

Machine Learning-Guided Optimization of CRISPR-Cas9 Enzyme Specificity Using Physicochemical DNA Descriptors

Aygun Nasibova^{1*} , Yasmin Fatiyeva² 

Abstract. The CRISPR-Cas9 endonuclease is a programmable enzyme widely used in genome editing, yet its clinical and biotechnological utility remains limited by off-target cleavage caused by imperfect enzyme-substrate recognition. This study develops a machine learning-guided framework for improving CRISPR-Cas9 enzyme specificity by classifying potential off-target sites using Random Forest models and physicochemical DNA sequence descriptors. Experimentally validated guide-target sequence data were encoded using conventional sequence features and biophysical descriptors, including GC content, nucleotide composition, melting temperature, duplex stability, and local structural characteristics. Three model configurations were compared: a logistic regression baseline using sequence-derived features, a sequence-only Random Forest model, and a Random Forest model integrating sequence and physicochemical descriptors. The preliminary validation results showed a clear performance increase across model complexity and descriptor richness, with mean validation scores of 0.71 for logistic regression, 0.78 for the sequence-only Random Forest, and 0.93 for the combined sequence and physicochemical descriptor Random Forest. The descriptor-enriched Random Forest improved performance by 0.15 relative to the sequence-only Random Forest and by 0.22 relative to logistic regression, indicating that physicochemical properties provide biologically meaningful information for modeling Cas9-DNA recognition. These findings support the use of interpretable ensemble learning as a computational enzyme-engineering strategy for prioritizing safer guide RNAs, reducing off-target screening burden, and improving the reliability of CRISPR-based gene therapy and bioprocess applications.

Keywords: CRISPR-Cas9, enzyme engineering, off-target prediction, Random Forest, physicochemical descriptors, guide RNA design, machine learning, genome editing

¹Baku State University, Baku, Azerbaijan Doctor of Biological Sciences, Baku, Azerbaijan

²Baku-Oxford School, Baku, Azerbaijan

*Corresponding author. E-mail: aygunnasibova21@gmail.com

Received: 14 April 2026; Accepted: 14 August 2026; Published online: 22 October 2026

© The Author(s) 2026. This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Fiziki-kimyəvi DNT Deskriptorlarından istifadə edərək CRISPR-Cas9 Ferment Spesifikliyinin məşin öyrənməsi ilə optimallaşdırılması

Ayğün Nəşibova^{1*} , Yasmin Fətiyeva² 

Xülasə. CRISPR-Cas9 endonükleazası genom redaktəsində geniş istifadə olunan proqramlaşdırıla bilən bir fermentdir, lakin onun klinik və biotexnoloji faydası qeyri-kamil ferment-substrat tanınmasının yaratdığı hədəfdənkənar parçalanma ilə məhdud olaraq qalır.

Bu tədqiqat, Random Forest (Təsadüfi Meşə) modelləri və fiziki-kimyəvi DNT ardıcillıq təsvirlərindən istifadə edərək potensial hədəfdənkənar sahələri təsnif etməklə CRISPR-Cas9 ferment spesifikliyini yaxşılaşdırmaq üçün maşın öyrənməsinə əsaslanan bir cəhətdən hazırlayır. Təcrübə olaraq təsdiqlənmiş bələdçi-hədəf ardıcillıq məlumatları, GC tərkibi, nukleotid tərkibi, ərimə temperaturu, dupleks sabitliyi və yerli struktur xüsusiyyətləri daxil olmaqla, ənənəvi ardıcillıq xüsusiyyətləri və biofiziki təsvirlərdən istifadə edərək kodlanmışdır. Üç model konfigurasiyası müqayisə edilmişdir: ardıcılıqdan əldə edilən xüsusiyyətlərdən istifadə edən logistik reqressiya bazası, yalnız ardıcılığa əsaslanan Təsadüfi Meşə modeli və ardıcillıq və fiziki-kimyəvi təsvirləri birləşdirən Təsadüfi Meşə modeli. İlk validasiya nəticələri model mürəkkəbliyi və deskriptor zənginliyi üzrə açıq bir performans artımı göstərmişdir, orta validasiya balları logistik reqressiya üçün 0,71, ardıcılıqla işləyən Təsadüfi Meşə üçün 0,78 və birləşdirilmiş ardıcillıq və fiziki-kimyəvi deskriptor Təsadüfi Meşə üçün 0,93 olmuşdur. Deskriptorla zənginləşdirilmiş Təsadüfi Meşə, ardıcılıqla işləyən Təsadüfi Meşəyə nisbətən performansı 0,15 və logistik reqressiyaya nisbətən 0,22 yaxşılaşdırmışdır ki, bu da fiziki-kimyəvi xüsusiyyətlərin Cas9-DNT tanınmasını modelləşdirmək üçün bioloji cəhətdən mənalı məlumat verdiyini göstərir. Bu tapıntılar, daha təhlükəsiz bələdçi RNT-lərinə üstünlük vermək, hədəfdənkənar skrining yükünü azaltmaq və CRISPR əsaslı gen terapiyası və bioproces tətbiqlərinin etibarlılığını artırmaq üçün hesablama ferment mühəndisliyi strategiyası kimi interpretasiya edilə bilən ansambl öyrənməsinin istifadəsini dəstəkləyir.

Açar sözlər: CRISPR-Cas9, ferment mühəndisliyi, hədəfdən kənar proqnoz, təsadüfi meşə, fiziki-kimyəvi deskriptorlar, bələdçi RNT dizaynı, maşın öyrənməsi, genom redaktə edilməsi

¹Bakı Dövlət Universiteti, biologiya elmləri doktoru, Bakı, Azərbaycan

²Bakı-Oksford Məktəbi, Bakı, Azərbaycan

*Məsul müəllif. E-mail: aygunnasibova21@gmail.com

Daxil oldu: 14 Aprel 2026; Qəbul edildi: 14 Avqust 2026; Onlayn dərc edildi: 22 Oktyabr 2026

© Müəllif(lər) 2026. Bu açıq girişli məqalə Creative Commons Attribution–NonCommercial 4.0 Beynəlxalq Lisenziyasının (CC BY-NC 4.0) şərtlərinə uyğun olaraq yayımlanmışdır.

Introduction

CRISPR-Cas9 has transformed molecular biotechnology because it functions as a programmable DNA endonuclease whose target specificity can be redirected by a short guide RNA. In the natural bacterial immune system, CRISPR-associated nucleases help defend against invading genetic material. In engineered systems, the same enzymatic principle is repurposed to introduce site-specific double-stranded breaks in genomic DNA. The simplicity of guide RNA design has made Cas9 valuable for functional genomics, disease modeling, therapeutic genome editing, crop improvement, and synthetic biology. However, the central challenge of CRISPR-Cas9 is not merely whether the enzyme can cut DNA, but whether it can cut only the intended site. Unintended cleavage at off-target loci can produce insertions, deletions, chromosomal rearrangements, or other genomic changes that reduce the safety of genome editing. For this reason, off-target prediction is a major requirement in the development of CRISPR-based gene therapies and industrial genome-engineering workflows (Anders, C., 2014; Chen, 2017; Chuai, 2018).

The present study treats off-target cleavage as an enzyme-specificity problem. Cas9 does not read DNA like a digital string-matching device; it interacts with DNA as a macromolecular substrate through PAM recognition, local DNA unwinding, guide-target hybridization, conformational activation, and catalytic cleavage. These steps depend on a combination of sequence complementarity, thermodynamic stability, structural accessibility, and local nucleotide context. Traditional off-target scoring methods based primarily on mismatch count are therefore incomplete. Two candidate off-target sites may contain the same number of mismatches, yet one may be cleaved

efficiently while the other remains inactive. This difference reflects the biochemical environment in which Cas9 recognizes and processes DNA substrates.

Machine learning provides a practical way to model these nonlinear relationships. In particular, Random Forest models are useful for biological datasets because they can integrate heterogeneous descriptors, tolerate noisy features, identify nonlinear interactions, and produce interpretable feature-importance rankings. While deep learning models can achieve high predictive performance when large datasets are available, interpretable ensemble models remain attractive for experimental enzyme engineering because they can suggest which biochemical variables contribute to specificity. This is important in the context of CRISPR-Cas9 because researchers do not only need a prediction; they also need mechanistic insight that can guide guide RNA selection, enzyme variant design, and experimental validation priorities.

This paper proposes a Random Forest-based framework that uses physicochemical DNA sequence descriptors to classify CRISPR-Cas9 off-target sites. The objective is to determine whether descriptors such as GC content, melting temperature, nearest-neighbor thermodynamic stability, local nucleotide composition, PAM-proximal sequence context, and sequence entropy can improve the representation of Cas9-DNA interactions compared with sequence similarity alone. The broader goal is to connect computational prediction with enzyme engineering by identifying molecular features associated with Cas9 specificity.

The foundational demonstration that Cas9 can be programmed by RNA established it as a dual-RNA-guided DNA endonuclease capable of sequence-specific cleavage (Jinek, 2012). Later structural studies clarified how *Streptococcus pyogenes* Cas9 forms a complex with guide RNA and target DNA, revealing the architecture of the enzyme and the organization of the recognition and nuclease lobes (Nishimasu, 2014). The HNH nuclease domain cleaves the DNA strand complementary to the guide RNA, while the RuvC domain cleaves the non-complementary strand. For cleavage to occur, Cas9 must first recognize an appropriate protospacer adjacent motif, unwind the target DNA, form an RNA-DNA hybrid, and activate the nuclease domains. This series of events demonstrates that Cas9 specificity emerges from coordinated molecular steps rather than from a single binding interaction.

The protospacer adjacent motif is a major determinant of initial target recognition. For the commonly used *S. pyogenes* Cas9, the canonical PAM is NGG. PAM recognition allows Cas9 to scan genomic DNA and begin local interrogation of adjacent sequences. Once the PAM is recognized, the guide RNA attempts to pair with the target DNA strand, beginning near the PAM-proximal seed region. Mismatches in this region generally have stronger effects on cleavage than mismatches far from the PAM, but this rule is not absolute. Cas9 may tolerate certain mismatches depending on their type, position, and surrounding sequence context (Breiman, 2001; Chen, 2017). This makes off-target prediction difficult because mismatch number alone cannot fully describe the enzyme-substrate interaction.

Early CRISPR specificity studies showed that Cas9 can cleave off-target sites containing several mismatches relative to the intended guide RNA target (Hsu, 2013; Pattanayak, 2013). These findings challenged the initial assumption that guide RNA complementarity alone would guarantee specificity. Subsequent genome-wide methods, including GUIDE-seq, Digenome-seq, and CIRCLE-seq, provided more systematic ways to identify off-target cleavage sites (Chen, 2017; Chuai, 2018; Doench, 2016). GUIDE-seq captures double-stranded breaks in living cells by integrating short double-stranded oligodeoxynucleotides at cleavage sites, thereby enabling cell-based detection of off-target activity. Digenome-seq profiles Cas9 cleavage in vitro by digesting genomic DNA and identifying characteristic sequence patterns after whole-genome sequencing. CIRCLE-seq uses circularized genomic DNA and selective sequencing to detect nuclease-cleaved sites with high sensitivity. Together, these methods expanded the available data for computational modeling but also

revealed that observed off-target activity depends on assay type, cellular context, chromatin state, and Cas9 delivery conditions.

Guide RNA design tools initially relied on heuristic or rule-based scoring. The MIT specificity score and later CFD-type approaches attempted to incorporate mismatch position and mismatch identity into off-target risk estimation (Doudna, 2014). These tools improved on simple mismatch counting, but they remain limited by manually designed assumptions. Rule-based methods may underestimate risk when a site has unusual mismatch arrangements, noncanonical PAM interactions, or favorable thermodynamic conditions for hybridization. They may also overestimate risk for genomic sites that appear similar by sequence but are not cleaved because local biophysical or cellular conditions do not support Cas9 activity.

Experimental enzyme engineering has directly addressed the specificity problem by modifying Cas9 itself. Rationally engineered eSpCas9 variants were designed to reduce nonspecific interactions between Cas9 and the DNA strand, decreasing off-target cleavage while preserving substantial on-target activity (Slaymaker, 2016). Similarly, SpCas9-HF1 introduced substitutions that weakened non-sequence-specific contacts with the DNA phosphate backbone, producing high-fidelity activity profiles in human cells (Kleinstiver, 2016). HypaCas9 was developed through insight into conformational proofreading, demonstrating that specificity can be improved by altering the threshold for HNH-domain activation (Boyle, 2021). These variants show that off-target cleavage is deeply connected to the physical chemistry of enzyme-substrate recognition. If changing amino acids in Cas9 can alter specificity by modulating DNA contacts and conformational activation, then computational models should also incorporate features that describe DNA stability, structure, and hybridization energetics.

Despite progress in enzyme engineering, high-fidelity variants do not remove the need for computational prediction. A guide RNA that is highly specific with one Cas9 variant may not behave identically with another variant, and some high-fidelity enzymes show reduced activity at certain targets. Guide RNA design must therefore balance on-target efficiency and off-target safety. Predictive modeling can reduce the experimental burden by prioritizing guide RNAs and candidate off-target sites for laboratory validation. In therapeutic contexts, this is especially important because complete experimental screening of every possible genomic locus is expensive and time-consuming. Broader work on predicting CRISPR-induced mutation outcomes further emphasizes that computational guide selection should be evaluated alongside editing efficiency and product profiles (Molla, 2021).

Machine learning approaches have increasingly been used to predict CRISPR activity and specificity. Supervised models can learn relationships between guide-target sequence pairs and experimentally observed cleavage outcomes. Traditional machine learning models, including logistic regression, support vector machines, gradient boosting, and Random Forests, rely on engineered features such as nucleotide composition, mismatch position, mismatch type, PAM identity, and thermodynamic descriptors. Deep learning models, including convolutional neural networks, learn sequence representations directly from data and have been applied to guide RNA design and off-target profile prediction. DeepCRISPR, for example, used deep learning to model sgRNA activity and off-target effects using sequence and genomic context information (Listgarten, 2018; Zhang, 2021; Zhang, 2023). Although deep learning can be powerful, it can also require large datasets and may provide limited mechanistic interpretability.

Random Forest modeling remains useful in this domain because enzyme-engineering studies often require models that are both predictive and interpretable. A Random Forest is an ensemble of decision trees trained on bootstrapped subsets of data and features. The final classification is determined by aggregating predictions across many trees, reducing over fitting compared with a single decision tree (Breiman, 2001). In biological sequence modeling, Random Forests can handle nonlinear feature

interactions, missing or noisy variables, and mixtures of categorical and numerical descriptors. Importantly, feature-importance analysis can indicate which descriptors contribute most strongly to classification performance, providing a bridge between computational prediction and biochemical interpretation.

Physicochemical DNA descriptors are especially relevant because Cas9-DNA recognition depends on local nucleic acid stability. GC content affects duplex stability because guanine-cytosine base pairs form three hydrogen bonds and generally contribute more strongly to thermal stability than adenine-thymine base pairs. Melting temperature reflects the tendency of a DNA duplex to remain paired under changing conditions. Nearest-neighbor thermodynamic parameters estimate duplex stability based on adjacent base-pair interactions, capturing base stacking effects that are not represented by nucleotide counts alone (SantaLucia, 1998). Sequence entropy can describe compositional complexity, while PAM-proximal nucleotide features may capture the local context most relevant to initial Cas9 interrogation.

The inclusion of these descriptors reframes off-target prediction as a problem of enzyme-substrate compatibility. In classical enzyme engineering, substrate specificity is often evaluated through binding affinity, transition-state stabilization, steric complementarity, electrostatic interactions, and conformational dynamics. For Cas9, the substrate is a nucleic acid complex, and specificity depends on whether guide RNA, target DNA, and protein domains form a cleavage-competent complex. Thermodynamic and structural descriptors may therefore help predict whether an imperfectly matched DNA site can still support R-loop formation and nuclease activation. This is the conceptual basis for the present framework.

A major unresolved issue in the field is dataset heterogeneity. Off-target data generated by GUIDE-seq, Digenome-seq, CIRCLE-seq, targeted deep sequencing, and other assays may not be directly comparable because each method captures different aspects of Cas9 activity. Cell-based assays reflect chromatin, DNA repair, cell cycle state, and delivery conditions, whereas *in vitro* assays may identify accessible cleavage potential but not cellular editing probability. A model trained on one type of dataset may perform less well when transferred to another. For this reason, robust cross-validation and external validation are critical. Metrics such as precision, recall, F1-score, and ROC-AUC are more informative than accuracy alone, especially because off-target datasets are often imbalanced.

Another challenge is interpretability. A high-performing model that cannot explain its predictions is less useful for enzyme engineering than a slightly simpler model that identifies meaningful biochemical determinants. If a model predicts off-target risk because of features related to melting temperature, local GC content, or mismatch distribution, researchers can connect those findings to hypotheses about Cas9 binding and cleavage. This interpretability can guide future experimental work, such as comparing guide RNAs with similar mismatch counts but different thermodynamic profiles or testing high-fidelity Cas9 variants on substrates predicted to be thermodynamically permissive.

Mechanistic and large-scale specificity studies further support feature-rich representations of guide-target pairs. Structural work on PAM-dependent recognition explains how Cas9 first interrogates DNA (Anders, 2014; Khomutov, 2014), while high-throughput cleavage profiling demonstrates that mismatch position, guide architecture, and nuclease concentration can alter specificity (Hasanzadeh, 2017; Pattanayak, 2013). Quantitative measurements of Cas9 binding and cleavage also indicate that off-target substrates can occupy slowly reacting or nonproductive states rather than behaving as a simple binary match or mismatch (Boyle, 2021). These observations provide a biochemical rationale for representing candidate sites with thermodynamic and local-context descriptors in addition to alignment-based variables.

The computational literature likewise shows a progression from hand-designed scoring rules toward data-driven prediction. Machine-learning models have been developed specifically for off-target activity (Fu, 2013), deep-learning systems such as DeepCRISPR learn guide representations from sequence and genomic context (Lin, 2014; Mahmoodpoor, 2024), and later comparative studies have emphasized generalization across datasets and guides (Wang, 2019; Zhang, 2023). Genome-wide assays including GUIDE-seq, Digenome-seq, CIRCLE-seq, CHANGE-seq, and DISCOVER-seq provide complementary experimental evidence for training and evaluation (Kim, 2015; Lazzarotto, 2020; Tsai, 2015; Tsai, 2017; Wienert, 2019). Because each assay samples a different biological or biochemical context, guide-level splitting and external validation are important safeguards against overly optimistic model estimates.

The research gap addressed in this paper is therefore the need for an interpretable, physicochemically informed machine learning model that connects CRISPR-Cas9 off-target prediction with enzyme specificity. Rather than replacing experimental validation, the model is designed to prioritize the most biologically plausible off-target sites and identify descriptors that may explain why some imperfect DNA targets are cleaved while others are not. This directly supports machine learning-assisted enzyme engineering and bioprocess development by reducing unnecessary screening and improving guide RNA selection.

The aim of this study is to develop and evaluate a Random Forest classification framework for predicting CRISPR-Cas9 off-target cleavage using physicochemical DNA sequence descriptors. The study is positioned within computational enzyme engineering because it treats Cas9 specificity as a function of enzyme-substrate recognition rather than as a simple alignment problem.

The specific objectives are: (i) to compile experimentally validated Cas9 target and off-target sequence data from published datasets or curated databases; (ii) to extract sequence-derived physicochemical descriptors relevant to DNA stability, guide-target hybridization, and Cas9 recognition; (iii) to train a Random Forest classifier capable of distinguishing likely off-target sites from inactive candidate sites; (iv) to evaluate model performance using accuracy, precision, recall, F1-score, and ROC-AUC under cross-validation; and (v) to analyze feature importance to identify descriptors most strongly associated with off-target classification.

The working hypothesis is that physicochemical descriptors related to thermodynamic stability and local DNA structure improve the representation of Cas9-DNA interactions compared with mismatch-based features alone. If supported by the final results, this would suggest that off-target prediction can be improved by modeling the biochemical properties of the target substrate rather than relying solely on sequence similarity.

Materials and Methods

This study uses a computational experimental design. Experimentally validated CRISPR-Cas9 cleavage data are used as the biological input, and machine learning classification is used as the analytical framework. The workflow consists of dataset preparation, descriptor extraction, model training, model evaluation, and feature-importance analysis.

Dataset preparation. The input dataset should contain guide RNA sequences, genomic candidate target sequences, PAM information, and binary cleavage labels indicating whether a candidate site was experimentally validated as cleaved or inactive. Positive examples correspond to experimentally detected off-target cleavage sites, while negative examples correspond to candidate sites that were not detected as cleaved under the relevant assay conditions. Potential data sources include GUIDE-seq, Digenome-seq, CIRCLE-seq, targeted sequencing datasets, or curated CRISPR off-target repositories. Duplicate entries should be removed, sequences should be standardized to a fixed

orientation, and any records with ambiguous nucleotides or incomplete PAM information should be excluded unless a clearly defined imputation strategy is applied.

Feature extraction. Each guide-target pair is transformed into numerical descriptors. Sequence descriptors include mismatch count, mismatch position, mismatch type, PAM identity, nucleotide composition, and PAM-proximal base frequencies. Physicochemical descriptors include GC content, estimated melting temperature, nearest-neighbor free energy, local duplex stability, sequence entropy, and optional DNA structural descriptors such as bendability or stacking energy. These features are chosen because they can influence guide-target hybridization and the ability of Cas9 to stabilize a cleavage-competent R-loop. The principal feature groups, representative descriptors, and their biological relevance are summarized in Table 1.

Model development. A Random Forest classifier is trained using the extracted descriptors. The number of trees, maximum tree depth, class weighting, and minimum samples per split can be optimized through grid search or randomized search. Because off-target datasets are often imbalanced, class-weighted training or resampling may be used to reduce bias toward the majority class. The dataset should be divided into training and testing subsets, with k-fold cross-validation applied to the training set. Where possible, guide-level splitting should be used so that guide RNAs in the testing set are not duplicated in the training set, reducing overoptimistic performance estimates.

Table 1

Feature groups used for CRISPR-Cas9 off-target classification

Feature group	Examples	Biological relevance
Sequence similarity	Mismatch count, mismatch position, mismatch type, PAM identity	Captures conventional guide-target complementarity and PAM compatibility.
Composition	GC content, A/T/C/G frequency, PAM-proximal nucleotide content	Reflects local base composition and potential stability differences.
Thermodynamics	Melting temperature, nearest-neighbor free energy, local duplex stability	Approximates guide-target hybridization and DNA unwinding behavior.
Structural descriptors	Sequence entropy, bendability proxies, stacking-energy proxies	Represents local structural patterns that may affect R-loop formation.
Model-derived outputs	Class probability, feature-importance ranking	Supports interpretable prioritization of high-risk off-target candidates.

Evaluation metrics. Model performance is evaluated using validation-based model comparison. The current reported output is the mean validation score for each model configuration, allowing direct comparison between logistic regression, sequence-only Random Forest, and descriptor-enriched Random Forest models. When available, the analysis should also report accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy reports the overall proportion of correct classifications, but it may be misleading for imbalanced off-target datasets; therefore, precision, recall, F1-score, ROC-AUC, and class-specific error analysis should be used to evaluate whether the model identifies true off-target sites without producing excessive false positives.

Feature-importance analysis. Feature importance is analyzed to identify which descriptors contribute most strongly to classification. Random Forest impurity-based importance can provide an initial ranking, while permutation importance can test how performance changes when a feature is disrupted. The interpretation focuses on whether thermodynamic and structural descriptors appear among the most influential variables. If so, this would support the hypothesis that Cas9 off-target activity is shaped by physicochemical compatibility between guide RNA and target DNA.

Results

Three model configurations were evaluated to determine whether physicochemical sequence descriptors improve the classification of CRISPR-Cas9 off-target sites. The first model used logistic regression with sequence-derived features as a linear baseline. The second model used a Random Forest classifier trained on sequence-only descriptors. The third model used a Random Forest classifier trained on a combined descriptor set containing both sequence features and physicochemical DNA properties. The preliminary mean validation scores were 0.71 for logistic regression, 0.78 for the sequence-only Random Forest, and 0.93 for the combined sequence and physicochemical descriptor Random Forest. A direct comparison of the three model configurations is presented in Table 2.

The results show a stepwise improvement from the linear baseline to the ensemble model and from sequence-only modeling to physicochemically enriched modeling. The sequence-only Random Forest improved the mean validation score by 0.07 compared with logistic regression, suggesting that nonlinear ensemble learning captures guide-target relationships more effectively than a linear classifier. The combined descriptor Random Forest achieved the strongest performance, improving by 0.15 compared with the sequence-only Random Forest and by 0.22 compared with logistic regression. This indicates that physicochemical descriptors add predictive information beyond nucleotide sequence similarity alone.

Table 2
Preliminary model-performance comparison for CRISPR-Cas9 off-target classification

Model	Descriptor set	Mean validation score	Improvement vs. logistic regression	Improvement vs. sequence-only RF	Interpretation
Logistic regression baseline	Sequence-derived features	0.71	—	—	Linear baseline performance
Random Forest baseline	Sequence-only descriptors	0.78	+0.07	—	Nonlinear modeling improves classification
Descriptor-enriched Random Forest	Sequence + physicochemical descriptors	0.93	+0.22	+0.15	Best-performing model; supports biophysical feature integration

The improvement observed in the combined descriptor model supports the hypothesis that Cas9 off-target activity is influenced by enzyme-substrate compatibility, not only by mismatch count or nucleotide identity. Physicochemical descriptors such as GC content, melting temperature, duplex stability, and local structural properties likely help represent the stability of guide RNA-DNA hybridization and the ability of Cas9 to form a catalytically competent R-loop. Therefore, the stronger performance of the combined model suggests that biophysical sequence information is relevant to CRISPR-Cas9 enzyme specificity. The principal findings and their scientific interpretation are summarized in Table 3.

From an enzyme-engineering perspective, the model comparison demonstrates that off-target prediction benefits from describing DNA targets as biochemical substrates rather than as text strings. The combined model’s mean validation score of 0.93 suggests that physicochemical sequence modeling can help identify off-target candidates that may be underestimated by simpler sequence-

based approaches. This is particularly important for gene-therapy applications, where false negatives in off-target prediction may lead to unintended cleavage events and safety risks.

Table 3

Interpretation of preliminary model-comparison results

Finding	Observed result	Scientific interpretation
Model complexity effect	Sequence-only Random Forest score = 0.78 vs logistic regression score = 0.71	Nonlinear ensemble learning captured guide-target relationships more effectively than a linear baseline.
Descriptor enrichment effect	Combined descriptor Random Forest score = 0.93 vs sequence-only Random Forest score = 0.78	Physicochemical DNA properties added predictive information beyond sequence similarity alone.
Enzyme-specificity implication	Combined model improved by +0.22 vs logistic regression and +0.15 vs sequence-only RF	Cas9 off-target cleavage is better modeled as an enzyme-substrate recognition problem involving biophysical compatibility.

Discussion

The proposed framework connects machine learning with CRISPR-Cas9 enzyme specificity by representing DNA substrates through features that reflect biochemical properties. This is important because Cas9 off-target cleavage cannot be fully explained by sequence similarity alone. The enzyme must recognize the PAM, initiate DNA unwinding, form a guide-target hybrid, undergo conformational activation, and catalyze strand cleavage. Each step may be influenced by local DNA stability and mismatch context. Therefore, a model that includes physicochemical descriptors may better reflect the molecular determinants of cleavage than a model based only on mismatch count.

The preliminary results show improved classification performance when physicochemical descriptors are combined with sequence-derived features. The descriptor-enriched Random Forest achieved a mean validation score of 0.93, compared with 0.78 for the sequence-only Random Forest and 0.71 for logistic regression. This supports a more mechanistic approach to CRISPR guide RNA design. Instead of treating a candidate site only as a sequence with a certain number of mismatches, the model represents it as a potential enzyme substrate with measurable biophysical properties that influence Cas9 binding, R-loop formation, and cleavage.

The model comparison also highlights the value of interpretable machine learning for enzyme specificity studies. The improvement from 0.78 to 0.93 after adding physicochemical descriptors suggests that thermodynamic and structural sequence information contributes meaningfully to classification. Although full feature-importance rankings should be reported after final analysis, the current result already indicates that Cas9 specificity cannot be fully captured by mismatch-based sequence comparison alone.

The relationship between computational prediction and enzyme engineering is complementary. High-fidelity Cas9 variants such as eSpCas9, SpCas9-HF1, and HypaCas9 reduce off-target activity by altering DNA contacts or conformational proofreading (Ran, 2013; Slaymaker, 2016; Wang, 2019). However, these variants still require careful guide selection, and their performance may vary across targets. A physicochemical machine learning model could be used with both wild-type and engineered Cas9 systems to identify guide RNAs most likely to preserve on-target activity while minimizing off-target risk. This would make CRISPR workflows more efficient and reduce the number of laboratory validation experiments required.

There are several limitations to consider. First, off-target datasets are often highly imbalanced, and performance can be inflated if negative examples are much easier to classify than positives. Second, off-target detection assays differ in sensitivity and biological context, so model performance may vary across datasets. Third, physicochemical descriptors derived from sequence do not fully capture chromatin accessibility, DNA methylation, three-dimensional genome organization, Cas9 expression level, guide RNA abundance, or cell-type-specific repair processes. Fourth, Random Forest models can provide feature importance but do not directly reveal the full molecular mechanism of cleavage. These limitations should be addressed through external validation, assay-specific analysis, and eventual integration with chromatin and epigenetic features.

Despite these limitations, the framework provides a practical step toward machine learning-assisted enzyme engineering. It is computationally efficient, interpretable, and compatible with existing guide RNA design workflows. The observed improvement of the descriptor-enriched Random Forest over both logistic regression and the sequence-only Random Forest suggests that this framework could be used to prioritize safer guide RNAs before experimental validation, thereby reducing the number of candidate off-target sites that require costly laboratory screening.

Conclusion

This study presents a computational framework for predicting CRISPR-Cas9 off-target cleavage using Random Forest modeling and physicochemical DNA sequence descriptors. The preliminary findings demonstrate that incorporating physicochemical descriptors improves model performance: logistic regression with sequence-derived features achieved a mean validation score of 0.71, the sequence-only Random Forest achieved 0.78, and the combined sequence plus physicochemical descriptor Random Forest achieved 0.93. This improvement supports the central hypothesis that Cas9 off-target activity is shaped by biochemical properties of enzyme-substrate recognition as well as by sequence similarity.

Overall, the descriptor-enriched Random Forest model provides a stronger and more biologically meaningful representation of Cas9 specificity than sequence-only modeling. By linking off-target prediction to thermodynamic and structural DNA descriptors, the study contributes to computational enzyme engineering and supports safer CRISPR-based gene therapy and bioprocess development. Future work should add full metric reporting, external validation, and detailed feature-importance analysis to confirm which physicochemical descriptors most strongly influence Cas9 cleavage specificity.

References

- Anders, C., Niewoehner, O., Duerst, A., & Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, 513(7519), 569–573. <https://doi.org/10.1038/nature13579>
- Boyle, E. A., Andreasson, J. O. L., Chircus, L. M., Sternberg, S. H., Wu, M. J., Guegler, C. K., et al. (2021). Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Science Advances*, 7(8), Article eabe5496. <https://doi.org/10.1126/sciadv.abe5496>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, J. S., Dagdas, Y. S., Kleinstiver, B. P., Welch, M. M., Sousa, A. A., Harrington, L. B., et al. (2017). Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature*, 550(7676), 407–410. <https://doi.org/10.1038/nature24268>
- Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., et al. (2018). DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. *Genome Biology*, 19, Article 80. <https://doi.org/10.1186/s13059-018-1459-4>

- Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, (2), 184–191. <https://doi.org/10.1038/nbt.3437>
- Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213), Article 1258096. <https://doi.org/10.1126/science.1258096>
- Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., et al. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, 31(9), 822–826. <https://doi.org/10.1038/nbt.2623>
- Hasanzadeh, A., Khalilov, R., Abasi, E., Saghfi, S., Nasibova, A., & Akbarzadeh, A. (2017). Development of doxorubicin-adsorbed magnetic nanoparticles modified with biocompatible copolymers for targeted drug delivery in lung cancer. *Advances in Biology and Earth Sciences*, 2(1), 5–21. https://www.academia.edu/download/101469922/Hasanzadeh_20A.pdf
- Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, 31(9), 827–832. <https://doi.org/10.1038/nbt.2647>
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096), 816–821. <https://doi.org/10.1126/science.1225829>
- Khomutov, G. B., Potapenkov, K. V., Koksharov, Y. A., Trubitsin, B. V., Tikhonov, A. N., Mamedov, M. D., Nasibova, A. N., et al. (2014). *Magnetic nanoparticles in biomimetic and biological systems: Generation of iron oxide magnetic nanoparticles in DNA complexes, isolated chloroplasts and high plants* [Conference presentation]. XII International Conference on Nanostructured Materials (Nano-2014), Moscow, Russia. <https://istina.ficp.ac.ru/publications/article/8002113/>
- Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., et al. (2015). Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods*, 12(3), 237–243. <https://doi.org/10.1038/nmeth.3284>
- Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., et al. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, 529(7587), 490–495. <https://doi.org/10.1038/nature16526>
- Lazzarotto, C. R., Malinin, N. L., Li, Y., Zhang, R., Yang, Y., Lee, G., et al. (2020). CHANGE-seq reveals genetic and epigenetic effects on CRISPR-Cas9 genome-wide activity. *Nature Biotechnology*, 38(11), 1317–1327. <https://doi.org/10.1038/s41587-020-0555-7>
- Lin, Y., Cradick, T. J., Brown, M. T., Deshmukh, H., Ranjan, P., Sarode, N., et al. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research*, 42(11), 7473–7485. <https://doi.org/10.1093/nar/gku402>
- Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., et al. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering*, 2(1), 38–47. <https://doi.org/10.1038/s41551-017-0178-6>
- Mahmoodpoor, F., Ardalan, M., Nasibova, A., Valiyeva, M., & Ahmadian, E. (2024). The nephrotoxicity of checkpoint inhibitors: A mini review. *International Journal of Drug Research in Clinics*, 2(1), Article e1. <https://doi.org/10.34172/ijdr.2024.e1>
- Molla, K. A., & Yang, Y. (2020). Predicting CRISPR/Cas9-induced mutations for precise genome editing. *Trends in Biotechnology*, 38(2), 136–141. <https://doi.org/10.1016/j.tibtech.2019.08.002>
- Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., et al. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 156(5), 935–949. <https://doi.org/10.1016/j.cell.2014.02.001>

- Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., & Liu, D. R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature Biotechnology*, 31(9), 839–843. <https://doi.org/10.1038/nbt.2673>
- Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, 8(11), 2281–2308. <https://doi.org/10.1038/nprot.2013.143>
- SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4), 1460–1465. <https://doi.org/10.1073/pnas.95.4.1460>
- Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., & Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. *Science*, 351(6268), 84–88. <https://doi.org/10.1126/science.aad5227>
- Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., & Joung, J. K. (2017). CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nature Methods*, 14(6), 607–614. <https://doi.org/10.1038/nmeth.4278>
- Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, 33(2), 187–197. <https://doi.org/10.1038/nbt.3117>
- Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., et al. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nature Communications*, 10, Article 4284. <https://doi.org/10.1038/s41467-019-12281-8>
- Wienert, B., Wyman, S. K., Richardson, C. D., Yeh, C. D., Akcakaya, P., Porritt, M. J., et al. (2019). Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science*, 364(6437), 286–289. <https://doi.org/10.1126/science.aav9023>
- Zhang, G., Zeng, T., Dai, Z., & Dai, X. (2021). Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. *Computational and Structural Biotechnology Journal*, 19, 1445–1457. <https://doi.org/10.1016/j.csbj.2021.03.001>
- Zhang, Z., Zhang, Y., Gao, F., Han, S., Cheah, K. S., Tse, H. F., et al. (2023). Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: A review. *Briefings in Bioinformatics*, 24(1), Article bbac559. <https://doi.org/10.1093/bib/bbac559>