Deep learning of Cas13 guide activity from high-throughput gene essentiality screening

Jingyi Wei[1,2], Peter Lotfy[3], Kian Faizi[3], Hugo Kitano[4], Patrick D. Hsu[5,6,*], Silvana Konermann[2,*]

[1]Department of Bioengineering, Stanford University, Stanford, CA

[2]Department of Biochemistry, Stanford University, Stanford, CA

[3]Laboratory of Molecular and Cell Biology, Salk Institute for Biological Studies, La Jolla, CA

[4]Department of Computer Science, Stanford University, Stanford, CA

[5]Department of Bioengineering, University of California, Berkeley, Berkeley, CA

[6]Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA

[*]Corresponding author

Correspondence: silvanak@stanford.edu (S.K.), pdhsu@berkeley.edu (P.D.H.)

Abstract

Transcriptome engineering requires flexible RNA-targeting technologies that can perturb mammalian transcripts in a robust and scalable manner. CRISPR systems that natively target RNA molecules, such as Cas13 enzymes, are enabling rapid progress in the investigation of RNA biology and advancement of RNA therapeutics. Here, we sought to develop a Cas13 platform for high-throughput phenotypic screening and elucidate the design principles underpinning its RNA targeting efficiency. We employed the RfxCas13d (CasRx) system in a positive selection screen by tiling 55 known essential genes with single nucleotide resolution. Leveraging this dataset of over 127,000 guide RNAs, we systematically compared a series of linear regression and machine learning algorithms to train a convolutional neural network (CNN) model that is able to robustly predict guide RNA performance based on guide sequence alone. We further incorporated secondary features including secondary structure, free energy, target site position, and target isoform percent. To evaluate model performance, we conducted orthogonal screens via cell surface protein knockdown. The final CNN model is able to predict highly effective guide RNAs (gRNAs) within each transcript with >90% accuracy in this independent test set. To provide user interpretability, we evaluate feature contributions using both integrated gradients and SHapley Additive exPlanations (SHAP). We identify a specific sequence motif at guide position 15-24 along with selected secondary features to be predictive of highly efficient guides. Taken together, we derive Cas13d guide design rules from large-scale screen data, release a guide design tool (http://RNAtargeting.org) to advance the RNA targeting toolbox, and describe a path for systematic development of deep learning models to predict CRISPR activity.

* * *

Introduction

The ability to robustly perturb specific RNA molecules in the cell is required for functional elucidation of the transcriptome and its diverse phenotypes. Despite rapid progress in effective technologies for genome engineering, analogous systems for transcriptome engineering lag behind their DNA counterparts. Recent advances in the discovery and development of RNA-targeting CRISPR systems, such as Cas13 enzymes that can be programmed with a guide RNA to directly target RNA substrates, are beginning to address this gap (Abudayyeh et al., 2016; East-Seletsky et al., 2016). Because CRISPR proteins are exogenous to eukaryotes, they can be flexibly engineered to mediate RNA cleavage or binding activities in desired subcellular compartments. Further, their modular nature enables the facile fusion of effector domains to expand the RNA targeting toolbox. As a result, a broadening suite of Cas13-based tools is now able to perturb RNA expression (Abudayyeh et al., 2017; Konermann et al., 2018), splicing (Konermann et al., 2018), nucleotide sequence (Abudayyeh et al., 2019; Cox et al., 2017; Xu et al., 2021), and methylation (Wilson et al., 2020), as well as profile RNA-protein interactions (Han et al., 2020). These capabilities are now accelerating applications across the study of fundamental RNA biology, RNA-based therapeutics, and molecular diagnostics.

The Cas13 family is unified by the presence of two conserved HEPN ribonuclease motifs and a common RNA cleavage mechanism, yet divided into several subtypes on the basis of sequence diversity and coding sequence length. Cas13d enzymes, in particular the engineered Cas13d-NLS from *R. flavefaciens* strain *XPD3002* (CasRx) (Konermann et al., 2018), are highly compact RNA targeting effectors with robust activity in mammalian and plant cells relative to other subtypes (Wessels et al., 2020; Li et al., 2021; Mahas et al., 2019), motivating their further development as RNA targeting tools.

The most well-established approaches for RNA targeting, including RNA interference and antisense oligonucleotides, have been collectively challenged by poor specificity, low throughput, or variable efficiency. Although CasRx has been shown to be highly specific in mammalian cells, it remains to be adapted into a platform for high-throughput, pooled genetic screening with phenotypic readouts. Furthermore, the ability to select highly effective guide RNAs requires a better understanding of the rules underpinning Cas13d guide efficiency. Recent approaches to understand and predict Cas13d activity have been limited by relatively small experimental datasets and manual selection of primary sequence features (Wessels et al., 2020).

We therefore sought to adapt Cas13d into a scalable platform for high-throughput phenotypic screening. First, we designed a library of >127,000 guide RNAs tiling 55 essential human transcripts with single-nucleotide resolution and assayed the effect of each guide on cell proliferation. After quantifying guide RNA abundance over a 14-day time period in K562 cells, we evaluated guide efficiency based on spacer depletion ratios and systematically compared linear, ensemble, and deep learning models to predict guide activity. A deep learning convolutional neural network (CNN) model based on guide sequence alone was able to achieve

surprisingly high accuracy at classifying effective guides (80% true positive rate at a 0.9 model threshold) for held-out transcripts. We refined this initial model by adding secondary features relating to RNA structure and mRNA target characteristics. Validation against an orthogonal dataset based on cell surface protein knockdown successfully predicted effective guides with 95% accuracy across the top 10 selected guides for each gene.

Model feature interpretation using integrated gradients (IG) (Sundararajan et al., 2017), Shapley additive explanations (SHAP) (Lundberg et al., 2020), and transcription factor motif discovery (TF-MoDISco) (Shrikumar et al., 2018) revealed strong preferences for unstructured guides and target regions. Optimal guide activity was achieved by targeting the beginning of the coding portion of a transcript as well as a newly discovered core sequence motif from position 15-24 of the spacer sequence.

Overall, we demonstrate the utility of CasRx for performing large-scale screens with phenotypic readouts such as cell proliferation and survival, and outline a strategy to systematically develop robust deep learning models for predicting guide RNA activity. We make our best-performing model available to the research community at http://RNAtargeting.org for all known coding and non-coding transcripts in the human and mouse transcriptome, as well as any custom target RNA sequence.

Results

Development of CasRx as a platform for high-throughput phenotypic screening

In order to systematically understand CasRx efficiency, we sought to generate a large library of diverse guide RNAs and screen them based on cell proliferation. Reasoning that CasRx knockdown of essential transcripts would lead to the depletion of highly effective guides, we selected a set of 55 essential genes with high confidence in their essentiality based on the overlap between three previously reported survival screens performed with RNAi and CRISPR interference (CRISPRi) in K562 cells (Hart et al., 2015; Horlbeck et al., 2016; Luo et al., 2008). We selected K562 cells as a model system due to their ease of use in pooled screens and variable CasRx mediated protein knock-down activity in this cell type (**Figure S1A, B**).

First, we produced stable cell lines via transfection of an all-in-one plasmid containing the CasRx effector, PiggyBac transposase, and an antibiotic selection cassette. Because we previously discovered that Cas13d does not have any flanking sequence requirements, unlike other Cas13 subtypes, we elected to tile 55 essential transcripts with single nucleotide resolution. Guide RNA (gRNA) spacers were designed to target the 5′ UTR, gene body, and 3′ UTR for each target transcript. As controls, we additionally targeted 5 non-essential genes that are not expected to have any effect on cell viability. After selection of stable cell lines, we transduced the effector cell line with a pooled lentiviral library containing these 144,745 guide RNAs at low MOI to enable independent infection events. Cells were cultured for 14 days to

allow for depletion of cells containing effective guides targeting the essential transcripts followed by gDNA extraction and NGS analysis of the depleted gRNA cassettes (**Figure 1A)**.

We analyzed the depletion of gRNAs 14 days post guide transduction compared to the original distribution of gRNAs across essential, non-essential, and non-targeting spacer categories. Analysis of the cumulative distribution function of guide RNAs post-selection demonstrated that the top 20th percentile of essential gene guides are clearly separated from guides targeting non-essential genes or non-targeting guides **(Figure 1B)**. We noticed that this corresponded to a 0.45 depletion ratio, which we set as an efficiency cut-off for further analysis. This initial analysis confirmed that targeting of essential transcripts by Cas13d-based survival screening is able to capture a robust signal for effective guides. In addition, we observed a wide spread of guide depletion ratios beyond the most clearly separated top 20% of guides. This highlights the need for a highly reliable predictive model of Cas13d guide effectiveness in order to both simplify smaller scale RNA-targeting experiments as well as enable direct RNA-targeted transcriptome-scale screening.

Different essential transcripts can have highly variable effects on cell survival and proliferation. To address this, we performed a transcript-level analysis of guide depletion and ranked both essential and non-essential transcripts based on the percentage of guides below our efficiency cut-off (0.45 ratio) (**Figure 1C**). As expected, we were able to effectively separate essential from non-essential transcripts. Within the set of essential transcripts, we further observed that ribosomal genes are more robustly depleted relative to other essential genes, suggesting that they may have a stronger impact on cell proliferation or survival.

Next, we analyzed the distribution of effective spacers by their location on the target transcript. Heat map representation of the top 20th percentile guides that pass the efficiency cut-off revealed a striking degree of clustering among the top 20th percentile guides, leading to spacer hot spots and deserts along the transcript length **(Figure 1D**). Across all transcripts, the observed distribution of effective guides along transcripts was clearly distinct from a non-clustered random distribution (**Figure 1E**). Multiple factors could be responsible for the observed clustering of highly effective guides, including enrichment for specific target positions, sequences, or sequence contexts.

Prediction of CasRx activity based on guide RNA sequence alone

We therefore sought to systematically analyze the elements that distinguish effective Cas13d gRNAs, starting with the contribution of spacer sequence alone. To investigate the presence of potential sequence preferences at the nucleotide level, we first examined the correlation of nucleotide identity with guide efficiency at each position along the 30 nt spacer (**Figure 2A**). We observed increased base preference for G and C at the DR-proximal spacer positions 15-24. To integrate sequence information across the entire spacer, we developed computational algorithms that predict guide efficiency (**Figure 2B**). Following one-hot encoding of each 30 nt spacer into four binary vectors representing each base, we systematically evaluated linear, ensemble, and deep learning models to predict effective guides based on sequence alone. We

compared 3 methods employing logistic regression, 2 ensemble methods (Random forest (RF) and Gradient-boosted tree (GBT)) and 2 deep learning models (convolutional neural network (CNN) and bidirectional long short-term memory neural network (LSTM)). After filtering to remove guides with off-target matches in essential transcripts, these models were trained to predict and select high efficiency guides within each target transcript (**Figure S1C-H**).

Due to the high degree of spacer clustering that we observed, models that are trained and tested on guide RNAs from the same sets of transcripts would be expected to potentially be subject to overfitting by learning the targeting hotspots specific to those transcripts and therefore result in overestimated model performance. To ensure model generalizability to other genes, we therefore partitioned 54 target transcripts into 9 folds for cross-validation. One target transcript, RPS19BP1, was excluded from this analysis because its guides were not effectively depleted in our screen and clustered with non-essential transcripts (**Figure 1D**).

We observed high model performance for the gradient-boosting tree (GBT) and the two deep learning models, based on Area under the receiver operating characteristic curve (AUROC) and Area under precision-recall curve (AUPRC) metrics across all 9 gene splits (**Figure 2C**). Overall, the CNN model performed best based on AUROC (0.845 relative to a baseline of 0.5), AUPRC (0.541 relative to a baseline of 0.18) and a true positive rate of 80% effective guides at a 0.9 model threshold. The relatively high prediction accuracy indicates that spacer sequence is a primary factor determining guide efficiency.

Between the two deep learning models tested, we picked the CNN model for further refinement and evaluation as it specifically performed best on held-out transcripts, indicating an advantage if applied to entirely new targets compared to the LSTM model (**Figure S2A**). First, we sought to understand if Cas13d had any flanking sequence preferences across this large dataset. Adding flanking sequences of varying length from 1-7 nt to the model input did not meaningfully improve model performance (**Figure 2D**), consistent with our previous biochemical studies suggesting a lack of strong flanking sequence requirements (Konermann et al., 2018).

We next sought to understand the relationship of spacer length and Cas13d activity. The full-length spacer in the native RfxCas13d array is 30 nt, a length that is generally consistent across the Cas13d family (Konermann et al., 2018). We previously showed that truncating the spacer below 22 nt led to loss of knockdown activity (Zhang et al. 2018), which we further investigated by designing a panel of truncations ranging from the 30 nt full-length spacer down to 12 nt (**Figure S2B**). Across two different direct repeat lengths and two spacers targeting the cell surface marker CD81, we observed a gradual decrease in knockdown efficiency with spacers shorter than 24 nt. This pattern replicated across our guide library, characterized by a steep drop in AUPRC below a spacer length of 24 nt (**Figure 2E**) and only a very minor decrease from 30 nt to 24 nt. This is further consistent with the per-base correlation data from **Figure 2A** that suggests a window of increased sequence sensitivity from 15-24 nt.

Taken together, we were able to achieve robust predictive performance of guide efficiency by training a CNN model on guide sequence alone. This approach revealed a marked sequence preference for nucleotide positions 15 - 24 in the spacer region.

Addition of secondary features improves guide prediction accuracy

Next, we reasoned that RNA structure, target site expression levels, and coding property may be correlated with guide efficiency. We therefore developed a rational list of secondary RNA features that largely cannot be easily obtained from spacer sequence alone.

We found that guides with predicted higher guide unfolding energy, implying a more highly structured RNA, were less likely to be effective (**Figure 3A and S3A**). Consistent with the preference for a less structured spacer region in the guide RNA, we found that highly structured target RNA regions were disfavored (**Figure 3B and S3B**). In contrast, guide RNAs where the conserved structure of the direct repeat was predicted to be disrupted had only a marginal decrease in guide efficacy (**Figure 3C and S3C**). Finally, analysis of the spacer GC content revealed a preference for a balanced composition of pyrimidines and purines in the spacer sequence (45- 55% GC content (**Figure 3D and S3D**). Moving onto analysis of larger scale regions of the target RNA, we determined the impact of guide location in the coding region (CDS) of the target compared to the 5′ and 3′ untranslated regions (UTRs). We found that guides targeting the coding portion had a higher likelihood of being effective **(Figure 3E and S3E**), with the lowest fractions of effectives guides at either 3′ or 5′ ends of the transcript within the UTRs (**Figure 3F and S3F**).

Finally, we expected that guides targeting exons conserved across transcript isoforms would have a higher chance of showing a phenotype, which we confirmed by analyzing the percent of target mRNA isoforms targeted by each guide RNA (**Figure 3G and S3G**). A similar pattern was observed when directly analyzing the relative abundance of the guide target region across all isoforms by RNA-seq reads (**Figure 3H and S3H**).

Because most of these secondary features showed a relatively modest correlation with guide efficiency, we tested if they would be able to improve model performance when added individually to the sequence-based CNN model (**Figure 3I, S4A,B**). We found that the position of the guide on the target transcript had the most prominent effect, followed by a more modest increase when adding target and guide RNA folding energy predictions respectively. In contrast, addition of spacer GC content and predicted DR folding disruption features did not significantly change model performance, consistent with our expectation that spacer GC content would have been successfully captured by the spacer sequence-only CNN model.

To build our final CNN model, we sequentially included each of the five features that had successfully improved model performance when added individually. AUROC and AUPRC were evaluated at each cumulative secondary feature addition (**Figure 3J**). Addition of each of the 5 selected features improved model performance at least to a modest degree, and our final model achieved a very high average AUROC of 0.875 and a high average AUPRC of 0.638 **(Figure**

3K and **S4C,D,E** for feature variations**)**. As a comparison, we tested the addition of the same set of secondary features to the GBT model, which was the best performing model not based on deep learning (**Figure S5A**). All features included in the final CNN model, along with spacer GC content, were found to be important for the GBT model as well (**Figure S5B, C**).

One of the key applications of a predictive model like this one would be to accurately predict the most effective guides in order to aid in guide and library design. To evaluate the model in this context, we set a target score threshold of 0.8 and plotted the guide percentile rank distribution of the guides predicted to have high efficacy by the model. As hoped for, guides were heavily skewed towards the highest efficiency ranks, with a true positive rate of 0.83. Setting a higher target score threshold to 0.9 further increased the true positive ratio to 0.92 (**Figure 3L**).

Model validation on an orthogonal dataset based on cell surface protein knockdown

Next, we sought to validate our model on CasRx knockdown of cell surface markers, reasoning that an orthogonal readout to gene essentiality and cell survival would ensure generalizability of our model predictions to multiple screen modalities. Analogous to the survival screen, we designed 3,218 guides tiling two transcripts, CD58 and CD81, with single-nucleotide resolution. 10 days after lentiviral transduction of the guide library, target knockdown at the protein level was evaluated via FACS sorting into 4 bins on the basis of residual target expression level (**Figure 4A**). Following NGS quantification of guide representation, guide efficiency was calculated as a ratio of guide percentage in the bin exhibiting greatest knockdown (bin 1) to the sum of its percentage in bin 1 and the bin exhibiting the highest level of target expression (bin 4).

Analysis of the cumulative distribution of guide ratios demonstrated that the majority of targeting guides (>=60%) were clearly separated from non-targeting guides (**Figure 4B**). As expected, we did not observe any significant non-targeting guide enrichment in the top 20th percentile of targeting guides, the cutoff we had previously established for effective guides from the survival screen.

To evaluate our model's performance on this new dataset, we tested an ensemble CNN model based on our survival screen data on each CD transcript. We found that the ensemble model outperformed all individual models (**Figure S6C**) and achieved highly robust prediction accuracy for both CD58 (AUROC of 0.88 and AUPRC of 0.66) and CD81 (AUROC of 0.86 and AUPRC of 0.62) (**Figure 4C**). This performance is comparable to the model accuracy on held-out essential genes (**Figure 3K**), highlighting its generalizability. To further confirm the robustness of our model to experimental variables such as cell line, delivery method, guide length, and experimental lab, we evaluated our CNN model on a published CasRx guide tiling screen dataset on CD46, CD55, and CD71 (Wessels et al., 2020) where all these variables were distinct from our dataset. The prediction of all three CD genes proved to be very accurate, with high AUROCs ranging from 0.85 to 0.89 (**Figure S6D**), further supporting the utility of our model.

In practice, our model is likely to be used to predict the top 3-10 guides for each target transcript, both for applications involving targeting of individual selected transcripts as well as for library design for larger-scale screens across thousands of target RNAs. To simulate this test case, we examined the true percentile rank of the top 10 predicted high efficiency guides from our model. We found that all top 10 predicted high efficiency guides for CD58 fell in the top 20th percentile, and 9 out of 10 predicted high efficiency guides for CD81 fell in the top 20th percentile (**Figure 4D**). Taken together, 95% of guides selected by the model were highly efficient, indicating its high precision and utility.

Finally, we evaluated our best non-deep learning model, the GBT model, as a comparison. Consistent with our observation on the held-out transcripts of our survival screen, we observed a slightly worse performance compared to the CNN approach (AUROC of ~0.84 and AUPRC ~0.59 for both genes) (**Figure 4E).**

Feature interpretation and discovery of a core sequence motif

Having confirmed the robustness of our model across two distinct datasets, we analyzed the contribution of individual model features to CasRx guide activity using three distinct methods for model interpretation. To achieve a better understanding of CasRx targeting requirements and preferences, we used an integrated gradients approach (IG) (Sundararajan et al., 2017) to provide observability for our CNN model. IGs revealed that targeting the beginning of the 5′ UTR and end of the 3′ UTR was the most disfavored, with an overall preference for targeting the beginning of the coding region (CDS) (**Figure S7A**). Targeting regions of a transcript that are conserved across isoforms was also preferred. Finally, guide and target unfolding energy values had a relatively high impact on predicted guide efficacy, with stronger predicted RNA folding generally contributing to a classification as a less effective guide.

A downside of the integrated gradients approach to ML model interpretability is that it does not provide a straightforward way to rank the importance of features relative to each other. SHapley Additive exPlanations (SHAP), a game theoretic approach, is designed to enable feature ranking (Lundberg et al., 2020). As a comparison to the IG approach, we employed SHAP analysis on our GBT model (**Figure S7B**). We found that the direction of feature contribution to guide classification was generally consistent between both models (CNN and GBT) and both methods of feature evaluation (IG and SHAP). SHAP ranking was consistent with spacer sequence composition as the most important secondary feature, with a clear preference for intermediate GC content (40-60% GC). Taken together with our initial observation that the sequence-only GBT and CNN models performed surprisingly well, we decided to further investigate the target sequence preferences of Cas13d as learned by our models.

IG (**Figure 5A, B**) and SHAP (**Figure 5C, D**) analysis on each nucleotide in the guide sequence nominated a core region of position 15-24 in the spacer sequence as a major contributor to guide performance. Consistent with our original correlation analysis (**Figure 2A**), the CNN and GBT models had a clear preference for an alternating stretch of guanines, cytosines and guanines ($G_{15\text{-}18}C_{19\text{-}22}G_{23\text{-}24}$) in this core region (**Figure 5B and 5D**). This unique core motif was

not found for Cas13a when we performed a correlational analysis of available datasets (Abudayyeh et al., 2017; Metsky et al., 2021) (**Figure S8**). Indeed, no consistent sequence motif or core region emerged across the Cas13a datasets analyzed, which could be due to intrinsic enzymatic properties of Cas13a or limitations in the size of available datasets.

As our IG and SHAP analyses investigated each sequence nucleotide position independently, we further sought to determine the role of specific motifs (nucleotide combinations) in guide efficacy. We employed Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco), an algorithm that identifies sequence motifs incorporated in deep learning models by clustering sequence segments and relating them to their predicted performance (Shrikumar et al., 2018). We discovered a total of 14 distinct patterns associated with high efficiency guides, with the top 5 patterns shown in **Figure 5E**. As TF-MoDISco is normally applied to genomic regions for the identification of transcription factor binding sites, it is designed to identify motifs in a position-independent manner across distinct genomic sites. In our analysis, we noticed that all identified high efficiency motifs were anchored to a specific guide position centered around spacer nucleotides 18-20 (**Figure S9A**), consistent with the prior observation of a core region using individual base scores alone**.**

Strikingly, all top 5 patterns contained a cytosine at position 21, with a single guanine at varying positions in the core region across the different patterns. Taken together, the identified motifs can be summarized as $\mathbf{GN_xC_{21}}$ or $\mathbf{N_xC_{21}G}$ within the core region. Generally, the patterns were sparse and characterized by just two dominant bases (one G and one C), in contrast to the longer 9-base motif comprised of $G_{15\text{-}18}C_{19\text{-}22}G_{23\text{-}24}$ that the individual base-level analysis would have suggested **(Figures 5A and 5C)**.

Given these sparser core region motifs revealed by the TF-MoDISco analysis, we wanted to gain a better understanding of the identity of the variable intervening (N) bases in the combinatorial motifs. To this end, we performed a straightforward analysis of enriched and depleted 3-mers across the whole guide position in high efficiency guides. Consistent with prior analyses, enriched 3-mers were again clustered in the core region (position 15-24) of the guide **(Figure 5F**). In addition to the consistent finding of a prominent enrichment of C at position 21, they revealed a strong preference for A or T intercalated with G and C **(Figures 5F-G**), a finding that was obscured in the base-by-base analysis shown in Figure 5B and D. Analysis of enriched and depleted 4-mers had a similar finding (**Figure S9B)**. Further analysis of the GC content of the core region confirmed a preference for the presence of key $\mathbf{GN_xC_{21}}$ or $\mathbf{N_xC_{21}G}$ motif bases, along with a narrow preference for a medium GC content via intercalating A/T nucleotides at the N positions of the motif (**Fig. 5H and I**).

Discussion

In this study, we successfully applied CasRx for large scale screening with a survival readout, demonstrating the feasibility of large-scale phenotypic direct RNA-targeting screening with Cas13d. While our data was able to clearly separate guides targeting known essential genes

from control guides, we also observed wide variability in the efficacy of spacers within a given transcript. The ability to effectively identify the most effective guides *a priori* is a key step enabling further RNA targeting applications with more compact libraries across larger gene sets. In order to understand key features that determine guide activity, we took advantage of our large-scale dataset employing over 127,000 distinct spacer sequences against 55 target transcripts.

Using ensemble and deep learning methods, we discovered that guide sequence alone was sufficient to build a surprisingly accurate model for Cas13d guide classification. This demonstrates the utility of our approach to use deep learning based on spacer sequence, without manual sequence feature selection or feature engineering. In contrast, most prior models of CRISPR guide efficacy have relied on manual selection of a limited set of guide sequence features combined with simpler ML models, such as elastic nets (Horlbeck et al., 2016), SVM plus logistic regression (Doench et al., 2016), or random forest approaches (Wessels et al., 2020). This report therefore provides a blueprint for a streamlined deep-learning workflow for developing accurate sequence-based models given a sufficiently large dataset (>100,000 guides).

Systematic addition of secondary features including structural prediction of both guide and target RNA, RNA target position, and conservation across isoforms further improved our model, yielding a highly robust AUROC of 0.875 and AUPRC of 0.638 on held-out genes across 9-fold cross-validation. To ensure that our model would perform well across screen modalities, we performed a validation screen using FACS readout of cell surface protein knockdown. Across this orthogonal dataset as well as a previously published dataset (Wessels et al., 2020) targeting three transcripts using a different cell line, delivery method, and guide length, our model reliably predicted highly effective guides for each gene with up to 95% accuracy. This indicates that model performance is preserved across cell types, Cas13d dosage, and screen readout modalities.

One common downside of deep learning models for biological applications is the lack of observability of feature contributions to the model output. Prior deep learning models for both Cas9 (Chuai et al., 2018; Kim et al., 2019); (Xue et al., 2019) and Cpf1 (Kim et al., 2018) have partially begun to address this limitation through neuron visualizations or feature saliency maps in some cases. In this report, we successfully applied three different model interpretation approaches, including integrated gradients, SHAP and TF-MoDISco for comprehensive model feature interpretation across both CNN and GBT models. Importantly, through initial correlation analysis as well as model feature interpretation, we discovered a core region at guide position 15-24 with a specific sequence composition predictive of high efficiency guides. Further analysis of motifs in this region revealed a distinct preference for cytosine at position 21 as part of a $GW_{1\text{-}4}C_{21}$ or $C_{21}W_{0\text{-}2}G$ motif with a narrow preference for 50-60% GC content in this window. Importantly, analysis of base preference at the individual nucleotide level only obscured this motif, underscoring the importance of motif-level approaches to model interpretation such as TF-MoDISco used here - the first time to our knowledge such a motif-level approach has been

applied to CRISPR guide activity prediction. Analysis of available Cas13a datasets (Abudayyeh et al., 2017; Metsky et al., 2021) did not indicate a similar motif, suggesting that it may be unique to Cas13d.

Future work should incorporate more accurate structural prediction for long RNAs, including mRNAs and lincRNAs as those are developed further. We evaluated RNA secondary structure prediction algorithms and found that the LinearFold implementation of the contrafold model (Huang et al., 2019) performed best when compared to Vienna (Lorenz et al., 2011) and Eternafold (Wayment-Steele et al., 2020) in the context of our model (**Figure S4C**). When applied to the prediction of local target unfolding energy, however, we found that performance declined when adding longer flanking sequences to the target RNA (>60 nt total). This is likely due to limited prediction accuracy for longer RNA sequences. Other factors, including RNA-protein interactions, also impact target site accessibility. Approaches such as *in vivo* SHAPE analysis (Spitale et al., 2013) for target transcripts could provide additional experimental data on protein occupancy or local accessibility, once higher coverage datasets are developed. We anticipate incorporation of these features will enable further improvements to highly accurate prediction of guide efficiency for direct RNA targeting.

Beyond its specific application to CasRx activity prediction, we envision that the deep learning model architecture, systematic feature addition and model training workflow as well as the model interpretation approach outlined in this paper will be broadly applicable to other sequence-based models, such as the prediction of gRNA activities for newly discovered CRISPR enzymes, DNA/RNA modifications and DNA/RNA-protein interactions.

Finally, we make our model for CasRx guide prediction available at http://RNAtargeting.org. Based on the most comprehensive Cas13 screening dataset to date, we created this webtool for guide selection across model organism transcriptomes as well as prediction of guide efficacy on custom RNA sequences.

Data and Code Availability

The model is freely accessible at http://RNAtargeting.org. The underlying data and code for this manuscript will be available on Github.

Acknowledgments

We thank the Konermann laboratory and Hsu laboratory for support and advice; A. Kundaje and J. Zou for advice on deep learning models; C. Duffy for advice on linear and ensemble models; J. Zou for the recommendation of the LinearFold package; W. Zhuk for building the deep learning model architecture; HK. Wayment-Steele for advice on guide free energy calculation and the recommendation of Arnie; A. Shrikumar for instructions on TF-MoDISco; and B. Hsu for helping build the CasRx guide design website. S.K. is a Hanna Gray Fellow of the Howard Hughes Medical Institute and Chan Zuckerberg Biohub Investigator. P.D.H. is supported by the NIH (DP5 OD021369, R01 GM131073, R01 GM132465), DARPA, Emergent Ventures, the Shurl and Kay Curci Foundation, and the Rainwater Charitable Foundation.

Author Contributions

S.K. and P.D.H. conceived this study and supervised the design and analysis of all experiments. J.W. and H.K. built the computational models, performed feature engineering, and implemented model interpretation. S.K. and J.W. analyzed the NGS data from the screens and calculated secondary features. J.W. performed the validation screening with supervision from S.K. J.W. created the guide efficiency prediction tool. P.L. S.K., and P.D.H. adapted CasRx for high-throughput screening. S.K., K.F., P.L., and P.D.H. performed the cell proliferation screen. S.K., P.D.H., and J.W. wrote the manuscript with input from all authors.

Competing Interest Statement

P.D.H. is a cofounder of Spotlight Therapeutics and Moment Biosciences and serves on the board of directors and scientific advisory boards, and is a scientific advisory board member to Vial Health and Serotiny. P.D.H. and S.K. are inventors on patents relating to CRISPR technologies.

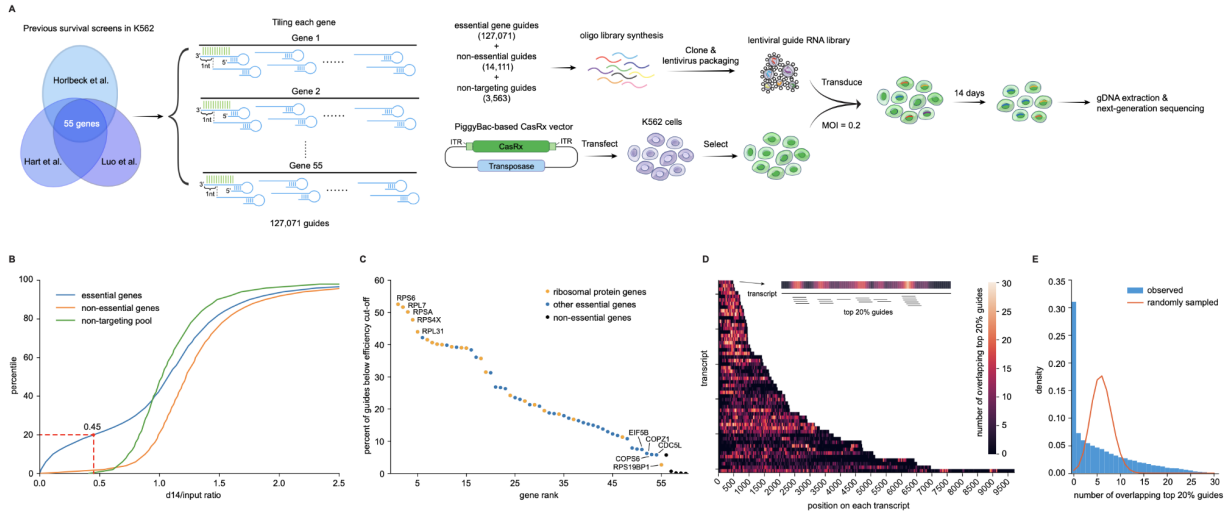


Figure 1: Development of *Rfx*Cas13d (CasRx) for high-throughput phenotypic screening. **A.** Schematic of a CasRx pooled library screen for essential gene knockdown. **B.** Cumulative distribution of the guide RNA depletion ratio at day 14 compared to the input library across essential, non-essential, and non-targeting spacer categories. Red dashed lines indicate the ratio at the top 20th percentile of essential transcript targeting guides. The corresponding 0.45 ratio was set as an efficiency cut-off for further analysis. **C**. Gene ranking based on the fraction of highly depleted guides per transcript. Individual transcripts were ranked based on the percentage of its guides below the efficiency cut-off (day14/input ratio <0.45). Orange dots denote ribosomal protein genes; blue dots denote other essential genes; black dots denote non-essential genes. The top 5 and bottom 5 essential genes are annotated. **D.** Heat map of the positional distribution of top 20th percentile guides on each transcript. Heat map color indicates the number of overlapping top 20% guides for each position on the transcript. **E.** Frequency distribution of the number of overlapping top 20% guides across all transcripts. The blue histogram shows the observed distribution of the number of overlapping top 20% guides. The orange curve represents the randomly sampled distribution of nucleotide-level overlap of 20% guides in the library. The observed distribution indicates a higher level of clustering than expected if the top 20% guides were randomly distributed along the transcripts.

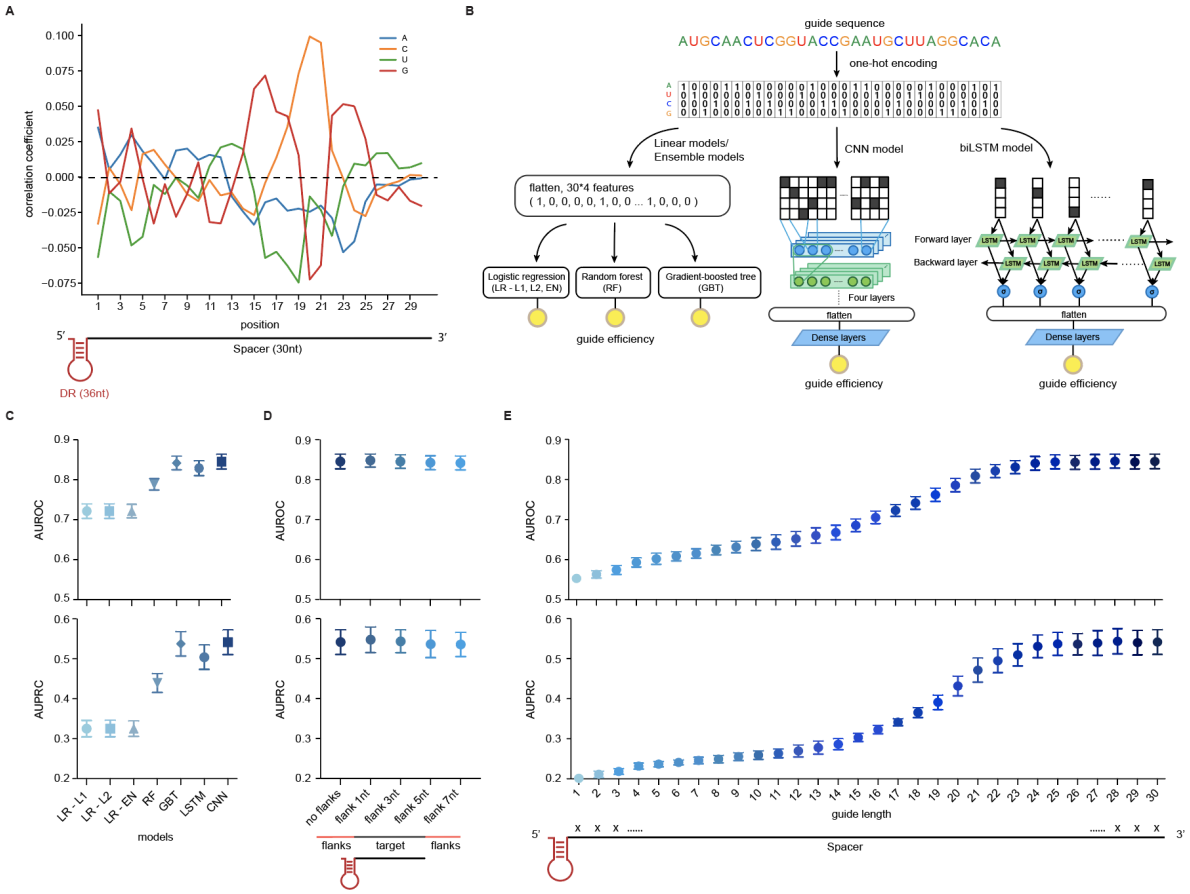


Figure 2: Prediction of CasRx activity based on guide RNA sequence alone. A. Correlation of each nucleotide with guide efficiency at each position along the 30 nt spacer. Pearson correlation coefficient for each nucleotide identity with guide efficiency at each position is shown. **B.** Schematic of computational algorithms to predict guide efficiency based on guide sequence only. **C.** Comparison of prediction accuracy between linear, ensemble and deep learning models based on 9-fold cross-validation split by transcript. Averages of Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under Precision-Recall Curve (AUPRC) across test sets from all 9 splits are shown ± SD. LR - L1, logistic regression with L1 regularization; LR - L2, logistic regression with L2 regularization; LR - EN, logistic regression with elastic net regularization; GBT, gradient-boosted tree; RF, random forest classifier; CNN, convolutional neural network; biLSTM, Bidirectional long short-term memory neural network. The CNN model achieved the highest average AUROC and AUPRC. Note that the baseline for AUPRC is equal to the fraction of positive class (high efficiency guide percent), in this case 0.18. **D.** The effect of target RNA flanking sequence on model prediction accuracy. RNA target flanking sequences of various lengths (1-7 nt) were added to the 30 nt guide target sequence in the CNN model to evaluate impact on guide efficiency prediction. Model AUROC and AUPRC (mean ± SD) are shown. **E.** The effect of guide length on model prediction accuracy. The guide spacer sequence was truncated from the 3′ end from the full-length 30 nt sequence down to 1 nt and a CNN model was trained for guide efficiency prediction. Resulting model AUROCs and AUPRCs (mean ± SD) are shown.

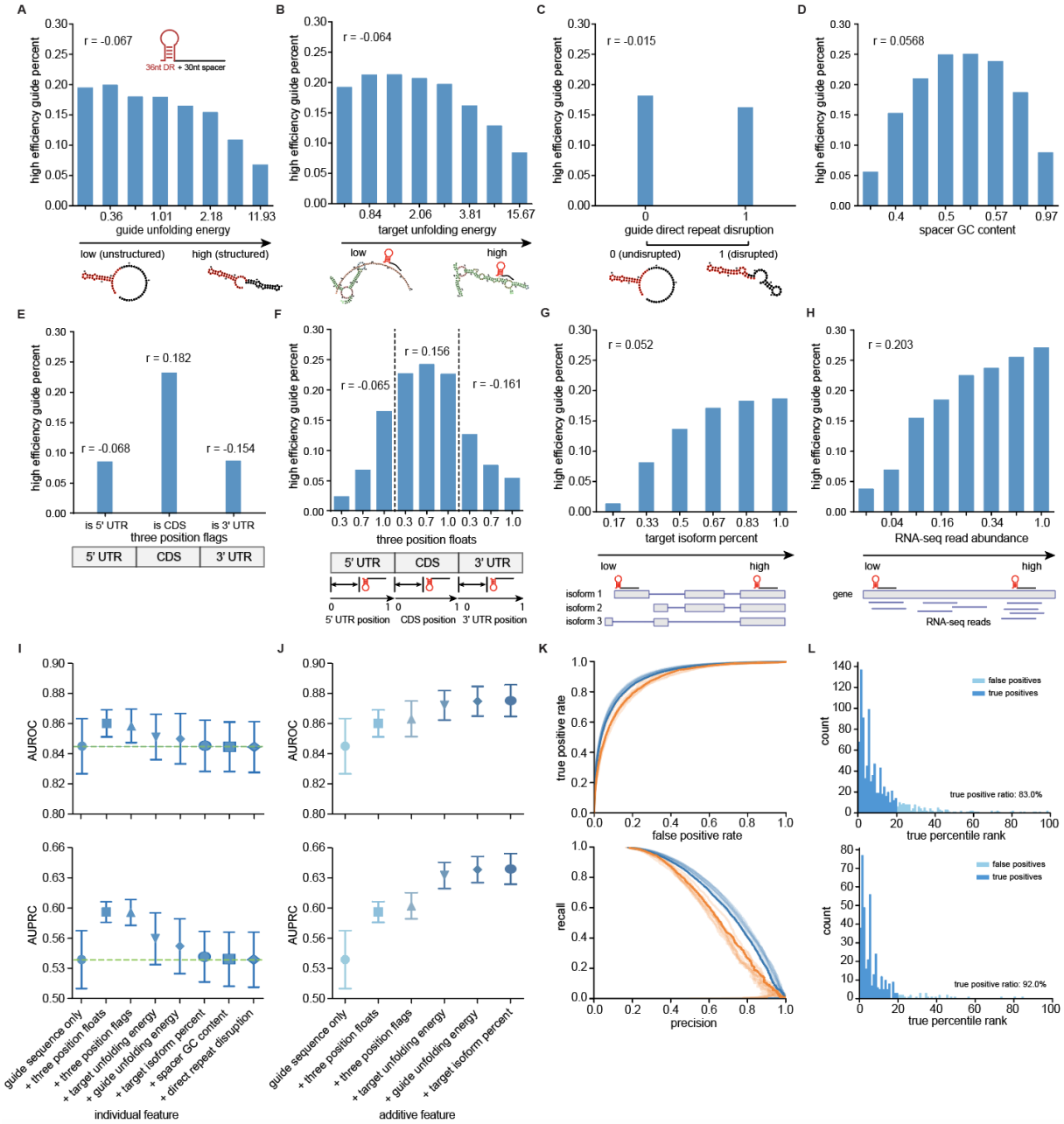


Figure 3: Addition of secondary features improves guide prediction accuracy. A. Fraction of high efficiency guides across different predicted guide RNA unfolding energies based on the LinearFold algorithm. Higher unfolding energy corresponds to a more highly structured spacer region of the guide RNA. **B.** Fraction of high efficiency guides across different predicted target unfolding energies based on the LinearFold algorithm. Lower target unfolding energies correspond to a less structured and more accessible target site. **C.** Fraction of high efficiency guides for guides with and without predicted disruption of the canonical direct repeat secondary structure. **D.** Fraction of high efficiency guides based on different spacer GC compositions. **E.** Fraction of high efficiency guides based on target position in 5′ UTR, coding sequence (CDS) or 3′ UTR. **F.** Fraction of high efficiency guides based on relative target position in the 5′ UTR,

CDS, or 3′ UTR region. **G.** Fraction of high efficiency guides based on RNA target conservation across transcript isoforms. **H.** Percent of high efficiency guides depending on relative RNA target abundance within each transcript based on RNA-seq read mapping. **I.** Model performance following addition of individual secondary features. Each secondary feature (or feature group) was added to the sequence-only CNN model individually, and the model performance was evaluated by AUROC and AUPRC (mean ± SD) of all test sets in 9-fold split of genes. Features were ordered based on final model performance. Green dashed lines denote the average AUROC and AUPRC for the guide sequence only model. **J.** Model performance following sequential addition of secondary features. Each secondary feature (or feature group) was added to the CNN model sequentially, ordered by its individual contribution to model performance from Fig. 3I. **K.** ROC (receiver operating characteristic curve) and PRC (precision-recall curve) for the final model shown in 3J. The ROC displays the true positive rate (TPR) against the false positive rate (FPR). The PRC displays the recall (true positive rate) against precision (positive predictive value). Blue curves denote model performance on the training data across all 9 data splits. Orange curves denote model performance on the held-out transcripts across all 9 data splits. Darker lines indicate the medium split. **L.** Distribution of the true percentile rank of predicted high efficiency guides. High efficiency guides for held-out transcripts are selected by the model using different target score thresholds (0.8, upper plot and 0.9, lower plot). True positives are plotted in dark blue, and false positives are plotted in sky blue.

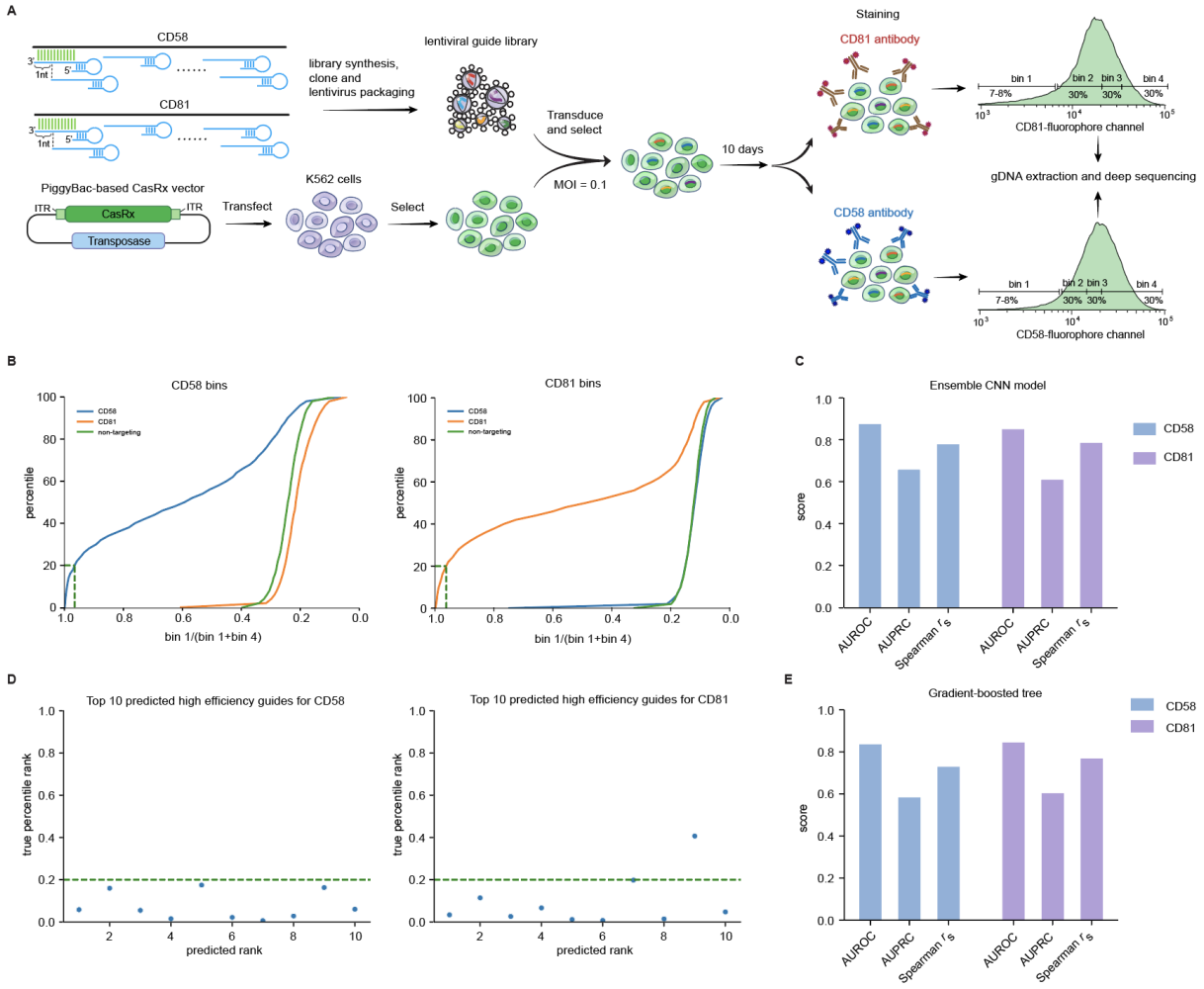


Figure 4: Model validation on an orthogonal dataset based on cell surface protein knockdown. A. Schematic of a pooled CRISPR-CasRx guide tiling screen on CD58 and CD81 transcripts in K562 cells. **B.** Cumulative distribution of guide ratio for CD58, CD81 and non-targeting pool. Guide ratio was calculated as the ratio of guide percentage in bin 1 (greatest knockdown) relative to bin 1 + bin 4 (highest level of target expression). Green dashed lines indicate the ratio for the top 20th percentile of targeting guides. **C.** Prediction of high efficiency guides for CD58 and CD81 protein knockdown using the ensemble CNN model based on survival screen data. Model prediction accuracy evaluated by AUROC, AUPRC, and Spearman's correlation coefficient (r_s) is shown for CD58 and CD81, respectively. **D.** True percentile rank of the top 10 guides for both CD58 and CD81 predicted by the CNN model. **E.** GBT model performance on CD58 and CD81. Model prediction accuracy, evaluated by AUROC, AUPRC and r_s is shown for CD58 and CD81, respectively.

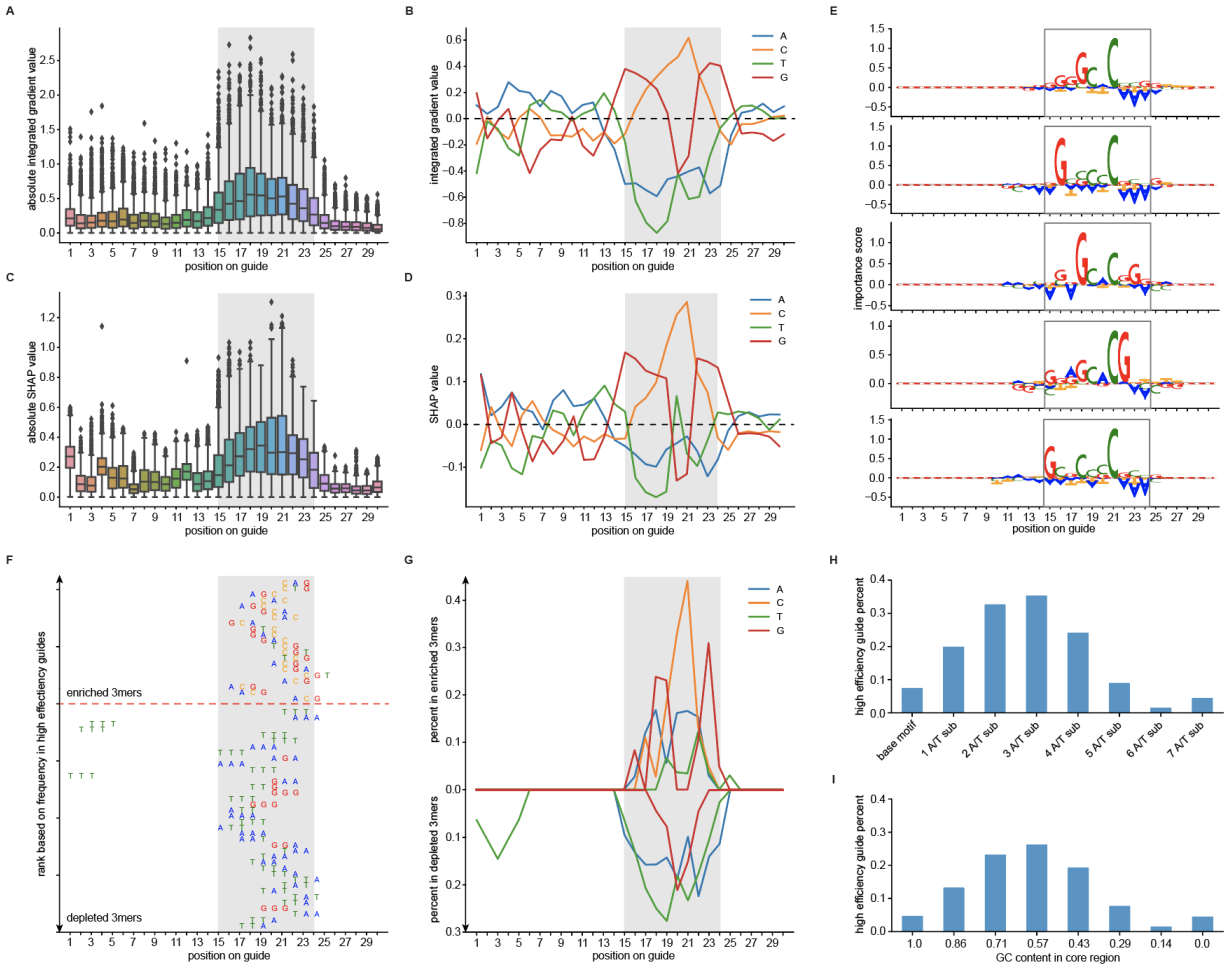
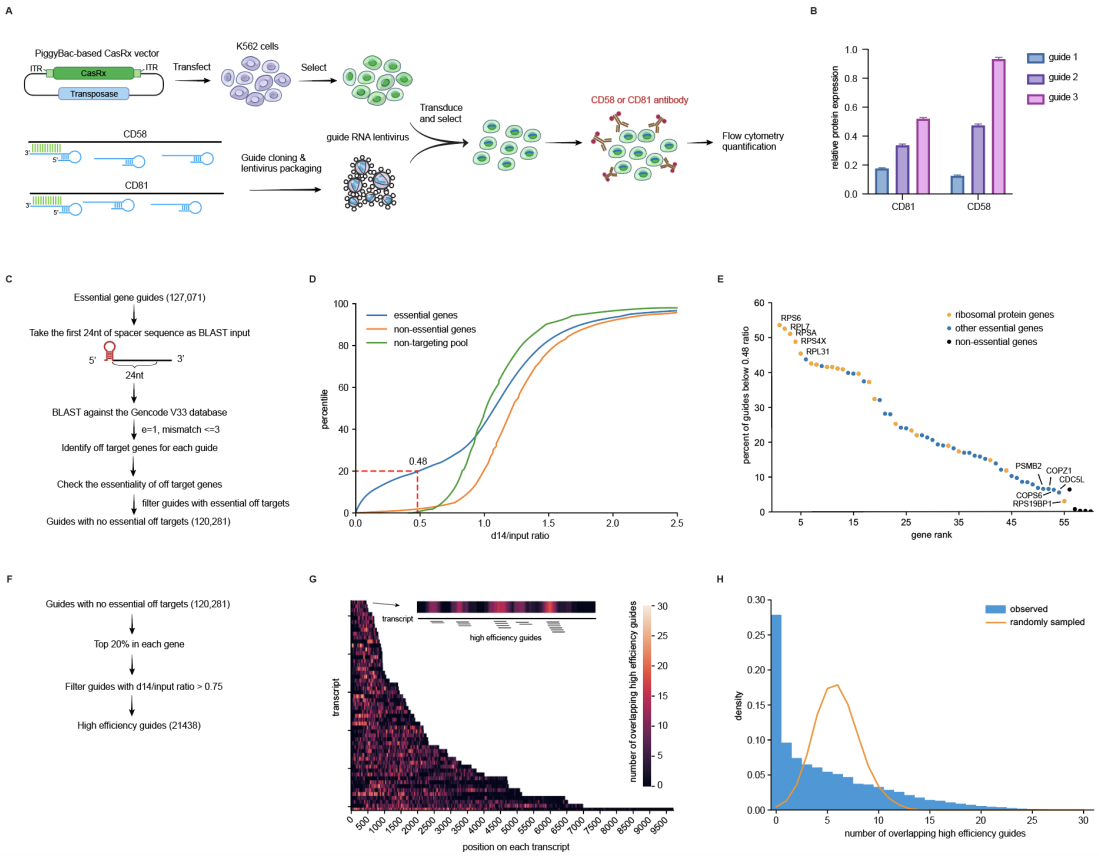
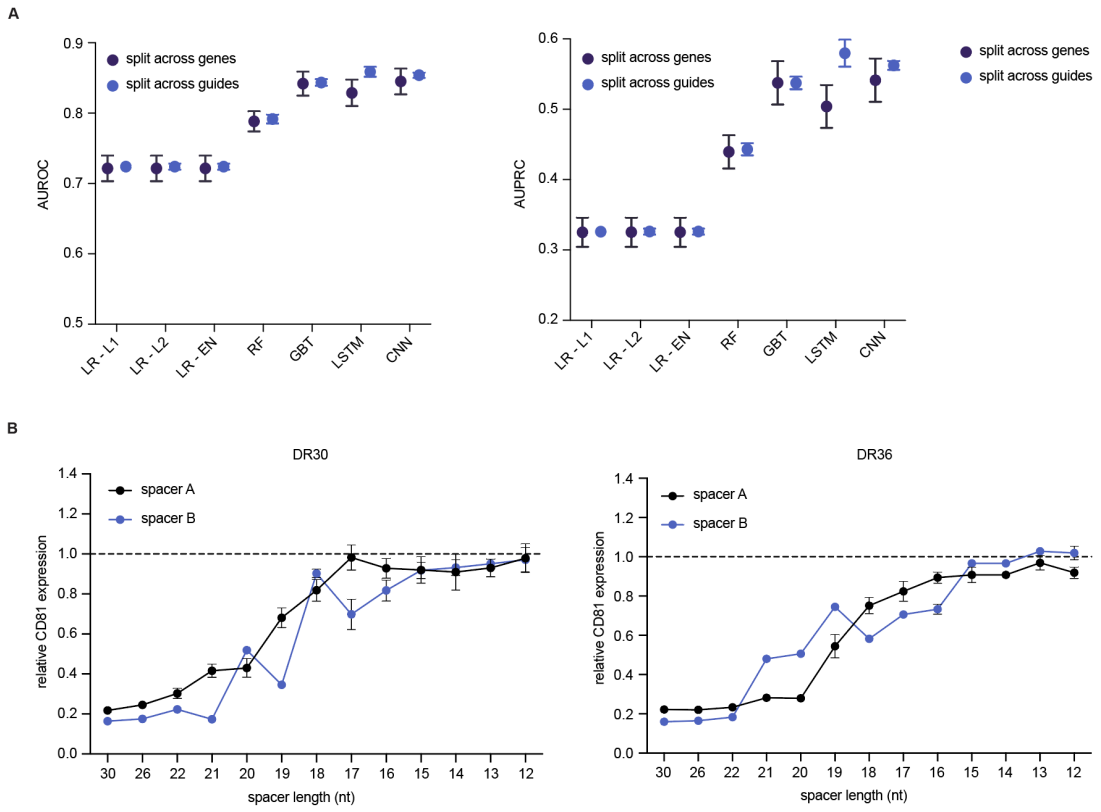


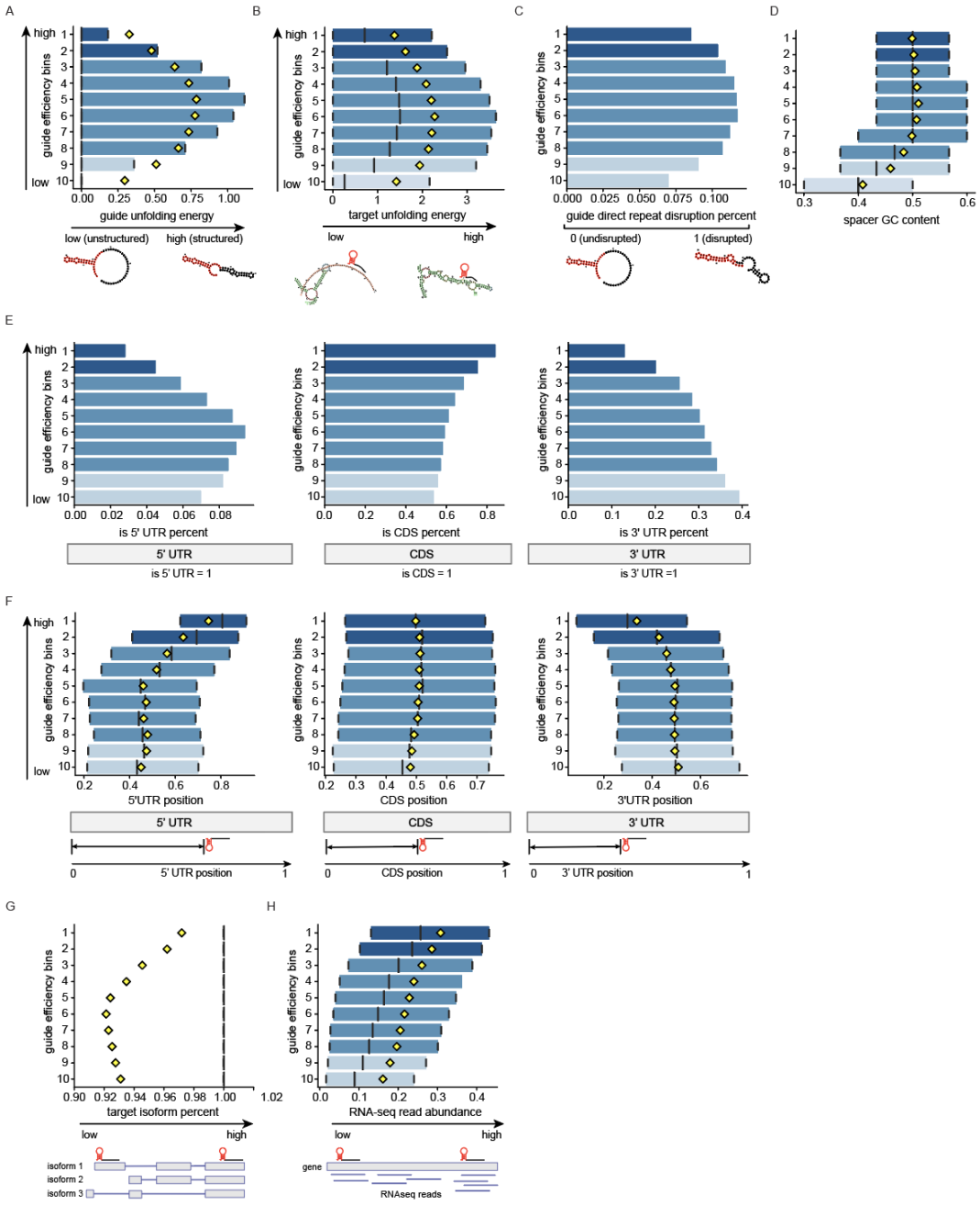
Figure 5: Feature interpretation and discovery of a core CasRx sequence motif.
A. Evaluation of the importance of each position in the guide sequence in the CNN mode using Integrated Gradients (IG). Higher absolute gradient values indicate greater importance for predicting a high efficiency guide. The grey box highlights the identified core region (position 15-24). **B.** Evaluation of the importance of each nucleotide at each position in the CNN model by IG. **C.** Evaluation of the importance of each position in the guide sequence in the GBT model. SHAP (SHapley Additive exPlanations) was applied to the GBT model to calculate the positional nucleotide importance for all test guides. **D.** Evaluation of the contribution of each nucleotide at each position in the GBT model by SHAP value. **E.** Top 5 sequence motifs identified by TF-MoDISco (Transcription Factor Motif Discovery from Importance Scores) as applied to the CNN model. Patterns are aligned to the 30 nt spacer according to the mode position of the sequences in each pattern (Figure S9A). **F.** Top enriched and depleted positional 3-mers ranked by their frequency in high efficiency guides. **G.** Summary of base composition from top enriched and depleted 3-mers at each position. **H.** Fraction of high efficiency guides in guides with the positional base motif based on Figure 5B and D and A/T substitutions within the base motif. **I.** Fraction of high efficiency guides based on core region GC content. Guides are divided to eight bins based on their GC content in the core region (position 17-23), and the percent of high efficiency guides is plotted for each bin.



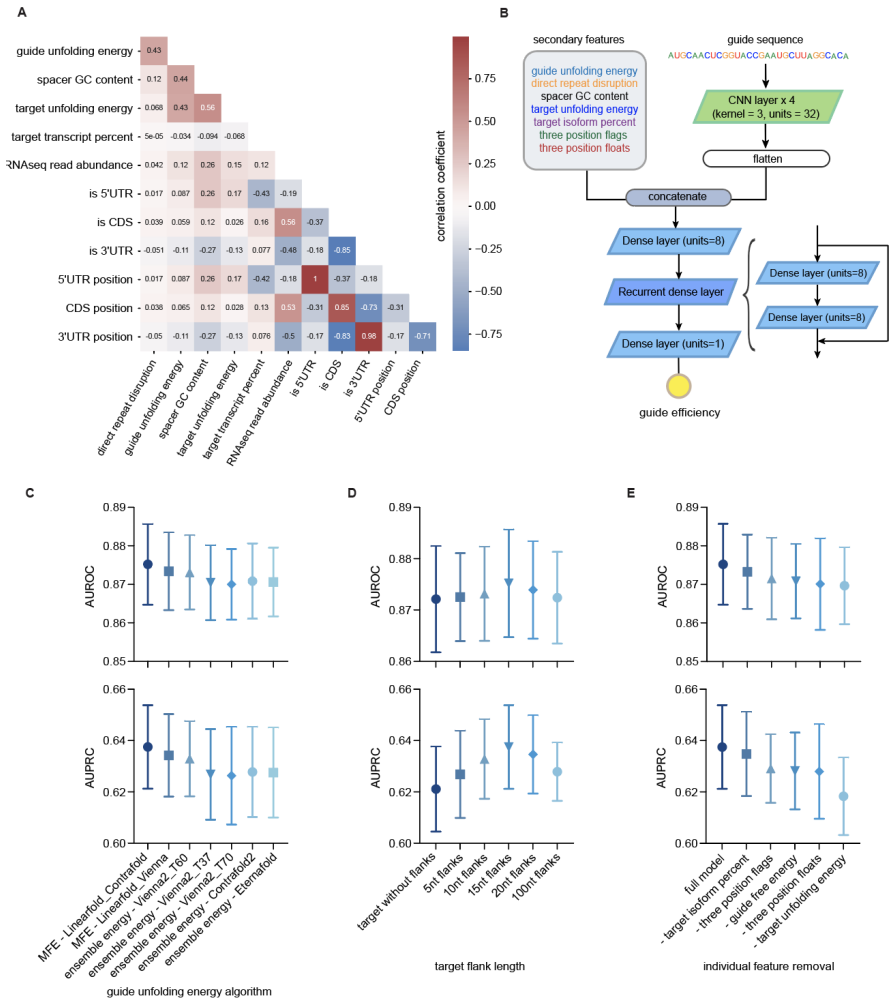
Supplementary Figure 1: CasRx knockdown in K562 cells and screen data processing. **A.** Schematic of CasRx-mediated knockdown of CD58 and CD81 proteins in K562 cells upon CasRx effector and guide delivery. **B.** Comparison of CasRx-mediated CD58 and CD81 protein knockdown in K562 cells across individual guides. Relative protein expression levels compared to the non-targeting guide are shown for each guide. Mean ± SEM for n = 3 replicates. **C.** Schematic of essential gene off-target filtering. **D.** Cumulative distribution of guide depletion ratios across essential, non-essential, and non-targeting guide categories after essential gene off-target filtering. The dashed red line indicates the ratio at the top 20th percentile of essential gene-targeting guides. **E.** Gene ranking post-filtering based on the fraction of highly depleted guides per transcript analogous to Figure 1C. Individual transcripts were ranked based on the percentage of guides below a ratio of 0.48. Orange dots denote ribosomal protein genes; Blue dots denote other essential genes; Black dots denote non-essential genes. The top 5 and bottom 5 essential genes are annotated. **F.** Definition of high efficiency guides for predictive model development. The top 20% filtered guides within each essential gene were selected, and an absolute ratio cut-off of 0.75 was applied to generate a final set of 21,438 high efficiency guides across all essential transcripts. **G.** Heat map of the positional distribution of high efficiency guides on each transcript. The heat map color indicates the number of overlapping final high efficiency guides on each position along the transcript, analogous to Figure 1G. **H.** Frequency distribution of the number of overlapping high efficiency guides across all transcripts, analogous to Figure 1E.



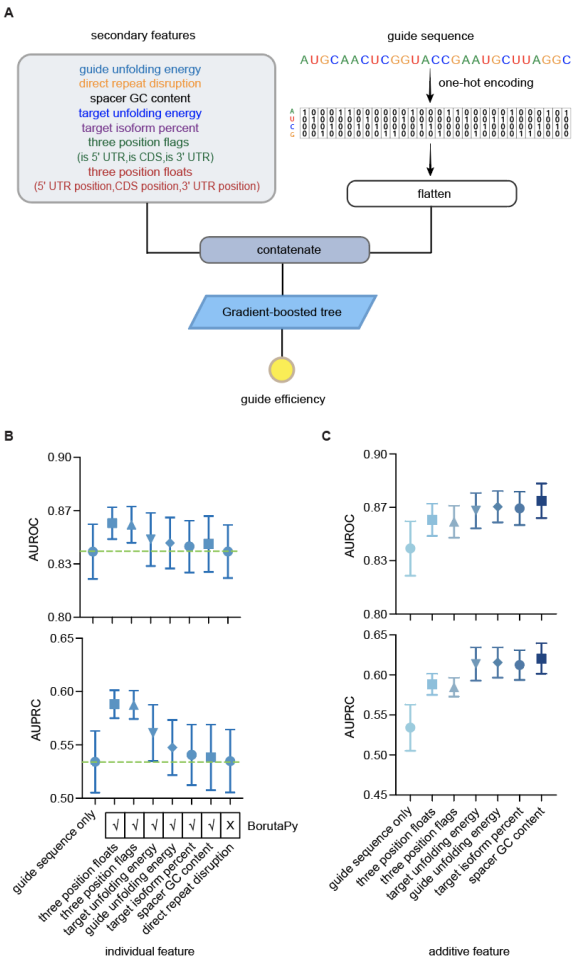
Supplementary Figure 2: Test set method comparison and CasRx guide truncation experiments. A. Comparison of model performance for different training and test set splitting methods. The survival screen data was either split into test and training sets across guides or across target transcripts to generate 9 splits for cross-validation. For each model, AUROC and AUPRC (mean ± SD) are shown. **B.** CasRx spacer truncation experiment. For two different direct repeat lengths (DR30 and DR36), two spacers targeting the cell surface marker CD81 were truncated from the 30 nt full-length spacer down to 12 nt. Cell surface protein knockdown in K562s was evaluated after 48h by flow cytometry. Mean ± SEM for n = 3 replicates is shown.



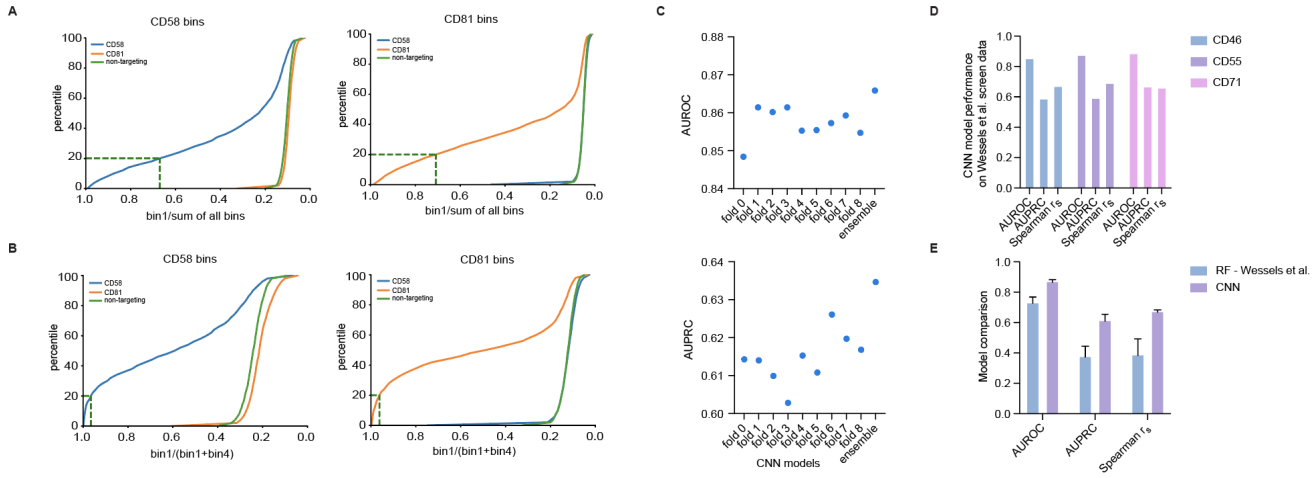
Supplementary Figure 3: Secondary feature distribution across different guide efficiency scores. A - H. Feature distribution in guide efficiency bins. Guides are divided to 10 bins based on their efficiency (d14 depletion ratio) and the distribution of each secondary feature is plotted in each bin. Bin 1: most efficient; bin 10: least efficient. For binary/flag features, bar plots are applied to show the percentage of guides with feature values of 1. For the other features, boxplots are applied to summarize the 25th, 50th and 75th percentiles of feature values. The yellow diamonds on the box plots denote the mean of feature values.



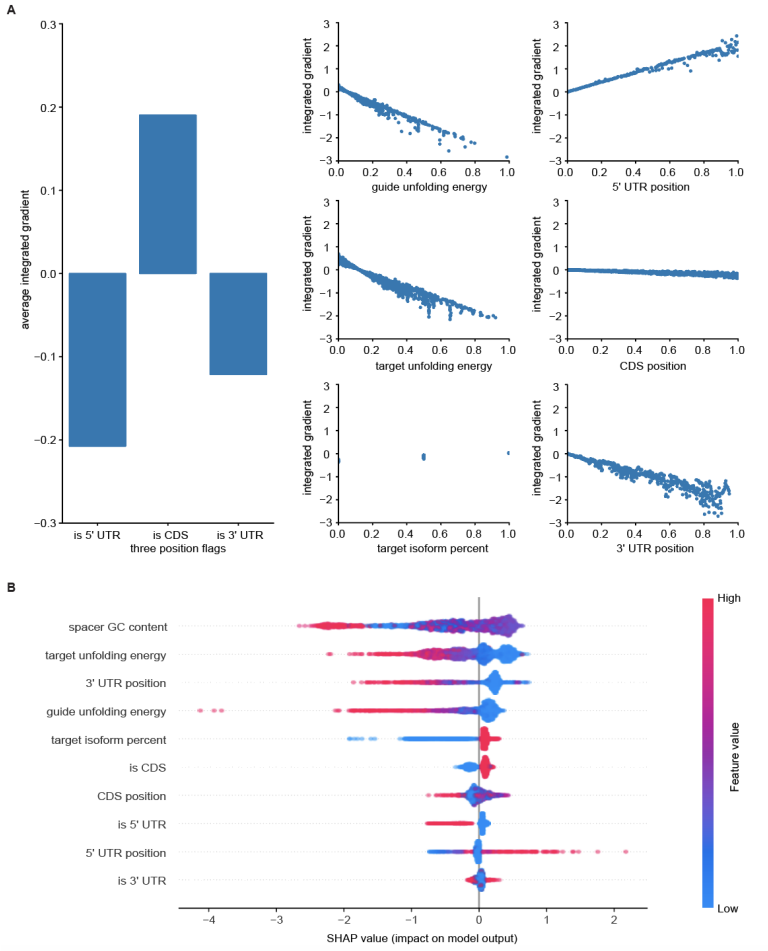
Supplementary Figure 4: Secondary feature interactions and substitutions. A. Pairwise correlation between secondary features. **B.** Schematic of the integration of secondary features into the full CNN model. **C.** Comparison of model performance based on different RNA secondary structure algorithms. The guide MFE (minimum free energy) was calculated using the CONTRAfold or Vienna model using the LinearFold algorithm and the ensemble unfolding energy was calculated using Contrafold2, Eternafold, and Vienna2. Model AUROC and AUPRC (mean ± SD) are shown for each algorithm. **D.** Model performance upon adjustment of target flank length for target RNA unfolding energy calculation. Target flanks with different lengths (0, 5, 10, 15, 20 or 100 nt) were added to the 30 nt guide-binding site to calculate the local target unfolding energy. Model AUROC and AUPRC (mean ± SD) are shown for each target flank length. **E.** Model performance upon removal of individual secondary features. Each secondary feature (or feature group) was removed individually from the final CNN model and the model AUROC and AUPRC (mean ± SD) are shown. Removal of each secondary feature reduced model accuracy, supporting each of their inclusion in the final model.



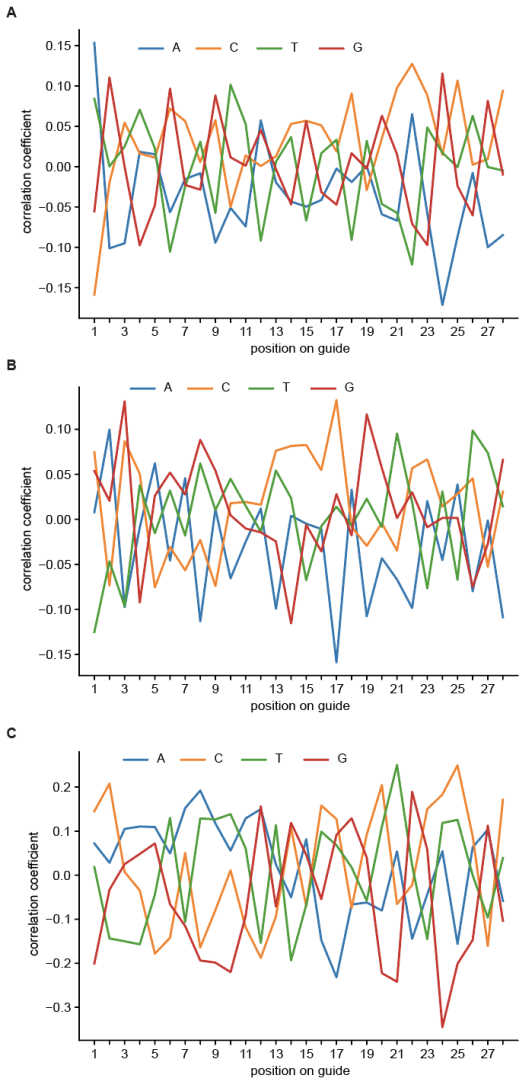
Supplementary Figure 5: Secondary feature selection for the GBT model. A. Schematic of the Gradient-boosted tree (GBT) model with secondary features. The GBT model was our best performing model amongst those not based on deep learning and was therefore selected for secondary feature addition. **B.** GBT model performance upon addition of individual features. Each secondary feature (or feature group) was added to the model individually, and the model performance was evaluated by the average AUROC and AUPRC for held-out transcripts across all 9 data splits. Features were sorted based on final model performance. The table below the plots indicates the results of the evaluation of feature contribution using BorutaPy (Kursa et al., 2010). Green dashed lines denote the average AUROC and AUPRC for the guide sequence only model. **C.** GBT model performance upon sequential feature addition. Each secondary feature (or feature group) was added to the model sequentially, ordered by its individual contribution to model performance from Figure S4B. Model AUROC and AUPRC (mean ± SD) are shown. Overall, the full GBT model performed slightly worse than the CNN model (shown in Figure 3J).



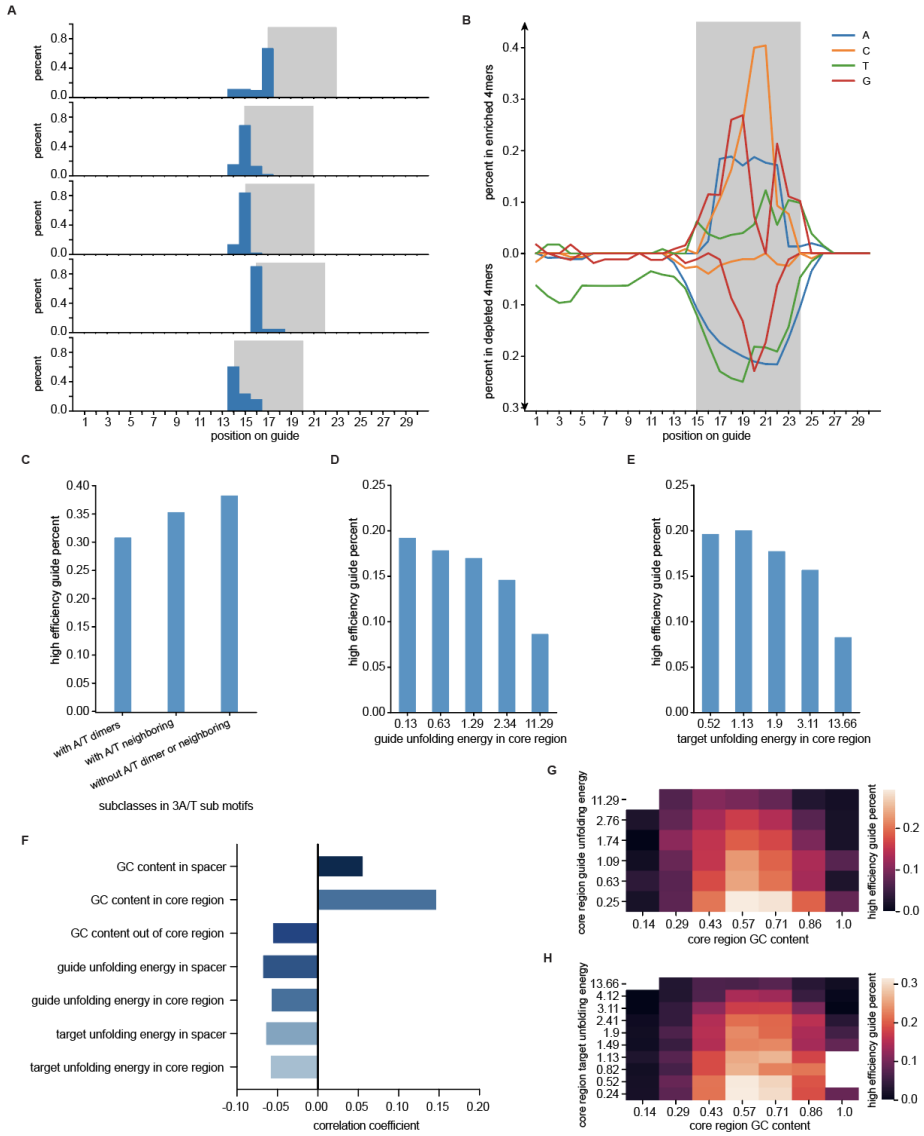
Supplementary Figure 6: Selection of screen guide efficiency metrics and model comparisons. A. Cumulative distribution of the ratio of guide percentage in bin 1 relative to all bins for CD58, CD81 and non-targeting guides. Green dashed lines indicate the ratio for the top 20th percentile of targeting guides. **B.** Cumulative distribution of the ratio of guide percentage in bin 1 relative to bin 4 for CD58, CD81 and non-targeting guides. Green dashed lines indicate the ratio for the top 20th percentile of targeting guides. The final selection of the ratio of bin 1 relative to the sum of bin 1 + 4 (shown in Figure 4B) for model evaluation was based on its superior separation of targeting guides from non-targeting controls. **C.** Comparison of model performance of individual CNN models for each training-test split of survival screen data relative to the ensemble model based on averaging the prediction of individual models. Model AUROC and AUPRC on the two validation genes (CD58 and CD81) are shown. **D.** Performance of the ensemble CNN model on a previously published CasRx guide tiling dataset for three CD genes in HEK293T cells (Wessels et al., 2020). Model AUROC, AUPRC and Spearman's correlation coefficient (r_s), are shown for CD46, CD55, and CD71 respectively. **E.** Comparison of a previously published Random forest model (Wessels et al., 2020) to the CNN model described here on opposing datasets. Model AUROC, AUPRC and Spearman's correlation coefficient (mean ± SD) are shown across genes.



Supplementary Figure 7: Secondary feature contribution for the CNN and GBT model. A. Contribution of secondary features to guide efficiency in the CNN model. The Integrated gradients approach (IG) (Sundararajan et al., 2017) was applied to evaluate feature contribution in the CNN model for all test guides. In this approach, a positive gradient value represents a positive contribution of a given feature to the prediction of a highly efficient guide and a greater absolute magnitude of the gradient indicates a stronger impact of a given feature to the classification of the guide. For binary features (position flags), bar plots indicate average IGs of all test samples. For other secondary features, scatter plots indicate IGs against input feature values for each test sample. **B.** Contribution of secondary features to guide efficiency in the GBT model. SHAP (SHapley Additive exPlanations) (Lundberg et al., 2020) was applied to evaluate feature contributions in the GBT model for all test samples. The beeswarm plot displays the SHAP value (impact on model output) against feature input value for all secondary features across all test guides. Positive SHAP scores, which drive the prediction toward the positive class, are indicated in red while negative SHAP values, which drive the prediction toward the negative class, are indicated in blue. The features are ranked based on the sum of their absolute SHAP scores across all test guides.



Supplementary Figure 8: Cas13a guide sequence nucleotide correlation with guide efficiency. A. Correlation of each nucleotide with guide efficiency at each guide position in the *Lwa*Cas13a luciferase knockdown dataset (Abudayyeh et al., 2017). 186 LwaCas13a guides for Gluc and 93 guides for Cluc were analyzed and the Pearson correlation coefficient for each positional nucleotide with guide efficiency is shown. **B.** Correlation of each nucleotide with guide efficiency at each guide position in the LwaCas13a endogenous gene knockdown dataset (Abudayyeh et al., 2017). 279 LwaCas13a guides for KRAS, PPIB and MALAT1 were analyzed and the Pearson correlation coefficient for each positional nucleotide with guide efficiency is shown. **C.** Correlation of each nucleotide with guide efficiency at each guide position in the LwaCas13a ADAPT dataset (Metsky et al., 2021). 85 perfect match guides from LwaCas13a ADAPT data were analyzed and the Pearson correlation coefficient for each positional nucleotide with guide efficiency is shown.



Supplementary Figure 9: Guide sequence motifs and core region feature importance. A. The positional distribution of the seqlets in the top 5 motifs identified by TF-MoDISco. The histogram summarizes the start position distribution of the seqlets for each motif. The grey box highlights the motif window starting at the mode position in each motif. Identified motifs are highly positional. **B.** Summary of base composition from enriched and depleted 4-mers at each position. **C.** Comparison of high efficiency guide percentages between subclasses of the 3 A/T substituted motifs. Subclasses: motifs with A/T dimers; motifs with no neighboring A/T bases; motifs with no A/T dimers or neighboring A/T bases. **D.** Fraction of high efficiency guides based on core region guide unfolding energy. **E.** Fraction of high efficiency guides based on core region target unfolding energy. **F.** Comparison of core region feature correlation and whole guide feature correlation with high efficiency guides. Spearman correlation is shown. **G.** 2D heat map of the impact of core region guide features on guide efficiency. Bins with fewer than 50 guides are shown as empty. **H.** 2D heat map of the impact of core region target features on guide efficiency. Bins with fewer than 50 guides are shown as empty.

Methods

Plasmid design

For the CasRx expression vector, we designed a piggyBac-based all-in-one plasmid containing the CasRx effector, piggyBac transposase, and antibiotic selection cassette: PB_EF1a-CasRx-msfGFP-2A-Blast. The CasRx effector is fused to msfGFP and under the control of a constitutive EF1a promoter. A nuclear localization signal SV40 NLS was added to both the N and C terminus of CasRx-msfGFP. The antibiotic selection cassette, blasticidin S deaminase is linked with CasRx-msfGFP via a P2A self-cleaving peptide. For the CasRx guide cloning vector, we designed a lentiviral vector: hU6-(CasRx DR)-EF1a-Puro-WPRE. The CasRx DR is a 36-base direct repeat (CAAGTAAACCCCTACCAACTGGTCGGGGTTTGAAAC) for CasRx pre-gRNA (Konermann et al., 2018). The 30 nt guide spacer sequence is cloned into the vector through Gibson cloning using two BsmBI cleavage sites. For individual guide truncation experiments, we designed a piggyBac-based all-in-one plasmid containing the CasRx effector, guide DR, piggyBac transposase, and antibiotic selection cassette: hU6-(CasRx DR)-TRE-CasRx-msfGFP-EF1a-rtTA-2A-Puro-CMV-transposase.

Guide library design

For the survival screening, we picked 55 essential genes from the intersection of the essential hits in three previous survival screens performed in K562 cells (Hart et al., 2015; Horlbeck et al., 2016; Luo et al., 2008). We selected the major transcript of these genes from the Refseq database and designed guides that tile these transcripts with single nucleotide resolution. All the transcripts are mature transcripts with introns removed. A total of 127,071 targeting guides were generated for the 55 essential genes. We also designed 14111 guides tiling 5 non-essential control genes (CTCFL, SAGE1, TLX1, DTX2, OR2C3). Along with 3563 non-targeting guides, we constructed a pooled library of 144745 guides.

For the validation screening on cell surface markers, 3218 guides were designed that tiled CD58 transcripts (NM_001779.3, NM_001144822.2) and CD81 transcripts (NM_004356.4,NM_001297649.2) with single nucleotide resolution. The targeting guides were pooled with 1186 non-targeting guides as the final library.

Guide library synthesis, cloning, and library amplification

For each guide spacer sequence in the guide library, we added a constant left overhang (“AACCCCTACCAACTGGTCGGGGTTTGAAAC”) and a right overhang (“TTTTTTTTGAATTCAAGCTTGGCGTAACTAGA”) to facilitate cloning. The resulting libraries were synthesized as oligo pools by Twist Biosciences, and then PCR amplified using the primer pair: Lib_F (“TCTTGTGGAAAGGACGAAACACCGCAAGTAAACCCCTACCAACTGGTCGGGGTTTG”) and Lib_R (“AGAGCTAGCCAGACGTGTGCTCTTCCGATCNNNNNNNNNTCTAGTTACGCCAAGCTTGAATTC”). The PCR reaction was performed using NEBNext High Fidelity PCR Master Mix (NEB,

catalog no. M0541L) for 20 cycles. The amplified library was gel-purified and cloned into the BsmBI digested guide cloning vector (hU6-(CasRx DR)-EF1a-Puro-WPRE) through Gibson assembly. The cloned guide library was then purified and concentrated by isopropanol precipitation.

For guide library amplification, the library plasmid was electroporated to Endura electrocompetent cells (Lucigen, catalog no. 60242-2) at 50–100 ng/ul. After electroporation, cells were recovered in LB medium for 1h, and then plated on LB agar plates with 100 ug/mL carbenicillin at 37°C for 12-14h. The colonies were then harvested at a coverage of > 500 colonies per guide. The amplified guide library plasmid was extracted using the Macherey-Nagel NucleoBond Xtra Maxi EF Kit (Macherey-Nagel, catalog no. 740424.10). To determine guide RNA representation, we PCR amplified the guide region using customized NGS primers containing Illumina adaptor sequences. NextSeq sequencing was performed to determine guide RNA representation in the guide library. We checked that the library had >70% perfectly matching guides, <0.5% undetected guides, and a skew ratio (90th percentile:10th percentile read number) of less than 10.

Lentivirus production

To produce lentivirus for the guide library, HEK 293FT cells, purchased from Thermo Fisher (Cat # R70007) were grown in DMEM supplemented with 10% FBS (D10 media) at 37 °C with 5% CO2. The cells were passaged at a ratio of 1:2 using TrypLE (Gibco) and seeded 20–24 h before transfection at 1.8×10^7 cells per T225 flask. For lentiviral plasmid transfection, the guide library plasmid was mixed with psPAX2 (Addgene, catalog no. 12260) and pMD2.G (Addgene, catalog no. 12259) in Opti-MEM, and transfected to HEK 293FT using Lipofectamine 2000 (Thermo Fisher, catalog no. 11668027) and PLUS reagent (Thermo Fisher, catalog no. 11514015). The medium was changed 4 hours after transfection with fresh, prewarmed D10 medium. Two days after the start of lentiviral transfection, the supernatant from the HEK293FT cells were harvested and filtered using a 0.45um Stericup filter. The lentiviral titer was determined through spinfection on K562 cells prior to the screen experiments.

Cell culture and CasRx cell line generation

K562 cells were purchased from ATCC (CCL-243), and cultured in RPMI 1640 medium supplemented with 10% FBS at 37 °C with 5% CO2. To generate a stable CasRx-expressing K562 cell line, we transfected K562 cells with the piggyBac-based all-in-one CasRx expression vector (PB_EF1a-CasRx-msfGFP-2A-Blast) using Lipofectamine 3000 Transfection Reagent (Thermo Fisher, catalog no. L3000001). Two days after transfection, we selected the cells with 10 µg/ml blasticidin S (Thermo Fisher, catalog no. A1113903). After selection for 1-2 weeks, we checked the percent of CasRx-expressing cells using flow cytometry and confirmed that more than 95% of cells expressed CasRx-GFP.

Survival screen

The guide library for the survival screening was lentivirally transduced at MOI=0.2 by spinfection into the stable CasRx-expressing K562 cell line. We ensured the guide library had a coverage of

>1000 cells expressing each guide. Two days after transduction, we selected the cells with 1 μg/ml puromycin to ensure guide expression and further cultured for 14 days. We harvested cells at day 14 (end of the screen), and we extracted the genomic DNA using Zymo Research Quick-gDNA MidiPrep (Zymo Research, cat. no. D4075). The guide region was PCR amplified using customized NGS primers containing Illumina adaptor sequences. The PCR products were then gel purified and quantified with Nanodrop and Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, cat. no. Q32851). Different guide libraries were pooled according to their concentration. The pooled guide libraries were sequenced on the Illumina NextSeq, with 80 cycles of read 1 (forward) and 8 cycles of index 1. Three replicates were performed for the survival screening.

Validation screen on CD58 and CD81

The guide library for the validation screening was lentivirally transduced at MOI=0.1 by spinfection into the stable CasRx-expressing K562 cell line. We started with 45 million cells to make sure the guide library has a coverage of >1000 cells for each guide. After spinfection, we selected the cells with 1 μg/ml puromycin to ensure guide expression and further cultured for 10 days. At the end of the screening, we divided the cells into two pools and stained them with CD58 antibody (BD Biosciences, catalog no. 564363) and CD81 antibody ((BD Biosciences, catalog no. 561958) respectively. After taking out a small proportion as unsorted control, we FACS sorted each cell pool into four bins based on target gene expression level indicated by antibody-conjugated fluorescence intensity. Specifically, cells were first gated by forward and side scatter to select for live, single cells. Next, cells were gated by GFP to select for those with high expression of CasRx. Within the high CasRx-expressing cells, we sorted the cells into four bins based on the intensity of CD58 or CD81-conjugated fluorescence intensity. As high efficiency guides were defined as the top 20% for each gene, we set the bin with the lowest target gene expression (bin 1) at 7-8%, which is the fraction of the target gene's high efficiency guide number in the whole library:1600*20%/4401). The rest of the population was divided into three bins of similar size (~30%). The genomic DNA for cells in each bin was extracted and sequenced as in the survival screening. Four replicates were performed for the validation screen.

Data preprocessing and analysis

On a per guide RNA basis, we calculated its percentage in the day 14 guide pool and the input library pool. Guide efficiency was evaluated by the depletion ratio of guide percentage in day 14 pool to the input pool. We ranked guides within each gene based on their average ratio of the three replicates, and we defined the top 20% guides within each gene and with ratio < 0.75 as high efficiency guides. We also excluded all guides from the transcript RPS19BP1 because most guides were not depleted from the screening and clustered with non-essential gene guides.

For the validation screening, we first filtered guides with less than 200 counts in all CD58 bins and CD81 bins. We then calculated each guide's percentage in each bin and calculated the relative ratio of guide percentage between bins. After comparing different ratios (including the

ratio of guide percentage in bin 1 to the sum of all bins and the ratio of guide percentage in bin 1 to the sum of bin 1 and bin 4), we decided to use the ratio of guide percentage in bin 1 to the sum of its percentage in bin 1 and bin 4 for evaluation of guide efficiency. We then ranked guides within each gene based on their average ratio of the four replicates, and we defined the top 20% guides for each gene as high efficiency guides.

Off-target filtering

We performed BLAST to identify the potential off target genes for our guides. As the first 24 nucleotides in CasRx guides are shown to be most effective (**Figure S2B**), we took the first 24 nucleotides of each guide as BLAST input. BLAST was performed using a generous E value of 1 (e=1) against the Gencode V33 database. BLAST results were parsed and off target genes were identified as those with up to three mismatches to the guide input. To check the essentiality of the off target genes, we made an essential gene list by combining the essential gene hits from the three previous survival screens in K562 cells and we compared the off target genes with the essential gene list. Guides with essential off-target genes were filtered. For our survival screening, 6790 guides were filtered and 120281 guides were left.

Analysis of the positional distribution of effective guides

For each transcript, we calculated the number of top 20th percentile guides (with ratio below the efficiency cut-off (0.45 ratio)) at each position on the transcript, and plotted the results with a heatmap. We further summarized the distribution of effective guide numbers across all positions with a histogram. In theory, a position would have at most 30 guides covering it, so the number of effective guides ranges from 0 to 30 for each position. We compared the results with a randomly sampled distribution, which is simulated from random sampling of 20% guides in the library for 100 times. In theory, the randomly sampled distribution would show a peak at 6 (30*20%), which agrees with our simulation results.

Data splits

For model hyperparameter tuning and evaluation, we split our 54 essential transcripts into 9 folds, each containing a unique set of 6 test transcripts. The 54 transcripts were distributed evenly across the 9 folds according to their high efficiency guide percent to make the 9-fold split relatively balanced. Using the predefined transcript splits, we performed 9-fold cross-validation to tune model hyperparameters and compare prediction accuracy between models.

Feature calculation and model inputs

For the sequence input, each 30 nt guide spacer was one-hot encoded into four binary vectors of length 30 to represent the nucleotide identity at each position.

To calculate guide unfolding energy, we used LinearFold, a linear-time RNA secondary structure prediction algorithm (Huang et al., 2019) on the full-length guide sequence (36nt DR +30nt spacer). We started with the default parameters and the CONTRAfold v2.0 model (Do et al., 2006; Lorenz et al., 2011; Wayment-Steele et al., 2020) provided by the LinearFold software at https://github.com/LinearFold/LinearFold. We subtracted the predicted MFE energy with the

baseline energy (MFE of the unstructured guide with the 30 nt spacer unfolded) to calculate guide unfolding energy. We also tested the Vienna RNAfold model in LinearFold as a comparison. To determine whether using the ensemble guide unfolding energy instead of MFE could improve model prediction, we further tested three RNA structure prediction algorithms (Contrafold2, Eternafold, Vienna) wrapped by Arnie (https://github.com/DasLab/arnie) to calculate the ensemble guide unfolding energy with the partition function (Do et al., 2006; Lorenz et al., 2011; Wayment-Steele et al., 2020). For the Vienna package, we tested different temperature(T) settings: 37°C , 60 °C, and 70 °C. In our final model, we used the guide unfolding energy calculated by LinearFold's default CONTRAfold v2.0 model as it improved model prediction accuracy to the greatest extent.

To calculate the "direct repeat disruption" feature, we used the guide secondary structure predicted by LinearFold's CONTRAfold v2.0 model to determine whether the 36 nt direct repeat region structure is different from the canonical reference structure. A feature value of "1" indicates that the guide direct repeat structure is predicted to be disrupted.

To calculate target unfolding energy, we first used LinearFold's CONTRAfold v2.0 model to predict MFE of the native local target region using the local target sequence. We then predicted MFE of the guide unwound local target region by supplying the algorithm with the constraint that the 30 nt guide-binding site is unpaired. (This can be achieved by feeding in an additional constraint structure with the guide-binding site annotated with "."). We then subtracted the former MFE (MFE of the native target region) by the latter (MFE of the guide unwound target region) to estimate local target unfolding energy. The local target region was defined as the 30 nt guide-binding site with 15 nt flanking sequence on both sides. Flanking sequences of different lengths were compared, and the length 15 was chosen for the final model as it improved model prediction accuracy to the greatest extent.

To calculate target isoform percent, we obtained all transcript isoforms for each gene from the Refseq database, and calculated the percent of isoforms that a guide targets.

To calculate RNA-seq read abundance, we used a polyA plus RNA-seq dataset on K562 nuclear fraction from ENCODE (https://www.encodeproject.org/experiments/ENCSR000CPS/). We obtained the raw RNA-seq read data in Fastq format and counted the occurrence of each guide's 30 nt target region. We then normalized the target region's count within each gene.

To calculate the three position flags, we obtained Refseq's annotations of the 5′ UTR, CDS, or 3′ UTR region for our target transcripts. Guides that target the 5′ UTR, CDS, or 3′ UTR region have a flag value of 1 for that correspondent feature, and 0 for the other two flag features. To calculate the three position floats (5′ UTR position,CDS position,3′ UTR position), we calculated the relative position of the guide target site in the 5′ UTR, CDS, or 3′ UTR region. Guides located out of the region have a flag value of 0 for the correspondent feature.

Model architecture

Sequence-only models
For linear models and ensemble models, the one-hot encoded guide sequence was flattened and converted to 30*4= 120 flag features. The features are then fed into the models to generate the output. For the CNN model, the one-hot encoded guide was treated as a 4-channel image, and a few 1D convolutional layers were applied to generate a feature map, which was flattened and passed to a dense layer to generate the final output. For the biLSTM model, the guide sequence was treated as a sentence with four characters, and two LSTMs, each processing the input sequence in one direction (forward or backward), were applied to generate sequence representations. The resulting vectors were merged, flattened, and passed to a dense layer to generate the final output.

Full model with secondary features
For the CNN model with secondary features, the one-hot encoded guide was passed to a few convolutional layers as in the sequence-only model. The output from the CNN layers was flattened and concatenated with the normalized secondary features. The concatenated feature vector was sequentially passed to a dense layer, a recurrent dense layer and a final dense layer of 1 unit to generate the output. All dense layers use leaky ReLU as the activation function. The CNN layer kernel size, unit number, layer number and the dense layer unit number were defined after hyperparameter tuning.
For the Gradient-boosted classification tree, the one-hot encoded guide sequence was flattened and converted to 30*4= 120 flag features. The sequence features are concatenated with the normalized secondary features, and then fed into the model to generate output.

Model training, hyperparameter tuning and evaluation
All models were trained to solve a binary classification task – predicting high efficiency guides, and the model output is the probability that a guide is a high efficiency guide.
The linear models and ensemble models were trained in scikit-learn 0.24 and the deep learning models (LSTM and CNN) were trained in TensorFlow 2.3.1. For the deep learning models, we used binary cross-entropy as the loss function and applied the Adam optimizer for model training. Early stopping was used to prevent model overfitting.
For all models, the prediction accuracy is evaluated by AUROC (Area Under the Receiver Operating Characteristic curve) and AUPRC (The Area Under Precision-Recall Curve).

To tune hyperparameters and evaluate model performance, we used 9-fold cross-validation over the hyperparameter space. For linear models and ensemble models, we used the "GridSearchCV" function in scikit-learn to perform a grid search over the hyperparameter set. For deep learning models, we used the Hyperband tuner in TensorFlow to select top models quickly by filtering poor models during training.

The hyperparameter sets for all models are listed below:

- logistic regression with L1 regularization: regularization strength - logarithmic in (10^{-5}, 10^{5}))
- logistic regression with L2 regularization: regularization strength - logarithmic in (10^{-5}, 10^{5}))

- logistic regression with elastic net regularization: regularization strength - logarithmic in (10^{-4}, 10^{4})), L1 ratio - equally spaced from 0.1 to 1.
- Gradient-boosted classification trees: number of trees – [100,200,400,800,1000,1200,1500,1800,2000], maximum depth of a tree – [2,4,8], the number of features to consider when looking for the best split - all, sqrt(n_features), log2(n_features).
- Random forest (RF): number of trees – [100,200,400,800,1000,1200,1500,1800,2000], number of features to consider when looking for the best split - all, sqrt(n_features), log2(n_features).
- Long short-term memory recurrent neural network (LSTM): LSTM units - [16, 32,64,128], dense layer units – [8, 16, 32], recurrent dense layer number – [0,1,2,3], dropout rate - [0.0, 0.1, 0.25]
- Convolutional neural network (CNN): CNN layer kernel size – [3,4,5], CNN units- [8,16,32,64], CNN layer number – [3,4,5], dense layer units - [8,16,32,64], recurrent dense layer number – [0,1,2,3]

For all models, we chose the hyperparameter set with the highest average AUROC across all test sets in the 9-fold splits, and evaluated the final model performance using both the average AUROC and average AUPRC across test sets.

Secondary feature selection

For the CNN model, we added each secondary feature individually to guide sequence features and calculated the change in model performance. We selected features that successfully improved model performance, and added these features sequentially upon guide sequence features to check feature redundancy. We also tried removing individual features from the final model to confirm the necessity of the features.

For the Gradient-boosted tree, besides the above methods, we also used Boruta, an all-relevant feature selection method that aims to find all features useful for prediction (Kursa et al., 2010). We implemented it using BorutaPy, the Python implementation of Boruta (https://github.com/scikit-learn-contrib/boruta_py) on our Gradient-boosted tree.

Final model and model testing on the validation screens

We chose the CNN model as our final model after hyperparameter tuning and model comparison. We re-trained the model using all the survival screen data. To prevent overfitting, we split out a validation set during model training as in the previous 9-fold cross-validation split. We built 9 individual models using different validation sets from the 9-fold split of essential transcripts, and we compared their performance on the two cell surface markers, CD58 and CD81. We further built an ensemble model that averaged the prediction of all the individual models. We found that the ensemble model outperformed all individual models on the two CD genes, so we set the ensemble CNN model as our final model. As a comparison, we also retrained the best non-deep learning model, the Gradient-boosted tree using all the survival screen data. We tested the model on the two CD genes and evaluated model performance using AUROC and AUPRC.

Model comparison with Wessels et al. model

We tested the performance of the Random forest model from Wessels et al. on our CD genes and essential genes using the web server https://cas13design.nygenome.org. We evaluated the

model performance using AUROC, AUPRC and Spearman's correlation coefficient, r_s. As the Random forest model is designed for 23 nt long guides, we extended the guides from their model output to 30 nt (extends toward the 3′ end) to be in accordance with our screen data. For comparison, we retrieved the CasRx guide tiling screen dataset on three CD genes, CD46, CD55, and CD71, from Wessels et al. and tested our model's performance. We adjusted the guide length to 23 nt in our model to be in accordance with their screen data, and we set the top 20% guides for each gene as "high efficiency guide". The model performance was also evaluated by AUROC, AUPRC and Spearman's correlation coefficient, r_s.

Model interpretation and feature contributions

For the CNN model, we applied "Integrated Gradients" (IG) to investigate feature contributions in the model. "Integrated Gradients" is an attribution method that evaluates feature importance by integrating the gradient of output to input features along the straightline path from the baseline input to the actual input value (Sundararajan et al., 2017). Due to the non-linearity of the deep learning model, we applied "Integrated Gradients" to the best-performing individual CNN model on CD genes rather than the ensemble model. To compute integrated gradients, we first set all-zero baselines for the sequence input, position flags and position floats, and used average baselines for other features. Next, we generated a linear interpolation between the baselines and the inputs using 50 steps. We then computed gradients using the "tf.GradientTape" function in TensorFlow for the interpolated points, and approximated the gradients integral with the trapezoidal rule. To evaluate the relative importance of each position on the guide, we averaged the absolute integrated gradient values at each position across all test sequences. To evaluate the contribution of each nucleotide at each position, we averaged the integrated gradients for that nucleotide across all test sequences.

For the Gradient-boosted tree, we applied SHAP (SHapley Additive exPlanations) to investigate feature contributions in the model. SHAP is a game theoretic approach that estimates how each feature contributes to the model output by providing the SHAP value for each input feature (Lundberg et al., 2020). We implemented the SHAP package from https://github.com/slundberg/shap, and applied it to our Gradient-boosted tree. To evaluate the relative importance of each position on the guide, we averaged the SHAP values at each position across test sequences. To evaluate the contribution of each nucleotide at each position, we averaged the SHAP values for that nucleotide across test sequences.

Cas13a guide sequence contribution to guide efficiency

We analyzed three Cas13a guide efficiency datasets: 1) the Luciferase knockdown dataset containing 186 LwaCas13a guides for Gaussia luciferase (Gluc) and 93 guides for Cypridina Luciferase (Cluc) (Abudayyeh et al., 2017); 2) the endogenous gene knockdown dataset containing 93 LwaCas13a guides for each of KRAS, PPIB and MALAT1 (Abudayyeh et al., 2017); and 3) the ADAPT dataset containing 85 perfect match LwaCas13a guides for virus detection (Metsky et al., 2021). We calculated the Pearson correlation between each nucleotide at each position with guide efficiency to evaluate the sequence contribution.

Motif discovery

For motif discovery, we used TF-MoDISco (Transcription Factor Motif Discovery from Importance Scores), an algorithm that discovers motifs by clustering important regions in sequences using per-base importance scores (Shrikumar et al., 2018). We implemented TF-MoDISco from https://github.com/kundajelab/tfmodisco using the integrated gradients of all high efficiency guides in our training data as input. We ran TF-MoDISco with a sliding window size of 7 and a flank length of 2. For final motif processing, we trimmed the clustered motifs to a window size of 6, added an initial flank length of 2 and a final flank length of 3 to get the final motifs. The top 5 active motifs are picked and aligned to the 30 nt spacer according to the mode position of sequences in each motif.

Nmer analysis

To identify enriched or depleted positional nmers, we divided our survival screen data to 9 folds as in the model training workflow and calculated the ratio of all possible positional nmers' percentage in high efficiency guides to non-high efficiency guides in the training set and test set respectively for each fold. We identified enriched (or depleted) nmers based on their ratio in the training set with a predefined ratio cut-off. We selected the nmers identified as enriched (or depleted) across all folds, and ranked them by their average percent in high efficiency guides in the test sets across all folds. The initial ratio cut-off is set as 2 for enriched nmers and 0.5 for depleted nmers. The cut-off is adjusted during the nmer identification process so that the percent of guides with enriched nmers are ~20% and the percent of guides with depleted nmers are ~40%. We mainly focused on 3 mers and 4 mers in this paper.

Core region feature calculation

For the guide unfolding energy in the core region, we used LinearFold to predict MFE of the whole length guide (DR + spacer) with the constraint that the 7 nt core region is unpaired. We then subtracted the calculated MFE from the MFE of the native guide to estimate guide unfolding energy in the core region. For the target unfolding energy in the core region, we used LinearFold to predict MFE of the local target region with the constraint that the 7 nt core region is unpaired. We then subtracted the calculated MFE from the MFE of the native local target region to estimate target unfolding energy in the core region. The local target region was defined as the 30 nt guide-binding site with 15 nt flanking sequence on both sides.

References

1. Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., Verdine, V., Cox, D. B. T., Kellner, M. J., Regev, A., Lander, E. S., Voytas, D. F., Ting, A. Y., & Zhang, F. (2017). RNA targeting with CRISPR–Cas13. *Nature*, *550*(7675), 280–284.
2. Abudayyeh, O. O., Gootenberg, J. S., Franklin, B., Koob, J., Kellner, M. J., Ladha, A., Joung, J., Kirchgatterer, P., Cox, D. B. T., & Zhang, F. (2019). A cytosine deaminase for programmable single-base RNA editing. *Science*, *365*(6451), 382–386.
3. Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B. T., Shmakov, S., Makarova, K. S., Semenova, E., Minakhin, L., Severinov, K., Regev, A., Lander, E. S., Koonin, E. V., & Zhang, F. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, *353*(6299). https://doi.org/10.1126/science.aaf5573
4. Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B., Gu, F., Qu, S., Huang, D., Wei, J., & Liu, Q. (2018). DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology*, *19*(1), 80.
5. Cox, D. B. T., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J., Joung, J., & Zhang, F. (2017). RNA editing with CRISPR-Cas13. *Science*, *358*(6366), 1019–1027.
6. Do, C. B., Woods, D. A., & Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* , *22*(14), e90–e98.
7. Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, *34*(2), 184–191.
8. East-Seletsky, A., O'Connell, M. R., Knight, S. C., Burstein, D., Cate, J. H. D., Tjian, R., & Doudna, J. A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*, *538*(7624), 270–273.
9. Han, S., Zhao, B. S., Myers, S. A., Carr, S. A., He, C., & Ting, A. Y. (2020). RNA-protein interaction mapping via MS2- or Cas13-based APEX targeting. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(36), 22068–22079.
10. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., & Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, *163*(6), 1515–1526.
11. Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M., & Weissman, J. S. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife*, *5*. https://doi.org/10.7554/eLife.19760

12. Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., & Mathews, D. H. (2019). LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. In *Bioinformatics* (Vol. 35, Issue 14, pp. i295–i304). https://doi.org/10.1093/bioinformatics/btz375
13. Kim, H. K., Kim, Y., Lee, S., Min, S., Bae, J. Y., Choi, J. W., Park, J., Jung, D., Yoon, S., & Kim, H. H. (2019). SpCas9 activity prediction by DeepSpCas9, a deep learning–based model with high generalization performance. In *Science Advances* (Vol. 5, Issue 11, p. eaax9249). https://doi.org/10.1126/sciadv.aax9249
14. Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S., & Kim, H. (henry). (2018). Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nature Biotechnology*, *36*(3), 239–241.
15. Konermann, S., Lotfy, P., Brideau, N. J., Oki, J., Shokhirev, M. N., & Hsu, P. D. (2018). Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell*, *173*(3), 665–676.e14.
16. Kursa, M. B., Rudnicki, W. R., & Others. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13.
17. Li, S., Li, X., Xue, W., Zhang, L., Yang, L.-Z., Cao, S.-M., Lei, Y.-N., Liu, C.-X., Guo, S.-K., Shan, L., Wu, M., Tao, X., Zhang, J.-L., Gao, X., Zhang, J., Wei, J., Li, J., Yang, L., & Chen, L.-L. (2021). Screening for functional circular RNAs using the CRISPR–Cas13 system. In *Nature Methods* (Vol. 18, Issue 1, pp. 51–59). https://doi.org/10.1038/s41592-020-01011-4
18. Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology: AMB*, *6*, 26.
19. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, *2*(1), 56–67.
20. Luo, B., Cheung, H. W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J. S., Beroukhim, R., Weir, B. A., Mermel, C., Barbie, D. A., Awad, T., Zhou, X., Nguyen, T., Piqani, B., Li, C., Golub, T. R., Meyerson, M., … Root, D. E. (2008). Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(51), 20380–20385.
21. Mahas, A., Aman, R., & Mahfouz, M. (2019). CRISPR-Cas13d mediates robust RNA virus interference in plants. *Genome Biology*, *20*(1), 263.
22. Metsky, H. C., Welch, N. L., Haradhvala, N. J., Rumker, L., Zhang, Y. B., Pillai, P. P., Yang, D. K., Ackerman, C. M., Weller, J., Blainey, P. C., Myhrvold, C., Mitzenmacher, M., & Sabeti, P. C. (2021). Designing viral diagnostics with model-based optimization. In *bioRxiv* (p. 2020.11.28.401877). https://doi.org/10.1101/2020.11.28.401877
23. Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., & Kundaje, A. (2018). Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1811.00416
24. Spitale, R. C., Crisalli, P., Flynn, R. A., Torre, E. A., Kool, E. T., & Chang, H. Y. (2013).

RNA SHAPE analysis in living cells. *Nature Chemical Biology*, *9*(1), 18–20.
25. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 3319–3328). PMLR.
26. Wayment-Steele, H. K., Kladwang, W., Participants, E., & Das, R. (2020). RNA secondary structure packages ranked and improved by high-throughput experiments. In *bioRxiv* (p. 2020.05.29.124511). https://doi.org/10.1101/2020.05.29.124511
27. Wessels, H.-H., Méndez-Mancilla, A., Guo, X., Legut, M., Daniloski, Z., & Sanjana, N. E. (2020). Massively parallel Cas13 screens reveal principles for guide RNA design. *Nature Biotechnology*, *38*(6), 722–727.
28. Wilson, C., Chen, P. J., Miao, Z., & Liu, D. R. (2020). Programmable m6A modification of cellular RNAs with a Cas13-directed methyltransferase. In *Nature Biotechnology* (Vol. 38, Issue 12, pp. 1431–1440). https://doi.org/10.1038/s41587-020-0572-6
29. Xu, C., Zhou, Y., Xiao, Q., He, B., Geng, G., Wang, Z., Cao, B., Dong, X., Bai, W., Wang, Y., Wang, X., Zhou, D., Yuan, T., Huo, X., Lai, J., & Yang, H. (2021). Programmable RNA editing with compact CRISPR-Cas13 systems from uncultivated microbes. *Nature Methods*, *18*(5), 499–506.
30. Xue, L., Tang, B., Chen, W., & Luo, J. (2019). Prediction of CRISPR sgRNA Activity Using a Deep Convolutional Neural Network. *Journal of Chemical Information and Modeling*, *59*(1), 615–624.