A generalizable Cas9/sgRNA prediction model using machine transfer learning with small high-quality datasets

Dalton T. Ham[1,+], Tyler S. Browne[1,+], Pooja N. Banglorewala[1], Tyler Wilson[2], Richard Michael[2], Gregory B. Gloor[1,*], and David R. Edgell[1,*]

[1] Department of Biochemistry, Schulich School of Medicine and Dentistry, London, ON, N6A5C1, Canada
[2] Tesseraqt Optimization Inc, Toronto, ON, Canada
[+] These authors contributed equally to this work
[*] Corresponding authors: dedgell@uwo.ca, ggloor@uwo.ca

ABSTRACT

The CRISPR/Cas9 nuclease from *Streptococcus pyogenes* (SpCas9) can be used with single guide RNAs (sgRNAs) as a sequence-specific antimicrobial agent and as a genome-engineering tool. However, current bacterial sgRNA activity models poorly predict SpCas9/sgRNA activity and are not generalizable, possibly because the underlying datasets used to train the models do not accurately measure SpCas9/sgRNA cleavage activity and cannot distinguish cleavage activity from toxicity. We solved this problem by using a two-plasmid positive selection system to generate high-quality biologically-relevant data that more accurately reports on SpCas9/sgRNA cleavage activity and that separates activity from toxicity. We developed a new machine transfer learning architecture (crisprHAL) that can be trained on existing datasets and that shows marked improvements in sgRNA activity prediction accuracy when transfer learning is used with small amounts of high-quality data. The crisprHAL model recapitulates known SpCas9/sgRNA-target DNA interactions and provides a pathway to a generalizable sgRNA bacterial activity prediction tool.

INTRODUCTION

The Cas9 nucleases from the type II-A clustered regularly interspaced short palindromic repeat (CRISPR) system have gene-editing applications in both bacteria and eukaryotes[1,2]. Cas9 cleavage of DNA templates requires an associated CRISPR RNA (crRNA) that is complementary to the target site, and a trans-activating CRISPR RNA (tracrRNA) that is required for crRNA assembly with Cas9[3]; in most applications these two RNAs are genetically fused into a single guide RNA (sgRNA)[4]. In bacteria, Cas9 nucleases can be used as sequence-specific antimicrobial agents to target distinct bacterial species for elimination[5–11] because many bacteria lack appropriate DNA repair pathways to repair double-strand breaks (DSB). Cleavage by Cas9 causes replication fork collapse and cell death[12]. Alternatively, Cas9 cleavage can eliminate plasmids through the cellular RecBCD exonuclease pathway that degrades linearized DNA. Cas9 can also be used for bacterial genome engineering[13–15], or for transcriptional modulation with catalytically inactive dCas9 variants[16–18].

A major unsolved problem when using Cas9 is the inability to accurately select sgRNA/ Cas9 combinations that lead to high on-target activity in both eukaryotic and prokaryotic systems. Selection of sgRNAs typically involves computational prediction of activity where the underlying models are usually trained on data of *in vitro* or *in vivo* Cas9/sgRNA activity, and may also include biochemical parameters of Cas9 activity, biophysical calculations of sgRNA:DNA stability, and chromatin accessibility information[19–24]. However, as recently reported[25,26], most computational models poorly predict sgRNA activity outside of the dataset on which they are trained. This lack of generalizability could be because the underlying data are sparse and not independently validated, because the datasets may not accurately represent Cas9/sgRNA cleavage activity and instead report a secondary DNA repair outcome of DSB generation, because the deep learning algorithms are not optimal, or a combination of all three[25].

In spite of the conceptual simplicity in targeting sgRNAs to small bacterial genomes, eukaryotic-based computational models fail to accurately predict activity in bacteria[27]. One issue for sgRNA activity predictions in bacteria is that there are few bacteria-specific large-scale datasets of Cas9/sgRNA activity[28,29]. In each case, deep sequencing was used to readout sgRNA abundance of a pooled sgRNA library targeting the *Escherichia coli* genome, with the assumption that sgRNA depletion was correlated with active Cas9/sgRNA combinations. A complicating factor in assessing Cas9/sgRNA activity in bacteria is that expression of Cas9 (and dCas9) alone can result in cellular toxicity and slow growth[30–33]. Thus, experimental strategies that only use bacterial killing as a measure of Cas9/sgRNA activity cannot separate toxicity from activity because both will result in depletion of sgRNAs from a pooled high-throughput experiment. Two sgRNA prediction models have been developed based on this data, sgRNA-cleavage-activity-prediction[28] and DeepSgRNAbacteria[34], but we found poor correlation between predicted Cas9/sgRNA activity and killing of *Salmonella typhimurium*[5]. Other factors that possibly impact sgRNA activity in bacteria include sub-optimal secondary structures in the crRNA and tracrRNA[35], and similarity between the crRNA seed region and so-called "non-targets" in bacterial genomes. In contrast, DNA modifications do not impact activity of type II CRISPR systems (from which Cas9 is derived)[36,37]. Similarly, there is no bias in activity for Cas9/sgRNAs targeting the template or non-template strand of transcribed genes, or in targeting the leading or lagging strands relative to DNA replication origins[5].

Taken together, the evidence indicates that there is a pressing need for additional high-quality bacterial sgRNA activity data sets to validate and generalize previous findings, and to provide training data for predictive machine learning models. Here, we develop a paired experimental design in *E. coli* that compares behaviours of sgRNA/Cas9 combinations in repressed and induced conditions to provide a readout of activity where active sgRNAs are enriched in a pooled library. This approach differs from previous depletion studies by accounting for initial sgRNA abundance in the pooled library, and does not rely on end-of-experiment sgRNA abundance as the sole indicator of sgRNA activity. Additionally, this setup distinguishes highly active Cas9/sgRNA combinations from toxic ones with poor growth, even in repressed conditions. We used this approach with the SpCas9 nuclease[4] and the TevSpCas9 dual-nuclease[38] to generate robust sgRNA activity datasets to train a sgRNA prediction model, crisprHAL (crispr mac**H**ine tr**A**nsfer **L**earning) that recapitulates the known biology of the Cas9/sgRNA-target DNA interaction surface. Significantly, we found that transfer learning from existing datasets with a small amount of sgRNA activity data (279 sgRNAs) from our new assays improved bacterial sgRNA predictions relative to previous models. Collectively, our study highlights the importance of accurate sgRNA activity data and transfer learning as being crucial for computational modelling.

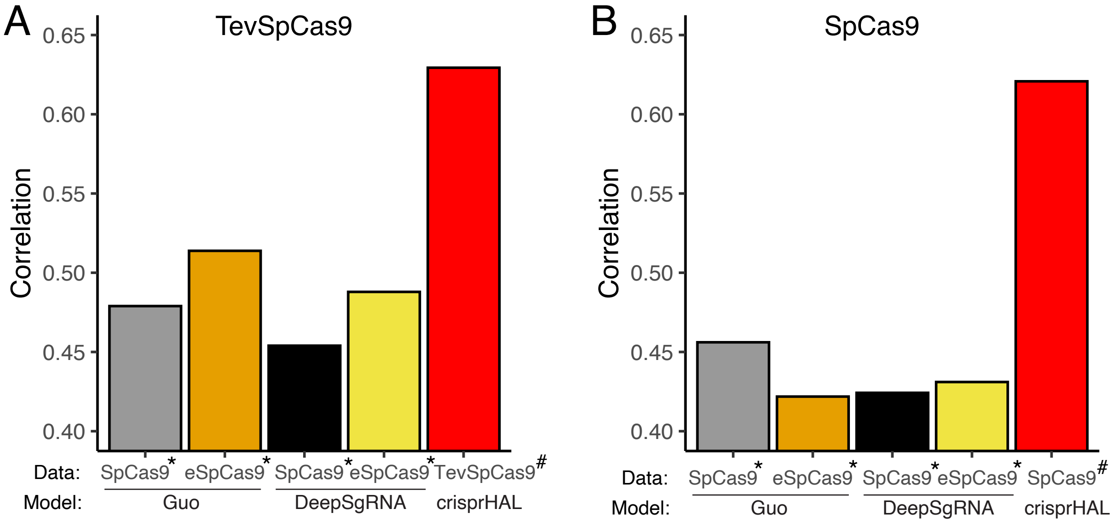

Figure 1. Spearman ranked correlation of predicted versus measured activity for sgRNA prediction models. Barcharts are Spearman Rank correlations between the **A)** TevSpCas9 dataset (n=279) and **B)** the SpCas9 dataset (n=303) generated in this study and predictions from bacterial sgRNA activity models including crisprHAL. The crisprHAL values are reported as the average rank correlation from 5-fold cross validation. For both panels, asterisks (*) indicate datasets from Guo et al.[28] and hash marks (#) indicate datasets generated in this study.

RESULTS

Current bacterial sgRNA prediction models are poorly generalizable

We were interested in understanding why existing sgRNA prediction models poorly correlate with *in vivo* activity[5]. Thus, we tested whether current bacterial sgRNA prediction models were generalizable to different SpCas9 activity datasets (Figure 1). For this, we used a two-plasmid positive selection system (Figure 2) to generate two high-quality activity datasets for the SpCas9 and the TevSpCas9 dual nuclease (as described in detail in the following sections). When the TevSpCas9 dataset was used as an input for the sgRNA-cleavage-activity-prediction model (hereafter referred to as the Guo model) and the DeepSgRNAbacteria model (hereafter referred to as the DeepSgRNA model), we found only modest predictive performance of either model by Spearman correlation of rank order between experimentally determined and predicted activity (Figure 1). Modest predictive power was observed regardless of which of the two published sgRNA depletion datasets the Guo or DeepSgRNA models were trained on; one dataset used SpCas9 and the other used an enhanced high-fidelity SpCas9 variant (eSpCas9). These results emphasize a major issue with Cas9/sgRNA activity predictions, namely the lack of generalizability and accuracy when models are used with data outside of the initial training data, and highlight the need for high-quality datasets that accurately report on Cas9/sgRNA cleavage activity.

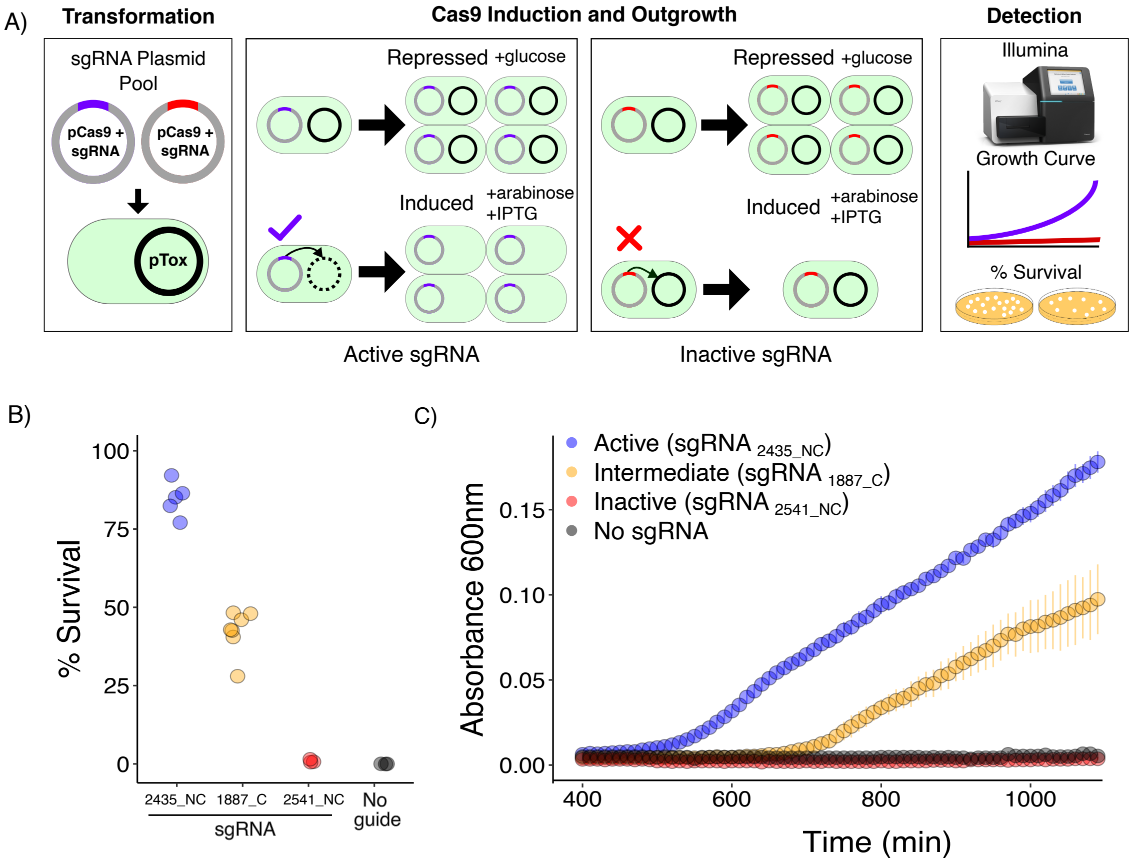

Figure 2. Two-plasmid survival assay. A) Experimental workflow of the two-plasmid system. *Transformation*, the pCas9 plasmid expressing SpCas9 or TevSpCas9 from an arabinose-inducible promoter and a sgRNA from a constitutive tetracycline resistance gene promoter is transformed into *E. coli* harbouring pTox. *Induction and Outgrowth*, transformed cells are split into repressed (0.2% D-glucose) or induced (0.2% L-arabinose and 0.1 mM IPTG) conditions and grown for 18 hrs. *Active sgRNAs, blue* promote robust cleavage of the toxic plasmid and cell growth while *inactive sgRNAs, red* do not cleave pTox preventing cell growth. *Detection*, SpCas9/sgRNA activity can be read out by i) deep-sequencing the pCas9 sgRNA cassette, ii) growth curves that measure optical density of induced and repressed cultures, or iii) plating on solid media to determine a percent survival based on the ratio of colonies on induced media (chloramphenicol and IPTG) and repressed media (chloramphenicol and D-glucose). **B)** Different TevSpCas9/sgRNA combinations promote a range of survival. Plot of survival percentage for three different sgRNAs targeted to pTox (2435_NC,1887_C,2541_NC) identified as active (blue), intermediate (orange), inactive (red) as well as a no-sgRNA(NG) control (black). Individual data points represent independent experiments. **D)** Growth curve of *E. coli* harbouring the SpCas9/sgRNA combinations used in panel **B** plotted as time versus absorbance at 600 nM. Data points represent the mean of three biological replicates and the whiskers representing the standard deviation from the mean.

Profiling sgRNA activity using a two-plasmid system

To increase the accuracy of SpCas9 and TevSpCas9 targeting predictions, we started with an improved assay in which we used an integrated approach to assess SpCas9/sgRNA activity in *E. coli* (Figure 2A). We adapted a two-plasmid system used for *in vivo* selection experiments[39–41] that is known to correlate with enzymatic activity *in vitro*[42] and expressed the SpCas9 or dual-nuclease TevSpCas9 protein (arabinose inducible) and a sgRNA (constitutive expression) from one plasmid (pCas9) in combination with a second plasmid (pTox) harbouring the *ccdB* DNA gyrase toxin controlled by an IPTG inducible promoter (Figure 2A). Cleavage of the pTox plasmid by an active SpCas9/sgRNA combination or TevSpCas9/sgRNA combination (Figure 2B) leads to degradation of the pTox plasmid and subsequent cell growth and enrichment of cells lacking the pTox plasmid in the population. Inactive SpCas9/sgRNA or TevSpCas9/sgRNA combinations do not eliminate the pTox plasmid and are unable to grow under toxin-inducing conditions. Importantly, the activity of the (Tev)SpCas9/sgRNA combination is related to the rate of pTox plasmid clearance, and so partially active combinations will have intermediate outgrowth and lethality characteristics. With this system, (Tev)SpCas9/sgRNA activity can be analyzed by deep sequencing of the sgRNA expression cassette following competetive growth in liquid media, or by growth rate in liquid media, or by counting colonies grown on solid media (Figure 2A). The dual-active-side nuclease TevSpCas9 has an extended targeting requirement that includes the 5'-CNNNG-3' I-TevI cleavage motif (Supplementary Figure S1)[38]. Thus, all TevSpCas9 sites are also Cas9 sites, and cleavage by an active TevSpCas9/sgRNA combination will create an additional DSB with the potential to enhance killing efficiency.

We validated this system by targeting three TevSpCas9/sgRNA combinations to a unique region of pTox; sgRNA$_{2435\_NC}$, sgRNA$_{1887\_C}$, and sgRNA$_{2541\_NC}$ (in this naming scheme sgRNAs are identified by the position of the first PAM-distal nucleotide of the sgRNA target in pTox and whether they target the coding or non-coding strand, as all genes are in the same orientation). We plated the transformed *E. coli* cells on solid media and calculated percent survival by comparing the proportion of colony forming units (CFUs) on toxin-inducing or toxin-repressing agar plates. When expressed in combination with the TevSpCas9 protein, the three sgRNAs tested showed survival ranging from 88.2±4.1% (standard error of the mean) for sgRNA$_{2435\_NC}$ to 0.9±0.29% for sgRNA$_{2541\_NC}$ (Figure 2B). When no sgRNA was present (NG, no guide), we observed 0% survival (Figure 2B). We conducted a similar experiment in liquid media by measuring absorbance at 600 nm over 18 hours to detect growth under inducing and non-inducing SpCas9 conditions in combination with the same three sgRNAs (Figure 2C). The resulting growth curves are consistent with the survival values on solid media, with sgRNA$_{2435\_NC}$ promoting robust growth, sgRNA$_{1887\_C}$ promoting intermediate growth and sgRNA$_{2541\_NC}$ and the NG control showing no growth (Figure 2C). Collectively, these results show that bacterial growth is dependent on cleavage of the pTox plasmid by TevSpCas9/sgRNA, agreeing with previous results using SpCas9[41], and that differential TevSpCas9/sgRNA activity results in distinct growth differences over a large and consistent range.

Sensitivity of the two-plasmid system

We next tested the ability of the two-plasmid system to detect changes in SpCas9/sgRNA or TevSpCas9/sgRNA activity when read out via a multiplexed high-throughput sequencing experiment. This experiment was designed to validate the sensitivity of the two-plasmid system when reporting on a range of TevSpCas9/sgRNA activities, and to assess the effect of mismatches between the sgRNAs relative to their cognate target site. For this, we designed an oliognucleotide pool where single and double nucleotide transversions were tiled along the length of 28 different sgRNAs that were targeted to a unique 3.2 kb region of

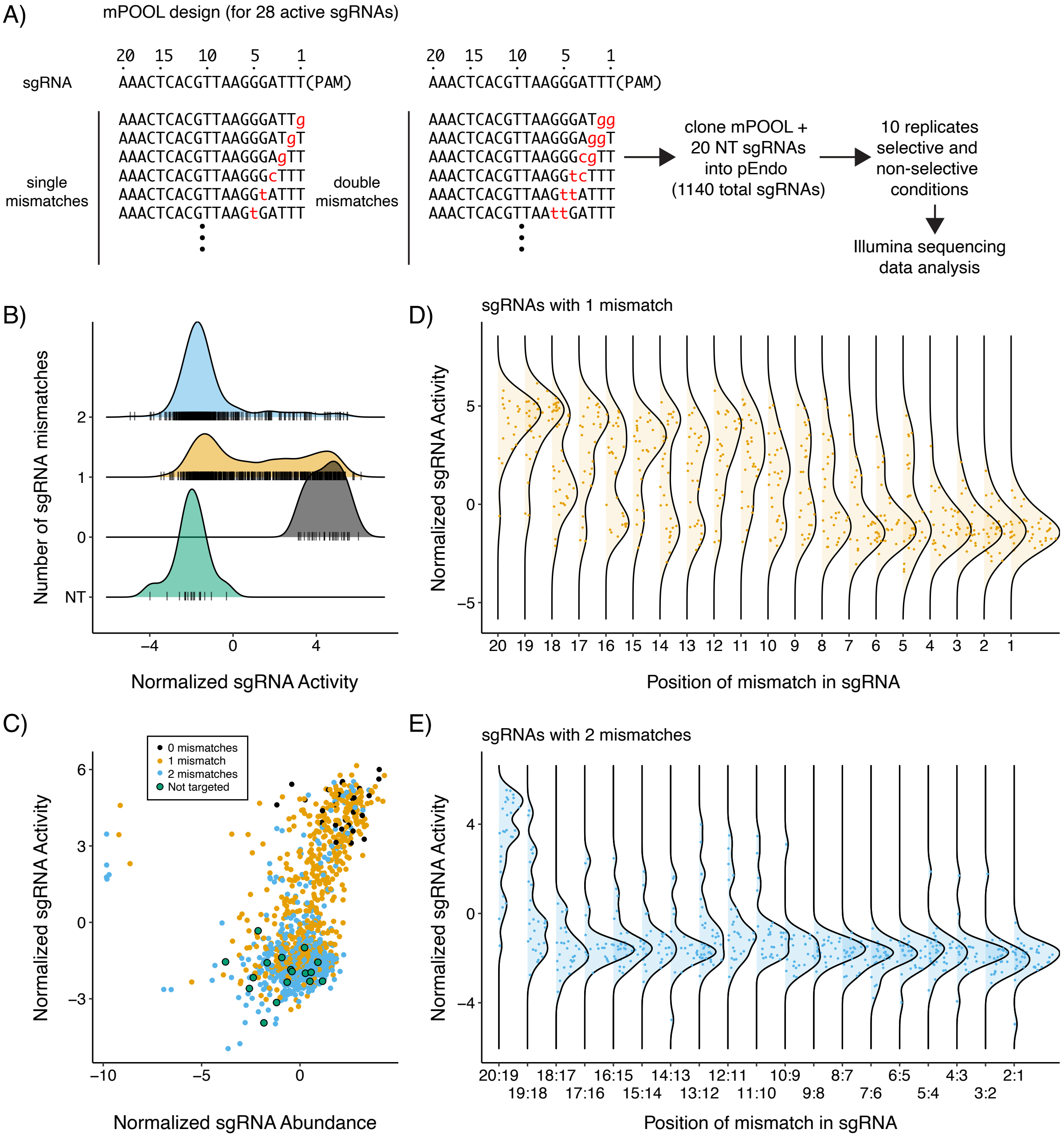

Figure 3. Activity of sgRNAs with single and double mismatches. A) Schematic of the mutant pool (mPool) design and experimental approach. Single and dinucleotide transversions are indicated by lower case red letters, with sgRNAs numbered from PAM proximal (postion 1) to PAM distal (position 20). **B)** Ridge plots of normalized sgRNA activity scores for non-targeted sgRNAs (NT, green) perfectly matching sgRNAs (black), sgRNAS with single nucleotide mismatches sgRNAs (yellow), and sgRNAs with dinucleotde mismatches (cyan). **C)** Bland-Altmann plot comparing the normalized abundance and normalalized activity scores for sgRNAs in the mPOOL with the colours representing the same sgRNAs categories as panel B. **D and E** Ridge plots of normalized sgRNA activity scores by position of mismatch for sgRNAs with **(D)** single or **(E)** 2 mismatches.

the pTox plasmid (Figure 3A, Supplementary Table S1). The mutated oligonucleotide pool (mPool) also contained 20 sgRNAs not targeted to pTox and 28 exactly matching sgRNAs as internal controls, for a total of 1140 sgRNAs. The mPool was cloned into pTevSpCas9 and we performed 10 independent transformations into *E. coli* harbouring the pTox plasmid. Each transformation culture was split and then grown under conditions that repressed or induced TevSpCas9 and CcdB. We anticipated that active TevSpCas9/sgRNA combinations would become enriched under the inducing conditions relative to the pool grown under non-induced conditions. Our output score (reported as normalized activity) was the log2 difference in relative sgRNA abundance between the induced and uninduced conditions (Materials and Methods). Given the solid and liquid culture results, we anticipated that the assay would report a distribution of activities that depended on the underlying activity of the sgRNA/pTevSpCas9 combination. After Illumina sequencing of the sgRNA cassette from both conditions and data analyses, active combinations were identified by a higher normalized activity score (Supplementary Table S2).

As expected, when co-expressed with TevSpCas9, sgRNAs that exactly matched their target sequences (black) (Figure 3B and 3C) tended to exhibit high normalized activity scores, sgRNAs with single mismatches to their target site (orange) showed a broad range of activities, and sgRNAs with double mismatches to their target site (blue) generally had low activity scores that were similar to non-targeted sgRNAs (green). Also as expected, the ability of the sgRNA to confer activity was most impacted by mismatches in the seed region corresponding to positions 1-10 relative to the PAM proximal end (Figure 3D)[43,44]. The impact of double transversions was more pronounced than that of single transversions. In the former, mismatches in all positions except 20 and 19 severely reduced activity (Figure 3E), while in the latter there was a broader range of activity conferred (Figure 3D). These results agree with previous studies on mismatch tolerance of Cas9/sgRNA from *in vitro* data and eukaryotic systems[45–47]. The data also show that our experimental system can report a gradient of sgRNA activities across an ~1000-fold normalized activity range and a ~2000-fold range in relative abundance; although the relative abundance range was more clustered except for a few outlier sgRNA sequences.

High-throughput profiling of a pooled sgRNA library

We next synthesized an oligonucleotide pool (oPool) to interrogate the activities of 304 exact match sgRNAs targeted to the pTox plasmid, with all sgRNA sites having a 5'-NGG-3' PAM sequence (Figure 4A, Supplementary Table S3). The oPool also contained 15 sgRNAs with nucleotide mismatches that had varying degrees of target complementarity to the pTox plasmid, and 48 sgRNAs that did not have any complementarity to the pTox plasmid. In addition, 73 of the 304 sgRNAs that exactly matched their target sequence also contained an exact match with a consensus I-TevI cleavage site at the correct spacing from the SpCas9 binding site. In total, the oPool contained 367 sgRNAs (Supplemental Table S3). The oPool was cloned into pTevSpCas9 and pSpCas9, and 10 transformation replicates for each was generated. Following induction and outgrowth, the result was read out by Illumina sequencing and data analysis to assign normalized activity and relative abundance scores for each sgRNA in combination with both SpCas9 and TevSpCas9 (Supplementary Table S4). The major findings from these experiments are:

1. Of the 304 sgRNAs with perfect complementarity to the pTox plasmid, 174 had sig-

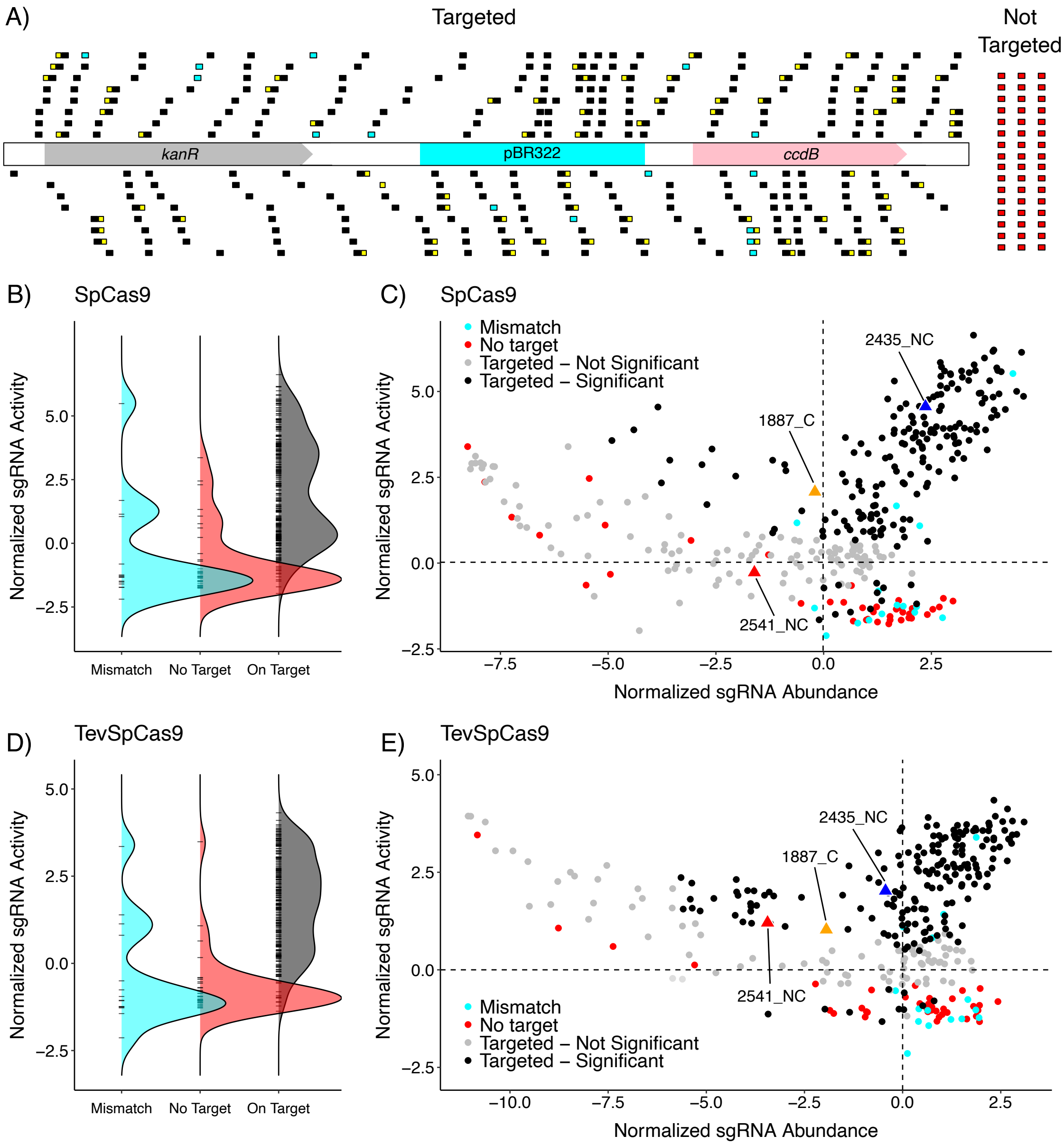

Figure 4. High-throughput pooled screen detects distribution of SpCas9/sgRNA and TevSpCas9/sgRNAs activity. A) Schematic of target sites for the sgRNAs pPool with black boxes representing sgRNA target site (304), cyan boxes representing target sites with mismatches (15), red boxes representing non-targeting sgRNAs (48) and yellow boxes representing TevCas9 sites (75). **B) and D)** Distribution of normalized activity scores for mismatched (cyan), non targeting (red), and on target (black) sgRNAs for Cas9 (B) and TevCas9 (D) experiments. **C and E)** Bland-Altmann plots comparing the normalized abundance and activity scores for individual sgRNAs in the Cas9 (C) and TevCas9 (E) pooled experiments. sgRNAs with a false-discovery rate (FDR) < 0.01 are highlighted black and sgRNAs with a FDR > 0.01 are coloured grey. Cyan and red points represent mismatched sgRNAs and non-targeting sgRNAs respectively. sgRNAs that were tested individually in Figure 2B and 2C are shown as triangles where 2435_NC is blue, 1887_C is orange and 2541_NC is red.

nificant positive normalized activity scores in the SpCas9 data set and 178 in the TevSpCas9 data set using an FDR < 0.01 (Figure 4C and 4E).

2. The non-targeted (red) and mismatched (cyan) sgRNAs generally had negative normalized activity scores indicating that they did not cleave the pTox plasmid efficiently (Figure 4B-E).
3. We found no nucleotide preference in the first position of the 5'-NGG-3' PAM for either SpCas9 or TevSpCas9 (Supplementary Figure X).
4. sgRNA relative abundance alone was misleading as a measure of activity as the vast majority of sgRNA sequences were highly abundant, and both mismatched and non-targeted sgRNA sequences tended to be more abundant than average (Figure 4C and 4E).

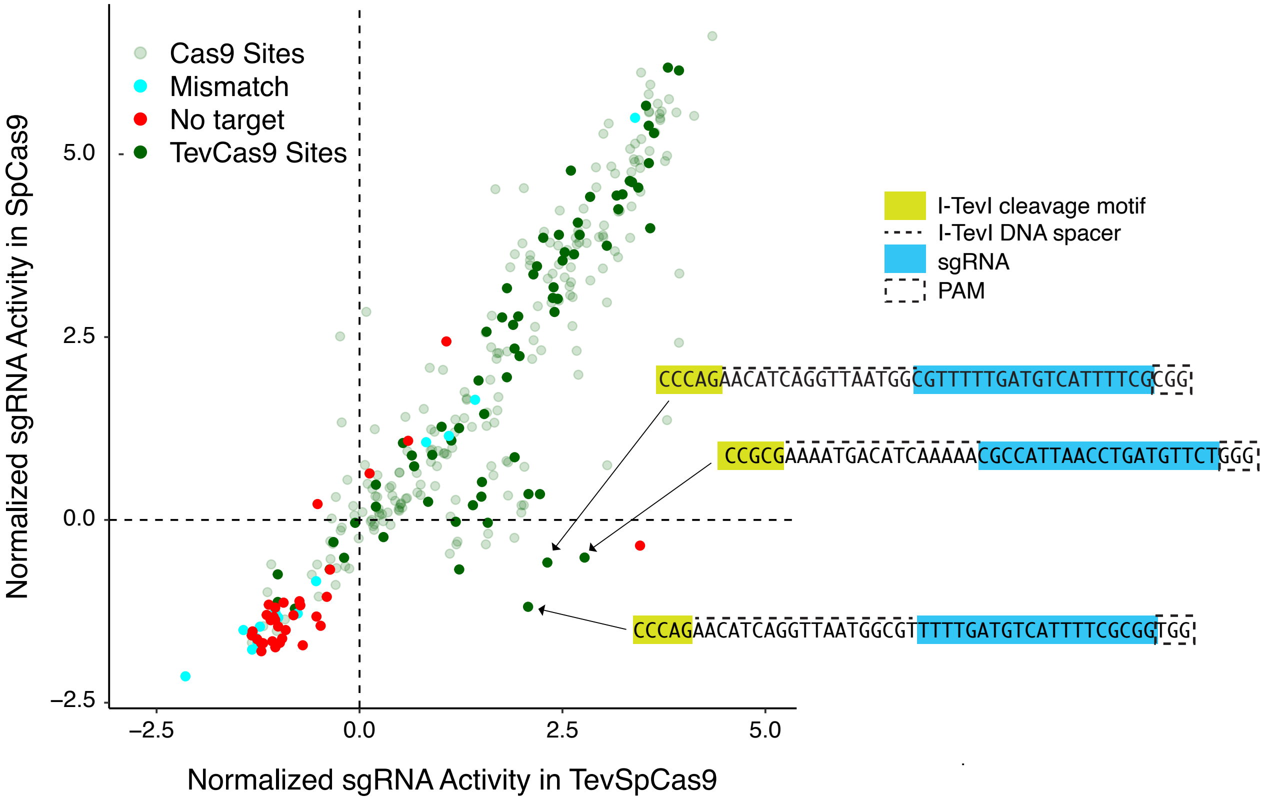

Figure 5. Activity of TevSpCas9 versus SpCas9 with the pooled sgRNA library. Comparing the difference between condition values for sgRNAs present in both TevSpCas9 and SpCas9 pooled experiments where dark green dots represent sgRNAs with upstream I-TevI recognition sites and light green dots representing sgRNAs with Cas9 sites only. Non-targeting and mismatched sgRNAs are highlighted as red and cyan respectively. Three sgRNAs that target TevSpCas9 sites are indicated.

One interesting finding from the oPool experiment was the activity of sgRNAs in the SpCas9 versus the TevSpCas9 experiment. Overall, the readouts from the same sgRNAs in both assays behaved similarly (Figure 5, Pearson correlation 0.90, p-value < 2.2 x 10^{-16}), but we found 22 sgRNAs that promoted higher activity with TevSpCas9 than with SpCas9.

In the SpCas9 experiment, these sgRNAs had low normalized activity scores ranging from -1.21 to 1.45 versus -1.01 to 2.77 in the TevSpCas9 experiment. The single non-targeted sgRNA (NT42) with a high activity of 3.4 in the TevSpCas9 experiment also showed high replicate-to-replicate variability suggesting that this was an outlier (Supplemental Table 4). One explanation for the increased activity of sgRNAs in the TevSpCas9 experiment was the presence of the I-TevI 5'-CNNNG-3' cleavage motif at an appropriate distance upstream of the sgRNA binding site (Figure 2B and Figure 5). This observation suggests that SpCas9 binding is necessary but not sufficient for cleavage, and that low SpCas9 cleavage can be rescued by the I-TevI nuclease domain to promote elimination of the pTox plasmid.

We also noted a large dynamic range for the normalized activity scores (∼1000-fold) and relative abundances of the sgRNA sequences (∼2000-fold) (Figure 4C and 4E). The dynamic range allowed us to identify sgRNAs with low abundances but large activity scores (upper left quadrant of Figure 4C and 4E). Conversely, we identified sgRNAs with high abundance but negative activity scores (lower right quadrants of Figure 4C and 4E); 58.7% and 73.5% of these sgRNAs are non-targeting (red) or mismatched (cyan) guides with respect to the pTox plasmid. Taken together, the data highlight the importance of conducting an experiment where the paired design allows the readout of relative enrichment with multiple replicates to accurately measure the ability of sgRNA to confer activity on the complex. Moreover, the approach demonstrates that using sgRNA relative abundance alone as an indicator of activity can lead to false identification of the abilty of sgRNAs to confer activity.

Growth curves of individual sgRNAs identifies toxic guides

The pooled sgRNA experiments in Figures 3 and 4 revealed a wide range of sgRNA activity. To cross validate these activity measurements we blindly picked 77 colonies from the transformed pTevSpCas9/sgRNA-oPool library to test using individual growth experiments as shown in Figure 2D; the identity of each sgRNA was confirmed by sequencing of isolated plasmids. We rationalized that growth curves performed with individual sgRNAs would better resolve measure the properties of sgRNA species independent of their behaviour in a sgRNA pool where we could only measure relative changes. These experiments were performed when the TevSpCas9 protein and the CcdB proteins were induced or repressed, and we found three different classes of sgRNA sequences (Figure 6A-C, Supplementary Figure X). Those sgRNAs that conferred a high level of activity when complexed with TevSpCas9 (20 of 77) grew in both induced and repressed conditions (Figure 6A) whereas inactive sgRNAs (12 of 77) only grew in the repressed condition (Figure 6B). Surprisingly, we found a number of sgRNAs that we classified as toxic (12 of 77) because they grew poorly in both the induced and repressed condition (Figure 6C) as compared to a non-targeting sgRNA (Figure 6D). The growth curves for the remaining 33 sgRNAs did not clearly fit in any category but showed intermediate activity. For active sgRNAs, we consistently found that maximal optical density values were lower in the induced than the repressed condition. We attribute this difference to the presence of glucose in the media used for the repressed condition, which is a preferred carbon source to the arabinose present in the media for the induced condition.

For each sgRNA, we calculated the area under the curve (AUC) for the induced and repressed conditions and normalized them relative to the average AUC for all sgRNAs for each condition (Figure 6E, Supplementary Table S5). This plot emphasizes that many guides conferring activity grew well in both induced and repressed conditions (20 of 77, black dots Figure 6E). Conversely, a subset of sgRNAs showed poor or no growth in induced conditions, but robust growth in repressed conditions, and thus were considered inactive (12 of 77, red dots in Figure 6E), although there was no clear separation between these two groups. This analysis also revealed that toxic sgRNAs grew poorly in both repressed conditions and induced conditions (12 of 77, cyan dots Figure 6E). We considered that toxicity could be due to off-target sgRNA sites in the *E. coli* genome, however none of these sgRNAs have off-target sites with 3 or fewer mismatches. This suggests that toxicity is either an intrinsic property of the sgRNA or that these sgRNAs confer some unwanted property on the TevSpCas9 protein when complexed with the toxic sgRNA

To address parallels between individual and pooled experiments, we mapped the different classes of sgRNAs from the growth experiments back to the analyses of the deep sequencing experiments (Figure 6F and 6G). This revealed that sgRNAs that were classed as inactive in the growth experiments had poor activity in the pooled experiments, with a mean normalized activity score of -0.309 and relative abundance value that were suggestive of minimal or modest enrichment (Figure 6F). In contrast, sgRNAs conferring activity in the growth experiments largely had positive activity scores and relative abundance values (Figure 6F). Interestingly, guides determine to be toxic by the growth curves had activity scores ranging from -1.33 to 1.81 in the pooled experiment (6F, mean value of -0.0431) and many of these sgRNAs had positive relative abundance values (Figure 6F). One explanation for this apparent discrepancy between toxicity and activity is that toxic sgRNAs vary in how they promote bacterial growth in the repressed and induced conditions. For instance, a toxic sgRNA may still be active on the intended pTox plasmid target site under inducing conditions (thus promoting growth), but show toxicity under repressive conditions (thus preventing growth) in turn altering the relative difference calculation that is used to infer activity.

Collectively, these data emphasize the importance of independent validation of sgRNA activity using different methods of activity assessment. Our analyses revealed that many sgRNAs that would be considered active solely by their relative abundance in deep sequencing experiments demonstrated high levels of toxicity when analyzed individually. Thus, toxicity and high activity are not mutually exclusive, but cannot be distinguished if sgRNAs are classified based on a single line of experimental evidence. Further studies are needed to directly identify toxic sgRNAs and determine the mechanism of toxicity.

Transfer learning is required for suitable TevSpCas9 predictive ability

With this data in hand, we next concentrated on building a model, crisprHAL (Figure 7A, Supplementary Table S6), that could more accurately predict sgRNA target site sequence-associated TevSpCas9 and SpCas9 activity in *E. coli*. For this, we constructed a dual branch deep learning architecture utilizing transfer learning. To select our model architecture and evaluate transfer learning performance, we used 5-fold cross validation, measured by Spearman ranked correlation coefficient, hereafter referred to as rank correlation.

Our initial model tests used only the TevSpCas9 dataset which, unsurprisingly given the dataset size, resulted in poor performance, with a rank correlation of 0.308 (Figure 7B). Thus, we chose to pursue transfer learning to improve performance and constructed new base models on SpCas9 (n=40308) and eSpCas9 (n=45010) datasets derived from an

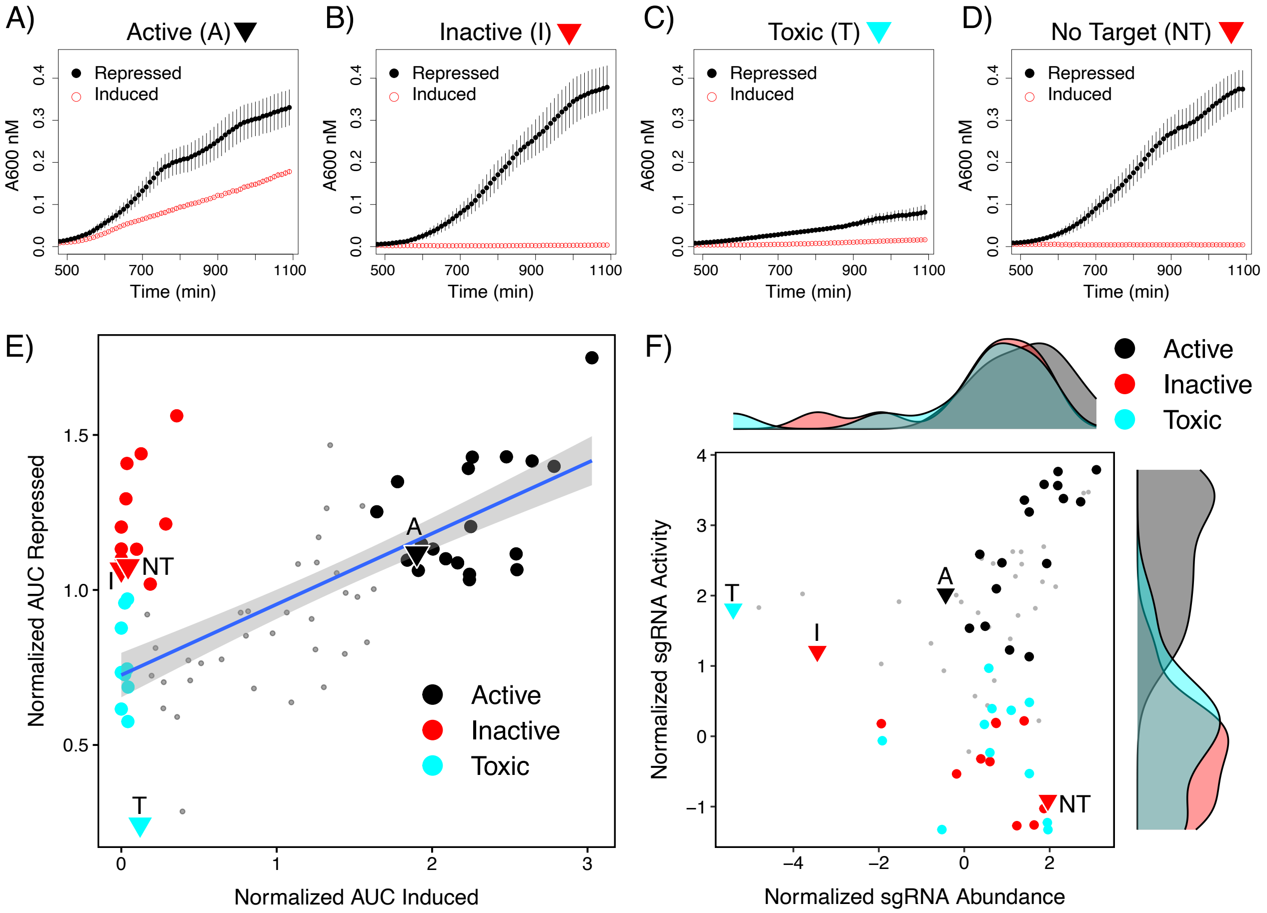

Figure 6. Assaying sgRNAs individually identifies distinct phenotypes. Representative growth curves of active **(A)**, inactive **(B)**, toxic **(C)** and non-targeted **(D)** sgRNAs under induced (red dots and line) and repressed (black dots and line) conditions. Points are the mean of three biological replicates and whiskers represent the mean plus or minus the standard deviation. Growth curves for all tested sgRNAs are in Supplemental Figure 4 and Supplementary Table 5. **E)** Plot of AUC for induced and repressed conditions for all sgRNAs. sgRNAs were classified as active (black dots, AUC > 1.64) or toxic (cyan dots, AUC < 0.121) based on quantiles of the AUC for the induced condition. Gray points are sgRNAs with an intermediate phenotype. **F)** Marginal density plot of normalized activity and abundance for each sgRNA using data shown in Figure 4E. Gray points are sgRNAs with an intermediate phenotype.

sgRNA depletion experiment in *E. coli*[28]. Hereafter, this data is referred to as the Guo SpCas9 or Guo eSpCas9 dataset to distinguish it from our SpCas9 dataset generated in this study. A major difference of our model compared to prior models is the use of a log ratio-based relative difference metric for scoring sgRNA-associated nuclease activity[48]. This scoring method resulted in mean scores near 0 across all datasets, however, we noted differences in their dynamic ranges. The Guo SpCas9 and eSpCas9 datasets contained the widest range of scores, with standard deviations of 2.280 and 2.492, respectively. The TevSpCas9 dataset contained the smallest range, with a standard deviation of 1.305. To compensate for the variations in activity score ranges, we standardized the scores for each dataset by dividing by the standard deviation (Supplementary Table S6). This improved

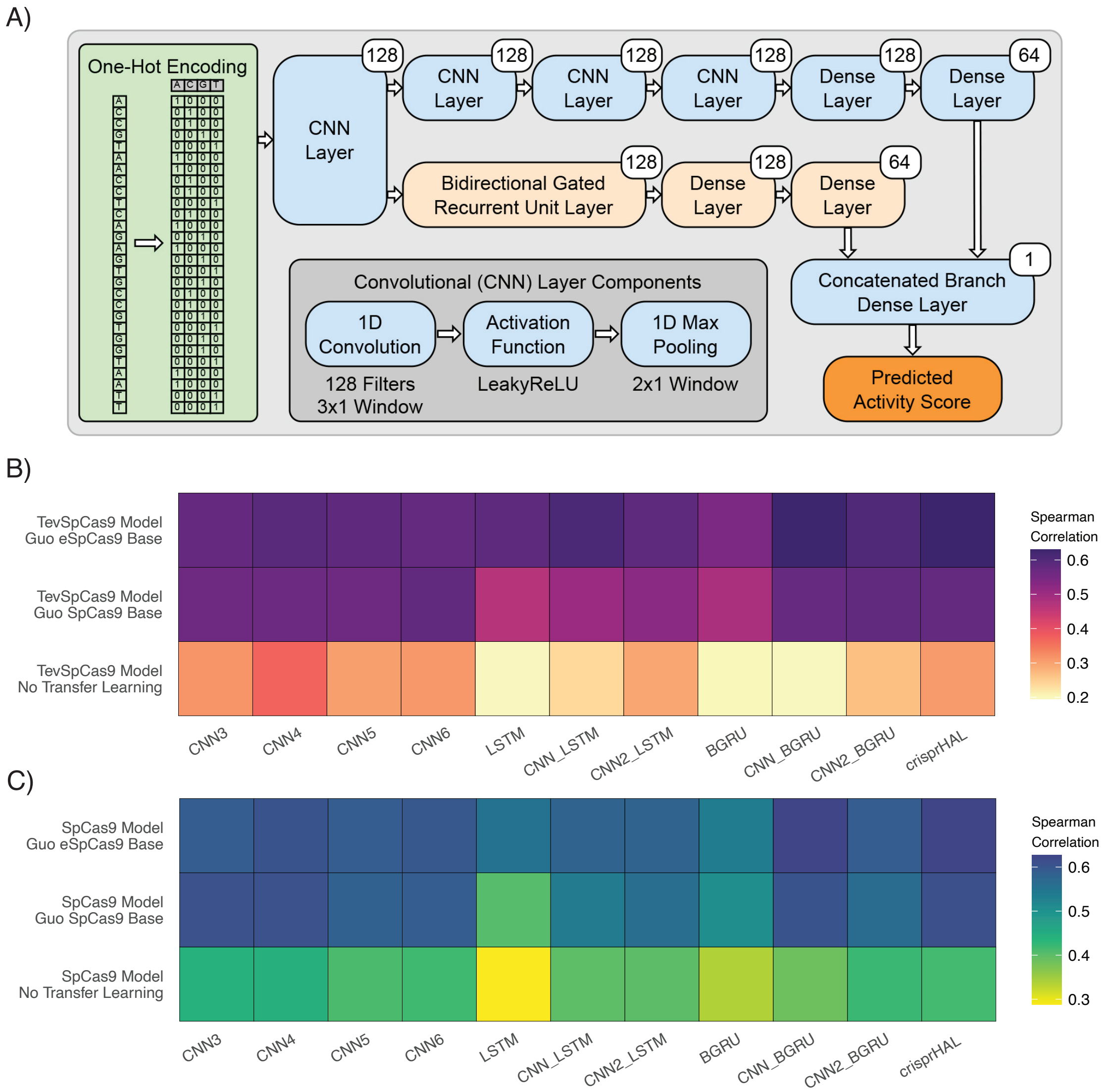

Figure 7. crisprHAL model architecture and branch architecture tests. A) Model architecture of crisprHAL showing the one-hot encoding of an input sgRNA sequence (green), the dual branch 4-layer CNN and hybrid CNN-BGRU RNN architecture with frozen layers (blue) and unfrozen layers (light orange), and the output predicted activity score (dark orange). The first CNN layer is connected to both branches of the model. The number of neurons per layer are indicated in the top right corner of each layer. Testing model performance on **B)** TevSpCas9 (n=279) and **C)** our SpCas9 (n=302) datasets without transfer learning versus transfer learning from a base model constructed on the Guo SpCas9 or eSpCas9 dataset. Model architectures tested are: 3, 4, 5, and 6 layer CNNs; long-short term memory RNN (LSTM) or bidirectional gated recurrent unit RNN with 0, 1, or 2 preceding CNN layers; and crisprHAL – composed of the two best branches – a 4 layer CNN and a hybrid CNN-BGRU which share the first CNN layer. All architectures contain two fully connected dense layers following their respective CNN and/or RNN layers. Performance is measured by average rank correlation from 5-fold cross validation.

transfer learning performance since each dataset was on the same scale. Base model performance was unaltered as expected since this is a simple linear scaling.

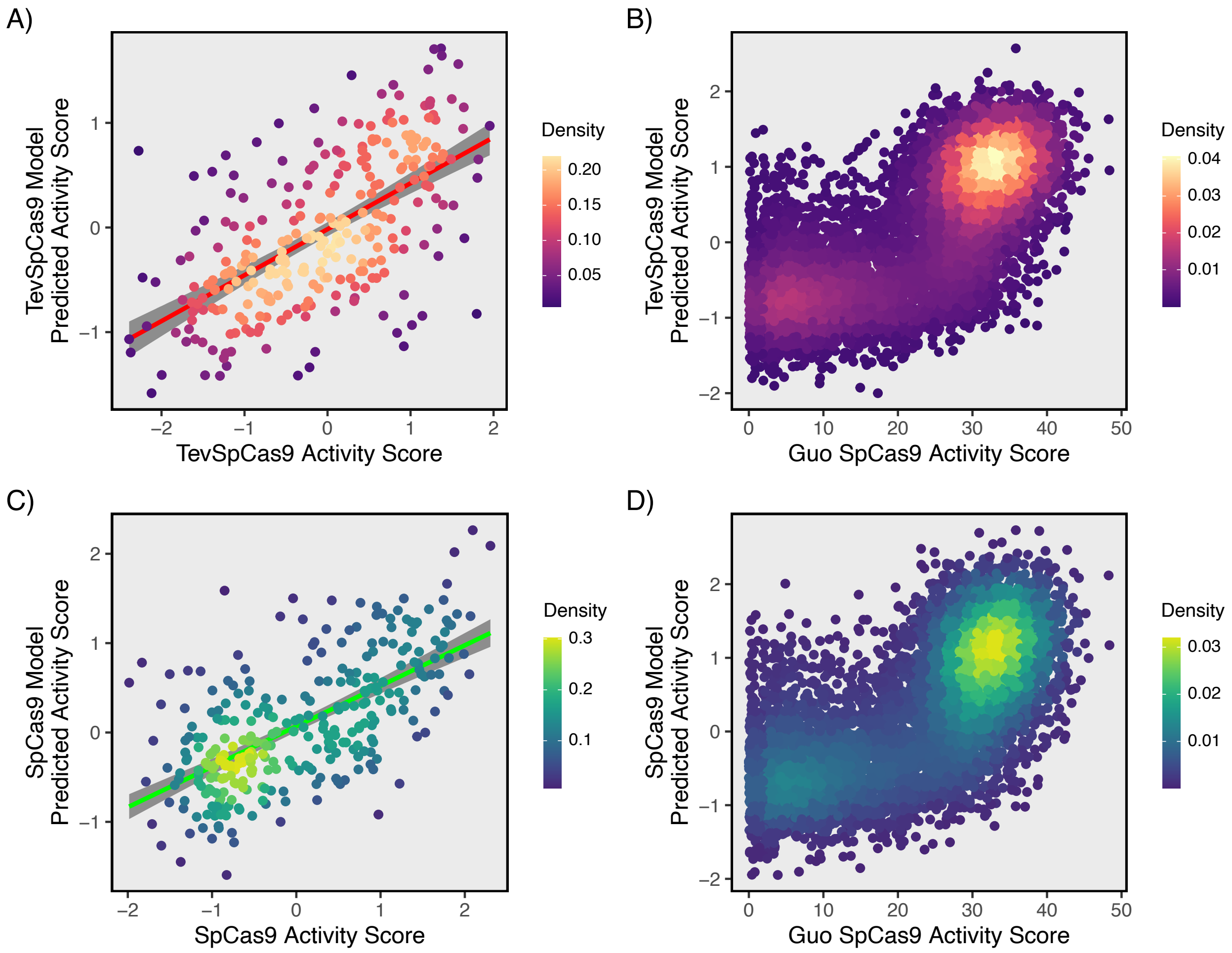

Figure 8. Predictive performance of crisprHAL. A) Correlation between the TevSpCas9 dataset (n=279) and the TevSpCas9 model 5-fold cross validation predictions (mean rank correlation of 0.630 across 5-folds). **B)** Correlation between the unique Guo SpCas9 dataset (n=7821) and predictions from the TevSpCas9 model (rank correlation of 0.682). **C)** Correlation between the SpCas9 dataset (n=302) and the SpCas9 model 5-fold cross validation predictions (mean rank correlation of 0.627 across 5-folds). **D)** Correlation between the unique Guo SpCas9 dataset and predictions from the SpCas9 model (rank correlation of 0.657). Model predictions are compared to original Z-score sgRNA activity scores for the unique Guo SpCas9 dataset. In panels A and C the linear line of best fit and standard error of the fit are shown.

Following base model construction, we tested variations in freezing parameter weights for specific layers within the model to optimize for transfer learning performance. We found that freezing the multi-layer CNN branch and leaving all layers of the CNN-BGRU branch – except for the initial CNN layer shared by both branches of the model and the final output layer – resulted in the best performance, as shown in Figure 7B. Additionally, the performance of the model was higher when the final layers of the model, which concatenate the outputs of both branches of the model, were frozen.

We found that base models constructed on the Guo eSpCas9 dataset had better transfer learning performance than base models constructed with the Guo SpCas9 dataset. TevSpCas9 average 5-fold cross validation performance improved by 0.053 rank correlation when transfer learning with the eSpCas9 base model versus the Guo SpCas9 base model (Figure 7B). We found that crisprHAL with a dual branch architecture performed well on our TevSpCas9 dataset (n=279) after transfer learning from a base model built on eSpCas9 data (n=45010), with a rank correlation of 0.630 (Figure 7B). This exceeds the best prior bacterial model, built for SpCas9 by Guo et al., which had a rank correlation in this dataset of 0.52 (Figure 1A). Within the 4 prior bacterial models tested, we found that the gradient boosting regression tree (GBR) models for SpCas9 and eSpCas9 by Guo et al. generalize to TevSpCas9 data better than the deep learning based models for SpCas9 and eSpCas9 from DeepSgRNA, respectively[28,34]. This contrasts with the improved performance from the deep learning models versus the GBR models on their own SpCas9 and eSpCas9 data; both Guo and DeepSgRNA construct their models on the same data.

To validate model performance, we tested crisprHAL on a set of unique sequences from the Guo et al. SpCas9 dataset (n=8728)[28]. This set of sequences was curated to remove any overlap with the Guo eSpCas9 dataset used to construct the base model, a process which removed 36342 sgRNAs. As shown in Figure 8B, crisprHAL performs well on this dataset with a rank correlation of 0.682.

Applying transfer learning model to the SpCas9 dataset

We also tested the model with the SpCas9 dataset (n=303) in place of TevSpCas9 for transfer learning from the eSpCas9 base model while leaving all other aspects of the model unaltered. The model performs well with the SpCas9 data, resulting in a 5-fold cross validation average rank correlation of 0.627 (Figure 8C). This performance exceeds that of all existing models, with the best prior model, built for SpCas9 by Guo et al., attaining a rank correlation of 0.456 (Figure 1B)[28,34]. We noted transfer learning to be an essential component of the SpCas9 performance. Without transfer learning the model reaches a rank correlation of only 0.417.

In line with our TevSpCas9 model performance, our SpCas9 model performs best when using the dual branch model, with performance only marginally exceeding that of the hybrid CNN-RNN alone (Figure 7C). Although performance is optimal when using the eSpCas9 dataset base model, we found models used our SpCas9 dataset to perform well when transfer learning from the Guo SpCas9 dataset base model, with a 5-fold cross validation average rank correlation of 0.609, notably higher than the TevSpCas9 results (Fig 7B-C). When testing SpCas9 model generalization, we found it to perform well on the unique Guo SpCas9 dataset, with a rank correlation of 0.657 (Figure 8D)[28]. Testing could not be performed with the TevSpCas9 dataset due to the presence of all sgRNAs being cross-listed.

Downstream target site nucleotides impact predictive performance

Prior models have proposed various input sequence lengths for optimal predictive performance[28,34]. For example, the DeepSgRNABacteria model suggests that 43nt input sequences may be optimal for eSpCas9 and SpCas9 performance based on calculated importance scores, with nucleotides downstream of the sgRNA binding site containing more information than upstream nucleotides[34]. Biologically, it is implausible that sequences outside of those contacted by either the sgRNA or the SpCas9 nuclease[49–53] should have a large effect on a machine learning model, and these models may thus be overparameterized.

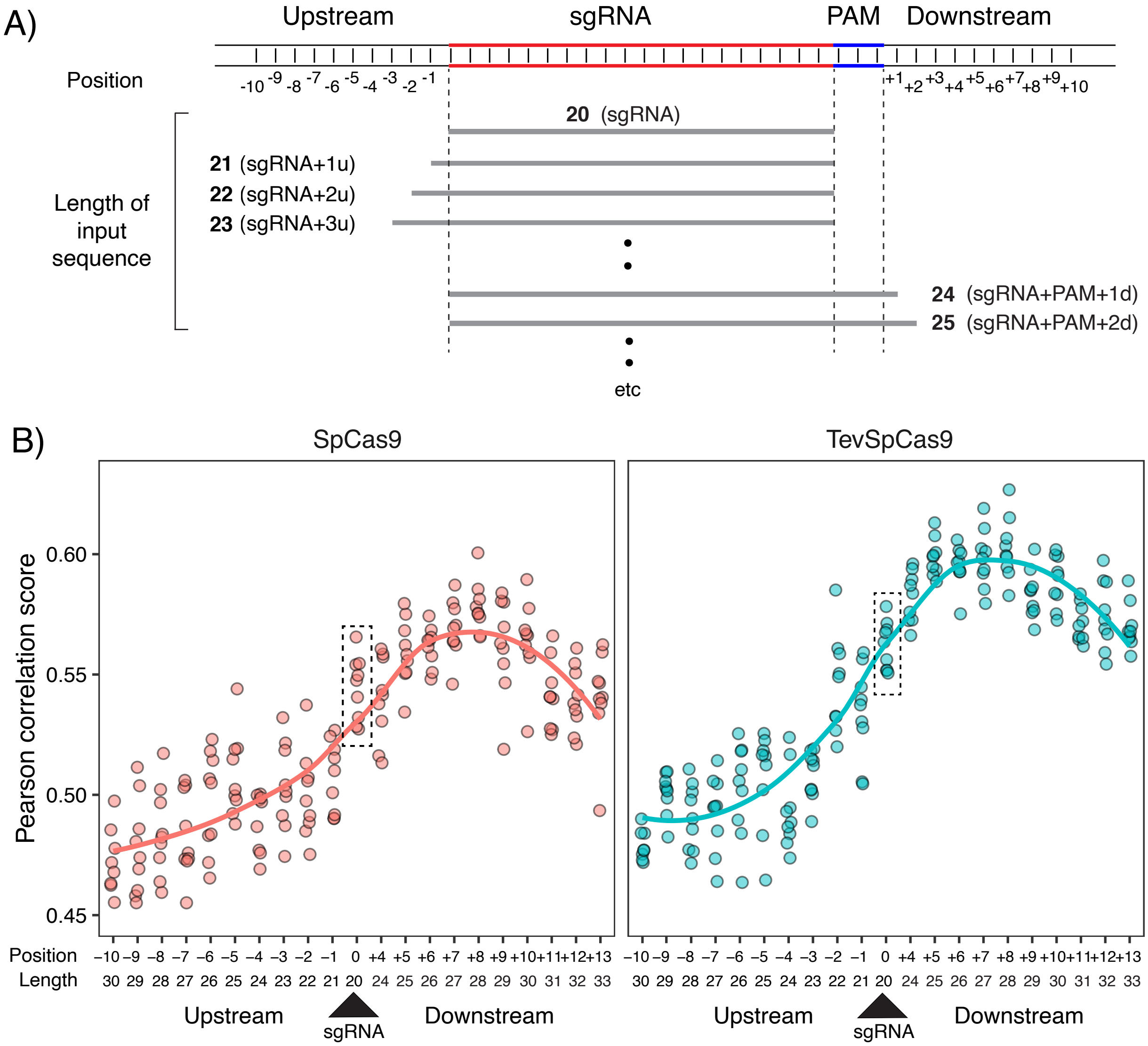

Figure 9. Impact of sgRNA target site sequence length on model performance. A) Length of sgRNA target site sequences tested as model inputs. All tested sequences include the 20nt sgRNA target site. Upstream (-) tests extend the model input sequence by 1nt 10 times. Downstream (+) tests begin with the addition of the NGG PAM to the model input, then extend by 1nt 10 times. **B)** TevSpCas9 model (left) and SpCas9 model (right) performance on their respective dataset versus the sgRNA sequence length input. Performance is measured by average rank correlation produced by 5-fold cross validation. Both models utilize transfer learning from our base model trained on the eSpCas9 dataset. The boxed regions in **B)** represent the 20 nt sgRNA sequence.

To identify the optimal input sequence length to use for our model, we constructed versions of our model with input sequences extending upstream and downstream of the 20nt sgRNA target site (Figure 9A). The 20-nt base input was extended upstream 10 positions in 1-nt increments upstream and downstream 11 positions in 1-nt increments. A single increment of 3 nt covered the PAM sequence. Predictive performance of these incremental models was measured by 5-fold CV across the TevSpCas9 dataset using rank correlation (Figure 9B).

We noted that nucleotide additions upstream of the sgRNA target site immediately decreased the predictive ability of the model (Figure 9B). In contrast, nucleotide additions downstream of the sgRNA target site improved predictive performance, up to the limit of 8nt downstream. Based upon these results we chose an input sequence of 28nt, comprising the 20 nt sgRNA target site, the 3nt PAM, and 5 additional downstream nucleotides. No upstream nucleotides were included in our input sequence for the crisprHAL model.

DISCUSSION

Although targeting SpCas9/sgRNA to desired sequences in small-sized bacterial genomes appears straightforward because it relies on apparent nucleotide complementarity, there are significant limitations in our ability to reliably identify highly active SpCas9/sgRNA combinations. Ideally, a predictive model of SpCas9/sgRNA activity should be agnostic to different datasets, generalize to different organismal systems, and recapitulate the known biology of SpCas9/sgRNA target interactions. Current prediction models do not meet all of these criteria. Here, we identify three areas that improve computational models of SpCas9/sgRNA activity; collection of biological data that accurately assesses SpCas9/sgRNA cleavage, appropriate treatment of high-througput Illumina data for model training, and machine transfer learning to capitalize on existing and additional datasets.

Accurate computational predictions of sgRNA activity rely on biological data that reports on SpCas9/sgRNA cleavage activity and not secondary outcomes of DNA cleavage. This is particularly relevant in mammalian systems where many Cas9/sgRNA datasets report on non-homologous end joining (NHEJ) DNA repair outcomes of cleavage rather than directly assessing Cas9 cleavage. While bacteria generally lack NHEJ pathways, Cas9/sgRNA cleavage can be enhanced in *recA* deficient strains, or strains expressing dominant negative *recA* variants, to suppress DNA repair through the SOS response[29]. Our strategy to assess SpCas9/sgRNA cleavage was to use a two-plasmid enrichment assay with pooled libraries read out by Illumina sequencing that agrees well with the kinetics of *in vitro* DNA cleavage[42,54]. Crucially, this system also helps distinguish SpCas9/sgRNA on-target activity from toxicity that can be a confounding issue in bacterial systems where overexpression of Cas9 (or dCas9) can cause cellular toxicity, or at the very least to reduce the growth rate significantly. With the enrichment assay, we found that about one in seven SpCas9/sgRNA combinations showed evidence of toxicity. Some toxicity is likely to due to SpCas9/sgRNA cleavage of the bacterial chromosome at off-target sites, which can be avoided by selection of appropriate sgRNAs[8]. It is also possible that toxicity could result from partial matches between the sgRNA and functionally critical genes on the chromosome that preclude DNA cleavage but facilitate transcriptional repression[55,56]. In large-scale pooled sgRNA depletion experiments, these sgRNAs would mistakenly be classified as having high on-target activity and add noise when training machine learning models. In more directed applications like the one implemented in this study, these sgRNAs are easily identifiable and no longer reported as false positives.

Significantly, we found that the small amounts of high quality data generated in this study that while insufficient to train a model on their own, improved sgRNA activity predictions with the new crisprHAL model that utilized machine transfer learning as well as a unique dual-brand CNN and RNN architecture. Our work corroborates prior findings that hybrid CNN-RNN architectures are well suited for transfer learning[57–59]. We found that the multi-layer CNN was the primary contributor to base model performance on the eSpCas9 data, reaffirming its use by models such as DeepSgRNA[34]. Inclusion of this multi-layer CNN branch, in addition to the hybrid CNN-BGRU, improved base model performance on eSpCas9 while retaining transfer learning capacity. Our dual branch structure provided a significant boost to model generalization performance on the unique Guo SpCas9 dataset as compared to the hybrid CNN-BGRU only architecture. Additionally, since all parameters in the multi-layer CNN and branch concatenation layers were frozen, nullification of the multi-layer CNN branch's contribution to the output prediction was unlikely. We attained the best model performance when using the same scoring method across the datasets, while compensating for variations in dynamic range through scaling by the standard deviation of scores from each dataset. With this treatment, crisprHAL predictions showed a linear correlation with measured sgRNA activity, Moreover, we did not utilize negative control sgRNAs in our process for sgRNA activity score calculations. Given that our eSpCas9 base model performs at least as well as the prior Guo and DeepSgRNA models constructed on that dataset, we suggest that negative control sgRNAs are unnecessary for the scoring of SpCas9/sgRNA activity.

One parameter for model inclusion that we explored in detail was the length of up- and down-stream sequence flanking the 20-nt sgRNA target site. Inclusion of flanking DNA sequence in prior models was justified by factors such as chromatin accessibility, consideration of DNA unwinding, and Cas9 activity data that indicated nucleotide preference in flanking regions (although it is possible this reflects DNA repair and not Cas9 cleavage preference). However, outside of the 20-nt sgRNA-target strand interaction, Cas9-target DNA contacts occur exclusively downstream of the PAM sequence, including a transient interaction 14-nt downstream that impacts binding and dissociation[49–53]. Thus, the biological data argue against inclusion of upstream DNA sequences. Indeed, our data show the best model performance with a 28-nt input sequence that includes the sgRNA binding site, the PAM and 5 downstream nucleotides, and that inclusion of upstream sequence is uninformative.

In summary, we have generated novel datasets for the activity of several hundred SpCas9/sgRNA and TevSpCas9/sgRNA combinations in a bacterial environment. The experimental setup detects activity over a large dynamic range and is able to distinguish toxicity from on-target cleavage activity. The datasets were then used in conjunction with machine transfer learning and novel model architecture to produce crisprHAL, the most accurate TevSpCas9 and SpCas9 activity prediction model for bacteria to date. Our results show that small amounts of high-quality data can improve predictions of sgRNA activity and represent a step towards a generalizable model for bacteria. In principle, the approach outlined here to improve sgRNA activity predictions could be applied to any biological system where it is possible to collect high-quality Cas9/sgRNA cleavage data.

Materials and Methods

Bacterial strains

E. coli EPI300 (F'λ^-mcrAΔ(mrr-hsdRMS-mcrBC)ϕ80d*lacZ*Δ*M15*Δ*(lac)X74 recA1 endA1 araD139*Δ*(ara,leu)7697 galU galK rpsL* (Str^R) *nupG trfA dhfr*) (Epicenter) was used for cloning the sgRNA pools. Screening sgRNA activity was done in NEB 5-alpha F'I^q *E.*

coli (F' *proA*$^+$*B*$^+$ *lacI*^{q}*Δ(lacZ)M15 zzf::Tn10* (Tet^R) */fhuA2Δ(argF-lacZ)U169 phoA glnV44 ϕ80Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*) strain harbouring pTox.

Construction of sgRNA pools

The pTox plasmid was screened for 5'-NGG-3' PAM sequences in a unique 3.2kb region that included the kanamycin acetlytransferase (*kan*^{R}) coding region, the pBR322 origin of replication, and the *ccdB* DNA gyrase toxin coding region. The DNA sequence corresponding to 20 nts upstream of each PAM sequence was computationally extracted to create a pool of 304 sgRNA with exact matches to pTox. Sequences that contain BsaI-HF-V2 restriction sites that generate proper overhangs for Golden Gate Cloning were added to the ends of the sgRNA sequences for subsequent cloning. The sequence 5'-CCTGGTTCTTGGTCTCTCACG-3' was added upstream of the sgRNA and 5'-GTTTTAGAGACCGCTGCCAGTTCATTTCTTAGGG-3' was added downstream to allow for efficient and directional cloning. The pool of 304 sgRNAs (oPool), including 15 sgRNAs with internal mismatches and 48 non-targeting sgRNAs, was ordered as single-stranded fragments at 1 pmol/oligo from Integrated DNA technologies (IDT) (Supplementary Tables S1 and S3). To create the library of sgRNAs with nucleotide transversions (mPool), 28 sgRNAs were picked and single and double nucleotide transversions were tiled along the length of each oligo. BsaI sites were added computationally, and ordered from IDT as above. For each library, second strand synthesis was performed using 1 μg of single stranded pool DNA and equimolar amounts of primer DE-5224 in NEB buffer 2 (50 nM NaCl, 10 mM Tris-HCl, 10 mM MgCl$_2$, 1 mM DTT, pH 7.9) by denaturing at 94°C for 5 minutes. Primers were annealed by decreasing temperature 0.1°C/second to 56°C and holding for 5 minutes, and followed by decreasing temperature 0.1°C/second to 37°C. To the annealed oligonucleotides, 1 μL of Klenow polymerase (New England Biolabs) and 1 μL of 10 mM dNTPs were added and incubated for 1 hour at 37°C, followed by a 20 minute incubation at 75°C before being held at 4°C. The resulting dsDNA fragments were purified using a Zymogen DNA Clean & Concentrator-5 kit following manufacturer specifications. Golden Gate cloning was used to clone the oPool and mPool into SpCas9 and TevSpCas9 by combining 6 pmol of oPool or mPool, 100 ng of backbone plasmid, 0.002 mg BSA, 2 μL T4 DNA ligase buffer (50 mM Tris-HCl, 10 mM MgCl$_2$, 1 mM ATP, 10 mM DTT, pH 7.5), 160 units T4 DNA ligase (New England Biolabs) and 20 units of BsaI-HF-V2 (New England Biolabs) with the following thermocycler conditions: 37°C for 5 min then 22°C for 5 min for 10 cycles, 37°C for 30 min, 80°C for 20 min, 12°C inf. The resulting pool was then transformed by heatshock into *E. coli* EPI300 and plated on LB plates (10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride, 1% agar) supplemented with 25 mg/mL chloramphenicol and 0.2% w/v D-glucose.

Pooled sgRNA two-plasmid enrichment experiment

A two-plasmid enrichment experiment was used to assay sgRNA activity as previously described[39,40]. For liquid selections, 50 ng of the sgRNA plasmid pool was transformed into 50 μL *E. coli* NEB 5-alpha F'I^q competent cells harbouring pTox by heat shock. Cells were allowed to recover in 1 mL of non-selective 2xYT media (16 g/L, 10 g/L yeast extract, and 5 g/L NaCl) for 30 minutes at 37°C with shaking at 225 rpm. The recovery was then split and 500 μL was added to 500 μL of inducing 2xYT (0.04% (w/v) L-arabinose and 50 mg/mL chloramphenicol) or to 500 μL of repressive 2xYT (0.4% (w/v) D-glucose and 50 mg/mL chloramphenicol) and incubated for 90 min at 37°C with shaking at 225 rpm. The two cultures were washed with 1 mL of inducing media (1x M9, 0.8% (w/v) tryptone, 1% v/v glycerol, 1 mM $MgSO_4$, 1mM $CaCl_2$, 0.2% (w/v) thiamine, 10 mg/mL tetracycline, 25 mg/mL chloramphenicol, 0.4 mM IPTG) or repressed media (1x M9, 0.8% (w/v) tryptone, 1% v/v glycerol, 1 mM $MgSO_4$, 1mM $CaCl_2$, 0.2% (w/v) thiamine, 10 mg/mL tetracycline, 25 mg/mL chloramphenicol, 0.2% (w/v) D-glucose) respectively before addition to 50 mL of the same media that was used in the wash in a 250 mL baffled flask. These cells were grown overnight at 37°C with shaking at 225 rpm. Plasmids were then isolated using the Monarch Plasmid Miniprep Kit (NEB) according to manufactuerers specifications. The sgRNA locus was then PCR amplified using primers (Supplementary Table S7) containing Ilumina adapter sequence, 4 random nucleotides, 12-mer barcodes to specify the replicate, and plasmid-specific nucleotides at the 3' end. The resulting amplicons were sent for 150 bp paired-end Illumina MiSeq sequencing at the London Regional Genomics Center (London, ON).

Growth-curve experiments with individual sgRNAs

The pool of cells containing pTevSpCas9+sgRNA was grown overnight in selective LB (25 mg/mL chloramphenicol and 0.2% (w/v) D-glucose), diluted and plated on agar plates (10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride, 1.5% agar (w/v) supplemented with 25 /mL chloramphenicol and 0.2% w/v D-glucose). Individual colonies were selected and grown overnight in selective LB (25 mg/mL chloramphenicol and 0.2% (w/v) D-glucose) before plasmids were isolated using the Monarch Plasmid Miniprep Kit (NEB) according to manufactuer specificatons. The sgRNA locus of each plasmid was Sanger sequenced at London Regional Genomics Center (London, ON) to determine the sgRNA identity. In three independent transformations, 20 ng of each plasmid, isolated oPool DNA, and pTevSpCas9 with no sgRNA were transformed into 20 μL *E. coli* competent NEB 5-alpha F'I^q cells harbouring the pTox. Cells were allowed to recover in 1 mL of non-selective 2xYT media (16 g/L, 10 g/L yeast extract, and 5 g/L NaCl) for 30 minutes at 37°C with shaking at 225 rpm. The recovery was then split and 500 μL was added to 500 μL of inducing 2xYT (0.04% (w/v) L-arabinose and 50 mg/mL chloramphenicol) or to 500 μL of repressive 2xYT (0.4% (w/v) D-glucose and 50 mg/mL chloramphenicol) and incubated for 40 min at 37°C with shaking at 225 rpm. These cultures were then plated on inducing or repressing M9 plates and grown overnight at 37°C. At the same time, 20 μL was added to 180 μL of inducing and repressing M9 liquid media in a 96-well plate for growth curves. Plates were grown at 37°C in the BioTek Epoch 2 Microplate Spectrophotometer measuring the absorbance at 600 nm every 10 minutes for 18 hours with double orbital shaking. Raw data was collected, processed, and analyzed using the Growthcurver R package[60].

Datasets and input sequence encoding

Two distinct groups of data are utilized in model development: in-house data generated with the nuclease TevSpCas9, and those created from 70,000 sgRNA activity maps in *E. coli* developed by Guo et al. with the nucleases eSpCas9 and SpCas9[28]. Due to their method of generation, these datasets contain overlapping sgRNA sequences. To generate a unique sgRNA testing set for model testing, sgRNAs in the Guo SpCas9 dataset that are cross-listed with the eSpCas9 dataset were removed. This is the set referred to as the unique sgRNA Guo SpCas9 dataset.

All sgRNAs were mapped to the *E. coli* genome and pTox plasmid. Target sites with 15nt PAM proximal matches with a separate site were excluded from our datasets. Based upon mapping results, 43nt target site sequences were obtained for each sgRNA, containing the 20nt sgRNA target site, the 3nt PAM, and 10nt upstream and downstream. These extended inputs provided the ability to test sequence length versus predictive performance.

The nucleotides comprising the sgRNA target site sequences are commonly represented with strings of single characters (A, C, G, T) each representing a nucleotide. However, alphabetic encoding of nucleotides is not a friendly format for deep learning models. We converted our input sequences with one-hot encoding, where the input sequence is represented as a 4-by-N matrix – 4 nucleotide options across an N-length input sequence. The nucleotides, A, C, G, and T, are encoded as [1 0 0 1], [0 1 0 0], [0 0 1 0], and [0 0 0 1] respectively.

Data processing and activity score calculation

Reads from the Illumina sequencign were parsed using a custom script that deconvolute the barcoded sequences into a table that contained replicates of induced or repressed conditions. The bacterial sgRNA read counts from these datasets representing on-target activity scores are compositional in nature[48], and therefore require some form of normalization or transformation to become interpretable[61]. All sgRNAs in the Guo et al. datasets having a read count less than 20 in either replicate of the catalytically dead Cas9 (eSpdCas9 and SpdCas9) samples were removed. Relative abundance (‘rab.all‘) and difference values (‘diff.btw‘) for each guide were calculated using the ‘aldex.effect‘ function of ALDEx2[48]. These scores were then normalized to Z-scores by dividing each score by the standard deviation its respective dataset (Supplementary Figure S4). To obtain an untouched dataset for model generalization testing, we used the original, Z-score based normalization, sgRNA activity scoring by Guo et al. for the unique Guo SpCas9 dataset[28]. Data were plotted using R.

Model establishment and transfer learning

During model development we tested various architectures, including those with multiple branches, to test the performance of CNN and RNN neural networks. A CNN is an artificial neural network which excels at capturing spatial information from an input. This capability results in the frequent application of CNNs to image recognition problems. Similar to pixels in an image, one-hot encoded nucleotide sequences can be used as inputs to a CNN, whereby local nucleotide preferences can be extracted[34,58,59].

Contrasting the CNNs local information capture capabilities, RNNs excel at learning sequential information. RNNs contain an internal memory state which are updated to learn important interactions within a sequence. Prior work has shown the benefit of utilizing a combination of CNN and RNN layers within a model to improve performance[58,59]. Spatial information captured by CNN layers can be fed to the RNN, whereby sequential information is then deduced, increasing performance[57].

We developed models on prior datasets to optimize for transfer learning – referred to as base models. Transfer learning is a method whereby a model utilizes information transferred from a similar domain to improve performance[62]. In practice, the base model is commonly constructed on datasets larger than those to which the transfer learning will be applied. For our context, we test models constructed on either the Guo SpCas9 or eSpCas9 dataset, and apply those base models as the starting point for training on our smaller datasets. To maximize the benefit of the pre-learned information from the base model, we tested variations in model layer freezing, where parameters in specific layers of the model are fixed before transfer learning model training occurs.

Model training and tuning

We constructed crisprHAL with Tensorflow Keras[63]. This network was trained using the optimizer Adam, with mean squared error used as the loss function. All variations of the base models used a batch size of 200, with transfer learning being performed with a batch size of 20. The base model and transfer learning model were tuned using 5-fold cross validation with a 80% training set and a 20% test set for each fold for each stages respective dataset – base data and transfer learning data. Hyperparameter tuning was performed for a number of factors affecting the model, including: number of CNN layers, number of dense layers, channel sizes, CNN window sizes, RNN size, dense layer sizes, dropout rates, and activation functions between layers.

Installation and testing of other models

We installed and ran the Guo and DeepSgRNA models (downloaded from Github sites https://github.com/zhangchonglab/sgRNA-cleavage-activity-prediction and https://github.com/biomedBit/DeepSgrnaBacteria)[28,34]. To test the Guo SpCas9 and eSpCas9 models, we converted our sgRNA-associated target site sequence inputs to the required 30nt length, containing the 20nt sgRNA target, 4nt upstream, and 6nt downstream including the NGG PAM. To test the DeepSgRNA SpCas9 and eSpCas9 models, we converted our sgRNA-associated target site sequence inputs to the required 43nt length, containing the 20nt sgRNA target, 10nt upstream, and 13nt downstream including the NGG PAM.

Performance and evaluation of models

To evaluate our models we used Spearman rank correlation coefficient, referred to as rank correlation. We chose this metric rather than Pearson correlation coefficient as it does not depend on a linear association between variables. Additionally, given its past use, it provides a clear metric from which to compare our models' performance to prior models[28,34,58,59]. We calculated rank correlation with the "spearmanr" function from the Scipy stats Python package[64].

Code availability

Our model to predict TevSpCas9 and SpCas9 target site activity is available for download at https://github.com/tbrowne5/crisprHAL without restriction.

References

1. Vigouroux, A. & Bikard, D. CRISPR tools to control gene expression in bacteria. Microbiol. Mol. Biol. Rev. **84**, e00077–19 (2020).
2. Adli, M. The CRISPR tool kit for genome editing and beyond. Nat. Commun. **9**, 1–13 (2018).

3. Deltcheva, E. et al. CRISPR RNA maturation by *trans*-encoded small RNA and host factor RNase III. Nature **471**, 602–607 (2011).
4. Jinek, M. et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science **337**, 816–821 (2012).
5. Hamilton, T. A. et al. Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. Nat. Commun. **10**, 1–9 (2019).
6. Neil, K. et al. High-efficiency delivery of CRISPR-Cas9 by engineered probiotics enables precise microbiome editing. Mol. Syst. Biol. **17**, e10335 (2021).
7. Bikard, D. et al. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat. Biotechnol. **32**, 1146–1150 (2014).
8. Reuter, A. et al. Targeted-antibacterial-plasmids (TAPs) combining conjugation and CRISPR/Cas systems achieve strain-specific antibacterial activity. Nucleic Acids Res. **49**, 3584–3598 (2021).
9. Citorik, R. J., Mimee, M. & Lu, T. K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat. Biotechnol. **32**, 1141–1145 (2014).
10. Gomaa, A. A. et al. Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. MBio **5**, e00928–13 (2014).
11. Lam, K. N. et al. Phage-delivered CRISPR-Cas9 for strain-specific depletion and genomic deletions in the gut microbiome. Cell Reports **37**, 109930 (2021).
12. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of *Escherichia coli*. Nucleic Acids Res. **44**, 4243–4251 (2016).
13. Pyne, M. E., Moo-Young, M., Chung, D. A. & Chou, C. P. Coupling the CRISPR/Cas9 system with lambda red recombineering enables simplified chromosomal gene replacement in *Escherichia coli*. Appl. Environ. Microbiol. **81**, 5103–5114 (2015).
14. Jiang, Y. et al. Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. Appl. Environ. Microbiol. **81**, 2506–2514 (2015).
15. Zerbini, F. et al. Large scale validation of an efficient CRISPR/Cas-based multi gene editing protocol in *Escherichia coli*. Microb. Cell Factories **16**, 1–18 (2017).
16. Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell **152**, 1173–1183 (2013).
17. Bikard, D. et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. **41**, 7429–7437 (2013).
18. Pellegrino, G. M. et al. Metabolically-targeted dCas9 expression in bacteria. Nucleic Acids Res. **51**, 982–996 (2023).
19. Farasat, I. & Salis, H. M. A biophysical model of CRISPR/Cas9 activity for rational design of genome editing and gene regulation. PLoS Comput. Biol. **12**, e1004724 (2016).
20. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. **34**, 184–191 (2016).

21. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*. Nat. Methods **12**, 982–988 (2015).

22. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol. **19**, 1–18 (2018).

23. Chari, R., Yeo, N. C., Chavez, A. & Church, G. M. sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity. ACS Synth. Biol. **6**, 902–904 (2017).

24. Singh, R., Kuscu, C., Quinlan, A., Qi, Y. & Adli, M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Res. **43**, e118–e118 (2015).

25. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. Nucleic Acids Res. **50**, 3616–3637 (2022).

26. Moreb, E. & Lynch, M. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. Nat. Commun. **12**, 1–13 (2021).

27. Shen, J., Zhou, J., Chen, G.-Q. & Xiu, Z.-L. Efficient genome engineering of a virulent *Klebsiella* bacteriophage using CRISPR-Cas9. J. Virol. **92**, e00534–18 (2018).

28. Guo, J. et al. Improved sgRNA design in bacteria via genome-wide activity profiling. Nucleic acids research **46**, 7052–7069 (2018).

29. Moreb, E. A. et al. Managing the sos response for enhanced crispr-cas-based recombineering in e. coli through transient inhibition of host reca activity. ACS Synth. Biol. **6**, 2209–2218 (2017).

30. Ye, S., Enghiad, B., Zhao, H. & Takano, E. Fine-tuning the regulation of cas9 expression levels for efficient CRISPR-Cas9 mediated recombination in *Streptomyces*. J. Ind. Microbiol. Biotechnol. **47**, 413–423 (2020).

31. Peters, J. M. et al. Bacterial CRISPR: accomplishments and prospects. Curr. Opin. Microbiol. **27**, 121–126 (2015).

32. Zhao, J., Fang, H. & Zhang, D. Expanding application of CRISPR-Cas9 system in microorganisms. Synth. Syst. Biotechnol. **5**, 269–276 (2020).

33. Misra, C. S. et al. Determination of Cas9/dCas9 associated toxicity in microbes. BioRxiv 848135 (2019).

34. Wang, L. & Zhang, J. Prediction of sgRNA on-target activity in bacteria by deep learning. BMC Bioinforma. **20**, 1–14 (2019).

35. Moreb, E. A. & Lynch, M. D. A meta-analysis of gRNA library screens enables an improved understanding of the impact of gRNA folding and structural stability on CRISPR-Cas9 activity. The CRISPR J. **5**, 146–154 (2022).

36. Dupuis, M.-È., Villion, M., Magadán, A. H. & Moineau, S. CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. Nat. Commun. **4**, 1–7 (2013).

37. Strotskaya, A. et al. The action of *Escherichia coli* CRISPR–Cas system on lytic bacteriophages with different lifestyles and development strategies. Nucleic Acids Res. **45**, 1946–1957 (2017).

38. Wolfs, J. M. et al. Biasing genome-editing events toward precise length deletions with an RNA-guided TevCas9 dual nuclease. Proc. Natl. Acad. Sci. **113**, 14988–14993 (2016).

39. Chen, Z. & Zhao, H. A highly sensitive selection method for directed evolution of homing endonucleases. Nucleic Acids Res. **33**, e154–e154 (2005).

40. Kleinstiver, B. P., Fernandes, A. D., Gloor, G. B. & Edgell, D. R. A unified genetic, computational and experimental framework identifies functionally relevant residues of the homing endonuclease I-BmoI. Nucleic Acids Res. **38**, 2411–2427 (2010).

41. Kleinstiver, B. P. et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature **523**, 481–485 (2015).

42. McMurrough, T. A., Dickson, R. J., Thibert, S. M., Gloor, G. B. & Edgell, D. R. Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. Proc. Natl. Acad. Sci. **111**, E2376–E2383 (2014).

43. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. **31**, 233–239 (2013).

44. Semenova, E. et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. **108**, 10098–10103 (2011).

45. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. **31**, 822–826 (2013).

46. Anderson, E. M. et al. Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of off-target activity. J. Biotechnol. **211**, 56–65 (2015).

47. Fu, B. X., St. Onge, R. P., Fire, A. Z. & Smith, J. D. Distinct patterns of Cas9 mismatch tolerance *in vitro* and *in vivo*. Nucleic Acids Res. **44**, 5365–5377 (2016).

48. Fernandes, A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome **2**, 1–13 (2014).

49. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of pam-dependent target dna recognition by the cas9 endonuclease. Nature **513**, 569–573 (2014).

50. Qian, Z. et al. The post-PAM interaction of RNA-guided spCas9 with DNA dictates its target binding and dissociation. Sci. Adv. **5**, eaaw9807, DOI: 10.1126/sciadv.aaw9807 (2019).

51. Jiang, F. et al. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science **351**, 867–871 (2016).

52. Zhang, Q. et al. Efficient DNA interrogation of Spcas9 governed by its electrostatic interaction with DNA beyond the PAM and protospacer. Nucleic Acids Res. **49**, 12433–12444 (2021).

53. Yang, M. et al. Nonspecific interactions between SpCas9 and dsDNA sites located downstream of the PAM mediate facilitated diffusion to accelerate target search. Chem. Sci. **12**, 12776–12784 (2021).

54. McMurrough, T. A. et al. Active site residue identity regulates cleavage preference of LAGLIDADG homing endonucleases. Nucleic Acids Res. **46**, 11990–12007 (2018).

55. Ratner, H. K. et al. Catalytically active Cas9 mediates transcriptional interference to facilitate bacterial virulence. Mol. cell **75**, 498–510 (2019).

56. Cui, L. et al. A CRISPRi screen in *E. coli* reveals sequence-specific toxicity of dCas9. Nat. communications **9**, 1912 (2018).

57. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. **44**, e107–e107 (2016).

58. C-RNNCrispr: prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. Comput. Struct. Biotechnol. J. **18**, 344–354 (2020).

59. 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.

60. Sprouffske, K. & Wagner, A. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. BMC Bioinforma. **17**, 1–4 (2016).

61. Gloor, G. B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can. J. Microbiol. **62**, 692–703 (2016).

62. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. J. Big Data **3** (2016).

63. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.

64. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods **17**, 261–272 (2020).

Acknowledgements

Supported by a CIHR Project Grant (PJT-159708) to D.R.E. and G.B.G. D.T.H. and T.S.B. were supported by a MITACS Accelerate Award. We thank Michael Hallet for discussions on machine learning.

Author contributions statement

D.T.H, T.S.B, G.B.G and D.R.E conceived the experiments, D.T.H, P.N.B, T.S.B conducted the experiments, D.T.H, T.S.B, G.B.G and D.R.E analyzed the results. All authors reviewed the paper.