical properties drive predictions, facilitating biological insight and model refinement.
- **Future potential:** As larger Cas13b efficiency datasets become available (e.g., from high-throughput screens), fine-tuning foundation models may become competitive. Our negative results establish a baseline for comparison.

C. Comparison with Existing Guide Design Tools

To our knowledge, this is the **first computational framework specifically tailored for Cas13b crRNA design against viral genomes**. Existing tools like CASowary [12] target Cas13a/d for mammalian transcript knockdown, while platforms such as CHOPCHOP [18] focus on Cas9. Our pipeline addresses the unique constraints of antiviral applications:

- Conservation analysis across viral serotypes (not applicable to fixed genomes)
- Mismatch tolerance and optimization (viral diversity)
- Integration of virus-specific and enzyme-specific design rules

D. Limitations and Future Work

- **Experimental validation:** The predicted candidates require wet-lab testing to confirm efficacy in Dengue-infected cells.
- **Off-target analysis:** While our spacers are specific to Dengue genomes, computational screening against the human transcriptome is recommended before therapeutic application.
- **Expanded training data:** Acquiring more Cas13b efficiency data (especially for viral targets) would enable more robust ML models and potentially unlock the advantages of foundation models.
- **Delivery mechanisms:** Translation to therapy requires development of efficient Cas13b delivery systems, which is beyond the scope of this work.

VI. CONCLUSION

We have presented a comprehensive computational pipeline for the rational design of Cas13b crRNAs targeting Dengue Virus, addressing a critical gap in antiviral therapeutic development. By integrating genomic data extraction, sequence optimization, and machine learning-based efficacy prediction, we identified a catalog of high-confidence candidate guide RNAs suitable for immediate experimental validation.

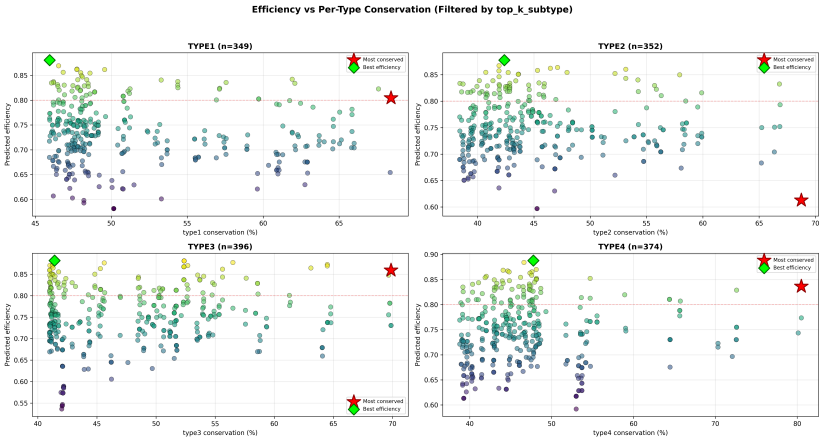

Fig. 3. Predicted efficiency versus serotype-specific conservation for the top candidate crRNAs. Each point represents a unique spacer optimized for DENV-1 ($n = 349$), DENV-2 ($n = 352$), DENV-3 ($n = 396$), and DENV-4 ($n = 374$). The **Green Diamond** marks the candidate with the highest predicted efficiency, while the **Red Star** marks the candidate with the highest conservation coverage for that serotype. The red dashed line indicates the high-efficiency threshold (≥ 0.8).

Our work establishes the first baseline for Cas13b efficiency prediction using classical machine learning, demonstrating superior performance over large language model embeddings in small-data settings. The identified candidates offer a promising starting point for developing broad-spectrum Cas13b-based antivirals against Dengue, with potential applicability to other RNA viruses.

Future efforts would focus on experimental validation of top candidates, expansion of the training dataset to improve model performance, and extension of the pipeline to other RNA viruses such as Chikungunya, and SARS-CoV-2 relevant to public health.

Acknowledgment

We extend our sincere gratitude to Dr. Syed Faraz Ahmed, Department of Electrical and Electronic Engineering, University of Melbourne, for his careful review of this manuscript, constructive suggestions for improvement, and invaluable support in data curation and analysis. His insightful feedback was valuable in this work

References

[1] S. Ali *et al.*, "Dengue outbreak and fragile healthcare system: Doctors at the verge of mental and physical stress," *Frontiers in Public Health*, vol. 10, p. 966042, 2022.

[2] J. D. Stanaway *et al.*, "The global burden of dengue: an analysis from the global burden of disease study 2013," *The Lancet Infectious Diseases*, vol. 16, no. 6, pp. 712–723, 2016.

[3] L. C. Katzelnick, L. Gresh, M. E. Halloran, J. C. Mercado, G. Kuan, A. Gordon, A. Balmaseda, and E. Harris, "Antibody-dependent enhancement of severe dengue disease in humans," *Science*, vol. 358, no. 6365, pp. 929–932, 2017.

[4] G. Chuai, H. W. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan *et al.*, "Deepcrispr: optimized crispr guide rna design by deep learning," *Genome biology*, vol. 19, no. 1, pp. 1–18, 2018.

[5] X. Cheng, S. Zhang, J. Zhang, W. Chen, Z. Zhu, X. Liu, J. Chen, J. Wang, L. Wang, G. Zhao *et al.*, "Prediction of on-target and off-target activity of crispr–cas13d guide rnas using deep learning," *Nature Communications*, vol. 14, no. 1, pp. 1–14, 2023.

[6] Y. Pan *et al.*, "Transitioning from wet lab to artificial intelligence: a systematic review of ai predictors in crispr," *Briefings in Bioinformatics*, 2025, pMCID: PMC11796103.

[7] O. O. Abudayyeh, J. S. Gootenberg, S. Konermann, J. Joung, I. M. Slaymaker, D. B. T. Cox, S. Shmakov, K. S. Makarova, E. Semenova, L. Minakhin *et al.*, "C2c2 is a single-component programmable rna-guided rna-targeting crispr effector," *Science*, vol. 353, no. 6299, p. aaf5573, 2016.

[8] D. B. T. Cox, J. S. Gootenberg, O. O. Abudayyeh, B. Franklin, M. J. Kellner, J. Joung, and F. Zhang, "Rna editing with crispr-cas13," *Science*, vol. 358, no. 6366, pp. 1019–1027, 2017.

[9] C. A. Freije, C. Myhrvold, M. Boehm, A. E. Lin, N. L. Welch, A. Carter, H. C. Metsky, Y. Luo, O. O. Abudayyeh, J. S. Gootenberg *et al.*, "Programmable inhibition and detection of rna viruses using cas13," *Molecular cell*, vol. 76, no. 5, pp. 826–837, 2019.

[10] W. Hu, A. Kumar, S. F. Ahmed, S. Qi, D. K. G. Ma, H. Chen, G. J. Singh, J. M. L. Casan, M. Haber, I. Voskoboinik, M. R. McKay, J. A. Trapani, P. G. Ekert, and M. Fareh, "Single-base tiled screen unveils design principles of pspcas13b for potent and off-target-free rna silencing," *Nature Structural & Molecular Biology*, vol. 31, pp. 1702–1716, 2024. [Online]. Available: https://doi.org/10.1038/s41594-024-01336-0

[11] M. A. Moreno-Mateos, C. E. Vejnar, J.-D. Beaudoin, J. P. Fernandez, E. K. Mis, M. K. Khokha, and A. J. Giraldez, "Crisprscan: designing highly efficient sgrnas for crispr-cas9 targeting in vivo," *Nature methods*, vol. 12, no. 10, pp. 982–988, 2015.

[12] H.-K. Kim, S. Min, M. Song, S. Jung, J. W. Choi, Y. Kim, S. Lee, S. Yoon, and H. H. Kim, "Casowary: Crispr-cas13 guide rna predictor for transcript depletion," *Nucleic acids research*, vol. 49, no. 3, pp. e16–e16, 2021.

[13] H.-H. Wessels, A. Méndez-Mancilla, X. Guo, M. Legut, Z. Daniloski, and N. E. Sanjana, "Massively parallel cas13 screens reveal principles for guide rna design," *Nature biotechnology*, vol. 38, no. 6, pp. 722–727, 2020.

[14] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. L. Carranza, A. H. Grzak, F. Oteri, T. Vatanen, K. Harris, J. K. Pritchard *et al.*, "The nucleotide transformer: Building and evaluating robust foundation models for human genomics," *bioRxiv*, pp. 2023–01, 2023.

[15] J. Chen, Z. Hu, S. Sun, Q. Tan, Y. Wang, Z. Yu, L. Zong, L. Hong, J. Xiao, I. King *et al.*, "Rna-fm: a foundation model for rna structure and function prediction," *arXiv preprint arXiv:2204.00300*, 2022.

[16] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in Neural Information Processing Systems*, vol. 35, pp. 507–520, 2022.

[17] L. Brigato and E. Iocchi, "A close look at deep learning with small data," *25th International Conference on Pattern Recognition (ICPR)*, pp. 2490–2497, 2021.

[18] K. Labun, T. G. Montague, M. Krause, Y. N. Torres Cleuren, H. Beskemeier, and E. Valen, "Chopchop v3: expanding the crispr web toolbox beyond genome editing," *Nucleic acids research*, vol. 47, no. W1, pp. W171–W174, 2019.

VII. Supplementary Material

TABLE IV: List of 245 crRNA spacers used to train the Cas13b efficiency models obtained from source data of [10]. The spacers can be found in the supplementary table 2 of [10] which lists the spacers along with the figures reporting their experimental efficacy values.

Name	Spacer (5'–3')	Source Figure/Table (Hu et al. (2024)
PspmCherry1 (matching position 2)	CCAUGUUAUCCUCCUCGCCCUUGCUCACCA	Fig.1d; Extended Data Fig.1a-1b
PspmCherry2 (matching position 3)	GGCCATGTTATCCTCCTCGTCCATGCTCAC	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
PspmCherry3 (matching position 49)	GGAGCCCUCCAUGUGCACCUUGAAGCGCAU	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
PspmCherry4 (matching position 52)	CACGGAGCCCUCCAUGUGCACCUUGAAGCG	Fig.1d; Extended Data Fig.1a-1b
PspmCherry5 (matching position 118)	GGCGGUCUGGGUGCCCUCGUAGGGGCGGCC	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
PspmCherry6 (matching position 193)	GGAGCCGUACAUGAACUGAGGGGACAGGAU	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
PspmCherry7 (matching position 250)	GGGGAAGGACAGCUUCAAGUAGUCGGGGAU	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
PspmCherry8 (matching position 252)	UCGGGGAAGGACAGCUUCAAGUAGUCGGGG	Fig.1d; Extended Data Fig.1a-1b
PspmCherry10 (matching position 319)	GGAGUCCUGGGUCACGGUCACCACGCCGCC	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
PspmCherry9 (matching position 315)	UCCUGGGUCACGGUCACCACGCCGCCGUCC	Fig.1d; Extended Data Fig.1a-1b
PspmCherry11 (matching position 406)	GGUCUUCUUCUGCAUUACGGGGCCGUCGGA	Fig.1d; Extended Data Fig.1a-1b; Extended Data Fig. 5a
Pspmcherry12(-15nt)	UGAUCUCGCCCUUCAGGGCGCCGUCCUCGG	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(-12nt)	UCUCGCCCUUCAGGGCGCCGUCCUCGGGGU	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(-9nt)	CGCCCUUCAGGGCGCCGUCCUCGGGGUACA	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(-6nt)	CCUUCAGGGCGCCGUCCUCGGGGUACAUCC	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(-3nt)	UCAGGGCGCCGUCCUCGGGGUACAUCCGCU	Fig. 2a; Extended Data Fig. 2b
PspmCherry12 (matching position 455)	GGGCGCCGUCCUCGGGGUACAUCCGCUCGG	Fig.1d-1g; Extended Data Fig.1a-1d; Extended Data Fig. 5a-5c; Extended Data Fig. 8c-8d; Extended Data Fig. 9a-9c
Pspmcherry12(+3nt)	CGCCGUCCUCGGGGUACAUCCGCUCGGAGG	Fig. 2a; Extended Data Fig. 2b
Pspmcherr12(+6nt)	CGUCCUCGGGGUACAUCCGCUCGGAGGAGG	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(+9nt)	CCUCGGGGUACAUCCGCUCGGAGGAGGCCU	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(+12nt)	CGGGGUACAUCCGCUCGGAGGAGGCCUCCC	Fig. 2a; Extended Data Fig. 2b
Pspmcherry12(+15nt)	GGUACAUCCGCUCGGAGGAGGCCUCCCAGC	Fig. 2a; Extended Data Fig. 2b
PspmCherry13 (matching position 526)	GGUGGUCUUGACCUCAGCGUCGUAGUGGCC	Fig.1d; Extended Data Fig.1a-1b
PspmCherry14 (matching position 593)	UCCAACUUGAUGUUGACGUUGUAGGCGCCG	Fig.1d; Extended Data Fig.1a-1b
PspmCherry15 (matching position 585)	GGUGAUGUCCAACUUGAUGUUGACGUUGUA	Fig.1d; Extended Data Fig.1a-1b
Pspmcherry16(-15nt)	CUCGUCCAUGCCGCCGGUGGAGUGGCGGCC	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(-12nt)	GUCCAUGCCGCCGGUGGAGUGGCGGCCCUC	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(-9nt)	CAUGCCGCCGGUGGAGUGGCGGCCCUCGGC	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(-6nt))	GCCGCCGGUGGAGUGGCGGCCCUCGGCGCG	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(-3nt)	GCCGGUGGAGUGGCGGCCCUCGGCGCGUUC	Fig. 2a; Extended Data Fig. 2c
PspmCherry16 (matching position 655)	GGUGGAGUGGCGGCCCUCGGCGCGUUCGUA	Fig.1d; Extended Data Fig.1a-1b
Pspmcherry16(+3nt)	GGAGUGGCGGCCCUCGGCGCGUUCGUACUG	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(+6nt)	GUGGCGGCCCUCGGCGCGUUCGUACUGUUC	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(+9nt)	GCGGCCCUCGGCGCGUUCGUACUGUUCCAC	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(+12nt)	GCCCUCGGCGCGUUCGUACUGUUCCACGAU	Fig. 2a; Extended Data Fig. 2c
Pspmcherry16(+15nt)	CUCGGCGCGUUCGUACUGUUCCACGAUGGU	Fig. 2a; Extended Data Fig. 2c
BCR-ABL1P190 crRNA1	GGCUCAAAGUCAGAUGCUACUGGCCGCUGA	Fig. 2c
BCR-ABL1P190 crRNA2	GCUCAAAGUCAGAUGCUACUGGCCGCUGAA	Fig. 2c
BCR-ABL1P190 crRNA3	CUCAAAGUCAGAUGCUACUGGCCGCUGAAG	Fig. 2c
BCR-ABL1P190 crRNA4	UCAAAGUCAGAUGCUACUGGCCGCUGAAGG	Fig. 2c
BCR-ABL1P190 crRNA5	CAAAGUCAGAUGCUACUGGCCGCUGAAGGG	Fig. 2c
BCR-ABL1P190 crRNA6	AAAGUCAGAUGCUACUGGCCGCUGAAGGGC	Fig. 2c

Continued on next page

Name	Spacer (5'–3')	Source Figure (Hu et al. (2024)
BCR-ABL1P190 crRNA7	AAGUCAGAUGCUACUGGCCGCUGAAGGGCU	Fig. 2c
BCR-ABL1P190 crRNA8	AGUCAGAUGCUACUGGCCGCUGAAGGGCUU	Fig. 2c
BCR-ABL1P190 crRNA9	GUCAGAUGCUACUGGCCGCUGAAGGGCUUC	Fig. 2c
BCR-ABL1P190 crRNA10	UCAGAUGCUACUGGCCGCUGAAGGGCUUCU	Fig. 2c
BCR-ABL1P190 crRNA11	CAGAUGCUACUGGCCGCUGAAGGGCUUCUG	Fig. 2c
BCR-ABL1P190 crRNA12	AGAUGCUACUGGCCGCUGAAGGGCUUCUGC	Fig. 2c
BCR-ABL1P190 crRNA13	GAUGCUACUGGCCGCUGAAGGGCUUCUGCG	Fig. 2c
BCR-ABL1P190 crRNA14	AUGCUACUGGCCGCUGAAGGGCUUCUGCGU	Fig. 2c
BCR-ABL1P190 crRNA15	UGCUACUGGCCGCUGAAGGGCUUCUGCGUC	Fig. 2c
BCR-ABL1P190 crRNA16	GCUACUGGCCGCUGAAGGGCUUCUGCGUCU	Fig. 2c
BCR-ABL1P190 crRNA17	CUACUGGCCGCUGAAGGGCUUCUGCGUCUC	Fig. 2c
BCR-ABL1P190 crRNA18	UACUGGCCGCUGAAGGGCUUCUGCGUCUCC	Fig. 2c
BCR-ABL1P190 crRNA19	ACUGGCCGCUGAAGGGCUUCUGCGUCUCCA	Fig. 2c
BCR-ABL1P190 crRNA20	CUGGCCGCUGAAGGGCUUCUGCGUCUCCAU	Fig. 2c
BCR-ABL1P190 crRNA21	UGGCCGCUGAAGGGCUUCUGCGUCUCCAUG	Fig. 2c;Fig. 7b-7m
BCR-ABL1P190 crRNA22	GGCCGCUGAAGGGCUUCUGCGUCUCCAUGG	Fig. 2c
BCR-ABL1P190 crRNA23	GCCGCUGAAGGGCUUCUGCGUCUCCAUGGA	Fig. 2c
BCR-ABL1P190 crRNA24	CCGCUGAAGGGCUUCUGCGUCUCCAUGGAA	Fig. 2c
BCR-ABL1P190 crRNA25	CGCUGAAGGGCUUCUGCGUCUCCAUGGAAG	Fig. 2c
BCR-ABL1P190 crRNA26	GCUGAAGGGCUUCUGCGUCUCCAUGGAAGG	Fig. 2c
BCR-ABL1P190 crRNA27	CUGAAGGGCUUCUGCGUCUCCAUGGAAGGC	Fig. 2c
BCR-ABL1P190 crRNA28	UGAAGGGCUUCUGCGUCUCCAUGGAAGGCG	Fig. 2c
BCR-ABL1P190 crRNA29	GAAGGGCUUCUGCGUCUCCAUGGAAGGCGC	Fig. 2c
BCR-ABL1P190 crRNA30	AAGGGCUUCUGCGUCUCCAUGGAAGGCGCC	Fig. 2c
BCR-ABL1P190 crRNA31	AGGGCUUCUGCGUCUCCAUGGAAGGCGCCC	Fig. 2c
BCR-ABL1P190 crRNA32	GGGCUUCUGCGUCUCCAUGGAAGGCGCCCU	Fig. 2c
BCR-ABL1P190 crRNA33	GGCUUCUGCGUCUCCAUGGAAGGCGCCCUC	Fig. 2c
BCR-ABL1P190 crRNA34	GCUUCUGCGUCUCCAUGGAAGGCGCCCUCG	Fig. 2c
BCR-ABL1P190 crRNA35	CUUCUGCGUCUCCAUGGAAGGCGCCCUCGC	Fig. 2c
BCR-ABL1P190 crRNA36	UUCUGCGUCUCCAUGGAAGGCGCCCUCGCC	Fig. 2c
BCR-ABL1P190 crRNA37	UCUGCGUCUCCAUGGAAGGCGCCCUCGCCA	Fig. 2c
BCR-ABL1P190 crRNA38	CUGCGUCUCCAUGGAAGGCGCCCUCGCCAU	Fig. 2c
BCR-ABL1P190 crRNA39	UGCGUCUCCAUGGAAGGCGCCCUCGCCAUC	Fig. 2c
BCR-ABL1P190 crRNA40	GCGUCUCCAUGGAAGGCGCCCUCGCCAUCG	Fig. 2c
BCR-ABL1P190 crRNA41	CGUCUCCAUGGAAGGCGCCCUCGCCAUCGU	Fig. 2c
PspmCherry tiled1 (matching position 485)	GCUUCAGCCUCUGCUUGAUCUCGCCCUUCA	Fig. 2b
PspmCherry tiled2 (matching position 484)	CUUCAGCCUCUGCUUGAUCUCGCCCUUCAG	Fig. 2b
PspmCherry tiled3 (matching position 483)	UUCAGCCUCUGCUUGAUCUCGCCCUUCAGG	Fig. 2b
PspmCherry tiled4 (matching position 482)	UCAGCCUCUGCUUGAUCUCGCCCUUCAGGG	Fig. 2b
PspmCherry tiled5 (matching position 481)	CAGCCUCUGCUUGAUCUCGCCCUUCAGGGC	Fig. 2b
PspmCherry tiled6 (matching position 480)	AGCCUCUGCUUGAUCUCGCCCUUCAGGGCG	Fig. 2b
PspmCherry tiled7 (matching position 479)	GCCUCUGCUUGAUCUCGCCCUUCAGGGCGC	Fig. 2b
PspmCherry tiled8 (matching position 478)	CCUCUGCUUGAUCUCGCCCUUCAGGGCGCC	Fig. 2b
PspmCherry tiled9 (matching position 477)	CUCUGCUUGAUCUCGCCCUUCAGGGCGCCG	Fig. 2b
PspmCherry tiled10 (matching position 476)	UCUGCUUGAUCUCGCCCUUCAGGGCGCCGU	Fig. 2b
PspmCherry tiled11 (matching position 475)	CUGCUUGAUCUCGCCCUUCAGGGCGCCGUC	Fig. 2b
PspmCherry tiled12 (matching position 474)	UGCUUGAUCUCGCCCUUCAGGGCGCCGUCC	Fig. 2b
PspmCherry tiled13 (matching position 473)	GCUUGAUCUCGCCCUUCAGGGCGCCGUCCU	Fig. 2b
PspmCherry tiled14 (matching position 472)	CUUGAUCUCGCCCUUCAGGGCGCCGUCCUC	Fig. 2b
PspmCherry titling15 (matching position 471)	UUGAUCUCGCCCUUCAGGGCGCCGUCCUCG	Fig. 2b
PspmCherry tiled17 (matching position 469)	GAUCUCGCCCUUCAGGGCGCCGUCCUCGGG	Fig. 2b
PspmCherry tiled18 (matching position 468)	AUCUCGCCCUUCAGGGCGCCGUCCUCGGGG	Fig. 2b
PspmCherry tiled20 (matching position 466)	CUCGCCCUUCAGGGCGCCGUCCUCGGGGUA	Fig. 2b
PspmCherry tiled21 (matching position 465)	UCGCCCUUCAGGGCGCCGUCCUCGGGGUAC	Fig. 2b
PspmCherry tiled23 (matching position 463)	GCCCUUCAGGGCGCCGUCCUCGGGGUACAU	Fig. 2b
PspmCherry tiled24 (matching position 462)	CCCUUCAGGGCGCCGUCCUCGGGGUACAUC	Fig. 2b; Fig. 4c 4i; Extended Data Fig. 6f
PspmCherry tiled26 (matching position 460)	CUUCAGGGCGCCGUCCUCGGGGUACAUCCG	Fig. 2b; Fig. 4e 4i Extended Data Fig. 6c 6f
PspmCherry tiled27 (matching position 459)	UUCAGGGCGCCGUCCUCGGGGUACAUCCGC	Fig. 2b; Fig. 4f 4i; Extended Data Fig. 6d 6f
PspmCherry tiled29 (matching position 457)	CAGGGCGCCGUCCUCGGGGUACAUCCGCUC	Fig. 2b
PspmCherry tiled30 (matching position 456)	AGGGCGCCGUCCUCGGGGUACAUCCGCUCG	Fig. 2b
PspmCherry tiled32 (matching position 454)	GGCGCCGUCCUCGGGGUACAUCCGCUCGGA	Fig. 2b
PspmCherry tiled33 (matching position 453)	GCGCCGUCCUCGGGGUACAUCCGCUCGGAG	Fig. 2b
PspmCherry tiled35 (matching position 451)	GCCGUCCUCGGGGUACAUCCGCUCGGAGGA	Fig. 2b

Continued on next page

Name	Spacer (5'–3')	Source Figure (Hu et al. (2024)
PspmCherry tiled36 (matching position 450)	CCGUCCUCGGGGUACAUCCGCUCGGAGGAG	Fig. 2b
PspmCherry tiled38 (matching position 448)	GUCCUCGGGGUACAUCCGCUCGGAGGAGGC	Fig. 2b
PspmCherry tiled39 (matching position 447)	UCCUCGGGGUACAUCCGCUCGGAGGAGGCC	Fig. 2b; Fig. 4a 4i; Extended Data Fig. 6b 6f
PspmCherry tiled41 (matching position 445)	CUCGGGGUACAUCCGCUCGGAGGAGGCCUC	Fig. 2b
PspmCherry tiled42 (matching position 444)	UCGGGGUACAUCCGCUCGGAGGAGGCCUCC	Fig. 2b
PspmCherry tiled44 (matching position 442)	GGGGUACAUCCGCUCGGAGGAGGCCUCCCA	Fig. 2b
PspmCherry tiled 45 (matching position 441)	GGGUACAUCCGCUCGGAGGAGGCCUCCCAG	Fig. 2b
PspmCherry tiled47 (matching position 439)	GUACAUCCGCUCGGAGGAGGCCUCCCAGCC	Fig. 2b
PspmCherry tiled48 (matching position 438)	UACAUCCGCUCGGAGGAGGCCUCCCAGCCC	Fig. 2b
PspmCherry tiled49 (matching position 437)	ACAUCCGCUCGGAGGAGGCCUCCCAGCCCA	Fig. 2b
PspmCherry tiled50 (matching position 436)	CAUCCGCUCGGAGGAGGCCUCCCAGCCCAU	Fig. 2b
PspmCherry tiled51 (matching position 435)	ATCCGCTCGGAGGAGGCCTCCCAGCCCATG	Fig. 2b
PspmCherry tiled52 (matching position 434)	TCCGCTCGGAGGAGGCCTCCCAGCCCATGG	Fig. 2b
PspmCherry tiled53 (matching position 432)	CCGCTCGGAGGAGGCCTCCCAGCCCATGGT	Fig. 2b
PspmCherry tiled54 (matching position 431)	CGCTCGGAGGAGGCCTCCCAGCCCATGGTC	Fig. 2b
PspmCherry tiled55 (matching position 430)	GCTCGGAGGAGGCCTCCCAGCCCATGGTCT	Fig. 2b
PspmCherry tiled56 (matching position 429)	CTCGGAGGAGGCCTCCCAGCCCATGGTCTT	Fig. 2b
PspmCherry tiled57 (matching position 428)	TCGGAGGAGGCCTCCCAGCCCATGGTCTTC	Fig. 2b
PspmCherry tiled58 (matching position 427)	CGGAGGAGGCCTCCCAGCCCATGGTCTTCT	Fig. 2b
PspmCherry tiled59 (matching position 426)	GGAGGAGGCCTCCCAGCCCATGGTCTTCTT	Fig. 2b
PspmCherry tiled60 (matching position 425)	GAGGAGGCCTCCCAGCCCATGGTCTTCTTC	Fig. 2b
PspmCherry tiled61 (matching position 424)	AGGAGGCCTCCCAGCCCATGGTCTTCTTCT	Fig. 2b
SFPQ-ABL1-12	GAAGCGGCUCUCGGAGGAGACGUAGAGCAU	Fig. 3a
SFPQ-ABL1-9	GCGGCUCUCGGAGGAGACGUAGAGCAUGUC	Fig. 3a
SFPQ-ABL1-6	GCUCUCGGAGGAGACGUAGAGCAUGUCACU	Fig. 3a
SFPQ-ABL1-3	CUCGGAGGAGACGUAGAGCAUGUCACUUCC	Fig. 3a
SFPQ-ABL1+0	GGAGGAGACGUAGAGCAUGUCACUUCCCAU	Fig. 3a
SFPQ-ABL1+3	GGAGACGUAGAGCAUGUCACUUCCCAUCAU	Fig. 3a
SFPQ-ABL1+6	GACGUAGAGCAUGUCACUUCCCAUCAUGGA	Fig. 3a
SFPQ-ABL1+9	GUAGAGCAUGUCACUUCCCAUCAUGGAACC	Fig. 3a
SFPQ-ABL1+12	GAGCAUGUCACUUCCCAUCAUGGAACCACU	Fig. 3a
SNX2-ABL1-12	GAAGCGGCUCUCGGAGGAGACGUAGAGCUC	Fig. 3a
SNX2-ABL1-9	GCGGCUCUCGGAGGAGACGUAGAGCUCUUC	Fig. 3a
SNX2-ABL1-6	GCUCUCGGAGGAGACGUAGAGCUCUUCCCU	Fig. 3a
SNX2-ABL1-3	CUCGGAGGAGACGUAGAGCUCUUCCCUGGA	Fig. 3a
SNX2-ABL1+0	GGAGGAGACGUAGAGCUCUUCCCUGGAUCU	Fig. 3a
SNX2-ABL1+3	GGAGACGUAGAGCUCUUCCCUGGAUCUAUC	Fig. 3a
SNX2-ABL1+6	GACGUAGAGCUCUUCCCUGGAUCUAUCAAA	Fig. 3a
SNX2-ABL1+9	GUAGAGCUCUUCCCUGGAUCUAUCAAAGAU	Fig. 3a
SNX2-ABL1+12	GAGCUCUUCCCUGGAUCUAUCAAAGAUCAC	Fig. 3a
Spike crRNA -TIL1	GUCGCGCACCAGGUUGAUUGGGGUGUGCUU	Fig. 3a
Spike crRNA-TIL2	UCGCGCACCAGGUUGAUUGGGGUGUGCUUG	Fig. 3a
Spike crRNA-TIL3	CGCGCACCAGGUUGAUUGGGGUGUGCUUGG	Fig. 3a
Spike crRNA-TIL4	GCGCACCAGGUUGAUUGGGGUGUGCUUGGA	Fig. 3a
Spike crRNA-TILF5	CGCACCAGGUUGAUUGGGGUGUGCUUGGAG	Fig. 3a
Spike crRNA-TIL6	GCACCAGGUUGAUUGGGGUGUGCUUGGAGU	Fig. 3a
Spike crRNA-TIL7	CACCAGGUUGAUUGGGGUGUGCUUGGAGUA	Fig. 3a
Spike crRNA-TIL8	ACCAGGUUGAUUGGGGUGUGCUUGGAGUAG	Fig. 3a
Spike crRNA-TILF9	CCAGGUUGAUUGGGGUGUGCUUGGAGUAGA	Fig. 3a
Spike crRNA-TIL10	CAGGUUGAUUGGGGUGUGCUUGGAGUAGAU	Fig. 3a
Spike crRNA-TIL11	AGGUUGAUUGGGGUGUGCUUGGAGUAGAUC	Fig. 3a
Spike crRNA-TIL12	GGUUGAUUGGGGUGUGCUUGGAGUAGAUCU	Fig. 3a
Spike crRNA-TIL13	GUUGAUUGGGGUGUGCUUGGAGUAGAUCUU	Fig. 3a
Spike crRNA-TIL14	UUGAUUGGGGUGUGCUUGGAGUAGAUCUUA	Fig. 3a
Spike crRNA-TIL15	UGAUUGGGGUGUGCUUGGAGUAGAUCUUAA	Fig. 3a
Spike crRNA-TIL16	GAUUGGGGUGUGCUUGGAGUAGAUCUUAAA	Fig. 3a
Spike crRNA-TIL17	AUUGGGGUGUGCUUGGAGUAGAUCUUAAAG	Fig. 3a
Spike crRNA-TIL18	UUGGGGUGUGCUUGGAGUAGAUCUUAAAGU	Fig. 3a
Spike crRNA-TIL19	UGGGGUGUGCUUGGAGUAGAUCUUAAAGUA	Fig. 3a
Spike crRNA-TIL20	GGGGUGUGCUUGGAGUAGAUCUUAAAGUAG	Fig. 3a
Spike crRNA-TIL21	GGGUGUGCUUGGAGUAGAUCUUAAAGUAGC	Fig. 3a
Spike crRNA-TIL22	GGUGUGCUUGGAGUAGAUCUUAAAGUAGCC	Fig. 3a
Spike crRNA-TIL23	GUGUGCUUGGAGUAGAUCUUAAAGUAGCCA	Fig. 3a

Continued on next page

Name	Spacer (5'–3')	Source Figure (Hu et al. (2024)
Spike crRNA-TIL24	UGUGCUUGGAGUAGAUCUUAAAGUAGCCAU	Fig. 3a
Spike crRNA-TIL25	GUGCUUGGAGUAGAUCUUAAAGUAGCCAUC	Fig. 3a
Spike crRNA-TIL26	UGCUUGGAGUAGAUCUUAAAGUAGCCAUCG	Fig. 3a
Spike crRNA-TIL27	GCUUGGAGUAGAUCUUAAAGUAGCCAUCGA	Fig. 3a
Spike crRNA-TIL28	CUUGGAGUAGAUCUUAAAGUAGCCAUCGAU	Fig. 3a
Spike crRNA-TIL29	UUGGAGUAGAUCUUAAAGUAGCCAUCGAUG	Fig. 3a
Spike crRNA-TIL30	UGGAGUAGAUCUUAAAGUAGCCAUCGAUGU	Fig. 3a
Spike crRNA-TIL31	GGAGUAGAUCUUAAAGUAGCCAUCGAUGUU	Fig. 3a
Spike crRNA-TIL32	GAGUAGAUCUUAAAGUAGCCAUCGAUGUUC	Fig. 3a
Spike crRNA-TIL33	AGUAGAUCUUAAAGUAGCCAUCGAUGUUCU	Fig. 3a
Spike crRNA-TIL34	GUAGAUCUUAAAGUAGCCAUCGAUGUUCUU	Fig. 3a
Spike crRNA-TIL35	UAGAUCUUAAAGUAGCCAUCGAUGUUCUUA	Fig. 3a
Spike crRNA-TIL36	AGAUCUUAAAGUAGCCAUCGAUGUUCUUAA	Fig. 3a
Spike crRNA-TIL37	GAUCUUAAAGUAGCCAUCGAUGUUCUUAAA	Fig. 3a
Spike crRNA-TIL38	AUCUUAAAGUAGCCAUCGAUGUUCUUAAAC	Fig. 3a
Spike crRNA-TIL39	UCUUAAAGUAGCCAUCGAUGUUCUUAAACA	Fig. 3a
Spike crRNA-TIL40	CUUAAAGUAGCCAUCGAUGUUCUUAAACAC	Fig. 3a
Spike crRNA-TIL41	UUAAAGUAGCCAUCGAUGUUCUUAAACACG	Fig. 3a
Spike crRNA-TIL42	UAAAGUAGCCAUCGAUGUUCUUAAACACGA	Fig. 3a
Spike crRNA-TIL43	AAAGUAGCCAUCGAUGUUCUUAAACACGAA	Fig. 3a
Spike crRNA-TIL44	AAGUAGCCAUCGAUGUUCUUAAACACGAAC	Fig. 3a
Spike crRNA-TIL45	AGUAGCCAUCGAUGUUCUUAAACACGAACU	Fig. 3a
Spike crRNA-TIL46	GUAGCCAUCGAUGUUCUUAAACACGAACUC	Fig. 3a
Spike crRNA-TIL47	UAGCCAUCGAUGUUCUUAAACACGAACUCC	Fig. 3a
Spike crRNA-TIL48	AGCCAUCGAUGUUCUUAAACACGAACUCCC	Fig. 3a
Spike crRNA-TIL49	GCCAUCGAUGUUCUUAAACACGAACUCCCG	Fig. 3a
Spike crRNA-TIL50	CCAUCGAUGUUCUUAAACACGAACUCCCGC	Fig. 3a
Spike crRNA-TIL51	CAUCGAUGUUCUUAAACACGAACUCCCGCA	Fig. 3a
Spike crRNA-TIL52	AUCGAUGUUCUUAAACACGAACUCCCGCAG	Fig. 3a
Spike crRNA-TIL53	UCGAUGUUCUUAAACACGAACUCCCGCAGG	Fig. 3a
Spike crRNA-TIL54	CGAUGUUCUUAAACACGAACUCCCGCAGGU	Fig. 3a
Spike crRNA-TIL55	GAUGUUCUUAAACACGAACUCCCGCAGGUU	Fig. 3a
Spike crRNA-TIL56	AUGUUCUUAAACACGAACUCCCGCAGGUUC	Fig. 3a
Spike crRNA-TIL57	UGUUCUUAAACACGAACUCCCGCAGGUUCU	Fig. 3a
Spike crRNA-TIL58	GUUCUUAAACACGAACUCCCGCAGGUUCUU	Fig. 3a
Spike crRNA-TIL59	UUCUUAAACACGAACUCCCGCAGGUUCUUG	Fig. 3a
Spike crRNA-TIL60	UCUUAAACACGAACUCCCGCAGGUUCUUGA	Fig. 3a
Spike crRNA-TIL61	CUUAAACACGAACUCCCGCAGGUUCUUGAA	Fig. 3a
Ineffective EGFP crRNA1	CCGGUGAACAGCUCCUCGCCCUUGCUCACC	Fig. 3f
Ineffective EGFP crRNA2	CAGGAUGGGCACCACCCCGGUGAACAGCUC	Fig. 3f
Ineffective EGFP crRNA3	CUUGUAGUUGCCGUCGUCCUUGAAGAAGAU	Fig. 3f
Ineffective EGFP crRNA4	CACCAGGGUGUCGCCCUCGAACUUCACCUC	Fig. 3f
Ineffective EGFP crRNA5	CAGGAUGUUGCCGUCCUCCUUGAAGUCGAU	Fig. 3f
Ineffective EGFP crRNA6	CUUCUGCUUGUCGGCCAUGAUAUAGACGUU	Fig. 3f
Ineffective EGFP crRNA7	CUGCACGCUGCCGUCCUCGAUGUUGUGGCG	Fig. 3f
Ineffective EGFP crRNA8	CUUGUACAGCUCGUCCAUGCCGAGAGUGAU	Fig. 3f
Ineffective EGFP crRNA9	UACUCCAGCUUGUGCCCCAGGAUGUUGCCG	Fig. 3f
Potent EGFP crRNA1	GGUCAGCUUGCCGUAGGUGGCAUCGCCCUC	Fig. 3e
Potent EGFP crRNA2	GGUCACGAGGGUGGGCCAGGGCACGGGCAG	Fig. 3e
Potent EGFP crRNA3	GGUAGCGGCUGAAGCACUGCACGCCGUAGG	Fig. 3e
Potent EGFP crRNA4	GGGCAUGGCGGACUUGAAGAAGUCGUGCUG	Fig. 3e
Potent EGFP crRNA5	GGUGCGCUCCUGGACGUAGCCUUCGGGCAU	Fig. 3e
Potent EGFP crRNA6	GGCGCGGGUCUUGUAGUUGCCGUCGUCCUU	Fig. 3e
Potent EGFP crRNA7	GGCGAGCUGCACGCUGCCGUCCUCGAUGUU	Fig. 3e
Potent EGFP crRNA8	GGUAGUGGUCGGCGAGCUGCACGCUGCCGU	Fig. 3e
Potent EGFP crRNA9	GGGCCGUCGCCGAUGGGGGUGUUCUGCUGG	Fig. 3e
Potent EGFP crRNA10	GGGCAGCAGCACGGGGCCGUCGCCGAUGGG	Fig. 3e
Potent EGFP crRNA11	GGGGGUGUUCUGCUGGUAGUGGUCGGCGAG	Fig. 3i
Ineffective Tagbfp crRNA1	CUCGGUGAAGGCCUCCCAGCCGAGUGUUUU	Fig. 3i
Ineffective Tagbfp crRNA2	CGGUCAGCACGCCCCCGUCCUCGUAUGUGG	Fig. 3i
Ineffective Tagbfp crRNA3	CAGAUGGCUCCCGCCCACGAGCUUCAGGGC	Fig. 3i
Ineffective Tagbfp crRNA4	CCAUGUCGUUUCUGCCUUCCAGGCCGCCGU	Fig. 3i
Ineffective Tagbfp crRNA5	CGAUCAGAUGGCUCCCGCCCACGAGCUUCA	Fig. 3i

Continued on next page

Name	Spacer (5'–3')	Source Figure (Hu et al. (2024)
Ineffective Tagbfp crRNA6	CUCGUUGUUGGCCUCCUUGAUUCUUUCCAG	Fig. 3i
Ineffective Tagbfp crRNA7	CAAUUAAGCUUGUGCCCCAGUUUGCUAGGG	Fig. 3i
Ineffective Tagbfp crRNA8	CCCAUGUGAAGCCCUCAGGGAAGGACUGCU	Fig. 3i
Ineffective Tagbfp crRNA9	AUGUGAAGUUCACCCCUCUGAUCUUGACGU	Fig. 3i
Potent Tagbfp crRNA1	GGUGCCCUCGUAGGGCUUGCCUUCGCCCUC	Fig. 3h
Potent Tagbfp crRNA2	GGCGAAGGGGAGAGGGCCGCCCUCGACCAC	Fig. 3h
Potent Tagbfp crRNA3	GGUGACUCUCUCCCAUGUGAAGCCCUCAGG	Fig. 3h
Potent Tagbfp crRNA4	GGGGUACAGGGUCUCGGUGAAGGCCUCCCA	Fig. 3h
Potent Tagbfp crRNA5	GGUUUCUUGGAUCUAUAUGUGGUCUUGAUG	Fig. 3h
Potent Tagbfp crRNA6	GGUCUGGGUGCCCUCGUAGGGCUUGCCUUC	Fig. 3h
Potent Tagbfp crRNA7	GGAUGUCGAAGGCGAAGGGGAGAGGGCCGC	Fig. 3h
Potent Tagbfp crRNA8	GGUCUUGCUGCCGUAGAGGAAGCUAGUAGC	Fig. 3h
Potent Tagbfp crRNA9	GGGAUGCCCUGGGUGUGGUUGAUGAAGGUC	Fig. 3h
Potent Tagbfp crRNA10	GGGAGGUCGCAGUAUCUGGCCACUGCCACC	Fig. 3h

TABLE V
SUMMARY OF 157 ENGINEERED FEATURES USED IN ML MODEL

Category	Feature Description	Count
Positional One-Hot	Binary indicator for A, C, G, T at each of 30 positions	120
Nucleotide Counts	Total count of A, C, G, T in spacer	4
GC Content	Percentage of G+C nucleotides	1
Dinucleotides	Frequency of all 16 pairs (AA, AC, ..., TT)	16
Trinucleotides	Frequency of key trimers (e.g., GGG, CCC)	10
Motif: 5' End	'GG' at positions 1-2 (Binary)	1
Motif: Penalties	Count of 'C' at positions 1-4	1
	Count of 'C' at positions 11-17 (central region)	1
Thermodynamics	Minimum Free Energy (MFE)	1
Homopolymers	Presence of >3 consecutive identical bases	1
Conservation	Pre-computed conservation score	1
Total		**157**

TABLE VI
TOP 5 CANDIDATE CRRNAS FOR DENV-1, DENV-2, DENV-3, AND DENV-4

Serotype	Rank	Optimized Spacer Sequence (5' $\rightarrow$ 3')	Pred. Efficiency	Coverage (%)
DENV-1	1	GGAGTAGTCGTACTAGGATCACAAGAAGGA	0.881	45.94
	2	GGAGTAGTCGTACTAGGATCACAAGAAGGT	0.875	44.12
	3	GGTGTAGTCGTACTAGGATCACAAGAAGGA	0.872	43.80
	4	GGAAGAACAAGACGGGAACTTTGTGTGTCG	0.804	68.42
	5	GGAGCAACAAGACGGGAACTTTGTGTGTCG	0.798	65.10
DENV-2	1	GGAGACAGAGGATGGGGAAATGGATGTGGA	0.878	42.36
	2	GGAGACAGAGGATGGGGAAATGGATGTGGG	0.871	40.15
	3	GGAGACAGAGGATGGGGAAATGGATGTGTA	0.865	38.90
	4	GGTGTCAACATAGAAGCAGAACCTCCATTC	0.832	55.20
	5	GGGAGTCAACATAGAAGCAGAACCTCCATT	0.613	68.75
DENV-3	1	GGGGTTGGATCGCAAGAGGGAGCAATGCAC	0.882	41.41
	2	GGGGTTGGATCGCAAGAGGGAGCAATGCAT	0.879	40.85
	3	GGTGCAAGGCGGATGGGCATCTTGGGAGAC	0.859	69.88
	4	GGTGCAAGGCGGATGGGCATCTTGGGAGAT	0.855	68.10
	5	GGTGCAAGGCGGATGGGCATCTTGGGAGAA	0.848	65.50
DENV-4	1	GGGGTAGAACATGGAGGATGCGTCACAACC	0.888	47.77
	2	GGGGTAGAACATGGAGGATGCGTCACAACA	0.885	46.90
	3	GGAACAAGATGTCCAAGGCAAGGAGAGCCT	0.836	80.51
	4	GGAACAAGATGTCCAAGGCAAGGAGAGCCT	0.831	78.20
	5	GGAACAAGATGTCCAAGGCAAGGAGAGCCA	0.828	75.15