nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01830-8

Prediction of on-target and off-target activity of CRISPR–Cas13d guide RNAs using deep learning

Received: 8 December 2021

Accepted: 16 May 2023

Published online: 03 July 2023

Check for updates

Hans-Hermann Wessels[1,2,6], Andrew Stirn[1,3,6], Alejandro Méndez-Mancilla[1,2], Eric J. Kim[3], Sydney K. Hart[1,2], David A. Knowles[1,3,4,5] ✉ & Neville E. Sanjana[1,2] ✉

Transcriptome engineering applications in living cells with RNA-targeting CRISPR effectors depend on accurate prediction of on-target activity and off-target avoidance. Here we design and test ~200,000 *Rfx*Cas13d guide RNAs targeting essential genes in human cells with systematically designed mismatches and insertions and deletions (indels). We find that mismatches and indels have a position- and context-dependent impact on Cas13d activity, and mismatches that result in G–U wobble pairings are better tolerated than other single-base mismatches. Using this large-scale dataset, we train a convolutional neural network that we term targeted inhibition of gene expression via gRNA design (TIGER) to predict efficacy from guide sequence and context. TIGER outperforms the existing models at predicting on-target and off-target activity on our dataset and published datasets. We show that TIGER scoring combined with specific mismatches yields the first general framework to modulate transcript expression, enabling the use of RNA-targeting CRISPRs to precisely control gene dosage.

Programmable RNA-guided, RNA-targeting type VI clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated proteins (Cas; Cas13) enable direct manipulation of cellular RNAs with high precision compared to previous RNA-targeting technologies[1–5]. A growing number of RNA-engineering technologies have been developed using nuclease active Cas13 or inactive dCas13 effector proteins[6]. These methods critically rely on the ability of Cas13 to distinguish between binding sites in target RNAs and closely related secondary (off-target) binding sites based on the complementarity between guide RNA (gRNA) sequence and bound RNA sequence. In general, the goal is to maximize on-target gRNA activity while minimizing off-target effects. While progress has been made in understanding Cas13 gRNA design rules for nuclease activation and on-target activity[3,7–11], relatively less is known about Cas13 off-target binding and activation. Our understanding is currently limited to Cas13a that has been used in diagnostics[10,11] with only a few studies of off-target activity for Cas13d systems[7,9], which are more commonly used for in vivo perturbation. Several recent examples use Cas13 effectors in preclinical model systems[12–17], emphasizing the need for high precision for any potential human therapeutics.

Understanding the determinants of gRNA activity can not only improve target specificity but also enable controlled/precise attenuation of gene dosage. Notably, biological systems often rely on relative gene dosage as opposed to a binary on-off state. Dose-dependent gene expression is crucial to maintain balanced stoichiometry between members of multiprotein complexes[18] or during embryonic development[19], and underlies X-chromosome inactivation[20]. Moreover, somatic copy number variation and subsequent gene amplification

[1]New York Genome Center, New York City, NY, USA. [2]Department of Biology, New York University, New York City, NY, USA. [3]Department of Computer Science, Columbia University, New York City, NY, USA. [4]Data Science Institute, Columbia University, New York City, NY, USA. [5]Department of Systems Biology, Columbia University, New York City, NY, USA. [6]These authors contributed equally: Hans-Hermann Wessels, Andrew Stirn.
✉e-mail: daknowles@nygenome.org; nsanjana@nygenome.org

have been associated with cancer[21] and a large number of human genetic diseases[22].

Precise modulation of gene expression in mammalian systems can be achieved in multiple ways. For example, synthetic promoter sequences[23] or tetracycline-dependent promoter constructs[24] can be used to modulate gene expression. Similarly, the insertion of *cis*-regulatory elements such as miRNA binding sites in the 3′UTRs of endogenous genes renders them susceptible to the recruitment of the endogenous RNA surveillance and silencing machinery[25]. Such approaches, however, require a considerable amount of engineering on an individual target basis. In contrast, programmable nuclease-null (dCas9) CRISPR systems provide a flexible and scalable alternative for systematic titration of gene expression[26]. One caveat is that epigenetic effector domains commonly fused to dCas9 (for example, KRAB domain) may act more in a switch-like fashion[27].

Here we set out to comprehensively investigate the effects of closely related gRNA variants on target knockdown and predict the on-target and off-target activity of Cas13d in human cells. We are able to achieve strong performance predicting on-target efficacy with a deep learning prediction model trained on both perfect match (PM) and mismatched gRNAs. Leveraging the model's insights into target specificity and activity, we propose and validate a new RNA-targeting CRISPR-based method for titration of gene dosage.

Results

*Rfx*Cas13d screens for perfect match and variant guide RNAs

To systematically assess the efficacy of *Rfx*Cas13d gRNAs, we designed ~120,000 Cas13d gRNAs with a diverse set of mismatches and indels to target known essential genes (Fig. 1a,b and Supplementary Data 1 and 2). This gRNA pool contains 10,000 PM gRNAs for 16 genes and 108,600 gRNAs with designed mismatches for 600 PM gRNAs for six of these genes. In this manner, we can compare how each engineered gRNA mutation impacts Cas13d activity related to its cognate PM gRNA. We designed the mismatch gRNAs to contain 1, 2 or 3 nucleotide mismatches and the indel gRNAs to contain 1 or 2 nucleotide indels. For both mismatches and indels, we designed separate groups of gRNAs with adjacent placement of mismatches/indels or random spacing of the mismatches/indels.

Targeting known essential genes in human cells, we performed cellular fitness (dropout) screens with the expectation that cells will drop out of the population over time depending on the relative gRNA activity and a corresponding degree of essential gene depletion (Fig. 1c). We lentivirally transduced a library of 120,000 crRNAs into a monoclonal HEK293FT cell line with doxycycline-inducible nuclear-localized *Rfx*Cas13d nuclease[7]. We found that gRNA counts and gRNA depletion (fold change (FC) relative to the gRNA abundance at an early time point) were highly reproducible between replicate screens, time points and gRNA categories (Fig. 1d, Supplementary Fig. 1a–d and Supplementary Data 3 and 4).

We first validated the performance of our previous random forest on-target model (RF_{on})[7] for PM gRNAs. For PM gRNAs, the 23 nt of the spacer region contains the reverse complement of the intended RNA target site. Specifically, 70.1% of predicted most-active quartile (Q4) gRNAs depleted more strongly than expected based on a false-discovery rate (FDR) calculated using the nontargeting (NT; negative control) gRNAs (FDR < 0.01; Fig. 1e and Supplementary Fig. 1e). This fraction of active Q4 gRNAs ranged from ~95% (104 of 110 gRNAs) for EIF3B to ~49% for NUP133 (53 of 109 gRNAs; Fig. 1f and Supplementary Fig. 1f). We found that active gRNAs were distributed in clusters along the target transcript sequence, yielding significant similarities between neighboring gRNA efficacies (autocorrelations of $r = 0.13–0.60$; Fig. 1g,h), in agreement with our previous study (CD46 $r = 0.66$; CD55 $r = 0.65$ and CD71 $r = 0.4$)[7].

Indels are more deleterious than substitutions in guide RNAs

For 600 PM gRNAs predicted to have high activity by RF_{on} (quartile Q3 or Q4), we designed 108,600 gRNA variants (18,100 per gene). These variant gRNAs include 83,400 gRNAs with single, double or triple base substitutions. We also included 25,200 gRNAs containing single or double indels (Fig. 1a,b). We found 66.1% of PM gRNAs to be active ($\log_2(FC) < -0.5$, FDR < 0.01; Fig. 2a; Methods). Accumulating base substitutions gradually decreased gRNA efficacy from single mismatches (SMs, 34.4% active gRNAs) to random triple mismatches (RTMs, 3.3% active gRNAs; Fig. 2b). Overall, base substitutions were better tolerated than indel variants such as single nucleotide deletions (SD) or single nucleotide insertions (SI) within the gRNA sequence (guides active—SM, 34.4% > SD and 20.7% > SI 17.9%).

Next, we calculated the relative activity for all gRNA variants relative to their cognate PM gRNAs (Methods). Most SM variants resulted in modest decreases in activity compared to the cognate PM gRNA (Fig. 2c). In contrast, single indels resulted in a greater loss of activity compared to SM variants with the greatest loss of activity for insertions in the central region of the gRNA (Fig. 2d,e). Unlike base substitutions, indels introduce bulges on either gRNA or target sides, respectively. Because the gRNA is embedded within the Cas13d enzyme[28], gRNA positioning is likely more constrained than target sequence arrangement within the ternary Cas13–gRNA–target complex. This may explain why RNA bulges on the gRNA side (as introduced by nucleotide insertions) are the most disruptive.

We also confirmed the presence of the SM-intolerant seed sequence centered on guide nucleotide positions 18 (ref. 7; Fig. 2f). Single-base substitutions outside the seed region led to a milder attenuation of relative gRNA efficacy for all SM gRNAs. We noted that A-to-G and to a lesser extent C-to-U substitutions within the gRNA had a milder effect than other substitutions, including in the seed region (Fig. 2f). A-to-G substitutions lead to G–U wobble pairing with the G on the gRNA side, while C-to-U substitutions have the G in the target site. We found that for all mismatch types, the contribution of G–U wobbles ameliorated the relative decrease in efficacy from the mismatch (Fig. 2g and Supplementary Fig. 1h).

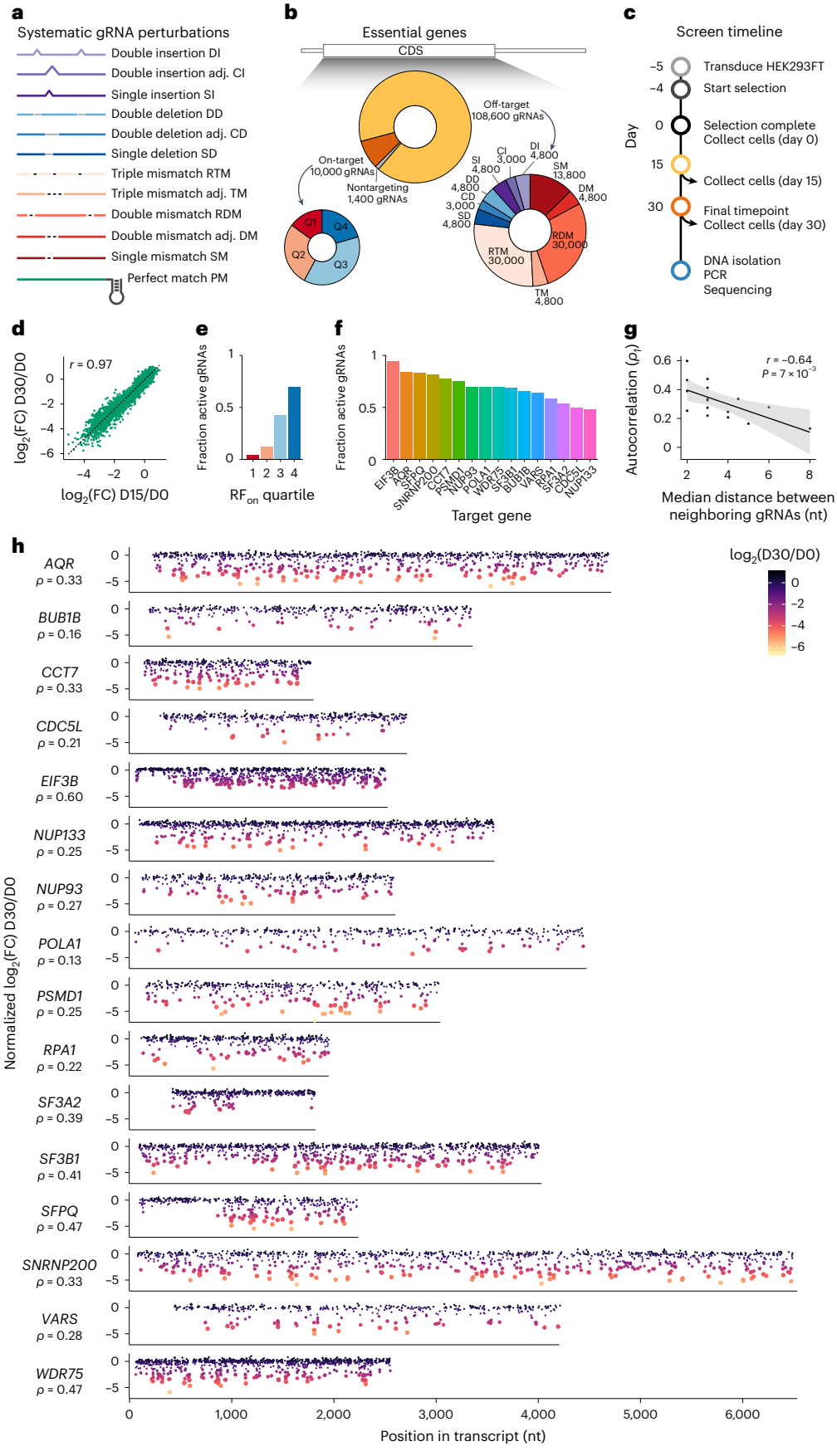
A deep learning model to predict guide RNA efficacy

Existing approaches to model Cas13d efficacy[7–9] predict knockdown efficiencies only for gRNAs perfectly matching their target site. Although large-scale off-target screens have empowered predictive modeling for DNA-targeting CRISPRs like Cas9 (refs. 26,29–36) and Cas13a-mediated RNA diagnostics in vitro[10], there have not been systematic efforts to learn an off-target model for RNA-targeting CRISPRs in vivo. Toward this end, we adopted a convolutional neural network

Fig. 1 | Pooled CRISPR–Cas13 essentiality screen assaying Cas13d gRNA efficacy. a, Design of pooled CRISPR–Cas13d screen for mapping gRNA variants with mismatch and indel changes to PM gRNAs. **b**, Composition of gRNA library containing 120,000 perfectly matching and mismatched gRNA sequences targeting the coding region of essential genes. **c**, Abundance of individual gRNAs was measured in TetO-*Rfx*Cas13d-HEK293FT cells over 30 d ($n = 3$ independent transduction replicates). **d**, The Pearson correlation of gRNA abundance as $\log_2(FC)$ on day 15 and day 30 relative to the day 0 input representation showing PM gRNAs as a mean of three replicates ($n = 13,782$). **e**, Fraction of active gRNAs ($\log_2(FC) < -0.5$) for PM gRNAs separated by RF_{on} quartile predictions. **f**, Fraction of active ($\log_2(FC) < -0.5$) predicted quartile 4 (Q4) PM gRNAs for all 16 essential gene targets. **g**, Relationship between median distance between neighboring PM gRNAs and autocorrelation of $\log_2(FC)$ at lag = 1 ($n = 16$ target gene transcripts). Line indicates linear regression and 95% confidence interval with the Pearson correlation (r) and P value (two-sided t-test). **h**, Distribution of PM gRNAs along the coding region of the 16 target gene transcripts and their $\log_2(FC)$ enrichments. Negative $\log_2(FC)$ values indicate better transcript knockdown. ρ, autocorrelation of $\log_2(FC)$ with lag = 1.

https://doi.org/10.1038/s41587-023-01830-8

(CNN) architecture similar to those pioneered in computer vision[37,38] to predict a guide's efficacy from a one-hot encoding of its sequence. Similar to a previous CNN for Cas9 off-target prediction[26], our model has two convolution layers followed by a max-pooling layer and then interleaves three dropout and dense layers for a total of two hidden layers plus an output layer.



https://doi.org/10.1038/s41587-023-01830-8

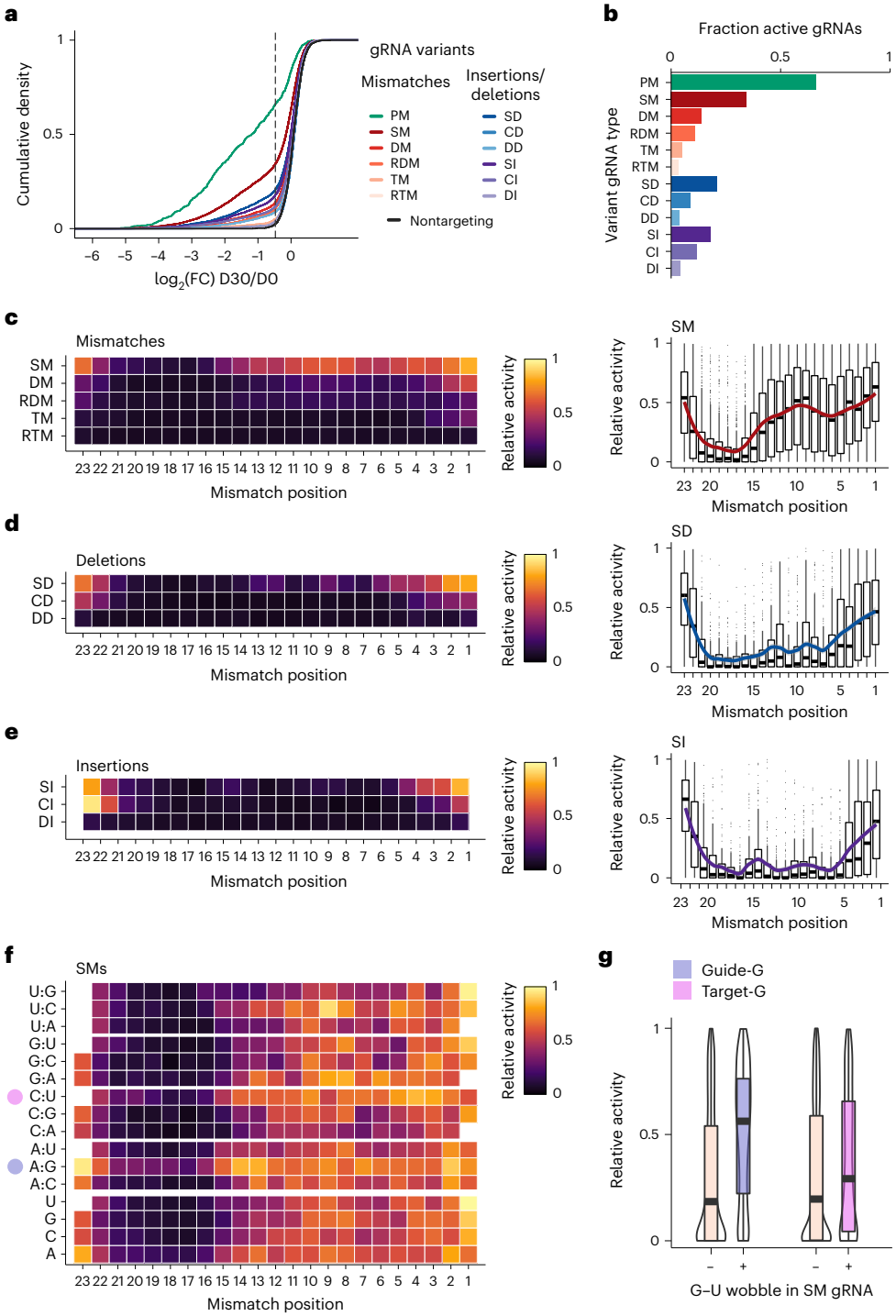


Fig. 2 | Large-scale mapping of Cas13d gRNA mismatch activity. a, Empirical cumulative distribution of gRNA depletion for all gRNAs by introduced mutation type. Vertical line indicates the cutoff for active gRNAs ($\log_2$(FC) < −0.5). **b**, Fraction of active gRNAs for gRNAs as shown in **a**. **c**–**f**, Relative targeting activity (fraction of parental PM gRNA $\log_2$(FC)) of gRNAs for all PM gRNA derivatives with mismatches or indels at the indicated position relative to their cognate PM gRNAs (n = 388 reference PM gRNAs with $\log_2$(FC) < −0.5. Each cell indicates the mean of gRNAs). **c**, Left: heatmap depicting all mismatch types. Right: boxplot highlighting the full distribution of relative gRNA activity for SM gRNAs. **d**, Left: heatmap depicting all deletion types. Right: boxplot highlighting the SD gRNAs. **e**, Left: heatmap depicting all insertion types. Right: boxplot highlighting the SI gRNAs. **f**, Detailed representation of relative activity for SM gRNA separated by reference guide nucleotide (bottom) or substitution identity (top) for each mismatch position. **g**, Relative targeting efficacy for gRNAs containing 0 or 1 G–U wobble base pairs compared to unpaired mismatches for all mismatched gRNA types (n = 388 reference PM gRNAs with $\log_2$(FC) < −0.5). All gRNA abbreviations are defined in Fig. 1a. For boxplots in **c**–**e** and **g**, the boxes indicate the median and interquartile range (IQR) with whiskers indicating 1.5× IQR.

Our model has the following two architectural augmentations beyond those used in a previous study: additional sequence context flanking the 23 nt target site and the flexibility to input a vector of nonsequence features at our first dense layer. We considered six groups of nonsequence features as follows: (1) crRNA folding minimum free energy (MFE), (2) the RNA–RNA hybridization MFE between spacer and

https://doi.org/10.1038/s41587-023-01830-8

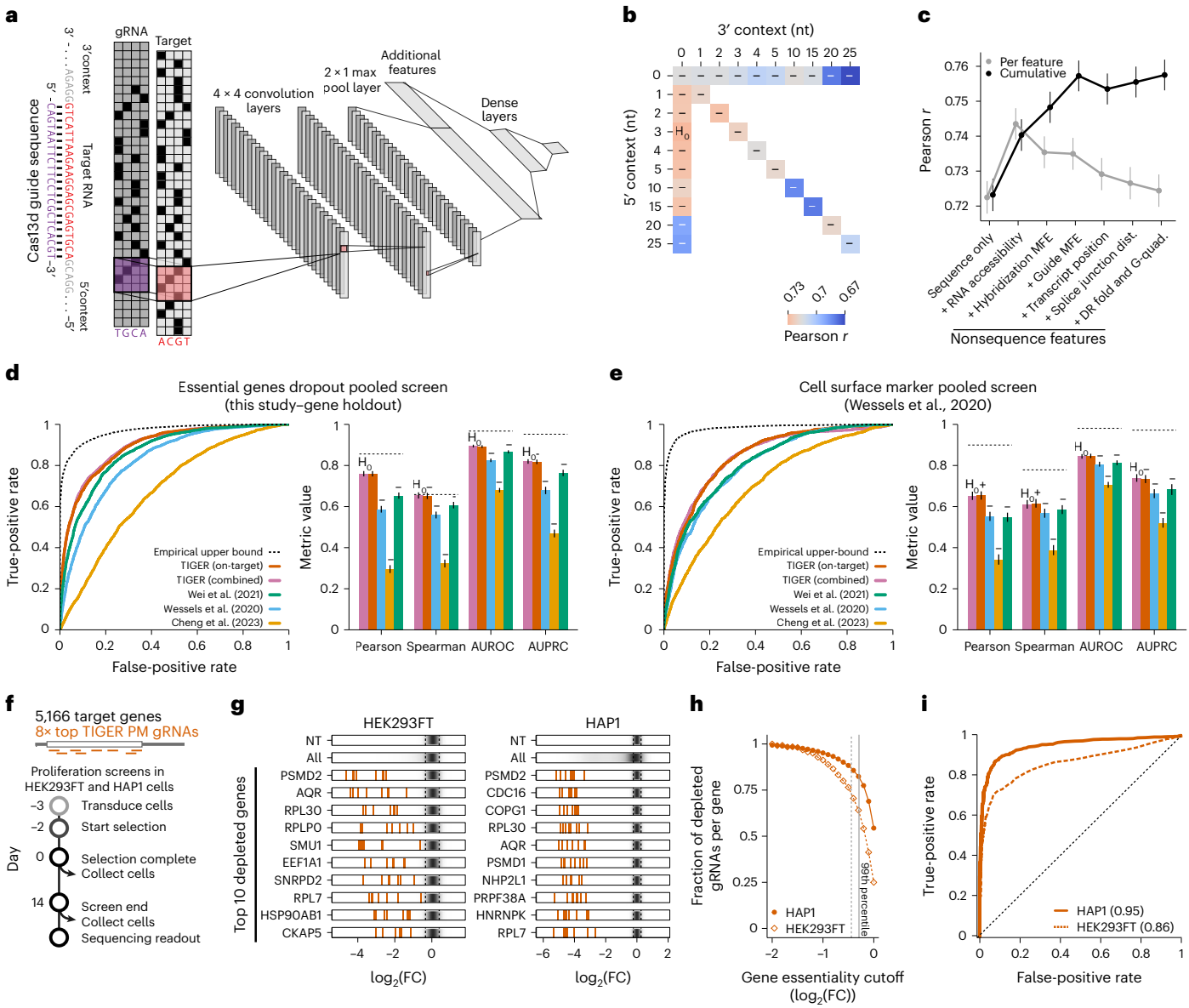


Fig. 3 | A deep learning model to predict optimal Cas13d gRNAs. a, TIGER combines one-hot-encoded guide and target sequences for sequence input, following an *AlexNet* architecture but allowing for nonsequence features as inputs to the first dense layer. **b**, Correlation of predictions with additional sequence context (5′ only, 3′ only and combined 5′ and 3′) to the 23nt gRNA target site (tenfold CV randomized at the target site level) using a sequence-only model. H_0 denotes the best-performing condition and all differences between other conditions and H_0 are significant ($P < 0.05$, Steiger's test[53]). **c**, The effect of including different feature groups (individually and cumulatively) on the correlation of predictions aggregated from the heldout target sites ($n = 10$ random folds). We present feature groups in descending order of increased correlation (individually). **d**, ROC curve and other performance metrics for predictions aggregated from the heldout genes ($n = 16$) of the survival screen of essential genes. **e**, ROC curve and other performance metrics for all gRNAs from a previously published screen using flow cytometry of cell surface proteins[7]. We employ a Steiger's test[53] for the Pearson and Spearman comparisons, DeLong's test[54,55] for AUROC comparisons and a bootstrapped Kolmogorov–Smirnov test[56] for AUPRC comparisons (**d**,**e**). Values denote aggregate performance over CV folds and error bars denote ±2 s.e. **f**, Design of pooled CRISPR–Cas13d screen targeting 5,166 genes with eight high-efficacy gRNAs from $\text{TIGER}_{\text{combined}}$ predictions. **g**, The top ten most depleted genes show consistent depletion in each cell line ($n = 8$ gRNAs per target gene). Dotted lines (black) indicate the 1st and 99th percentiles for NT gRNA distribution. **h**, Fraction of active gRNAs (more depleted than the 99th percentile of the NT gRNAs) as a function of gene depletion. Gray lines indicate 99th percentile of NT gRNA distribution for HAP1 (solid) and HEK293FT (dashed) cells. **i**, ROC curve for DepMap essential genes for screens depicted in **f**–**h** ($n = 1{,}082$ essential genes, $n = 458$ nonessential genes). Numbers in parenthesis next to each cell line name indicate AUROC.

target site (multiple positions), (3) target accessibility (that is, a lack of predicted secondary structure in multiple windows), (4) the target site's proximity to an exon–exon junction (5′ distance, 3′ distance), (5) the target site's location within the transcript (relative position in CDS) and 6) binary features related to the gRNA's secondary structure (folded repeat and G-quadruplex). We termed this deep learning approach as Targeted Inhibition of Gene Expression via gRNA design (TIGER; Fig. 3a).

We first sought to determine the optimal flanking target sequence context when using only nucleotide sequence or all features (Fig. 3b and Supplementary Fig. 2a). We added additional context to just the 5′-end, just the 3′-end and equally to both ends and predicted gRNA efficacy using tenfold cross-validation (CV) over target sites (Supplementary Note). Consistent with findings from another recent study[8], we found that additional 5′ target site context of three nucleotides

was optimal for the sequence-only model. However, the impact of this additional context is reduced when including nonsequence features, which capture 5′-end target accessibility and likely make the extra 5′ context redundant. When analyzing the impact of each nonsequence feature in TIGER, we find that target RNA accessibility yields the greatest increase in performance (Fig. 3c).

Although our CNN is a regression model that predicts FC in guide abundance, we can also threshold our predictions to classify active guides versus inactive ones. We again identified active gRNAs based on an empirical FDR calculated from the NT gRNAs (FDR < 0.01). To best assess generalization across the transcriptome, we aggregated predictions across gene-level CV (holding all gRNAs for each of the 16 genes in turn) to compute correlations between predictions and observations (Pearson and Spearman correlations), area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC; Fig. 3d and Supplementary Fig. 2b). For each of these metrics, we estimated an upper bound by taking advantage of the three biological replicates of our Cas13d essential gene screen, quantifying the error when using a heldout replicate to predict the mean of the other two replicates. Across all four evaluation metrics, we find that TIGER trained on PM gRNAs in combination with mismatched gRNAs ($TIGER_{combined}$) outperforms or matches TIGER trained on PM alone ($TIGER_{on\text{-}target}$) and yields superior performance compared to our previously published model (RF_{on})[7] and two recent deep learning models[8,9]. The gene-level CV (all gRNAs for a target gene are held out from the training set) ensures that, during training, TIGER does not have access to any data related to the particular gene used for evaluation.

In addition to holding out all gRNAs for individual genes, we repeated these experiments with target site-level or gRNA-level CV (Supplementary Fig. 2c–h). Holding out target sites ensures that a PM gRNA's mismatched variants will not be used in training (for example, active SM gRNAs are highly informative of their PM parents; Supplementary Note). Under these CV strategies, the on-target and combined models observe a boost in performance relative to gene-level CV. Here too, the combined model outperforms all other considered models. Both TIGER models perform close to this estimated upper bound for the Spearman correlation, suggesting that they are adept at filtering variation due to technical noise.

In addition to the survival screen with essential gene-targeting gRNAs, we further evaluated model performance using a separate gRNA dataset with a different phenotypic selection (flow cytometry of cell surface proteins[7], which has been used by other groups to assess generalization performance[8,9]. Specifically, we use PM gRNAs from our survival screen as training data (PM and mismatched gRNAs in case of $TIGER_{combined}$), holding out PM gRNAs from the cell surface protein screen as test data[7]. To compare with our previous RF_{on} model, we retrained it solely on the larger essentiality screen dataset (Fig. 1a). Notably, this validation dataset contains no genes or target sites in common with our survival screen. When testing generalization performance on this validation dataset, we find that $TIGER_{combined}$ and $TIGER_{on\text{-}target}$ yield best-in-class predictions (Fig. 3e and Supplementary Fig. 2i). To further benchmark TIGER, we also trained a linear regression model and a recurrent neural network with Bidirectional Gated Recurrent Unit (BiGRU) model and, across both datasets, we found that TIGER was superior to linear regression and comparable to the recurrent neural network (Supplementary Fig. 3). Taken together, we find that TIGER's predictions generalize over different screen modalities (cell proliferation and surface marker expression) and target genes (essential and nonessential).

Feature importance by Shapley additive explanations

To determine TIGER's learned gRNA design rules, we performed a ten-fold CV of our combined model ($TIGER_{combined}$) with target site-level CV. For each holdout, we collected Shapley additive explanations (SHAP)[39] values for sequence and nonsequence features. For sequence features of PM gRNAs, we observed a strong contribution of G and C nucleotides in the seed region (nucleotides 15–21) of the gRNA (Supplementary Fig. 4a,b), reflecting the local importance of G and C nucleotides[7,8]. Similarly, we examined the Pearson correlations and SHAP values for the gRNA substitutions (Supplementary Fig. 4c,d). We found that the CNN model correctly learned the increased importance of SMs in the seed region including the differential contribution of G–U mismatched base pairing. We also examined SHAP values for nonsequence features for PM guides alone or all PM plus mismatched guides (Supplementary Fig. 4e,f). Among nonsequence features, we found that RNA–RNA hybridization (as in target site accessibility, gRNA-RNA hybridization MFE and crRNA folding MFE) had the largest contributions to model predictions, consistent with our earlier findings (Fig. 3c).

TIGER consistently predicts highly active gRNAs

Next, we sought to test the generalizability of our $TIGER_{combined}$ model at scale. We predicted eight high-efficacy gRNAs for 5,166 genes and performed proliferation screens in two different cell lines (HEK293FT and HAP1; Fig. 3f and Supplementary Data 5–8). We noticed high consistency between gRNAs that target the same gene (Fig. 3g). Extending this analysis, we found that $TIGER_{combined}$ correctly predicted active gRNAs (defined as $\log_2(FC) < -1$) in the HEK293 and HAP1 screens for 91% and 95%, respectively, of the chosen gRNAs (Fig. 3h), highlighting the robustness and generalizability of our on-target model across cell lines and thousands of unseen genes.

Having confirmed robust targeting, we next assessed whether RNA-targeting *Rfx*Cas13d CRISPR screens could discriminate essential genes from nonessential genes. Among the 5,166 target genes included in these screens, we embedded a set of 1,082 common essential genes and 458 nonessential genes based on DepMap classifications (Methods). We found that we could successfully discriminate essential genes against control genes using gRNA efficacy predictions using the $TIGER_{combined}$ model (Fig. 3i, AUROC 0.86 and 0.95 in HEK293FT and HAP1 cells). Gene depletion in HAP1 cells was generally more pronounced compared to that in HEK293FT cells (Supplementary Fig. 5a).

Previous reports have suggested that cell fitness (and therefore cell proliferation) may be affected by nonspecific collateral *Rfx*Cas13d activity as a function of target gene expression[40,41]. However, it is unclear if the collateral activity affects cell fitness in a controlled setting with single integration of *Rfx*Cas13d effector protein and gRNA expression cassettes. To explore this relationship, we analyzed gene dropout across a large span of expression levels (1 to >1,000 transcripts per million (TPM)). For common essential genes, we also observed a known dependency of gene expression and gene depletion[42,43], which was much reduced for the nonessential gene group (Supplementary Fig. 5b). For nonessential genes, we noticed no evidence of dropout as a function of gene expression in HEK293FT. For the haploid HAP1 cells, we noticed that a relationship between gene expression and control gene depletion was mainly driven by genes expressed >100 TPM. It is worth noting that only ~5% of 19,177 protein-coding genes are expressed at levels above 100 TPM in human cell lines ($n = 1{,}377$ Cancer Cell Line Encyclopedia cell lines), and the majority of these genes are classified as common essential genes in DepMap. Nonspecific collateral activity may only affect a small subset (<6%) of very highly expressed genes (>100 TPM) in sensitive cells with single-integration Cas13d screens. Despite this, we found that HAP1 cells showed higher sensitivity (AUROC) for the identification of essential genes (Fig. 3i).

Taken together, we find that $TIGER_{combined}$ gRNA predictions are robust and generalizable across thousands of genes in multiple cell lines and that undesired viability defects based on target gene expression may be cell-type-dependent and usually only occur for a subset of highly expressed genes.

https://doi.org/10.1038/s41587-023-01830-8

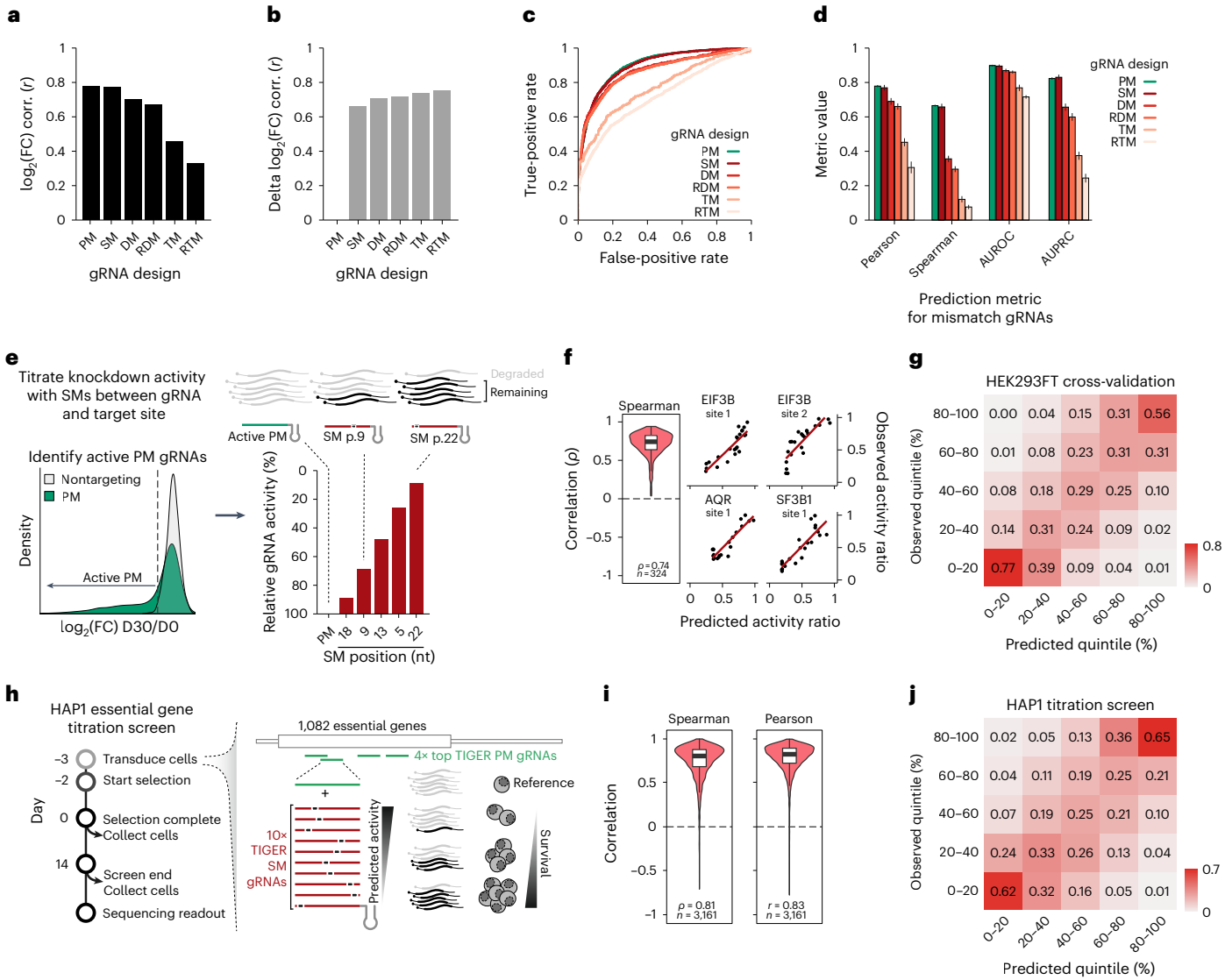


Fig. 4 | Training TIGER using gRNAs with mismatches enables prediction of off-target activity and transcript modulation using gRNAs with SMs. a, The correlation between observed and TIGER-predicted gRNA abundance by gRNA design type. **b**, The correlation between change in observed and TIGER-predicted gRNA abundance by gRNA design type. The change in gRNA abundance is defined as the difference in $\log_2$(FC) for a particular gRNA with mismatches and its cognate PM gRNA. **c**, ROC curves for each gRNA design type from tenfold target-site CV. **d**, Aggregate correlation (Pearson and Spearman) and aggregated areas under the ROC and precision–recall curves for each gRNA design type from tenfold target-site CV. Values denote aggregate performance over CV folds and error bars denote ±2 s.e. **e**, A framework for using gRNA with SMs to modulate Cas13 targeting activity. **f**, Correlation of predicted and observed relative activity ratio for all 23 SM gRNAs per highly active ($\log_2$(FC) < −1) target sites (n = 324 target sites). **g**, Confusion matrix for efficiency ratios between a gRNA with an SM to the intended target RNA binned by quintiles for all active target sites (FDR < 0.01; n = 393 target sites with n = 9,032 SM gRNA variants). Each column is normalized. **h**, Design of pooled CRISPR–Cas13d screen for $\text{TIGER}_{\text{combined}}$ gRNA predictions targeting 1,082 common essential genes with four high-efficacy PM gRNAs, and ten SM gRNAs with varying relative activity. **i**, Correlation of predicted and observed relative activity ratio for all ten SM gRNAs per highly active PM gRNA (n = 3,161 gRNAs with $\log_2$(FC) < −1). **j**, Confusion matrix for efficiency ratios between a gRNA with an SM to the intended target RNA binned by quintiles (n = 30,582 SM gRNA variants). Each column is normalized. For boxplots in **f** and **i**, the boxes indicate the median and IQRs with whiskers indicating 1.5× IQR.

Off-target prediction and gene knockdown titration

Although multiple groups have developed predictive models of on-target Cas13d activity, there has been comparatively less work on off-target activity and no predictive models exist for Cas13d. Similarly, for CRISPR-Cas9 gRNAs, nearly all deep learning approaches focus on predicting on-activity and do not include separate inputs for gRNA and target site sequences[26,29–36,44–46]. TIGER's architecture easily accommodates mismatches between target and gRNA (Fig. 3a). In addition to the ability to predict on-target efficacy for PM gRNA when trained on PM and mismatched data ($\text{TIGER}_{\text{combined}}$), we sought to extend the usability of TIGER to enable precise off-target predictions via target site-level CV (Supplementary Note). This validation strategy avoids PM and SM guides for the same target appearing in training and validation.

When predicting changes in abundance for mismatch variant gRNAs, we find that the correlation between TIGER's predictions and observed values decreases as the number of mismatches and the distance between them increases (Fig. 4a and Supplementary Fig. 6a). We wondered whether this might be due in part to the variability in effect size between different PM gRNAs (Fig. 2d,e). To test this, we instead measured the difference in the predicted gRNA abundance between

variant gRNAs and their cognate PM gRNA (Fig. 4b and Supplementary Fig. 6b). Here we find that the correlation no longer decreases with increasing mismatches, suggesting that when we explicitly account for the variability in PM gRNAs, TIGER is able to predict the effect of different mismatches. As we did for PM gRNAs, we computed four performance metrics for each category of mismatch variant gRNAs (Fig. 4c,d and Supplementary Fig. 6c). This time, however, we compared our $TIGER_{combined}$ model (trained on PM and mismatches) to a $TIGER_{off\text{-}target}$ model (trained on mismatched gRNAs only). As is the case with on-target prediction, we find that the combined model is superior (Supplementary Fig. 6d). Using an independent pooled screen (surface protein expression)[7], which also contains gRNAs with SM and double mismatch (DM) variants, we further confirmed our model's ability to predict gRNA efficacy after being trained on the survival screen dataset (Supplementary Fig. 6e).

Given that TIGER can predict a PM gRNA's efficacy and how this efficacy changes when mismatches are introduced, we can both identify mismatched target sites with off-target activity and engineer mismatches to precisely reduce gRNA efficacy and titer knockdown (Fig. 4e). To this end, we defined the 'efficacy ratio' as the ratio of the FC of an SM gRNA to the FC of its PM cognate gRNA. Using target site-level CV, we compared predicted and observed relative gRNA activity for the 23 SM gRNA variants designed for each individual target site. We found a high correlation between predicted and observed relative activities (median, $r = 0.76$; median, $\rho = 0.74$, $n = 324$ active target sites with PM $\log_2(FC) < -1$; Fig. 4f). We binned all observed and predicted efficacy ratios for SM gRNAs with active cognate PM gRNAs into quintiles to compute a confusion matrix (Fig. 4g and Supplementary Data 9**)**. For each quintile, we found that SM gRNAs were most often correctly classified. In particular, TIGER achieves the best performance at the extremes (0–20 and 80–100) and can determine those SMs with high accuracy that minimally impact or maximally disrupt activity.

Finally, we sought to test the generalizability of our TIGER off-target model for gene essentiality titration across thousands of genes and target sites. Specifically, we predicted four high-efficacy gRNAs for 1,082 common essential genes and designed ten gRNA variants with single nucleotide mismatches for each PM gRNA target site ($n = 47{,}608$ PM and SM gRNAs). Using a pooled proliferation screen in a different cell line (HAP1), we measured relative activity loss compared to the cognate PM gRNA via depletion (Fig. 4h and Supplementary Data 10–12). We found a high correlation between predicted and observed relative activities (median, $r = 0.83$; median, $\rho = 0.81$, $n = 3{,}161$ target sites with cognate PM $\log_2(FC) < -1$; Fig. 4i, Supplementary Fig. 6f and Supplementary Data 13). In this independent screen, we found strong agreement across quintiles (Fig. 4j). This suggests that our model is able to predict gRNA variants with a defined relative activity with high accuracy for unseen target sites and generalizes across cell lines.

Discussion

In this study, we generated a large Cas13d dataset that measures the activity of ~200,000 gRNAs across multiple human cell lines and performed a comprehensive assessment of Cas13d gRNA on-target and off-target activity. Specifically, we sought to characterize PM gRNA activity determinants and gRNAs permutations across a large set of nucleotide mismatches and indels relative to their cognate target sites. We found that a gRNA's ability to trigger Cas13d nuclease activity depends on the permutation position within the gRNA, the nucleotide identity and the target site context. Previous studies have not characterized certain gRNA permutations such as indels. Our analysis shows that mismatches are generally better tolerated compared to more disruptive indels in gRNA or target RNA sequences. Using this unique dataset, we trained the TIGER CNN model for on-target activity and off-target activity. We find that TIGER has strong performance for Cas13d on-target activity compared to existing Cas13d on-target models including those with larger training sets. Of relevance for understanding impacts across the transcriptome, our TIGER model is a compelling attempt to understand and model Cas13d off-target binding and nuclease activation. Finally, we apply our TIGER platform to develop an approach for precise and massively parallel interrogation of gene dosage.

New CRISPR technologies hold great promise for a new generation of therapeutic agents. Among these, RNA-targeting CRISPR proteins have recently been shown to provide therapeutic values in disease models[12–17]. High precision is key to the safety of therapeutic RNA-targeting CRISPR agents. We believe that TIGER predictions will enable ranking and ultimately avoidance of undesired off-target binding sites and nuclease activation, and further spur the development of RNA-targeting therapeutics. The ability to distinguish between closely related target sites may enable the targeting of allelic variants and other nearly undruggable targets like fusion gene products[47].

Furthermore, our model can be used for precise modulation of target RNA knockdown at scale. Specifically, our study suggests that RNA-targeting CRISPR perturbations can be used to systematically study the effect of gene dosage at the RNA levels. This platform fundamentally extends on previous microRNA-based platforms[25] that on the one hand, a lack of scalability due to laborious target site engineering and, on the other hand, a lack of target-specificity if engineered microRNAs are provided exogenously due to their short target site recognition sequence. In addition, tuning of gene expression at the RNA level may be beneficial compared to modulation at the DNA level, as gene expression initiation is inherently stochastic[48] and biological systems have evolved in a way to fine-tune gene expression post-transcriptionally[49,50]. Other DNA-targeting (for example, dCas9-based) CRISPR approaches have been proposed for gene expression modulation[26]. However, it is unclear if epigenetic effector domains (for example, KRAB) fused dCas9 proteins are well suited as they may act more in a binary on-off fashion[27,51], and may lack precision for closely spaced genes due to the spreading of chromatin modifications and DNA methylation[52]. Here we provide precise measurements for the titration of 1,082 essential genes across thousands of target sites confirming the TIGER model's prediction accuracy. While we observe a high degree of concordance between measured essentiality and predicted gRNA efficacy, our methods, similar to previous CRISPR-based methods[26], may be limited by the assumption that RNA knockdown and gene essentiality scale linearly. Therefore, future experiments are needed to specifically evaluate differences between genes with a linear relationship between expression and essentiality and those that require threshold expression levels.

Taken together, we believe that the ability to model the effect of nucleotide mismatches not only allows for an enhanced understanding of gRNA on-target specificity and off-target avoidance but also enables precise target knockdown to a defined degree that will be useful for diverse transcriptome engineering applications.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-01830-8.

References

1. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
2. Abudayyeh, O. O. et al. RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).
3. Smargon, A. A. et al. Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Mol. Cell* **65**, 618–630 (2017).

4. Konermann, S. et al. Transcriptome engineering with RNA-targeting article transcriptome engineering with RNA-targeting. *Cell* **173**, 1–12 (2018).
5. Yan, W. X. et al. Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Mol. Cell* **70**, 327–339 (2018).
6. Smargon, A. A., Shi, Y. J. & Yeo, G. W. RNA-targeting CRISPR systems from metagenomic discovery to transcriptomic engineering. *Nat. Cell Biol.* **22**, 143–150 (2020).
7. Wessels, H. H. et al. Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat. Biotechnol.* **38**, 722–727 (2020).
8. Wei, J. et al. Deep learning and CRISPR–Cas13d ortholog discovery for optimized RNA targeting. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.14.460134 (2022).
9. Cheng, X. et al. Modeling CRISPR–Cas13d on-target and off-target effects using machine learning approaches. *Nat. Commun.* **14**, 752 (2023).
10. Metsky, H. C. et al. Designing sensitive viral diagnostics with machine learning. *Nat. Biotechnol.* **40**, 1123–1131 (2022).
11. Tambe, A., East-seletsky, A., Knott, G. J., Connell, M. R. O. & Doudna, J. A. RNA binding and HEPN-nuclease activation are decoupled in CRISPR–Cas13a. *Cell Rep.* **24**, 1025–1036 (2018).
12. Powell, J. E. et al. Targeted gene silencing in the nervous system with CRISPR–Cas13. *Sci. Adv.* **8**, eabk2485 (2022).
13. Morelli, K. H. et al. An RNA-targeting CRISPR–Cas13d system alleviates disease-related phenotypes in Huntington's disease models. *Nat. Neurosci.* **26**, 27–38 (2023).
14. Méndez-Mancilla, A. et al. Chemically modified guide RNAs enhance CRISPR–Cas13 knockdown in human cells. *Cell Chem Biol.* **29**, 321–327 (2022).
15. Rotolo, L. et al. Species-agnostic polymeric formulations for inhalable messenger RNA delivery to the lung. *Nat. Mater.* **22**, 369–379 (2023).
16. Fan, N. et al. Hierarchical self-uncloaking CRISPR–Cas13a-customized RNA nanococoons for spatial-controlled genome editing and precise cancer therapy. *Sci. Adv.* **8**, eabn7382 (2022).
17. Guo, Y. et al. Specific knockdown of Htra2 by CRISPR–CasRx prevents acquired sensorineural hearing loss in mice. *Mol. Ther. Nucleic Acids* **28**, 643–655 (2022).
18. Nasim, M. T. et al. Stoichiometric imbalance in the receptor complex contributes to dysfunctional BMPR-II mediated signalling in pulmonary arterial hypertension. *Hum. Mol. Genet.* **17**, 1683–1694 (2008).
19. Gurdon, J. B. & Bourillot, P. Y. Morphogen gradient interpretation. *Nature* **413**, 797–803 (2001).
20. McHugh, C. A. et al. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).
21. Fehrmann, R. S. N. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
22. Collins, R. L. et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041–3055 (2022).
23. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
24. Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl Acad. Sci. USA* **89**, 5547–5551 (1992).
25. Michaels, Y. S. et al. Precise tuning of gene expression levels in mammalian cells. *Nat. Commun.* **10**, 818 (2019).
26. Jost, M. et al. Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.* **38**, 355–364 (2020).
27. Bintu, L. et al. Dynamics of epigenetic regulation at the single-cell level. *Science* **351**, 720–724 (2016).
28. Zhang, C. et al. Structural basis for the RNA-guided ribonuclease activity of CRISPR–Cas13d. *Cell* **175**, 212–223 (2018).
29. Charlier, J., Nadon, R. & Makarenkov, V. Accurate deep learning off-target prediction with novel sgRNA–DNA sequence encoding in CRISPR–Cas9 gene editing. *Bioinformatics* **37**, 2299–2307 (2021).
30. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci. Adv.* **5**, eaax9249 (2019).
31. Lin, J. & Wong, K. C. Off-target predictions in CRISPR–Cas9 gene editing using deep learning. *Bioinformatics* **34**, i656–i663 (2018).
32. Lin, J., Zhang, Z., Zhang, S., Chen, J. & Wong, K. C. CRISPR-Net: a recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels. *Adv. Sci.* **7**, 1903562 (2020).
33. Liu, Q., Cheng, X., Liu, G., Li, B. & Liu, X. Deep learning improves the ability of sgRNA off-target propensity prediction. *BMC Bioinf.* **21**, 51 (2020).
34. Luo, J., Chen, W., Xue, L. & Tang, B. Prediction of activity and specificity of CRISPR-Cpf1 using convolutional deep learning neural networks. *BMC Bioinformatics* **20**, 332 (2019).
35. Niu, R., Peng, J., Zhang, Z. & Shang, X. R-CRISPR: a deep learning network to predict off-target activities with mismatch, insertion and deletion in CRISPR-Cas9 system. *Genes (Basel).* **12**, 1878 (2021).
36. Zhang, G., Zeng, T., Dai, Z. & Dai, X. Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. *Comput. Struct. Biotechnol. J.* **19**, 1445–1457 (2021).
37. LeCun, Y. et al. Backpropagation applied to digit recognition. *Neural Comput.* **1**, 541–551 (1989).
38. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
39. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
40. Shi, P. et al. Collateral activity of the CRISPR/RfxCas13d system in human cells. *Commun. Biol.* **6**, 334 (2023).
41. Kelley, C. P., Haerle, M. C. & Wang, E. T. Negative autoregulation mitigates collateral RNase activity of repeat-targeting CRISPR–Cas13d in mammalian cells. *Cell Rep.* **40**, 111226 (2022).
42. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
43. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733 (2014).
44. Kim, H. K. et al. High-throughput analysis of the activities of xCas9, SpCas9-NG and SpCas9 at matched and mismatched target sequences in human cells. *Nat. Biomed. Eng.* **4**, 111–124 (2020).
45. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
46. Xiang, X. et al. Enhancing CRISPR–Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat. Commun.* **12**, 3238 (2021).
47. Hu, W. et al. Single-base precision design of CRISPR–Cas13b enables systematic silencing of oncogenic fusions. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.22.497105 (2022).
48. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).

49. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of Drosophila microRNA genes. *Genome Biol.* **4**, R42 (2003).
50. Stoeger, T., Battich, N. & Pelkmans, L. Passive noise filtering by cellular compartmentalization. *Cell* **164**, 1151–1161 (2016).
51. Noviello, G., Gjaltema, R.A.F. & Schulz, E.G. CasTuner is a degron and CRISPR/Cas-based toolkit for analog tuning of endogenous gene expression. *Nat. Commun.* **14**, 3225 (2023).
52. Lensch, S. et al. Dynamic spreading of chromatin-mediated gene silencing and reactivation between neighboring genes in single cells. *eLife* **11**, e75115 (2022).
53. Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87**, 245–251 (1980).
54. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
55. Sun, X. & Xu, W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett.* **21**, 1389–1393 (2014).
56. Massey, F. J. J. The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).

Methods

Cell culture

Doxycycline-inducible *Rfx*Cas13d-NLS HEK293FT and HAP1 cells were generated via lentiviral transgenesis (Addgene,138149)[7]. *Rfx*Cas13d-NLS HEK293FT and HAP1 cells were maintained at 37 °C with 5% carbon dioxide in D10 media or I10 media, respectively. Dulbecco's Modified Eagle Medium with high glucose and stabilized L-glutamine (Caisson, DML23) or Iscove's Modified Dulbecco's Medium (Caisson, IML02) supplemented with 10% fetal bovine serum (Serum Plus II; Sigma-Aldrich, 14009C) and 5 μg ml^{-1} Blasticidin S (Thermo Fisher Scientific, A1113903).

Pooled lentiviral production and screening

Lentivirus was produced via transfection of library plasmid pool and appropriate packaging plasmids (psPAX2—Addgene, 12260; pMD2.G—Addgene, 12259) using linear polyethylenimine MW25000 (Polysciences, 23966). We seeded ten million HEK293FT cells per 10 cm dish and transfected with 60 μl polyethylenimine, 9.2 μg plasmid pool, 6.4 μg psPAX2 and 4.4 μg pMD2.G. At 3 d post-transfection, viral supernatant was collected and passed through a 0.45-μm filter and stored at −80 °C until further use.

Doxycycline-inducible *Rfx*Cas13d-NLS HEK293FT and HAP1 cells were transduced with the pooled library lentivirus in separate infection replicates, ensuring at least 1,000× guide representation in the selected cell pool per infection replicate using spinfection. We performed three or four independent infection replicate screens/experiments. After 24 h, cells were selected with 1 μg ml^{-1} puromycin (Thermo Fisher Scientific; A1113803), resulting in ~30% cell survival. Puromycin selection was performed ~48 h to 72 h after the addition of puromycin. Assuming independent infection events (Poisson), we determined that ~83% of surviving cells received a single sgRNA construct[57]. *Rfx*Cas13d expression was induced by the addition of 1 μg ml^{-1} doxycycline (Sigma-Aldrich, D9891) upon complete puromycin selection at the time of input sample collection (day 0). Cells were passed every 2–3 d (maintaining full representation) and supplemented with fresh doxycycline. For our initial screen, we collected genomic DNA (gDNA; at least 1,000 cells per construct representation) from each sample on day 0, day 15 and day 30. For the TIGER on-target and titration screens, we collected samples on day 0, day 7 and day 14.

Screen library design and pooled oligo cloning

For this study, we designed the following three CRISPR–Cas13d gRNA libraries: (1) a total of 120,000 gRNA libraries tiling 16 essential genes with PM, mismatch and indel gRNAs (Supplementary Data 1 and 2), (2) a total of 42,326 gRNA on-target libraries were designed using our TIGER$_{combined}$ model targeting 5,166 target genes with eight PM gRNAs (Supplementary Data 5) and (3) a total of 48,608 gRNA titration libraries targeting 1,082 essential genes with four PMs and ten SM gRNAs (Supplementary Data 10). For each target gene, we designed gRNAs against the most abundant isoform using publicly available transcript isoform quantifications from the Cancer Cell Line Encyclopedia for HEK-TE quantification (https://sites.broadinstitute.org/ccle/).

For the first library, we selected 16 genes (Supplementary Data 1) previously found to be essential using RNA-targeting CRISPR screens[7]. We predicted gRNA efficacies for all possible 23-mer gRNAs using our RF$_{on}$ model with minimal constraints (T-homopolymer < 4, V-homopolymer < 5)[7]. We only considered gRNAs falling within the coding region boundaries, which have been shown to be the most active previously[7]. For ten genes, we selected, if possible, ~100 evenly spaced PM gRNAs for each prediction quartile to a total of 400 gRNAs. For six genes, we selected 1,000 PM gRNAs in the same way across all prediction quartiles. For one of the six genes, we selected 100 PM gRNAs in the most effective quartiles Q3 (30 gRNAs) and Q4 (70 gRNAs) and designed 181 gRNA variants per parental reference PM gRNA. These 181 variant gRNAs included 23 SM gRNAs, 50 random DM gRNAs, eight consecutive DMs, 50 RTMs, eight consecutive TMs, eight SDs, eight random double nucleotide deletions (DDs), five consecutive double deletions (CDs), eight SIs, eight random double nucleotide insertions (DIs) and five consecutive double insertions (CIs). For nucleotide substitution or insertions, we chose a random base avoiding self-substitutions and terminal thymidine/uridine (T/U) bases. We filtered PM gRNAs and derivatives to not end with a T/U at position 23, as it would be interpreted as the start of the Pol III-terminator sequence directly downstream of the gRNAs and would lead to gRNAs truncations. For SM gRNA, we randomly sampled one nucleotide substitution per position to evenly cover all 23 positions. For all other categories, we randomly sampled from all possible variants. In total, we designed 120,000 gRNAs (Supplementary Data 2, 10,000 PM gRNAs, 108,600 PM gRNA variants and 1,400 NT control guides with more than three mismatches to the hg19 transcriptome).

For the second and third libraries, we predicted PM and SM gRNA efficacies using our TIGER$_{combined}$ model. In total, we designed gRNAs for 5,166 genes. We included several (not mutually exclusive) gene groups as follows: 1,082 common essential genes (DepMap release 08-2021; DepMap score < −1 in ≥500 cell lines or scored in all five screens in ref. 58), 1,052 other genes (DepMap score < −0.5 in ≥100 cell lines or scored in at least three screens in ref. 58), 1,477 RNA binding protein genes[59], 1,706 transcription factor genes[60] and 458 control genes (DepMap score between −0.1 and +0.1 in ≥700 cell lines, spanning a wide range of gene expression values from 1 to >1,000 TPM). For all 5,166 target genes, we designed eight top PM gRNAs targeting the genes coding region. For the 1,082 essential genes, we designed four top PM gRNAs for ten SM gRNA variants, randomly sampling across the 69 possible variants per PM gRNA with a roughly even spread of predicted low-activity to high-activity SM gRNA variants. We added 1,000 NT control guides with more than three mismatches to the hg19 transcriptome. On-target and titration libraries were designed together, padded by library-specific priming sites enabling separate PCR amplification and plasmid library cloning.

The pooled crRNA libraries were synthesized as single-stranded oligonucleotides (Twist Biosciences) and then PCR amplified in one reaction per 10,000 gRNAs with a 50-μl reaction volume—0.5 μl Q5 polymerase (NEB), 10 μl 5× reaction buffer, 2 μl oligo pool (1 ng μl^{-1}), 2.5 μl of each forward and reverse primer (10 μM), 2.5 μl dNTPs (10 mM) and 30 μl water. PCR conditions were 98 °C/30 s, 8× or 9× (98 °C/10 s, 63° C/10 s and 72 °C/15 s) and 72 °C/3 min. The PCR product was either gel-purified or purified using the Zymo Clean and Concentrator 25 kit and then Gibson-cloned into BsmBI-digested pLenti*Rfx*Guide-Puro (Addgene, 138151) using eight Gibson reactions with a 20-μl reaction volume each time—500 ng digested plasmid (0.088 pmol), 123.15 ng purified oligo pool (1.3245 pmol, 15:1 molar ratio), 10 μl 2× Gibson Assembly Master Mix (NEB), incubated for 1 h at 50 °C. Each gRNA was represented by >200 colonies. Complete library representation with minimal bias (90th percentile/10th percentile gRNA read ratios of ~2:5 for all libraries) was verified by Illumina sequencing (MiSeq).

Screen readout and read analysis

We used a two-step PCR protocol (PCR1 and PCR2) to amplify the gRNA cassette for Illumina sequencing from gDNA. The gDNA was extracted from screen cells using the following protocol[57]: for 100 million cells, 12 ml of NK lysis buffer (50 mM Tris, 50 mM ethylenediaminetetraacetic acid, 1% SDS and pH 8) was used for cell lysis. Once cells were resuspended, 60 μl of 20 mg ml^{-1} Proteinase K (Qiagen) was added and the sample was incubated at 55 °C overnight. The next day, 60 μl of 20 mg ml^{-1} RNase A (Qiagen) was added and mixed, and samples were incubated at 37 °C for 30 min. Then, 4 ml of prechilled 7.5 M ammonium acetate was added, and samples were vortexed and spun at 4,000*g* for 10 min. The supernatant was placed in a new tube, mixed well with 12 ml isopropanol and spun at 4,000*g* for 10 min. DNA pellets were washed with 12 ml of 70% ethanol, spun and dried, and pellets were resuspended

with 0.2× TE buffer (Sigma-Aldrich). In addition, we also generated linearized plasmid library input and diluted it down to mimic similar copy number conditions as the gDNA samples.

For the PCR1 reaction, we used 960 µg (screen 1) or 880 µg (screens 2 and 3) gDNA for each sample. For each sample, we performed 96 (screen 1) or 88 (screens 2 and 3) 100 µl PCR1 reactions with a 100-µl reaction volume—10 µl 10× Taq buffer, 0.02 U μl^{-1} Taq-B enzyme (Enzymatics, P7250L), 0.2 mM dNTPs, 0.2 µM forward and reverse primers and 100 ng gDNA per µl; thermocycling conditions for PCR1 were 94 °C/30 s, 10× (screen 1) or 18× (screens 2 and 3; 94 °C/10 s, 55 °C/30 s, 68 °C/45 s) and 68 °C/3 min. Because our screen 1 library contained a large number of gRNAs with a hamming distance of one to one another, we decided to only perform ten cycles of PCR1. For each sample, all PCR1 products were pooled and mixed.

For each sample, we performed 24 (screen 1) or 6 (screens 2 and 3) PCR2 reactions with a 100-µl reaction volume—20 µl 5× Q5 buffer (NEB), 0.01 U μl^{-1} Q5 enzyme, 20 µl PCR1 product, 0.2 mM dNTPs and 0.4 µM forward and reverse PCR2 primers in 100 µl. Thermocycling conditions for PCR2 were 98 °C/30 s, 18× (screen 1) or 8× (screens 2 and 3; 98 °C/10 s, 63 °C/30 s and 72 °C/45 s) and 72 °C/5 min. For screen 1, we performed an additional PCR2 on the linearized plasmid pool sample with either Q5 or Taq-B polymerase. We found raw counts to be highly correlated with no obvious influence due to the choice of polymerase. PCR primers can be found in Supplementary Data 14.

For each sample, PCR2 products were pooled, followed by normalization (gel-based band densitometry quantification), before combining equal amount of uniquely barcoded samples. The pooled product was then purified using SPRI beads. First, we performed a 0.6× vol/vol SPRI to remove gDNA carryover, followed by the addition of a 0.3× vol/vol SPRI (0.6 + 0.3 = 0.9× final) to the supernatant to purify the ~260 bp PCR product. Oligonucleotides can be found in Supplementary Data 14. The final amplicons were sequenced on Illumina NextSeq 500—II MidOutput 1 × 150 v2.5 (screen 1) and Illumina NextSeq 500—II HighOutput 1 × 150 v2.5 (screens 2 and 3).

Reads were first demultiplexed based on Illumina i7 barcodes present in PCR2 reverse primers using *bcl2fastq* and then by their custom in-read 5′ barcode allowing for one mismatch. They were trimmed to the expected gRNA length by searching for known anchor sequences relative to the guide sequence. They were collapsed (*FASTX-Toolkit*) to count perfect duplicates followed by exact string-match intersection with the reference to retain only perfectly matching and unique alignments. The raw gRNA counts (Supplementary Data 3, 6 and 11) were normalized using a median of ratio method[61] and then batch-corrected for biological replicates using combat implemented in the *SVA* R package[62]. Nonreproducible technical outliers were removed by flagging individual values with high variance within replicate samples of each time point (D0, D15 and D30 for screen 1 and D0, D7 and D14 for screens 2 and 3). These outlier counts are a common contaminant in early-passage input samples due to plasmid carryover from virus production in the viral supernatant used for infection[63]. Specifically, we calculated the $\log_2$-transformed variance across all samples for each gRNA. Then, we calculated the variance within each time point across all replicate samples and flagged individual gRNA counts within the upper 0.6% variance percentile (for example, cutoff = −1.366 for screen 1). We only flagged individual counts if those were present in the upper half of the count distribution to avoid masking variance within depleting gRNAs. Because we had three replicates, we replaced the flagged count with NA but kept the other two unflagged replicates. For screen 1, we removed 154 gRNAs due to all filtering steps (28 gRNAs were not detected; 126 gRNAs were only lowly represented in the plasmid library with less than 60 normalized counts).

gRNA enrichments (Supplementary Data 4, 7 and 12) were calculated by building the count ratios between a time point and the corresponding input (day 0) sample for each replicate followed by $\log_2$-transformation ($\log_2$(FC)). Consistency between replicates was estimated using Pearson correlations and robust rank aggregation[64]. For data representation and modeling, we used the mean $\log_2$(FC) across replicates. Delta $\log_2$(FC) for mismatching guides was calculated by subtracting the $\log_2$(FC) of the permuted gRNA from the PM reference guide. For data representations in Figs. 1 and 2, we normalized the observed $\log_2$(FC) guide values in the following way: for each gene D in the dataset, we computed the upper and lower quartiles of the guide $\log_2$(FC) (UQ_D and LQ_D, respectively) as well as the corresponding quartiles for the $\log_2$(FC) among all datasets pooled together (UQ_P and LQ_P). We then updated each FC x as follows:

$$\hat{x} = \left(\frac{x - LQ_D}{(UQ_D - LQ_D)}(UQ_P - LQ_P) + LQ_P\right)$$

Gene essentiality normalization and determination of active guides

Given that essentiality varies across the 16 genes in our first pooled library (that is, some genes are more essential than others), we experimented with different per-gene normalization approaches (including no normalization) to equalize survival effects across all genes in our survival screen before data modeling. All of our considered transformations are location–scale transformations, where the location and scale were derived separately for each gene. Data were transformed as follows: (x − location)/scale. We found that the best generalization performance was obtained using median $\log_2$(FC) as the location and the distance between the 10th and 90th percentiles as the scale.

We used the distribution of NT gRNAs to determine which gRNAs targeting essential genes to consider as being active. We selected the most depleted 1% of NT gRNAs as the threshold for activity after testing for normality of the NT gRNA distribution (Lilliefors test, $P < 0.001$). This threshold corresponds to a $\log_2$(FC) < −0.50 for screen 1 (HEK293FT) as well as $\log_2$(FC) < −0.44 and $\log_2$(FC) < −0.29 for screen 2 in HEK293FT and HAP1, respectively.

Cell surface marker flow cytometry pooled screens

For certain analyses (for example, entire dataset holdout), we used a set of published pooled gene-tiling Cas13d screens using flow cytometry of cell surface markers from our previous study[7]. In these screens, library-transduced cells were sorted based on the expression of cell surface markers (CD46, CD55 or CD71). Specifically, we used the processed count data (available in Supplementary Data 8; ref. 7) for the CD46, CD55 and CD71 tiling screens and calculated gRNA FC by computing the count ratios between sorted bins (high versus low) for each replicate followed by $\log_2$-transformation ($\log_2$(FC)).

Predicting RNA secondary structures and RNA–RNA hybridization energies

crRNA secondary structure and MFEs were derived using RNAfold (--gquad) on the full-length crRNA (DR + guide) sequence[65]. Target RNA-pairing probability (accessibility) was calculated using RNAplfold (-L 40 -W 80 -u 50) as described previously[7]. These parameters specify a moving window of 80 nucleotides and a maximal base pairing span of up to 40 nucleotides. We chose these parameters because previous studies[3,66,67] both in the context of Cas13 and RNA interference have found optimal performance for a local window around the target site.

We performed a grid search calculating the RNA accessibility for each target nucleotide in a window of minus 20 bases downstream of the target site to plus 20 bases upstream of the target site assessing the unpaired probability of each nucleotide over 1–50 bases for all perfectly matching guides. Then, we calculated the Pearson correlation coefficient between the $\log_{10}$-transformed unpaired probabilities and the observed gRNA $\log_2$(FC) for each point and window relative to the gRNA. We selected four centers of high correlation to feed into the model. Target RNA accessibility features are (1) position −11 upstream

to the first spacer nucleotide with a width of 23 nt; (2) position −11 with a width of 4 nt; (3) position −19 with a width of 4 nt and (4) position −25 with a width of 4 nt.

RNA–RNA hybridization between the gRNA (PM and gRNA with nucleotide substitutions but not for indel gRNAs) and its target site was calculated using RNAhybrid (-s -c)[68]. We calculated the RNA-hybridization MFE for each gRNA nucleotide position p over the distance d to position $p + d$ with its cognate target sequence. All measures were either directly correlated with the observed gRNA $\log_2$(FC) or using partial correlation to account for the crRNA folding MFE. In each case, we computed the Pearson correlation. We selected three centers of high correlation to feed into the model. Hybridization MFE features are (1) position $p = 1$ and $d = 23$; (2) position $p = 3$ and $d = 12$; and (3) position $p = 15$ and $d = 9$.

Assessing target RNA context

To assess the target RNA context, we calculated the nucleotide probability at each position (p) over a window (w) of 1–50 nucleotides centered around the position of interest (for example, $p = -18$ with $w = 11$ summarizes the nucleotide content in a window from −23 to −13 with +1 being the first base of the crRNA). We evaluated p for all positions within 75 nucleotides upstream and downstream of the gRNA. The nucleotide context of each point was then correlated with the observed $\log_2$(FC) crRNA enrichments for all PM crRNAs, either directly or using partial correlation accounting for crRNA folding MFE. In each case, we used the Pearson correlation. We used the same positions p and window sizes that have been used before in RF_{on}[7]. These RNA nucleotide context features were only used in the RF_{on} model but not in the CNN models.

Convolutional neural network for deep learning

The sequence input to the CNN consists of 23 nt target and gRNA sequences with 2 nt of upstream and downstream target context. Initially, the input is processed with two consecutive convolution layers each with 32 4 × 4 kernels followed by a rectified linear unit. At the next stage, we perform max pooling with a pooling size of two. We flattened the resulting 32 channels of 16 × 4 learned feature representations to a vector of length 2,048 to which we concatenate any nonsequence features. To regularize, we use a 25% dropout before our dense layers, which consist of three layers with 128, 32 and 1 neurons, respectively. Between each dense layer, we use a 10% dropout. The first two dense (hidden) layers apply sigmoid activations. The single output neuron does not apply any activation function. This design is similar to previous deep learning models in computer vision[38] and sequence analysis[26].

We use a log hyperbolic cosine (log-cosh) loss function, which is similar to an L1 loss but is continuously differentiable. We optimize our models with *Adam* (adaptive moment estimation), an adaptive stochastic optimization algorithm that requires only first-order gradients[69], with a learning rate of 0.001. We employ early stopping with patience of 100 epochs and restore model parameters of the best epoch. We implemented our models using TensorFlow via the Docker image tensorflow: 2.11.0 with GPU support. In addition, our code imports the following Python packages: biopython (v1.80), pandas (v1.5.2), tensorflow-probability (v0.19.0), matplotlib (v3.6.1), seaborn (v0.12.1), shap (v0.41.0), statsmodels (v0.12.2) and sklearn.

Cross-validation across genes, target sites, gRNAs and datasets

We consider the following three CV approaches: gene level, target level and individual gRNAs. For gene-level CV, we create 16 folds where each fold contains all (gRNA and target) tuples specific to that gene. This CV approach promotes transcriptome-wide generalization by holding out entire genes (and all of the corresponding target sites and gRNAs for that gene). The second approach (target-level CV) randomly divides target sites into ten, nonoverlapping folds. Holding out a target site places all PM and mismatched gRNAs designed for that target site into the holdout set, ensuring that a target sequence never appears both in training and validation sets. This is important as active PM and SM gRNAs for the same target sequence can have similar activity. The third approach (gRNA-level CV) holds out individual gRNAs; notably, related gRNAs such as mismatch gRNAs (for a particular PM gRNA in the holdout group) may still be included in the training set. As expected, we observe better performance with gRNA-level CV than target- or gene-level CV.

In some experiments, we train on the 120,000 gRNA library and then test on a separately collected pooled screen–flow cytometry of cell surface proteins from three genes (entire dataset holdout)[7]. For the entire dataset holdout, we used the $\log_2$-transformed gRNA depletion in the fluorophore-low bin divided by the fluorophore-high bin. Supplementary Note contains a formal description of all of these CV strategies. All CV folds used in the study are presented in Supplementary Data 2.

Comparison with linear regression, random forest and recurrent neural networks

To quantify the RF_{on} performance for predicting PM gRNA efficacy on the essentiality screen, we used the same 16 gene holdouts described above and iteratively retrained the model with its previously described architecture[7] on 15 genes to predict the 16th gene. In addition, we retrained RF_{on} on the entire essentiality dataset to evaluate its performance on the phenotype selection data (flow cytometry). For the recent Cas13 deep learning models[8,9], we uploaded each of the 16 transcripts for our targeted essential genes to their respective web portals to generate predictions for all potential target sites along the transcript (1 nt tiling). We only compute performance at target sites for which we measured gRNA activity. Cheng et al. transformed FC data via a parameterized sigmoid function[9]. We exactly applied this transformation to our FCs before computing their performance metrics. Wei et al. generated predictions for 30 nt spacers[8]. When computing their performance, we use the first 23 nt of their 30 nt spacer sequence to match their predictions to our 23 nt gRNA sequences.

We also designed a BiGRU[70] model as an alternative deep learning approach that is based on recurrent units instead of convolution units. We designed the BiGRU architecture to closely mimic the dense layers of TIGER. The BiGRU model uses a convolution kernel of length one to learn a 32-dimensional embedding from the one-hot-encoded target and gRNA sequences (16 unique 32-dimensional embeddings for each possible guide-target pair). This embedding feeds a BiGRU layer, which outputs a 32-dimensional representation for each sequence position. We concatenate the 32-dimensional outputs for both directions and all sequence positions and, thereafter, flatten it to a vector of length 64× sequence length and concatenate nonsequence features. To regularize, we apply a 25% dropout on this vector before feeding it to our dense layers, which consist of three layers with 128, 32 and 1 neurons. Between each dense layer, we use a 10% dropout. The first two dense (hidden) layers apply sigmoid activations. The single output neuron does not apply any activation function.

For linear regression, we used one-hot-encoded sequences (flattened into a vector) and concatenate nonsequence features to the same vector.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data generated in this study have been deposited at NCBI Gene Expression Omnibus (GEO) with the accession number GSE232228. Flow cytometry screen data from ref. 7 is available under the accession number GSE142675.

Code availability

Code to run Cas13d on-target and off-target TIGER models has been deposited on Github (https://github.com/daklab/tiger). A web-accessible version of TIGER is available at https://tiger.nygenome.org/.

References

57. Chen, S. et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246–1260 (2015).
58. Hart, T. et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).
59. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
60. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
62. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
63. Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. & Elledge, S. J. Sources of error in mammalian genetic screens. *G3 (Bethesda).* **6**, 2781–2790 (2016).
64. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
65. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
66. Agarwal, V., Subtelny, A. O., Thiru, P., Ulitsky, I. & Bartel, D. P. Predicting microRNA targeting efficacy in Drosophila. *Genome Biol.* **19**, 152 (2018).
67. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
68. Krueger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* **34**, 451–454 (2006).
69. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
70. Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: encoder–decoder approaches. In *Proc. SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (eds Wu, D. et al.) 103–111 (Association for Computational Linguistics, 2014).

Acknowledgements

We thank the entire Sanjana and Knowles Labs for their support and advice. D.A.K. is supported by Columbia and NYGC startup funds, NIH/NCI (R21CA272345) and an NSF CAREER (DBI2146398). N.E.S. is supported by NYU and NYGC startup funds, NIH/NHGRI (DP2HG010099), NIH/NCI (R01CA218668), NIH/NIGMS (R01GM138635), DARPA (D18AP00053), Cancer Research Institute and the Simons Foundation for Autism Research Initiative.

Author contributions

H.W. and N.E.S. conceived the study. H.W., A.S., D.A.K. and N.E.S. designed the experiments. H.W. and A.M. cloned libraries and performed the CRISPR screens. S.K.H. assisted with cell culture for pooled screens. H.W., A.S., D.A.K. and N.E.S. analyzed the data and developed the deep learning model. A.S. and E.J.K. implemented the web-based online TIGER tool. H.W., A.S., D.A.K. and N.E.S. wrote the paper with input from all authors.

Competing interests

The New York Genome Center and New York University have applied for patents relating to the work in this article. H.W. is a cofounder of Neptune Biotech. N.E.S. is an advisor to Qiagen and is a cofounder of OverT Bio. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01830-8.

Correspondence and requests for materials should be addressed to David A. Knowles or Neville E. Sanjana.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature research

Corresponding author(s): Neville Sanjana and David Knowles

Last updated by author(s): May 10, 2023

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	A full data collection description is provided in the method section: Sequencing reads were trimmed using a custom python (v.2.7.10) script (available upon request). gRNA read counts were generated using FASTX-Toolkit (v0.0.14). gRNA and target RNA feature extraction: RNAfold and RNAplfold ViennaRNA (v2.4.10) RNAhybrid (v2.1.2) gRNA prediction from alternative guide RNA prediction algorithm were collected here: http://deepcas13.weililab.org (No version ID available) https://www.rnatargeting.org (No version ID available) https://cas13design.nygenome.org (v1.1, re-trained on data generated in this study where indicated in the text)
Data analysis	The method section describes in detail data processing and analysis: Pooled screen data processing: R v3.6.0 RobustRankAggreg v1.1 (R package) SVA v3.34.0 (R package) Deep Learning: We use TensorFlow's official Docker image for version 2.11.0 with GPU support: https://hub.docker.com/layers/tensorflow/tensorflow/2.11.0-gpu/images/ sha256-67f1a7b[...]1c0cd311655be7477f2bc1b6f27e014b9a57231bd55b3?context=explore Additional python packages not included in Docker image: bbiopython (version 1.80) pandas (version 1.5.2) matplotlib (version 3.6.1)

seaborn (version 0.12.1)
shap (version 0.41.0)
statsmodels (version 0.13.5)
tensorflow-probability (version 0.19.0)
statsmodels (version 0.12.2)
sklearn
We have made all code essential to this study available on https://github.com/daklab/tiger as noted in the Code Availability Statement.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated in this study has been deposited at NCBI Gene Expression Omnibus (GEO) with the accession number GSE232228. Raw flow cytometry screen data from Wessels, Méndez-Mancilla et al. (ref. 7) is available under the accession number GSE142675.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A full description of Cas13d guide RNA library size and composition is provided in the method section. All screens have been conducted in at least three independent replicate experiments. Detailed statistics for included guide numbers and guide classes can be found in Supplementary Data 2, 5 and 10. The final predictive guide model (TIGERcombined) was constructed using features for 93,145 guides RNAs (Types: PM, SM, DM, TM, RDM, RTM) (255 filtered gRNAs of the same gRNA types removed) that match to coding sequences of the 16 target genes presented in Supplementary Data 1 and 2.
Data exclusions	The initial HEK293FT off-target screen was conducted in three replicates. In total, we removed 333 targeting gRNAs due to filtering steps described in the method section. In addition, non-reproducible technical outliers were masked by flagging individual values with high variance within replicate samples of each timepoint (D0, D15, D30). Masking did not lead to gRNA removal.
Replication	All screens have been conducted in at least three replicate experiments. We confirm that all screen have been conducted independently. Replicate screens were highly correlated.
Randomization	We consider three cross-validation methods: one at the gene level, one at the target level and one for technical holdouts. For gene holdouts, we create sixteen folds where each fold contains all (guide, target) tuples specific to that gene. We feel this method is generally the most challenging method to succeed on as it requires and promotes transcriptome-wide generalization. The second method randomly folds target sites into ten, non-overlapping folds. Holding out target site places all perfectly matched and mismatched gRNAs for a target site into the holdout set, ensuring that a target sequence never appears both in training and validation sets. Without this restriction (i.e. randomly folding at the gRNA level), we found that more sophisticated model architectures, specifically those with recurrent units, that perform extremely well on gRNA holdouts (Pearson r = 0.88) failed to generalize to held out genes (Pearson r = 0.36). For the technical holdout, we hold out gRNAs from the flow cytometry of cell surface proteins from three genes. Thus, our technical holdout is a multi-gene holdout. For the technical holdout, we used the log2-transformed gRNA depletion in the fluorophore-low bin divided by the fluorophore-high bin.
Blinding	Not applicable. All experiments have been processed and analyzed in an unbiased way. Analysis did not require blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293FT cells (also denoted to as HEK293 cells) were acquired from Thermo Fisher (R70007). HAP1 cells were acquired from Horizon Discovery.
Authentication	The cell lines were not authenticated in the lab after purchase from the vendors.
Mycoplasma contamination	All cell lines were tested as mycoplasma-free using Lonza MycoAlert (#LT07-518).
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.