OXFORD

Briefings in Bioinformatics, 2023, **24(3)**, 1–25
https://doi.org/10.1093/bib/bbad131
Advance access publication date 20 April 2023
Problem Solving Protocol

Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review

Zeinab Sherkatghanad, Moloud Abdar, Jeremy Charlier and Vladimir Makarenkov

Corresponding author: Vladimir Makarenkov, Departement of Computer Science, Université du Québec à Montréal, Montreal, QC, Canada.
E-mail: makarenkov.vladimir@uqam.ca

Abstract

CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein 9) is a popular and effective two-component technology used for targeted genetic manipulation. It is currently the most versatile and accurate method of gene and genome editing, which benefits from a large variety of practical applications. For example, in biomedicine, it has been used in research related to cancer, virus infections, pathogen detection, and genetic diseases. Current CRISPR/Cas9 research is based on data-driven models for on- and off-target prediction as a cleavage may occur at non-target sequence locations. Nowadays, conventional machine learning and deep learning methods are applied on a regular basis to accurately predict on-target knockout efficacy and off-target profile of given single-guide RNAs (sgRNAs). In this paper, we present an overview and a comparative analysis of traditional machine learning and deep learning models used in CRISPR/Cas9. We highlight the key research challenges and directions associated with target activity prediction. We discuss recent advances in the sgRNA–DNA sequence encoding used in state-of-the-art on- and off-target prediction models. Furthermore, we present the most popular deep learning neural network architectures used in CRISPR/Cas9 prediction models. Finally, we summarize the existing challenges and discuss possible future investigations in the field of on- and off-target prediction. Our paper provides valuable support for academic and industrial researchers interested in the application of machine learning methods in the field of CRISPR/Cas9 genome editing.

Keywords: CRISPR-Cas9, Genome Editing, Machine Learning, Deep Learning, On-Targets, Off-Targets

INTRODUCTION

Advances in the area of genome editing (also called gene editing) in the 2010s revolutionized molecular biology, genetics, and biomedicine. Genome editing techniques allow precise manipulation, deletion, and insertion of sequence fragments within the DNA of living organisms. In recent years, three effective types of genome editing toolsets, called Zinc Finger Nucleases (ZFNs) [1], Transcription Activator-Like Effector Nucleases (TALENs), and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), have been developed to study the process of target modifications in gene sequences [1–5].

Highly effective CRISPR/Cas9 gene editing system, co-invented in 2012 by Emmanuelle Charpentier and Jennifer Doudna [6], has been used in various fields, ranging from basic research on genetic therapies at the cellular level to applied biomedical research [6–11]. CRISPR/Cas9 demonstrated important clinical potential for treating human diseases such as cancer and genetic disorders [12–14], for plant genetic engineering [15–17], as well as for animal disease treatment [18, 19]. Figure 1 presents a schematic view of the CRISPR/Cas9 gene editing system and its practical applications.

The CRISPR/Cas9 genetic engineering system is an adapted version of the bacterial CRISPR-Cas9 antiviral defense system. CRISPR is a family of DNA sequences present in prokaryotic genomes that stems from DNA fragments of bacteriophages, which had infected prokaryotic genomes in the past. These DNAs are used as antiviral defense elements to recognize and eliminate DNA from similar bacteriophages during eventual infections [20]. Cas9 is a type of nuclease enzyme that uses CRISPR sequences as a guide to identify and cleave specific DNA fragments that are complementary to a given CRISPR sequence.

Zeinab Sherkatghanad is a PhD student at the Department of Computer Science of Université du Québec à Montréal (Montreal, Canada). Her research interests are in the fields of data mining, deep learning and computer vision.
Moloud Abdar is an Associate Research Fellow with the Institute for Intelligent Systems Research and Innovation at Deakin University, Australia. His research interests include machine learning, deep learning, computer vision, uncertainty quantification, data mining and medical image analysis.
Jeremy Charlier is a Research Fellow at the Department of Computer Science of Université du Québec à Montréal (Montreal, Canada). His research interests include the application of deep learning methods in the fields of bioinformatics and finance.
Vladimir Makarenkov is a Full Professor and Director of a graduate Bioinformatics program at the Department of Computer Science of Université du Québec à Montréal (Montreal, Canada). His research interests are in the fields of bioinformatics, artificial intelligence, and data mining. They include design and development of new bioinformatics algorithms and software, such as practical tools for analysis of CRISPR-Cas9 and high-throughput screening (HTS) data.
Received: November 2, 2022. **Revised:** March 7, 2023. **Accepted:** March 13, 2023

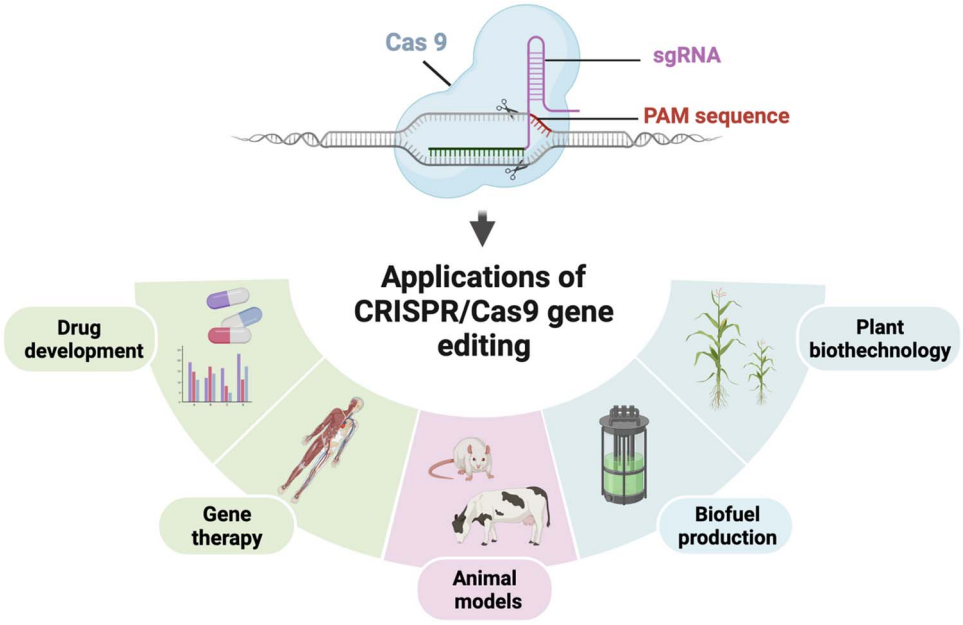

Figure 1. Schematic view of CRISPR/Cas9 gene editing system and its practical applications.

In the CRISPR/Cas9 editing system, the Cas9 nuclease combined with a guide RNA (gRNA) is delivered into a cell, allowing the cell's genome to be cut in a specific location, some targeted genes to be removed from it, and some other added to it, *in vivo* [21]. Guide RNA, or artificially programmed single-guide RNA (sgRNA) used in type II CRISPR/Cas9 systems, is responsible for identifying the target DNA sequence in the cell's genome and ensuring that the cutting takes place at the desired sequence location. The Protospacer-Adjacent Motif (PAM), located at the end of the DNA target site, is a short three-to-five nucleobase sequence serving as the binding signal of the Cas protein [22].

The CRISPR/Cas9 system is nonetheless prone to unintended off-targets: a cleavage may occur at non-target locations [23–25]. Thus, the safety aspect of the use of CRISPR/Cas9 on humans remains an open issue. The main challenge in the effective application of the CRISPR/Cas9 system is to maximize on-target activity (i.e. guide efficiency) and minimize the number of potential off-targets (i.e. guide specificity).

Over the last few years, data-driven machine learning methods emerged as a new modeling approach which outperforms the common scoring prediction methods, such as MIT CRISPR Design Tool2 [26], CCTop algorithm [27], CRISPR Design [28], E-CRISP [29], and CHOPCHOP [30]. One of the main drawbacks of the latter methods is the lack of capacity to increase the prediction accuracy when the number of samples increases. In contrast, one of the main advantages of modern data-driven models relying on deep learning is their ability to improve the predictive performance as the number of samples grows. Current state-of-the-art research aiming at designing robust clinical CRISPR/Cas9 applications looks for enhancing data-driven models by: (i) increasing on-target efficiency, (ii) improving off-target specificity, and (iii) simultaneously maximizing on-target activity and minimizing off-target effects.

Among recent reviews and comparative studies discussing the use of computational methods for evaluating gRNA efficiency and predicting gRNA specificity, we need to mention the following works. Wang *et al.* [31] highlighted new advances in CRISPR-Cas systems in terms of RNA targeting, tracking, and editing. The authors compared Cas protein-based technologies with traditional technologies intended for these goals. Liu *et al.* [32] provided an up-to-date overview of computational methods for gRNA design, including web-based platforms, to help researchers select optimal tools for their CRISPR-Cas experiments. However, among 14 methods for evaluating gRNA efficiency and 13 methods for predicting gRNA specificity considered by Liu *et al.*, only one employs deep learning. Chen *et al.* [33] first briefly reviewed the main properties of CRISPR systems and their use in genome editing. Then, the authors discussed feasible methods for detecting potential off-targets during CRISPR/Cas9 genetic manipulations. Yan *et al.* [34] compared 17 *in silico* off-target prediction tools in order to evaluate their genome-wide CRISPR performances, and introduced an integrated Genome-Wide Off-target cleavage Search (iGWOS) platform designed for optimal genome-wide off-target predictions. The main goal of the study by Yaish *et al.* [35] was to systematically evaluate data pre-processing and formulation of the CRISPR off-target prediction problem. The authors pointed out that data transformation is a crucial data pre-processing step which should be applied prior to model training. They highlighted the importance of considering as model's features both inactive off-target sites and the number of mismatches between gRNAs and their off-target sites. Moreover, Yaish *et al.* introduced predictive off-target *in cellula* models based on gradient boosting (i.e. the XGBoost decision tree-based ensemble learning framework was implemented by these authors) and compared them with state-of-the-art off-target prediction methods. The paper of O'Brien *et al.* [36] presents the main machine learning approaches and pitfalls in the context of CRISPR-Cas9 experiments. The authors consider the related computational problems, including algorithm choice, accuracy overestimation, and data interoperability. They concluded that due to the broad availability of machine learning-based tools, *in silico* optimization can successfully replace *in vitro* CRISPR-Cas9 designs. Thus, algorithmic solutions can be used for maximizing gRNA editing efficiency and minimizing gRNA specificity. Finally, a recent review of Konstantakos *et al.* [37] addresses the problem of CRISPR–Cas9 gRNA efficiency prediction and evaluates the role of deep learning

in this context. The authors discussed the main computational approaches for on-target activity prediction, focusing on the selection of optimal features and algorithms. In their comparative experiments, Konstantakos *et al*. assessed the performances of 10 deep learning and one conventional machine learning methods on six benchmark data sets, and provided recommendations for their use. The authors pointed out that the existing on-target prediction approaches still have some flaws, including their sensitivity to data heterogeneity, unclear decision making mechanism, and inability to produce general gRNA design rules.

In this review paper, we provide a summary of studies that have examined the effectiveness of artificial intelligence (AI) methods for on- and off-target activity prediction related to CRISPR/Cas9. In contrast to some previous reviews [32, 34, 37–42], our study discusses the use of both traditional machine learning and deep learning methods, focusing on the latest state-of-the-art on- and off-target prediction models. We describe the main advantages and disadvantage of existing prediction models, while highlighting noticeable progress that has been made in sequence encoding.

Our main contributions are as follows:

- To the best of our knowledge, this is the first comprehensive review of both traditional machine learning and deep learning methods, used for both on-target (guide efficiency) and off-target (guide specificity) outcome prediction in CRISPR/-Cas9 genome editing;
- A description of the benchmark data sets used for on- and off-target prediction in CRISPR/Cas9 is provided. Most of the discussed data sets were curated and made available for researchers on our GitHub repository;
- The main sequence encoding techniques used in CRISPR/-Cas9, applying to both traditional machine learning and deep learning methods, are discussed along with their intrinsic properties;
- The main deep learning models used for on- and off-target prediction in CRISPR/Cas9 are presented and their advantages and limitations are highlighted;
- Research challenges and avenues of future investigation regarding the application of traditional machine learning and deep learning methods in the field of CRISPR/Cas9 genome editing are discussed.

DATA DESCRIPTION

In this section, we present the most popular CRISPR/Cas9 benchmark data sets used in the literature for on- and off-target prediction. These data sets can be divided into three categories: data sets including off-targets only, on-targets only, and both off- and on-targets.

The first group of benchmark data sets consists of off-targets. GUIDE-seq was one of the first off-target data repositories, based on the results of the GUIDE-seq technique developed by Tsai *et al*. [43]. It can serve as an accurate framework for genome-wide identification of off-target effects. The sgRNAs used in GUIDE-seq target the following sites: VEGFA site 1, VEGFA site 2, VEGFA site 3, FANCF, HEK293 site 2, HEK293 site 3, and HEK293 site 4, in which 28 off-targets with a minimum modification frequency of 0.1 were identified (among 403 potential off-targets). The CIRCLE-Seq (Circularization for *in vitro* Reporting of CLeavage Effects by sequencing) screening strategy introduced by Tsai *et al*. [44] was used to analyze the related data set that includes gRNA–DNA pairs for 10 gRNA sequences with the corresponding mismatch, insertion, and deletion information; 7371 of these sequence pairs were identified as active off-targets. Cameron *et al*. [45] proposed the SITE-Seq biochemical method that uses Cas9 programmed with sgRNAs to recognize cut sites within genomic DNA. The related data set contains sgRNA–DNA sequence pairs for nine guide sequences; 3767 of these sequence pairs correspond to active off-targets. Abadi *et al*. [46] collected a training data set based on three genome-wide methods for unbiased CRISPR-Cas9 cleavage site profiling, which are as follows: (i) Genome-wide unbiased identification of DSBs enabled by sequencing (GUIDE-Seq) [43, 47], (ii) High-throughput genome-wide translocation sequencing (HTGTS) [48], and (iii) Breaks labeling, enrichment on streptavidin and next-generation sequencing (BLESS) [49, 50]. The resulting data set was assembled from the five following studies: [43, 47–50]. It includes 33 collections of sgRNAs with their respective targets. Altogether, these sgRNAs cleaved 872 genomic targets across human genome. Lazzarotto *et al*. [51] applied their CHANGE-seq automatable tagmentation-based method to analyze the related *in vitro* Cas9 genome-wide nuclease activity data set. CHANGE-seq was carried out to analyze 110 sgRNA targets across 13 therapeutically relevant loci in human primary T cells. A total of 201 934 off-target sites were identified with variable numbers of off-target sites, ranging from 19 to 61 415, for an individual sgRNA.

The second group of benchmark data sets consists of on-targets. Wang *et al*. [52] used a practical library containing 73 000 sgRNAs to generate knockout collections and to investigate screens in human cell lines HL-60 and KBM7. The authors tested both ribosomal and non-ribosomal protein coding genes with all possible sgRNAs gathered in the library. Koike-Yusa *et al*. [53] conducted their investigation on a data set that consisted of 87 897 gRNAs targeting 19 150 mouse protein-coding genes. They designed genome-wide mutant mouse embryonic stem cell libraries to identify unknown host factors that modulate toxin susceptibility. The Doench V1 data set [54] consists of 1831 guides targeting three human (CD13, CD15, and CD33) and six mouse (Cd5, Cd28, H2-K, Cd45, Thy1, and Cd43) genes, all producing cell-surface markers, which could be assayed by flow cytometry. GenomeCRISPR is a well-formatted data repository, organized by Rauscher *et al*. [55], which was designed for high-throughput CRISPR screening studies. GenomeCRISPR contains over 550 000 sgRNAs on-targets derived from 84 different experiments. Wang *et al*. [56] used the DeepHf (Deep learning for High-Fidelity Cas9) method to perform a genome-scale screen measuring gRNA activity of two highly specific SpCas9 variants (eSpCas9(1.1) and SpCas9-HF1), and a wild-type SpCas9 (WT-SpCas9) in human cells. The obtained data set contains indel rates for over 50 000 gRNAs for each nuclease, covering about 20 000 genes. It is the largest gRNA on-target activity data set reported to date for mammalian cells. Kim *et al*. [57] generated a data set of SpCas9 activities at 12 832 target sequences from a human cell library using the deep learning–based DeepSpCas9 model. The DeepSpCas9 target sequences were chosen from the human genome and synthetic sequences without using any information related to the activity of the associated sgRNAs. The sgDesigner data are a unique plasmid target library expressed in human cells that was used by Hiranniramol *et al*. [58] for experimental quantification of sgRNA CRISPR/Cas9 efficiency. A pool of 12 472 oligonucleotides was used to train a machine learning algorithm for assay design.

The third group of benchmark data sets considered in the literature includes data containing both off- and on-target information. First, we need to mention the RESistance assays (RES) data set, i.e. Doench V2, made available by Doench *et al*. [59]. It consists of 2549 unique guides targeting eight genes (i.e. CCDC101, MED12, TADA2B, TADA1, HPRT, CUL3, NF1, and NF2) from Human A375

cells. The well-known CRISPOR database organized and maintained by Haeussler *et al.* [60] aggregates different public data sets that have been widely used to quantify on-target guide efficiency and detect off-target cleavage sites, including the Wang-Xu *et al.* data set (2076 guides targeting 221 genes in Human HL-60 cells) [52, 61], Koike-Yusa *et al.* data set [53], Doench V1 and V2 data sets [54, 59], Hart *et al.* data set (4239 guides targeting 829 genes in Human Hct116 cells) [62], Z_fish MM data set (1020 guides targeting 128 genes in Zebrafish genome) [63], Z_fish VZ data set (102 guides targeting different genes in Zebrafish genome) [64], Z_fish GZ data set (111 guides targeting different genes in Zebrafish genome) [65], Drosophila data set [66], Chari *et al.* data set (1234 guides targeting Human 293T cells) [67], Ciona data set (72 guides targeting different genes in Ciona genome) [68], and Farboud *et al.* data set (50 guides targeting different genes in Caenorhabditis elegans genome) [69]. Moreover, Haeussler *et al.* [60] developed the CRISPOR web tool (available at: crispor.org) that is intended to design, evaluate and clone guide sequences for the CRISPR/Cas9 system. This web tool incorporates several on- and off-target scoring algorithms. It also displays pre-calculated results for all human exons from the UCSC Genome Browser tracks. On the first page of crispor.org, the user enters three pieces of information: (i) a single genomic sequence, typically an exon under 2300 bp; (ii) a genome (from the list of more than 150 genomes, including plants and emerging model organisms); (iii) a PAM motif. The main output of CRISPOR is a web page that shows the annotated input sequence and the list of possible guides in this sequence. Furthermore, CRISPOR also generates a list of primers related to a selected guide. The related GitHub data repository organized by Haeussler comprises direct links to 22 experimental data sets along with some necessary data conversion scripts written in Python. Munoz *et al.* [70] designed the CRISPR tiling library, which is a large tiling-sgRNA data set containing data for 139 genes with an average of 364 sgRNAs/gene for three cancer cell lines DLD1, RKO, and NCI-H1299. Furthermore, the DeepCRISPR data set generated by Chuai *et al.* [71] includes approximately 0.68 billion sgRNA sequences derived from 13 human cell lines, including HEK293, MCF-7, K562, HL60, NB4, BE2C, Caco-2, GM06990, Hela, HCT116, LNCap, HepG2, and GM12878. This large data set comprises epigenetic information for different cell types, providing a unified feature space which combines the data from various experiments and cell types. The related DeepCRISPR software includes the models for both sgRNA on-target knockout efficacy and genome-wide off-target cleavage profile prediction.

In Table 1, we present a summary of the main features of the CRISPR/Cas9 benchmark data sets used for on- and off-target prediction, including the original study describing the data set in question, the URL link to the data set and the target type. We curated the benchmark data reported in Table 1 and made them available for researchers on our GitHub repository at the following URL address: https://github.com/dagrate/public_data_crisprCas9. Moreover, several data sets presented here have been one-hot encoded and prepared for use in machine learning and deep learning experiments. The related Python scripts have been also made available to the scientific community.

In conclusion, we think that using the latest benchmark data containing large amounts of samples and features is likely to facilitate future work in CRISPR/Cas9, since such data can provide wide and complete coverage of intrinsic properties of both off- and on-targets under study, and thus be effectively exploited by state-of-the-art machine learning and deep learning methods.

sgRNA–DNA SEQUENCE ENCODING

Before building AI models intended for on- and off-target prediction, the sgRNA–DNA sequence data must be pre-processed to be used as input. Data pre-processing, or data encoding, allows converting the sgRNA–DNA sequences of letters into sequences of numbers that AI models can read and interpret to build their predictions. Data pre-processing is an important milestone when trying to boost the predictive performance of AI models. The two most popular encoding techniques used in CRISPR-Cas9 are: (a) One-hot encoding and (b) Word embedding. Figure 2 highlights the differences between the two techniques. In one-hot encoding, each possible channel A, C, G or T is represented by a one-hot vector such as [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1]. In embedding, a particular word, or string, is represented using a unique vector representation. A sgRNA–DNA sequence, which can be subdivided into substrings of length k, called k-mers, can be thus transformed into a vector representation. The most popular embedding technique is Word2Vec [72]. This natural language processing technique relies on the use of neural networks. In this review, we first discuss some recent papers that use one-hot encoding schemes in CRISPR-Cas9, followed by a brief overview of papers dealing with word embedding, and by the section presenting further sequence characteristics often used as explanatory features in ML and DL models.

It is worth noting that most of the sequence encoding schemes discussed in this section apply to both traditional machine learning and deep learning models. For example, the recent works of Lin *et al.*, Wang *et al.*, and Charlier *et al.* [56, 73–75] use the same sequence encoding schemes to provide the input of machine learning and deep learning models compared in these papers.

In this paragraph, we present recent studies using novel one-hot encoding techniques in the field of genome editing. When applying a one-hot encoding, each sgRNA–DNA sequence pair of length L can be encoded in a one-hot matrix of four rows and L columns. Each row corresponds to the nucleotide type, i.e. A, C, G, and T (or U). Each base in the sgRNA and the target DNA is then encoded in the form of a one-hot vector according to a particular method. For example, Chuai *et al.* [71] proposed the deepCRISPR model for sgRNA on- and off-target prediction. DeepCRISPR relies on a deep convolutionary denoising neural network and one-hot data pre-processing. The nucleotide sequence is a 20-bp sgRNA sequence with an NGG PAM across the human genome. It is represented by four channels, (A, C, G, and T), and each epigenetic feature is considered as one channel. Thus, the encoded matrix used by Chuai *et al.* is of size $(4+n)\times 23$, where 4 corresponds to the number of channels and n to the number of epigenetic features. Lin *et al.* [73] introduced a one-hot sequence encoding method that converts each sgRNA–DNA sequence pair into a matrix to be used as a convolutional input. In their encoding the four channels are used to represent both sgRNA and target DNA. Thus, each character in the sgRNA and target DNA sequences is represented by a single one-hot vector. Consequently, every sgRNA–DNA sequence pair is encoded in a matrix of size 4×23, where 23 corresponds to the 3-bp PAM adjacent to the 20 bases. The use of such 4×23 input matrices allowed the authors to apply for the first time deep Feedforward Neural Networks (FNNs) and deep Convolutional Neural Networks (CNNs) for off-target prediction in CRISPR-Cas9 gene editing. Charlier *et al.* [75] described a different novel one-hot encoding method. Their main idea was to build a data encoding procedure that relies on a bijective mapping for sgRNA–DNA sequence pairs. It allows for encoding, and decoding, of the sgRNA–DNA sequence pairs without any information loss that

Table 1. A summary of the most popular CRISPR/Cas9 benchmark data sets and databases used for on- and off-target prediction.

Source	Year	Data description	Target	Data link
Wang *et al.* data [52]	2014	A library containing 73 000 sgRNAs	On-targets	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3972032/#SD2
Koike-Yusa *et al.* data [53]	2014	87,897 gRNAs targeting 19 150 mouse protein-coding genes	On-targets	Deposited at the European Nucleotide Archive under accession number ERP003292.
Doench V1 data [54]	2014	1831 guides targeting three human (CD13, CD15 and CD33) and six mouse genes (Cd5, Cd28, H2-K, Cd45, Thy1 and Cd43)	On-targets	broadinstitute.org/rnai/public/analysis-tools/sgrna-design
GUIDE-seq data [43]	2015	CRISPR RNA-guided nucleases (RGNs) from two human cell lines: U2OS and HEK293; different sites such as VEGFA sites 1, 2 and 3, and HEK293 sites 2, 3 and 4 were studied	Off-targets	https://github.com/tsailabSJ/guideseq
Doench V2 data [59]	2016	2549 unique guides targeting eight genes (CCDC101, MED12, TADA2B, TADA1, HPRT, CUL3, NF1 and NF2) from human A375 cells	Off-targets and on-targets	https://www.nature.com/articles/nbt.3437
CRISPOR program + data repository [60]	2016	Aggregate data for more than 150 genomes, including the following public data sets: Wang-Xu [52, 61], Koike-Yusa [53], Doench V1 and V2 [54, 59], Hart [62], Z_fish MM [63], Z_fish VZ [64], Z_fish GZ [65], osophila [66], Chari [67], Dr Ciona [68], Farboud [69]	Off-targets and on-targets	http://crispor.org + https://github.com/maximilianh/crisporPaper/tree/master/effData#readme
GenomeCRISPR database [55]	2016	Aggregate data for more than 550 000 sgRNAs derived from 84 experiments	On-targets	http://genomecrispr.org
CIRCLE-Seq data [44]	2017	Contains mismatch, insertion and deletion information, and includes sgRNA–DNA pairs from 10 guide sequences, 7371 of which are off-targets (430 with bulges)	Off-targets	https://github.com/tsailabSJ/circleseq
SITE-Seq data [45]	2017	gRNA–DNA pairs from nine guide sequences, 3767 of which are active off-targets (no bulges)	Off-targets	https://experiments.springernature.com/articles/10.1038/nmeth.4284
Abadi *et al.* [46]	2017	A data set based on three genome-wide methods for unbiased CRISPR-Cas9 cleavage sites profiling: GUIDE-Seq, HTGTS and BLESS. It includes 33 collections of sgRNAs with their respective targets	Off-targets	https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005807
DeepCRISPR [71] platform	2018	Includes approximately 0.68 billion sgRNA sequences derived from 13 human cell lines	Off-targets and on-targets	https://github.com/bm2-lab/DeepCRISPR
DeepHf data [56]	2019	Includes indel rates of over 50 000 gRNAs for each nuclease, covering about 20 000 genes. It is the largest gRNA on-target activity set reported for mammalian cells	On-targets	http://www.DeepHF.com
DeepSpCas9 data [57]	2019	A dataset of SpCas9 activities at 12 832 integrated target sequences for a human cell library	On-targets	http://deepcrispr.info/DeepSpCas9
sgDesigner data [58]	2020	A unique plasmid library expressed in human cells was used to quantify the potency of thousands of CRISPR/Cas9 sgRNAs (a pool of 12 472 oligonucleotides was analyzed)	On-targets	https://academic.oup.com/bioinformatics/article/36/9/2684/5714741?login=false#supplementary-data
CHANGE-seq data [51]	2020	110 sgRNA targets across 13 therapeutically relevant loci in human primary T-cells were studied to identify 201 934 off-target sites across the human genome	Off-targets	https://github.com/tsailabSJ/changeseq

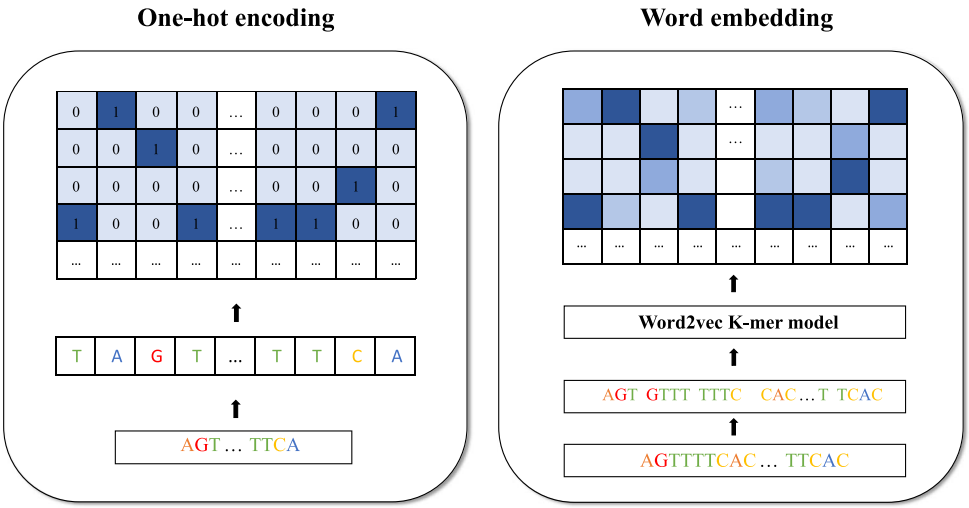

Figure 2. Two sequence encoding models used in CRISPR/Cas9: one-hot encoding and word embedding.

can occur in the encoding scheme adopted in [73]. Specifically, Charlier *et al.* combined a 4 × 23 matrix used for sgRNA encoding and a 4 × 23 matrix used for DNA encoding, resulting in a 8 × 23 matrix used as a convolutional input. The authors applied FNNs, CNNs, and Recurrent Neural Networks (RNNs) to generate accurate off-target predictions. Lin *et al.* [74] have recently introduced an encoding technique capable of incorporating in the input data base mismatch, missing base (RNA bulge or insertion), and extra-base (DNA bulge or deletion) information. Each sequence pair was considered as a fixed length vector with the following five-bit channel: (A, C, G, T, _). Additionally, the authors introduced a two-bit direction channel that was used to identify the indel and mismatch directions. Thus, a combined seven-bit channel, encoded as seven one-hot encoded vectors, allowed them to take into account not only sgRNA–DNA sequence mismatches, but insertions and deletions as well. Precisely, Lin *et al.* [74] used a 7 × 23 matrix encoding scheme (see Figure 3), where 23 is the length of the sgRNA–DNA sequence pairs. This encoding scheme is a perfect example of feature engineering, i.e. new feature construction process that is explicitly defined and manually or automatically applied. Feature engineering is common in machine learning. Moreover, some machine learning methods, e.g. Support Vector Machine (SVMs), incorporate feature engineering as part of their operation [76]. In the case of the sgRNA–DNA sequence encoding proposed by Lin *et al.* the created seven-bit-long features allow one to take into account all possible correspondences existing between the original sgRNA–DNA pairs of features taking the values (A, C, G, T, _). This innovative encoding scheme was used with different deep learning models for off-target prediction on CIRCLE-Seq and GUIDE-Seq data sets, and demonstrated state-of-the-art prediction performance. Zhang *et al.* [77] designed an encoding scheme consisting of a matrix of size 20 × L, with L being the sequence length. In the encoding process, the authors used a four-bit channel (A, C, G, T) for sgRNA encoding, a four-bit channel (A, C, G, T) for DNA encoding, and a 12-bit channel to one-hot encode all possible mismatches. They regrouped the three corresponding matrices, resulting in a final matrix of size $20 \times L$. This extended matrix was then used for data augmentation to reduce the class imbalance between off-targets and on-targets, while a CNN model was used for on-target activity prediction. Zhang *et al.* [78] proposed another encoding scheme with a similar objective to incorporate mismatch, DNA, and RNA bulge information into different off-target prediction models. The authors first considered a four-bit channel (A, C, G, T) and a one-hot vector encoding scheme. Furthermore, they used a two-bit channel to indicate a base deletion on RNA and DNA, and another one-bit function channel to indicate if the location is part of the guide sequence (0) or the PAM sequence (1). The encoded matrix was thus of size 7 × 23. Finally, an 'OR' operation was carried out to indicate when two bases in a base pair were identical. Zhang *et al.* tested their encoding scheme with different FNN, CNN, and RNN models. They demonstrated performance on par with state-of-the-art. Overall, among the different one-hot encoding schemes found in the literature, the highest potential has been recently demonstrated by those relying on the use of indels.

The second common data pre-processing strategy used by several researchers is a Natural Language Processing (NLP) technique, called word embedding. The idea of applying word embedding on sgRNAs is that off-targets are encoded closer to each other in the vector space than are on-targets. Liu *et al.* [79] combined the word embedding with a transformer to convey sgRNA sequences to a deep neural network (DNN) model consisting of CNNs and FNNs. The authors demonstrated that the word embedding approach had similar predictive performance as the latest one-hot encoding-based deep learning models. Later on, Liu *et al.* [80] proposed to use a trained unsupervised learning algorithm, GloVe [81], designed to aggregate word-to-word occurrence statistics outputting linear substructures of the word vector space. The authors applied GloVe to convert sgRNA sequences into substructures of the word vector space. They forwarded the sgRNA word vectors to a bidirectional LSTM and a CNN with five convolutional layers to predict the sgRNA off-target propensity, and demonstrated state-of-the-art predictive performance of their models. Zhang *et al.* [82] proposed to label-encode and word-embed sgRNA sequences. Each sgRNA sequence was transformed into a numerical vector using the Tokenizer module from the Keras library [83]. The encoded sequences were then passed to pooling and convolutional layers, and to three convolutional layers to obtain sgRNA cleavage efficiency predictions. Using word-embedding techniques from NLP is fairly recent, but we are confident that it opens new perspectives for future work in the field of genome editing.

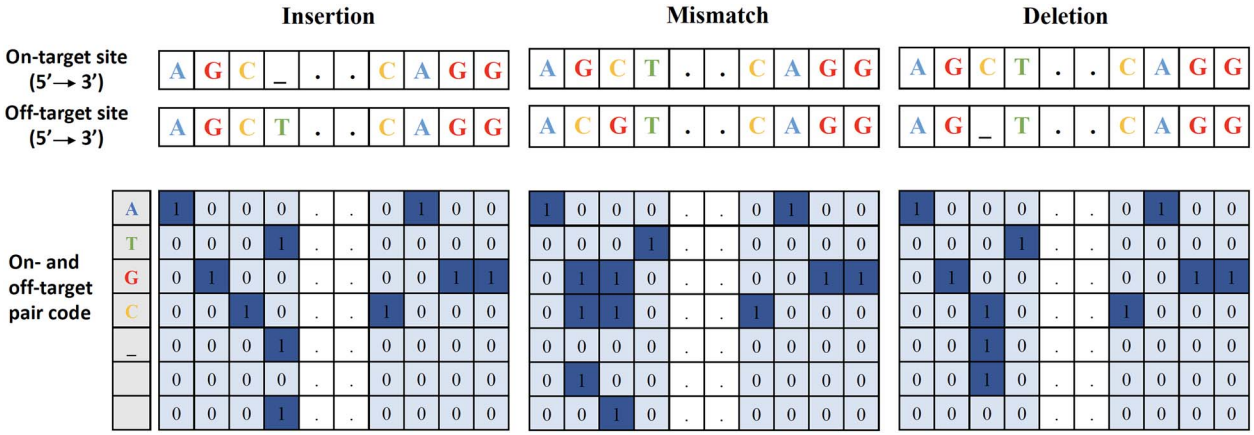

Figure 3. A novel effective sgRNA–DNA one-hot sequence encoding scheme used by Lin *et al.* [74]. A seven-bit encoding example is shown. Here, '_' symbol indicates the DNA or RNA bulge position. Each sgRNA–DNA sequence pair is encoded as a fixed-length seven-row matrix that includes a five-bit character channel (A, G, C, T, _) and a two-bit direction channel. The five-bit channel is used to encode the on- and off-target site nucleotides, whereas the direction channel is used to indicate the mismatch and indel locations.

One-hot encoding and word embedding are not the only ways to represent a sequence in machine learning models. The majority of conventional machine learning (ML) models used in CRISPR-Cas9 consider additional sequence features such as position-specific features, different counts, and structural/thermodynamic characteristics to best capture sequence information (see Table 2). In contrast to conventional machine learning models, deep learning (DL) models can automatically learn the sequence features by generating new internal features that are crucial for accurate outcome predictions (see, for example, the convolutional deep learning models used by Lin *et al.* [74] and Shrawgi *et al.* [84], and the discussion therein). Here, we briefly recall some important works in the field and highlight sequence features that helped to boost the performance of the related machine learning models. Doench *et al.* [54], Xu *et al.*[61], and Peng *et al.* [85] identified several sgRNA and target DNA features allowing one to improve the prediction results, including position-specific nucleotide composition for individual (order 1) nucleotides (A/C/T/G) and pairwise (order 2) nucleotides (AA/AT/AG/...) in 30mer sequences (i.e. the 20mer guide plus context on either side) and GC counts for each guide sequence. Afterwards, Doench *et al.* [59] used thermodynamic features and position-independent features in addition to previously considered position-specific features and the GC counts (Doench *et al.* [54]). Position-independent features included individual nucleotide counts (order 1) and adjacent pairwise nucleotide counts (order 2), ignoring their position in the sgRNA. Thermodynamic features were computed from the melting temperatures of the DNA version of the RNA guide sequence, or portions thereof, using the Biopython Tm_staluc function [59]. Rahman *et al.* [86] considered position-specific features, position-independent features, as well as sgRNAs secondary structures to create additional features for their ML models. The structural features used by Rahman *et al.*, which are also known as thermodynamic properties, included Minimum Free Energy, the most favorable thermodynamic RNA–RNA interaction energy, local pair probabilities and specific sgRNA heat parameters. In their CRISPRO experiments, Schoonenberg *et al.* [87] considered several categorical and numerical features. The categorical features used in this work include targeted amino acids 1 and 2, domain occupancy status (InterPro), exon multiple of 3, the ability of targeted transcript to escape nonsense-mediated decay, single nucleotide and dinucleotide positional identities within sgRNA spacer, and orientation of sgRNA relative to gene. All categorical features were one-hot encoded. Numerical features considered by Schoonenberg *et al.* include the PROVEAN (Protein Variation Effect Analyzer) deletion score of the targeted amino acids 1 and 2, position in the gene, predicted disorder score of amino acids 1 and 2, GC content of the 20-mer guide, length of the targeted exon, and off-target score of the gRNA. Chen *et al.* [88] used two categorical features, i.e. the mismatched bases on both the gRNA (called 'Ref allele') and donor DNA (called 'Alt allele'), and one numerical feature, i.e. mismatch position relative to the gRNA. These features were then converted into binary vectors using one-hot encoding and employed as input of several traditional ML models tested by the authors. Rafid *et al.* [89] proposed an accurate SVM-based tool using sequence-based features only. The authors experimented with three types of features, including position-independent features, position-specific features, and n-gapped dinucleotides (nGD). The nGD features count the number of times that two given nucleotides appear at a certain distance in a sgRNA sequence. Dhanjal *et al.* [90] explored 11 categories of sequence features that possibly govern the specificity of sgRNAs. The main finding of their work consists in the identification of the four most important sequence features, including accessibility of target sequence in the genome, mismatch count between the off-target and target sequences, position-specific occurrence of nucleotides in the spacer and regions flanking it on both sides, and, finally, GC count in the target and off-target sequences. Hiranniramol *et al.* [58] used sequence and structural features of the most and the least potent sgRNAs to train their sgDesigner model. Significance levels for numerical features were computed using Student's t-test, and for binary features using the χ^2 test. The authors considered only the high- and low-efficiency sgRNAs groups to emphasize the most predictive features affecting sgRNA efficiency. He *et al.* [91] demonstrated that sequence-specific sgRNA activity, frameshift probability and amino acid features could significantly improve the selection of efficient sgRNAs in protein knockouts. They highlighted the importance of amino acid sensitivity as one of the critical factors that govern the efficiency prediction, in addition to the use of effective sequence models that predict sgRNA activity.

In conclusion, we need to point out that feature comparison between different experiential CRISPR-Cas9 data sets uncover substantial discordance and that further research is warranted to

Table 2. A summary of studies applying traditional machine learning methods for on and off-target prediction in CRISPR/Cas9.

Study	Year	Target prediction	ML model(s) used	Encoding	Data	Link for data/ software	Prediction metric/ results
Wang *et al.* [52]	2014	On-target	SVM	One-hot + GC content	A library of 73 000 sgRNAs was used to generate knockout collections for two human cell lines	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3972032	log2 fold change estimations
Doench *et al.* [54]	2014	On-target	SVM and Logistic regression	One-hot + GC counts + position-specific features	1831 gRNAs targeting three human genes and six mouse genes were used to generate screening data; Doench V1	http://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design	AUROC: 0.8
Xu *et al.* [61]	2015	On-target	Logistic regression	One-hot + GC counts + position-specific features	Wang [52], Koike-Yusa [53], Shalem [105], Zhou [106] Gilbert [107] Konermann [108]	http://crispr.dfci.harvard.edu/SSC	AUROC: 0.73
Fusi *et al.* [100]	2015	On-target	SVM, L1 regression, L2 regression, RF regression, SVM+logistic regression, L1 logistic regression, linear regression, GBRT	One-hot + GC counts + position-specific features	Wang ribosomal, Wang non-ribosomal [52], Koike-Yusa [53], Doench V1 [54]	http://research.microsoft.com/enus/projects/azimuth	Spearman: 0.52 AUROC: 0.75
Doench *et al.* [59]	2016	Off-target and on-target	Boosted RT, L1 regression, L2 regression, SVM+logistic regression, RF, linear regression	One-hot + GC counts + position-specific features + position-independent features + thermodynamic features	2549 unique guides were used to generate Doench V2 data set	http://www.broadinstitute.org/rnai/public/analysistools/sgrna-design, http://research.microsoft.com/enus/projects/azimuth	Spearman: 0.54 for on-target AUROC: 0.8 for off-target
Rahman *et al.* [86]	2017	On-target	CRISPRpred (SVM, RF, linear regression)	One-hot + position-specific features + position-independent features + structural/ thermodynamic features	Doench V1 [54]	https://github.com/khaled-buet/CRISPRpred	AUROC: 0.85 AUPRC: 0.56 MCC: 0.4
Abadi *et al.* [46]	2017	Off-target	CRISTA (CRISPR Target Assessment using RF regression)	GC content + sgRNA secondary structure features	GUIDE-Seq [43, 47], HTGTS [48], BLESS [49, 50]	http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005807	Spearman: 0.81 AUROC: 0.96 AUPRC: 0.96 R^2 = 0.8.
Peng *et al.* [85]	2018	Off-target	Ensemble SVM	One-hot + GC content + position-specific features	CRISPOR [60]	https://github.com/penn-hui/OfftargetPredict, Cas-OFFinder [109]	AUROC: 0.99 AUPRC: 0.45
Schoonenberg *et al.* [87]	2018	On-target	CRISPRO (GBDT, Ridge, RF, Lasso, SVM)	One-hot + GC content + position-specific features	Doench V2 [59], Munoz [70], Donovan [110], Brenan [111]	http://gitlab.com/bauerlab/crispro	Spearman: 0.57

(*continued*)

Table 2. Continued.

Study	Year	Target prediction	ML model(s) used	Encoding	Data	Link for data/ software	Prediction metric/ results
Listgarten *et al.* [112]	2018	Off-target	Elevation (Boosted regression Trees, L1 regression, Naïve Bayes)	One-hot	GUIDE-Seq [43], Doench V2 [59], CRISPOR [60]	http://research.microsoft.com/en-us/projects/crispr	AUROC: 0.98
Chen *et al.* [88]	2019	Off-target	Logistic regression, SVM, RF, NN	One-hot + Ref allele + Alt allele + mismatch position in guide features	Unpublished data of Sharon *et al.* were used to generate CRISPEY (Cas9-Retron precISe Parallel Editing via homologY) data consisted of 23,936 samples (18,717 training set and 4,680 testing set)	https://github.com/elizapandabella/CRISPEY_ML_Public	Accuracy: 94%.
Zhang *et al.* [95]	2019	Off-target	Ensemble learning framework of scoring features using AdaBoost	One-hot + GC counts + position-specific features	GUIDE-Seq [43], CRISPOR [60]	https://github.com/Alexzsx/CRISPR	AUROC: 0.938 AUPRC: 0.299
Lazzarotto *et al.* [51]	2020	Off-target	CHANGE-seq (GTB)	One-hot + sequence features	High-throughput sequencing data generated (CHANGE-seq), GUIDE-Seq [43], CIRCLE-seq [44]	https://github.com/tsailabSJ/changeseq, Cas-OFFinder [109], DeepTools [113]	AUROC: 0.995 AUPRC:0.881
Rafid *et al.* [89]	2020	On-target	CRISPR pred(SEQ), (SVM)	Position-independent + position-specific features + n-gapped di-nucleotide	CRISPOR [60], DeepHF [56]	https://github.com/Rafid013/CRISPRpredSEQ	Spearman:0.829 AUROC: 0.893
He *et al.* [91]	2020	On-target	GuidePro (two-layer ensemble, SVM and RF)	Sequence-specific features	Doench V1 [54], Doench V2 [59], CRISPOR [60], Munoz [70], Schoonenberg [87], Aguirre [114], Evers [115], Bertomeu [117],	https://bioinformatics.mdanderson.org/apps/GuidePro, https://github.com/MDhewei/GuidePro, inDelphi [116], Lindel [118], FORECasT [119]	Spearman:0.523
Wang *et al.* [101]	2020	On-target	GNL-Scorer Eight models; (GBRT, DT, linear regression, L2 regression, L1 regression, BRR, RF, NN)	One-hot + GC count + position-independent + position-dependent features + thermodynamic features	10 public data sets: Doench V1 [54], Doench V2 [59], HCT116 [62], Hela [62], Z_fish MM [63], Z_fish VZ [64], Z_fish GZ [65], Drosophila [66], HEK293T [67], Ciona [68]	https://github.com/TerminatorJ/GNL_Scorer	Spearman: 0.502
Dhanjal *et al.* [90]	2020	Off-target	L1 logistic regression, L2 logistic regression, RF, xgboost	One-hot + GC content + position-specific features	CIRCLE-seq [44], CRISPcut [120]	http://web.iitd.ac.in/crispcut/off-targets	Accuracy: 91.49% AUROC: 0.97

(*continued*)

Table 2. Continued.

Study	Year	Target prediction	ML model(s) used	Encoding	Data	Link for data/ software	Prediction metric/ results
X. Liu *et al.* [104]	2020	Off-target and on-target	SeqCor (open-source software bundle to correct the experimental data using random forest-based)	A general-purpose hash function	DeepCRISPR [71]	https://github.com/wangyi-fudan/SeqCor	Spearman: 0.4 for off-targets, and 0.369 for on-targets
Hiranniramol *et al.* [58]	2020	On-target	sgDesigner (stacking SVM and XGBoost using logistic regression)	GC content + structural features	Wang [52], Koike-Yusa [53], Doench V1 [54], Chari [67], Shalem [105]	https://github.com/wang-lab/sgDesigner, RNAfold [121]	Spearman: 0.75 AUROC: 0.934 Accuracy: 86.3%
Konstantakos *et al.* [102]	2022	On-target	CRISPRedict (linear regression, binomial regression, logistic regression)	Overall and position-specific nucleotide composition + structural properties of sRNAs	Koike-Yusa [53], DeepSpCas9 [57], CRISPOR [60], Labuhn [71, 122],	https://github.com/VKonstantakos/CRISPRedict, http://www.crispredict.org	Spearman: 0.380 for U6 data sets, and 0.355 for T7 data sets, nDCG: 0.805 for U6 data sets, and 0.554 for T7 data sets
Zarate *et al.* [103]	2022	On-target	BoostMEC (Boosting Model for Efficient CRISPR)	GC content + position-specific features + thermodynamic features	DeepSpCas9 [57], CRISPRon [123]	https://github.com/oazarate/BoostMEC	Spearman: 0.78

RF: Random forest, GBRT: Gradient-boosted regression tree, SVM: Support Vector machine, MCC: Matthews correlation coefficient, GB: Gradient boosting, NN: Neural networks, KNN: K-nearest neighbors, GTB: Gradient tree boosting, DT: Decision tree, BRR: Bayesian ridge regression, nDCG: Normalized discounted cumulative gain.

identify the most significant generic predictors (i.e. explanatory features) in case of both guide efficiency (i.e. on-target activity) and guide specificity (i.e. off-target effects).

TRADITIONAL MACHINE LEARNING MODELS AND THEIR APPLICATIONS IN CRISPR/Cas9

In this section, we present different conventional machine learning models for on- and off-target prediction found in the genome editing literature related to CRISPR/Cas9. The presentation is organized based on the target categories: (1) off-target prediction only, (2) on-target prediction only, and finally (3) both on- and off-target prediction. Within each category, the works follow chronological order.

First, we discuss papers dealing with off-target activity prediction. Abadi *et al.* [46] proposed the CRISPR Target Assessment (CRISTA) algorithm that relies on a random forest ensemble machine learning framework to determine the propensity of a genomic site to be cleaved by a given sgRNA. The authors determined that the system attributes representing spatial structure and rigidity of the entire genomic site, as well as those related to the PAM region have the main impact on the prediction capabilities. Peng *et al.* [85] were among the first authors to capitalize on the recent advances in CRISPR/Cas9 data availability. They experimented with two positive sample sets, comprising both on- and off-targets. The first of them contains 215 sequence pairs related to 29 sgRNAs' on-target and off-target editing sites. The second data set includes 527 sequence pairs obtained using high-throughput sequencing techniques—Digenome-seq [25], GUIDE-seq [43], HTGTS [48], CIRCLE-seq [44], and multiplex Digenome-seq [92]. The authors randomly under-sampled the data to compensate for the class imbalance between on- and off-targets. Then, they trained an ensemble SVM classifier to detect the off-target sites. The authors demonstrated the ability of their model to outperform state-of-the-art predictive methods by aggregating larger numbers of sgRNA–DNA sequence pairs. The work of Peng *et al.* opened new directions for data aggregation in the field of genome editing. Chen *et al.* [88] generated the CRISPEY data set consisting of 23 936 samples, each of which contains a 20-nucleotide gRNA sequence and a 100-basepair donor DNA sequence. In this data set, 306 samples are labeled as effect samples and 23 630 samples are labeled as no-effect samples. To predict eventual off-targets, the authors applied three conventional machine learning algorithms including logistic regression, SVM [93], and random forest [94], as well as a simple DNN. The SVM model with a recall rate of 64% and the logistic regression with an accuracy of 94% provided the best prediction results overall. Zhang *et al.* [95] explored the ensemble learning potential for off-target prediction by synergizing multiple tools. The input of their ensemble learning model included five scores calculated by the following scoring methods: CCTop [27], MIT Website [28], CFD [54], MIT [60], and Cropit [96], as well as evolutionary conservation data and Chromatin state segmentation data. The authors considered an

imbalanced data set containing 25 332 putative off-target DNA sequences, with 152 verified positive off-targets. They compared the five following machine learning algorithms—AdaBoost [97], random forest, a multi-layer perceptron [98], SVM, and decision trees [99]. In their experiments, Zhang *et al.* demonstrated that the ensemble-based AdaBoost algorithm was able to outperform the other predictive algorithms in terms of the area under the precision recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) metrics. Lazzarotto *et al.* [51] proposed an approach targeting the fast pace of changes in genome editing with a scalable, automatable tagmentation-based model for estimating the genome-wide Cas9 *in vitro* activity. Their CHANGE-Seq model was designed to better understand the specificity of genome editors. In their experiments, Lazzarotto *et al.* used the encoded one-dimensional vectors to train a gradient tree boosting model to predict off-target activities. The authors highlighted the importance of the protospacer and the PAM position to ensure accurate off-target predictions. Moreover, they showed that CHANGE-Seq generally outperforms the well-known GUIDE-seq [43] and CIRCLE-seq [44] models.

Second, we present a summary of recent papers addressing the problem of machine learning prediction of on-target activities. Wang *et al.* [52] were among the first authors to use SVMs to predict sgRNA efficacy. The authors used log2 fold change of sgRNAs targeting ribosomal protein genes as their efficacy indicator. Precisely, the log2 fold change was applied to build a binary classification, where ribosomal protein gene-targeting sgRNAs were designated either as weak or as strong. Doench *et al.* [54] trained a logistic regression classifier to differentiate the highest activity quintile of sgRNAs from their lowest activity quintile. The authors used sequence features from nine mouse and human genes with cross-validation to ensure the generalization across genes. Xu *et al.* [61] applied a regularized regression technique that linearly combines the penalties of the Lasso and Ridge methods, and Elastic-Net, to predict sgRNA efficiency in CRISPR/Cas9 knockout experiments. The authors demonstrated that Elastic-Net outperforms existing models on different independent data sets. Fusi *et al.* [100] investigated how to achieve the best optimal predicative performance in CRISPR/Cas9 gene editing. The authors relied on two different primary data sets composed of mouse and human genes. They built and trained five traditional machine learning classifiers to predict the knockout efficacy, and observed that the gradient-boosted regression trees yielded the best performance overall. Rahman *et al.* [86] introduced the CRISPRpred model aiming at providing accurate *in silico* predictions of sgRNA on-target activity. CRISPRpred is capable to extract relevant features in order to use them in an SVM-based machine learning framework. The work of Rahman *et al.* emphasizes the importance of feature engineering in boosting the predictive performance of sgRNA on-target prediction models. Furthermore, Rafid *et al.* [89] demonstrated the importance of feature engineering and data pre-processing to ensure effective sgRNA on-target activity prediction. The authors proposed a novel SVM-based machine learning tool, named CRISPRpred(SEQ), which is capable to challenge the effective DeepCRISPR model [71] based on deep learning. The authors demonstrated that due to designing better explanatory features, CRISPRpred(SEQ), which used a simpler model architecture, was able to outperform DeepCRISPR in three out of four cell lines. Wang *et al.* [101] proposed a novel methodology targeting cross-species generalization of on-target activities. The authors developed the GNL-Scorer software computing two cross-species generalization scores, GNL and GNL-Human. GNL-Scorer also combines different data sets, features and models for sgRNA activity prediction, agnostic to the species. The authors claimed that GNL-Scorer facilitates the current *in silico* design of sgRNAs. Konstantakos *et al.* [102] introduced a new interpretable gRNA efficiency prediction model and the related web tool, called CRISPRedict, including various regression and classification models for gRNA scoring. This web tool offers accurate efficiency predictions under different experimental conditions (e.g. U6/T7 transcription) and the related visualizations facilitating the explanation of the obtained results. As explanatory features, the authors considered overall and position-specific nucleotide composition, as well as variables reflecting the structural properties of gRNAs. They conducted a multi-step feature selection analysis to infer a minimal relevant feature subset. Then, they used a binomial and a linear regression models to predict the percentage of successful edits for the U6 and T7 variants, and trained two logistic regression models by labeling the top 20% and the bottom 20% of gRNAs as efficient and inefficient, respectively. Konstantakos *et al.* evaluated the performance of CRISPRedict, comparing it with state-of-the-art gRNAs design tools, including some deep learning models, and concluded that despite its simplicity, CRISPRedict provides interpretable efficiency predictions with comparable performance. Zarate *et al.* [103] developed a new machine learning model, called BoostMEC (Boosting Model for Efficient CRISPR), to predict CRISPR-Cas9 editing efficiency. BoostMEC is based on a gradient boosting technique and LightGBM (Light Gradient-Boosting Machine). The LightGBM hyperparameters were tuned using tenfold cross-validation and Bayesian hyperparameter optimization. The authors compared BoostMec with 10 state-of-the-art on-target prediction models on 13 benchmark data sets. They concluded that BoostMEC, which relies on direct and derived sgRNA features and traditional machine learning, has an advantage over state-of-the-art prediction models based on deep learning because of its ability to produce more interpretable feature insights and predictions.

Finally, we discuss papers addressing the problem of both on- and off-target activity prediction by means of conventional machine learning models. Doench *et al.* [59] designed and tested their novel sgRNA design rules to create human and mouse genome-wide libraries and carry out the corresponding positive and negative selection screens. The authors proposed a new metric to predict off-target sites, and designed optimized sgRNA libraries with maximized on-target activity and minimized off-target effects. In order to identify an optimal classifier, they compared the performance of eight conventional machine learning models, including linear regression, L1-regularized linear regression, L2-regularized linear regression, a hybrid SVM plus logistic regression, random forest, gradient-boosted regression trees, L1 logistic regression (a classifier), and SVM (with linear kernel with default L2 regularization). Liu *et al.* [104] proposed an open-source software, called SeqCor, which relies on the application of the random forest algorithm to extract sequence features that influence gRNA knockout efficiency as well as gRNA off-target activity at specific sites. The aim of their work was to facilitate the extraction of the sequence features and to minimize possible bias effects that may be present in a library used in CRISPR/Cas9-based screening.

Although the use of traditional machine learning algorithms, whose main advantages are their relative simplicity and fast training, led to some impressive on- and off-target activity prediction results, recent studies conducted using deep learning methods (see the next section) often demonstrated a superior

performance. We are convinced that, in general, deep learning models are better suited for both on- and off-target activity prediction than conventional machine learning models, since modern CRISPR/Cas9 data sets contain hundreds of thousands, and sometimes millions, of samples, and state-of-the-art deep learning algorithms can be effectively used on such huge volumes of data encompassing complex non-linear patterns. However, for benchmark purpose, the results provided by deep learning algorithms should be always compared with those yielded by some well-performing traditional machine learning methods such as SVM, random forest and XGBoost, as well as their ensemble frameworks, which are capable of increasing the accuracy of individual methods.

Table 2 reports the main traditional machine learning classifiers and regressors used for on- and off-target prediction in CRISPR/Cas9.

A BRIEF REVIEW OF DEEP NEURAL NETWORKS

Deep learning applications across all research fields have recently gained popularity due to easier access to data, boosted computing power, and recent theoretical progress in supervised learning. DNNs are at the core of deep learning. They are capable of learning complex patterns from the data using multiple layers of interconnected neurons. Nonetheless, their training and optimization are still very challenging problems. This section is divided into two parts. First, we present the main properties of existing deep learning network architectures. Second, we discuss their applications in CRISPR/Cas9.

In this section, we describe succinctly the three main types of DNNs used for on- and off-target activity prediction in CRISPR/-Cas9. They are FNNs, CNNs, and RNNs. Figure 4 illustrates three typical deep learning network architectures used in CRISPR/Cas9. Finally, we present some popular activation functions of neurons used in deep learning models.

FNNs are one of the most popular DNN models [76]. In a standard FNN architecture, the information always moves forward between different layers of interconnected neurons. The two main types of FNN are as follows: a single-layer FNN and a multi-layer FNN. In a single-layer FNN, the input layer, i.e. the first layer of neurons receiving the data as input, is directly fully connected to the output layer. The output layer is the last layer outputting the predictions. In a multi-layer FNN, the input layer and the output layer are fully connected to hidden layers. Thus, a multi-layer FNN has at least three layers of neurons. Figure 4 presents a typical multi-layer FNN architecture used for off-target predictions in CRISPR/Cas9 (for more details, see [75]).

CNNs, introduced by Lecun *et al.* [124], rely, as FNNs, on fully connected layers, but also include convolutional and pooling layers. A convolutional layer consists of a collection of convolutional filters used to extract spatial features from the input image. Within the convolutional layer, different filters, or kernels, can be applied to process the data and generate feature maps. These feature maps help the neural network to better regress or classify the input data. Following the convolutional layers, hidden fully connected layers are often used to further improve the predictive performance of a CNN. An example of a CNN architecture used for off-target prediction in CRISPR/Cas9 is presented in Figure 4 (for more details, see [75]).

RNNs are another type of neural networks that can be used effectively for on- and off-target prediction in CRISPR/Cas9. In contrast to FNNs and CNNs, the information in RNNs does not always move forward; it can also move backward. The aim of RNNs is to replicate a memory process through a recurrent learning mechanism. RNNs aggregate information of the past inputs and that of the current input in order to produce the current output. An example of an RNN architecture used for off-target prediction in CRISPR/Cas9 is presented in Figure 4 (see also [75]). Two popular types of RNNs are Long Short-Term Memory (LSTM) models, which are designed to learn order dependence in sequence prediction problems [125, 126] and Gated Recurrent Unit (GRU) models, which use a similar to an LSTM prediction mechanism, but require less memory and are usually faster than LSTMs as they have no output gate [127]. Both LSTM and GRU model architectures have the ability to forget the information using a forget gate. One of the advantages of LSTMs is that they are able to overcome the vanishing gradient problem that occurs while training networks with backpropagation and gradient-based learning methods, preventing undesirable weight updates in the RNN. For the input sequence $< x_1, x_2, \ldots, x_T >$, the key mathematical equations of the forward pass of an LSTM unit are as follows [76]:

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \\
h_t &= o_t \circ \sigma_h(c_t),
\end{aligned}
\tag{1}
$$

where $\circ$ denotes the Hadamard product, x_t is the unit's input at time t, h_t is the corresponding unit's output, c_t is the hidden unit's memory, and i_t, f_t and o_t are, respectively, the activation vectors of the input gate, of the forget gate, and of the output gate. The variables W, U and b are, respectively, the weight matrices and the bias parameter, and σ_g denotes a sigmoid function. Finally, both σ_c and σ_h denote hyperbolic tangent functions.

For all deep learning network architectures presented here, the neurons rely on an activation function that is used to determine whether a given neuron should be activated or not. Thus, the activation function being used determines whether the neuron's contribution to the network is important or not in the prediction process. An activation function allows the model to transform the weighted sum of the neuron's input signals into an output signal. Table 3 summarizes the most common activation functions used with artificial neural networks.

DNNs, by their nature, have a large number of parameters to fine-tune during their training. Special attention must be given to the problem of overfitting in the model's training. Overfitting occurs when the model learns patterns only present in the training set and cannot generalize its predictive performance on the test set. Different techniques exist to limit the overfitting while training DNNs. The most important of them are early-stopping, network-reduction, regularization, and dropout [76, 128]. We invite the reader to consult the aforementioned literature for more details.

DEEP LEARNING MODELS AND THEIR APPLICATIONS IN CRISPR/Cas9

This section includes four thematic subsections. First, we discuss the studies emphasizing the use of novel sequence encoding strategies along with deep learning models. Second, we present works using different feature engineering approaches. Third, we highlight the papers applying class rebalancing techniques prior

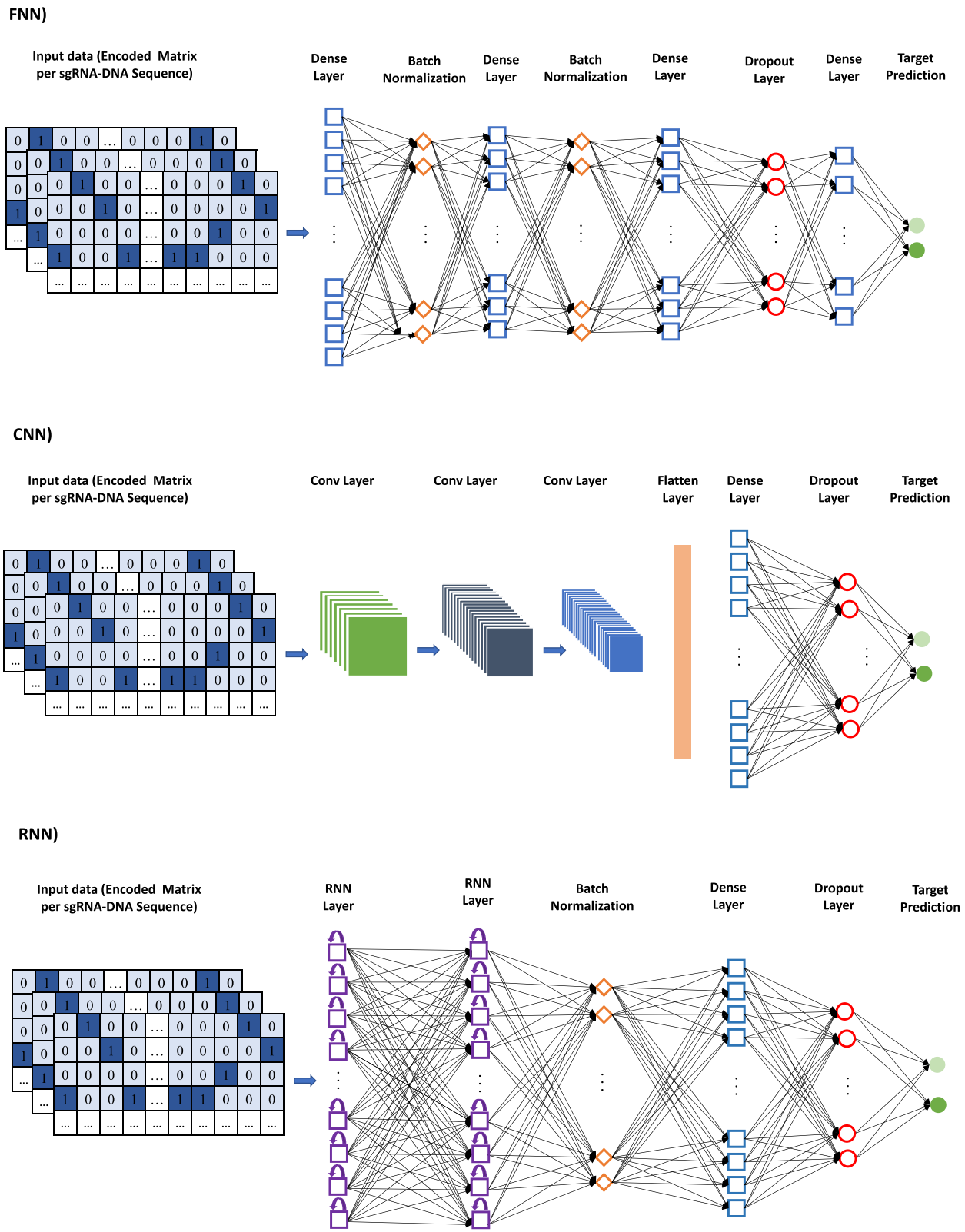

Figure 4. Typical FNN, CNN, and RNN architectures used to predict on-and off-targets in CRISPR/Cas9. For each network, the encoded matrix containing the sgRNA–DNA sequence pair information is used as input (for more details, see Charlier *et al.* [75]).

to carrying out deep learning algorithms. Forth, we describe some recent works relying on the use of attention mechanism. Within each category, the works follow the off-target, on-target and both on- and off-target prediction order.

Models relying on novel sequence encoding strategies

As regards the off-target prediction, Lin *et al.* [73] proposed a novel sequence encoding scheme based on the use of DNNs. They

Table 3. Activation functions commonly used with artificial neural networks.

Name of the function	Formula
ReLU	$\sigma(x) = \mathbf{max}(0, x)$
Leaky ReLU	$\sigma(x) = \mathbf{max}(\alpha x, x)$
Randomized leaky ReLU (RReLU)	$\sigma(x) = \mathbf{max}(0, x) + \alpha \times \mathbf{min}(0, x)$
Parametric leaky ReLU (PReLU)	$\sigma(x) = \mathbf{max}(0, x) - \alpha \times \mathbf{max}(0, -x)$
Scaled exponential Linear Units (SeLU)	$\sigma(x) = \lambda \begin{cases} x, & x > 0 \\ \alpha\, e^{x} - \alpha, & x \leq 0 \end{cases}$
Logistic (Sigmoid)	$\sigma(x) = \frac{1}{1 + e^{-x}}$
Hyperbolic Tangent (Tanh)	$\sigma(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$
Softmax	$\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$, where K is the number of classes

investigated several network architectures, including CNNs and FNNs. In their experiments, the authors used publicly available CRISPOR (18 236 samples) and GUIDE-Seq (430 samples) data sets. The transfer learning strategy was applied to obtain the off-target predictions for a much smaller GUIDE-Seq data set as the model trained on the CRISPOR data set was used to predict the GUIDE-Seq off-targets. Despite the fact that the encoding strategy proposed by Lin *et al.* [73] could lead to some information loss as the 4 × 23 matrix considered by the authors cannot fully represent the unique match-mismatch information of a given sgRNA–DNA sequence pair (it was then corrected by Lin *et al.* [74], who considered a loss-free model based on a 7 × 23 matrix encoding), very encouraging AUCROC results were obtained by these authors (i.e. an area under the curve values up to 0.972 were generated). The main conclusion of this work was that CNN and FNN deep learning networks steadily outperform state-of-the-art off-target scoring prediction methods (i.e. CFD, MIT, CROP-IT, and CCTOP) as well as some traditional machine learning classifiers (i.e. random forest, gradient boosting trees, and logistic regression). Afterwards, Lin *et al.* [74] proposed an original one-hot sequence encoding scheme (see Figure 3) and an effective CRISPR-Net model, where a Long-term Recurrent Convolutional Neural Network (LRCN) was playing the role of a feature extractor. The authors noticed that convolutional layers of LRCN were able to discover useful features from sequences directly and independently, preventing possible biases introduced by hand-crafted sequence features. The convolutional kernels within the LRCN were replaced by an inception module and a bidirectional LSTM was used for scoring the off-target activity of each potential sgRNA-target. Charlier *et al.* [75] proposed a novel sgRNA–DNA sequence encoding technique, which was applied in a deep learning off-target prediction framework. The loss-free encoding model introduced by the authors relied on 8 × 23 matrix representations. Charlier *et al.* compared the prediction performance of different FNN, CNN, and RNN network architectures (see Figure 4), as well as of several machine learning classifiers (i.e. random forest, naive Bayes, and logistic regression). The predictions were performed on two well-known gene editing data sets, CRISPOR and GUIDE-Seq, which were previously considered by Lin *et al.* [73]. The transfer learning approach was used as well to predict the off-targets on the much smaller GUIDE-Seq data set. The proposed prediction framework led to more accurate off-target prediction results, compared with those obtained by Lin *et al.* [73], yielding an improvement of the AUCROC metric up to 35%.

Regarding the on-target prediction, Xue *et al.* [129] introduced DeepCas9, an effective deep-learning framework based on CNNs. The authors proposed a network architecture to automatically learn the sequence determinants, conducting their experiments with 10 CRISPR/Cas9 data sets of different sizes. Xue *et al.* demonstrated that DeepCas9 is capable of outperforming some traditional machine learning methods such as random forest and logistic regression in terms of on-target activity prediction. In their timely review, Konstantakos *et al.* [37] evaluated 11 tools for gRNA efficiency prediction, which were applied to analyze six benchmark data sets. The evaluated tools included 10 deep learning models and one conventional machine learning model, called Azimuth 2.0 [59] and based on gradient-boosted regression trees. Most of the considered deep learning models represented the target sequence by means of different one-hot encoding strategies, while a few of them captured the epigenetic features as well. Comparison of gRNA efficiency prediction was carried out using the Spearman correlation. The authors observed that the correlation between the predicted and the true efficiency varied a lot, depending on the data set being analyzed and the model being used. Overall, the DeepHF [56] and DeepSpCas9 [57] models were consistently among the top performers along with the Average prediction strategy that consisted of the mean of the DeepCRISPR [71], DeepCas9 [129], DeepSpCas9 [57], and Azimuth 2.0 [59] predictions. Konstantakos *et al.* pointed out that simpler machine learning tools can sometimes outperform their much more sophisticated deep learning counterparts, and that large data sets that benefit from unbiased experimental measurements play a crucial role in training a generalizable model such as DeepHF or DeepSpCas9.

Concerning the prediction of both on- and off-targets, we need to mention the work of Chuai *et al.* [71], who were among the first to tackle the problem of sgRNA–DNA sequence encoding adapted to the input of deep learning models. The authors implemented a deep learning framework, called DeepCRISPR, predicting simultaneously the sgRNA on-target knockout efficacy and the off-target cleavage. An original one-hot encoding strategy used by the authors consisted of a four-channel-based sgRNA–DNA sequence encoding and epigenetic feature encoding in which each feature was considered as an independent channel. Chuai *et al.* introduced an unsupervised representation learning strategy to train a Deep Convolutional Denoising Neural Network (DCDNN) auto-encoder to learn the underlying representation of sgRNAs. The unsupervised deep representation learning approach was then used to transfer the encoded data to a hybrid DNN. The proposed network included a softmax activation function and an identity function for the classification and regression tasks, respectively. DeepCRISPR provided good prediction results in terms of both AUPRC and AUCROC, compared with the CFD prediction method [54].

Most of the aforementioned studies demonstrated impressive prediction results using novel sgRNA–DNA one-hot sequence encoding schemes combined with deep learning models. We think, however, that further prediction improvements will not anymore result from the use of sophisticated data encoding schemes, nor from the models complexity, but rather from an effective use of additional biological or physical features as well as from the application of the feature engineering and class rebalancing techniques.

Models relying on feature engineering

In addition to conventional nucleotide sequence data, some plausible biological and physical features, such as gene melting temperature, molecular weight or microhomology features, can be

also used as input of CRISPR/Cas9 predictive models. Moreover, some new informative features can be created, or reengineered, from the existing ones. Feature engineering has proven to be an important milestone when developing and optimizing predictive models [76].

As regards the off-target prediction, Liu *et al.* [80] introduced a deep learning architecture, called CnnCrispr, to predict the off-target propensity of sgRNAs at specific DNA fragments. The approach proposed by these researchers relies on the GloVe embedding model [81] to extract the global statistical information from genes. The constructed word vector matrix was then embedded into the considered deep learning model including a bidirectional LSTM and a CNN. The authors demonstrated that the proposed approach outperforms state-of-the-art classification and regression algorithms. Stortz *et al.* [130] introduced the piCRISPR deep learning model intended for off-target prediction using physically informed features. The authors designed four different feature encoding schemes to incorporate the following physically informed features: the target-guide encoding, the target-mismatch encoding, the target-mismatch-type encoding, and the target-OR-guide encoding. Moreover, they assessed the feature importance using the model-agnostic SHAP (SHapley Additive exPlanations) technique [131]. In their experiments conducted with the crisprSQL data set, Stortz *et al.* demonstrated the importance of both the sequence context and the chromatin accessibility for effective cleavage prediction. Niu et. al. [132] proposed the R-CRISPR deep learning model that encodes sgRNA target sequences into a binary matrix and then uses a CNN model as a feature extractor. Precisely, the authors applied a Rep-VGG inference time body composed of a stack of 3×3 convolutions and ReLUs [133] in the convolutional layers to extract relevant features. The CNN output was then passed to a bi-directional recurrent layers using an LSTM to get accurate off-target predictions. Fu *et al.* [134] introduced the MOFF off-target predictor based on the MOFF-target score function that is the sum of the multiplication of individual mismatch effect (IME), the combinatorial effect (CE), and the guide-intrinsic mismatch tolerance effect (GMTE), where GMTE is estimated from a dinucleotide CNN regression model. Two different encoding strategies: (1) mononucleotide encoding and (2) dinucleotide encoding were used to vectorize the input sequences. The tests conducted on a high-throughput allele-editing screen of 18 cancer hotspot mutations confirmed that MOFF significantly improves the selectivity and expands the application domain of Cas9-based allele-specific editing.

Concerning the on-target prediction, Shrawgi *et al.* [84] introduced DeepSgRNA, a deep learning architecture relying on CNN, to identify and predict RNA guides. The aim of their approach was to eliminate the need in manual feature construction, which improved the scalability of the approach. DeepSgRNA relies on hierarchical feature generation abilities of CNNs. In their experiments with the GenomeCRISPR data, the authors proved that DeepSgRNA was able to achieve state-of-the-art sgRNA prediction efficiency. Furthermore, we need to mention the pioneering work of Wang *et al.* [56], who compared several deep learning and conventional machine learning models to provide gRNA activity predictions for SpCas9 (eSpCas9(1.1)), Cas9-High Fidelity (SpCas9-HF1), and wild-type SpCas9 (WT-SpCas9) data. Feature engineering was performed using the effective Tree SHAP technique [131] that combines the SHAP values [135] with the XGBoost algorithm. The authors built two RNN and one CNN deep learning models, and trained them along with a linear regression, a L2-regularized linear regression, an XGBoost, and a multilayer perceptron. In their experiments, Wang *et al.* demonstrated that the RNN that received as input the sequence data with added biological features was able to outperform the other competing models. Their second best-performing model was the RNN that received as input sequence data only. Furthermore, Wang *et al.* developed a DeepHF website to ease the access to WT-SpCas9, eSpCas9(1.1), and SpCas9-HF1 indel data. The authors also applied SHAP with XGBoost and RNNs to assess the feature importance of the sequence input. Moreover, they used the Deep SHAP algorithm [135] to estimate the position-dependent gRNA nucleotide contribution to deep learning predictions. Based on the results of Wang *et al.*, we can conclude that additional biological information and plausible reengineered features increase the predictive performance of deep learning models, leaving opportunities for future model enhancement. Elkayam and Orenstein [136] proposed the DeepCRISTL on-target prediction model, which can be considered as an improvement of DeepHF [56]. It is based on the use of the BLSTM (Bidirectional Long Short-Term Memory) and transfer learning techniques. To improve the prediction performance of DeepHF, the authors also considered some plausible biological features of the DeepHF data set. Elkayam and Orenstein used random hyperparameter search to carry out hyperparameter optimization of their DeepCRISTL-pre-train model, which allowed them to outperform the DeepHF model proposed by Wang *et al.* [56].

Finally, the CRISPRon and CRISPRoff deep learning models, based on CNN and gradient boosting regression trees, and the related interactive webserver were developed by Xiang *et al.* [123]. For their prediction, the authors used different position-specific sequence, position-independent, and thermodynamic features. They established that the gRNA–DNA binding energy is a major contributor in predicting the on-target activity of gRNAs. Using the CRISPRon model, one can compute the on-target efficiency of all possible gRNAs with NGG PAM sequences, for genes, genomic regions, or custom sequences. The CRISPRoff model allows users to compute the specificity of gRNAs with NGG PAM sequences. Moreover, by means of CRISPRoff, one can compute the relative likelihood of cleavage by CRISPR-Cas9 at off-target sites compared with the likelihood of cleavage at on-target sites.

Feature engineering has proven to be essential in many research fields involving machine learning predictions. We think that this aspect has not yet been fully explored in CRISPR/Cas9 and that future investigations should contribute to higher classification performance of both traditional machine learning and deep learning models applied to predict on- and off-target activities.

Models relying on class rebalancing techniques

Training deep learning models with real-world CRISPR-Cas9 data is challenging because of a large natural imbalance existing between positive and negative samples. This leads to an imbalanced data classification problem with a much larger majority class and a much smaller minority class. Thus, the predictive models observe and learn more from samples of the majority class and, as a consequence, can fail to identify accurately samples from the minority class. This can negatively impact their overall predictive performance. In CRISPR/Cas9, the problem of data imbalance mainly applies to the task of off-target prediction.

In this context, we need to mention the work of Zhang *et al.* [77] and their DL-CRISPR deep learning model of off-target activity prediction, with data augmentation as a solution for the class imbalance problem. The authors first gathered data from two source types (i.e. from *in vitro* and cell-based experiments) to increase the size of the positive class samples (i.e. off-targets),

and thus improve the model's competency. Precisely, Zhang *et al*. proposed to increase synthetically the number of positive samples by rotating the sgRNA–DNA encoded images by 90, 180, and 270 degrees, respectively. Hence, the number of positive samples in the data set was quadrupled. The data were then passed to a four-layer CNN to perform the off-target prediction. The main finding of this work was that data augmentation was a critical step for improving the predictive performance of DL-CRISPR. Class rebalancing and data augmentation are still fast evolving domains. Alleviating data imbalance should boost the predictive performance of off-target prediction models as class imbalance remains one of the intrinsic properties of CRISPR-Cas9 data.

Models relying on attention mechanism

Recent progress in the development of deep learning models using attention mechanism [137] has triggered interest in many research fields, including CRISPR-Cas9, where it has already provided some promising off-target specificity and on-target efficiency prediction results.

Recently, Zhang *et al*. [78] have implemented the CRISPR-IP off-target prediction model based on a CNN, a BLSTM, and an attention layer learning the sgRNA–DNA sequence pair features. CRISPR-IP combines the four following types of network layers: (i) the convolutional layer to learn local features, (ii) the recurrent layer to learn the context features of the sequences, (iii) the attention layer to learn global features from the attention score, and (iv) the dense layer to map the features to the sample label space. The authors also used a new type of encoding scheme to overcome the problem of information loss in the sequence encoding. Xiao *et al*. [138] designed AttCRISPR, a deep learning framework based on the attention mechanism for predicting on-target activity. The proposed approach relies on two attention modules, one spatial and one temporal, facilitating the model's interpretability. AttCRISPR uses an ensemble learning strategy stacking encoding-based and embedding-based methods to improve its predictive performance. Liu *et al*. [79] analyzed two transformer-based neural networks, AttnToMismatch_CNN and AttnToCrispr_CNN, using cell-specific information of genes. Both models are similar, except that AttnToCrispr_CNN employs a linear regression at the final layer. AttnToMismatch_CNN and AttnToCrispr_CNN demonstrated competitive performance for both off-target sgRNA specificity prediction and on-target efficiency prediction. Furthermore, Liu *et al*. introduced a third model, called seqCrispr, which harbors an LSTM component and a CNN component in parallel to provide accurate on-target efficiency predictions. Finally, Zhang *et al*. [82] proposed two novel interpretable attention-based CNN models, called CRISPR-ONT and CRISPR-OFFT, designed for predicting sgRNA's on- and off-target activities, respectively. Their methodology emphasizes the importance of the feature explainability for obtaining accurate on- and off-target predictions. Interpretable attention-based CNNs were used to highlight how RNA-guide Cas9 nucleases could be used to investigate mammalian genomes.

We are confident that future promising methods involving attention mechanism and deep learning should soon emerge in the field.

Table 4 summarizes the most important recent deep learning models used for on- and off-target target prediction in CRISPR/-Cas9.

CONCLUSIONS AND OUTLOOK

AI methods have emerged as state-of-the-art approach in the field of genome editing. We reviewed recent applications of traditional machine learning and deep learning algorithms for prediction of on- and off-target activity in CRISPR/Cas9. We believe that our review paper can serve as a guideline for CRISPR/Cas9 practitioners willing to apply AI methods in genome editing.

Main Conclusions

The main conclusions of our study are as follows:

- First, we highlighted the importance of sequence encoding for sgRNA–DNA on- and off-target prediction. Initial models implied straightforward one-hot sequence encoding of the sgRNA–DNA sequence pairs [73]. Subsequent sequence encoding schemes were introduced [75] demonstrating higher predictive performance. The latest efforts have been focusing on the supplementary information embedding with different channels reflecting insertions, deletions, and mismatches [74].
- Second, some recent works have demonstrated that the ensemble ML methods have generally outperformed non-ensemble ML methods [46, 95]. For instance, AdaBoost [95] and random forest [46] led to superior predictive performance than a standard logistic regression or an SVM [100].
- Third, recent studies have highlighted the importance of feature selection and feature engineering for accurate activity prediction in CRISPR/Cas9. New methodologies have been introduced to incorporate sequence information, such as gene melting temperature, molecular weight, or microhomology features, into the model's input [56]. Some works emphasize the need of automated feature learning and automated feature engineering to boost the performance of deep learning models [78].
- Fourth, we observed that most of publicly available data sets have incomparable numbers of positive and negative samples, thus leading to a class imbalance problem that has a negative impact on the performance of both traditional ML and DL models, especially when predicting off-targets. Recent papers propose to use data augmentation to increase the number of samples of the minority class [77] or to apply some standard re-sampling techniques, such as under-sampling [80], to mitigate the impact of data imbalance [76].
- Fifth, for sufficiently large data sets, DNNs have demonstrated their superior predictive performance in comparison with both scoring methods and conventional ML algorithms such as SVM, random forest, and XGBoost [56, 73–75]. However, for smaller data sets, simpler ML tools were sometimes able to outperform some of their DL counterparts [37].
- Sixth, attention-based DL models have been extensively used in some recent works in the field [82, 137, 138]. The attention mechanisms have been shown to increase the efficacy of the deep learning process [156–159]. The latest DL models that rely on recurrent neural networks and attention-based mechanism have demonstrated very promising prediction performances [79, 138].

Research Gaps and Future Research Directions

Research gaps and future research directions related to the application of AI methods in genome editing include:

- Powerful deep learning models have a huge number of parameters that need to be tuned in the training process. Thus, to be effective, these models require large amounts of input data. Transfer learning started to be used in the field to leverage the problem of deep learning training on data sets with insufficient amount of training samples. For instance,

Table 4. A summary of studies applying deep learning models for on and off-target prediction in CRISPR/Cas9.

Study	Year	Target prediction	ML model(s) used	Encoding	Data	Link for data/ software	Prediction metric/ results
Chuai *et al.* [71]	2018	Off-target and on-target	DeepCRISPR (DCDNN)	One-hot	0.68 billion sgRNA sequences derived from 13 human cell lines were considered to generate DeepCRISPR data	https://github.com/bm2-lab/DeepCRISPR	Spearman: 0.246 AUROC: 0.804 AUPRC: 0.303
Lin *et al.* [73]	2018	Off-target	CNN and FNN	One-hot	GUIDE-seq [43], CRISPOR [60]	https://github.com/MichaelLinn/off_target_prediction	AUROC: 97.2% for CNN AUROC: 97% for FNN
Xue *et al.* [129]	2018	On-target	DeepCas9 (1D CNN)	One-hot	Wang [52], Doench V1 [54], Doench V2 [59], HCT116 [62], Z_fish MM [63], Z_fish VZ [64], Z_fish GZ [65], Chari [67], Ciona[68], Farboud [69]	https://github.com/lje00006/DeepCas9	Spearman: 0.23-0.61
Liu *et al.* [139]	2018	On-target	SeqCrispr (RNN+CNN +transfer learning)	Embedding, sgRNA–DNA binding melting temperature + DNase, CTCF, RRBS, and H3K4me3 peaks + global gene network properties	DeepCRISPR [71] CRISPR-Cpf1 [140]	https://github.com/qiaoliuhub/seqCrispr	Spearman: 0.77
Wang *et al.* [56]	2019	On-target	DeepHF (RNN)	Embedding + GC content + position-specific features + position-independent features + thermodynamic features	DeepHF - the largest gRNA on-target activity set for mammalian cells	github.com/izhangcd/DeepHF, https://bio.tools/DeepHF	Spearman: 0.867, 0.862, and 0.860
Shrawgi *et al.* [84]	2019	On-target	DeepSgRNA (CNN, with Hierarchical feature generation abilities)	One-hot	GenomeCRISPR [55]	http://genomecrispr.org	Spearman: 0.82 AUROC: 0.85
Liu *et al.* [79]	2019	Off-target	AttnToMis match_CNN (Transformer +2dCNN)	Embedding	DeepCRISPR [71]	https://github.com/qiaoliuhub/AttnToCrispr	AUROC: 0.961, AUPRC: 0.071
		Off-target	AttnToCrispr _CNN (Transformer +CNN)	Embedding			Spearman: 0.778 Pearson: 0.781 MSE: 412 ± 27 MSE: 412 ± 27
		On-target	seqCrispr (LSTM+CNN).	One-hot			Spearman: 0.765 Pearson: 0.760 MSE: 442 ± 33

(*continued*)

Table 4. Continued.

Study	Year	Target prediction	ML model(s) used	Encoding	Data	Link for data/ software	Prediction metric/ results
Dimauro *et al.* [141]	2019	On-target	CRISPRLearner (deep CNN and data augmentation)	One-hot	Farboud [69] Wang [52], Doench V1 [54], Doench V2 [59], HCT116 [62], Z_fish MM [63], Z_fish VZ [64], Z_fish GZ [65], Chari [67], Ciona[68]	https://github.com/pierclgr/crisprlearner	Spearman:0.23 -0.69
Wang *et al.* [142]	2019	On-target	CNN with 5layers+ transfer learning	One-hot	Cas9, eSpCas9, Cas9 (ΔrecA) [143]	https://github.com/biomedBit/DeepSgrnaBacteria	Spearman: 0.582, 0.7105, 0.360
Aktas *et al.* [144]	2019	Off-target and on-target	CNN, MLP, BLSTM	One-hot	DeepCRISPR [71]	https://github.com/bm2-lab/DeepCRISPR	Accuracy: 96.7%
Kim *et al.* [57]	2019	On-target	DeepSpCas9 (3 1D-CNN)	One-hot	DeepSpCas9	http://deepcrispr.info/DeepSpCas9	Spearman: 0.73
Liu *et al.* [80]	2020	Off-target	CnnCrispr (BLSTM and CNN)	Embedding (GloVe embedding model [81])	DeepCRISPR [71]	https://github.com/LQYoLH/CnnCrispr	AUROC: 0.957 AUPRC: 0.429
Zhang *et al.* [77]	2020	Off-target	DL-CRISPR (Data augmentation)	One-hot	A series of *in vitro* and cell-based assays were collected to generate data using a new data augmentation method	https://github.com/yuuuuzhang/DL-CRISPR_offtarget_prediction	Accuracy: 98.57%, Sensitivity: 95.13%
Zhang *et al.* [145]	2020	On-target	CNN-SVR	One-hot	DeepCRISPR [71]	https://github.com/Peppags/CNN-SVR	AUROC: 0.94 Spearman: 0.7
Chen *et al.* [146]	2020	Off-target	DNA-BERT and LightGBM	Embedding	DeepCRISPR [71]	https://github.com/bm2-lab/DeepCRISPR	AUROC: 0.993, AUPRC: 0.594 Spearman: 0.276
Zhang *et al.* [147]	2020	On-target	C-RNNCrispr (CNN+RNN)	One-hot	DeepCRISPR [71]	https://github.com/Peppags/C-RNNCrispr	AUROC: 0.976 Spearman: 0.877
Trivedi *et al.* [148]	2020	Off-target	Crispr2vec (logistic regression, SVM and DNN)	One-hot	GUIDE-seq [43], CIRCLE-seq [44]	http://www.rgenome.net/cas-offinder	Spearman: 0.60 AUROC: 0.91 on unseen sgRNAs, 0.88 on the smart
Lin *et al.* [74]	2020	Off-target	CRISPR-Net (LRCN)	One-hot	GUIDE-seq [43], Doench V2 [59], CRISPOR [60], CIRCLE-seq [44], SITE-Seq [45]	https://codeocean.com/capsule/9553651/tree/v1	AUROC: 0.995 AUPRC: 0.317
Zhang *et al.* [82]	2021	Off-target	CRISPR-OFFT (1d-CNN, attention)	Embedding	Off-target data sets: Digenome-seq [25], GUIDE-seq [43], BLESS [49, 50],	https://github.com/Peppags/CRISPRont-CRISPRofft	AUROC: 0.97 AUPRC: 0.79
		On-target	CRISPR-ONT	Embedding	On-target data sets: DeepHF [56], Sniper-Cas9 [149], SpCas9-NG [150], xCas9 [151]		AUROC: 0.865

(*continued*)

Table 4. Continued.

Study	Year	Target prediction	ML model(s) used	Encoding	Data	Link for data/ software	Prediction metric/ results
Xiao *et al.* [138]	2021	On-target	AttCRISPR (Embedding -based Method)	One-hot and embedding	DeepHF [56]	https://github.com/South-Walker/AttCRISPR	Spearman: 0.872
Charlier *et al.* [75]	2021	Off-target	FNN, CNN, RNN, RF, NB, LR.	One-hot	GUIDE-seq [43], CRISPOR [60]	https://github.com/dagrate/dl-offtarget	AUROC: 0.995 AUC PR1: 0.949 Accuracy: 99.9%
Stortz *et al.* [130]	2021	Off-target	piCRISPR (RNN, CNN)	One-hot + physically informed features: the target-guide, the target-mismatch, the target-mismatch-type and the target-OR-guide encoding	crisprSQL [152]	https://github.com/florianst/picrispr	AUROC: 0.983 AUPRC: 0.978 Spearman: 0.1
Xiang *et al.* [123]	2021	On-target and off-target	CRISPRon and CRISPRoff: Gradient boosting, regression trees, CNN	One-hot + GC content + position-specific features position-independent features + thermodynamic features	A pool of 12,000 gRNA oligos, targeting 3,834 human protein-coding genes included in CRISPRon database, Kim *et al.* data set [153]	https://rth.dk/resources/crispr/crisproff and https://rth.dk/resources/crispr/crispron	CRISPRon: Spearman: 0.91
Niu *et al.* [132]	2021	Off-target	R-CRISPR (bi-directional recurrent network)	One-hot	GUIDE-seq [43], Doench V2 [59], CRISPOR [60], CIRCLE-seq [44], SITE-Seq [45]	https://codeocean.com/capsule/9553651/tree/v1	AUROC: 0.991 AUPRC: 0.319
Vinod kumar *et al.* [154]	2021	Off-target	GCN-CRISPR Graph Convolution Network	One-hot	CRISPOR [60]	DeepCrispr: https://doi.org/10.1186/s13059-018-1459-4, CnnCrispr: https://doi.org/10.1186/s12859-020-3395-z	AUROC: 0.987
Zhang *et al.* [78]	2022	Off-target	CRISPR-IP (CNN, BLSTM)	One-hot	CIRCLE-Seq [44], SITE-Seq [45]	https://github.com/BioinfoVirgo/CRISPR-IP	AUROC: 0.982 AUPRC: 0.751 Accuracy: 0.990
Elkayam and Orenstein [136]	2022	On-target	DeepCRISTL (BLSTM + transfer learning)	Embedding + GC content + position-specific features + position-independent features + thermodynamic features	DeepHF [56], CRISPRon [123]	https://github.com/OrensteinLab/DeepCRISTL	Spearman: 0.878
Fu *et al.* [134]	2022	Off-target	MOFF (two CNN regression models)	One-hot	GUIDE-seq data [43], CHANGE-seq [51], TTISS [155]	https://github.com/MDhewei/MOFF	Spearman: 0.5

RNN: Recurrent Neural Network, CNN: Convolutional Neural Network, FNN: Feedforward Neural Network, DCDNN: Deep Convolutional Denoising Neural Network, LSTM: Long Short-Term Memory, BLSTM: Bidirectional Long Short-Term Memory, LRCN: Long-term Recurrent Convolutional Network, SVR: Support Vector Regression.

Lin *et al.* and Charlier *et al.* [73, 75] have successively employed transfer learning to predict the off-target sequences in small data sets. Further experiments should show how the most appropriate larger data sets used for training could be selected. Moreover, it would be interesting to see whether some pretrained deep learning models could be effectively used for transfer learning in CRISPR/Cas9.

- DNNs usually stack several layers of different types, each of which often containing dozens of neurons. Thus, designing efficient deep learning network architectures and finding optimal sets of hyperparameters remain extremely important and challenging tasks [160, 161]. Various hyperparameter tuning techniques, which should be extensively tested with CRISPR/Cas9 data, include evolutionary strategies, random grid search, exhaustive grid search, and Bayesian optimization [76].
- Explainability and interpretability of DNNs have been a topic of interest for the past few years [76, 162]. Recent methodologies have been introduced to further address the lack of human-level explainability and interpretability in the field [163–165]. Future research in genome editing could fill the gap in understanding the nature of on- and off-target activities, which would be an important milestone in clinical applications.
- As we pointed out, some recent works in the field have focused on the use of features engineering to boost the predictive performance of machine learning models [56, 84]. Informative features such as epigenetic features, microhomology properties, or RNA fold score can be further exploited to increase the models accuracy. Convolutional layers of CNN and LRCN deep learning networks are able to discover useful features from sequences directly and independently, avoiding eventual biases introduced by hand-crafted features [74, 84, 132]. The use of the SHAP [135] (SHapley Additive exPlanations—this algorithm gives an explanation to the model's behavior, connecting optimal credit allocation with local explanations using the classic Shapley values from game theory), Tree SHAP [131] (this algorithm calculates SHAP values for tree-based models), and Deep SHAP [135] algorithms (this is a high-speed approximation algorithm for SHAP values) is highly recommended to assess how each feature impacts the selected model.
- Uncertainty quantification is a key technique to improve the trustworthiness of predictions made by a trained network. This technique has become popular for evaluating uncertainty in various research fields [166–170]. There are two types of uncertainty: the aleatoric uncertainty that is an inherent property of the data distribution and the epistemic uncertainty that refers to the model's uncertainty. This technique could be effectively applied in genome editing to improve the trustworthiness of on- and off-target predictions. Kirillov *et al.* [171] have recently designed one of the first methods that incorporates uncertainty into the final prediction. This method provides interpretable evaluation of Cas9–gRNA and Cas12a–gRNA specificity using deep kernel learning, predicting the cleavage efficiency of a gRNA with a corresponding confidence interval.
- Active learning is a semi-supervised technique in which a learning algorithm is used to label unlabeled data. An active learning algorithm uses an initial subset of labeled data for training. The algorithm then predicts the most appropriate labels for unlabeled data. This technique is of particular interest in biology because obtaining labeled data is often costly and time-consuming [172–174]. Active learning can be employed in genome editing in situations when unlabeled data are abundant, while accurate automatic or manual labeling is impossible.

Key Points

- We reviewed current knowledge regarding the use of supervised machine learning methods for on- and off-target prediction in CRISPR/Cas9.
- We highlighted the importance of the data preprocessing step including encoding of the sgRNA–DNA sequence pairs without any information loss, embedding supplementary data with different channels reflecting insertions, deletions and mismatches, and considering some additional sequence information such as gene melting temperature, molecular weight, or microhomology features.
- Most of CRISPR/Cas9 data sets have incomparable numbers of positive and negative samples, thus leading to a class imbalance situation that should be mitigated using either data augmentation or data re-sampling techniques, especially in the case of off-target prediction.
- When training data sets were large enough, DNNs have demonstrated their superior predictive performance in comparison with scoring methods and traditional machine learning algorithms. However, for benchmark purpose, the results obtained using state-of-the-art deep learning methods should be compared with those provided by some effective conventional machine learning algorithms, such as SVM, random forest, and XGBoost.
- We emphasized the importance of feature selection for accurate on- and off-target prediction in CRISPR/Cas9. Thus, the automated feature learning and automated feature engineering techniques should be used to boost the performance of deep learning models.

ACKNOWLEDGMENTS

We thank Dr Robert Nadon (McGill University) and the three anonymous reviewers of this manuscript for their helpful comments and suggestions.

FUNDING

This work was supported by Le Fonds Québécois de la Recherche sur la Nature et les Technologies [173878], and Natural Sciences and Engineering Research Council of Canada [249644].

REFERENCES

1. Esvelt KM, Wang HH. Genome-scale engineering for systems and synthetic biology. *Mol Syst Biol* 2013; **9**(1): 641.
2. Puchta H, Fauser F. Gene targeting in plants: 25 years later. *Int J Dev Biol* 2013; **57**(6–7-8): 629–37.
3. Barrangou R, Doudna JA. Applications of crispr technologies in research and beyond. *Nat Biotechnol* 2016; **34**(9): 933–41.
4. Manghwar H, Lindsey K, Zhang X, Jin S. Crispr/cas system: recent advances and future prospects for genome editing. *Trends Plant Sci* 2019; **24**(12): 1102–25.

5. Bogdanove AJ, Bohm A, Miller JC, et al. Engineering altered protein–dna recognition specificity. *Nucleic Acids Res* 2018; **46**(10): 4845–71.
6. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *Science* 2012; **337**(6096): 816–21.
7. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the cas9 rna-guided endonuclease. *Nat Biotechnol* 2013; **31**(3): 230–2.
8. Le Cong F, Ran A, Cox D, et al. Multiplex genome engineering using crispr/cas systems. *Science* 2013; **339**(6121): 819–23.
9. Mali P, John Aach P, Stranges B, et al. Cas9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* 2013; **31**(9): 833–8.
10. Chang N, Changhong Sun L, Gao DZ, et al. Genome editing with rna-guided cas9 nuclease in zebrafish embryos. *Cell Res* 2013; **23**(4): 465–72.
11. Hsu PD, Lander ES, Zhang F. Development and applications of crispr-cas9 for genome engineering. *Cell* 2014; **157**(6): 1262–78.
12. Kang XJ, Caparas CIN, Soh BS, Fan Y. Addressing challenges in the clinical applications associated with crispr/cas9 technology and ethical questions to prevent its misuse. *Protein Cell* 2017; **8**(11): 791–5.
13. Liang P, Yanwen X, Zhang X, et al. Crispr/cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell* 2015; **6**(5): 363–72.
14. Ma H, Marti-Gutierrez N, Park S-W, et al. Correction of a pathogenic gene mutation in human embryos. *Nature* 2017; **548**(7668): 413–9.
15. Liu H, Ding Y, Zhou Y, et al. Crispr-p 2.0: an improved crispr-cas9 tool for genome editing in plants. *Mol Plant* 2017; **10**(3): 530–2.
16. Tang X, Zheng X, Qi Y, et al. A single transcript crispr-cas9 system for efficient genome editing in plants. *Mol Plant* 2016; **9**(7): 1088–91.
17. Raitskin O, Patron NJ. Multi-gene engineering in plants with rna-guided cas9 nuclease. *Curr Opin Biotechnol* 2016; **37**: 69–75.
18. Zarei A, Razban V, Hosseini SE, Tabei SMB. Creating cell and animal models of human disease by genome editing using crispr/cas9. *J Gene Med* 2019; **21**(4): e3082.
19. Wang D, Huang J, Wang X, et al. The eradication of breast cancer cells and stem cells by 8-hydroxyquinoline-loaded hyaluronan modified mesoporous silica nanoparticle-supported lipid bilayers containing docetaxel. *Biomaterials* 2013; **34**(31): 7662–73.
20. Barrangou R, Fremaux C, Deveau H, et al. Crispr provides acquired resistance against viruses in prokaryotes. *Science* 2007; **315**(5819): 1709–12.
21. Bak RO, Gomez-Ospina N, Porteus MH. Gene editing on center stage. *Trends Genet* 2018; **34**(8): 600–11.
22. Shah SA, Erdmann S, Mojica FJM, Garrett RA. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* 2013; **10**(5): 891–9.
23. Zhang X-H, Tee LY, Wang X-G, et al. Off-target effects in crispr/cas9-mediated genome engineering. *Mol Ther Nucleic Acids* 2015; **4**:e264.
24. Chen JS, Dagdas YS, Kleinstiver BP, et al. Enhanced proofreading governs crispr–cas9 targeting accuracy. *Nature* 2017; **550**(7676): 407–10.
25. Kim D, Bae S, Park J, et al. Digenome-seq: genome-wide profiling of crispr-cas9 off-target effects in human cells. *Nat Methods* 2015; **12**(3): 237–43.
26. Ran FAFA, Hsu PD, Wright J, et al. Genome engineering using the crispr-cas9 system. *Nat Protoc* 2013; **8**(11): 2281–308.
27. Stemmer M, Thumberger T, del Sol M, et al. Cctop: an intuitive, flexible and reliable crispr/cas9 target prediction tool. *PloS One* 2015; **10**(4): e0124633.
28. Hsu PD, Scott DA, Weinstein JA, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nat Biotechnol* 2013; **31**(9): 827–32.
29. Heigwer F, Kerr G, Boutros M. E-crisp: fast crispr target site identification. *Nat Methods* 2014; **11**(2): 122–3.
30. Montague TG, Cruz JM, Gagnon JA, et al. Chopchop: a crispr/-cas9 and talen web tool for genome editing. *Nucleic Acids Res* 2014; **42**(W1): W401–7.
31. Wang F, Wang L, Zou X, et al. Advances in crispr-cas systems for rna targeting, tracking and editing. *Biotechnol Adv* 2019; **37**(5): 708–29.
32. Liu G, Zhang Y, Zhang T. Computational approaches for effective crispr guide rna design and evaluation. *Comput Struct Biotechnol J* 2020; **18**:35–44.
33. Chen S, Yao Y, Zhang Y, Fan G. Crispr system: discovery, development and off-target detection. *Cell Signal* 2020; **70**:109577.
34. Yan J, Xue D, Chuai G, et al. Benchmarking and integrating genome-wide crispr off-target detection and prediction. *Nucleic Acids Res* 2020; **48**(20): 11370–9.
35. Yaish O, Asif M, Orenstein Y. A systematic evaluation of data processing and problem formulation of crispr off-target site prediction. *Brief Bioinform* 2022; **23**(5): bbac157.
36. O'Brien AR, Burgio G, Bauer DC. Domain-specific introduction to machine learning terminology, pitfalls and opportunities in crispr-based gene editing. *Brief Bioinform* 2021; **22**(1): 308–14.
37. Konstantakos V, Nentidis A, Krithara A, Paliouras G. Crispr–cas9 grna efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res* 2022; **50**(7): 3616–37.
38. Naeem M, Majeed S, Hoque MZ, Ahmad I. Latest developed strategies to minimize the off-target effects in crispr-cas-mediated genome editing. *Cell* 2020; **9**(7): 1608.
39. Almutiri T, Saeed F, Alassaf M. A survey of machine learning and deep learning applications in genome editing. In: *Advances on Smart and Soft Computing*. Springer, 2022, 145–62.
40. Wilson LOW, O'Brien AR, Bauer DC. The current state and future of crispr-cas9 grna design tools. *Front Pharmacol* 2018; **9**:749.
41. Wang J, Zhang X, Cheng L, Luo Y. An overview and metanalysis of machine and deep learning-based crispr grna design tools. *RNA Biol* 2020; **17**(1): 13–22.
42. Newman A, Starrs L, Burgio G. Cas9 cuts and consequences; detecting, predicting, and mitigating crispr/cas9 on-and off-target damage: techniques for detecting, predicting, and mitigating the on-and off-target effects of cas9 editing. *Bioessays* 2020; **42**(9): 2000047.
43. Tsai SQ, Zheng Z, Nguyen NT, et al. Guide-seq enables genome-wide profiling of off-target cleavage by crispr-cas nucleases. *Nat Biotechnol* 2015; **33**(2): 187–97.
44. Tsai SQ, Nguyen NT, Malagon-Lopez J, et al. Circle-seq: a highly sensitive in vitro screen for genome-wide crispr–cas9 nuclease off-targets. *Nat Methods* 2017; **14**(6): 607–14.
45. Cameron P, Fuller CK, Donohoue PD, et al. Mapping the genomic landscape of crispr–cas9 cleavage. *Nat Methods* 2017; **14**(6): 600–6.

46. Abadi S, Yan WX, 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**(10): e1005807.
47. Kleinstiver BP, Pattanayak V, Prew MS, et al. High-fidelity crispr–cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 2016; **529**(7587): 490–5.
48. Frock RL, Jiazhi H, Meyers RM, et al. Genome-wide detection of dna double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* 2015; **33**(2): 179–86.
49. Ran FACL, Cong L, Yan WX, et al. In vivo genome editing using staphylococcus aureus cas9. *Nature* 2015; **520**(7546): 186–91.
50. Slaymaker IM, Gao L, Zetsche B, et al. Rationally engineered cas9 nucleases with improved specificity. *Science* 2016; **351**(6268): 84–8.
51. Lazzarotto CR, Malinin NL, Li Y, et al. Change-seq reveals genetic and epigenetic effects on crispr–cas9 genome-wide activity. *Nat Biotechnol* 2020; **38**(11): 1317–27.
52. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the crispr-cas9 system. *Science* 2014; **343**(6166): 80–4.
53. Koike-Yusa H, Li Y, Tan E-P, et al. Genome-wide recessive genetic screening in mammalian cells with a lentiviral crispr-guide rna library. *Nat Biotechnol* 2014; **32**(3): 267–73.
54. Doench JG, Hartenian E, Graham DB, et al. Rational design of highly active sgrnas for crispr-cas9–mediated gene inactivation. *Nat Biotechnol* 2014; **32**(12): 1262–7.
55. Rauscher B, Heigwer F, Breinig M, et al. Genomecrispr-a database for high-throughput crispr/cas9 screens. *Nucleic Acids Res* 2016;gkw997.
56. Wang D, Zhang C, Wang B, et al. Optimized crispr guide rna design for two high-fidelity cas9 variants by deep learning. *Nat Commun* 2019; **10**(1): 1–14.
57. Kim HK, Kim Y, Lee S, et al. Spcas9 activity prediction by deepspcas9, a deep learning–based model with high generalization performance. *Sci Adv* 2019; **5**(11): eaax9249.
58. Hiranniramol K, Chen Y, Liu W, Wang X. Generalizable sgrna design for improved crispr/cas9 editing efficiency. *Bioinformatics* 2020; **36**(9): 2684–9.
59. Doench JG, Fusi N, Sullender M, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nat Biotechnol* 2016; **34**(2): 184–91.
60. Haeussler M, Schönig K, Eckert H, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide rna selection tool crispor. *Genome Biol* 2016; **17**(1): 1–12.
61. Han X, Xiao T, Chen C-H, et al. Sequence determinants of improved crispr sgrna design. *Genome Res* 2015; **25**(8): 1147–57.
62. Hart T, Chandrashekhar M, Aregger M, et al. High-resolution crispr screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 2015; **163**(6): 1515–26.
63. Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, et al. Crisprscan: designing highly efficient sgrnas for crispr-cas9 targeting in vivo. *Nat Methods* 2015; **12**(10): 982–8.
64. Varshney GK, Pei W, LaFave MC, et al. High-throughput gene targeting and phenotyping in zebrafish using crispr/cas9. *Genome Res* 2015; **25**(7): 1030–42.
65. Gagnon JA, Valen E, Thyme SB, et al. Efficient mutagenesis by cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide rnas. *PloS One* 2014; **9**(5): e98186.
66. Ren X, Yang Z, Jiang X, et al. Enhanced specificity and efficiency of the crispr/cas9 system with optimized sgrna parameters in drosophila. *Cell Rep* 2014; **9**(3): 1151–62.
67. Chari R, Mali P, Moosburner M, Church GM. Unraveling crispr-cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* 2015; **12**(9): 823–6.
68. Gandhi S, Christiaen L, Stolfi A. Rational design and whole-genome predictions of single guide rnas for efficient crispr/cas9-mediated genome editing in ciona. 2016.
69. Farboud B, Meyer BJ. Dramatic enhancement of genome editing by crispr/cas9 through improved guide rna design. *Genetics* 2015; **199**(4): 959–71.
70. Munoz DM, Cassiani PJ, Li L, et al. Crispr screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regionscrispr screens for the discovery of cancer vulnerabilities. *Cancer Discov* 2016; **6**(8): 900–13.
71. Chuai G, Ma H, Yan J, et al. Deepcrispr: optimized crispr guide rna design by deep learning. *Genome Biol* 2018; **19**(1): 1–18.
72. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Processing Syst* 2013; **26**.
73. Lin J, Wong K-C. Off-target predictions in crispr-cas9 gene editing using deep learning. *Bioinformatics* 2018; **34**(17): i656–63.
74. Lin J, Zhang Z, Zhang S, et al. Crispr-net: a recurrent convolutional network quantifies crispr off-target activities with mismatches and indels. *Adv Sci* 2020; **7**(13): 1903562.
75. Charlier J, Nadon R, Makarenkov V. Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in crispr-cas9 gene editing. *Bioinformatics* 2021; **37**(16): 2299–307.
76. Goodfellow I, Bengio Y, Courville A. Deep learning. 2016.
77. Zhang Y, Long Y, Yin R, Kwoh CK. Dl-crispr: a deep learning method for off-target activity prediction in crispr/cas9 with data augmentation. *IEEE. Access* 2020; **8**:76610–7.
78. Zhang Z-R, Jiang Z-R. Effective use of sequence information to predict crispr-cas9 off-target. *Comput Struct Biotechnol J* 2022; **20**: 650–61.
79. Liu Q, He D, Xie L. Prediction of off-target specificity and cell-specific fitness of crispr-cas system using attention boosted deep learning and network-based gene feature. *PLoS Comput Biol* 2019; **15**(10): e1007480.
80. Liu Q, Cheng X, Liu G, et al. Deep learning improves the ability of sgrna off-target propensity prediction. *BMC Bioinform* 2020; **21**(1): 1–15.
81. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–43, 2014.
82. Zhang G, Zeng T, Dai Z, Dai X. Prediction of crispr/cas9 single guide rna cleavage efficiency and specificity by attention-based convolutional neural networks. *Comput Struct Biotechnol J* 2021; **19**:1445–57.
83. Chollet F, et al. *Keras*, 2015.
84. Shrawgi H, Sisodia DS. Convolution neural network model for predicting single guide rna efficiency in crispr/cas9 system. *Chemom Intel Lab Syst* 2019; **189**:149–54.
85. Peng H, Zheng Y, Zhao Z, et al. Recognition of crispr/cas9 off-target sites through ensemble learning of uneven mismatch distributions. *Bioinformatics* 2018; **34**(17): i757–65.
86. Rahman MK, Sohel M, Rahman.Crisprpred: a flexible and efficient tool for sgrnas on-target activity prediction in crispr/cas9 systems. *PloS One* 2017; **12**(8): e0181943.
87. Schoonenberg VAC, Cole MA, Yao Q, et al. Crispro: identification of functional protein coding sequences based on genome editing dense mutagenesis. *Genome Biol* 2018; **19**(1): 1–19.

88. Chen S-A A, Tran E. Optimizing precision genome editing through machine learning. *Forest (C= 001, l2)* 2019; **85**(15.78): 1–39.
89. Rafid AHM, Toufikuzzaman M, Rahman MS, et al. Crisprpred (seq): a sequence-based method for sgrna on target activity prediction using traditional machine learning. *BMC Bioinform* 2020; **21**(1): 1–13.
90. Dhanjal JK, Dammalapati S, Pal S, Sundar D. Evaluation of off-targets predicted by sgrna design tools. *Genomics* 2020; **112**(5): 3609–14.
91. He W, Wang H, Wei Y, et al. Guidepro: a multi-source ensemble predictor for prioritizing sgrnas in crispr/cas9 protein knock-outs. *Bioinformatics* 2021; **37**(1): 134–6.
92. Kim D, Kim S, Kim S, et al. Genome-wide target specificities of crispr-cas9 nucleases revealed by multiplex digenome-seq. *Genome Res* 2016; **26**(3): 406–15.
93. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discovery* 1998; **2**(2): 121–67.
94. Breiman L. Random forests. *Mach Learn* 2001; **45**(1): 5–32.
95. Zhang S, Li X, Lin Q, Wong K-C. Synergizing crispr/cas9 off-target predictions for ensemble insights and practical applications. *Bioinformatics* 2019; **35**(7): 1108–15.
96. Singh R, Kuscu C, Quinlan A, et al. Cas9-chromatin binding information enables more accurate crispr off-target prediction. *Nucleic Acids Res* 2015; **43**(18): e118–8.
97. Freund Y, Schapire RE, et al. Experiments with a new boosting algorithm. In: *icml*, Vol. **96**. Citeseer, 1996, 148–56.
98. Bishop CM, et al. *Neural networks for pattern recognition*. Oxford university press, 1995.
99. Ross Quinlan J. Simplifying decision trees. *Int J Man Mach Stud* 1987; **27**(3): 221–34.
100. Fusi N, Smith I, Doench J, Listgarten J. In silico predictive modeling of crispr/cas9 guide efficiency. *BioRxiv* 2015;021568.
101. Wang J, Xiang X, Bolund L, et al. Gnl-scorer: a generalized model for predicting crispr on-target activity by machine learning and featurization. *J Mol Cell Biol* 2020; **12**(11): 909–11.
102. Konstantakos V, Nentidis A, Krithara A, Paliouras G. CRISPRedict: a CRISPR-Cas9 web tool for interpretable efficiency predictions. *Nucleic Acids Res* 2022; **50**(W1): W191–8.
103. Zarate OA, Yang Y, Wang X, Wang J-P. Boostmec: predicting crispr-cas9 cleavage efficiency through boosting models. *BMC Bioinform* 2022; **23**(1): 1–14.
104. Liu X, Yang Y, Qiu Y, et al. Seqcor: correct the effect of guide rna sequences in clustered regularly interspaced short palindromic repeats/cas9 screening by machine learning algorithm. *J Genet Genomics* 2020; **47**(11): 672–80.
105. Shalem O, Sanjana NE, Hartenian E, et al. Genome-scale crispr-cas9 knockout screening in human cells. *Science* 2014; **343**(6166): 84–7.
106. Zhou Y, Zhu S, Cai C, et al. High-throughput screening of a crispr/cas9 library for functional genomics in human cells. *Nature* 2014; **509**(7501): 487–91.
107. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-scale crispr-mediated control of gene repression and activation. *Cell* 2014; **159**(3): 647–61.
108. Konermann S, Brigham MD, Trevino AE, et al. Genome-scale transcriptional activation by an engineered crispr-cas9 complex. *Nature* 2015; **517**(7536): 583–8.
109. 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**(10): 1473–5.
110. Donovan KF, Hegde M, Sullender M, et al. Creation of novel protein variants with crispr/cas9-mediated mutagenesis: turning a screening by-product into a discovery tool. *PloS One* 2017; **12**(1): e0170445.
111. Brenan L, Andreev A, Cohen O, et al. Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants. *Cell Rep* 2016; **17**(4): 1171–83.
112. Listgarten J, Weinstein M, Kleinstiver BP, et al. Prediction of off-target activities for the end-to-end design of crispr guide rnas. *Nat Biomed Eng* 2018; **2**(1): 38–47.
113. Ramírez F, Dündar F, Diehl S. Björn a Grüning, and Thomas Manke. Deeptools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 2014; **42**(W1): W187–91.
114. Aguirre AJ, Meyers RM, Weir BA, et al. Genomic copy number dictates a gene-independent cell response to crispr/cas9 targetinggenomic copy number affects crispr/cas9 screens. *Cancer Discov* 2016; **6**(8): 914–29.
115. Evers B, Jastrzebski K, Heijmans JPM, et al. Crispr knockout screening outperforms shrna and crispri in identifying essential genes. *Nat Biotechnol* 2016; **34**(6): 631–3.
116. Shen MW, Arbab M, Hsu JY, et al. Predictable and precise template-free crispr editing of pathogenic variants. *Nature* 2018; **563**(7733): 646–51.
117. Bertomeu T, Coulombe-Huntington J, Chatr-Aryamontri A, et al. A high-resolution genome-wide crispr/cas9 viability screen reveals structural features and contextual diversity of the human cell-essential proteome. *Mol Cell Biol* 2018; **38**(1): e00302–17.
118. Chen W, McKenna A, Schreiber J, et al. Massively parallel profiling and predictive modeling of the outcomes of crispr/cas9-mediated double-strand break repair. *Nucleic Acids Res* 2019; **47**(15): 7989–8003.
119. Allen F, Crepaldi L, Alsinet C, et al. Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nat Biotechnol* 2019; **37**(1): 64–72.
120. Dhanjal JK, Radhakrishnan N, Sundar D. Crispcut: a novel tool for designing optimal sgrnas for crispr/cas9 based experiments in human cells. *Genomics* 2019; **111**(4): 560–6.
121. Hofacker IL. Vienna rna secondary structure server. *Nucleic Acids Res* 2003; **31**(13): 3429–31.
122. Labuhn M, Adams FF, Ng M, et al. Refined sgrna efficacy prediction improves large-and small-scale crispr–cas9 applications. *Nucleic Acids Res* 2018; **46**(3): 1375–85.
123. Xiang X, Corsi GI, Anthon C, et al. Enhancing crispr-cas9 grna efficiency prediction by data integration and deep learning. *Nat Commun* 2021; **12**(1): 1–9.
124. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; **86**(11): 2278–324.
125. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; **9**(8): 1735–80.
126. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with lstm. *Neural Comput* 2000; **12**(10): 2451–71.
127. Chung J, Gulcehre C, Cho KH, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:14123555* 2014.
128. Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, **1168**, 022022. IOP Publishing, 2019.
129. Xue L, Tang B, Chen W, Luo J. Prediction of crispr sgrna activity using a deep convolutional neural network. *J Chem Inf Model* 2018; **59**(1): 615–24.

130. Störtz F, Mak J, Minary P. Picrispr: physically informed features improve deep learning models for crispr/cas9 off-target cleavage prediction. *bioRxiv* 2021.
131. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:180203888* 2018.
132. 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**(12): 1878.
133. Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733–42, 2021.
134. Rongjie F, He W, Dou J, et al. Systematic decomposition of sequence determinants governing crispr/cas9 specificity. *Nat Commun* 2022; **13**(1): 474.
135. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30**.
136. Elkayam S, Orenstein Y. Deepcristl: deep transfer learning to predict crispr/cas9 functional and endogenous on-target editing efficiency. *Bioinformatics* 2022; **38**(Supplement_1): i161–8.
137. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; **30**.
138. Xiao L-M, Wan Y-Q, Jiang Z-R. Attcrispr: a spacetime interpretable model for prediction of sgrna on-target activity. *BMC Bioinform* 2021; **22**(1): 1–17.
139. Liu Q, He D, Xie L. Identifying context-specific network features for crispr-cas9 targeting efficiency using accurate and interpretable deep neural network. *bioRxiv* 2018;505602.
140. Kim HK, Min S, Song M, et al. Deep learning improves prediction of crispr–cpf1 guide rna activity. *Nat Biotechnol* 2018; **36**(3): 239–41.
141. Dimauro G, Colagrande P, Carlucci R, et al. Crisprlearner: a deep learning-based system to predict crispr/cas9 sgrna on-target cleavage efficiency. *Electronics* 2019; **8**(12): 1478.
142. Wang L, Zhang J. Prediction of sgrna on-target activity in bacteria by deep learning. *BMC Bioinform* 2019; **20**(1): 1–14.
143. Guo J, Wang T, Guan C, et al. Improved sgrna design in bacteria via genome-wide activity profiling. *Nucleic Acids Res* 2018; **46**(14): 7052–69.
144. Özlem Aktas, Elif Dogan, and Tolga Ensari. Crispr/cas9 target prediction with deep learning. *In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 1–5. IEEE, 2019.
145. Zhang G, Dai Z, Dai X. A novel hybrid cnn-svr for crispr/cas9 guide rna activity prediction. *Front Genet* 2020; **10**:1303.
146. Dong Chen, Wenjie Shu, and Shaoliang Peng. Predicting crispr-cas9 off-target with self-supervised neural networks. *In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 245–50. IEEE, 2020.
147. Zhang G, Dai Z, Dai X. C-rnncrispr: prediction of crispr/cas9 sgrna activity using convolutional and recurrent neural networks. *Comput Struct Biotechnol J* 2020; **18**:344–54.
148. Trivedi TB, Boger R, Kamath GM, et al. Crispr2vec: machine learning model predicts off-target cuts of crispr systems. *bioRxiv* 2020.
149. Lee JK, Jeong E, Lee J, et al. Directed evolution of crispr-cas9 to increase its specificity. *Nat Commun* 2018; **9**(1): 1–10.
150. Nishimasu H, Shi X, Ishiguro S, et al. Engineered crispr-cas9 nuclease with expanded targeting space. *Science* 2018; **361**(6408): 1259–62.
151. Hu JH, Miller SM, Geurts MH, et al. Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature* 2018; **556**(7699): 57–63.
152. Störtz F, Minary P. Crisprsql: a novel database platform for crispr/cas off-target cleavage assays. *Nucleic Acids Res* 2021; **49**(D1): D855–61.
153. Kim N, Kim HK, Lee S, et al. Prediction of the sequence-specific cleavage activity of cas9 variants. *Nat Biotechnol* 2020; **38**(11): 1328–36.
154. Vinodkumar PK, Ozcinar C, Anbarjafari G. Prediction of sgrna off-target activity in crispr/cas9 gene editing using graph convolution network. *Entropy* 2021; **23**(5): 608.
155. Schmid-Burgk JL, Gao L, Li D, et al. Highly parallel profiling of cas9 variant specificity. *Mol Cell* 2020; **78**(4): 794–800.
156. Chen W, Ouyang S, Tong W, et al. Gcsanet: a global context spatial attention deep learning network for remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2022; **15**: 1150–62.
157. Shen L, Zheng J, Lee EH, et al. Attention-guided deep learning for gestational age prediction using fetal brain mri. *Sci Rep* 2022; **12**(1): 1–10.
158. Santana Correia A de, Colombini EL. Attention, please! A survey of neural attention models in deep learning. *Artif Intell Rev* 2022; 1–88.
159. Basiri ME, Nemati S, Abdar M, et al. Abcdm: an attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Gener Comput Syst* 2021; **115**:279–94.
160. Cho H, Kim Y, Lee E, et al. Basic enhancement strategies when using bayesian optimization for hyperparameter tuning of deep neural networks. *IEEE Access* 2020; **8**:52588–608.
161. Jian W, Toscano-Palmerin S, Frazier PI, Wilson AG. Practical multi-fidelity bayesian optimization for hyperparameter tuning. In: *Uncertainty in Artificial Intelligence*. PMLR, 2020, 788–98.
162. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–44, 2016.
163. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell* 2019; **267**:1–38.
164. Chou Y-L, Moreira C, Bruza P, et al. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inf Fusion* 2022; **81**:59–83.
165. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion* 2021; **76**:89–106.
166. Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion* 2021; **76**:243–97.
167. Abdar M, Samami M, Mahmoodabad SD, et al. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. *Comput Biol Med* 2021; **135**:104418.
168. Abdar M, Salari S, Qahremani S, et al. Uncertaintyfusenet: robust uncertainty-aware hierarchical feature fusion model with ensemble Monte Carlo dropout for covid-19 detection. *Inf Fusion* 2023; **90**:364–81.
169. Hoffmann L, Fortmeier I, Elster C. Uncertainty quantification by ensemble learning for computational optical form measurements. *Mach Learn: Sci Technol* 2021; **2**(3): 035030.
170. Mazoure B, Mazoure A, Bédard J, Makarenkov V. Dunescan: a web server for uncertainty estimation in skin cancer detection with deep neural networks. *Sci Rep* 2022; **12**(1): 1–10.

171. Kirillov B, Savitskaya E, Panov M, et al. Uncertainty-aware and interpretable evaluation of cas9–grna and cas12a–grna specificity for fully matched and partially mismatched targets with deep kernel learning. *Nucleic Acids Res* 2022; **50**(2): e11–1.
172. Gordon J, Bronskill J, Bauer M, et al. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:180509921* 2018.
173. Lee HB, Lee H, Na D, et al. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. *arXiv preprint arXiv:190512917* 2019.
174. Nguyen C, Do T-T, Carneiro G. Uncertainty in model-agnostic meta-learning using variational inference. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3090–100, 2020.