



Review

Advances in AI-Guided CRISPR-Cas9 Engineering Strategies for Microbial Biotechnology

Javier Alejandro Delgado-Nungaray ^{1,*} , Dulce Alitzel Pérez-Ponce ¹, Luis Joel Figueroa-Yáñez ^{2,*} ,
Eire Reynaga-Delgado ³ , Mario Alberto García-Ramírez ⁴ and Orfil Gonzalez-Reynoso ¹

¹ Chemical Engineering Department, University Center for Exact and Engineering Sciences, University of Guadalajara, Blvd. M. García Barragán # 1451, Guadalajara 44430, Mexico; dulce.perez5599@alumnos.udg.mx (D.A.P.-P.); orfil.gonzalez@academicos.udg.mx (O.G.-R.)

² Industrial Biotechnology Unit, Center for Research and Assistance in Technology and Design of the State of Jalisco, A.C. (CIATEJ), Zapopan 45019, Mexico

³ Pharmacobiology Department, University Center for Exact and Engineering Sciences, University of Guadalajara, Blvd. M. García Barragán # 1451, Guadalajara 44430, Mexico; eire.rdelgado@academicos.udg.mx

⁴ Electronics Department, University Center for Exact and Engineering Sciences, University of Guadalajara, Blvd. M. García Barragán # 1451, Guadalajara 44430, Mexico; mario.garcia@academicos.udg.mx

* Correspondence: javier.dnugaray@alumnos.udg.mx (J.A.D.-N.); lfigueroa@ciatej.mx (L.J.F.-Y.)

Abstract

CRISPR-Cas9 has transformed microbial biotechnology by enabling precise genome modifications; however, achieving high editing efficiency remains a challenge due to multiple determinants, including on-target specificity, off-target events, PAM sequence, sgRNA scaffold composition, and RNA secondary structure. Our review foresees how artificial intelligence (AI) can address those challenges by enabling automated identification as well as highly active guide RNA (gRNA) optimisation. We highlight the influence of a data-driven training strategy that is focused on high-quality, diverse, and accurately labelled microbial datasets—mainly, given the limitations of models derived from mammalian systems that are not directly transferable to microbial organisms. Moreover, we discuss the key role of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and centralised, curated CRISPR-Cas databases as foundational elements for developing robust and predictive frameworks. Emerging directions are also explored, including generative AI approaches capable of supporting automated experimental planning. By considering the potential dual use of such technologies, the review further addresses bioethical considerations and regulatory frameworks necessary to ensure responsible genome engineering as a milestone, as well as the implementation of safeguards against misuse, particularly in pathogenic microorganisms. Furthermore, the convergence of standardised experimental data, specialised microbial datasets, and advanced AI architectures is paving the way to transform microbial biotechnology by accelerating metabolic engineering and synthetic biology applications.



Academic Editor: Ming Chen

Received: 25 March 2026

Revised: 25 April 2026

Accepted: 5 June 2026

Published: 24 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: AI-driven; CRISPR-Cas9; gRNA optimisation; artificial intelligence; machine learning; deep learning; microbial biotechnology; GPT; biosecurity

1. Introduction

CRISPR-Cas9 remains the most widely used genome engineering technology and has been applied across numerous microbial species due to its low cost, simple design, and experimental versatility [1–4]. The CRISPR-Cas9 system features three components: a single

guide RNA (sgRNA), composed of a guide RNA (gRNA) that specifies the target site, fused to a scaffold region, followed by a transcription termination residue; the Cas9 endonuclease; and a protospacer adjacent motif (PAM) located in the target DNA [5]. Each one of those components introduces specific challenges to achieve efficient and precise genome modification. The sgRNA activity is strongly linked to careful gRNA design to maximise on-target efficiency whilst minimising off-target events [6]. The performance is further influenced by the secondary (2D) structure of both gRNA and full sgRNA that can enhance or decrease Cas9 binding and DNA cleavage [7]. Collectively, those factors underscore the complexity of optimising CRISPR-Cas9 editing outcomes in microbial systems.

Artificial intelligence (AI) has revolutionised numerous scientific disciplines, ranging from drug discovery to biological conservation [8,9]. In the context of genome engineering, experimental design often involves repetitive and time-consuming tasks, creating an opportunity for AI-guided approaches to significantly improve the efficiency [10]. The search for swift, robust, and optimised protocols for microbial engineering has therefore increasingly relied on advanced computational methods, particularly deep learning (DL) [11]. The DL algorithms' application to large CRISPR-Cas experimental datasets has enabled the gRNAs identification with enhanced target specificity and improved editing performance [12].

AI integration into CRISPR-Cas gRNA optimisation has the potential to address major challenges in the pharmaceutical and chemical industries. One of the most pressing issues is the growing demand for valuable chemicals produced through microbial cell factories. Bacteria and yeast are widely used in metabolic engineering due to their rapid growth, relatively simple cultivation requirements, and scalability in industrial bioprocesses (Figure 1) [13,14]. However, microorganisms are considerably genetically diverse and show distinct metabolic characteristics that result in reduced efficiency or even CRISPR-Cas9 system failure in certain strains. Consequently, ongoing developments focus on improving precise on-target genome editing by minimising off-target effects and Cas9-associated toxicity, and enhancing homologous recombination efficiency in microbial systems [15].

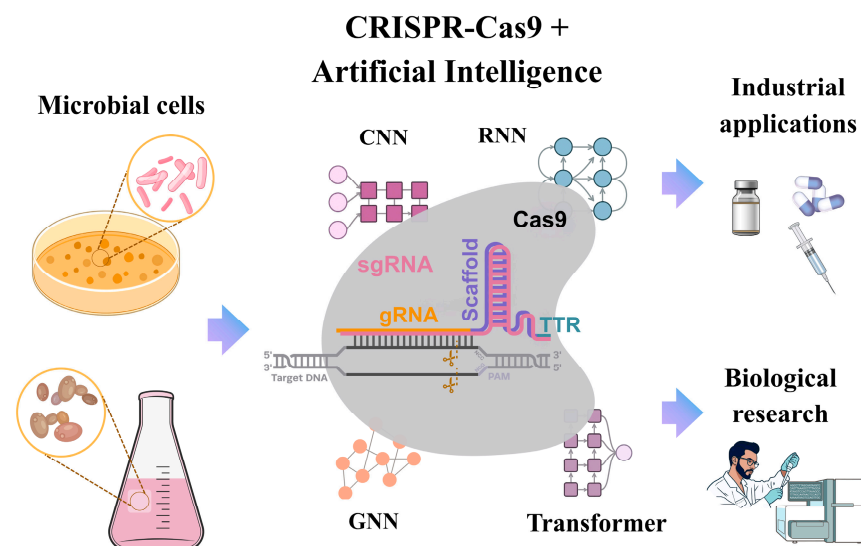


Figure 1. AI-guided CRISPR-Cas9 optimisation in microbial cells facilitates industrial applications and fundamental biological research. CNN: convolutional neural network; RNN: recurrent neural network; GNN: graph neural network; TTR: transcriptional termination residual.

The convergence of CRISPR-Cas technologies with AI has significantly advanced microbial strain optimisation by increasing the efficiency and productivity of biopharmaceutical manufacturing processes. Interdisciplinary approaches that integrate genetic engineering and bioinformatics can pose challenges related to scaling up microbial produc-

tion. In this context, AI-guided CRISPR-Cas optimisation can accelerate strain engineering strategies, reducing experimental time and improving the efficiency of microbial cell factories [16,17].

Beyond industrial applications, CRISPR-Cas also uncover key roles in fundamental biological research such as evolutionary processes, complexity of biological systems, and synthetic genomes with tailored characteristics by considering that there exist technical difficulties in gRNA design, including high GC content in many microbial genomes [18]. In addition, CRISPR-based genome-wide knockout screens using gRNA libraries have become valuable tools for identifying genes involved in antimicrobial resistance (AMR) pathways. AI-driven approaches can further enhance these strategies by improving the design and predictive performance of CRISPR technologies aimed at combating AMR [19,20].

Nowadays, AI is increasingly adopted as a core tool in the design and optimisation of CRISPR components, including on-target gRNA selection, off-target prediction, RNA 2D structure analysis, and simulation of genome editing outcomes. However, a challenge in applying these tools to microorganisms is that most existing models were trained on eukaryotic data. The untransferability of mammalian CRISPR-Cas9 activity models to prokaryotic systems stems from divergences in DNA repair pathways, genomic architecture, and cellular survival dynamics. Eukaryotic models primarily exhibit hybrid outcomes of Cas9 cleavage and non-homologous end-joining (NHEJ) repair, further confounded by complex chromatin structures and blocking effects of nucleosomes, which vary even between yeasts and humans. Prokaryotes, on the other hand, lack robust NHEJ machinery and rely primarily on homologous recombination [21]; CRISPR-Cas9 efficiency is manifested as double-strand break (DSB)-induced toxicity, which provides a cleavage activity measurement devoid of mammalian repair biases. Furthermore, bacterial DNA accessibility is thought to be influenced by distinct chromosomal factors, including DNA supercoiling, nucleoid-associated proteins, and torsional constraints. Consequently, the reliance of mammalian models on eukaryotic-specific features makes them mechanistically incompatible with the prokaryotes, leading to the low predictive performance seen in organisms like *Escherichia coli* [22,23].

AI integration into CRISPR-Cas workflows enables key rational experimental planning whilst reducing development time, cost, and trial-and-error approaches. In this context, our review examines the current challenges associated with CRISPR-Cas gRNA design and highlights how AI-driven approaches can improve genome editing performance in microbial systems. Particular emphasis is placed on the need for microbial organism-specific datasets, standardised data frameworks, and FAIR-compliant repositories to support robust predictive models. By focusing on microbial genome engineering, the review outlines how the convergence of AI, CRISPR-Cas technologies, and high-quality data infrastructures can accelerate the development of predictive and scalable strategies for microbial biotechnology.

2. CRISPR-Cas gRNAs: Opportunities for AI Optimisation

2.1. gRNA: On- and Off-Target Events

The gRNA ranges from 18 to 22 nucleotides (nt) in length, with a 20-nt sequence most commonly used to balance editing efficiency and target specificity. gRNAs longer than 20 nt exhibit reduced effectiveness, whilst the five nt at the 3' end of the gRNA (positions 16–20 nt), known as the seed sequence, play a key role in target recognition [24–26].

The rational gRNA design started for application in mammalian cells, as shown by Doench et al. [27] who created a predictive model for efficient gRNAs using a base of 1841 sequences to obtain highly active gRNAs for human and mouse genes. The linear-sparse-modelling implementation approach by using an L1-regularised Support Vector Machine (SVM) for feature selection, followed by a logistic regression classifier, including

GC content and position-specific nt. Those are the backbone of actual experimentation and are in continuous use as an efficiency score in CRISPR-Cas tools [28]. Another SVM application was conducted to predict the efficacy of sgRNAs targeting 400 essential non-ribosomal genes with a 73,000 sgRNA library for human cells [29].

Nucleotide preferences along the gRNA sequence have been reported to exhibit position-dependent biases associated with Cas9 activity (Figure 2). In particular, guanine (G) is strongly preferred at positions N-19 and N-20, whilst cytosine (C) is strongly unfavourable at N-20 but favoured at N-18 and N-19 [30]. Uracil (U) is disfavoured at N-18 and N-19 due to its potential to trigger premature transcription termination. N-16 favours C over G, whereas C is commonly unfavoured at N-3. In contrast, adenine (A) is more frequently observed in the central gRNA region [27]. Observations of position-specific nucleotides from microbial data, specifically from *E. coli*, have shed light on the seed region, where UA, CU, CG, and GC dinucleotides have been reported to have beneficial effects on gRNA activity, whereas GG and CC dinucleotides display inhibitory effects [22].

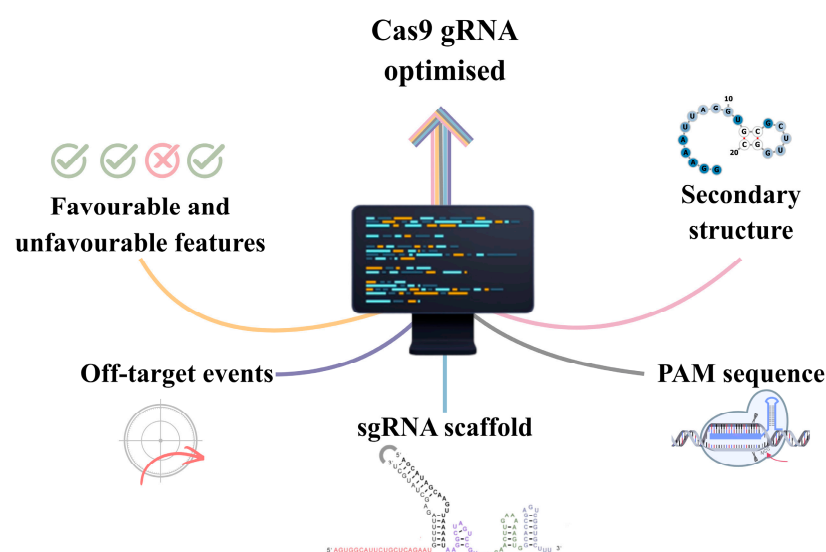


Figure 2. Features to be considered for Cas9 gRNA optimisation when using AI models.

The mechanistic explanation of position-specific nucleotide preferences relies on biophysical features, such as melting temperature, self-folding free energy, and quantum chemical properties, which directly govern gRNA efficiency. These properties include the highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) energy gap and hydrogen-bonding energy. Such associations between AI outputs and underlying biological mechanisms are increasingly established as CRISPR-Cas models move beyond correlation-only interpretations through explainable AI (XAI) methods. For instance, employing iterative Random Forest (iRF) has allowed the identification of traits for sgRNA design in *E. coli*. Specifically, G at the N-20 position has been associated with gRNA efficiency due to its HOMO–LUMO energy gap and hydrogen bonding energy. Mechanistically, once the PAM sequence is identified and bound, the DNA kinking enables helix unwinding to permit DNA–gRNA binding. These structural events are stabilised by a phosphate lock loop proximal to the PAM; a high HOMO–LUMO gap in this region indicates molecular stability, whilst weaker hydrogen bonding facilitates the subsequent DNA double helix unwinding [31].

Furthermore, the use of Tree SHAP (SHapley Additive exPlanations), an approach integrating the SHAP algorithm with XGBoost, has allowed for the quantification of the exact contribution of each position-dependent nucleotide to the final activity score. This methodology revealed that the preference for purines over pyrimidines is primarily driven

by the intrinsic binding energy of the Cas9-gRNA complex [32]. Within the PAM-proximal region, specific nucleotides promote R-loop formation and stabilise the gRNA-DNA heteroduplex through interactions with the arginine-rich bridge helix of Cas9 [33]. These PAM-proximal effects are mediated by the unidirectional nature of R-loop zipping, whereas PAM-distal effects are caused by Cas9 conformational changes and kinetic barriers during the final stages of activation, as identified via random forest (RF) analysis and free energy landscape [34]. Local features like GC content impact both heteroduplex stability and gRNA 2D structure. Extreme GC content and complex 2D structures can physically block the seed region, preventing the gRNA from recognising its target, whilst a GC content below 30% results in duplex instability [33,35].

Off-target activity is quite keen to follow reproducible, learnable patterns rather than those that randomly occur. Cas9, in particular, tends to tolerate mismatches at the 5' end of the gRNA, whereas those near the 3' end adjacent to the PAM are less tolerated and more disruptive to cleavage. The AI-driven integration methodology, thus, enables genome-wide screening for potential unintended targets [36].

The DL robustness approaches have proven to surpass traditional ML algorithms, including SVM, regularised linear regression models, elastic net, and RF, to predict gRNA activity [37]. Hybrid architectures combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been widely adopted, with predicted scores showing strong correlations with sgRNA activity [38]. Whilst RNNs are effective at capturing long-range dependencies within nucleotide sequences, which are commonly applied in genomics, CNN-based models have been reported to use alternative approaches to identify highly active sgRNAs and to accurately predict target efficacy [33,39]. Within the RNN family, bidirectional long short-term memory (BiLSTM) networks have also been applied to predict gRNA activity across wild-type *Streptococcus pyogenes* Cas9 (SpCas9) and high-fidelity variants such as eSpCas9(1.1) and SpCas9-HF1. The integration of these diverse DL architectures into predictive tools underscores their versatility in facilitating efficient gRNA design, as detailed in Table 1 [32].

Table 1. ML/DL Tools for gRNA design.

Tool Name	Associated Algorithms	Core Features & Training Strategy	Strengths	Limitations	Target Species	Study
CRISPRon	CNN + Feedforward Layers	Appends thermodynamic parameters (RNA-DNA binding energy) directly into the network	Capture complex sequence features	Limited to on-target prediction	Human, Mouse, and Zebrafish	[11]
DeepSpCas9	CNN (Multiple Filter Sizes)	End-to-end learning trained on a massive, uniformly generated high-throughput dataset	High generalisation performance; trained on direct DNA cleavage frequencies	Restricted to wild-type SpCas9 on-target activity	Human	[37]
DeepCas9	1D-CNN	One-hot encoding of 30-nt sequences; automated spatial feature extraction	Reliable capture of complex sgRNA sequence patterns	Susceptible to biases introduced by data inconsistencies	Human and Mouse	[33]
DeepHF	BiLSTM + Dense Layers	Combines sequential memory embeddings with hand-crafted biological/thermodynamic features	Demonstrates high accuracy for high-fidelity Cas9 variants	Lacks generalisation to other Cas orthologs	Human	[32]

Table 1. Cont.

Tool Name	Associated Algorithms	Core Features & Training Strategy	Strengths	Limitations	Target Species	Study
CRISPR-Learner	CNN	Transfer learning capabilities; dynamic zero-padding for variable-length sequences	Supports custom dataset training for gRNA design	Currently restricted to the assessment of on-target cleavage efficiency	Human and Mouse	[40]
DeepCRISPR	Deep Belief Network (DBN)/ Autoencoder + CNN	Unsupervised pre-training on unlabelled data; integration of epigenetic features	Unifies on-target and off-target prediction within a single computational framework	Strictly limited to NGG-based SpCas9 systems. Currently validated only for human genomic data	Human and Mouse	[41]

The DL approaches can assist in identifying key features associated with gRNA efficiency by enabling the single-nucleotide position analysis as well as di-, tri-, and higher-order combinatorial nucleotide patterns that collectively influence gRNA performance; however, the biological relevance of those predictions requires functional validation in the CRISPR-Cas activity context [42]. Persistent limitations in gRNA datasets availability further constrain model development, as reliable discrimination between true functional hits and false positives often depends on the multiple gRNAs' inclusion per target gene, whilst the exclusion of ineffective gRNAs does not compromise predictive accuracy—thereby constraining the development of robust predictive models [43]. Moreover, many established gRNA design rules have been derived predominantly from human, mouse and zebrafish genomic data. Those are raising concerns regarding the straight transferability to microbial genome engineering, where distinct genomic architectures and regulatory contexts may negatively affect predictive performance [11,27,41,44,45].

2.2. sgRNA Scaffold

In CRISPR-Cas genome engineering, the sgRNA scaffold comprises 76 nucleotides (GTTTGTAGAGCTAGAAATAGCAAGTTAAATAAGGCTAGTCCGTTATCAACTT-GAA AAAGTGGCACCGAGTCGGTGC). Those are organised into five functional modules: lower stem, bulge, upper stem, nexus, and hairpins (Figure 3). The bulge and nexus are highly conserved, whereas engineering efforts have predominantly focused on the lower stem, upper stem, and hairpin modules [46,47]. It represents a critical determinant of Cas9 activity, as it mediates the interaction with the Cas9 nuclease [48]. Extensive sgRNA scaffold variants have been shown to retain DNA cleavage and editing efficiencies comparable to the wild-type Cas9 scaffold despite substantial sequence alterations. It does indicate that only a limited number of nucleotides within the 2D structure are highly conserved by highlighting the intrinsic structural malleability that enables the rational design of scaffold variants to enhance genome editing performance [5].

Systematically, sgRNA scaffold variants and modification strategies have been comprehensively examined in the review by De Saeger from an engineering point of view [49]. In here, the first dedicated sgRNA scaffold resource is introduced, comprising 230 scaffold variants, providing a foundational dataset for the emerging application of AI-guided optimisation of CRISPR-Cas efficiency. Despite this progress, formalised design rules for the scaffold region remain largely undefined, representing a clear DL approach opportunity.

2.3. RNA Secondary Structure

RNA secondary (2D) structure is as fundamental as nucleotide sequence composition in gRNA design, as it determines gRNA–DNA interactions and genome editing efficiency.

Key structural features must be considered at two levels: the gRNA 2D structure and the overall sgRNA folding. For active gRNAs, reduced self-folding potential during the design and selection process is essential, as the formation of stable hairpin structures can impede hybridisation with the target DNA sequence [7]. gRNA accessibility has a leading role in target recognition, with favourable minimum free energy (MFE) values within the range of -3.30 to 0.00 kcal·mol⁻¹, whereas MFE values below -5 kcal·mol⁻¹ have been associated with reduced gRNA activity [30,35].

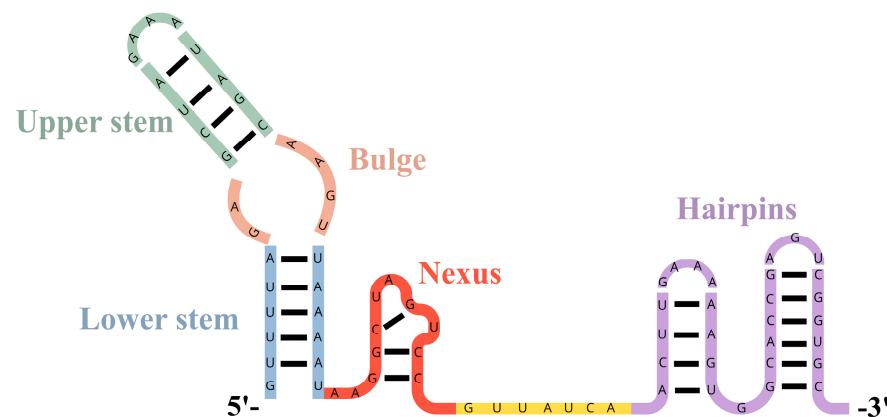


Figure 3. Structural architecture and functional modules of the sgRNA scaffold. Colour-coding: lower stem (blue), bulge (orange), upper stem (green), the nexus (red), and hairpins (purple).

In addition to gRNA folding, sgRNA 2D structure has been shown to affect CRISPR activity, particularly through interactions between the gRNA seed region and scaffold sequence. Unpaired configurations between seed positions 18–20 and scaffold positions 51–53 are considered favourable, and those are incorporated as predictive features in SVM models for CRISPR-Cas activity prediction [50].

For this reason, sgRNA 2D structure engineering has emerged as an additional strategy to enhance Cas9 specificity. As studies have shown that the hairpin design structure at the 5' end of the sgRNA can impede R-loop formation, a process required for the conformational Cas9 activation, whilst simultaneously suppressing off-target activity [51].

Given the pivotal role of 2D structure in gRNA design, optimisation approaches such as Elitist Genetic Algorithms (EGAs) have been applied to improve gRNA arrays in *E. coli*, where best-performing solution preservation across generations, combined with selection based on the lowest MFE values and 2D structure visualisation, has demonstrated the capacity of algorithmic methods to enhance gRNA design, thereby motivating the adoption of more advanced ML and DL frameworks for systematic and scalable optimisation [52].

By building on this expanded understanding, DL has emerged as a key driver for next-generation gRNA design by capturing complex, non-linear relationships between gRNA sequence, 2D structure, and editing outcomes. As shown elsewhere, graph-based representations combined with DL, particularly graph neural networks (GNNs) and graph attention networks (GATs), have demonstrated improved gene-editing efficiency prediction by jointly integrating structural and sequence information, often outperforming baseline approaches in CRISPR-Cas9 [53]. Collectively, those advances highlight the structure-aware potential of graph-based DL frameworks to enhance predictive accuracy and adaptability in genome engineering.

2.4. PAM Sequence

Another key determinant of gRNA efficacy is the presence of a protospacer-adjacent motif (PAM), a short DNA sequence flanking the target site and required for Cas9 recognition [54]. For SpCas9, the canonical 5'-NGG-3' PAM (N denotes any nucleotide) is

associated with the highest proportion of active sgRNAs, whereas non-canonical PAMs such as NAG, NCG, and NGA exhibit substantially lower activity. Although PAMs are not part of the gRNA sequence—nucleotide preferences at the N position—it does influence well recognition, where C is favoured and T is not [27,43]. Importantly, alternative PAMs must be considered during gRNA selection, as Cas9 preferentially localises to PAM-rich genomic regions through 3-D collisions and lateral diffusion along DNA, thereby contributing to potential off-target interactions [55].

Efforts to characterise PAM recombination at the level of the PAM-interacting domain of Cas9 have highlighted the potential for machine learning-driven protein engineering. Restricted Boltzmann machine (RBM) coupled with constrained Langevin dynamics has been used to model Cas9 sequence-function relationships, generating variants with improved activity compared to wild-type SpCas9 [56]. Those findings depict the integrating machine learning promise of structure-informed, physics-based modelling to engineer Cas9 variants with altered PAM specificities whilst maintaining, or even enhancing, catalytic efficiency. By supporting such direction, large-scale resources such as CRISPR-Cas Atlas have emerged, providing a comprehensive dataset that includes PAM information alongside other CRISPR-Cas features by enabling the development of predictive models for PAM preference [57].

2.5. Cas9 mRNA with Shine-Dalgarno Sequence

Following gRNA design, the Cas9 delivery choice strategy represents a critical step in microbial genome engineering. Plasmid-based CRISPR-Cas9 systems remain the most widely used approach due to the stability and low cost; however, an emerging and unexplored alternative is delivery via messenger RNA (mRNA) that enables transient Cas9 expression and potentially reduces off-target events [58].

When Cas9 is delivered as mRNA, the translation process becomes a central consideration. Whilst eukaryotic translation initiation is primarily driven by 5'-cap recognition through the eIF4F complex, with the 3'-poly(A) tail contributing indirectly through interactions between eIF4G and the poly(A)-binding protein (PABP; Pab1 in yeast or PABPC1 in humans), bacterial translation relies on the presence of a Shine-Dalgarno (SD) sequence, with the canonical 5'-UAAGGAGGU, to direct ribosome recruitment [59–62]. This distinction is critical because commercially available Cas9 mRNA products are predominantly optimised for eukaryotic expression, especially for human and mouse cells, and those are, therefore, not directly transferable to bacterial hosts.

ML approaches have been applied to predict translation initiation rates through the rational design of ribosome-binding sites (RBSs). Methods based on Gaussian process regression and upper confidence bound multi-armed bandit algorithms have been used to model interactions between engineered RBS sequences and the anti-Shine-Dalgarno (aSD) region of the 16S rRNA, enabling the design of RNA sequences tailored to host-specific translational requirements [63].

3. Data-Driven Training for Microbial Systems and AI Tools for sgRNA Design

3.1. Data Collection and Preprocessing

In CRISPR-Cas9 research, the predictive accuracy of learned functions depends directly on the volume and quality of the training datasets, whilst AI's capacity to iteratively learn from these data further enhances its utility in addressing complex biological problems. Unlike other disciplines, large-scale data acquisition in microbial engineering is frequently quite expensive and experimentally demanding. These constraints are particularly relevant for DL approaches that require extensive data. Consequently, DL applications in micro-

bial CRISPR-Cas research must address challenges such as data imbalance, uncertainty, overfitting, vanishing gradients, and scalability limitations [10,64].

Data collection can be challenging because relevant information is frequently dispersed across scientific publications and heterogeneous experimental reports. CRISPR-Cas experiments often generate complex datasets that can impact feature extraction because information can be represented in categorical or numerical data [65]. The process involves systematic filtering, annotation, and raw data organisation to ensure that curated datasets accurately represent microbial genetic and phenotypic variability [66].

Before model training, careful preprocessing is required by removing low-quality or ambiguous sequences, normalising experimental measurements, and encoding relevant sequence and contextual features [65]. Those procedures are essential for reducing noise, correcting inconsistencies, and improving the reliability of downstream analyses and AI model performance [67,68].

For CRISPR-Cas applications, preprocessing starts by converting gRNA sequences into fixed-dimensional feature vectors that can be interpreted by predictive models. Common encoding strategies include one-hot encoding and word-embedding approaches. In one-hot encoding, each nucleotide is represented by a binary matrix with rows linked to the four nitrogenous bases and the columns representing sequence positions. Alternatively, embedding methods transform nucleotide sequences into higher-dimensional vectors that capture contextual relationships between bases [69].

Furthermore, the preprocessing steps may involve aligning gRNA sequences to reference genomes, scripting genomic features, and gathering key information related to potential off-target sites. Dataset normalisation as well as filtering procedures are applied to reduce experimental biases and improve dataset consistency. For example, preprocessing pipelines may exclude high variability in efficiency across experimental conditions, genes with insufficient gRNA representation, or lack the canonical 5'-NGG-3' PAM required for SpCas9 activity [11,70].

3.2. Data Labelling and Objective Definition

In supervised learning strategies, every data instance must be associated with a defined output label that determines the model's predictive objective. Within CRISPR-Cas applications, those labels normally represent gRNA editing efficiency, quantified as the proportion of insertions and deletions (indels) generated at the target locus, or phenotypic outcomes, such as growth rate and fitness under specific conditions [71].

Editing efficiency is typically categorised by using predictive models as both "effective" or "ineffective" ones based on predefined activity thresholds through binary classification. Labels are derived, primarily, from the high-throughput quantification of surrogate target sites through targeted deep sequencing [11,70]. However, to better capture the specific repair-outcome frequency, labels should be redefined as multi-dimensional vectors for capturing the frequency of specific repair outcomes, by considering the DNA repair pathway following the introduction of a double-strand break (DSB) by Cas9 [72]. Lately, the label choice dictates the modelling strategy: regression algorithms are employed for continuous efficiency prediction, whereas classification models are applied to assign guides to discrete activity categories. Precise objective definition is therefore key to ensuring that AI-guided gRNA design tools align to the intended application in microbial biotechnology [73].

3.3. Data Diversity and Representativeness

A key challenge in developing AI models for microbial CRISPR-Cas applications is ensuring adequate biological diversity along training data. Nowadays, a large majority of training datasets for predictive tools are harvested from a limited number of cell lines due to

a “genomic representativeness” pipeline; however, the approach must shift towards the use of specific genomes for off-target analysis [70,74]. The limitation affects cell-type specificity, as microorganisms vary in genome composition and cellular contexts. Consequently, models trained exclusively on data from a single species exhibit limited extrapolation for optimisation. Expanding datasets to include multiple species is therefore essential to improve model robustness [11].

To reduce the non-scientific overfitting tool to specific experimental conditions, independent test sets should be reserved for external validation. It is well established that models that demonstrate strong internal cross-validation performance may show reduced accuracy when applied to entirely new datasets. Such observations underscore the importance of using broad, diverse training data and performing rigorous evaluations through independent datasets [27,45].

3.4. Balancing Quantity and Quality

Effective AI model development for CRISPR-Cas applications requires a careful balance between dataset size and data reliability. Predictive tool development often involves a trade-off: whilst massive high-throughput screening (HTS) datasets provide thousands of data points, those frequently suffer from significant noise. HTS data, although voluminous, are often highly imbalanced and prone to experimental artefacts such as errors introduced during library synthesis or the inclusion of low-efficiency gRNAs [74]. Models trained on noisy HTS data without proper denoising may learn that these artefacts are a sort of biological signal, leading to elevated variance and poor generalisation to unseen, high-quality data [70].

Data curation requires taking into consideration the input feature selection. Although DL models can extract patterns directly from raw sequence data, incorporating biologically informative features can enhance performance, particularly when training datasets are size-limited—such as the situation encountered in microbial systems [75]. High-quality datasets should satisfy specific quality dimensions by including accuracy, completeness, consistency, timeliness, uniqueness, and validity. These dimensions in CRISPR-Cas may correspond to parameters such as indel frequency, PAM sequence and cell type, gRNA efficiency, genome annotation updates, duplicate gRNAs, and syntax rules (e.g., the canonical 20 bp length). Furthermore, AI-driven data quality assessment methods, such as Isolation Forest and data-centric AI approaches, can be applied to identify inconsistencies and improve dataset reliability across multiple quality dimensions [11,41,43,76,77].

When there is a lack of experimental data, an available option is to expand the datasets to improve model learning and performance artificially. For example, DeepCRISPR generated approximately 200,000 additional synthetic gRNA sequences from an initial set of 15,000 experimentally validated gRNAs by introducing controlled sequence variations. The approach increased dataset diversity without requiring additional laboratory experiments, enabling the model to capture more generalisable sequence patterns [41]. Despite these advantages, data quality must remain a priority. Augmented data should be biologically plausible and consistent with established molecular constraints to prevent noise or bias introduction that might influence the predictive accuracy [65].

In the microbial context, ensuring such biological plausibility is particularly demanding, as microbial CRISPR-Cas data acquisition is fundamentally bottlenecked by DSB-induced lethality. The general absence of NHEJ in most prokaryotes results in a binary survival/death outcome, rather than the diverse indel repair patterns observed in eukaryotes [21–23,31]. This “all-or-nothing” data is further complicated by transformation recalcitrance in non-model species, where low transformation efficiencies lead to library

dropout and skewed datasets. This necessitates the use of compact libraries or inducible systems to prevent the premature loss of essential gene targets during cloning phases [78–80].

Training data accuracy is further eroded by escape mutants and survivor bias; cells can evade CRISPR-mediated death through PAM-escape mutations, plasmid loss, or the RecA-mediated SOS response, which increases hypermutation rates and leads to false-negative labels for active gRNAs [22,81]. Additionally, extreme taxonomic heterogeneity, including differences in GC content, DNA supercoiling, and epigenetic modifications like adenine methylation, creates resistant genomic loci [22,80,82]. Consequently, whilst mammalian pipelines analyse a “hybrid output” of cutting and repair, microbial data generation must overcome the confounding overlap between true targeting activity and gRNA-intrinsic cellular toxicity, both of which complicate the development of robust microbial CRISPR-Cas datasets [22,23].

3.5. Model Selection and Training Data

For relatively simple prediction tasks or limited datasets, classical ML methods such as linear regression, SVM, and GBRT have been widely applied in CRISPR-Cas efficiency prediction. Such models effectively drive structured, tabular inputs and can capture interactions among engineered sequence items [26,43].

As the dataset size and complexity get bigger, DL approaches become advantageous. Multilayer neural networks can automatically learn hierarchical representations from raw input data by reducing the need for extensive manually improved engineering [83]. CNNs, for example, can process encoded nucleotide sequences directly and identify relevant patterns such as sequence motifs and positional dependencies [37].

A set of different strategies is used for training data selection and optimisation, including the use of full data, random and stratified sampling, information-theoretic feature selection, data augmentation, active learning, and transfer learning [84]. Those approaches have been applied in the prediction and optimisation of gRNA. For instance, CRISPRon [11] and CRISPRon-ABE/CBE [74] combined full data utilisation with the integration of newly generated datasets and previously published experimental data. Random and stratified sampling strategies have been implemented in model training; Yang et al. [69] applied a StratifiedKFold cross-validation strategy featuring five folds to mitigate dataset variability when training a hybrid neural network model, whereas Kimata et al. [85] addressed class imbalance in CRISPR-Cas9 off-target datasets by performing random down-sampling of the negative class along model training.

Trivedi et al. [86] showed that augmenting imbalanced CRISPR-Cas training sets with synthetic sgRNAs through random single-nucleotide substitutions in the non-seed region did improve sgRNA activity prediction for yeast species. In addition, the transfer learning strategy has been used to leverage knowledge from large experimental datasets; Yaish and Orenstein [87] trained models on large in vitro CHANGE-seq data and then fine-tuned them using GUIDE-seq datasets to predict CRISPR-Cas9 off-target sites with bulges.

3.6. Microbial Datasets and AI Tools-Associated

A major challenge in gRNA design for microbial engineering is that widely available datasets have been generated within non-microbial systems, which results in a limited microbial-specific training data [27,29,43,88]. A retrospective analysis by Moreb and Lynch [89] evaluated sequence features across 44 published CRISPR-Cas datasets spanning multiple species (including human, mouse, zebrafish, *E. coli*, and *Yarrowia lipolytica*). The analysis showed that variability in gRNA activity cannot be attributed to any single sequence feature, but differences in species, cell type, genomic context, and screening method introduce batch effects that are not readily captured by simple predictive tools.

Consequently, the direct application of generalised human-derived models to microbial systems is scientifically unsound, as the apparent “rules” governing CRISPR-Cas activity are biologically context-dependent, with editing outcomes influenced by factors beyond the immediate target sequence. Whilst species-specific prediction models may improve accuracy, cross-species generalisation remains a challenging opportunity and requires careful validation of sequence-based features. In this context, DL-based approaches offer promise for improving species-aware gRNA prediction by capturing complex genomic context effects [90].

The rapid development of CRISPR-Cas technologies has led to the generation of numerous gRNA design tools aimed at maximising on-target efficiency whilst minimising off-target events, including CHOPCHOP, WU-CRISPR, E-CRISP, CRISPR-ERA, CRISPOR, GuideScan, Cas-OFFinder, CRISTA, CRISPR-P, and CRISPRz [72,91]. However, those tools have been developed, validated, and optimised for use in human, mouse, zebrafish, plant, and crop experiments, thereby limiting their direct applicability to microbial genome engineering. Although microbial CRISPR-Cas datasets remain comparatively scarce, recent efforts have begun to establish foundational resources for species such as *Citrobacter rodentium*, *E. coli*, *Komagataella phaffii*, and *Y. lipolytica* (Table 2).

Table 2. Microbial CRISPR-Cas datasets and associated AI tools.

Microorganism Species	Library Size	Associated Algorithms	Associated AI Tools	AI Tool Strengths	AI Tool Limitations	Study
<i>Citrobacter rodentium</i>	31,796	Transfer learning/ CNN-RNN	crisprHAL	Distinguishes on-target cleavage from cellular toxicity and generalises predictive performance across diverse bacterial species	Necessitates transfer learning from larger datasets to ensure high performance	[23]
<i>E. coli</i> K12 MG1655	65,928	Gradient boosting regression tree (GBRT)	sgRNA-cleavage-activity-prediction	Eliminates mammalian-specific biases like chromatin noise and DNA repair preferences	Displays reduced resolution in DNA-repair-deficient backgrounds	[22]
		CNN	DeepSgRNABacteria	Captures critical flanking sequence information	Lacks cross-domain validity between prokaryotes and eukaryotes	[21]
<i>E. coli</i> strains MG1655 and BW25113	~10,000	N/A	N/A	N/A	N/A	[81]
<i>E. coli</i> W (ATCC 9637)	6044	N/A	N/A	N/A	N/A	[92]
<i>K. phaffii</i> GS115	31,984	N/A	N/A	N/A	N/A	[93]
<i>Y. lipolytica</i> strain PO1f	46,234	CNN	DeepGuide	Accurately predicts high-activity Cas9 and Cas12a gRNAs in specific organisms like <i>Y. lipolytica</i> . Incorporates genomic context and epigenetic features to improve targeting precision	Struggles to identify low-activity gRNAs for Cas9 with accuracy	[94,95]

N/A: Not applicable.

Notably, only two AI-guided gRNA prediction tools, crisprHAL and DeepGuide, have been specifically developed for microbial systems, targeting bacterial and fungal species, respectively, by leveraging organism-specific libraries for model training [23,95]. The limited landscape underscores both the current constraints and the substantial opportunity to expand data-driven for microbial biotechnology. The novel tools' development presents a significant opportunity to integrate scientific efforts in diverse fields, from biotechnology to computer science. Initiating studies focused on data generation will ultimately enable AI tool optimisation, providing deeper biological insights into CRISPR-Cas9 in bacteria or yeast, whilst generating translational knowledge. Such improvements might enhance biotechnological processes, fostering mutually beneficial collaborations between academia and industry.

4. Applied Examples of CRISPR-Cas9–AI in Microbial Biotechnology

The field of microbial biotechnology has been significantly advanced by the synergy between CRISPR-Cas9 and Artificial Intelligence (AI). This integration facilitates the sophisticated engineering of biological systems, transforming microorganisms into efficient cell factories. By steering microbial metabolism toward the production of specific high-value products, including biofuels; pharmaceuticals; and fine chemicals such as fatty acids, amino acids, and flavanones; targeted gene knockouts can enhance the productivity and industrial robustness of these strains [96]. Currently, multiplexed genome editing has accelerated the development of strains with multiple desirable traits, allowing for the swift creation of optimised microbial cell factories through the simultaneous targeting of multiple genes [97].

For instance, *Pseudomonas putida* was engineered for isoprenol production, a precursor to the sustainable aviation fuel 1,4-dimethylcyclooctane, using CRISPR with a catalytically inactive Cas9 (dCas9). This system employed multiplexed sgRNA arrays to simultaneously downregulate up to four gene targets, with gRNAs designed using CRISPOR. An AI-driven process identified 67 validated genes across energy, lipid, and carbohydrate/amino acid metabolisms. Specifically, the study utilised the Automated Recommendation Tool (ART), an ensemble modelling framework incorporating seven models, including Neural Regressor, Random Forest Regressor, Support Vector Regressor, Kernel Ridge Regressor, k-Nearest Neighbour Regressor, Gaussian Process Regressor, and Gradient Boosting Regressor. This approach identified beneficial perturbations and recommended new sgRNA combinations, resulting in a 5-fold increase in isoprenol titer and demonstrating AI precision in target identification for genome engineering [98].

In *Saccharomyces cerevisiae*, CRISPR-Cas9 was employed for the marker-less integration of multiple gene cassettes into the chromosome to design polyhydroxyalkanoate (PHA) synthases. Two novel AI-designed PHA enzymes (PhaC), PhaC_{VAE1} and PhaC_{VAE2}, were confirmed active and yielded poly(hydroxybutyrate) levels of 6.2% and 4.5% of cell dry weight, respectively. The AI component utilised a conditional variational autoencoder architecture, incorporating BiLSTM layers and multi-head self-attention blocks to design functional PhaC variants capable of polymerising R-hydroxyacids. This AI-guided protein engineering enabled a high success rate for functional enzyme design, allowing for precise, low-throughput CRISPR-mediated integration that bypassed the labour-intensive trial-and-error characteristic of traditional experimental approaches [99].

S. cerevisiae has also been engineered to enhance bioethanol production through AI-driven metabolic flux optimisation. By correlating metabolic flux data from a genome-scale metabolic model with empirical ethanol yields, gene targets for knockout were identified and ranked. An ML-based prediction model was established using flux balance analysis, where linear models, specifically Automatic Relevance Determination and Bayesian Ridge regression, showed superior predictive performance. CRISPR-Cas9 was then used to per-

form single and multi-gene knockouts, targeting subunits of succinate dehydrogenase and glycerol-3-phosphate dehydrogenase. The resulting double-knockout strain $\Delta gpd2\Delta sdh6$ showed a 27.9% improvement over the wild-type, achieving an ethanol concentration of 6.06 g/L [100].

Finally, the oleaginous yeast *Y. lipolytica* was modified using a CRISPR-Cas9 and scarless promoter replacement strategy. Using gRNAs designed using CHOPCHOP, the system enabled high-throughput tuning of gene expression by replacing native promoters with a library of varying strengths or by deleting them entirely. This strategy was applied to optimise the production of betanin, a red-violet food colourant. By targeting 56 transcription factors, modifications were identified that increased betanin titers to 188 mg/L, exceeding the 99 mg/L produced by the control strain [101].

To bridge the gap between these AI-guided sequence designs and their phenotypic realisation, droplet microfluidics has emerged for high-throughput functional validation. These automated systems enable the parallelisation of hundreds of CRISPR-mediated modifications by interfacing with robotic liquid handling. In *E. coli*, such platforms facilitate the fast assessment of complex physiological phenotypes, including the enzymatic activity of galactokinase and the production of indigoidine. Ultimately, the integration of microfluidic hardware with AI-guided design enhances the predictability of microbial engineering, allowing for the quick scaling of genetic modifications [102].

As demonstrated by these cases, the application of CRISPR-Cas9-AI is expanding across the production of diverse metabolites. Whilst current applications remain specialised, the transition toward integrated, automated pipelines is poised to reduce development timelines and achieve higher product yields for industrial-scale biotechnology.

5. Challenges and Future Perspectives

5.1. FAIR Principles for CRISPR-Cas Data

Data-intensive scientific fields that include CRISPR-Cas research have an opportunity to accelerate discovery, evaluation, and reuse through robust data management. The foundation of effective data stewardship is encapsulated in the FAIR principles: Findability, Accessibility, Interoperability, and Reusability [103]. Transparent reporting of experimental conditions is essential to enhance reproducibility and comparability across studies [104]. However, adherence to FAIR principles must extend beyond technical accessibility to emphasise data quality and the appropriateness of reused datasets, as these factors directly influence interpretability [105].

The FAIR principles' implementation in CRISPR-Cas workflows strengthens genome engineering efforts by promoting standardised data acquisition, transparent analytical pipelines, and structured documentation of genetic components and experimental variables. Such practices enhance reproducibility, enable cross-platform comparability, and support long-term data reuse. Consequently, FAIR-compliant CRISPR-Cas data frameworks provide a foundation for robust AI model training, rational design strategies, and scalable microbial engineering applications [106].

However, the implementation of FAIR data principles in CRISPR-Cas research and microbial genome engineering remains challenged by metadata fragmentation and the absence of a universal ontology for hidden experimental variables. Whilst datasets may be findable, their reusability is often compromised because critical parameters, such as specific Cas9 nuclease variants, sgRNA scaffold sequences, and precise environmental conditions, are inconsistently reported [32,107,108]. This technical bottleneck is further exacerbated by poor phenotype interoperability, as CRISPR-Cas activity is quantified using incompatible metrics, ranging from indel frequencies to cellular survival rates in prokaryotic lethality assays or fluorescence intensity shifts in reporter systems [31,37,41]. Such

heterogeneity necessitates complex, ad hoc computational normalisation strategies which may inadvertently introduce biases and obscure the underlying mechanisms governing cleavage efficiency [31,108].

Furthermore, a negative result gap persists due to systemic publication bias that heavily favours successful genomic edits. As a result, most public repositories lack standardised data for failed gRNAs, depriving AI models of the key training instances needed to learn inhibitory features. In addition, challenges such as data heterogeneity, sparsity, and imbalance, and the influence of epigenetic features on sgRNA efficacy, complicate model development [35,41]. Collectively, these limitations restrict access to comprehensive data landscapes and hinder the development of a highly predictive and generalisable framework for microbial biotechnology.

5.2. CRISPR-Cas Databases for Precision Engineering

5.2.1. Dataset Standardisation

A major limitation in current CRISPR-Cas technologies is the lack of standardised approaches for detecting, quantifying, and reporting editing outcomes, particularly gRNA activity, along with associated confidence levels and experimental metadata. This lack of harmonisation directly affects the data comparability, reproducibility, and the performance of ML/DL models trained on heterogeneous datasets [109]. To address this challenge, future CRISPR-Cas repositories should adopt standardised data formats in accordance with FAIR principles, controlled vocabularies, and comprehensive metadata to ensure that archived datasets are machine-readable [110].

To achieve such a goal requires consensus on the minimal metadata fields necessary for AI-guided analyses. An illustrative example is the Biological General Repository for Interaction Datasets Open Repository of CRISPR Screens (BioGRID ORCS), which implements a structured reporting system known as Minimal Information About CRISPR Screens (MIACS) [111]. The metadata checklist includes the descriptors: screen name, organism, screen rationale, experimental setup, duration, condition name/condition dosage, multiplicity of infection (MOI), CRISPR library name/type, screen format, Cas variant, cell line/type, phenotype, analysis method, and significance threshold. Standardised reporting frameworks of this type provide a robust foundation for developing reliable AI tools and improving cross-study integration in CRISPR-Cas research.

5.2.2. Creation of Databases for CRISPR-Cas Technology

The value of specialised, purpose-driven databases (DBs) has already been demonstrated in other biological fields, where resources such as BioCyc [112], BRENDA [113], ExplorEnz [114], and KEGG [115] have substantially improved research efficiency tailored to distinct scientific applications.

The CRISPR-Cas bioinformatics maturation has progressed along two complementary trajectories: the biological classification of natural CRISPR-Cas immune systems in prokaryotes and the technological CRISPR-Cas tools optimisation for biological engineering. DBs dedicated to biological classification, such as CasPDB [116], CRISPRCasdb [117], and CasPEDIA [118], whilst established resources like UniProt [119] exemplify expert and community-driven curation processes that integrate and evaluate evidence from the scientific literature to maintain high-quality annotated records.

In contrast, standardised DBs specifically designed to support CRISPR-Cas biotechnological optimisation, particularly those centred on gRNA activity, design, and efficiency prediction, remain underdeveloped. At present, only a limited number of resources are available, including the BioGRID ORCS (<https://orcs.thebiogrid.org/>; v.2.0.18) [111], which flaunts 2217 screens and 94,219 genes, whilst crisprSQL (<https://www.crisprsql.com/>) [120]

provides a curated DB for CRISPR-Cas9 off-target assays. Although it is restricted to the human (hg38) and mouse (mm10) genomes. The development of dedicated DBs for CRISPR-Cas optimisation should rely on continuous curation, collaborative data integration, and the incorporation of high-value experimental datasets to support reproducibility and downstream analyses. Strengthening these domain-specific resources will be essential for advancing precision engineering in microbial systems.

5.3. Generative AI for CRISPR-Cas Technology

Artificial Intelligence has substantially transformed experimental planning in the life sciences, particularly following the rapid expansion of generative AI technologies in 2022. Generative pre-trained transformer (GPT) models, built on large language models (LLMs), are capable of processing vast volumes of scientific data and supporting tasks ranging from literature synthesis and text refinement to hypothesis generation and code debugging across diverse scientific disciplines [121]. As a result, the scientific community increasingly recognises generative AI as a powerful tool for accelerating and strengthening experimental research workflows [122].

Recent LLM applications extend beyond text-based assistance to the design of biological systems. Notably, AI-driven approaches have enabled the development of OpenCRISPR-1, a novel gene-editing system proposed as an alternative to SpCas9. This work demonstrates that generative LLMs can learn from natural sequence diversity and apply this knowledge toward precision genome editing, highlighting the potential of AI-designed gene editors [57].

In parallel, AI-assisted platforms for CRISPR-Cas experimental design, progressing from conceptual planning to bench-ready protocols, have emerged including the CRISPR AI Research Suite [123]. The GPT-based adaptability systems enable them to serve as intelligent agents that support the planning and execution of CRISPR-Cas experiments. A recent example is CRISPR-GPT, a domain-specific model trained on CRISPR-Cas knowledge that facilitates the automated design and analysis of gene-editing workflows [124]. By integrating tools such as CHOPCHOP and CRISPick, it supports multi-step CRISPR tasks, including system selection, gRNA design with off-target prediction, delivery strategy recommendations, protocol generation, validation planning, and data analysis. Importantly, the modular design allows interaction at different levels for the user's expertise in CRISPR-Cas technologies [125].

The development of improved CRISPR-Cas-based GPT models must extend beyond behaving as AI copilots and instead prioritise robust, microorganism-specific training to support microbial engineering across diverse biotechnological applications. Such systems should integrate structured databases with unstructured experimental literature, enabling the consolidation of CRISPR-Cas tools within an intelligent research ecosystem. Frameworks with a "Q&A mode" should use a robust corpus of CRISPR-related articles, given the rapid expansion of CRISPR-Cas publications, for broader and continuously updated datasets in future GPT-like architectures.

Integrating multiple platforms (e.g., CHOPCHOP, E-CRISP, CRISPick) would allow cross-platform comparison and filtering, enhancing confidence in candidate selection. Current design pipelines also tend to prioritise on-target and off-target scores whilst underrepresenting key biological determinants, including PAM site efficiency, position-specific nucleotide preferences, purine/pyrimidine distribution effects on nuclease activity, and gRNA 2D structure constraints, whilst incorporating these parameters would refine sequence prioritisation and improve functional predictability.

5.4. Ethical Concerns with CRISPR-Cas-Based GPT Models

A comprehensive regulatory framework that addresses social, ethical and legal considerations is essential to ensure the safe application of genome editing technologies, including microbial engineering. An initial step in microbial engineering research involves conducting a risk assessment. In this context, bioethics shed light on the role of evaluating whether a specific genome engineering application may create unacceptable biosafety or biosecurity risks. Biosafety focuses on preventing accidental exposure to or release of biological agents, whereas biosecurity addresses the prevention of deliberate misuse of dangerous biological systems. Risk assessment generally involves three stages: first, identifying potential risks and benefits; second, assigning probability estimates to these outcomes; and third, weighing these factors to determine whether the potential risks are ethically justifiable given the expected benefits [126].

Regulatory approaches to biotechnology vary considerably across regions. In the European Union, the primary focus is on the methodologies used to modify microorganisms rather than on the final product. In contrast, the United States regulatory system tends to evaluate the characteristics and risks of the final product rather than the process used to generate it [127]. In the Middle East, biosafety and biosecurity policies vary widely among countries and are often influenced by economic capacity, political stability, and financial support [128,129].

AI can enhance risk management in biotechnology by enabling a more systematic evaluation and monitoring of experimental designs; however, it may also introduce new risks if misused. Ethical considerations must accompany the development and deployment of CRISPR-Cas-oriented GPTs. For example, CRISPR-GPT incorporates biosafety safeguards by restricting queries related to pathogen engineering through keyword filtering and explicit warning mechanisms [124]. Future AI-assisted CRISPR-Cas GPTs may need to incorporate restrictions related to high-risk pathogens such as *Bacillus cereus* Biovar anthracis, *Coxiella burnetii*, *Francisella tularensis*, *Rickettsia prowazekii*, *Yersinia pestis*, *Burkholderia mallei*, and *Burkholderia pseudomallei*. These organisms are listed by international regulatory agencies such as the United States Federal Select Agent Program and the Australia Group, involving participants from North and Latin America, the European Union, Asia, and Oceania [130,131].

Beyond high-risk pathogens, attention must also be given to emerging infectious microorganisms that could potentially be engineered for dissemination due to their accessibility or ease of laboratory cultivation, including multidrug-resistant bacteria such as *Mycobacterium tuberculosis* [132]. These concerns highlight the need to establish clear governance frameworks that define the responsible boundaries of genome engineering, particularly in the context of AI-assisted CRISPR-Cas technologies that may have dual-use potential.

In regions where regulatory frameworks governing CRISPR-Cas biotechnology and AI remain limited or underdeveloped, such as Mexico under the “Principios de Chapultepec” and other Latin American countries, insufficient oversight may increase biosafety and biosecurity risks [125,133]. Many Latin American countries have yet to fully integrate dual use into national biosafety strategies. This situation may be attributed to a combination of structural, institutional, and regulatory gaps, as well as cultural factors. Strengthening governance structures, regulatory capacity, and bioethics education could help ensure the responsible adoption of emerging genome engineering technologies in those regions [134].

Responsible implementation of AI-assisted CRISPR-Cas GPTs, therefore, requires alignment with international biosecurity standards, transparency in model training and data sources, and the incorporation of safeguards to prevent inappropriate applications. Given that gene-editing technologies carry dual-use potential, including the theoretical

risk of harmful biological manipulation, governance structures must proactively address these concerns. Coordinated efforts among developed nations, international organisations, and countries in the Global South will be essential to establish harmonised regulations and ethical guidelines that support equitable and secure advancement of gene-editing research and its applications [135].

6. Conclusions

The Artificial Intelligence integration into genome engineering represents an irreversible trend that is reshaping microbial biotechnology. As discussed throughout this review, ML and DL enhance CRISPR-Cas9 workflows by enabling rational gRNA design based on multifactorial determinants, such as sequence-specific features, RNA secondary structure for both gRNA and sgRNA, and contextual genomic variables such as PAM preferences. Consequently, next-generation AI models must move beyond sequence-only approaches and adopt multi-feature modelling capable of capturing the complex relationships that govern CRISPR-Cas9 activity.

By reducing reliance on repetitive experimental screening, AI-assisted methods can accelerate experimental planning and increase the predictability of genome editing outcomes. However, a major limitation remains: many current gRNA design rules have been derived from mammalian cells and are often straightforwardly transferred to bacteria and yeast. At present, only a limited number of tools, such as crisprHAL, DeepGuide, DeepSgRN-Abacteria, and sgRNA-cleavage-activity-prediction, have been specifically developed for microbial applications. Addressing such a gap requires the generation of organism-specific, high-quality training datasets to support robust and context-aware predictive models. This effort must be accompanied by standardised experimental designs and consistent labelling strategies to minimise noise and prevent models from learning experimental artefacts.

Future advances in microbial genome engineering will also depend on the standardisation of data through the adoption of FAIR principles and the creation of centralised, standardised CRISPR-Cas gRNA repositories. Implementing structured reporting frameworks, such as MIACS metadata, is essential to ensure that datasets are machine-readable and reusable across research platforms. Furthermore, the emergence of generative AI, such as CRISPR-GPT, represents a promising frontier for automated experimental and high-efficiency gRNA design.

As the field progresses toward an increasingly intelligent research ecosystem, bioethical and biosecurity considerations must remain central, especially in regions where developing regulatory safeguards must ensure that advances in genome engineering align with international safety and governance standards to prevent misuse. Strengthening interdisciplinary collaboration between biotechnology, computational biology, and computer science will unlock the full potential of AI-guided CRISPR-Cas technologies.

Ultimately, AI integration, CRISPR-Cas9, and microbial systems biology is driving a transition toward a new era of predictive genome engineering. The convergence of high-quality experimental datasets, standardised CRISPR-Cas infrastructures for precise engineering, and advanced AI architectures will likely transform microbial genome engineering into a predictive and programmable discipline with broad implications for metabolic engineering and synthetic biology.

Author Contributions: J.A.D.-N.: Conceptualisation, Methodology, Formal analysis, Validation, and Visualisation. J.A.D.-N. and D.A.P.-P.: Investigation and Writing—original draft. J.A.D.-N., D.A.P.-P., L.J.F.-Y., O.G.-R., M.A.G.-R. and E.R.-D.: Writing—review and editing. J.A.D.-N., L.J.F.-Y. and O.G.-R.: Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SECIHTI and the Doctoral Program in Sciences in Biotechnological Processes at the University of Guadalajara, grant numbers 1267568 and 2163251.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2D	Secondary
AI	Artificial intelligence
AMR	Antimicrobial resistance
aSD	Anti Shine-Dalgarno
BiLSTM	Bidirectional long short-term memory
BioGRID ORCS	Biological General Repository for Interaction Datasets Open Repository of CRISPR Screens
Cas	CRISPR-associated protein
CNN	Convolutional neural network
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DB	Database
DBN	Deep Belief Network
DL	Deep learning
DNA	Deoxyribonucleic acid
DSB	Double-strand break
EGA	Elitist Genetic Algorithm
FAIR	Findability, Accessibility, Interoperability, and Reusability
GAT	Graph attention network
GBRT	Gradient boosting regression tree
GNN	Graph neural network
GPT	Generative pre-trained transformer
gRNA	Guide RNA
HTS	High-throughput screening
Indels	Insertions and deletions
iRF	iterative Random Forest
LLM	Large language model
MFE	Minimum free energy
MIACS	Minimal Information About CRISPR Screens
ML	Machine learning
mRNA	Messenger RNA
nt	Nucleotides
PAM	Protospacer adjacent motif
RBM	Restricted Boltzmann machine
RBS	Ribosome-binding site
RNA	Ribonucleic acid
RNN	Restricted Boltzmann machine
SD	Shine-Dalgarno
sgRNA	Single guide ribonucleic acid
SpCas9	<i>Streptococcus pyogenes</i> Cas9
SVM	Support Vector Machine
XAI	Explainable AI

References

- Li, X.; Liu, Y.; Ma, L.; Jiang, W.; Shi, T.; Li, L.; Li, C.; Chen, Z.; Fan, X.; Xu, Q. Metabolic Engineering of *Escherichia coli* for High-Yield Dopamine Production via Optimized Fermentation Strategies. *Appl. Environ. Microbiol.* **2025**, *91*, e00159–25. [[CrossRef](#)] [[PubMed](#)]
- Ye, C.; Zhang, Y.; Zhang, J.; Shi, M.; Nie, F.; Liu, Q. Metabolic Engineering of *Escherichia coli* BW25113 for the Production of Vitamin K2 Based on CRISPR/Cas9 Mediated Gene Knockout and Metabolic Pathway Modification. *J. Biol. Eng.* **2026**, *20*, 29. [[CrossRef](#)] [[PubMed](#)]
- Han, S.; Jang, H.W.; Park, S.; Kim, T.M.; Kim, H.J. Unlocking the Flavor Potential of Brewing Yeast with CRISPR/Cas9 Genome Editing. *LWT* **2025**, *230*, 118254. [[CrossRef](#)]
- Lee, H.-J.; Shin, D.J.; Nho, S.B.; Lee, K.W.; Kim, S.-K. Metabolic Engineering of *Saccharomyces cerevisiae* for Fermentative Production of Heme. *Biotechnol. J.* **2024**, *19*, e202400351. [[CrossRef](#)] [[PubMed](#)]
- Bush, K.; Corsi, G.I.; Yan, A.C.; Haynes, K.; Layzer, J.M.; Zhou, J.H.; Llanga, T.; Gorodkin, J.; Sullenger, B.A. Utilizing Directed Evolution to Interrogate and Optimize CRISPR/Cas Guide RNA Scaffolds. *Cell Chem. Biol.* **2023**, *30*, 879–892.e5. [[CrossRef](#)] [[PubMed](#)]
- Li, C.; Zou, Q.; Li, J.; Feng, H. Prediction of CRISPR-Cas9 on-Target Activity Based on a Hybrid Neural Network. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 2098–2106. [[CrossRef](#)] [[PubMed](#)]
- Riesenberg, S.; Helmbrecht, N.; Kanis, P.; Maricic, T.; Pääbo, S. Improved gRNA Secondary Structures Allow Editing of Target Sites Resistant to CRISPR-Cas9 Cleavage. *Nat. Commun.* **2022**, *13*, 489. [[CrossRef](#)] [[PubMed](#)]
- Reynolds, S.A.; Beery, S.; Burgess, N.; Burgman, M.; Butchart, S.H.M.; Cooke, S.J.; Coomes, D.; Danielsen, F.; Di Minin, E.; Durán, A.P.; et al. The Potential for AI to Revolutionize Conservation: A Horizon Scan. *Trends Ecol. Evol.* **2025**, *40*, 191–207. [[CrossRef](#)] [[PubMed](#)]
- Zhou, G.; Rusnac, D.-V.; Park, H.; Canzani, D.; Nguyen, H.M.; Stewart, L.; Bush, M.F.; Nguyen, P.T.; Wulff, H.; Yarov-Yarovoy, V.; et al. An Artificial Intelligence Accelerated Virtual Screening Platform for Drug Discovery. *Nat. Commun.* **2024**, *15*, 7761. [[CrossRef](#)] [[PubMed](#)]
- Yang, X.; Jia, J.; Zhou, X.; Wang, S. The Future of Artificial Intelligence: Time to Embrace More International Collaboration. *Innovation* **2024**, *5*, 100703. [[CrossRef](#)] [[PubMed](#)]
- Xiang, X.; Corsi, G.I.; Anthon, C.; Qu, K.; Pan, X.; Liang, X.; Han, P.; Dong, Z.; Liu, L.; Zhong, J.; et al. Enhancing CRISPR-Cas9 gRNA Efficiency Prediction by Data Integration and Deep Learning. *Nat. Commun.* **2021**, *12*, 3238. [[CrossRef](#)] [[PubMed](#)]
- Guha, D.; Avtaran, D.; Lenka, R.; Yang, T.; Wang, L.; Rathore, R.S. Leveraging a Smart AI-Controlled GRNA in Genome Editing for Identification and Replacement of Genetic Mutations. In *Proceedings of Fourth International Conference on Computing and Communication Networks*; Kumar, A., Swaroop, A., Shukla, P., Eds.; Springer Nature: Singapore, 2025; pp. 649–656.
- Wan, S.; Liu, X.; Sun, W.; Lv, B.; Li, C. Current Advances for Omics-Guided Process Optimization of Microbial Manufacturing. *Bioresour. Bioprocess.* **2023**, *10*, 30. [[CrossRef](#)] [[PubMed](#)]
- Abbate, E.; Andrion, J.; Apel, A.; Biggs, M.; Chaves, J.; Cheung, K.; Ciesla, A.; Clark-ElSayed, A.; Clay, M.; Contridas, R.; et al. Optimizing the Strain Engineering Process for Industrial-Scale Production of Bio-Based Molecules. *J. Ind. Microbiol. Biotechnol.* **2023**, *50*, kuad025. [[CrossRef](#)] [[PubMed](#)]
- Sun, X.; Zhang, H.; Jia, Y.; Li, J.; Jia, M. CRISPR-Cas9-Based Genome-Editing Technologies in Engineering Bacteria for the Production of Plant-Derived Terpenoids. *Eng. Microbiol.* **2024**, *4*, 100154. [[CrossRef](#)] [[PubMed](#)]
- Sadanov, A.K.; Baimakhanova, B.B.; Orasymbet, S.E.; Ratnikova, I.A.; Turlybaeva, Z.Z.; Baimakhanova, G.B.; Amitova, A.A.; Omirbekova, A.A.; Aitkaliyeva, G.S.; Kossalbayev, B.D.; et al. Engineering Useful Microbial Species for Pharmaceutical Applications. *Microorganisms* **2025**, *13*, 599. [[CrossRef](#)] [[PubMed](#)]
- Wu, Z.; Chen, T.; Sun, W.; Chen, Y.; Ying, H. Optimizing *Escherichia coli* Strains and Fermentation Processes for Enhanced L-Lysine Production: A Review. *Front. Microbiol.* **2024**, *15*, 1485624. [[CrossRef](#)] [[PubMed](#)]
- Gao, H.; Qiu, Z.; Wang, X.; Zhang, X.; Zhang, Y.; Dai, J.; Liang, Z. Recent Advances in Genome-Scale Engineering in *Escherichia coli* and Their Applications. *Eng. Microbiol.* **2024**, *4*, 100115. [[CrossRef](#)] [[PubMed](#)]
- Ahmed, M.; Kayode, H.; Okesanya, O.; Ukoaka, B.; Eshun, G.; Mourid, M.; Adigun, O.; Ogaya, J.; Mohamed, Z.; Lucero-Prisno, D. CRISPR-Cas Systems in the Fight Against Antimicrobial Resistance: Current Status, Potentials, and Future Directions. *Infect. Drug Resist.* **2024**, *17*, 5229–5245. [[CrossRef](#)] [[PubMed](#)]
- Okesanya, O.J.; Ahmed, M.M.; Ogaya, J.B.; Amisu, B.O.; Ukoaka, B.M.; Adigun, O.A.; Manirambona, E.; Adebusi, O.; Othman, Z.K.; Oluwakemi, O.G.; et al. Reinvigorating AMR Resilience: Leveraging CRISPR–Cas Technology Potentials to Combat the 2024 WHO Bacterial Priority Pathogens for Enhanced Global Health Security—A Systematic Review. *Trop. Med. Health* **2025**, *53*, 43. [[CrossRef](#)] [[PubMed](#)]
- Wang, L.; Zhang, J. Prediction of sgRNA On-Target Activity in Bacteria by Deep Learning. *BMC Bioinform.* **2019**, *20*, 517. [[CrossRef](#)] [[PubMed](#)]

22. Guo, J.; Wang, T.; Guan, C.; Liu, B.; Luo, C.; Xie, Z.; Zhang, C.; Xing, X.-H. Improved sgRNA Design in Bacteria via Genome-Wide Activity Profiling. *Nucleic Acids Res.* **2018**, *46*, 7052–7069. [[CrossRef](#)] [[PubMed](#)]
23. Ham, D.T.; Browne, T.S.; Banglorewala, P.N.; Wilson, T.L.; Michael, R.K.; Gloor, G.B.; Edgell, D.R. A Generalizable Cas9/sgRNA Prediction Model Using Machine Transfer Learning with Small High-Quality Datasets. *Nat. Commun.* **2023**, *14*, 5514. [[CrossRef](#)] [[PubMed](#)]
24. Wu, X.; Scott, D.A.; Kriz, A.J.; Chiu, A.C.; Hsu, P.D.; Dadon, D.B.; Cheng, A.W.; Trevino, A.E.; Konermann, S.; Chen, S.; et al. Genome-Wide Binding of the CRISPR Endonuclease Cas9 in Mammalian Cells. *Nat. Biotechnol.* **2014**, *32*, 670–676. [[CrossRef](#)] [[PubMed](#)]
25. Xu, H.; Xiao, T.; Chen, C.H.; Li, W.; Meyer, C.A.; Wu, Q.; Wu, D.; Cong, L.; Zhang, F.; Liu, J.S.; et al. Sequence Determinants of Improved CRISPR sgRNA Design. *Genome Res.* **2015**, *25*, 1147–1157. [[CrossRef](#)] [[PubMed](#)]
26. Moreno-Mateos, M.A.; Vejnar, C.E.; Beaudoin, J.-D.; Fernandez, J.P.; Mis, E.K.; Khokha, M.K.; Giraldez, A.J. CRISPRscan: Designing Highly Efficient sgRNAs for CRISPR-Cas9 Targeting In Vivo. *Nat. Methods* **2015**, *12*, 982–988. [[CrossRef](#)] [[PubMed](#)]
27. Doench, J.G.; Hartenian, E.; Graham, D.B.; Tothova, Z.; Hegde, M.; Smith, I.; Sullender, M.; Ebert, B.L.; Xavier, R.J.; Root, D.E. Rational Design of Highly Active sgRNAs for CRISPR-Cas9-Mediated Gene Inactivation. *Nat. Biotechnol.* **2014**, *32*, 1262–1267. [[CrossRef](#)] [[PubMed](#)]
28. Labun, K.; Montague, T.G.; Krause, M.; Torres Cleuren, Y.N.; Tjeldnes, H.; Valen, E. CHOPCHOP v3: Expanding the CRISPR Web Toolbox Beyond Genome Editing. *Nucleic Acids Res.* **2019**, *47*, W171–W174. [[CrossRef](#)] [[PubMed](#)]
29. Wang, T.; Wei, J.J.; Sabatini, D.M.; Lander, E.S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **2014**, *343*, 80–84. [[CrossRef](#)] [[PubMed](#)]
30. Corsi, G.I.; Qu, K.; Alkan, F.; Pan, X.; Luo, Y.; Gorodkin, J. CRISPR/Cas9 gRNA Activity Depends on Free Energy Changes and on the Target PAM Context. *Nat. Commun.* **2022**, *13*, 3006. [[CrossRef](#)] [[PubMed](#)]
31. Noshay, J.M.; Walker, T.; Alexander, W.G.; Klingeman, D.M.; Romero, J.; Walker, A.M.; Prates, E.; Eckert, C.; Irle, S.; Kainer, D.; et al. Quantum Biological Insights into CRISPR-Cas9 sgRNA Efficiency from Explainable-AI Driven Feature Engineering. *Nucleic Acids Res.* **2023**, *51*, 10147–10161. [[CrossRef](#)] [[PubMed](#)]
32. Wang, D.; Zhang, C.; Wang, B.; Li, B.; Wang, Q.; Liu, D.; Wang, H.; Zhou, Y.; Shi, L.; Lan, F.; et al. Optimized CRISPR Guide RNA Design for Two High-Fidelity Cas9 Variants by Deep Learning. *Nat. Commun.* **2019**, *10*, 4284. [[CrossRef](#)] [[PubMed](#)]
33. Xue, L.; Tang, B.; Chen, W.; Luo, J. Prediction of CRISPR SgRNA Activity Using a Deep Convolutional Neural Network. *J. Chem. Inf. Model.* **2019**, *59*, 615–624. [[CrossRef](#)] [[PubMed](#)]
34. Jin, L.; Liyanage, R.; Duan, D.; Chen, S.-J. Machine-Learning-Inferred and Energy-Landscape-Guided Analyses Reveal Kinetic Determinants of CRISPR/Cas9 Gene Editing. *PRX Life* **2026**, *4*, 013028. [[CrossRef](#)]
35. Moreb, E.A.; Lynch, M.D. A Meta-Analysis of gRNA Library Screens Enables an Improved Understanding of the Impact of GRNA Folding and Structural Stability on CRISPR-Cas9 Activity. *CRISPR J.* **2022**, *5*, 146–154. [[CrossRef](#)] [[PubMed](#)]
36. Peng, H.; Zheng, Y.; Blumenstein, M.; Tao, D.; Li, J. CRISPR/Cas9 Cleavage Efficiency Regression through Boosting Algorithms and Markov Sequence Profiling. *Bioinformatics* **2018**, *34*, 3069–3077. [[CrossRef](#)] [[PubMed](#)]
37. Kim, H.K.; Kim, Y.; Lee, S.; Min, S.; Bae, J.Y.; Choi, J.W.; Park, J.; Jung, D.; Yoon, S.; Kim, H.H. SpCas9 Activity Prediction by DeepSpCas9, a Deep Learning-Based Model with High Generalization Performance. *Sci. Adv.* **2019**, *5*, eaax9249. [[CrossRef](#)] [[PubMed](#)]
38. Zhu, W.; Xie, H.; Chen, Y.; Zhang, G. CrnnCrispr: An Interpretable Deep Learning Method for CRISPR/Cas9 sgRNA On-Target Activity Prediction. *Int. J. Mol. Sci.* **2024**, *25*, 4429. [[CrossRef](#)] [[PubMed](#)]
39. 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*, 1445–1457. [[CrossRef](#)] [[PubMed](#)]
40. Dimauro, G.; Colagrande, P.; Carlucci, R.; Ventura, M.; Bevilacqua, V.; Caivano, D. CRISPRLearner: A Deep Learning-Based System to Predict CRISPR/Cas9 SgRNA On-Target Cleavage Efficiency. *Electronics* **2019**, *8*, 1478. [[CrossRef](#)]
41. Chuai, G.; Ma, H.; Yan, J.; Chen, M.; Hong, N.; Xue, D.; Zhou, C.; Zhu, C.; Chen, K.; Duan, B.; et al. DeepCRISPR: Optimized CRISPR Guide RNA Design by Deep Learning. *Genome Biol.* **2018**, *19*, 80. [[CrossRef](#)] [[PubMed](#)]
42. 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*, 3616–3637. [[CrossRef](#)] [[PubMed](#)]
43. Doench, J.G.; Fusi, N.; Sullender, M.; Hegde, M.; Vaimberg, E.W.; Donovan, K.F.; Smith, I.; Tothova, Z.; Wilen, C.; Orchard, R.; et al. Optimized sgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34*, 184–191. [[CrossRef](#)] [[PubMed](#)]
44. Haeussler, M.; Schönig, K.; Eckert, H.; Eschstruth, A.; Mianné, J.; Renaud, J.-B.; Schneider-Maunoury, S.; Shkumatava, A.; Teboul, L.; Kent, J.; et al. Evaluation of Off-Target and on-Target Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR. *Genome Biol.* **2016**, *17*, 148. [[CrossRef](#)] [[PubMed](#)]

45. Labuhn, M.; Adams, F.F.; Ng, M.; Knoess, S.; Schambach, A.; Charpentier, E.M.; Schwarzer, A.; Mateo, J.L.; Klusmann, J.-H.; Heckl, D. Refined sgRNA Efficacy Prediction Improves Large- and Small-Scale CRISPR–Cas9 Applications. *Nucleic Acids Res.* **2018**, *46*, 1375–1385. [[CrossRef](#)] [[PubMed](#)]
46. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **2012**, *337*, 816–821. [[CrossRef](#)] [[PubMed](#)]
47. Dong, C.; Gou, Y.; Lian, J. SgRNA Engineering for Improved Genome Editing and Expanded Functional Assays. *Curr. Opin. Biotechnol.* **2022**, *75*, 102697. [[CrossRef](#)] [[PubMed](#)]
48. Briner, A.E.; Donohoue, P.D.; Gomaa, A.A.; Selle, K.; Slorach, E.M.; Nye, C.H.; Haurwitz, R.E.; Beisel, C.L.; May, A.P.; Barrangou, R. Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Mol. Cell* **2014**, *56*, 333–339. [[CrossRef](#)] [[PubMed](#)]
49. De Saeger, J. A Guide to Guides: An Overview of SpCas9 sgRNA Scaffold Variants and Modifications. *SynBio* **2025**, *3*, 19. [[CrossRef](#)]
50. Wong, N.; Liu, W.; Wang, X. WU-CRISPR: Characteristics of Functional Guide RNAs for the CRISPR/Cas9 System. *Genome Biol.* **2015**, *16*, 218. [[CrossRef](#)] [[PubMed](#)]
51. Kocak, D.D.; Josephs, E.A.; Bhandarkar, V.; Adkar, S.S.; Kwon, J.B.; Gersbach, C.A. Increasing the Specificity of CRISPR Systems with Engineered RNA Secondary Structures. *Nat. Biotechnol.* **2019**, *37*, 657–666. [[CrossRef](#)] [[PubMed](#)]
52. Zhang, Y.; Chen, G.; Liang, C.; Yang, B.; Lei, X.; Chen, T.; Jiang, H.; Xiong, W. MultiCRISPR-EGA: Optimizing Guide RNA Array Design for Multiplexed CRISPR Using the Elitist Genetic Algorithm. *ACS Synth. Biol.* **2025**, *14*, 919–930. [[CrossRef](#)] [[PubMed](#)]
53. Jiang, Y.; Li, B.; Xiong, J.; Liu, X. Graph-CRISPR: A Gene Editing Efficiency Prediction Model Based on Graph Neural Network with Integrated Sequence and Secondary Structure Feature Extraction. *Brief. Bioinform.* **2025**, *26*, bbaf410. [[CrossRef](#)] [[PubMed](#)]
54. Collias, D.; Beisel, C.L. CRISPR Technologies and the Search for the PAM-Free Nuclease. *Nat. Commun.* **2021**, *12*, 555. [[CrossRef](#)] [[PubMed](#)]
55. Globyte, V.; Lee, S.H.; Bae, T.; Kim, J.; Joo, C. CRISPR/Cas9 Searches for a Protospacer Adjacent Motif by Lateral Diffusion. *EMBO J.* **2018**, *38*, EMBJ201899466. [[CrossRef](#)] [[PubMed](#)]
56. Malbranke, C.; Rostain, W.; Depardieu, F.; Cocco, S.; Monasson, R.; Bikard, D. Computational Design of Novel Cas9 PAM-Interacting Domains Using Evolution-Based Modelling and Structural Quality Assessment. *PLoS Comput. Biol.* **2023**, *19*, e1011621. [[CrossRef](#)] [[PubMed](#)]
57. Ruffolo, J.A.; Nayfach, S.; Gallagher, J.; Bhatnagar, A.; Beazer, J.; Hussain, R.; Russ, J.; Yip, J.; Hill, E.; Pacesa, M.; et al. Design of Highly Functional Genome Editors by Modelling CRISPR–Cas Sequences. *Nature* **2025**, *645*, 518–525. [[CrossRef](#)] [[PubMed](#)]
58. Glass, Z.; Lee, M.; Li, Y.; Xu, Q. Engineering the Delivery System for CRISPR-Based Genome Editing. *Trends Biotechnol.* **2018**, *36*, 173–185. [[CrossRef](#)] [[PubMed](#)]
59. Tudek, A.; Krawczyk, P.S.; Mroczek, S.; Tomecki, R.; Turtola, M.; Matylla-Kulińska, K.; Jensen, T.H.; Dziembowski, A. Global View on the Metabolism of RNA Poly(A) Tails in Yeast *Saccharomyces cerevisiae*. *Nat. Commun.* **2021**, *12*, 4951. [[CrossRef](#)] [[PubMed](#)]
60. Wen, J.-D.; Kuo, S.-T.; Chou, H.-H.D. The Diversity of Shine-Dalgarno Sequences Sheds Light on the Evolution of Translation Initiation. *RNA Biol.* **2021**, *18*, 1489–1500. [[CrossRef](#)] [[PubMed](#)]
61. Passmore, L.A.; Collier, J. Roles of mRNA Poly(A) Tails in Regulation of Eukaryotic Gene Expression. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 93–106. [[CrossRef](#)] [[PubMed](#)]
62. Poonia, P.; Valabhoju, V.; Li, T.; Iben, J.; Niu, X.; Lin, Z.; Hinnebusch, A.G. Yeast Poly(A)-Binding Protein (Pab1) Controls Translation Initiation in Vivo Primarily by Blocking mRNA Decapping and Decay. *Nucleic Acids Res.* **2025**, *53*, gkaf143. [[CrossRef](#)] [[PubMed](#)]
63. Zhang, M.; Holowko, M.B.; Hayman Zumpe, H.; Ong, C.S. Machine Learning Guided Batched Design of a Bacterial Ribosome Binding Site. *ACS Synth. Biol.* **2022**, *11*, 2314–2326. [[CrossRef](#)] [[PubMed](#)]
64. Goshisht, M.K. Machine Learning and Deep Learning in Synthetic Biology: Key Architectures, Applications, and Challenges. *ACS Omega* **2024**, *9*, 9921–9945. [[CrossRef](#)] [[PubMed](#)]
65. Maharana, K.; Mondal, S.; Nemade, B. A Review: Data Pre-Processing and Data Augmentation Techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [[CrossRef](#)]
66. Dixit, S.; Kumar, A.; Srinivasan, K.; Vincent, P.M.D.R.; Ramu Krishnan, N. Advancing Genome Editing with Artificial Intelligence: Opportunities, Challenges, and Future Directions. *Front. Bioeng. Biotechnol.* **2024**, *11*, 1335901. [[CrossRef](#)] [[PubMed](#)]
67. Ortiz, B.L.; Gupta, V.; Kumar, R.; Jalin, A.; Cao, X.; Ziegenbein, C.; Singhal, A.; Tewari, M.; Choi, S.W. Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care. *JMIR mHealth uHealth* **2024**, *12*, e59587. [[CrossRef](#)] [[PubMed](#)]
68. Dagal, I.; Harrison, A.; Ibrahim, A.-W.; Mbaso, W.F. Comprehensive Evaluation of Data Preprocessing and Visualization Techniques for Enhanced Classification and Sampling. *Clust. Comput.* **2025**, *28*, 476. [[CrossRef](#)]
69. Yang, Y.; Li, J.; Zou, Q.; Ruan, Y.; Feng, H. Prediction of CRISPR-Cas9 off-Target Activities with Mismatches and Indels Based on Hybrid Neural Network. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 5039–5048. [[CrossRef](#)] [[PubMed](#)]

70. Cao, M.; Brennan, A.; Lee, C.M.; Park, S.; Bao, G. Deep Learning Based Models for CRISPR/Cas Off-Target Prediction. *Small Methods* **2025**, *9*, 2500122. [[CrossRef](#)] [[PubMed](#)]
71. Abbasi, A.F.; Asim, M.N.; Dengel, A. Transitioning from Wet Lab to Artificial Intelligence: A Systematic Review of AI Predictors in CRISPR. *J. Transl. Med.* **2025**, *23*, 153. [[CrossRef](#)] [[PubMed](#)]
72. Jasieniecka, A.; Domingues, I. CRISPR-Cas9 and Its Bioinformatics Tools: A Systematic Review. *Curr. Issues Mol. Biol.* **2025**, *47*, 307. [[CrossRef](#)] [[PubMed](#)]
73. Tyagi, S.; Kumar, R.; Das, A.; Won, S.Y.; Shukla, P. CRISPR-Cas9 System: A Genome-Editing Tool with Endless Possibilities. *J. Biotechnol.* **2020**, *319*, 36–53. [[CrossRef](#)] [[PubMed](#)]
74. Sun, Y.; Qu, K.; Corsi, G.I.; Anthon, C.; Pan, X.; Xiang, X.; Jensen, L.J.; Lin, L.; Luo, Y.; Gorodkin, J. Deep Learning Models Simultaneously Trained on Multiple Datasets Improve Base-Editing Activity Prediction. *Nat. Commun.* **2025**, *16*, 9821. [[CrossRef](#)] [[PubMed](#)]
75. Bhat, A.A.; Nisar, S.; Mukherjee, S.; Saha, N.; Yarravarapu, N.; Lone, S.N.; Masoodi, T.; Chauhan, R.; Maacha, S.; Bagga, P.; et al. Integration of CRISPR/Cas9 with Artificial Intelligence for Improved Cancer Therapeutics. *J. Transl. Med.* **2022**, *20*, 534. [[CrossRef](#)] [[PubMed](#)]
76. Agate, J. Artificial Intelligence Methods and Approaches to Improve Data Quality in Healthcare Data. *Artif. Intell. Life Sci.* **2025**, *8*, 100135. [[CrossRef](#)]
77. Aussel, C.; Cathomen, T.; Fuster-García, C. The Hidden Risks of CRISPR/Cas: Structural Variations and Genome Integrity. *Nat. Commun.* **2025**, *16*, 7208. [[CrossRef](#)] [[PubMed](#)]
78. Yan, Q.; Fong, S.S. Challenges and Advances for Genetic Engineering of Non-Model Bacteria and Uses in Consolidated Bioprocessing. *Front. Microbiol.* **2017**, *8*, 2060. [[CrossRef](#)] [[PubMed](#)]
79. Call, S.N.; Andrews, L.B. CRISPR-Based Approaches for Gene Regulation in Non-Model Bacteria. *Front. Genome Ed.* **2022**, *4*, 892304. [[CrossRef](#)] [[PubMed](#)]
80. Vercauteren, S.; Fiesack, S.; Maroc, L.; Verstraeten, N.; Dewachter, L.; Michiels, J.; Vonesch, S.C. The Rise and Future of CRISPR-Based Approaches for High-Throughput Genomics. *FEMS Microbiol. Rev.* **2024**, *48*, fuae020. [[CrossRef](#)] [[PubMed](#)]
81. Moreb, E.A.; Hoover, B.; Yaseen, A.; Valyasevi, N.; Roecker, Z.; Menacho-Melgar, R.; Lynch, M.D. Managing the SOS Response for Enhanced CRISPR-Cas-Based Recombineering in *E. coli* through Transient Inhibition of Host RecA Activity. *ACS Synth. Biol.* **2017**, *6*, 2209–2218. [[CrossRef](#)] [[PubMed](#)]
82. Ham, D.T.; Browne, T.S.; Zhang, C.Q.; Foo, G.W.; Uruthirapathy, A.S.; Gloor, G.B.; Edgell, D.R. Machine Learning Reveals Sequence and Methylation Determinants of SaCas9–PAM Interactions in Bacteria. *Nucleic Acids Res.* **2026**, *54*, gkaf1520. [[CrossRef](#)] [[PubMed](#)]
83. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
84. Xu, J.; Liu, C.; Tan, X.; Zhu, X.; Wu, A.; Wan, H.; Kong, W.; Li, C.; Xu, H.; Kuang, K.; et al. General Information Metrics for Improving AI Model Training Efficiency. *Artif. Intell. Rev.* **2025**, *58*, 289. [[CrossRef](#)]
85. Kimata, K.; Satou, K. Improved CRISPR/Cas9 off-Target Prediction with DNABERT and Epigenetic Features. *PLoS ONE* **2025**, *20*, e0335863. [[CrossRef](#)] [[PubMed](#)]
86. Trivedi, V.; Mohseni, A.; Lonardi, S.; Wheeldon, I. Balanced Training Sets Improve Deep Learning-Based Prediction of CRISPR sgRNA Activity. *ACS Synth. Biol.* **2024**, *13*, 3774–3781. [[CrossRef](#)] [[PubMed](#)]
87. Yaish, O.; Orenstein, Y. Generating, Modeling and Evaluating a Large-Scale Set of CRISPR/Cas9 off-Target Sites with Bulges. *Nucleic Acids Res.* **2024**, *52*, 6777–6790. [[CrossRef](#)] [[PubMed](#)]
88. Koike-Yusa, H.; Li, Y.; Tan, E.-P.; Velasco-Herrera, M.D.C.; Yusa, K. Genome-Wide Recessive Genetic Screening in Mammalian Cells with a Lentiviral CRISPR-guide RNA Library. *Nat. Biotechnol.* **2014**, *32*, 267–273. [[CrossRef](#)] [[PubMed](#)]
89. Moreb, E.A.; Lynch, M.D. Genome Dependent Cas9/gRNA Search Time Underlies Sequence Dependent gRNA Activity. *Nat. Commun.* **2021**, *12*, 5034. [[CrossRef](#)] [[PubMed](#)]
90. Trivedi, V.; Ramesh, A.; Wheeldon, I. Analyzing CRISPR Screens in Non-Conventional Microbes. *J. Ind. Microbiol. Biotechnol.* **2023**, *50*, kuad006. [[CrossRef](#)] [[PubMed](#)]
91. Manghwar, H.; Li, B.; Ding, X.; Hussain, A.; Lindsey, K.; Zhang, X.; Jin, S. CRISPR/Cas Systems in Genome Editing: Methodologies and Tools for sgRNA Design, Off-Target Evaluation, and Strategies to Mitigate Off-Target Effects. *Adv. Sci.* **2020**, *7*, 1902312. [[CrossRef](#)] [[PubMed](#)]
92. Moreb, E.A.; Hutmacher, M.; Lynch, M.D. CRISPR-Cas “Non-Target” Sites Inhibit On-Target Cutting Rates. *CRISPR J.* **2020**, *3*, 550–561. [[CrossRef](#)] [[PubMed](#)]
93. Tafrishi, A.; Trivedi, V.; Xing, Z.; Li, M.; Mewalal, R.; Cutler, S.R.; Blaby, I.; Wheeldon, I. Functional Genomic Screening in *Komagataella phaffii* Enabled by High-Activity CRISPR-Cas9 Library. *Metab. Eng.* **2024**, *85*, 73–83. [[CrossRef](#)] [[PubMed](#)]
94. Schwartz, C.; Cheng, J.-F.; Evans, R.; Schwartz, C.A.; Wagner, J.M.; Anglin, S.; Beitz, A.; Pan, W.; Lonardi, S.; Blenner, M.; et al. Validating Genome-Wide CRISPR-Cas9 Function Improves Screening in the Oleaginous Yeast *Yarrowia lipolytica*. *Metab. Eng.* **2019**, *55*, 102–110. [[CrossRef](#)] [[PubMed](#)]

95. Baisya, D.; Ramesh, A.; Schwartz, C.; Lonardi, S.; Wheeldon, I. Genome-Wide Functional Screens Enable the Prediction of High Activity CRISPR-Cas9 and -Cas12a Guides in *Yarrowia lipolytica*. *Nat. Commun.* **2022**, *13*, 922. [[CrossRef](#)] [[PubMed](#)]
96. Cho, S.; Shin, J.; Cho, B.-K. Applications of CRISPR/Cas System to Bacterial Metabolic Engineering. *Int. J. Mol. Sci.* **2018**, *19*, 1089. [[CrossRef](#)] [[PubMed](#)]
97. Yang, J.; Song, J.; Feng, Z.; Ma, Y. Application of CRISPR-Cas9 in Microbial Cell Factories. *Biotechnol. Lett.* **2025**, *47*, 46. [[CrossRef](#)] [[PubMed](#)]
98. Carruthers, D.N.; Kinnunen, P.C.; Li, Y.; Chen, Y.; Gin, J.W.; Yunus, I.S.; Galliard, W.R.; Tan, S.; Radiwojevic, T.; Adams, P.D.; et al. Automation and Machine Learning Drive Rapid Optimization of Isoprenol Production in *Pseudomonas putida*. *Nat. Commun.* **2025**, *16*, 11489. [[CrossRef](#)] [[PubMed](#)]
99. Tenkanen, T.; Ylinen, A.; Jouhten, P.; Penttilä, M.; Castillo, S. PHA Synthase Variant Design Using a Conditional Variational Autoencoder. *PLoS Comput. Biol.* **2026**, *22*, e1014087. [[CrossRef](#)] [[PubMed](#)]
100. Wu, D.; Xu, F.; Xu, Y.; Huang, M.; Li, Z.; Chu, J. Towards a Hybrid Model-Driven Platform Based on Flux Balance Analysis and a Machine Learning Pipeline for Biosystem Design. *Synth. Syst. Biotechnol.* **2024**, *9*, 33–42. [[CrossRef](#)] [[PubMed](#)]
101. Jiang, W.; Wang, S.; Ahlheit, D.; Fumagalli, T.; Yang, Z.; Ramanathan, S.; Jiang, X.; Weber, T.; Dahlin, J.; Borodina, I. High-Throughput Metabolic Engineering of *Yarrowia lipolytica* through Gene Expression Tuning. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2426686122. [[CrossRef](#)] [[PubMed](#)]
102. Iwai, K.; Wehrs, M.; Garber, M.; Sustarich, J.; Washburn, L.; Costello, Z.; Kim, P.W.; Ando, D.; Gaillard, W.R.; Hillson, N.J.; et al. Scalable and Automated CRISPR-Based Strain Engineering Using Droplet Microfluidics. *Microsyst. Nanoeng.* **2022**, *8*, 31. [[CrossRef](#)] [[PubMed](#)]
103. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [[CrossRef](#)] [[PubMed](#)]
104. Plewnia, A.; Hoenig, B.D.; Lötters, S.; Heine, C.; Erens, J.; Böning, P.; Bending, G.D.; Krehenwinkel, H.; Williams, M.A. The Emergence of a CRISPR-Cas Revolution in Ecology: Applications, Challenges, and an Ecologist's Overview of the Toolbox. *Mol. Ecol. Resour.* **2026**, *26*, e70086. [[CrossRef](#)] [[PubMed](#)]
105. David, R.; Mabile, L.; Specht, A.; Stryeck, S.; Thomsen, M.; Yahia, M.; Jonquet, C.; Dollé, L.; Jacob, D.; Bailo, D.; et al. FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles. *Data Sci. J.* **2020**, *19*, 1–11. [[CrossRef](#)]
106. D'Ambrosio, V.; Hansen, L.G.; Zhang, J.; Jensen, E.D.; Arsovska, D.; Laloux, M.; Jakociūnas, T.; Hjort, P.; De Lucrezia, D.; Marletta, S.; et al. A FAIR-Compliant Parts Catalogue for Genome Engineering and Expression Control in *Saccharomyces cerevisiae*. *Synth. Syst. Biotechnol.* **2022**, *7*, 657–663. [[CrossRef](#)] [[PubMed](#)]
107. Tao, J.; Bauer, D.E.; Chiarle, R. Assessing and Advancing the Safety of CRISPR-Cas Tools: From DNA to RNA Editing. *Nat. Commun.* **2023**, *14*, 212. [[CrossRef](#)] [[PubMed](#)]
108. Zhang, D.; Hurst, T.; Duan, D.; Chen, S.-J. Unified Energetics Analysis Unravels SpCas9 Cleavage Activity for Optimal gRNA Design. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8693–8698. [[CrossRef](#)] [[PubMed](#)]
109. Natural Methods Editorial. CRISPR Standards. *Nat. Methods* **2017**, *14*, 541. [[CrossRef](#)]
110. Takahashi, M.; Frøslev, T.G.; Paupério, J.; Thalinger, B.; Klymus, K.; Helbing, C.C.; Villacorta-Rath, C.; Silliman, K.; Thompson, L.R.; Jungbluth, S.P.; et al. A Metadata Checklist and Data Formatting Guidelines to Make EDNA FAIR (Findable, Accessible, Interoperable, and Reusable). *Environ. DNA* **2025**, *7*, e70100. [[CrossRef](#)]
111. Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B.-J.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F.; et al. The BioGRID Database: A Comprehensive Biomedical Resource of Curated Protein, Genetic, and Chemical Interactions. *Protein Sci.* **2021**, *30*, 187–200. [[CrossRef](#)] [[PubMed](#)]
112. Karp, P.D.; Billington, R.; Caspi, R.; Fulcher, C.A.; Latendresse, M.; Kothari, A.; Keseler, I.M.; Krummenacker, M.; Midford, P.E.; Ong, Q.; et al. The BioCyc Collection of Microbial Genomes and Metabolic Pathways. *Brief. Bioinform.* **2018**, *20*, 1085–1093. [[CrossRef](#)]
113. Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblit, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates. *Nucleic Acids Res.* **2021**, *49*, D498–D508. [[CrossRef](#)] [[PubMed](#)]
114. McDonald, A.G.; Boyce, S.; Tipton, K.F. ExplorEnz: The Primary Source of the IUBMB Enzyme List. *Nucleic Acids Res.* **2009**, *37*, D593–D597. [[CrossRef](#)] [[PubMed](#)]
115. Kanehisa, M.; Furumichi, M.; Sato, Y.; Kawashima, M.; Ishiguro-Watanabe, M. KEGG for Taxonomy-Based Analysis of Pathways and Genomes. *Nucleic Acids Res.* **2023**, *51*, D587–D592. [[CrossRef](#)] [[PubMed](#)]
116. Tang, Z.; Chen, S.; Chen, A.; He, B.; Zhou, Y.; Chai, G.; Guo, F.; Huang, J. CasPDB: An Integrated and Annotated Database for Cas Proteins from Bacteria and Archaea. *Database* **2019**, *2019*, baz093. [[CrossRef](#)] [[PubMed](#)]

117. Pourcel, C.; Touchon, M.; Villeriot, N.; Vernadet, J.P.; Couvin, D.; Toffano-Nioche, C.; Vergnaud, G. CRISPRCasdb a Successor of CRISPRdb Containing CRISPR Arrays and Cas Genes from Complete Genome Sequences, and Tools to Download and Query Lists of Repeats and Spacers. *Nucleic Acids Res.* **2020**, *48*, D535–D544. [CrossRef] [PubMed]
118. Adler, B.A.; Trinidad, M.I.; Bellieny-Rabelo, D.; Zhang, E.; Karp, H.M.; Skopintsev, P.; Thornton, B.W.; Weissman, R.F.; Yoon, P.H.; Chen, L.; et al. CasPEDIA Database: A Functional Classification System for Class 2 CRISPR-Cas Enzymes. *Nucleic Acids Res.* **2024**, *52*, D590–D596. [CrossRef] [PubMed]
119. Consortium, T.U. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2024**, *53*, D609–D617. [CrossRef] [PubMed]
120. Störtz, F.; Minary, P. CrisprSQL: A Novel Database Platform for CRISPR/Cas off-Target Cleavage Assays. *Nucleic Acids Res.* **2021**, *49*, D855–D861. [CrossRef] [PubMed]
121. Pu, Z.; Shi, C.-L.; Jeon, C.O.; Fu, J.; Liu, S.-J.; Lan, C.; Yao, Y.; Liu, Y.-X.; Jia, B. ChatGPT and Generative AI Are Revolutionizing the Scientific Community: A Janus-Faced Conundrum. *iMeta* **2024**, *3*, e178. [CrossRef] [PubMed]
122. Leiter, C.; Zhang, R.; Chen, Y.; Belouadi, J.; Larionov, D.; Fresen, V.; Eger, S. ChatGPT: A Meta-Analysis after 2.5 Months. *Mach. Learn. Appl.* **2024**, *16*, 100541. [CrossRef]
123. Moore, T. CRISPR AI Research Suite Version 0.1.0. 2026. Available online: <https://github.com/Tmmoore286/crispr-ai-research-suite> (accessed on 24 February 2026).
124. Qu, Y.; Huang, K.; Yin, M.; Zhan, K.; Liu, D.; Yin, D.; Cousins, H.C.; Johnson, W.A.; Wang, X.; Shah, M.; et al. CRISPR-GPT for Agentic Automation of Gene-Editing Experiments. *Nat. Biomed. Eng.* **2026**, *10*, 245–258. [CrossRef] [PubMed]
125. Secretaría de Ciencia, Humanidades, Tecnología e Innovación; Agencia de Transformación Digital y Telecomunicaciones. Declaración de Ética y Buenas Prácticas para el Uso y Desarrollo de la IA en México: SECIHTI y ATDT 2026. Available online: <https://seciht.mx/sala-de-prensa/presentan-declaracion-de-etica-y-buenas-practicas-para-el-uso-y-desarrollo-de-la-ia-en-mexico-seciht-y-atdt/> (accessed on 25 February 2026).
126. Resnik, D.B. Biosafety, Biosecurity, and Bioethics. *Monash Bioeth. Rev.* **2024**, *42*, 137–167. [CrossRef] [PubMed]
127. Ostos Ortiz, O.L. Edición Genética e Inteligencia Artificial: Desafíos Éticos Frente a Los Avances Biotecnológicos. *NOVA* **2024**, *22*, 43. [CrossRef]
128. AL-Eitan, L.; Alnemri, M. Biosafety and Biosecurity in the Era of Biotechnology: The Middle East Region. *J. Biosaf. Biosecur.* **2022**, *4*, 130–145. [CrossRef]
129. Soleimani Sasani, M. The Importance of Biosecurity in Emerging Biotechnologies and Synthetic Biology. *Avicenna J. Med. Biotechnol.* **2024**, *16*, 223–232. [CrossRef] [PubMed]
130. Federal Select Agent Program. Select Agents and Toxins List. 2025. Available online: <https://selectagents.gov/sat/list.htm> (accessed on 19 March 2026).
131. The Australia Group. List of Human and Animal Pathogens and Toxins for Export Control 2024. Available online: <https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/documents/common-control-lists/Common-Control-List-of-Dual-Use-Biological-Equipment.pdf> (accessed on 19 March 2026).
132. Bossi, P.; Garin, D.; Guihot, A.; Gay, F.; Crance, J.-M.; Debord, T.; Autran, B.; Bricaire, F. Biological Weapons. *Cell. Mol. Life Sci.* **2006**, *63*, 2196–2212. [CrossRef] [PubMed]
133. Zarate, S.; Cimadori, I.; Roca, M.M.; Jones, M.S.; Barnhill-Dilling, K. *Assessment of the Regulatory and Institutional Framework for Agricultural Gene Editing via CRISPR-Based Technologies in Latin America and the Caribbean*; Inter-American Development Bank: Washington, DC, USA, 2023. [CrossRef]
134. Flores-Coronado, J.A.; Alanis-Valdez, A.Y.; Herrera-Saldivar, M.F.; Flores-Flores, A.S.; Vazquez-Guillen, J.M.; Tamez-Guerra, R.S.; Rodriguez-Padilla, C. Awareness of the Dual-Use Dilemma in Scientific Research: Reflections and Challenges to Latin America. *Front. Bioeng. Biotechnol.* **2025**, *13*, 1649781. [CrossRef] [PubMed]
135. Li, X.; Gao, Y.; Zhang, Z.; Deng, W.; Cao, W.; Wei, X.; Gao, Z.; Yao, L.; Wang, S.; Xie, Y.; et al. Biosafety Considerations Triggered by Genome-Editing Technologies. *Biosaf. Health* **2025**, *7*, 141–151. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.