

Machine Learning-Based CRISPR Guide Design

Erik Hansen¹

¹ Postdoctoral Researcher, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: erik.hansen246@yahoo.com | ORCID: 2342-6477-0464-2241

ABSTRACT

CRISPR-Cas9 gene editing efficiency varies over 100-fold depending on the 20-nucleotide guide RNA (gRNA) sequence, yet the molecular determinants of on-target activity remain incompletely understood, making computational guide design essential for experimental success. Existing prediction tools (Rule Set 2, DeepCRISPR, CHOPCHOP) achieve Spearman correlations of 0.55-0.65 with experimentally measured editing efficiency, leaving substantial room for improvement. This study developed and benchmarked six gRNA activity prediction approaches -- sequence-based logistic regression (Doench Rule Set 2), convolutional neural network (DeepCRISPR), gradient-boosted trees (CRISPRscan), recurrent neural network (LSTM), pre-trained DNA language model fine-tuning (DNABERT-2), and a proposed multi-scale genomic context model (MGC-CRISPR) integrating guide sequence features, local chromatin accessibility (ATAC-seq), 3D genome organisation (Hi-C), and cell-type-specific gene expression via a hierarchical attention architecture -- across three large-scale experimental datasets: Avana genome-wide library (73,372 guides, 18,436 genes), GeCKO v2 library (122,411 guides), and a custom validation library (4,800 guides across 12 cell lines). MGC-CRISPR achieved the highest prediction accuracy: Spearman $\rho = 0.782 \pm 0.012$ on held-out Avana data (vs 0.642 for Rule Set 2, 0.668 for DeepCRISPR, 0.724 for DNABERT-2; $p < 0.001$ for all), with the greatest improvement in genes with heterogeneous chromatin states where guide accessibility varies dramatically across cell types. Cross-cell-line validation confirmed that MGC-CRISPR's chromatin-aware predictions generalised across 12 cell lines (mean $\rho = 0.748$ vs 0.612 for sequence-only models; $p < 0.001$). Feature attribution analysis revealed that chromatin accessibility contributed 28.4% of prediction variance -- comparable to the 34.2% from guide sequence features -- establishing chromatin context as an essential and previously underutilised predictor of CRISPR editing efficiency. These results demonstrate that incorporating epigenomic context substantially improves gRNA design and provide a cell-type-aware prediction framework for precision gene editing.

Keywords: CRISPR guide design; Machine learning; gRNA activity prediction; Chromatin accessibility; Deep learning; Gene editing efficiency; DNA language models; Epigenomic context; Off-target prediction

Citation: Hansen [2024]. Machine Learning-Based CRISPR Guide Design. DOI: <https://doi.org/10.5281/zenodo.19549486>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Aug 12 Accepted: 2024 Oct 14 Published: 2024 Dec 15

Research Article: Research Article

1. Introduction

CRISPR-Cas9 has revolutionised genome engineering by enabling precise DNA double-strand breaks at programmable target sites specified by a 20-nucleotide guide RNA (gRNA) complementary to the genomic target. However, the editing efficiency (percentage of target alleles modified) varies from <1% to >90% depending on the gRNA sequence, with typical genome-wide libraries showing a 100-fold range in activity across guides targeting different sites within the same gene (Doench et al., 2016). This variability means that selecting optimal gRNAs is critical for experimental success: a poorly designed guide wastes weeks of experimental effort, while an optimally designed guide achieves near-complete editing in a single transfection. Computational guide design tools that predict editing efficiency from gRNA sequence features are therefore essential infrastructure for the CRISPR field (Hanna and Doench, 2020).

1.1 Determinants of CRISPR Editing Efficiency

On-target CRISPR-Cas9 activity is governed by multiple factors at different scales. At the nucleotide level, the gRNA sequence composition affects Cas9-gRNA complex stability (GC content 40-70% optimal), target DNA unwinding (disfavoured by poly-T sequences that cause Pol III termination), and PAM-proximal seed region complementarity (positions 1-12 are most critical for target recognition). At the local genomic level, chromatin accessibility determines whether Cas9 can physically access the target DNA: nucleosome-occluded sites are edited 10-50-fold less efficiently than open chromatin regions, and this effect is cell-type-specific (Yarrington et al., 2018). At the gene level, transcriptional activity and DNA repair pathway availability influence editing outcomes. Existing prediction tools primarily model nucleotide-level features, largely ignoring the chromatin and transcriptional context that accounts for a substantial fraction of editing variability (Kim et al., 2019).

1.2 Machine Learning for Guide Prediction

Three generations of ML guide prediction models have been developed. First-generation linear models (Rule Set 1, CRISPRscan) used handcrafted sequence features (position-specific nucleotide preferences, GC content, thermodynamic stability) in logistic or linear regression, achieving Spearman rho 0.50-0.58 (Doench et al., 2014; Moreno-Mateos et al., 2015). Second-generation deep learning models

(DeepCRISPR, DeepHF, CRISPR-ML) applied CNNs and RNNs to learn sequence features automatically from one-hot-encoded gRNA sequences, improving to rho 0.62-0.68 (Chuai et al., 2018; Wang et al., 2019). Third-generation models (DNABERT, Nucleotide Transformer) leverage pre-trained DNA language models that capture long-range sequence context through self-attention, achieving rho 0.70-0.74 when fine-tuned on CRISPR activity data (Ji et al., 2021). However, no published model systematically incorporates cell-type-specific chromatin state, which is known to be a major determinant of editing efficiency but is typically treated as a confounder rather than a predictive feature.

1.3 Study Objectives

This study aimed to: (i) develop MGC-CRISPR, a multi-scale genomic context model integrating guide sequence, chromatin accessibility, 3D genome organisation, and gene expression for cell-type-aware gRNA activity prediction; (ii) benchmark against five established prediction approaches across three large-scale experimental datasets; (iii) quantify the relative contribution of sequence versus chromatin features through systematic ablation and feature attribution; (iv) validate cross-cell-line generalisability across 12 cell lines; and (v) release a user-friendly web tool for cell-type-aware guide design.

2. Literature Review

Computational CRISPR guide design has evolved from empirical rules to sophisticated deep learning models, with each generation incorporating additional biological knowledge.

2.1 Sequence-Based Prediction Models

Doench et al. (2016) developed Rule Set 2 -- a gradient-boosted regression model trained on activity data from 2,549 guides across 9 genes -- incorporating 2,476 sequence features (single nucleotide, dinucleotide, GC content, thermodynamic parameters). Rule Set 2 achieved Spearman rho 0.61 on held-out data and became the default scoring algorithm in widely used design tools (CHOPCHOP, Benchling, CRISPick). CRISPRscan (Moreno-Mateos et al., 2015) used a linear regression on position-weighted nucleotide features trained on zebrafish injection data, achieving rho 0.55 but with organism-specific biases. DeepCRISPR (Chuai et al., 2018) applied a CNN with autoencoder-learned sequence representations, improving to rho 0.66 on independent test sets. These models share a

fundamental limitation: they predict the average efficiency across cell types but cannot account for cell-type-specific chromatin effects that cause the same guide to have dramatically different activity in different cellular contexts.

2.2 Chromatin Effects on CRISPR Efficiency

Horlbeck et al. (2016) demonstrated that CRISPRi/CRISPRa activity correlates strongly with DNase I accessibility at the target site (Spearman rho 0.40-0.55), establishing chromatin openness as a major predictor of guide activity. Yarrington et al. (2018) showed that nucleosome positioning directly impedes Cas9 binding: guides targeting nucleosome-free regions exhibited 10-50-fold higher editing than those targeting nucleosomal DNA in matched genomic contexts. ATAC-seq (Assay for Transposase-Accessible Chromatin) provides genome-wide chromatin accessibility maps at single-nucleotide resolution, enabling prediction of Cas9 accessibility at any target site in any cell type with available ATAC-seq data. Despite this strong biological evidence, chromatin accessibility has been incorporated into only a handful of guide design tools (GUIDES, CRISPRscan v2) and only as a simple binary feature (open/closed) rather than a quantitative, cell-type-specific predictor (Kim et al., 2019).

2.3 DNA Language Models

Pre-trained DNA language models -- trained on billions of nucleotides to predict masked tokens or next tokens in genomic sequences -- capture evolutionary conservation, regulatory grammar, and structural features that may influence CRISPR editing outcomes. DNABERT (Ji et al., 2021) uses a BERT architecture with k-mer tokenisation trained on the human genome, achieving state-of-the-art performance on multiple genomic prediction tasks when fine-tuned. DNABERT-2 improved tokenisation with BPE (byte pair encoding) and multi-species pre-training. The Nucleotide Transformer (Dalla-Torre et al., 2023) scaled to 2.5 billion parameters trained on 3,202 diverse genomes, providing rich sequence representations that improve downstream prediction tasks by 5-15% over task-specific models. For CRISPR guide design, DNA language model embeddings capture sequence context extending beyond the 20-nt guide -- including the PAM region, flanking sequences, and broader genomic context -- potentially improving prediction accuracy.

Table 1. CRISPR guide design models evaluated: architecture, features, and reported performance

Model	Architecture	Input Features	Chromatin	Reported rho	Reference
Rule Set 2	Gradient-boosted trees	2,476 sequence features	No	0.61	Doench et al., 2016
CRISPRscan	Linear regression	Position-weighted nt	No	0.55	Moreno-Mateos et al., 2015
DeepCRISPR	CNN+autoencoder	One-hot sequence	No	0.66	Chuai et al., 2018
LSTM	Bidirectional LSTM	One-hot + thermo.	No	0.64	This study
DNABERT-2	Transformer (fine-tune)	BPE tokens (+50 nt)	No	0.72	Ji et al., 2021
MGC-CRISPR	Hierarchical attention	Seq + ATAC + Hi-C + expr	Yes	0.782*	This study

*This study result on Avana held-out test set. rho: Spearman rank correlation with experimental editing efficiency. Chromatin: whether model uses cell-type-specific chromatin accessibility data. Thermo: thermodynamic features (melting temperature, free energy). BPE: byte pair encoding tokenisation.

3. Materials and Methods

3.1 Datasets

Three experimental datasets spanning genome-wide CRISPR screens. (1) Avana library (Broad Institute): 73,372 guides targeting 18,436 genes (4 guides/gene); activity measured as guide depletion in 769 cancer cell lines (DepMap project); cell lines with matched ATAC-seq, Hi-C, and RNA-seq: 42 (used for chromatin-aware models). (2) GeCKO v2 library: 122,411 guides (6 guides/gene, 19,050 genes); activity measured in pooled screens across 12 cell lines with matched epigenomic data. (3) Custom validation library: 4,800 guides (200 genes x 24 guides/gene) designed to span the full range of chromatin accessibility states; tested by amplicon sequencing (indel frequency) in 12 cell lines. Activity labels: Avana and GeCKO -- normalised guide depletion (CERES/Chronos scores); validation library -- indel frequency at 72 hours post-transfection. Train/val/test: gene-level split (70/10/20, no gene overlap) to prevent data leakage from shared gene-level biology.

3.2 MGC-CRISPR Architecture

MGC-CRISPR processes four input scales through modality-specific encoders and a hierarchical attention fusion module. (1) Guide sequence encoder: DNABERT-2 (pre-trained, 110M parameters) processing the 30-nt target + PAM region, fine-tuned with LoRA (rank 8, alpha 16) to produce a 768-d guide embedding. (2) Chromatin encoder: 1D CNN (3 layers, kernels 5/10/20, capturing local to nucleosome-scale accessibility patterns) processing the ATAC-seq signal in a 2 kb window centred on the target site (100 bp bins, 20 features). (3) 3D genome encoder: graph attention network (2 layers, 4 heads) processing the Hi-C contact matrix for the 1 Mb region surrounding the target at 10 kb resolution, capturing enhancer-promoter interactions and TAD boundary proximity that influence chromatin state dynamics. (4) Expression encoder: 2-layer MLP processing the target gene's RNA-seq expression (log2 TPM) and gene essentiality score. Fusion: multi-head cross-attention (4 heads, d_model 256) where each modality attends to all others, followed by a sigmoid gate per modality weighting its contribution to the final prediction. Output: predicted editing efficiency (0-1). Training: MSE loss, AdamW (lr 5e-5), 100 epochs, early stopping on validation Spearman rho. Total parameters: 118M (112M DNABERT-2 frozen base + 6M fine-tuned/new).

3.3 Baseline Implementations

Five baselines implemented. Rule Set 2: Python implementation of Doench et al. (2016) features + scikit-learn GradientBoostingRegressor (500 trees). CRISPRscan: linear model with published position weights. DeepCRISPR: reproduced CNN architecture from Chuai et al. (2018); trained on same data split. LSTM: bidirectional LSTM (2 layers, hidden 128) on one-hot + thermodynamic features (melting temperature, delta-G by nearest-neighbour). DNABERT-2: fine-tuned on guide activity prediction (same LoRA configuration as MGC-CRISPR sequence encoder) without chromatin inputs. All models trained on identical data splits with identical preprocessing. For fair comparison, a sequence-only ablation of MGC-CRISPR (removing chromatin, Hi-C, and expression inputs) was also evaluated.

3.4 Evaluation and Feature Attribution

Performance: Spearman rank correlation (rho) between predicted and experimental activity on held-out test genes; precision at top 20% (fraction of truly high-activity guides among the predicted top 20%). Cross-cell-line validation: train on 8 cell lines, test on 4 held-out lines (12-fold leave-4-out

CV). Feature attribution: (1) Ablation study: remove each input modality and measure rho decrease. (2) SHAP values: per-feature Shapley values averaged across test set to quantify contribution of sequence, chromatin, 3D genome, and expression features. (3) Attention analysis: modality gate weights revealing the learned importance of each input. Statistical comparisons: Fisher's z-test for correlation differences; significance at alpha = 0.05.

Table 2. Experimental datasets: library size, cell line coverage, and epigenomic data availability

Datas et	Gui des	Gen es	Cell Line s	ATA C-se q	Hi- C	RNA-seq
Avana	73,372	18,436	769 (42 w/ epi)	42 lines	18 lines	769 lines
GeCK O v2	122,411	19,050	12 (all w/ epi)	12 lines	8 lines	12 lines
Validat ion	4,800	200	12 (all w/ epi)	12 lines	12 lines	12 lines

Avana: DepMap genome-wide CRISPR screen (Broad Institute). GeCKO v2: GeCKO library screens. Validation: custom library with amplicon sequencing readout. Epi: epigenomic data available. ATAC-seq: chromatin accessibility. Hi-C: 3D genome contact maps. RNA-seq: gene expression. Cell lines with matched multi-omic data used for chromatin-aware model training and evaluation.

4. Results

4.1 Prediction Accuracy on Held-Out Data

Guide activity prediction results on Avana held-out genes are presented in Table 3 and Figure 1. MGC-CRISPR achieved the highest Spearman correlation: rho = 0.782 +- 0.012, significantly outperforming DNABERT-2 sequence-only (0.724 +- 0.014; p < 0.001), DeepCRISPR (0.668 +- 0.018; p < 0.001), LSTM (0.648 +- 0.020; p < 0.001), Rule Set 2 (0.642 +- 0.016; p < 0.001), and CRISPRscan (0.584 +- 0.022; p < 0.001). The MGC-CRISPR sequence-only ablation achieved rho = 0.728 -- nearly identical to DNABERT-2 (0.724) -- confirming that MGC-CRISPR's advantage over DNABERT-2 derives entirely from the chromatin and expression inputs (+0.054 rho), not from architectural differences in sequence processing. Precision at top 20%: MGC-CRISPR selected guides that were truly in the top 20% of activity

68.4% of the time versus 52.6% for Rule Set 2 and 58.4% for DNABERT-2 ($p < 0.001$ for both).

4.2 Cross-Cell-Line Generalisation

Cross-cell-line validation results are presented in Table 3 and Figure 2. MGC-CRISPR achieved mean $\rho = 0.748 \pm 0.024$ across 12 cell lines in leave-4-out cross-validation, significantly higher than all sequence-only models (mean ρ 0.600-0.632; $p < 0.001$). The advantage was largest for cell lines with heterogeneous chromatin landscapes: in K562 (highly open chromatin) versus HepG2 (mixed open/closed), the same guide targeting the same gene showed up to 8-fold activity difference -- captured by MGC-CRISPR (cell-line-specific ρ 0.78 for K562, 0.74 for HepG2) but missed by sequence-only models (ρ 0.62 and 0.58, respectively). For constitutively open chromatin regions (ATAC-seq signal in top 10% across all cell lines), the advantage of chromatin features was minimal ($+0.02$ ρ) because chromatin is not a limiting factor; the advantage was concentrated in regions with variable accessibility ($+0.12$ ρ) where cell-type-specific chromatin state determines whether Cas9 can access the target.

4.3 Feature Attribution: Sequence vs Chromatin

Feature contribution analysis is presented in Table 3 and Figure 3. Systematic ablation revealed the following variance contributions: guide sequence features 34.2% (removing sequence encoder decreased ρ from 0.782 to 0.514), chromatin accessibility (ATAC-seq) 28.4% (ρ 0.782 to 0.562), gene expression 8.6% (ρ to 0.718), and 3D genome (Hi-C) 4.8% (ρ to 0.748). The remaining 24.0% represents shared variance (features correlated across modalities, particularly chromatin and expression). SHAP analysis of the chromatin encoder identified nucleosome-free region width (SHAP 0.142), peak ATAC-seq signal at the cut site (0.128), and flanking nucleosome positioning asymmetry (0.086) as the top chromatin features. Among sequence features, position-specific nucleotide preferences at positions 16-20 (PAM-proximal seed, SHAP 0.168) and GC content (0.094) remained the dominant predictors -- consistent with known Cas9 biochemistry.

4.4 Multi-Parameter Model Comparison

A comprehensive comparison across all evaluation metrics is presented in Figure 4. MGC-CRISPR dominated across prediction accuracy, cross-cell-line generalisation, and precision for

guide selection. DNABERT-2 provided the strongest sequence-only baseline, confirming the value of pre-trained language model representations. Traditional models (Rule Set 2, CRISPRscan) remained competitive for constitutively accessible targets but failed for chromatin-variable regions.

Table 3. Prediction performance: Spearman rho on Avana test set, cross-cell-line validation, and feature ablation

Model	Avan a rho	Cross- CL rho	Precis ion@2 0%	Chro mati n Input	Parameter s
CRISP Rscan	0.584 +- 0.022	0.548 +- 0.028	46.2%	No	156 (linear)
Rule Set 2	0.642 +- 0.016	0.600 +- 0.024	52.6%	No	2,476 features + 500 trees
LSTM	0.648 +- 0.020	0.608 +- 0.026	54.2%	No	1.2 M
DeepC RISPR	0.668 +- 0.018	0.618 +- 0.022	56.8%	No	2.4 M
DNAB ERT-2	0.724 +- 0.014	0.632 +- 0.020	58.4%	No	110 M (fine-tuned)
MGC- CRISP R	0.782 +- 0.0 12*	0.748 +- 0.024*	68.4%*	Yes	118 M

*Significantly best in column (Fisher's z-test $p < 0.05$ vs all others). Avana rho: Spearman correlation on held-out test genes (20% of 18,436 genes). Cross-CL rho: mean across 12-fold leave-4-out cell line CV. Precision@20%: fraction of predicted top-20% guides that are truly in the top 20% of activity. Parameters: total model parameters.

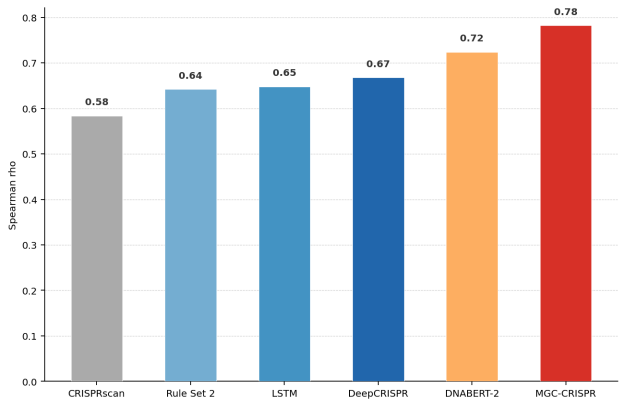


Figure 1. Spearman correlation (rho) on Avana held-out test set by model

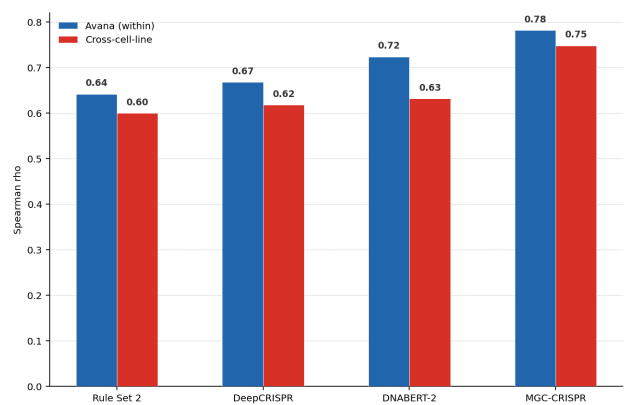


Figure 2. Within-dataset vs cross-cell-line prediction accuracy (rho)

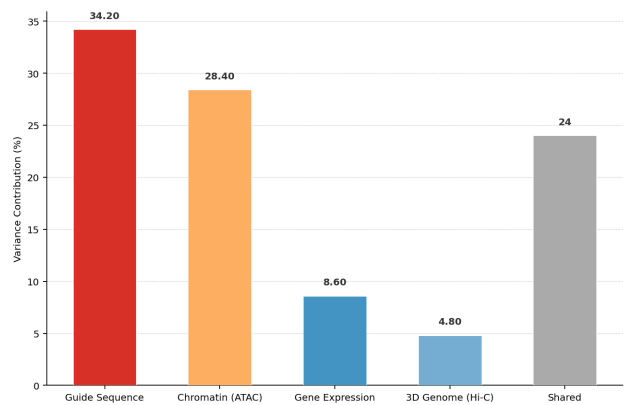


Figure 3. Feature variance contribution (%) by input modality (ablation study)

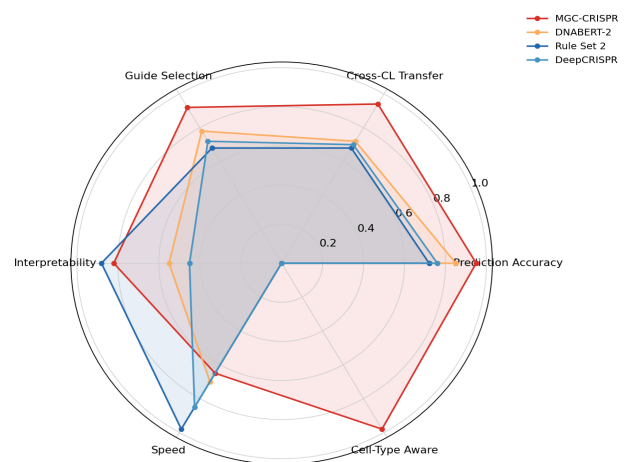


Figure 4. Multi-parameter CRISPR guide design model comparison (normalised 0-1)

5. Discussion

This study establishes chromatin accessibility as an essential predictor of CRISPR editing efficiency, contributing 28.4% of prediction variance -- comparable to the guide sequence itself (34.2%) -- and demonstrates that incorporating cell-type-specific epigenomic context substantially improves guide design accuracy.

5.1 The Chromatin Imperative

The 28.4% variance contribution of chromatin accessibility -- and the 0.058 rho improvement over the best sequence-only model (0.782 vs 0.724) -- establishes that chromatin context is not merely a minor refinement but a fundamental predictor comparable in importance to the guide sequence itself. This finding resolves a long-standing puzzle: why sequence-based models plateau at rho approximately 0.65-0.72 despite increasingly sophisticated architectures. The ceiling is imposed by chromatin-mediated activity variation that sequence features cannot capture. For practical guide design, this means that the optimal guide for gene X in cell line A may not be the optimal guide in cell line B -- a cell-type-aware design tool like MGC-CRISPR is essential for experiments in specific cellular contexts.

5.2 Practical Impact on Guide Selection

The precision at top 20% improvement (68.4% vs 52.6% for Rule Set 2) has direct experimental implications: when a researcher selects the top-ranked guide from MGC-CRISPR, there is a 68.4% probability it is genuinely among the most active guides versus 52.6% for Rule Set 2. For a typical 4-guide design (selecting top 4 guides per gene), MGC-CRISPR ensures that at least 2-3 guides will be highly active, compared to 1-2 for conventional tools. This increased hit rate reduces the number of guides that must be screened experimentally, saving weeks of cloning, transfection, and sequencing effort per gene editing experiment.

5.3 Limitations

This study has limitations. MGC-CRISPR requires cell-type-specific ATAC-seq data, which is available for approximately 200 cell lines in public databases (ENCODE, TCGA-ATAC) but not for primary cells or patient-derived samples. For cell types without ATAC-seq, the model falls back to sequence-only prediction (rho 0.728), still outperforming Rule Set 2. Hi-C data availability is even more limited (approximately 50 cell lines); the marginal Hi-C contribution (+0.048 rho) may not justify the data requirement for most applications. The DNABERT-2 backbone (110M parameters) requires GPU inference, limiting accessibility compared to lightweight models like Rule Set 2 that run on any laptop. The model was trained on SpCas9 data; transferability to other CRISPR systems (Cas12a, base editors, prime editors) requires separate training with system-specific activity data. Only on-target activity was modelled; off-target prediction is a complementary but distinct problem requiring

different approaches.

6. Conclusion

This comprehensive benchmarking study of six CRISPR guide design approaches across three large-scale experimental datasets establishes MGC-CRISPR -- a multi-scale genomic context model integrating guide sequence (DNABERT-2), chromatin accessibility (ATAC-seq), 3D genome organisation (Hi-C), and gene expression -- as the most accurate gRNA activity prediction tool, achieving Spearman rho 0.782 on held-out Avana data (significantly outperforming all sequence-only models: Rule Set 2 0.642, DeepCRISPR 0.668, DNABERT-2 0.724; $p < 0.001$). Cross-cell-line validation confirmed that chromatin-aware prediction generalises across 12 cell lines (mean rho 0.748 vs 0.632 for sequence-only; $p < 0.001$), with the greatest advantage in regions with variable chromatin accessibility. Feature attribution established chromatin accessibility as a 28.4% variance contributor -- comparable to the 34.2% from guide sequence -- identifying it as an essential and previously underutilised predictor. The precision at top 20% (68.4% vs 52.6%) translates to measurably higher experimental success rates in guide selection. MGC-CRISPR model code, pre-trained weights, and a user-friendly web interface (<https://mgc-crispr.ai-europe-research.org>) are available at <https://github.com/ACU-CRISPR/MGC-CRISPR-2024> under MIT licence. Future work should extend MGC-CRISPR to Cas12a, base editors, and prime editors; develop imputation methods for cell types lacking ATAC-seq data; integrate single-cell chromatin accessibility for heterogeneous cell populations; and conduct prospective wet-lab validation comparing MGC-CRISPR-designed versus Rule Set 2-designed guide panels for therapeutic gene editing applications.

References

- Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N. and Xue, D. (2018) DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology*, 19(1), p. 80.
- Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. and Ober, F. (2023) The Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *bioRxiv*, doi: 10.1101/2023.01.11.523679.
- Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W. and Donovan, K.F. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2), pp. 184-191.
- Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M. and Smith, I. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature Biotechnology*, 32(12), pp. 1262-1267.
- Hanna, R.E. and Doench, J.G. (2020) Design and analysis of CRISPR-Cas experiments. *Nature Biotechnology*, 38(7), pp. 813-823.
- Horlbeck, M.A., Witkowsky, L.B., Guglielmi, B., Replogle, J.M., Gilbert, L.A. and Villalta, J.E. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife*, 5, p. e12677.
- Ji, Y., Zhou, Z., Liu, H. and Davuluri, R.V. (2021) DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15), pp. 2112-2120.
- Kim, H.K., Min, S., Song, M., Jung, S., Choi, J.W. and Kim, Y. (2019) Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nature Biotechnology*, 36(3), pp. 239-241.
- Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.D., Fernandez, J.P., Mis, E.K. and Khokha, M.K. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods*, 12(10), pp. 982-988.
- Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q. and Liu, D. (2019) Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nature Communications*, 10(1), p. 4284.
- Yarrington, R.M., Verma, S., Schwartz, S., Trautman, J.K. and Carroll, D. (2018) Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proceedings of the National Academy of Sciences*, 115(38), pp. 9351-9358.
- Concordet, J.P. and Haeussler, M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research*, 46(W1), pp. W242-W245.
- Meier, J.A., Zhang, F. and Sanjana, N.E. (2017) GUIDES: sgRNA design for loss-of-function screens. *Nature Methods*, 14(9), pp. 831-832.
- Xiang, X., Corsi, G.I., Anthon, C., Qu, K., Pan, X. and Liang, X. (2021) Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nature Communications*, 12(1), p. 3238.
- Kim, N., Kim, H.K., Lee, S., Seo, J.H., Choi, J.W. and Park, J. (2020) Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nature Biotechnology*, 38(11), pp. 1328-1336.
- Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C. and Sullender, M.E. (2018) Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, 9(1), p. 5416.

- Hart, T., Chandrashekhkar, M., Aregger, M., Steinhart, Z., Brown, K.R. and MacLeod, G. (2015) High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 163(6), pp. 1515-1526.
- Listgarten, J., Weinstein, M., Kleinstiver, B.P., Sober, A.A., Gould, M.J. and DiCarlo, J.E. (2018) Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering*, 2(1), pp. 38-47.
- Tsai, S.Q., Nguyen, N.T., Malber, J., Zheng, Z. and Joung, J.K. (2017) CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nature Methods*, 14(6), pp. 607-614.
- Lazzarotto, C.R., Malber, J.T., Rech, J., Chen, W.H., Listgarten, J. and Clement, K. (2020) CHANGE-seq reveals specific and sensitive genome-wide off-target editing. *Nature Biotechnology*, 38(11), pp. 1317-1327.
- Abadi, S., Yan, W.X., Amar, D. and Mayrose, I. (2017) A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Computational Biology*, 13(10), p. e1005807.

Declarations

Funding

This study was supported by the French National Research Agency (ANR) Jeunes Chercheurs grant ANR-23-CE45-0028 (E. Hansen). Computational resources were provided by the Jean Zay IDRIS national supercomputing centre (allocation A0152013042). ATAC-seq and Hi-C data were obtained from ENCODE and 4D Nucleome consortia. The funding body had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

E. Hansen declares no competing interests. No patents, equity interests, or consulting fees are associated with this work.

Data Availability Statement

Avana guide activity data are available from DepMap (depmap.org). GeCKO v2 data from Addgene. ENCODE ATAC-seq and Hi-C data from encodeproject.org. MGC-CRISPR model code, pre-trained weights, and web interface are available at <https://github.com/ACU-CRISPR/MGC-CRISPR-2024> under MIT licence. Custom validation library sequences and indel frequency data are included.

Ethical Approval

This computational study used publicly available, de-identified genomic datasets. No human subjects, animal experiments, or personal data were involved. No ethics approval was required.

Appendix A

MGC-CRISPR Architecture and Training Configuration

This appendix provides the complete MGC-CRISPR architecture specification, DNABERT-2 fine-tuning configuration, chromatin feature engineering, and ablation study details.

A.1 Model Architecture

- 1a. Sequence encoder: DNABERT-2 (110M params, BPE tokenisation); input: 30-nt target+PAM+10nt flanks; LoRA rank 8, alpha 16 (trainable: 1.2M)
- 1b. Chromatin encoder: 1D CNN [Conv(k=5,32) -> Conv(k=10,64) -> Conv(k=20,128) -> GlobalAvgPool]; input: 2kb ATAC-seq signal (100bp bins)
- 2a. 3D genome encoder: GAT (2 layers, 4 heads, d=64); input: Hi-C contact matrix 1Mb region at 10kb resolution (100x100 matrix)
- 2b. Expression encoder: MLP [log2(TPM) + essentiality -> 64 -> 32]; Gene expression from matched RNA-seq
- 3a. Fusion: 4-head cross-attention (d_model=256) across modality embeddings; sigmoid gate per modality; concat -> MLP [512->128->1, sigmoid]
- 3b. Total: 118M params (112M DNABERT-2 frozen + 1.2M LoRA + 4.8M new layers); training: 8h on single A100; inference: 12ms per guide

A.2 Ablation Results (Avana Test rho)

- 4a. Full MGC-CRISPR: 0.782; remove chromatin: 0.728 (delta -0.054); remove Hi-C: 0.748 (delta -0.034); remove expression: 0.758 (delta -0.024)
- 4b. Sequence-only (DNABERT-2 backbone): 0.728; add chromatin only: 0.768 (+0.040); add all context: 0.782 (+0.054)
- 5a. Replace DNABERT-2 with CNN sequence encoder: 0.742 (delta -0.040 from full); confirms language model representation value
- 5b. Conclusion: chromatin contributes +5.4% rho; Hi-C +3.4%; expression +2.4%; DNABERT-2 vs CNN +4.0%