Multimodal Modeling of CRISPR-Cas12 Activity Using Foundation Models and Chromatin Accessibility Data

Azim Dehghani Amirabad [1] **Yanfei Zhang** [1] **Artem Moskalev** [1] **Sowmya Rajesh** [1] **Tommaso Mansi** [1] **Shuwei Li** [1] **Mangal Prakash*** [1] **Rui Liao*** [1]

Abstract

Predicting guide RNA (gRNA) activity is critical for effective CRISPR-Cas12 genome editing but remains challenging due to limited data, variation across protospacer adjacent motifs (PAMs—short sequence requirements for Cas binding), and reliance on large-scale training. We investigate whether pre-trained biological foundation model—originally trained on transcriptomic data—can improve gRNA activity estimation even without domain-specific pre-training. Using embeddings from existing RNA foundation model as input to lightweight regressor, we show substantial gains over traditional baselines. We also integrate chromatin accessibility data to capture regulatory context, improving performance further. Our results highlight the effectiveness of pre-trained foundation models and chromatin accessibility data for gRNA activity prediction.

1. Introduction

CRISPR-based genome editing enables programmable DNA modification, with Cas12 (Cpf1) offering advantages over the widely used Cas9 due to its distinct protospacer adjacent motif (PAMs—short sequence requirements for Cas binding) requirements and cleavage patterns (Zetsche et al., 2015; Kleinstiver et al., 2016). A core challenge in CRISPR applications is predicting the activity of guide RNAs (gRNAs), short sequences that direct the Cas protein to specific genomic loci. Effective gRNA selection is critical for both editing efficiency and safety (Doench et al., 2016).

Existing models use handcrafted features or one-hot encoded sequences to predict gRNA activity, achieving modest performance (Hsu et al., 2013; Xu et al., 2015). More recent deep learning approaches, including CNNs trained on large-scale screens, have improved prediction accuracy and incorporated additional features such as chromatin accessibility to account for regulatory context (Kim et al., 2018a). However, these models are typically sensitive to specific experimental settings, such as the variant of the genome-editing enzyme (Cas protein) used and the local sequence motifs required for binding (PAM sequence). As a result, they often fail to generalize across different enzyme variants or sequence constraints, limiting their applicability to broader genome engineering tasks.

In contrast, pre-trained biological Foundation Models (FMs) have shown strong generalization across a variety of genomic, transcriptomic, and proteomic tasks (Brandes et al., 2022; Chen et al., 2022; Frazer et al., 2021; Gao et al., 2022; Elnaggar et al., 2021). These models produce rich, contextualized embeddings that capture structural and regulatory signals beyond local sequence patterns. However, their utility for CRISPR gRNA prediction—particularly for Cas12—remains unexplored. Unlike DNA, RNA, and protein domains where pre-training such models is feasible due to abundant data, the gRNA space lacks sufficient data to support large-scale pre-training. This raises a critical question: Can we instead adapt existing DNA or RNA FMs to improve gRNA activity prediction, even though gRNAs are short (30-50 nucleotides) and may lie outside the distribution existing off-the-shelf FMs were trained on? From a biological standpoint, this is a plausible strategy—gRNA sequences are functionally embedded in broader genomic and transcriptomic contexts, and their activity depends on features such as sequence composition, secondary structure, and chromatin accessibility that are captured in part by nucleotide-based FMs.

In this work, we evaluate two different FMs: RNA-FM (Chen et al., 2022), trained on transcriptomes, and DNABERT-2 (Zhou et al., 2024), trained on genomic DNA. Using their embeddings as fixed input to lightweight CNN regressor, we evaluate their effectiveness for downstream gRNA on-target efficiency prediction. We incorporate chromatin accessibility profiles from ATAC-seq assay as an ad-

*Equal contribution [1]Johnson & Johnson Innovative Medicine. Correspondence to: Azim Dehghani Amirabad <azim.dehghani@gmail.com>, Mangal Prakash <mpraka12@its.jnj.com>, Rui Liao <rliao2@its.jnj.com>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

ditional input modality, capturing the openness of genomic regions and their broader regulatory context. This epigenetic signal complements sequence information and can enhance prediction accuracy, even when using strong pretrained embeddings. To sum up, we make the following contributions.

- We demonstrate that embeddings from a transcriptomic foundation model significantly outperform existing baselines in predicting CRISPR-Cas12 gRNAs, showcasing their ability to generalize to out-of-distribution sequences.
- We introduce a new dataset derived from public source curation, aligning ATAC-seq chromatin accessibility data to gRNA target loci. This structured epigenomic signal not only enhances predictions when integrated with sequence embeddings but also provides valuable resources for the community to advance research in this area.

Together, our results establish the first framework for CRISPR gRNA modeling that leverages pre-trained sequence representations and biologically grounded regulatory features.

2. Related Works

2.1. gRNA activity prediction

Early models for predicting gRNA activity relied on rule-based scoring or engineered sequence features derived from small datasets (Hsu et al., 2013; Xu et al., 2015; Moreno-Mateos et al., 2015), but offered limited generalization. DeepCpf1 (Kim et al., 2018a) introduced a CNN-based framework trained on one-hot encoded Cas12a sequences, integrating chromatin accessibility to improve prediction accuracy. Since then, deep learning models have been applied to Cas9 and Cas12 systems, but often remain PAM-specific and trained end-to-end on narrow datasets. Recently, gRNA-FM (Zhou et al., 2023) proposed a foundation model tailored for Cas9 guide design, but it does not address Cas12 or activity prediction. Our work fills this gap by leveraging general-purpose biological FMs for Cas12 gRNA activity prediction, independent of PAM motifs and with integrated epigenomic context.

2.2. Foundation models for DNA and RNA

Foundation models (FMs) trained on large-scale biological sequence data have shown strong generalization across a variety of genomic and transcriptomic tasks. FMs such as DNABERT (Ji et al., 2021), DNABERT-2 (Zhou et al., 2024), Evo (Nguyen et al., 2024), Evo-2 (Brixi et al., 2025), Nucleotide Transformer (Dalla-Torre et al., 2025) are pretrained on genomic DNA, and have been applied to promoter identification, variant effect prediction, and regulatory element classification. Similarly, RNA-FM (Chen et al., 2022), Rinalmo (Penić et al., 2024), SpliceBERT (Chen et al., 2023), UTR-BERT (Yang et al., 2024) and other transcriptomic FMs, capture regulatory and structural features relevant to RNA function and folding. These models generate rich, contextualized embeddings without task-specific supervision. In this work, we adopt RNA-FM and DNABERT-2 as strong, domain-representative FM backbones and evaluate their embeddings for Cas12 gRNA activity prediction.

3. Methods

3.1. Pretrained Sequence Embeddings

We model gRNA activity using pre-trained biological FMs to generate contextualized sequence representations. Specifically, we leverage two transformer-based models: **RNA-FM** (Chen et al., 2022), trained on transcriptomic sequences, and the **DNABERT-2** (Zhou et al., 2024), trained on large genomic DNA corpora owing to their reported SOTA performance in recent studies (Chen et al., 2022; Yazdani-Jahromi et al., 2024; Prakash et al., 2024; Zhou et al., 2024).

For each gRNA, we extracted not only the core 20-nt target binding sequence but also extended sequence contexts of lengths 34, and 50 nucleotides centered around the cleavage site. This enabled us to systematically investigate how varying amounts of surrounding sequence context influences predictive modeling of gRNA activity. These are passed through the frozen FM backbones to obtain fixed-length embeddings, which are used as inputs to a light-weight downstream prediction head. This setup, commonly referred as probing (Prakash et al., 2024; Lin et al., 2023) enables the use of rich, pre-trained embeddings while avoiding the need for retraining the base models.

To ensure fair comparison with prior work, we adopt the convolutional regression architecture from (Kim et al., 2018b)—a state-of-the-art model for Cas12 gRNA activity prediction—as our downstream head, training it on top of pretrained FM embeddings to predict Cas12 activity. This isolates the effect of the FM backbone, as all other components are kept identical to (Kim et al., 2018b). The architecture consists of convolutional and pooling layers followed by fully connected layers, terminating in a single scalar output (hyperparameters and training details reported in Appendix sections A.3 and A.4). This design using FM with CNN prediction head captures local dependencies while remaining efficient for low-data regimes. We refer to this setup—comprising the frozen foundation model backbone paired with the CNN regression head—as Cas-FM (RNA-FM) or Cas-FM (DNABERT-2), depending on the backbone employed.

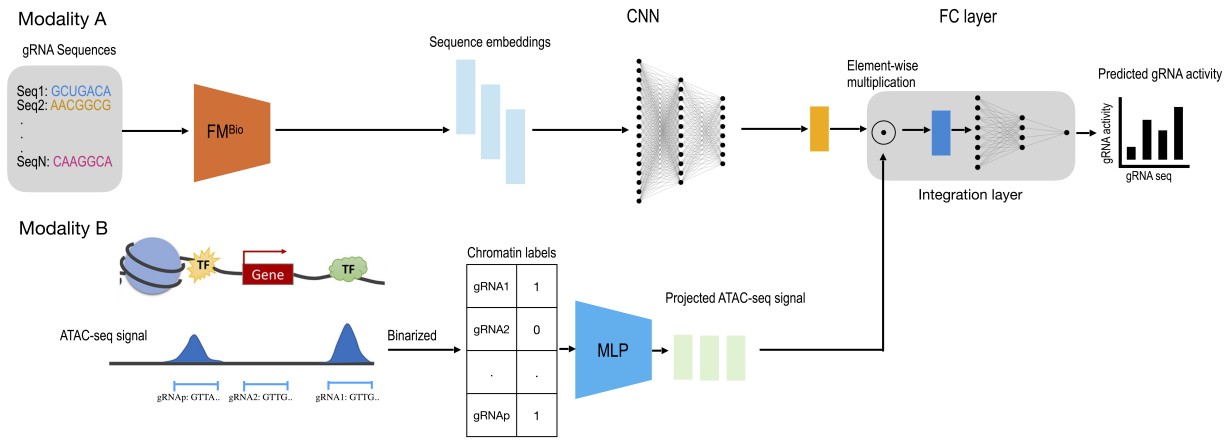

Figure 1. Multimodal framework for predicting CRISPR-Cas12 gRNA activity. gRNA sequences (Modality A) are encoded using pretrained biological foundation models (FM_{Bio}), and the resulting embeddings are processed by a CNN to learn sequence representations. Chromatin accessibility signals (Modality B), derived from binarized ATAC-seq peaks, are projected via an MLP into the same latent space as the sequence features. The two modalities are integrated through an element-wise multiplication layer, followed by a regression head to predict gRNA activity.

3.2. Chromatin Accessibility as an Epigenomic Modality

To incorporate regulatory context, we integrate chromatin accessibility (CA) information derived from ATAC-seq data (GEO accession: GSM2902624) for the gRNA sequences considered (details of corresponding gRNA data is provided in Sec. 4). Raw reads were aligned to the hg38 reference genome using `Bowtie2` (Langmead & Salzberg, 2012), and peaks were called using `MACS2` (Zhang et al., 2008). We then intersected the resulting peak coordinates with gRNA target loci using `BEDTools` (Quinlan & Hall, 2010) to assign binary accessibility labels—gRNAs overlapping a peak with normalized signal > 0.001 were labeled accessible; others were marked inaccessible.

These binary CA labels were projected into a dense embedding using a multilayer perceptron (MLP), forming the epigenomic input stream. This projected signal was integrated with the sequence-derived feature map from the pretrained FM (backbone) + CNN (downstream head) pipeline via element-wise multiplication in a dedicated integration layer. Keeping the downstream architecture fixed across unimodal (sequence-only) and multimodal (sequence + CA) settings allows us to isolate the contribution of chromatin accessibility to gRNA activity prediction (see Fig. 1). The exact architecture details are described in Appendix A.3.

4. Experiments and results

Evaluation Dataset. We evaluate our approach using the CRISPR-Cas12a (Cpf1) dataset introduced by Kim et al. (2018a), which contains approximately 15,000 guide RNA (gRNA) sequences paired with experimentally measured cleavage efficiencies. These activity scores reflect the capacity of each guide to induce gene disruption in human cells, making them a biologically grounded target for regression. The dataset spans a range of sequence contexts, including variation in PAM sequences, flanking sequences, and spacer regions, enabling robust model evaluation across heterogeneous genomic backgrounds. More details are provided in Appendix A.1.

Baselines. To rigorously benchmark our approach, we compare against a range of baselines commonly used in CRISPR gRNA activity prediction. These include CINDEL (Kim et al., 2017), a logistic regression model originally developed for indel frequency prediction; a boosted regression tree model (Boosted RT) implemented via XGBoost, using k-mer and positional features (Chen & Guestrin, 2016); and three regularized linear regression models—Lasso (L1)(Tibshirani, 1996), Ridge (L2)(Hoerl & Kennard, 1970), and ElasticNet (L1L2) (Zou & Hastie, 2005)—each trained on curated sets of biologically informed features such as nucleotide composition, GC content, and dinucleotide frequencies. These baselines typically use handcrafted sequence features or 1-hot representation of gRNA (see Table 1). Additionally we compared against the state-of-the-art deep learning model DeepCpf1 (Kim et al., 2018b) which has the same CNN architecture as our downstream head and also uses chromatin accessibility as additional modality. We use the same train and test splits as introduced in (Kim et al., 2018a). The numbers for baselines are directly taken from (Kim et al., 2018a).

Pretrained RNA-FM embeddings improve gRNA activity prediction. Table 1 summarizes the model performance. Cas-FM (RNA-FM) substantially outperforms all

Method	Correlation	CA	Features Used
CINDEL	0.61	No	Hand-crafted features
Lasso	0.64	No	One-hot encoding
L2	0.63	No	One-hot encoding
L1L2	0.64	No	One-hot encoding
Boosted RT	0.66	No	One-hot + positional features
DeepCpf1	0.71	Yes	One-hot + chromatin accessibility
Cas-FM (DNABERT-2)	0.49	No	Pretrained sequence embedding
Cas-FM (RNA-FM)	**0.76**	No	Pretrained RNA embedding
Cas-FM-CA (RNA-FM)	**0.78**	Yes	RNAFM embedding + CA

Table 1. Comparison of gRNA activity prediction methods. Cas-FM (RNA-FM) achieves the highest performance without using chromatic accessibility (CA) modality, while Cas-FM-CA (RNA-FM) further improves performance by incorporating chromatin accessibility as additional modality.

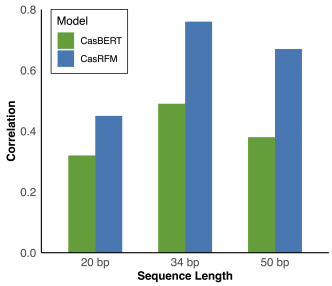

Figure 2. Performance of Cas-FM variants across varying gRNA sequence lengths, measured by Spearman rank correlation.

baselines while Cas-FM (DNABERT-2) significantly lags behind. Cas-FM (RNA-FM) works best likely due to the pre-training of RNA-FM backbone on transcriptomic sequences, which closely mirror the biological modality of guide RNAs. These results underscore the advantage of leveraging domain-aligned FMs: pretrained embeddings from RNA-FM capture complex, biologically meaningful dependencies beyond handcrafted features or 1-hot encodings.

Impact of gRNA sequence length/context on activity prediction. We evaluate Cas-FM (RNA-FM) and Cas-FM (DNABERT-2) with three context window sizes centered on the target site: 20 nucleotides (minimal core guide), 34 nucleotides (including flanking and PAM-adjacent bases), and 50 nucleotides (extended upstream and downstream context). Results in Fig 2 show that a 34-nt context consistently leads to the best performance. Cas-FM with either backbone underperforms with shorter 20-nt input, suggesting insufficient contextual information due to the lack of important flanking information that may influence folding, targeting, or interaction with Cas12, while the 50-nt context slightly degrades performance—likely due to the inclusion of less informative or noisy upstream/downstream sequence. These findings highlight the importance of context-aware input design and suggest a moderate-length context window—such as 34 nt—strike a balance between capturing local sequence features relevant to cleavage efficiency (e.g., nucleotide preferences near the PAM site) and avoiding overfitting or dilution from unrelated sequence regions.

Impact of chromatin accessibility on gRNA activity prediction Table 1 also assesses the contribution of epigenomic context for gRNA activity prediction. We evaluated whether incorporating chromatin accessibility (CA) improves the accuracy of gRNA activity prediction. Driven by the superior performance of Cas-FM with RNA-FM backbone, we limit our experiment with Cas-FM (RNA-FM) alone. Our results show that incorporating chromatin accessibility (denoted as Cas-FM-CA (RNA-FM)) improves predictive performance over using grRNA sequence embeddings alone (Cas-FM (RNA-FM)). This highlights the biological relevance of chromatin state in determining gRNA efficacy. Even when the sequence features are represented by powerful FM embeddings, the addition of CA features captures orthogonal information related to the physical availability of the genomic target site for Cas12-mediated cleavage. In practical terms, guides targeting open chromatin regions are more likely to be effective, and our model is better able to reflect that when explicitly given this contextual information.

These findings underscore the value of integrating multimodal biological data, combining large-scale pretrained sequence representations with relevant experimental measurements, to more accurately model the determinants of CRISPR guide activity.

5. Discussion

Our study demonstrates that the pre-trained transcriptomic foundational model RNA-FM significantly enhances the prediction of CRISPR-Cas12 gRNA activity compared to existing methods, even in out-of-distribution settings with short gRNA sequences that the foundation model was not pre-trained on. This is likely attributable to the fact that, from a biological perspective, gRNA sequences are functionally embedded within broader transcriptomic contexts. Their activity relies on factors such as sequence composition, secondary structure, and chromatin accessibility, which are partially captured by RNA-FM.

While prior work has noted the importance of incorporating extended sequence context around the gRNA (Kim et al., 2018b), our findings reinforce this insight in the context of foundation model-based learning. We confirm that a 34-nucleotide context strikes the best balance between informativeness and generalization, outperforming both shorter and longer input windows across different model variants.

Finally, we introduce a curated dataset of demonstration of chromatin accessibility features and show that integrating these in a multi-modal setting with gRNA sequences further enhances predictive accuracy. This suggests that epigenomic context introduces complementary information not captured by sequence alone, such as physical accessibility of the target site. Taken together, our work points toward a powerful direction for CRISPR modeling: leveraging domain-aware FMs while incorporating biologically grounded, orthogonal signals to better predict genome editing outcomes.

References

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Pre-trained language models for biological sequences at the scale of the human genome. *Nature Communications*, 13(1):1–13, 2022.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

Chen, J., Hou, T., Zhang, B., Zhou, X., Zhang, X., and Liu, X. Rna foundation model enables prediction of transcriptome and structural dynamics. *Nature Machine Intelligence*, 4(12):1046–1058, 2022.

Chen, K., Zhou, Y., Ding, M., Wang, Y., Ren, Z., and Yang, Y. Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *BioRxiv*, pp. 2023–01, 2023.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. pp. 785–794, 2016.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

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. *Nature Biotechnology*, 34(2):184–191, 2016.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J., Brock, K., Gal, Y., Birney, E., and Lunter, G. Deeper learning of human regulatory variants from large-scale genome sequencing data. *Nature Genetics*, 53(8):1141–1149, 2021.

Gao, R., Wang, Y., Shah, P., Zeng, Z., Liu, S.-W., Xie, J., Freemont, P. S., Wang, X., Li, H., and Zhou, J. Unfolded rna structures using deep learning reveal sequence-specific mechanisms of antisense oligonucleotide activity. *Bioinformatics*, 38(12):3231–3239, 2022.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. Z., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature Biotechnology*, 31(9):827–832, 2013.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Kent, W. J. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

Kim, H. K., Song, M., Lee, J., Menon, A. V., Jung, S., Kang, Y.-M., Choi, J. W., Woo, E., Koh, H. C., Nam, J.-W., et al. In vivo high-throughput profiling of crispr–cpf1 activity. *Nature methods*, 14(2):153–159, 2017.

Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S., and Kim, H. H. Deep learning improves prediction of crispr–cpf1 guide rna activity. *Nature biotechnology*, 36(3):239–241, 2018a.

Kim, H.-K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y.-H., Lee, S., Yoon, S., and Kim, H. Deep learning improves prediction of crispr–cpf1 guide rna activity. *Nature Biotechnology*, 36(3):239–241, 2018b.

Kleinstiver, B. P., Tsai, S. Q., Prew, M. S., Nguyen, N. T., Welch, M. M., Lopez, J. M., McCaw, Z. R., Aryee, M. J., and Joung, J. K. Genome-wide specificities of crispr-cpf1 nucleases in human cells. *Nature Biotechnology*, 34(8): 869–874, 2016.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K., and Giraldez, A. J. Crisprscan: designing highly efficient sgrnas for crispr-cas9 targeting in vivo. *Nature Methods*, 12(10): 982–988, 2015.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024.

Penić, R. J., Vlašić, T., Huber, R. G., Wan, Y., and Šikić, M. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*, 2024.

Prakash, M., Moskalev, A., DiMaggio, P. A., Combs, S., Mansi, T., Scheer, J., and Liao, R. Bridging biomolecular modalities for knowledge transfer in bio-language models. *bioRxiv*, pp. 2024–10, 2024.

Quinlan, A. R. and Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, X. S., et al. Sequence determinants of improved crispr sgrna design. *Genome Research*, 25(8):1147–1157, 2015.

Yang, Y., Li, G., Pang, K., Cao, W., Zhang, Z., and Li, X. Deciphering 3'utr mediated gene regulation using interpretable deep representation learning. *Advanced Science*, 11(39):2407013, 2024.

Yazdani-Jahromi, M., Prakash, M., Mansi, T., Moskalev, A., and Liao, R. Helm: Hierarchical encoding for mrna language modeling. *arXiv preprint arXiv:2410.12459*, 2024.

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., et al. Cpf1 is a single rna-guided endonuclease of a class 2 crispr-cas system. *Cell*, 163(3):759–771, 2015.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137, 2008.

Zhou, Z., Ji, Y., et al. grna-fm: Foundation models for crispr-cas9 grna representation. *bioRxiv*, 2023. Preprint: https://doi.org/10.1101/2023.10.17.562702.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oMLQB4EZE1.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

A. Appendix

A.1. gRNA Activity Dataset

The gRNA activity dataset used in this study originates from the HT1 high-throughput screening experiment described in the DeepCpf1 paper (Kim et al., 2018b). In this experiment, synthetic gRNA-target constructs were integrated into the genome of HEK293T cells using lentiviral delivery. Each construct encoded a 20-nt guide RNA sequence, a unique barcode, and a matching target site, allowing systematic evaluation of Cpf1 cleavage activity. The activity of each gRNA was quantified as the indel frequency measured via deep sequencing, corrected by background subtraction. For each guide, the dataset includes multiple representations of the target context: the core 20-nt gRNA sequence, an extended 34-bp target sequence (including PAM and flanking regions), and a 50-bp sequence window capturing broader genomic context around the cleavage site. This design supports modeling across multiple levels of sequence context.

Following the original data preparation strategy, we partitioned the dataset into two subsets: HT1-1 ($n = 15000$), which was used for model training, validation, and HT1-2 ($n = 1290$) for testing.

A.2. ATAC-seq data curation

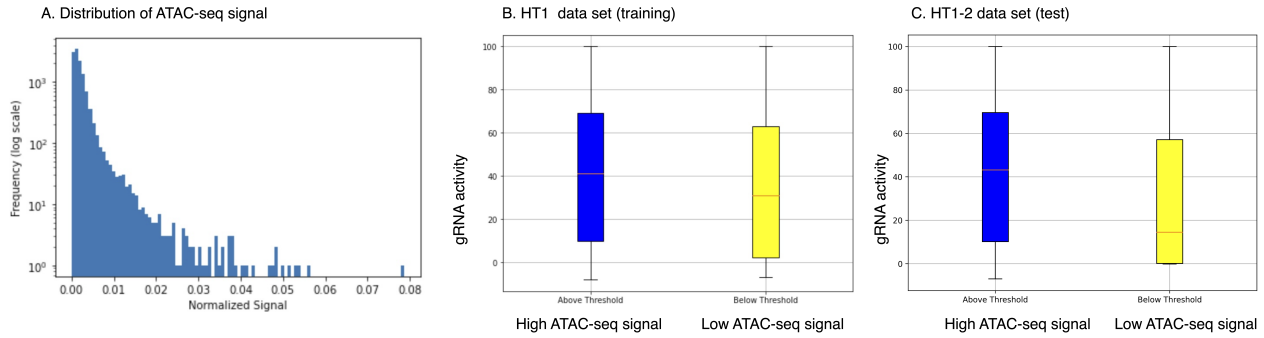

Figure 3. Association of continuous ATAC-seq signal with gRNA activity. A) Genome wide distribution of ATAC-seq signal in *HEK293T* cell line. B-C) ATAC-seq signal intensity is correlated with gRNA activity both in training and test data

To annotate gRNA target sites with chromatin accessibility (CA) status, we processed publicly available ATAC-seq data for the *HEK293T* cell line obtained from the GEO database (accession: GSM2902624). We specifically used the FASTQ files: `HEK293_ATAC_medium_depth_bio2_tech1_1.fastq.gz` and `HEK293_ATAC_medium_depth_bio2_tech1_2.fastq.gz`.

The raw data were preprocessed through a standard ATAC-seq pipeline. First, *alignment* of the reads to the human reference genome (*hg38*) was performed using `Bowtie2` (Langmead & Salzberg, 2012) with default parameters. Following alignment, *peak calling* was performed using `MACS2` (Zhang et al., 2008) to identify regions of open chromatin, reported as narrow peaks. Next, we performed normalization and thresholding of the `MACS2` signal intensities, applying a conservative cut-off of 0.001 . Genomic regions with a normalized signal above this threshold were considered accessible, while those below were treated as inaccessible.

To annotate each gRNA with a chromatin accessibility (CA) label, we first *aligned* the gRNA sequences to the human reference genome (*hg38*) using `BLAT` (Kent, 2002), a fast alignment tool suitable for short nucleotide queries. For each gRNA, we extracted the best-matching locus and recorded the corresponding genomic coordinates (chromosome, start, end, and strand). These coordinates were then converted into standard `BED` format to facilitate downstream processing. We *intersected* the gRNA BED file with ATAC-seq peak regions using `bedtools intersect` (Quinlan & Hall, 2010), and assigned an accessibility label of 1 (accessible) if a gRNA overlapped with a peak exceeding the signal threshold of 0.001, or 0 (inaccessible) otherwise. This resulted in a binary CA label for each gRNA, which

The raw data were preprocessed through a standard ATAC-seq pipeline. First, *alignment* of the reads to the human reference genome (*hg38*) was performed using `Bowtie2` with default parameters. Following alignment, *peak calling* was performed using `MACS2` to identify regions of open chromatin, reported as narrow peaks. Next, we performed normalization and

thresholding of the MACS2 signal intensities, applying a conservative cut-off of 0.001 . Genomic regions with a normalized signal above this threshold were considered accessible, while those below were treated as inaccessible. This resulted in a binary CA label for each gRNA, which was used as input feature in the downstream modeling and analysis.

We partitioned the genome into *highly accessible* and *low-accessibility* regions based on chromatin accessibility scores, using a threshold of 0.001 on the normalized ATAC-seq signal. This threshold was applied consistently across both training and test datasets to classify each gRNA target site as accessible or inaccessible.

To assess how accessibility influences gRNA efficacy, we compared the distribution of experimentally measured gRNA activity scores across the two accessibility categories. As shown in Figure 3, gRNAs targeting accessible regions exhibited significantly higher activity than those in inaccessible regions. This observation suggests a positive correlation between chromatin openness and Cpf1 cleavage efficiency, consistent with the biological intuition that accessible chromatin is more permissive to endonuclease activity.

A.3. Model architecture

Algorithm 1 Cas-FM (RNA-FM) architecture.

```
├─ Input modality
│  ├─ gRNA embedding ..................................................{RNA-FM T12 (640-dim)}
│  └─ Chromatin accessibility label ...............................{Binary label (1-dim)}
├─ CNN layer
│  ├─ Conv1D ......................................{kernel=5, filters=80, activation=ReLU}
│  ├─ AvgPool1D ...................................................................{pool size=2}
│  ├─ Flatten
│  ├─ Dropout ......................................................................{rate=0.3}
│  ├─ Dense layer ..............................................{units=80, activation=ReLU}
│  ├─ Dropout ......................................................................{rate=0.3}
│  ├─ Dense layer ..............................................{units=40, activation=ReLU}
│  ├─ Dropout ......................................................................{rate=0.3}
│  └─ Dense layer ..............................................{units=40, activation=ReLU}
├─ Chromatin encoder
│  └─ Dense layer ..............................................{units=40, activation=ReLU}
├─ Integration layer
│  └─ Element-wise multiplication ................{between gRNA sequence and chromatin
│     accessibility}
└─ Output head
   ├─ Dropout ......................................................................{rate=0.3}
   └─ Dense layer ............................................{units=1, activation=linear}
```

A.4. Model training hyperparameter

We trained the model for 100 epochs using the Adam optimizer with a learning rate of 5×10^{-5} and a batch size of 32. Model selection was based on validation loss, with early stopping applied after 10 epochs of no improvement. Mean Squared Error (MSE) was used as the training objective for the regression task, and Spearman rank correlation was computed on the test set for evaluation. Learning rate scheduling was performed using ReduceLROnPlateau with a patience of 5 epochs and a decay factor of 0.1.