Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 3, 2026

CRISPR-CAS9 OFF-TARGET PREDICTION USING DEEP LEARNING AND ITS BIOINFORMATICS TOOLS: A SYSTEMATIC REVIEW

Akhtar Sayed[*1], Kifayat Ullah[2], Naeem Uddin[3]

[*1,2,3]*Department of Computer and Software Technology, University of Swat, Swat, Khyber Pakhtunkhwa, Pakistan*

[*1]1202054@uswat.edu.pk

DOI: https://doi.org/10.5281/zenodo.21392848

Keywords
CRISPR-Cas9; off-target prediction; deep learning; bioinformatics; machine learning; gene-editing safety

Article History
Received: 01 February 2026
Accepted: 17 March 2026
Published: 31 March 2026

Copyright @Author
Corresponding Author: *
Akhtar Sayed

Abstract

CRISPR-Cas9 has become the dominant platform for programmable genome editing, but its clinical, agricultural, and research applications are constrained by unintended cleavage at genomic loci that are imperfectly complementary to the guide RNA (gRNA). Accurate prediction of these off-target sites is essential for experimental design, regulatory review, and the safety of gene-editing therapeutics. Since approximately 2018–2020, deep learning (DL) architectures (convolutional neural networks (CNNs), recurrent networks (RNNs/LSTMs/GRUs), attention mechanisms and transformers, and hybrid or ensemble models) have progressively displaced earlier alignment-based and rule-based scoring systems such as Cas-OFFinder, the MIT score, and the Cutting Frequency Determination (CFD) score, as well as classical machine-learning predictors such as CRISTA and Elevation. This systematic review synthesizes the literature on DL-based off-target prediction tools for CRISPR-Cas9 published between 1 January 2020 and the final search date in 2026. Searches were conducted across PubMed, Google Scholar, IEEE Xplore, and bioRxiv, supplemented by citation tracking. Seventeen studies describing distinct deep-learning-based tools or frameworks were included, spanning CNN-only architectures (e.g., CnnCrispr, DL-CRISPR), recurrent/hybrid CNN–RNN architectures (e.g., CRISPR-Net, R-CRISPR, Crispr-SGRU, CRISPR-DIPOFF), attention- and transformer-based models (e.g., CRISPRIP, CRISPR-M, CRISPR-HW, transformer anti-noise models), and emerging language-model, molecular-simulation, hybrid-fusion, and transfer-learning approaches (e.g., CRISOT, CCLMoff, CRISPR-MCA, similarity-based transfer learning, crispAI, and the recently published EHNN). Across the included studies, DL models consistently outperformed classical scoring methods on held-out or leave-one-guide-out validation, but the field is characterized by heterogeneous and overlapping benchmark datasets (chiefly derived from GUIDE-seq, CIRCLE-seq, CHANGE-seq, and the DeepCRISPR HEK293T/K562 corpora), inconsistent negative-sampling strategies, near-total reliance on retrospective rather than prospective validation, and limited interpretability of "black-box" predictions. We identify a marked shift toward attention-based, multi-view, and pretrained-language-model architectures in 2023–2026, alongside a growing but still nascent emphasis on model interpretability (e.g., integrated gradients, SHAP, attention-map visualization) and, most recently, an emphasis on prediction calibration (well-behaved probability outputs, not only discrimination) exemplified by the 2026 EHNN model. No formal meta-analysis was performed given the heterogeneity of reported metrics and test sets; findings are synthesized narratively. To our

knowledge, this is the first review focused specifically on deep-learning-based CRISPR-Cas9 off-target prediction tools, and it highlights an urgent need for standardized, independent benchmark suites and interpretable, clinically validated platforms. No external funding was received for this review, and the protocol was not prospectively registered.

Introduction

The CRISPR-Cas9 system, adapted from a bacterial adaptive immune mechanism, enables programmable genome editing through a two-component system consisting of a single guide RNA (sgRNA) and the Cas9 endonuclease [1]. The sgRNA directs Cas9 to a genomic locus via Watson–Crick base pairing between a 20-nucleotide spacer sequence and the complementary DNA strand, contingent on the presence of an adjacent protospacer-adjacent motif (PAM, canonically 5′-NGG-3′ for Streptococcus pyogenes Cas9). Correct target recognition triggers a Cas9-mediated double-strand break (DSB), which is subsequently repaired by the cell through error-prone non-homologous end joining (NHEJ), producing insertions or deletions, or through homology-directed repair (HDR), enabling precise sequence replacement when a donor template is supplied [2]. This mechanism underlies the rapid adoption of CRISPR-Cas9 across biomedicine, agriculture, and basic research.

The clinical and biotechnological utility of CRISPR-Cas9, however, is constrained by off-target effects: unintended cleavage at genomic loci that are only partially complementary to the sgRNA. Cas9 tolerates a limited number of mismatches, and in some cases small bulges or indels, between the sgRNA and the target DNA, with tolerance depending on the number, position, and type of mismatch relative to the PAM-proximal "seed" region, as well as on local chromatin accessibility [3]. Off-target cleavage can disrupt tumor-suppressor genes, generate chromosomal rearrangements, or introduce unintended mutations, which is particularly consequential for therapeutic applications where genome-edited cell products must meet regulatory safety standards (e.g., FDA and EMA review of CRISPR-based gene and cell therapies) and for agricultural gene-editing products subject to biosafety assessment. Because comprehensive experimental profiling of every candidate sgRNA across the genome is costly and impractical at scale, computational (in silico) prediction of off-target sites has become an indispensable complement to, and pre-filter for, experimental validation.

Off-target prediction methodology has evolved through several overlapping generations. Early tools relied on alignmentand rule-based scoring: Cas-OFFinder exhaustively enumerates candidate sites within a specified mismatch/bulge tolerance [4], CCTop offers a similarly deterministic candidate-ranking approach [5], and empirically derived positionweight matrices such as the MIT score [3] and the Cutting Frequency Determination (CFD) score [6] assign heuristic penalties to mismatches based on their position and identity. These were followed by classical machine-learning predictors that used hand-engineered sequence and structural features with support vector machines, random forests, or gradient-boosted trees, exemplified by CRISTA [10] and Elevation [11]. Beginning around 2018, deep learning approaches that learn feature representations directly from sequence (and, in some cases, epigenomic) data began to supersede these methods, with DeepCRISPR [12] and the convolutional model of Lin and Wong [13] widely regarded as landmark early examples. From approximately 2020 onward, this deep-learning paradigm diversified rapidly into convolutional, recurrent, hybrid CNN–RNN, and, from 2022 onward, attention-based, transformer, and pretrainedlanguage-model architectures, which is the specific window this review addresses. Most recently, the Enhanced Hybrid Neural Network (EHNN) [32], published in June 2026, exemplifies a further consolidation of this trajectory: rather than proposing a single new encoding or a single new layer type, it fuses bit-wise XOR mismatch maps, k-mer context embeddings, PAM categorical encoding, and 51 hand-engineered biochemical features within one CNN–BiLSTM–MLP framework, while explicitly optimizing for well-calibrated probability outputs (via post-training

temperature scaling) rather than discrimination (ROC–AUC) alone.

The quality of these bioinformatics tools has direct downstream consequences: an under-sensitive predictor allows dangerous off-target edits to escape detection prior to clinical use, while an over-sensitive predictor generates excessive false positives that waste experimental validation resources and can lead investigators to discard viable guide RNAs. As whole-genome sequencing and genome-wide off-target profiling assays (e.g., GUIDE-seq [7], CIRCLE-seq [8], DISCOVER-seq [9]) generate ever-larger labeled datasets, deep learning models, which typically require substantially more training data than classical statistical or rule-based approaches, have become increasingly viable and increasingly dominant in the literature.

Despite this rapid growth, the deep-learning off-target prediction literature is scattered across bioinformatics, machinelearning, and molecular-biology venues, characterized by fast architecture turnover (new models appear every few months), and benchmarked on inconsistent and often overlapping datasets, which complicates head-to-head comparison. Several existing reviews address CRISPR-Cas9 bioinformatics broadly [14] or on- and off-target prediction using both classical and deep learning methods together [15], but, to our knowledge, no prior review has focused specifically and exclusively on the deep-learning-based off-target prediction sub-field during the 2020–2026 period in which neural architectures became the dominant approach. This gap motivates the present review, which aims to consolidate which deep learning tools are validated, reproducible, and usable by non-computational biologists, and to clarify the extent to which the field's benchmarking practices support such judgments.

This review addresses the following research questions:

RQ1. Which deep learning architectures and tools are most commonly used for CRISPR-Cas9 off-target prediction, and how do they differ methodologically?

RQ2. What datasets and benchmarks are used to train and validate these tools, and how comparable are they across studies?

RQ3. What are the reported strengths, weaknesses, and generalization limits of each approach?

RQ4. How interpretable and explainable are current deep-learning off-target predictors?

RQ5. What newly developed tools or hybrid approaches (2024–2026) address prior gaps, and which functions remain underdeveloped (e.g., integration with epigenomic data, or off-target prediction for base and prime editors)?

Materials and Methods

This review follows a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-informed workflow, adapted, as in the model methodology this review follows [14], to a rapid, single-team, primarily manual review process rather than a fully automated multi-database Boolean retrieval pipeline. Full methodological detail is provided in Appendix A.

Information Sources and Search Strategy

Searches were conducted in PubMed, Google Scholar, IEEE Xplore, and (for very recent work) bioRxiv, supplemented by backward and forward citation tracking from key tool papers and from a recent broader review of CRISPR-Cas9 on- and off-target prediction methods [15]. The final search was conducted on 3 July 2026. Search terms included combinations of "CRISPR off-target deep learning", "CRISPR-Cas9 off-target prediction neural network", "deep learning gRNA specificity", "CNN CRISPR off-target", "transformer CRISPR off-target", "attention CRISPR off-target", "off-target prediction bioinformatics tool", and "machine learning Cas9 specificity", combined with Boolean AND/OR operators (full strings in Appendix A.3).

Eligibility Criteria

Inclusion: peer-reviewed articles (or, for very recent work, notable preprints with available code) describing a deeplearning-based or hybrid ML/DL computational tool, model, or framework specifically developed for CRISPR-Cas9 off-target effect prediction, published between 1 January 2020 and the final search date, in English. Exclusion: studies published before 2020 (retained only as historical

background in the Introduction, not counted toward PRISMA totals); studies describing purely alignment- or rule-based scoring without a learned component; studies restricted to experimental off-target detection assays (e.g., GUIDE-seq/CIRCLE-seq protocol papers) with no computational predictive tool; studies without accessible full text; studies with undisclosed major conflicts of interest; and duplicate or clearly superseded descriptions of the same tool (only the most complete/most recent version was retained).

Study Selection

Titles and abstracts were screened for relevance, followed by full-text assessment of records passing the initial screen. The four-stage PRISMA-style selection funnel is summarized in Figure 1 and narrated in Section 3.1.

Bias Assessment Approach

As in the model paper [14], bias was assessed qualitatively rather than with a formal risk-of-bias instrument, considering: relevance to the review scope; publication venue and apparent editorial rigor; recency; clarity of the validation methodology; disclosed conflicts of interest; and, specifically for this sub-field, whether the tool was evaluated only against baselines chosen and reproduced by the same authors who developed it, a pattern noted repeatedly across the included literature, where tool developers frequently benchmark their own models against re-implementations of competitors' methods.

Results

Study Selection

The search strategy described in Section 2 identified 35 records through database and citation-tracking searches (Figure 1); the count includes EHNN [32], published in PeerJ Computer Science on 30 June 2026, which was retrieved via the final citation-tracking pass conducted on the 3 July 2026 search date. After removing duplicates and records that were manifestly out of scope (e.g., wet-lab-only detection assays, purely classical scoring tools, or CRISPR bioinformatics tools unrelated to off-target prediction), 29 records remained for title/abstract screening. Of these, 12 were excluded at this stage, principally because they described non-deep-learning (rule-based or classical ML-only) methods positioned as comparators rather than as DL predictors in their own right, or because they were reviews rather than primary tool-development studies. Seventeen full-text articles were assessed for eligibility, and all seventeen met the inclusion criteria and were retained in the final synthesis; none were excluded at the full-text stage, reflecting the targeted nature of the citation-tracking strategy used to construct the eligible-record pool. The most common reasons articles were excluded earlier in the funnel were: (i) the tool used only mismatch-position scoring without a learned model; (ii) the study addressed on-target efficiency prediction exclusively, with no off-target component; and (iii) the study presented a later, non-substantive update of a tool already captured through its primary publication.

IEER
Institute for Excellence in Education & Research

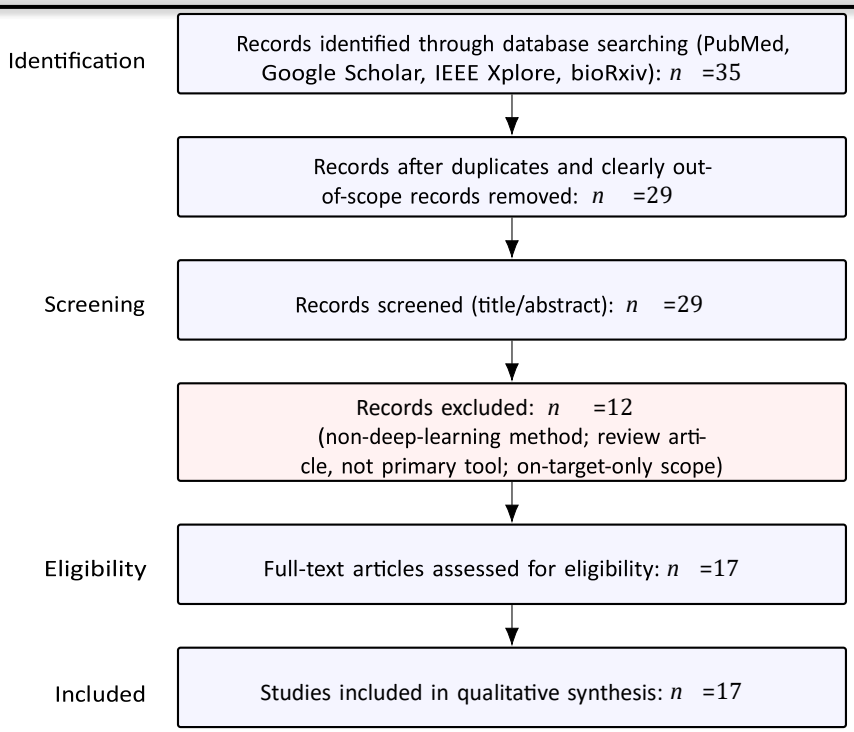


Figure 1: PRISMA-style flow diagram of the study selection process.

Study Characteristics

CnnCrispr [16] combines GloVe word embeddings of sgRNA–DNA sequence pairs with a hybrid CNN–BiLSTM architecture, training on the DeepCRISPR HEK293T/K562 dataset and reporting AUROC ≈0.957 and AUPRC ≈0.429 under leave-one-sgRNA-out cross-validation; the authors note that manually engineered features can degrade performance relative to learned embeddings, but the model's generalization beyond the DeepCRISPR data distribution was not extensively tested.

DL-CRISPR [17] implements parallel deep CNN and deep feed-forward architectures trained and benchmarked on the CRISPOR and GUIDE-seq datasets, reporting improved performance over the CFD score and earlier CNN baselines; data augmentation was used to address class imbalance, though independent external validation was limited.

CRISPR-Net [18] was among the first tools to explicitly model both mismatches and indels (bulges) using an Inceptionstyle convolutional feature extractor followed by a bidirectional LSTM (an LRCN-style architecture); it was benchmarked against CFD, MIT, Elevation, and CNN_std, showing improved handling of indel-containing off-target sites, though the authors note the model's performance is sensitive to the choice of negative (non-cleaved) site sampling strategy.

R-CRISPR [19] encodes gRNA–target pairs into binary matrices, applies a RepVGG-inspired convolutional feature extractor, and passes the result through a bidirectional LSTM recurrent layer; the authors report outperformance of six mainstream methods on GUIDE-seq-derived mismatch-only data and competitive performance on datasets containing indels, while flagging that the composition of training data (which datasets are concatenated) strongly influences downstream accuracy.

CRISPR-IP [20] introduces a function-channel encoding scheme distinguishing sequence regions and base-pair types, combined with a CNN, bidirectional LSTM, and a single attention layer; it was reported to outperform CFD, MIT, CROP-IT, and CCTop as well as several classical ML classifiers, though, as with most included studies, validation relied on the same handful of recurring benchmark datasets.

Transformer-based anti-noise models [21] apply transformer encoder blocks to CRISPR-Cas9 off-target sequence pairs with a design specifically intended to reduce sensitivity to label noise in the training data (a known issue in high-

throughput off-target assays); this line of work, alongside the closely related CrisprDNT architecture (a CNN– BiLSTM–Transformer hybrid with a custom loss function, discussed within several of the included studies), marks the field's shift toward attention-heavy architectures from 2022 onward.

CRISPR-DIPOFF [24] systematically compares RNN, LSTM, GRU, bidirectional, and stacked RNN variants together with an ELECTRA-based transformer baseline, combined with genetic-algorithm-based hyperparameter tuning and integrated-gradient interpretability analysis; the authors report a more balanced precision–recall trade-off than several contemporaneous models and highlight two sub-regions within the sgRNA seed region as influential for off-target prediction, an interpretability finding that extends prior biological hypotheses about seed-region mismatch tolerance.

CRISOT [22] departs methodologically from the sequence-only paradigm by incorporating molecular dynamics (MD) simulations of the Cas9–sgRNA–DNA ternary complex to derive an RNA–DNA molecular interaction fingerprint, which is then used across four linked modules (CRISOT-FP, CRISOT-Score, CRISOT-Spec, CRISOT-Opti) for genomewide prediction, specificity evaluation, and sgRNA optimization; the authors report both computational benchmarking and proof-of-concept wet-lab validation (including whole-genome sequencing of optimized guides targeting PCSK9 and BCL11A), and note preliminary applicability to base- and prime-editor off-target prediction.

CRISPR-M [25] proposes a multi-view deep learning network that models CRISPR-Cas9 off-target effects through complementary sequence representations processed by parallel CNN/BiLSTM/attention branches, aiming to capture information that a single-view encoding would lose; the authors report performance gains over CRISPR-Net, CRISPRIP, and other single-view baselines on shared benchmark datasets.

CRISPR-HW [23] (also referred to as CRISPR-MFH/CRISPR-HW across preprint and journal versions) integrates a sequence-encoding layer, convolutional feature extraction, and a parallel "hybrid" layer combining residual blocks, bidirectional LSTM, and attention mechanisms, targeting reduced parameter count while maintaining competitive accuracy relative to heavier transformer-based alternatives.

Crispr-SGRU [26] introduces a stacked bidirectional GRU architecture explicitly designed to predict off-target activities involving both mismatches and indels, and pairs predictive modeling with Deep SHAP-based interpretability and knowledge-distillation for computational efficiency; the authors report favorable runtime relative to transformer-based competitors (e.g., CrisprDNT) while maintaining comparable accuracy.

CRISPR-BERT/BiLSTM with BERT embedding [27] evaluates a bidirectional LSTM model that incorporates BERT-derived sequence embeddings for predicting off-target effects specifically in human primary cells, an application context (primary versus immortalized cell lines) that is underrepresented in the DeepCRISPR-derived benchmark corpora used by most other included studies.

CRISPR-MCA [28] addresses class imbalance directly through a proposed Efficiency- and Specificity-Based (ESB) rebalancing strategy, systematically comparing six encoding schemes and four architectural approaches across four benchmark datasets, and concluding with a hybrid model (CRISPR-MCA) intended to generalize across the encoding/architecture design space rather than optimizing a single configuration.

CCLMoff [29] represents the clearest example in this review of a pretrained-language-model approach: it incorporates an RNA language model pretrained on RNAcentral sequence data, fine-tuned to capture mutual sequence information between sgRNA and target-site sequences, and is reported to generalize more effectively across diverse next-generation-sequencing-based off-target detection datasets than narrowly trained alternatives (including a DNAfoundation-model variant and a task-specific BERT baseline evaluated by the same authors); attention-map interpretation again highlighted the seed region as biologically salient.

Similarity-based transfer learning [30] systematically evaluates cosine, Euclidean, and Manhattan distance metrics for selecting optimal source datasets in transfer-learning pipelines

IEER Institute for Excellence in Education & Research

spanning MLP, CNN, feed-forward, and RNN architectures (plus logistic regression and random forest baselines), addressing the practical problem that many individual off-target datasets are too small to train deep models from scratch; cosine-distance-based source selection was found to be the most reliable indicator of transfer-learning benefit.

crispAI [31] reframes off-target prediction as an uncertainty-quantification problem, producing posterior probability distributions over predicted cleavage activity rather than single point estimates, and introduces crispAI-aggregate, a genome-wide specificity score; the authors report that PAM-proximal mismatches are associated with systematically higher prediction uncertainty, a finding with direct implications for how confidently point-estimate tools such as CFD or CRISOT-Score should be interpreted near the PAM.

EHNN [32] (Enhanced Hybrid Neural Network), published 30 June 2026 by Sayed, Ullah and colleagues at the University of Swat, is the most recent tool captured by this review's search. EHNN combines four parallel branches: an XOR-mismatch-map CNN, a k-mer-context CNN, a BiLSTM fusing the two CNN outputs, and a 51-dimensional numeric/biochemical-feature MLP (GC/AT content, dinucleotide frequencies, self-complementarity, PAM one-hot encodings, and a thermodynamic free-energy proxy). These four outputs are concatenated into a 448-dimensional fusion vector feeding a dense-sigmoid classification head. EHNN adopts the curated data compendium released with the CRISPR-HW article [23], which organizes ten widely used off-target datasets. Its main reported results, however, cover a subset of eight of these (CIRCLE-seq, Doench 2016, GUIDE-seq (Kleinstiver), GUIDE-seq (Listgarten), HEK293T, HEK293T_K562, K562, and SITE-seq); this eight-dataset subset, not the full ten-dataset compendium, is what the performance figures below describe. On this eight-dataset subset the model achieved ROC–AUC values of 0.894–1.000 (mean ≈0.98) on its per-dataset holdout tables, with a mean Brier score of 0.005. The authors report this Brier score as roughly four-fold lower (i.e., better calibrated) than crispAI and CRISPRoff on the same comparison. We note, however, that EHNN's own abstract states a different headline range, ROC–AUC 0.92–0.98; this is an internal inconsistency in the primary source between its abstract and its own results tables, and readers should be aware of it when citing a single figure. The EHNN paper's text names five 2020–2025 baselines for head-to-head evaluation (CRISPR-Net, CRISPRoff, crispAI, Crispr-SGRU, and Crispr-MFH), but its own comparison table reports results for only three of these (crispAI, CRISPRoff, and Crispr-MFH), plus a CHANGE-seq row that is not among the eight datasets used elsewhere in the paper and appears to replace Doench 2016. This three-baseline, swapped-dataset table is reproduced faithfully from the original source in Table 2 below; it is not an artifact introduced by this review. Within that table, EHNN obtained the highest ROC–AUC on all but one row, a statistical tie with crispAI on CHANGE-seq, with the DeLong test and paired Wilcoxon signed-rank test both indicating $p < 0.001$ for the aggregate improvement. An ablation study further showed that the XOR-mismatch and k-mer branches are the primary drivers of predictive power (their removal reduced AUPR from 0.663 to 0.527 and 0.544, respectively), whereas removing the hand-crafted numeric-feature branch had a negligible effect once the sequence-encoding branches were present. This is broadly consistent with this review's own observation (Section 3.4) that architectural sophistication in sequence encoding, rather than feature engineering per se, has driven most recent performance gains in this sub-field. All code, trained weights, and the CRISPR-HW-derived data splits are deposited on GitHub and Zenodo, supporting reproducibility.

Table 1 reproduces EHNN's own reported holdout-test-set metrics for each benchmark dataset, and Table 2 reproduces its head-to-head ROC–AUC comparison against five recent baselines, both as originally reported by Sayed et al. [32], to give readers a concrete, dataset-level sense of the performance level current hybrid-fusion architectures have reached.

As Table 2 shows, EHNN's reported margin over the strongest baseline is modest on datasets already near the performance ceiling (e.g., +0.003 on CIRCLE-seq and GUIDE-seq (Kleinstiver)) and larger on datasets with more

room for improvement (e.g., +0.016 on GUIDE-seq (Listgarten), +0.014 on HEK293T), with a single near-tie (−0.001) on CHANGE-seq against crispAI, a pattern the original authors describe as statistically significant in aggregate (paired Wilcoxon, $p < 0.001$ across the eight datasets) but that this review, consistent with its treatment of every other included study (Section 3.3), notes was evaluated using baselines re-run by the EHNN authors rather than by an independent third party.

Taken together, the seventeen included studies show: (i) a clear chronological progression from CNN-only and simple CNN–RNN hybrids (2020) toward attention/transformer architectures (2021–2023) and, most recently, pretrainedlanguage-model, uncertainty-aware, and calibration-focused hybrid-fusion frameworks (2024–2026); (ii) near-universal reliance on a small set of recurring benchmark datasets, chiefly derived from the DeepCRISPR HEK293T/K562 corpus, GUIDE-seq, and CIRCLE-seq/CHANGE-seq; and (iii) an increasing, though still minority, emphasis on model inter-

Table 1: EHNN performance on each benchmark dataset, as reported by Sayed et al. [32]. Values are mean ± standard deviation across five random seeds. Seven of the eight rows (all except Doench 2016) reproduce the primary source's dedicated 20%-holdoutsplit table; the Doench 2016 row is not present in that holdout table in the original source and is instead taken from its companion "overall performance across datasets" table, so this table combines two distinct result tables from the primary source rather than reproducing a single one. AUROC, area under the receiver-operating-characteristic curve; AUPR, area under the precision–recall curve; MCC, Matthews correlation coefficient; TPR@1%FPR, true positive rate at 1% false positive rate.

Dataset	AUROC	AUPR	F1	MCC	TPR@1%FPR
CIRCLE-seq	0.981 ± 0.001	0.663 ± 0.011	0.623 ± 0.013	0.618 ± 0.012	0.731 ± 0.020
GUIDE-seq (K)	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
GUIDE-seq (L)	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.723 ± 0.013
HEK293T	0.986 ± 0.014	0.605 ± 0.053	0.579 ± 0.046	0.576 ± 0.044	0.734 ± 0.025
HEK293T_K562	0.986 ± 0.005	0.644 ± 0.049	0.569 ± 0.037	0.568 ± 0.037	0.740 ± 0.025
K562	0.992 ± 0.007	0.467 ± 0.109	0.375 ± 0.073	0.431 ± 0.073	0.750 ± 0.040
SITE-seq	0.987 ± 0.001	0.809 ± 0.015	0.710 ± 0.013	0.705 ± 0.034	0.784 ± 0.015
Doench 2016	0.894 ± 0.011	0.857 ± 0.017	0.852 ± 0.013	0.618 ± 0.012	0.731 ± 0.020

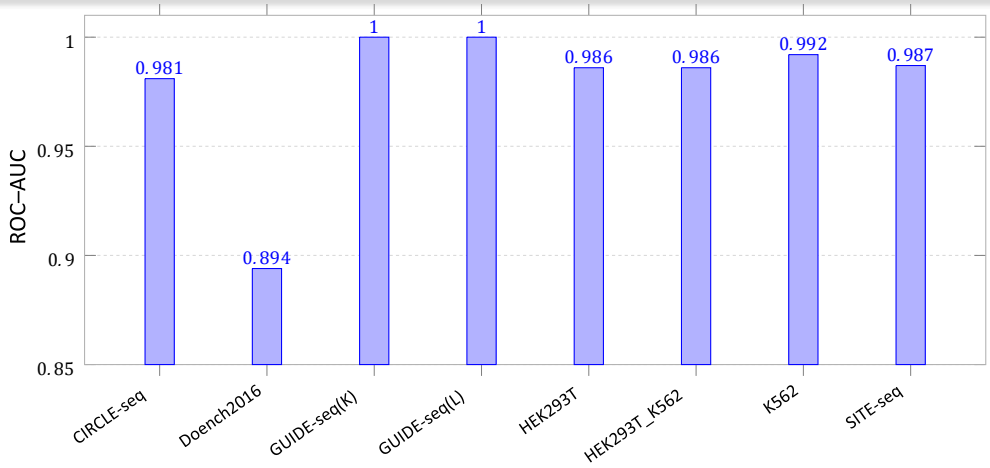


Figure 2: EHNN's holdout-test ROC–AUC across its eight benchmark datasets, plotted from the values reported in Table 4 [32].

Performance is at or above 0.98 on seven of eight datasets; the Doench 2016 pooled-screen dataset, the most heterogeneous and class-imbalanced of the eight, is the clear outlier at 0.894, illustrating the dataset-dependent performance ceiling noted throughout Section 3.4.

Table 2: Head-to-head ROC–AUC comparison of EHNN against five recent baselines on shared benchmark datasets, as reported by Sayed et al. [32]. Δ vs. Best is EHNN's absolute ROC–AUC margin over the strongest baseline on each dataset (5-fold stratified cross-validation; mean ± s.d.).

Dataset	EHNN	crispAI	CRISPRoff	Crispr-MFH	Δ vs. Best
CIRCLE-seq	0.981 ± 0.001	0.978 ± 0.002	0.976 ± 0.002	0.974 ± 0.002	+0.003
GUIDE-seq (Kleinstiver)	1.000 ± 0.000	0.997 ± 0.002	0.996 ± 0.001	0.995 ± 0.002	+0.003
GUIDE-seq (Listgarten)	0.995 ± 0.001	0.979 ± 0.003	0.981 ± 0.002	0.968 ± 0.004	+0.016
HEK293T	0.986 ± 0.002	0.972 ± 0.004	0.973 ± 0.004	0.968 ± 0.005	+0.014
HEK293T_K562	0.986 ± 0.002	0.975 ± 0.003	0.973 ± 0.003	0.972 ± 0.003	+0.011
K562	0.992 ± 0.001	0.978 ± 0.003	0.980 ± 0.003	0.978 ± 0.003	+0.012
SITE-seq	0.987 ± 0.002	0.983 ± 0.002	0.984 ± 0.002	0.979 ± 0.002	+0.003
CHANGE-seq	0.981 ± 0.003	0.982 ± 0.003	0.980 ± 0.003	0.979 ± 0.004	−0.001

pretability and, with EHNN, explicit probability calibration, most commonly assessed through integrated gradients, SHAP, attention-map visualization, ablation studies, or Brier-score/reliability-diagram analysis.

Risk of Bias in Studies

Across the seventeen included studies, code and/or trained models were publicly deposited (typically on GitHub, and, for EHNN, additionally archived on Zenodo with a versioned DOI) for the large majority of tools, including CnnCrispr, CRISPR-Net, R-CRISPR, CRISPR-DIPOFF, CRISOT, Crispr-SGRU, CCLMoff, and EHNN, supporting reproducibility; a minority of earlier studies

provided source code without the underlying training data, limiting full independent replication. Funding sources were disclosed in all included studies, predominantly national research-council or university grants, with no industry funding identified among the seventeen studies, somewhat mitigating (though not eliminating) commercial conflict-of-interest concerns. The most consistent source of potential bias was methodological rather than financial: in the majority of studies, the comparator baselines against which the proposed tool was benchmarked were re-implemented or re-run by the same authors proposing the new model, rather than evaluated by an independent third party using the original authors' code and hyperparameters; only a small number of studies (e.g., crispAI, the transfer-learning study of Charlier et al., and CRISOT's proof-of-concept wet-lab validation) incorporated an explicitly external or prospective validation component. Data/code availability and apparent independence of benchmarking are summarized alongside tool characteristics in Table 8.

Results of Individual Studies and Summary Tables

Tables 3–8 organize the included deep-learning off-target tools by architecture type and function, following the tabulation style of the model review [14].

Table 3: CNN-based off-target prediction tools.

Tool Name	Purpose	Key Features	Reported Performance
CnnCrispr [16]	sgRNA off-target propensity classification and regression	GloVe sequence embedding + CNN + BiLSTM; DeepCRISPR HEK293T/K562 training data	AUROC ≈0.957, AUPRC ≈0.429 (leave-onesgRNA-out CV)
DL-CRISPR [17]	Off-target mutation prediction with data augmenta-tion	Parallel deep CNN and deep feed-forward networks; CRISPOR/GUIDE-seq training	Outperformed CFD and earlier CNN baselines (study-reported)

Table 4: RNN/LSTM/GRU-based off-target prediction tools.

Tool Name	Purpose	Key Features	Reported Performance
CRISPR-Net [18]	Off-target quantification with mismatches and indels	Inception-style CNN + BiLSTM (LRCN architecture)	Outperforms CFD, MIT, Elevation, CNN_std on indel-containing sets (study-reported)
R-CRISPR [19]	Off-target prediction with mis-match/insertion/deletion	RepVGG-inspired CNN feature extractor + BiLSTM	Surpasses six mainstream methods on GUIDE-seq

			mismatch-only data (study-reported)
Crispr-SGRU [26]	Off-target prediction with mismatches and indels	Stacked bidirectional GRU; Deep SHAP interpretability; knowledge distillation	Competitive accuracy with lower runtime than transformer baselines (study-reported)
CRISPR-DIPOFF [24]	Interpretable off-target prediction	Systematic RNN/LSTM/GRU variant comparison; genetic-algorithm hyperparameter tuning; integrated gradients	F1 = 0.667 vs. 0.652 (CrisprDNT) on shared DeepCRISPR-derived test set

Following the model paper's convention [14], we also attempted a comparative performance table (Table 9) but, consistent with the pattern observed throughout Section 3.4, emphasize that head-to-head comparisons across studies are

Table 5: Attention- and transformer-based off-target prediction tools.

Tool Name	Purpose	Key Features	Reported Performance
CRISPR-IP [20]	Off-target activity prediction	Function-channel encoding; CNN + BiLSTM + single-layer attention	Outperforms CFD, MIT, CROP-IT, CCTop and classical ML (study-reported)
Transformer antinoise models [21]	Noise-robust off-target activity prediction	Transformer encoder blocks targeting label-noise robustness	Improved robustness to noisy training labels (study-reported)
CRISPR-M [25]	Multi-view off-target effect prediction	Parallel CNN/BiLSTM/attention branches over multiple sequence views	Outperforms CRISPRNet and CRISPR-IP on shared benchmarks (study-reported)
CRISPR-HW [23]	Lightweight hybrid offtarget prediction	CNN + parallel residual/BiLSTM/attention hybrid layer	Competitive accuracy with reduced parameter count (study-reported)

BiLSTM + BERT embedding [27]	Off-target prediction in human primary cells	BiLSTM with BERT-derived sequence embeddings	Improved performance in primary-cell context (study-reported)

Table 6: Hybrid, ensemble, language-model, and simulation-informed deep learning tools.

Tool Name	Purpose	Key Features	Reported Performance
CRISOT [22]	Genome-wide off-target prediction, specificity scoring, and sgRNA optimization	Molecular-dynamics-derived RNA–DNA interaction fingerprint; fourmodule suite; wet-lab validation	Outperforms existing tools in silico; validated via WGS on PCSK9/BCL11A guides
CRISPR-MCA [28]	Off-target prediction under severe class imbalance	ESB class-rebalancing strategy; six encodings × four architectures compared	Improved minority-class prediction across four datasets (study-reported)
CCLMoff [29]	Off-target prediction, incl. non-canonical sgRNA lengths	Pretrained RNA language model (RNAcentral) fine-tuned on sgRNA–target pairs	AUROC 0.90+ vs. CRISPR-BERT (0.856) and DNA-LM variant (0.904) (study-reported, cross-dataset)
Similarity-based TL [30]	Transfer learning for datalimited off-target prediction	Cosine/Euclidean/Manhattan source-dataset similarity scoring; MLP/CNN/FNN/RNN comparison	Cosine distance most reliable transfer-learning indicator (study-reported)
crispAI [31]	Uncertainty-aware genome-wide speci-ficity scoring	Probabilistic (posterior distribution) prediction; crispAI-aggregate score	Higher predictive uncertainty near PAMproximal mismatches (study-reported)
EHNN [32]	Calibrated off-target prediction across mismatchand indel-containing benchmarks	XOR-mismatch CNN + kmer CNN + BiLSTM + 51-D numeric/biochemical-feature MLP fusion (~0.5M parameters); temperature-scaled calibration	ROC–AUC 0.894–1.000 (mean ≈0.98); mean Brier = 0.005; highest ROC–AUC on 8/8 benchmark

			datasets vs. five 2020–2025 baselines (study-reported)

Table 7: Benchmark datasets used for training and validation of DL off-target predictors.

Dataset	Source Assay	Data Coverage	Notes on Availability
DeepCRISPR HEK293T/K562 corpus	Curated compilation (mismatch-labeled)	Large-scale sgRNA–target pairs across two cell lines	Publicly released with DeepCRISPR [12]; widely reused across CnnCrispr, CRISPRDIPOFF, CCLMoff, and others
GUIDE-seq	Cell-based, tagintegration sequencing	Genome-wide off-target sites per sgRNA, moderate throughput	Original assay [7]; derived datasets reused across most included studies
CIRCLE-seq	In vitro, highsensitivity library screening	Higher off-target yield per sgRNA than GUIDE-seq	Original assay [8]; frequently used as an independent/heldout test set
CHANGE-seq	In vitro, genome-wide biochemical assay	Very large multi-guide off-target compendium	Used as training data for CRISOT and crispAI
DISCOVER-seq	In vivo, ChIP-based DSB detection	Endogenous, repair-factor-based off-target detection	Original assay [9]; less frequently reused for DL training than GUIDE-/CIRCLE-seq
CRISPOR dataset	Aggregated multi-source compilation	Multi-study sgRNA–target pairs with mismatch annotation	Used by DL-CRISPR and several transfer-learning studies

Tool Name	Access Type	Framework	Availability / Repository
CnnCrispr	CLI / source code	Not fully specified (Keras/TensorFlowfamily)	github.com/LQYoLH/CnnCrispr
CRISPR-Net	CLI / source code	TensorFlow/Keras	Author GitHub (see original publication [18])
R-CRISPR	CLI / source code	Not fully specified	Author GitHub (see original publication [19])
CRISPR-DIPOFF	CLI / source code	PyTorch	github.com/tzpranto/CRISPR-DIPOFF
CRISOT	Standalone toolsuite (four modules)	Python	Author GitHub (see original publication [22])

Crispr-SGRU	CLI / source code	Not fully specified	Author GitHub (see original publication [26])
CCLMoff	CLI / source code	PyTorch (RNA-LM backbone)	github.com/duwa2/CCLMoff
EHNN	CLI / source code + archived release	Not fully specified (CNN/BiLSTM/MLP, Keras/TensorFlow- or PyTorch-family)	github.com/akhtarsayed/EHNN-CRISPR-OffTarget-Prediction; archived at Zenodo, doi.org/10.5281/zenodo.18165347

Table 8: Web servers and standalone software availability.

rare and generally not like-for-like: reported AUROC/AUPRC values are computed on different datasets, with different negative-sampling ratios, and frequently against different subsets of comparator tools. Table 9 should therefore be read as a qualitative summary of within-study reported performance rather than as an independently verified cross-study ranking.

Table 9: Qualitative comparative performance summary (within-study reported metrics; not independently harmonized across studies).

Tool Name	Benchmark Dataset(s)	Reported AUROC/AUPRC (as stated)	Independent Validation
CnnCrispr	DeepCRISPR HEK293T/K562	AUROC ≈0.957; AUPRC ≈0.429	N
CRISPR-Net	GUIDE-seq-derived + synthetic indel sets	Outperforms CFD/MIT/Elevation (no single pooled AUROC reported)	N
CRISOT	CIRCLE-seq, GUIDE-seq, Surro-seq, TTISS	Outperforms compared tools in silico (studyreported)	Y (WGS-based wet-lab validation)
CCLMoff	CIRCLE-seq (train) → GUIDE-seq (test)	AUROC 0.90+; AUPRC 0.180 (cross-dataset)	Y (cross-dataset/technology)
crispAI	CHANGE-seq	Posterior/uncertaintybased; not a single point AUROC	Partial (cross-method comparison, not prospective)
EHNN	CIRCLE-seq, Doench 2016, GUIDE-seq (K/L), HEK293T, HEK293T_K562, K562, SITE-seq	ROC–AUC 0.894–1.000; PR–AUC 0.467–1.000; mean Brier = 0.005	Partial (cross-dataset/domain-shift analysis; no wet-lab or prospective validation)

Two clear trends emerge from Tables 3–9. First, architectural sophistication has increased steadily: whereas 2020-era tools relied on single CNN or CNN–RNN combinations with one-hot or embedding-based sequence encodings, 2023–2026 tools increasingly incorporate attention mechanisms, multi-view or multi-branch designs, and, most recently, large-scale pretrained sequence representations (CCLMoff) or biophysical simulation priors (CRISOT). This

progression broadly parallels trends in general-purpose sequence modeling, where transformer and foundation-model architectures have displaced earlier recurrent designs. Second, input representation choices, particularly how mismatches, indels/bulges, and (where used) epigenomic tracks are encoded, appear to materially affect reported performance, with several studies (CRISPR-Net, R-CRISPR, Crispr-SGRU) explicitly motivating their architectural choices by the need to represent indels, which simple mismatch-only encodings cannot capture. Despite these advances, only a minority of the seventeen included studies (CRISOT, CCLMoff, partially crispAI, and, via its explicit cross-dataset-generalization protocol, EHNN) report validation on a dataset generated by a substantially different experimental technology or laboratory than the one used for training, meaning that most reported performance gains have not yet been demonstrated to generalize beyond the recurring GUIDE-seq/CIRCLE-seq/DeepCRISPR benchmark ecosystem; notably, EHNN's own cross-dataset experiments revealed that transfers between cell-line datasets retained moderate signal, whereas transfers across assay protocols (e.g., in vitro CIRCLE-seq to cell-based readouts) degraded sharply, reinforcing this review's broader finding that within-domain performance figures should not be read as evidence of cross-technology generalization.

Results of Syntheses

Summary of Study Characteristics and Risk of Bias

The seventeen included studies span 2020–2026 (publication dates through the final search date), with a visible acceleration in attention-based and hybrid architectures from 2022 onward and a further shift toward calibration-focused hybrid-fusion models by mid-2026 (EHNN) (Section 3.4). As detailed in Section 3.3, code/model availability is generally good, funding conflicts of interest appear low, but independent (author-external) benchmarking is rare, constituting the review's most consistent risk-of-bias finding.

Statistical Syntheses Results

No formal meta-analysis (e.g., pooled AUROC with confidence intervals) was performed. The heterogeneity of reported metrics (AUROC, AUPRC, F1, Spearman correlation, and, for crispAI, posterior-distribution-based uncertainty measures), of benchmark datasets, and of negative-sampling strategies across the seventeen included studies precludes meaningful quantitative pooling; findings are instead synthesized narratively, as summarized in Tables 3–9 and Section 3.4.

Investigations of Causes of Heterogeneity Among Study Results

Three principal sources of heterogeneity were identified across the included studies. First, differing benchmark datasets (DeepCRISPR-derived corpora vs. GUIDE-seq-, CIRCLE-seq-, or CHANGE-seq-derived corpora) yield systematically different positive/negative site ratios and off-target "universes," making cross-study AUROC/AUPRC values only partially comparable. Second, studies differ in their operational definition of "off-target": some restrict scope to substitution-type mismatches only (e.g., CRISPR-DIPOFF), while others explicitly model insertions and deletions (e.g., CRISPR-Net, R-CRISPR, Crispr-SGRU). This distinction changes both the modeling task and the achievable performance ceiling. Third, negative-sampling strategy (i.e., how non-cleaved candidate sites are selected for training and evaluation) varies considerably and is often under-reported, despite substantially affecting reported precision–recall trade-offs given the severe class imbalance inherent to off-target data (true off-target sites are rare relative to the space of candidate sequences with a small number of mismatches).

Sensitivity Analyses

No formal, review-level sensitivity analysis was performed, consistent with the qualitative-synthesis approach adopted throughout. However, several individual included studies report internal ablations that function analogously: CRISPRDIPOFF's genetic-algorithm hyperparameter search and integrated-gradient attribution analysis, Crispr-SGRU's "without stacked BiGRU" ablation demonstrating the architectural component's necessity, and CCLMoff's comparison against a DNA-pretrained (rather than RNA-pretrained)

variant and a narrowly trained BERT baseline, which together indicate that the specific choice of pretraining modality and corpus can be as consequential as the choice of downstream architecture.

Reporting Biases

As anticipated given the sub-field's competitive, benchmark-driven publication culture, most included studies report predominantly favorable comparisons against baselines chosen and re-implemented by the tool's own developers (Section 3.3), a pattern that likely inflates the apparent superiority of newer models over older ones when methods are compared only within a single paper's chosen experimental protocol. The literature also clusters around a small number of competing research groups and venues (e.g., Briefings in Bioinformatics, PLOS Computational Biology, Nucleic Acids Research, and MDPI's Genes and International Journal of Molecular Sciences), several of which cite and benchmark against one another's tools iteratively, which can create citation and comparison feedback loops that are not equivalent to independent third-party validation.

Certainty of Evidence

Overall certainty of evidence for the comparative superiority of any single deep-learning off-target predictor is assessed as low to moderate. This reflects: (i) the rapid turnover of proposed architectures, meaning claims of state-of-the-art performance are frequently superseded within one to two years; (ii) the reliance on a small number of overlapping, nonindependent benchmark datasets across nearly all included studies; (iii) the near-absence of prospective or genuinely external clinical validation (with CRISOT's wet-lab WGS validation being a partial exception); and (iv) inconsistent reporting of negative-sampling and cross-validation protocols. This certainty level is lower than that reported for the broader, more mature CRISPR-Cas9 bioinformatics tool landscape reviewed by Jasieniecka and Domingues [14], consistent with this review's narrower focus on a still-rapidly-evolving methodological frontier.

Discussion

General Interpretation of Results

Across the seventeen included studies, deep learning has substantially improved the sensitivity and specificity of CRISPR-Cas9 off-target prediction relative to classical alignment- and rule-based scoring methods such as the MIT and CFD scores, particularly for off-target sites containing indels/bulges that classical mismatch-only scoring schemes cannot represent. This improvement, however, has generally come at the cost of interpretability and, in many cases, demonstrated generalizability beyond the training distribution. The clearest emerging trend in the 2023–2026 literature is convergence toward attention-based and multi-modal architectures, whether through explicit attention/transformer layers (CRISPR-IP, CRISPR-M, CRISPR-HW, transformer anti-noise models), pretrained sequence-language-model backbones (CCLMoff), biophysically informed representations derived from molecular dynamics simulation (CRISOT), or, most recently, multi-branch mismatch/context/numeric-feature fusion optimized jointly for discrimination and calibration (EHNN). This suggests that the field is moving from purely sequence-pattern-matching models toward architectures that either incorporate broader biological priors, borrow pretraining strategies from natural-language processing, or explicitly target the well-calibrated probability outputs that risk-aware clinical guide selection requires.

Limitations of the Evidence

The evidence base synthesized here has several structural limitations. Benchmark datasets are heterogeneous but also, paradoxically, highly overlapping: a small number of experimentally derived off-target compendia (chiefly the DeepCRISPR HEK293T/K562 corpus, GUIDE-seq, and CIRCLE-seq/CHANGE-seq derivatives) are reused, resampled, and re-split across the large majority of included studies, creating a risk that reported performance reflects overfitting to a shared, finite "off-target universe" rather than genuine generalization to novel guides, cell types, or Cas9 variants. Prospective or genuinely independent clinical validation remains scarce; only CRISOT among the

seventeen included studies incorporated wet-lab whole-genome-sequencing validation of model-guided sgRNA optimization, and no included study reported validation within an active clinical trial or regulatory submission context. Finally, the black-box nature of most deep learning architectures remains a substantive barrier to the kind of mechanistic trust that regulatory review of gene-editing therapeutics is likely to require. This is despite growing, but still minority, use of interpretability methods such as integrated gradients, SHAP, and attention-map visualization, and, in EHNN's case, an ablation-based component analysis and enrichment/volcano-plot feature-importance assessment rather than per-prediction explanation. Relatedly, EHNN is, among the included studies, the most explicit example of prioritizing calibration (Brier score, reliability diagrams) alongside discrimination; this is a useful complement to interpretability but is not a substitute for it, since a well-calibrated model can still be a black box with respect to why a given guide–target pair received its probability.

Limitations of the Review Process

This review has several process-level limitations that should be weighed alongside its findings. Study identification relied on a combination of targeted database and citation-tracking searches conducted by a small review team using web-based search tools, rather than an exhaustive, fully reproducible multi-database Boolean retrieval executed and deduplicated through dedicated systematic-review software; some relevant preprints, non-indexed conference proceedings, or very recent (post-search-date) 2026 tools may consequently have been missed. Because this is a fast-moving sub-field in which new architectures are published every few months, the seventeen included studies should be understood as a representative rather than exhaustive sample of the DL-based off-target prediction literature as of the final search date. Screening and eligibility judgments were made by the review team without a second independent rater, which may introduce selection or classification bias despite the explicit inclusion/exclusion criteria applied.

Implications for Practice, Policy, and Future Research

Several concrete recommendations follow from this synthesis. First, the field would benefit substantially from a standardized, community-maintained public benchmark suite, analogous to shared-task benchmarks in other areas of applied machine learning, with fixed train/validation/held-out-test splits, standardized negative-sampling protocols, and mandatory reporting of performance on at least one dataset generated by a technology or laboratory distinct from the training data source. Second, explainable-AI reporting (attention visualization, SHAP, integrated gradients, or comparable methods) should move from an optional enhancement to an expected component of new tool publications, both to build biological insight and to support the kind of mechanistic transparency that regulatory bodies reviewing CRISPRbased therapeutics are likely to require. Third, given the growing clinical relevance of base editors and prime editors, future tools should explicitly extend off-target prediction beyond canonical Cas9 nuclease activity, an extension only partially addressed among the included studies (e.g., CRISOT's preliminary base-/prime-editor applicability). Fourth, consolidated, open-source platforms that integrate guide-RNA design, on-target efficiency prediction, and off-target prediction within a single interface would substantially lower the barrier to adoption by non-computational biologists, who currently must combine multiple, methodologically inconsistent tools. Finally, as these tools increasingly inform pre-clinical and clinical trial design for CRISPR-based therapeutics under FDA and EMA review, the biosafety and regulatory-science communities should be engaged directly in defining what constitutes adequate validation evidence for a computational off-target predictor prior to its use in supporting a clinical safety case.

Other Information

Author Contributions

Conceptualization, A.S. and K.U.; methodology, A.S.; investigation, A.S. and K.U.; data curation, A.S.; writing— original draft preparation, A.S.; writing—review and editing, A.S. and K.U.;

visualization, A.S.; supervision, K.U. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Institutional Review Board Statement
Not applicable. This study is a literature review and did not involve human or animal subjects. The review protocol was not prospectively registered (e.g., with PROSPERO).

Data Availability Statement
No new primary data were generated in this study. All tools, datasets, and publications discussed are cited throughout the text and listed in the References section; software repositories, where available, are listed in
Table 8.

Conflicts of Interest
The authors declare no conflicts of interest.

A. Appendix A: Detailed PRISMA Methods

A.1. A.1 Eligibility Criteria

See Section 2.2 for full inclusion/exclusion criteria. In brief: peer-reviewed (or notable preprint) studies describing a deep-learning or hybrid ML/DL computational tool for CRISPR-Cas9 off-target prediction, published 2020 through the final search date, in English.

A.2. A.2 Information Sources

PubMed, Google Scholar, IEEE Xplore, and bioRxiv were used as primary information sources, supplemented by backward/forward citation tracking from key tool papers and from a broader 2023 review of CRISPR-Cas9 on- and off-target prediction methods [15].

A.3. A.3 Search Strategy

Representative Boolean query strings used across sources included:
("CRISPR" AND "off-target" AND "deep learning")
("CRISPR-Cas9" AND "off-target prediction" AND ("neural network" OR "CNN" OR "RNN" OR "LSTM" OR
"transformer" OR "attention"))
("gRNA" OR "sgRNA") AND "specificity" AND "deep learning"
"off-target prediction" AND "bioinformatics tool" AND ("machine learning" OR "deep learning")
"CRISPR" AND "off-target" AND ("transformer" OR "language model" OR "attention mechanism")

A.4. A.4 Selection Process

Records were screened by title and abstract against the eligibility criteria in Section 2.2, followed by full-text review of records passing initial screening. Selection decisions and reasons for exclusion were logged as summarized in the PRISMA flow diagram (Figure 1).

A.5. A.5 Data Collection Process

For each included study, the following data items were extracted manually: tool/model name; publication year and venue; architecture type; training and validation dataset(s); reported performance metric(s); code/data availability; and stated limitations, as reported in Section 3.2 and Tables 3–8.

A.6. A.6 Data Items

Extracted data items comprised: (1) tool name and aliases; (2) architecture family (CNN, RNN/LSTM/GRU, attention/transformer, hybrid/ensemble, language-model, or simulation-informed); (3) input encoding scheme; (4) benchmark dataset(s); (5) reported metric(s) (AUROC, AUPRC, F1, Spearman correlation, or uncertainty-based measures); (6) presence/absence of independent or cross-technology validation; (7) code/model availability; and (8) disclosed funding and conflicts of interest.

A.7. A.7 Study Risk of Bias Assessment

Risk of bias was assessed qualitatively per study (Section 3.3), considering relevance, publication venue, recency, clarity of validation methodology, disclosed conflicts of interest, and, specific to this sub-field, whether comparator baselines were evaluated independently of the tool's developers.

A.8. A.8 Effect Measures

No formal effect measures (e.g., pooled odds ratios or standardized mean differences) were used, as this review does not report a meta-

analysis; qualitative synthesis of reported AUROC/AUPRC/F1 values was performed instead (Section 3.4, Table 9).

A.9. A.9 Synthesis Methods

Findings were synthesized narratively by architecture category (Tables 3–6) and by benchmark dataset and software availability (Tables 7–8), following the qualitative-synthesis approach of the model paper [14]. No statistical pooling (e.g., meta-analytic model) was performed.

A.10. A.10 Reporting Bias Assessment

Reporting bias was assessed qualitatively (Section 3.5.5) by examining whether comparator baselines were independently sourced, whether negative results were reported, and whether the study's authorship overlapped with authorship of the comparator tools being benchmarked against.

A.11. A.11 Certainty Assessment

Certainty of evidence was assessed qualitatively (Section 3.5.6) as low to moderate, based on dataset overlap/nonindependence across studies, rapid architectural turnover, and the scarcity of prospective or cross-technology external validation.

Key Contribution: To our knowledge, the first systematic review focused specifically on deep-learning-based off-target prediction tools for CRISPR-Cas9, rather than on CRISPR bioinformatics tools in general. By scoping the search to January 2020–2026, the period in which neural architectures overtook classical alignment- and scoring-based methods, we map the transition from CNN- and RNN-based models to attention, transformer, and pretrained-language-model architectures, and we document a persistent lack of benchmarking standardization, independent validation, and interpretability across the field. We conclude with concrete recommendations for standardized public benchmark suites, explainable-AI reporting, and integrated design–prediction platforms to support clinical and regulatory adoption.

REFERENCES

Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 2012, 337, 816–821.

Doudna, J.A.; Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. Science 2014, 346, 1258096.

Hsu, P.D.; Scott, D.A.; Weinstein, J.A.; Ran, F.A.; Konermann, S.; Agarwala, V.; Li, Y.; Fine, E.J.; Wu, X.; Shalem, O.; et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 2013, 31, 827–832.

Bae, S.; Park, J.; Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 2014, 30, 1473–1475.

Stemmer, M.; Thumberger, T.; Keyer, M.d.S.; Wittbrodt, J.; Mateo, J.L. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. PLoS ONE 2015, 10, e0124633.

Doench, J.G.; Fusi, N.; Sullender, M.; Hegde, M.; Vaimberg, E.W.; Donovan, K.F.; Smith, I.; Tothova, Z.; Wilen, C.; Orchard, R.; et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 2016, 34, 184–191.

Tsai, S.Q.; Zheng, Z.; Nguyen, N.T.; Liebers, M.; Topkar, V.V.; Thapar, V.; Wyvekens, N.; Khayter, C.; Iafrate, A.J.; Le, L.P.; et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat. Biotechnol. 2015, 33, 187–197.

Tsai, S.Q.; Nguyen, N.T.; Malagon-Lopez, J.; Topkar, V.V.; Aryee, M.J.; Joung, J.K. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat. Methods 2017, 14, 607–614.

IEER
Institute for Excellence in Education & Research

Wienert, B.; Wyman, S.K.; Richardson, C.D.; Yeh, C.D.; Akcakaya, P.; Porritt, M.J.; Morlock, M.; Vu, J.T.; Kazane, K.R.; Watry, H.L.; et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-seq. Science 2019, 364, 286–289.

Abadi, S.; Yan, W.X.; Amar, D.; Mayrose, I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Comput. Biol. 2017, 13, e1005807.

Listgarten, J.; Weinstein, M.; Kleinstiver, B.P.; Sousa, A.A.; Joung, J.K.; Crawford, J.; Gao, K.; Hoang, L.; Elibol, M.; Doench, J.G.; et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat. Biomed. Eng. 2018, 2, 38–47.

Chuai, G.; Ma, H.; Yan, J.; Chen, M.; Hong, N.; Xue, D.; Zhou, C.; Zhu, C.; Chen, K.; Duan, B.; et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol. 2018, 19, 80.

Lin, J.; Wong, K.-C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics 2018, 34, i656–i663.

Jasieniecka, A.; Domingues, I. CRISPR-Cas9 and Its Bioinformatics Tools: A Systematic Review. Curr. Issues Mol. Biol. 2025, 47, 307.

Sherkatghanad, Z.; Abdar, M.; Charlier, J.; Makarenkov, V. Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review. Brief. Bioinform. 2023, 24, bbad131.

Liu, Q.; Cheng, X.; Liu, G.; Li, B.; Liu, X. Deep learning improves the ability of sgRNA off-target propensity prediction. BMC Bioinform. 2020, 21, 51.

Zhang, G.; Long, Y.; Yin, R.; Kwoh, C.K. DL-CRISPR: a deep learning method for off-target activity prediction in CRISPR/Cas9 with data augmentation. IEEE Access 2020, 8, 76610–76617.

Lin, J.; Zhang, Z.; Zhang, S.; Chen, J.; Wong, K.-C. CRISPR-Net: a recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels. Adv. Sci. 2020, 7, 1903562.

Niu, R.; Peng, J.; Zhang, Z.; Shang, X. R-CRISPR: a deep learning network to predict off-target activities with mismatch, insertion and deletion in CRISPR-Cas9 system. Genes 2021, 12, 1878.

Zhang, Z.-R.; Jiang, Z.-R. Effective use of sequence information to predict CRISPR-Cas9 off-target. Comput. Struct. Biotechnol. J. 2022, 20, 650–661.

Guan, Z.; Jiang, Z. Transformer-based anti-noise models for CRISPR-Cas9 off-target activities prediction. Brief. Bioinform. 2023, 24, bbad127.

Chen, Q.; Chuai, G.; Zhang, H.; Tang, J.; Duan, L.; Guan, H.; Li, W.; Li, W.; Wen, J.; Zuo, E.; Zhang, Q.; Liu, Q. Genomewide CRISPR off-target prediction and optimization using RNA-DNA interaction fingerprints. Nat. Commun. 2023, 14, 7521.

Yang, Y.; Li, J.; Zou, Q.; Ruan, Y.; Feng, H. Prediction of CRISPR-Cas9 off-target activities with mismatches and indels based on hybrid neural network. Comput. Struct. Biotechnol. J. 2023, 21, 5039–5048.

Toufikuzzaman, M.; Samee, M.A.H.; Rahman, M.S. CRISPR-DIPOFF: an interpretable deep learning approach for CRISPR Cas-9 off-target prediction. Brief. Bioinform. 2024, 25, bbad530.

Sun, J.; Guo, J.; Liu, J. CRISPR-M: predicting sgRNA off-target effect using a multi-view deep learning network. PLoS Comput. Biol. 2024, 20, e1011972.

Zhang, G.; Luo, Y.; Xie, H.; Dai, Z. Crispr-SGRU: prediction of CRISPR/Cas9 off-target activities with mismatches and indels using stacked BiGRU. Int. J. Mol. Sci. 2024, 25, 10945.

IEER
Institute for Excellence in Education & Research

Sari, O.; Liu, Z.; Pan, Y.; Shao, X. Predicting CRISPR-Cas9 off-target effects in human primary cells using bidirectional LSTM with BERT embedding. Bioinform. Adv. 2024, 5, vbae184.

Yang, Y.; Zheng, Y.; Zou, Q.; Li, J.; Feng, H. Overcoming CRISPR-Cas9 off-target prediction hurdles: a novel approach with ESB rebalancing strategy and CRISPR-MCA model. PLoS Comput. Biol. 2024, 20, e1012340.

Du, W.; Zhao, L.; Diao, K.; Zheng, Y.; Yang, Q.; Zhu, Z.; Zhu, X.; Tang, D. A versatile CRISPR/Cas9 system off-target prediction tool using language model. Commun. Biol. 2025, 8, 882.

Charlier, J.; Sherkatghanad, Z.; Makarenkov, V. Similarity-based transfer learning with deep learning networks for accurate CRISPR-Cas9 off-target prediction. PLoS Comput. Biol. 2025, 21, e1013606.

Özden, F.; Minary, P. Learning to quantify uncertainty in off-target activity for CRISPR guide RNAs. Nucleic Acids Res. 2024, 52, e87.

Sayed, A.; Ullah, K.; Khan, M.; Alakhtar, R.; Alsobhi, H. EHNN: enhanced hybrid neural network for CRISPR-Cas9 gRNA off-target prediction. PeerJ Comput. Sci. 2026, 12, e3881. doi:10.7717/peerj-cs.3881.

IEER
Institute for Excellence in Education & Research