Applied machine learning for the analysis of CRISPR-Cas systems

Angewandtes maschinelles Lernen für die Analyse von CRISPR-Cas-Systemen

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Section: Infection and Immunity
submitted by

Yanying Yu
from
Guangzhou, China

Würzburg 2023

Submitted on:

Members of the Thesis Committee

Chairperson: Prof. Dr. Jörg Schultz

Primary Supervisor: Jun.-Prof. Dr. Lars Barquist

Supervisor (Second): Prof. Dr. Chase Beisel

Supervisor (Third): Dr. Ana Rita Brochado

Supervisor (Fourth): Dr. Tobias Müller

Supervisor (Fifth): Prof. Dr. Marco Galardini

“The others obey their own lead, follow their own impulses.
Don’t be distracted. Keep walking.
Follow your own nature, and follow Nature—along the road they share.”

Marcus Aurelius

Acknowledgments

I would like to give special thanks to my primary supervisor Lars Barquist for the opportunity to start my journey at Helmholtz Institute for RNA-based Infection Research (HIRI). His trust and advice have been a great comfort and guidance. I would also like to acknowledge my other committee members–Prof. Dr. Chase Beisel, Dr. Ana Rita Brochado, Dr. Tobias Müller, and Prof. Dr. Marco Galardini–for the feedback and support. I am also grateful to all my colleagues and collaborators for the discussions, the positive research experiences, and the good vibe in the academic environment. I thank Prof. Dr. Jörg Schultz for chairing my doctoral procedure. I thank Willow Kion-Crosby, Regan Hayward, Rituparno Sen, Sandra Gawlitt, and Jakob Jung for the proofreading. I thank my parents for their unconditional support. I thank my friends and family, here and there, now and then, for the love, care, and encouragement.

I really appreciate everyone in my life, who has, I believe, more or less, contributed to who I am and where I am standing. Thank you all for being in my life: Willow Kion-Crosby, Xiaopeng Liang, Jing Zhang, Longhui Lin, Yihuan Wang, Chunguang Liang, Gemma N. Villagómez G., and many more.

The past four years have not been easy. I have learned a lot from the good, the unexpected, and the unfortunate. Here, I have deep gratitude for what I have received and what I have been through.

客乡居不易，俏梅逆春寒。匆匆过南柯，徐徐梦平生。

22.02.2023
Würzburg, Germany

Table of contents

Summary

Among the defense strategies developed in microbes over millions of years, the innate adaptive CRISPR-Cas immune systems have spread across most of bacteria and archaea. The flexibility, simplicity, and specificity of CRISPR-Cas systems have laid the foundation for CRISPR-based genetic tools. Yet, the efficient administration of CRISPR-based tools demands rational designs to maximize the on-target efficiency and off-target specificity. Specifically, the selection of guide RNAs (gRNAs), which play a crucial role in the target recognition of CRISPR-Cas systems, is non-trivial. Despite the fact that the emerging machine learning techniques provide a solution to aid in gRNA design with prediction algorithms, design rules for many CRISPR-Cas systems are ill-defined, hindering their broader applications.

CRISPR interference (CRISPRi), an alternative gene silencing technique using a catalytically dead Cas protein to interfere with transcription, is a leading technique in bacteria for functional interrogation, pathway manipulation, and genome-wide screens. Although the application is promising, it also is hindered by under-investigated design rules. Therefore, in this work, I develop a state-of-art predictive machine learning model for guide silencing efficiency in bacteria leveraging the advantages of feature engineering, data integration, interpretable AI, and automated machine learning. I first systematically investigate the influential factors that attribute to the extent of depletion in multiple CRISPRi genome-wide essentiality screens in *Escherichia coli* and demonstrate the surprising dominant contribution of gene-specific effects, such as gene expression level. These observations allowed me to segregate the confounding gene-specific effects using a mixed-effect random forest (MERF) model to provide a better estimate of guide efficiency, together with the improvement led by integrating multiple screens. The MERF model outperformed existing tools in an independent high-throughput saturating screen. I next interpret the predictive model to extract the design rules for robust gene silencing, such as the preference for cytosine and disfavoring for guanine and thymine within and around the protospacer adjacent motif (PAM) sequence. I further incorporated the MERF model in a web-based tool that is freely accessible at www.ciao.helmholtz-hiri.de.

When comparing the MERF model with existing tools, the performance of the alternative gRNA design tool optimized for CRISPRi in eukaryotes when applied to bacteria was far from satisfying, questioning the robustness of prediction algorithms across organisms. In addition, the CRISPR-Cas systems exhibit diverse mechanisms albeit with some similarities. The captured predictive patterns from one dataset thereby are at risk of poor generalization when applied across organisms and CRISPR-Cas techniques. To fill the gap, the machine learning approach I present here for CRISPRi could serve as a blueprint for the effective development of prediction algorithms for specific organisms or CRISPR-Cas systems of interest. The explicit workflow includes three principle steps: 1) accommodating the feature set for the CRISPR-Cas system or technique; 2) optimizing a machine learning model using automated machine learning; 3) explaining the model using interpretable AI. To illustrate the applicability of the workflow and diversity of results when applied across different bacteria and CRISPR-Cas systems, I have applied this workflow to analyze three distinct CRISPR-Cas genome-wide screens. From the CRISPR base editor essentiality screen in *E. coli*, I have determined the PAM preference and sequence context in the editing window for efficient editing, such as A at the 2nd position of PAM, A/TT/TG downstream of PAM, and TC at the 4th to 5th position of gRNAs. From the CRISPR-Cas13a screen in *E. coli*, in addition to the strong correlation with the guide depletion, the target expression level is the strongest predictor in the model, supporting it as a main determinant of the activation of Cas13-induced immunity and better characterizing the CRISPR-Cas13 system. From the CRISPR-Cas12a screen in *Klebsiella pneumoniae*, I have extracted the design rules for robust antimicrobial activity across *K. pneumoniae* strains and provided a predictive algorithm for gRNA design, facilitating CRISPR-Cas12a as an alternative technique to tackle antibiotic resistance.

Overall, this thesis presents an accurate prediction algorithm for CRISPRi guide efficiency in bacteria, providing insights into the determinants of efficient silencing and guide designs. The systematic exploration has led to a robust machine learning approach for effective model development in other bacteria and CRISPR-Cas systems. Applying the approach in the analysis of independent CRISPR-Cas screens not only sheds light on the design rules but also the mechanisms of the CRISPR-Cas systems. Together, I demonstrate that applied machine learning paves the way to a deeper understanding and a broader application of CRISPR-Cas systems.

Zusammenfassung

Unter den Verteidigungsstrategien, welche sich über Millionen von Jahren in Mikroben entwickelt haben, hat sich das angeborene adaptive CRISPR-Cas Immunsystem in vielen Bakterien und den meisten Archaeen verbreitet. Flexibilität, Einfachheit und Spezifizität von CRISPR-Cas Systemen bilden die Grundlage für CRISPR-basierten genetischen Werkzeugen. Dennoch verlangt die effiziente Anwendung CRISPR-basierter genetischer Werkzeuge ein rationales Design, um die Effektivität zu maximieren und Spezifizität zu gewährleisten. Speziell die Auswahl an Leit-RNAs, oder auch „guide“ RNAs (gRNAs), welche eine essentielle Rolle in der Ziel-Erkennung des CRISPR-Cas Systems spielen, ist nicht trivial. Trotz aufkommender Techniken des maschinellen Lernens, die mit Hilfe von Vorhersage-Algorithmen eine Unterstützung im gRNA-Design darstellen, sind die Design-Regeln für viele CRISPR-Cas Systeme schlecht definiert und die breite Anwendung dadurch bisher gehindert.

CRISPR Interferenz (CRISPRi), eine Methode der Genrepression, nutzt ein katalytisch inaktives Cas-Protein, um die Gen-Transkription zu verhindern und ist eine führende Technik für Gen-Funktionsstudien, der Manipulation von Stoffwechselwegen und genomweiter Screens in Bakterien. Auch wenn viele der Anwendungen vielversprechend sind, ist die Umsetzung aufgrund der wenig untersuchten Design-Regeln schwierig. Daher entwickele ich in dieser Arbeit ein hochmodernes auf maschinellem Lernen basierendes Modell für die Vorhersage der gRNA Genrepressions-Effizienz in Bakterien, wobei die Merkmalskonstruktion, Datenintegration, interpretierbare künstliche Intelligenz (KI) und automatisiertes maschinelles Lernen genutzt wurden. Zuerst untersuche ich systematisch die Einflussfaktoren, welche zum Ausmaß der Depletion in genomweiten CRISPRi-Screens zur Gen-Essentialität in *Escherichia coli* beitragen und demonstriere den überraschend dominanten Beitrag genspezifischer Effekte, wie z. B. dem Genexpressionslevel. Diese Beobachtungen erlaubten mir die genspezifischen Störvariablen mit einem sogenannten mixed-effect random forest (MERF) Modell zu segregieren, um eine bessere Einschätzung der gRNA Effizienz zu erreichen und durch die Integration zusätzlicher Screen-Daten noch weiter zu verbessern. Das MERF Modell übertraf dabei bereits existierende Werkzeuge in einem unabhängigen Hochdurchsatz Sättigungs-Screen. Als nächstes interpretiere ich die Modell Vorhersage, um Design-Regeln für eine solide Genrepression zu extrahieren, wie z. B. eine Präferenz für Cytosin und eine Abneigung gegenüber Guanin und Thymin innerhalb und der „protospacer adjacent motif“ (PAM) direkt umgebenden Sequenz. Weiterhin integrierte ich das MERF Modell in einem Web-basierten Werkzeug, welches unter www.ciao.helmholtz-hiri.de frei zugänglich ist.

Ein Vergleich von existierenden Werkzeugen mit dem MERF Modell zeigt, dass alternative, für CRISPRi in Eukaryoten optimierte, gRNA Design-Werkzeuge schlecht abschneiden, sobald sie in Bakterien angewandt werden. Dies lässt Zweifel an einer robusten Übertragbarkeit dieser

Vorhersage-Algorithmen zwischen verschiedenen Organismen. Zusätzlich haben CRISPR-Cas Systeme, trotz einiger genereller Gemeinsamkeiten, höchst diverse Wirkungsmechanismen. Die Vorhersagemuster eines Datensets sind daher schlecht generalisierbar, sobald sie auf andere Organismen oder CRISPR-Cas Techniken angewandt werden. Diese Lücke kann mit dem hier präsentierten Ansatz des maschinellen Lernens für CRISPRi geschlossen werden und als eine Vorlage für die Entwicklung effektiver Vorhersage-Algorithmen für spezifische Organismen oder CRISPR-Cas Systeme dienen. Der explizite Arbeitsablauf beinhaltet drei Hauptschritte: 1) Aufnehmen des Merkmalsets des jeweiligen CRISPR-Cas Systems bzw. der CRISPR-Cas Technik; 2) Optimierung des maschinellen Lernen Modells durch automatisiertes maschinelles Lernen; 3) Erklärung des Modells mit interpretierbarer KI. Um die Anwendbarkeit des Arbeitsablaufs und die Diversität der Ergebnisse, im Zusammenhang mit unterschiedlichen Organismen und CRISPR-Cas Systemen, zu demonstrieren, habe ich diese Arbeitsschritte zur Analyse drei unterschiedlicher genomweiter Screens angewandt. Von dem CRISPR „base editor“ Essentialitäts-Screen in *E. coli*, konnten die PAM Präferenzen und der Sequenzkontext innerhalb des Editierungsfensters für eine effiziente Editierung abgeleitet werden. Beispielsweise tragen ein A an der zweiten PAM Position, ein A/TT/TG an der PAM direkt nachgeschalten Position und ein TC an der vierten oder fünften gRNA Position zur effizienten Editierung bei. Im CRISPR-Cas13a Screen in *E. coli*, stellten wir eine starke Korrelation zwischen dem Genexpressionslevel und der gRNA-Depletion fest. Zusätzlich ist das Expressionslevel des Ziel-Gens der stärkste Vorhersagefaktor des Modells, was das Expressionslevel als Hauptdeterminante für die Cas13-induzierte Immunität hervorhebt und die bessere Charakterisierung von CRISPR-Cas13 Systemen ermöglicht. Aus dem CRISPR-Cas12a Screen in *Klebsiella pneumoniae*, habe ich gRNA Design Regeln für die robuste antimikrobielle Aktivität über unterschiedliche *K. pneumonia*e Stämme hinweg extrahiert und einen Vorhersage-Algorithmus für das gRNA Design bereitgestellt. Dies ermöglicht die Nutzung von Cas12a als eine alternative Lösung, um Antibiotikaresistenzen zu bekämpfen.

Zusammengefasst präsentiert diese Thesis einen akkuraten Vorhersage-Algorithmus für die CRISPRi gRNA Effizienz in Bakterien und gibt Einblicke in die Determinanten für eine effiziente Genrepression und optimales gRNA Design. Die systematische Exploration führte zu einem robusten Ansatz des maschinellen Lernens für effektive Modell Entwicklungen in unterschiedlichen bakteriellen Spezies und CRISPR-Cas Systemen. Durch die Anwendung dieses Ansatzes auf unabhängige CRISPR-Cas Screens, konnte ich nicht nur wichtige Design Regeln ableiten, sondern auch die Mechanismen der jeweiligen CRISPR-Cas Systeme besser erleuchten. Zu guter Letzt demonstriere ich hier, dass angewandtes maschinelles Lernen den Weg zu einem tieferen Verständnis und einer breiteren Anwendung von CRISPR-Cas Systemen ebnen kann.

1. Introduction

The introduction section comprises five subsections. In section 1.1, I describe briefly the function and mechanism of the CRISPR-Cas system using CRISPR-Cas9 as an example. In section 1.2, I describe the classification of CRISPR-Cas systems and the mechanisms of Cas12a and Cas13a, given that my work involves these two systems in addition to Cas9. Section 1.3 includes examples of applications of CRISPR-based genetic tools, given that my work involves applications related to transcription regulation, base editing, and gene knockout. In section 1.4, I describe the importance of gRNA design and give examples of how design rules have been studied. Section 1.5 focuses on the machine learning methods in each main stage of model development involved in gRNA efficiency studies.

1.1 CRISPR-Cas system is an adaptive immune system in bacteria

Much like humans, bacteria often encounter unwelcome invaders, such as bacteriophages. In order to protect themselves and the population from these invaders, bacteria have developed various immune responses, including restriction-modification systems, argonaute (Ago) proteins, abortive initiation (Abi) systems, and CRISPR-Cas systems.

Despite the focus of my work being on CRISPR-Cas systems, other immune systems equally provide efficient protection and share some characteristics with CRISPR-Cas systems. Restriction-modification systems (Raleigh & Brooks, 1998) and Ago proteins (Höck & Meister, 2008; Niaz, 2018) can specifically recognize the alien sequences, DNA or RNA, and introduce cleavage in the alien sequences to prevent the reproduction of the invader and clear the infection. Toxin-antitoxin systems (Van Melderen & De Bast, 2009; Unterholzner et al., 2013), one of the Abi systems, can release toxins upon bacteriophage infection, which leads to individual cell death to protect the population. Amazingly, CRISPR-Cas systems possess both non-self sequence recognition specificity and altruistic characteristics. Apart from these, CRISPR-Cas systems are adaptive immune systems (Heler et al., 2014), which means the previous infections can be memorized to help the bacteria respond more efficiently to subsequent infections.

CRISPR, an abbreviation of clustered regularly interspaced short palindromic repeats, is an array of conserved sequences in bacteria genomes (Ishino et al., 1987). As suggested by its full name, the CRISPR sequence harbors short palindromic repeats separated by spacers. These spacers, which play a key role in non-self-recognition, originate from the invading alien sequences, termed protospacers, which are cleaved and inserted into the CRISPR array by CRISPR-associated (Cas) proteins in the adaptation step (Amitai & Sorek, 2016). Up-to-date, over 30 CRISPR-Cas systems have been identified (Koonin et al., 2017); these are described in detail in section 1.2. Among them, CRISPR-Cas9 system was the first commonly applied in genome editing given its structural simplicity (only requiring a single Cas protein to form the effector) and thereby the most well-studied given its popularity in numerous CRISPR-based tools (Ran, Hsu, Wright, et al., 2013; Charpentier & Marraffini, 2014) (see section 1.3 for a more detailed description). Using the CRISPR-Cas9 system as an example to briefly describe the mechanism of CRISPR-mediated immunity (**Figure 1.1**) (F. Jiang & Doudna, 2017), after the CRISPR array is transcribed, the repeat sequences in the premature CRISPR RNA (pre-crRNA) are bound by the trans-activating CRISPR RNA (tracrRNA), creating double-strand RNAs, which are subsequently processed by RNase III into a mature crRNA-tracrRNA duplex (Brouns et al., 2008; Deltcheva et al., 2011). Following the binding of Cas9 proteins to the duplex (Gasiunas et al., 2012; F. Jiang et al., 2015), the Cas9-RNA complex is guided by the spacer sequence to the target DNA, whose sequence is identical to the spacer sequence. Despite the existence of the spacer sequence in the bacteria genome, the Cas9-RNA complex additionally recognizes the protospacer adjacent motif (PAM) sequence, which is at the 3' end of the target sequence to identify the invader sequence as non-self (Marraffini & Sontheimer, 2010). Upon the successful recognition of the PAM, the Cas9-RNA complex opens up the target double-strand DNA to form an R-loop structure (F. Jiang et al., 2016), where the spacer sequence base pairs with the complementary strand of the target sequence. A double-strand break (DSB) in the target DNA introduced by the Cas9 protein removes the invader and leads to a successful defense (Jinek et al., 2012). The spacer sequence is often known as a guide RNA (gRNA), or simply a guide, based on its function in the interference step.

PAM plays a central role in target recognition of CRISPR-mediate immunity. In the absence of a PAM, the Cas9-RNA complex cannot cleave the target sequence, even if the gRNA sequence is perfectly complementary to the target. The dependence on a functional PAM avoids self-cleavage at the CRISPR locus. PAM mutation in viral sequences leads to the escape of CRISPR-mediate immunity (Paez-Espino et al., 2013). Diverse functional PAMs have been identified amongst Cas9 proteins (Leenay et al., 2016). For example, the primary PAM sequence of *Streptococcus pyogenes* Cas9 (SpCas9) is NGG (N=A/C/T/G) (Ma et al., 2016b; Sternberg et al., 2014), whereas that of *Streptococcus thermophilus* Cas9 from CRISPR1 locus (Sth1Cas9) is NNAGAAW (W=A/T) (Deveau et al., 2008; Horvath et al., 2008; Rock et al., 2017). This diversity in PAM sequences enables flexible targeting of invaders. Moreover, Cas proteins can recognize secondary PAM sequences with lower binding affinity, i.e. NAG for SpCas9. The candidate PAM sequences that can be recognized by a specific Cas protein can be identified using methods like PAM-SCANR and PAM-SEARCH (Leenay et al., 2016; Collias et al., 2020). The wheel-shaped graph, termed PAM wheel, was invented to convey the diversity and efficiency of each PAM sequence at the same time (Leenay et al., 2016). The probability of adapting and targeting a wider range of sequences with secondary PAMs indicates the flexibility of the immune response of CRISPR-Cas systems.

Apart from PAM, the flexibility of CRISPR-Cas systems lie in the tolerance of mismatches between gRNA and target sequence at certain positions (Fu et al., 2013; Hsu et al., 2013; Bravo et al., 2022). The common length of the mature gRNA for Cas9 is 20 nt (Gasiunas et al., 2012). The 8 to 12 nt PAM-proximal region, defining the specificity, is considered a 'seed' region (Semenova et al., 2011; Wiedenheft et al., 2011). One mismatch in the 'seed' region can abort the formation of the R-loop structure, whereas up to five mismatches or a short insertion and deletion in the PAM-distal region reduces the efficiency to different extents (Pattanayak et al., 2013). This tolerance promotes off-target binding and cleavage, which might lead to targeting unseen invaders.

The CRISPR-Cas system, specific-yet-flexible and simple-yet-efficient, stands out as an example of how dynamic and adaptive bacterial genome can be, which continues to surprise us with more CRISPR-Cas systems being discovered.

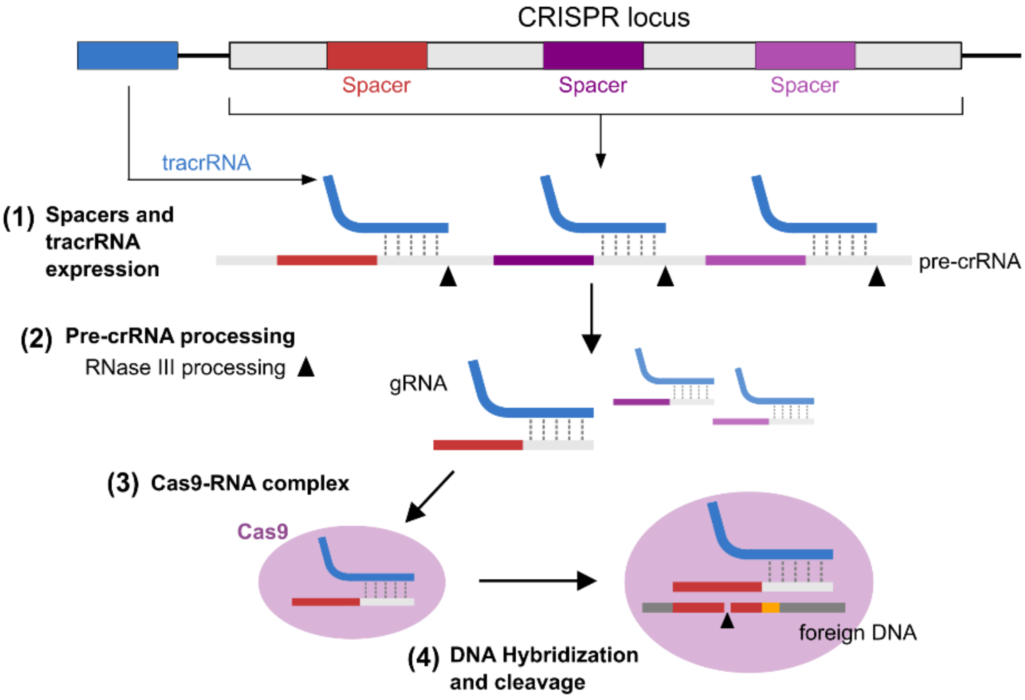

Figure 1.1. The mechanism of CRISPR-mediated immunity using CRISPR-Cas9 system as an example. Premature CRISPR RNA (pre-crRNA) was transcribed from the CRISPR locus, followed by the binding of trans-activating CRISPR RNA (tracrRNA, *blue*) to the repeat sequences in the pre-crRNA. The double-strand RNAs are subsequently processed by RNase III into a mature crRNA-tracrRNA duplex. Following the binding of Cas9 proteins to the duplex, the Cas9-RNA complex is guided by the spacer sequence to the foreign DNA. Upon the successful recognition of the protospacer adjacent motif (PAM, *orange*), gRNA hybridizes to the target DNA and the Cas9 protein cleaves the foreign DNA.

1.2 CRISPR-Cas systems show great diversity

The declining cost and surging practice of whole genome sequencing and metagenome sequencing allowed the discovery of novel CRISPR-Cas systems. Using specialized CRISPR identification approaches, computation tools such as PILER-CR (Edgar, 2007), CRT (Bland et al., 2007), CRISPRfinder (Grissa et al., 2007), CRISPRDetect (Biswas et al., 2016), and CRISPRCasfinder (Couvin et al., 2018), CRISPR-Cas systems have been identified in more than 40% of bacteria and 85% of archaea (Makarova et al., 2019), complementary to the other immune systems mentioned in section 1.1. Apart from the well-known CRISPR-Cas9 system, over 30 additional CRISPR-Cas systems have been identified (Koonin et al., 2017), which are categorized into two main classes (Class 1 and Class 2) and six subtypes (type I to VI).

Class 1 possesses CRISPR-Cas systems that require multiple Cas proteins to form an effector module in the interference step, while Class 2 systems require a single Cas protein. The subtype classification considers additionally whether the target sequence type is DNA or RNA. For example, the CRISPR-Cas9 system is a Class 2 type II system given that a single Cas protein is required to form an effector and it targets DNA, and the CRISPR-Cas13 system is a Class 2 type VI system given that a single Cas protein is required but it targets RNA.

Aside from target sequence type, other components of the mechanism among CRISPR-Cas systems might differ. These components include the pre-crRNA processing, PAM sequence, and the type of cleavage introduced by Cas proteins. Differences in pre-crRNA processing might simplify the steps of expressing and forming a function Cas-RNA complex in biotechnological applications. Diversity in PAMs enables the expansion of targetable sequences. The type of cleavage influences the efficiency of the CRISPR-Cas system as a genetic tool. The discrepancies in their mechanisms create opportunities for novel genetic tools but pose challenges to developing prediction algorithms for each system, which inspired my work in this thesis. Given that my work involves Cas9, Cas12a, and Cas13a systems, I describe their mechanisms in more detail here.

Class 2 consists of types II, V, and VI. Type II is represented by Cas9, whose mechanism was described in section 1.1. A representative of type V is Cas12a, previously known as Cpf1 (Zetsche et al., 2015). While targeting DNA, the CRISPR-Cas12a system is distinguished from the Cas9 system in several aspects. The maturation of crRNA is independent of tracrRNA and thus also RNase III. Additionally, Cas12a systems recognize T-rich PAMs at the 5’ end of the target sequence. For instance, the PAMs for the two most commonly used Cas12a in genetic tools are KYTV (K = T/G, Y = A/C, V = A/G/C) for *Francisella tularensis* Cas12a (FnCas12a) (Tu et al., 2017) and TTTV for *Acidaminococcus sp.* Cas12a (AsCas12a) and *Lachnospiraceae bacterium* Cas12a (LbCas12a) (Zetsche et al., 2015).

Compared to Cas9 which introduces a blunt-end DSB, Cas12a leaves a sticky-end break, leading to enhancing efficiency as a genome editing tool.

Another subtype, type VI in Class 2, instead includes the RNA-targeting CRISPR-Cas systems, which equip the bacteria to target RNA bacteriophages and the transcripts from DNA bacteriophages. Cas13a from *Leptotrichia shahii* (LshCas13a) (Abudayyeh et al., 2016) was the first characterized single-strand RNA-targeting type VI CRISPR-Cas system. Similar to Cas12a, tracrRNA is absent in the CRISPR-Cas13a system. Surprisingly, Cas13a also cleaves the bystander RNAs upon successful activation, thus inducing cell dormancy, preventing the dissemination of bacteriophages (Meeske et al., 2019). The criteria for successful activation of the CRISPR-Cas13a system differs from the other two subtypes. Besides the base pairing between gRNA and target RNA, Cas13a recognizes a protospacer flanking site (PFS) instead of PAM at the 3' end of the target sequence. Unlike Cas9 or Cas12, which rely on specific PAM sequences, Cas13a binds to the target when PFS is not G (Abudayyeh et al., 2016). The cleavage can however remain inactive upon binding when mismatches are present between gRNA and target RNA (Tambe et al., 2018) or when the expression level of the target RNA is below the threshold (Vialetto et al., 2022). The complex specificity and target expression threshold highlight the delicacy of CRISPR-Cas systems in the trade-off between efficient immunity and fitness cost.

Given the requirement of only a single Cas protein to form an effector, it is simpler to adapt Class 2 CRISPR-Cas systems for biotechnological applications compared to Class 1. While methods were developed for the rapid characterization of novel CRISPR-Cas systems (Karvelis et al., 2015; Leenay et al., 2016; Marshall et al., 2018; Wimmer et al., 2022), the accumulating knowledge of CRISPR-Cas systems unveiled other mysterious characteristics beyond adaptive immunity, such as naturally occurring self-targeting spacers, which target the endogenous genes or transcripts (Stern et al., 2010; Wimmer & Beisel, 2019). The self-targeting spacers hint at their multifunctional role in evolution. Databases, such as CRISPRCasdb (Pourcel et al. 2020), have been built around the classification and sequences of repeats and spacers, and allow mining of the spacer sequence dynamics to shed light on the phage-bacteria interaction (Dion et al., 2021). Moreover, The diversity of CRISPR-Cas systems potentiates the upgrade of the gene-editing tool kit with optimal protein size, cutting location, sequence recognition specificity, and numerous targeting mechanisms.

1.3 CRISPR-Cas systems enrich the genetic tool kit

CRISPR-Cas systems gained their fame owing to their versatility as genome editing tools. In the pre-CRISPR era, homologous recombination, zinc finger nucleases (ZFNs) (Carroll, 2011), and transcription activator-like effector (TALE) proteins (Sanjana et al., 2012) were among the methods developed for genome editing (H. Wang et al., 2016). However, these methods show limitations in either low efficiency or the necessity of protein redesign for each target, which were sufficiently amended by the CRISPR-Cas-based tools. CRISPR-Cas systems recognize the target simply with gRNA, which can be easily and cost-effectively redesigned for a new target compared to protein engineering. The requirement of tracrRNA for maturation in the CRISPR-Cas9 system can be further simplified by fusing crRNA and tracrRNA into one RNA sequence (single guide RNA, sgRNA) that still forms a functional hairpin structure for Cas9 protein binding to assemble a Cas-RNA complex (Jinek et al. 2012). Therefore, the CRISPR-Cas-based tools only require the expression of Cas proteins and programmed sgRNA, exhibiting greater simplicity. Given the simplicity of the CRISPR-Cas-based tools and the diversity of CRISPR-Cas systems, CRISPR-Cas systems have been repurposed for a large range of applications. Although my work only involves gene knockout, transcriptional regulation, and base editing, examples of other applications that include epigenetic modifications, genomic loci visualization, and RNA/DNA detection are also described briefly in the following sections to present a complete picture of the applications. Furthermore, these techniques have been extended to multiplex targeting and genome-wide screens, which was challenging for the previous genome editing tools given the time-consuming and expensive protein redesign for each target. Genome-wide screens are the main source of the data that I adopted for the work in this thesis.

1.3.1 Gene knockout

The nature of CRISPR-Cas systems is to clear the invaders by cleaving target sequences. Gene knockout assays were first tested in eukaryotic cells with the most well-characterized CRISPR-Cas9 system (Mali et al., 2013). The first steps of the gene knockout assay are similar to the natural mechanism described in section 1.1. Following the introduction of the DSB, the DSB is repaired by either non-homologous end joining (NHEJ) or homology-directed repair (HDR) (Chang et al., 2017). NHEJ and HDR both have their own advantages and disadvantages, and approaches have been developed to address the problems of NHEJ and HDR.

NHEJ is efficient but error-prone (Bétermier et al., 2014), resulting in a small insertion, deletion, and even large chromosomal translocation. The uncertainty in induced mutations could obscure the knockout outcome. In contrast, HDR shows high fidelity, albeit lower efficiency, the requirement of an ssDNA donor, and in competition with NHEJ. Attempts have been made to improve the low efficiency of HDR. For instance, considering that HDR only takes place during the S and G2 phases of the cell cycle, previous work (Lin et al., 2014) applied a timed delivery to synchronize the cells to improve the efficiency from 10% to 38%. Further improvement in HDR efficiency can be achieved by using stimulators of CRISPR-mediated HDR (Nambiar et al., 2019; Rees et al., 2019) and inhibitors of genes associated with NHEJ (Arnoult et al., 2017; Jayavaradhan et al., 2019; Ray et al., 2020).

To address the problem of either error-prone NHEJ or low-efficiency HDR, Cas9 nickase (nCas9) (Ran, Hsu, Lin, et al., 2013; B. Shen et al., 2014) was engineered to produce only single-stranded breaks (SSB). Therefore, using dual nickases for gene knockout (Ran, Hsu, Lin, et al., 2013) improves repair efficiency in the on-target position and reduces the unwanted mutations in the off-target positions.

1.3.2 Transcriptional regulation

Transcriptional regulation techniques based on CRISPR-Cas systems have been developed by engineering catalytically inactive Cas9. The cleavage activity of the CRISPR-Cas systems can be silenced by mutating their nuclease domains, such as by point mutations (D10A and H840A) in RuvC and HNH nuclease domains in SpCas9 (Qi et al., 2013). The catalytically inactive Cas9 is also called dead Cas9 (dCas9). Repressing and activating transcription are two primary ways of regulating transcription. Repressing or activating a gene can reveal its functional importance. The CRISPR-based gene repression technique is termed CRISPR interference (CRISPRi) whereas the gene activation technique is CRISPR activation (CRISPRa).

For CRISPRi, gene repression is caused by the inference with either transcription initiation or elongation. The loss of cleavage activity in dCas9 results in transient binding to the target DNA. The collision of the RNA polymerase with the bound Cas-RNA complex interferes with transcription initiation when the binding happens near the polymerase binding site and with the transcription elongation when in the coding region. The regression can be enhanced by a fusion of an extra repressor to dCas9, such as Kruppel-associated Box (KRAB) (Gilbert et al., 2013). For CRISPRa, contrary to repressors, an activator can be fused to activate the transcription. An example of the activator is VP64, four tandem copies of a 16-amino-acid-long transactivation domain (VP16) of the Herpes simplex virus (Maeder et al., 2013; Perez-Pinera et al., 2013). Both CRISPR interference and activation (CRISPRi/a) have been optimized by recruiting more proteins or domains, such as co-repressor proteins KRAB-box-associated protein-1 (KAP-1) for KRAB (Friedman et al., 1996), a tripartite activator VP64-p65-Rta (VPR) (Chavez et al., 2015), and multiple V64 domains using SunTag (Tanenbaum et al., 2014).

Although many studies focus on eukaryotic cells, CRISPRi/a has been reprogrammed for bacterial expression regulation (Bikard et al., 2013; Peters et al., 2019), which filled the gap in the application of CRISPR-Cas techniques in bacteria, given that many bacteria lack the necessary repair pathways (NHEJ or HDR) for the cleavages introduced by Cas proteins, thereby gene knockout using CRISPR-based techniques exhibits strong cytotoxicity. Applying CRISPRi/a in prokaryotes also paved the way to optimize synthetic gene circuits or metabolic networks for desired products in bacteria such as biofuel, pharmaceuticals, and biomaterials (Khalil & Collins, 2010; Jusiak et al., 2016; S. Cho et al., 2018; Mougiakos et al., 2018; M. Xie & Fussenegger, 2018), besides the functional interrogation.

Moreover, Mismatch-CRISPRi (Hawkins et al., 2020) leveraged the tolerance of mismatches and the lower efficiency caused by mismatches to reach a tunable expression regulation that gave insights into the expression-fitness relationships. The development of CRISPRi/a enables applications in transcriptional regulation and also prokaryotes.

1.3.3 Base editing

Besides CRISPRi and gene knockout, gene silencing can be achieved by introducing premature stop codons using CRISPR base editors (Billon et al., 2017; Kuscu et al., 2017). The CRISPR base editor used in CRISPR-STOP harbors a cytosine deaminase, which edits Cytosine (C) into Uridine (U) and then Thymine (T) via base excision repair. By editing CGA (Arg), CAG (Gln), and CAA (Gln) codons, stop codons TGA (opal), TAG (amber), or TAA (ochre) can be created respectively.

The application of CRISPR base editors however extends beyond gene silencing. Owing to its ability to introduce permanent edits in the genome, CRISPR base editors are a promising technique for treating genetic diseases and altering phenotypes of interest. Nevertheless, base editors were first developed to improve the efficiency and specificity of CRISPR-based gene knockout. APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) was the earliest cytosine deaminases fused to dCas9 in the first generation of CRISPR base editors, followed by later optimization with i.e. a fusion of uracil DNA glycosylase inhibitor (UGI) and replacement of dCas9 with nCas9 to improve editing efficiency and specificity (Komor et al., 2016; Komor, Zhao, et al., 2017; B. Yang et al., 2019). Besides cytosine base editors (CBEs), adenine base editors (ABEs), which edit Adenine (A) to Guanine (G), expand the horizon of this technique (Gaudelli et al., 2017). Including the consequent conversion on the opposition strand, all four transition mutations (C to T, T to C, A to G, and G to A) can be installed using CRISPR base editors.

Despite the improvement in efficiency and specificity, CRISPR base editors follow the original target recognition mechanism and are restrained in what is called the 'editing window'. While only the bases in the editing window can be edited, all the candidate bases (C for CBEs and A for ABEs), including the unwanted ones, stand a chance to be converted. To enlarge the pool of editable bases, Cas proteins and their variants with different PAMs or more flexible PAMs are adopted (Hu et al., 2018; X. Li et al., 2018; Nishimasu et al., 2018). Engineered deaminases with narrower editing windows or specific sequence contexts curtail unwanted editing (Y. B. Kim et al., 2017; Gehrke et al., 2018; X. Wang et al., 2018; Tan et al., 2019). Further, tools such as REPAIRx (Y. Liu et al., 2020) were developed to edit RNAs in the combination with Cas13 proteins. Similar to CRISPRi/a the applications of CRISPR base editors have been mainly investigated in eukaryotes but have been shown to achieve satisfactory performance in both gram-negative and -positive bacteria, though only a small quantity of genes were tested (Banno et al., 2018; Gu et al., 2018; Zheng et al., 2018).

Apart from base editors, other techniques have been developed to introduce permanent edits in the genome. For example, prime-editing combines prime editing guide RNA (pegRNA), a modified reverse-transcriptase, and nCas9 (Anzalone et al., 2019; Kantor et al., 2020; Scholefield & Harrison, 2021). Its capability to convey all 12 types of single base substitution instead of only four transition mutations revolutionizes the base editing technique. Furthermore, transposon-associated CRISRP-Cas systems potentiated accurate insertion of longer than 10kb sequences (Klompe et al., 2019; Strecker et al., 2019; Vo et al., 2020).

In summary, CRISPR-based base editing techniques offer an alternative for gene silencing, enable all 12 types of single base substitution, and allow the insertion of large sequences.

1.3.4 Epigenetic modification

Apart from deaminase, epigenetic modifiers have been fused to dCas proteins to investigate individual epigenetic marks and regulate gene expression.

Given that DNA methylation can repress gene expression, multiple examples using Cas proteins and methylation have been explored. These include the fusion of both KRAB and eukaryotic DNA methyltransferase (DNMT3A) to dCas9 allows efficient gene silencing (Amabile et al., 2016). The effect of methylation removal on modulating gene expression was also studied using ten-eleven translocation (TET) dioxygenases (X. S. Liu et al., 2016). In addition to DNA methylation, fusing histone acetyltransferase P300 can increase gene expression by increasing the acetylation at the Lysine 27 acetylation position (H3K27ac) in the enhancer elements (Hilton et al., 2015). The diversity of epigenetic modification also gave rise to extensive studies in other modifiers such as histone demethylase (LSD1) (Kearns et al., 2015), histone deacetylases (HDAC) (Kwon et al., 2017), endogenous chromatin regulators (Braun et al., 2017), and prokaryotic DNA methyltransferase (MQ3) (Xiong et al., 2017).

The focus of epigenetic modification has been on transcriptional regulation, which seems redundant to CRISPRi/a. But the effects on gene repression or activation induced by CRISPRi/a are reversible. In comparison, epigenetic changes are inheritable. Therefore, the effects induced by epigenetic modification can persist in the daughter cells, broadening the applications of CRISPR-Cas systems.

1.3.5 Genomic loci visualization

CRISPR-Cas systems have been also used in genomic loci visualization. When fusing with fluorescence such as GFP, dCas9 can label and visualize genomic loci in living cells without cell fixation and sample heating (B. Chen et al., 2013), overcoming the limitation of fluorescent in-situ hybridization

(FISH). Simultaneous tracking of multiple loci is possible by co-expressing dCas9 orthologs tagged with fluorescent proteins with different colors (Ma et al., 2015) or recruiting different fluorescent proteins attached to RNA binding proteins (RBPs) by gRNA-fused RNA aptamers (Ma et al., 2016a, 2018). To compete with the background fluorescent signal, dozens of sgRNAs are required to visualize a non-repetitive locus, whereas one sgRNA achieves comparable signal strength when targeting a repetitive locus.

1.3.6 RNA/DNA detection

CRISPR-based nucleic acid detection techniques have improved on the previous golden standard polymerase chain reaction (PCR) with the combination of single-nucleotide specificity, cost efficiency, and simplicity. In particular, PCR detection requires sophisticated analytical instruments that confine its comprehensive implementation. Plentiful tools based on Cas9, Cas12, or Cas13 have been established to detect both DNA and RNA for the diagnosis of infectious diseases (Kaminski et al., 2021; J. Li et al., 2022).

An example of a Cas9-based method is nucleic acid sequence-based amplification (NASBA)-CRISPR cleavage (Pardee et al., 2016). Its sensitivity at the femtomolar (10^{-15} M), much higher than the common sensitivity of PCR (10^{-13} M, based on 25 ng per 100 µl reaction and 200 bp template DNA), level empowered the detection of the Zika virus from plasma samples. NASBA combined RNA amplification and target-specific PAM sequence to achieve sensitivity and specificity. The truncated DNA pieces upon successful cleavage by Cas9 miss the trigger sequence to activate the toehold sensor, thus absent in color change. Another example of a Cas9-based method is LEOPARD (leveraging engineered tracrRNAs and on-target DNAs for parallel RNA detection) (Jiao et al., 2021). LEOPARD takes advantage of the reprogrammed tracrRNA (Rptr) to sense target RNA to form a non-canonical crRNA (ncrRNA). After binding to *Campylobacter jejuni* Cas9, which is capable of binding ncrRNA, ncrRNA guides the complex to the target DNA. Successful cleavage of DNA indicates the presence of target RNA with single nucleotide resolution. Due to the high resolution, the detection of target RNA using LEOPARD can be multiplexed, which was showcased by detecting several SARS-CoV-2 variants at once.

Cas12- and Cas13-based methods include DETECTR (DNA endonuclease-targeted CRISPR trans reporter) (J. S. Chen et al., 2018) and SHERLOCK (specific high-sensitivity enzymatic reporter unlocking) (Gootenberg et al., 2017). In these two methods, recombinase polymerase amplification (RPA) or reverse transcription RPA (RT-RPA) is first performed to amplify the sample, followed by the recognition of the target using Cas12 or Cas13 systems. A collateral cleavage is then activated upon the target recognition, which acts on the bystanding ssDNA or ssRNA for Cas12 and Cas13 respectively. The cleavage in ssDNA or ssRNA consequently separates the quencher from the fluorophore to obtain the fluorescent signal. These methods reached an attomolar (10^{-18} M) sensitivity at the detection of the Zika virus, the flavivirus dengue, and human papillomavirus. SHERLOCK also showed the capability of bacterial species differentiation, human genotyping, and cancer mutation detection. Version two of SHERLOCK further implemented multiplexed detection using multiple fluorescent labels and improved the sensitivity to the zeptomolar (10^{-21} M) level (Kellner et al., 2019).

Along with novel CRISPR-Cas system mining in metagenome data, the mutagenesis technique aimed to identify variants with desired properties (Tong et al., 2022), such as higher on-target efficiency or lower off-target cleavage activity, allowing the expansion of CRISPR-based diagnosis platform.

1.3.7 Multiplex targeting

Multiplexing in CRISPR-Cas methods involves expressing multiple gRNAs or Cas proteins at once. It is deemed fundamental to serve the purpose in some methods, such as double nCas9 for precise gene knockout and targeting non-repetitive regions for visualization. In other methods, it facilitates better performance, such as more complete silencing or, in diagnosis-relevant methods, higher efficiency and larger sample size. Moreover, multiplexing allows interrogation of multiple genes at once, untangling the higher-order genetic interaction. Other applications of multiplex CRISPR-Cas methods include sophisticated gene circuit construction, such as an AND gate controlled by co-expression of two genes (McCarty et al., 2020).

Numerous gRNAs can be expressed either as individual sgRNAs or as processed gRNAs from a CRISPR array. For the former one, the most common method is to express each sgRNA under an individual promotor. For the latter one, one promoter is used to express the CRISPR array carrying multiple gRNA sequences. The transcribed array is processed by the ribozyme, Csy4 (one type of CRISPR-associated protein), or RNase when the sgRNAs are flanked with self-cleaving sequences, Csy4 sites, or tRNAs respectively (K. Xie et al., 2015; L. Xu et al., 2017; Ferreira et al., 2018; Zhang et al., 2019; Yuan & Gao, 2022). To date, more than 20 gRNAs can be multiplexed by processing a single transcript crRNA array with Cas9, Cas12, and Cas13 under the mechanism of the native CRISPR-Cas systems (Campa et al., 2019; Liao, Ttofali, et al., 2019; Ellis et al., 2021).

1.3.8 Large-scale and genome-wide screens

Large-scale and genome-wide scale functional screenings benefit from the simplicity of gRNA design and CRISPR-Cas systems, probing genotype-phenotype to comprehensively understand functional cellular programs (Schuster et al., 2019; Todor et al., 2021). Understanding the genotype-phenotype relationship unveils the unknown gene function, the causal genes of diseases, and the potential drug targets. The genotype-phenotype relationship can also serve as a roadmap for synthetic gene circuits.

The workflow of a CRISPR genome-wide screen starts with the library design, in which the requisite number of gRNAs is designed for each target gene, which is typically three to six. gRNA oligos are synthesized, amplified, and delivered into target cells. Each gRNA becomes a marker of the cell given that it diminishes with cell death and proliferates with cell reproduction. Therefore, for a screen that focuses on understanding gene essentiality, the fitness effect of knocking out or knocking down a gene can be observed through the read count of the gRNAs targeting the corresponding gene compared to the initial time point. The lower read count of the gRNAs suggests a stronger fitness defect of the target gene, while the higher read count indicates a weaker effect on cell proliferation. The read counts of gRNAs are obtained by sequencing the extracted plasmids carrying the gRNAs or the amplified gRNA region integrated in the genome.

There are two main types of screens: gain-of-function and loss-of-function. The gain-of-function screen involves mainly CRISPRa, which promotes the overexpression of the target gene. The majority of CRISPR screens focus on the loss of function. CRISPR-Cas9 is frequently used in the loss-of-function screens to knock out the target genes in eukaryotes (He et al., 2019; Shalem et al., 2014; Yilmaz et al., 2018), whereas CRISPRi is more popular in prokaryotes (Rock et al., 2017; Rousset et al., 2018; T. Wang et al., 2018; Cui et al., 2018; H. H. Lee et al., 2019; McNeil et al., 2021) as mentioned above. The ability of the base editor to introduce a premature stop codon was also tested on a genome-wide scale as an alternative for loss-of-function screening (P. Xu et al., 2021; Y. Liu et al., 2022).

Prior to CRISPR screens, Transposon sequencing (Tn-Seq) (van Opijnen et al., 2009) and Transposon Directed Insertion Sequencing (TraDIS) (Langridge et al., 2009) were two simple-to-use transposon-based approaches for high-throughput loss-of-function screening. Gene perturbation relies on the random insertion of the transposon in these approaches. A large library size is therefore required to reach a genome-wide resolution. Moreover, bottleneck effects mitigate the diversity of their library, resulting in the tendency to miss small genes (Chao et al., 2016). Transposon-based approaches can identify essential genes based on missing read alignments at the corresponding genome loci but lack the flexibility to investigate the gene dosage effect, which occurs when the number of gene products changes by, for example, regulating gene expression (Cain et al., 2020). In contrast, gRNAs can be designed carefully to target the small genes and expressed in an inducible and titratable manner with CRISPRi (Qi et al., 2013; Fontana et al., 2018). The tunable gene expression using titratable CRISPRi further supported the investigation of gene vulnerability (Bosch et al., 2021), which describes gene essentiality as a quantitative trait instead of a binary measure. Gene vulnerability provides an alternative explanation for inconsistent essentiality across strains and paves the way for more robust drug design because genes with high vulnerability are better targets than strain-specific essential genes. In addition, when multiplexing targeting extends to a genome-wide scale, it allows systematic analysis of genetic interactions and synergistic drug combinations (Han et al., 2017; Shapiro et al., 2017; J. P. Shen et al., 2017; Rauscher et al., 2018; Chow et al., 2019). Moreover, droplet-based technology bridges CRISPR screening and single-cell RNA sequencing (scRNA-seq). Single-cell CRISPR (scCRISPR) screenings (Dixit et al., 2016; Datlinger et al., 2017) allow a single-cell resolution of the transcriptomic response to perturbation. Due to the growing frequency of CRISPR screens, databases have been built to compare screens from different studies and laboratories, such as DepMap (https://depmap.org/portal/).

However, complications exist in the design and analysis of genome-wide screens (Hanna & Doench, 2020). First of all, 3-6 gRNAs per gene are typically required for a robust readout, whose design is discussed in more detail in the following sections. Secondly, sequence deconvolution, the step of mapping sequence reads to individual gRNA in the library, is crucial for downstream analysis. After normalizing the read count of gRNAs, log-fold change is calculated between conditions or time points to recover the guide-level score. Given that the guide scores might vary, correctly and confidently converting the guide scores to a gene-level evaluation can be challenging. Algorithms have been dedicated to this problem (Bodapati et al., 2020), and include statistical testing-based methods (Rousset et al., 2018; T. Wang et al., 2018), the robust ranking algorithm (W. Li et al., 2014), maximum likelihood estimation (MLE)-based (W. Li et al., 2015), and Bayesian methods (Allen et al., 2019). A supplementary set of tools such as GEMINI and Orthrus (Zamanighomi et al., 2019; Ward et al., 2021) addressed the need for combinatorial screening analysis.

In summary, despite focusing largely on eukaryotic systems, the variety and flexible scales of CRISPR-Cas tools propelled the development of many other fields such as drug targets for cancers and infectious diseases. The continuous efforts not only aim to enlarge the tool kit in both eukaryotic and prokaryotic systems but also optimize the developed methods to improve their efficiency.

1.4 Rational design is crucial for a successful CRISPR-Cas experiment

Despite the power of CRISPR-Cas tools, conducting a successful CRISPR-Cas experiment requires rational design after selecting a suitable method from the ever-growing CRISPR-Cas tool kit. gRNA design is the first step in a CRISPR-Cas experimental setup but comes with challenges. In the design of genome-wide screens, multiple gRNAs are required for each gene, which stems from variability in gRNA efficiency when targeting the gene of interest. A poor selection of gRNAs is therefore prone to an unreliable read-out for the downstream analysis. gRNA design considers both on-target efficiency and off-target specificity, the two important characteristics of the CRISPR-Cas mechanism. CRISPR-Cas systems were evolved to defend against invading alien sequences. Both the tolerance of mismatches and the secondary PAM sequences play a significant role in the robust immune response, but this strength turns into a weakness when CRISPR-Cas systems extend to genetic tools. Low on-target efficiency and unwanted off-target mutations due to the robustness of CRISPR-Cas systems are the primary concerns in practice. On-target efficiency and off-target specificity are often evaluated separately and respective tools were developed based on diverse experimental approaches. For clarity, I describe on-target efficiency and off-target specificity in two individual sections below. Given that a large part of my work involves on-target efficiency prediction, this section describes the experimental settings to study both on-target efficiency and off-target specificity and how the guidelines of guide design have been concluded, whereas section 1.5 focuses on machine learning methods to predict on-target efficiency.

1.4.1 On-target efficiency

Dozens of tools were developed in the past ten years to predict gRNA on-target efficiency (Hanna & Doench, 2020). Investigations for on-target efficiency often exploited CRISPR large-scale screens to gain insights into the determinants of on-target efficiency and design rules were derived as guidelines for gRNA selection. In the screens, given multiple gRNAs are designed for each gene, the differences in the measured activity in the gRNAs targeting the same gene are the source of analyzing the differences in on-target efficiency. Unfortunately, the design rules extracted in different studies can be both consistent and inconsistent, which likely resulted from the differences in the experimental methods for the activity measurements. As follows, I describe a few examples of the gRNA activity measurements for SpCas9 gene knockout and compare the extracted design rules.

The first example is Rule Set 1 (Doench et al., 2014), which was extracted from one of the earliest large-scale SpCas9 knockout screens. In this screen, two pools of sgRNAs were designed to target mouse and human cell surface marker genes respectively. To have a comprehensive landscape of on-target efficiency, all potential sgRNAs with NGG PAM on both strands of all exons for mouse cell surface marker genes and coding sequences for human cell surface marker genes were included. Cells carrying sgRNAs were stained and analyzed using fluorescence-activated cell sorting (FACS) followed by DNA amplification and sequencing. In this setup, the successful knockout would eliminate the cell surface marker and the cell would be sorted as unstained. Therefore, the sorted unstained cells were expected to carry the sgRNAs with high knockout activity, which also indicated high on-target efficiency. The enrichment of each sgRNA was defined as the fold difference in abundance of the sgRNA between stained and unstained populations, providing a quantitative measure of gRNA activity, thereby on-target efficiency. By analyzing the relationship between the target position and activity of sgRNAs, lower activity was observed near C’ terminus and in both 5’ and 3’ untranslated regions (UTRs). After removing these sgRNAs targeting low-activity regions, the remaining 1,841 sgRNAs were ranked by the fold enrichment values for each gene and then divided by the number of sgRNA of the target gene. This transformed activity, percent-rank, was better in comparing sgRNA efficiency across genes and used to analyze the sequence composition. While no significant difference was detected between targeting coding and non-coding strand, sgRNAs with less than 6 or more than 16 guanine or cytosine bases exhibited lower activity. Besides the 20 nucleotides in sgRNA sequences and the N position in the PAM, 4 nucleotides upstream of the target sequence and 3 downstream of the PAM sequence were explored based on the probability of sgRNA with percent-rank activity higher than 0.8. Guanine was found to be favored at the 20th position of sgRNA, while it is strongly disfavored at the position immediately downstream of PAM.

The second example is Rule Set 2 (Doench et al., 2016). A total of 2,549 sgRNAs targeting the coding sequences of eight genes known to contribute to resistance of one type of small molecules (vemurafenib, selumetinib, or 6-thioguanine) were screened in melanoma cells with the addition of one of the three small molecules. An effective sgRNA was expected to cause substantial growth defects with the addition of the small molecules. Plasmids extracted from the harvested surviving cells were sequenced, providing the log-fold change of sgRNA compared to control cells receiving no small molecule treatment. Combining the 1,841 sgRNAs from the study of Rule Set 1, targeting position was found to have a limited effect on the on-target efficiency. Further, a machine learning model was built to predict the normalized rank of sgRNA in each gene. The addition of the melting temperature of the DNA version of the sgRNAs was demonstrated to improve the model performance, but the quantitative relationship was not mentioned in the study. Rule sets 1 and 2 presented general guidelines for gRNA design from two different experimental read-outs, and the integration of datasets arrived at a better prediction of guide efficiency.

The third example is CRISPRscan (Moreno-Mateos et al., 2015). Both Rule Set 1 and 2 were extracted from experiments in either mouse or human cell lines, leaving the generalization of the rule sets in other model organisms unproven. To fill the gap, CRISPRscan exploited a screen of 1,280 sgRNAs targeting 128 genes in zebrafish embryos. On-target efficiency was measured as the percentage of indel-containing reads sequenced from the amplified 1,280 target regions. Moreover, the expressed sgRNAs and Cas9-encoding mRNA were directly injected into the embryos, providing a read-out independent of the transcription rate of sgRNA or Cas9. In a total of 35 sequence positions, 6 nucleotides upstream of the target sequence to 6 downstream of the PAM sequence, were examined with the log-odds score of nucleotide frequency in the top 20% efficient sgRNAs in each gene. Consistent with Rule Set 1, guanine was favored at the 20th position of sgRNA and disfavored at the position immediately downstream of PAM, whereas both guanine and cytosine were preferred at the N position of PAM. However, there were also novel observations: guanine was commonly enriched at positions 1 to 14 of sgRNAs, whereas thymine and adenine were overall depleted. Of note, all sgRNAs were designed with GG at the 5' end of the sgRNA. This design is specific to T7-derived sgRNAs and the higher indel efficiency of GG was demonstrated in another large-scale screen in zebrafish zygotes targeting 122 loci (Gagnon et al., 2014). It was also shown to be partially caused by the *in vitro* transcription of T7 RNA polymerase. Aside from Rule Set 2, the Moreno-Mateo score from CRSIPRscan was incorporated in the majority of popular gRNA design platforms. The three studies described above are decent representatives of how design rules have been extracted from diverse screens and analytic methods.

The last example is an earlier study (T. Wang et al., 2014). Although the study centered on the establishment of the CRISPR-Cas9 large-scale screening method and yet attempts were made in understanding the variation in gRNA efficiency according to the distribution, average, or median value of measured activity values. In the study, the authors proposed the method for pooled loss-of-function screens in human cells and established a library with 73,000 gRNAs targeting near the beginning of 7,114 genes. By sequencing the sgRNA after incubation for 12 doublings, the depletion of sgRNAs compared to the initial time point uncovered the genes essential to cell proliferation. A comprehensive sgRNA set targeting 83 ribosomal genes, most of which are essential for cell growth, allowed the analysis of on-target efficiency: the sgRNA activity was lower with very high- or low-GC content and when targeting the last exon or template strand.

While most tools have been developed in eukaryotic systems for SpCas9, design rules for other CRISPR-Cas systems are emerging to enhance the utility of other CRISPR-based tools, such as CRISPRi/a (Gilbert et al., 2014; Konermann et al., 2014; Smith et al., 2016), *Staphylococcus aureus* Cas9 (SaCas9) (Najm et al., 2018), Cas12a (H. K. Kim et al., 2018), and base editor (Hwang et al., 2018). Given that my main project involves the development of a predictive model for CRISPRi and both CRISPRi and CRISPRa are means to regulate transcription, I describe a few examples of gRNA activity measurements for CRISPRi/a and the design rules here.

The first example is from Gilbert and collaborators (Gilbert et al., 2014). They investigated the gRNA activity using the same tiling library in both CRISPRi and CRISPRa screens in human cells. The comprehensive library, consisting of 54,810 sgRNAs, was designed to target 49 genes that convey resistance to ricin. The comparison of the sgRNA abundance between ricin treatment and standard conditions was then used to evaluate the on-target efficiency. While the sgRNA activity anticorrelated between CRISPRi and CRISPRa screens, the region exhibiting the highest target efficiency differed due to the machinery of the Cas variants. For CRISPRi, the active target window for dCas9-KRAB extended from -50 to +300 bp relative to the transcription start site (TSS) and peaked at 50 to 100 bp downstream from the TSS, wider than the active target window of dCas9. For CRISPRa, the peak shifted to -400 ~ -50 bp relative to the TSS for dCas9-SunTag with VP64, agreeing with the transcription activation machinery of VP16 domains in VP64.

The second and third examples are studies using a different repressor or activator for CRISPRi and CRISPRa, resulting in different active target windows despite similar experimental setups. Smith and collaborators (Smith et al., 2016) developed an inducible CRISPRi system in yeast using dCa9-Mxi1 and found that targeting up to 200 bp upstream of the TSS showed the maximal gRNA activity by comparing the median activity of around 600 gRNAs targeting 25 genes in 50 bp scanning windows with 25 bp overlapping. Konermann and collaborators (Konermann et al., 2014) instead suggested targeting the -200 to +1 bp window leads to the optimal level of transcription activation with the synergistic activation mediator (SAM) based on the correlation of the performance of 96 gRNAs targeting 12 genes.

The last example from Radzisheuskaya and collaborators (Radzisheuskaya et al., 2016) suggests that annotation has an impact on the target window recommendation and the sequence context affects the guide efficiency. The authors proposed a correct annotation after comparing various annotation methods and pointed out that a significant difference between functional and non-functional gRNAs was only observed for Cap Analysis of Gene Expression (CAGE)-predicted TSS, where a 2-fold mRNA depletion was used as a threshold to define functional gRNAs. The gRNAs targeting from -50 to +150 relative to CAGE-peak was demonstrated to be significantly more effective than targeting outside of the window, while the region for maximum efficiency is 0 to +50 bp. Using CAGE to reanalyze the data from the first example (Gilbert et al., 2014), a much higher proportion of functional gRNAs was identified in −50 to +250 bp relative to the CAGE-peak. In the first example (Gilbert et al., 2014), the author suggested that sequence features have little impact on the gRNA activity except for the negative effect of nucleotide homopolymers and extreme GC content. But in the CAGE-based reanalysis, the predicted efficiency of functional gRNAs in the recommended target window using Sequence Scan for CRISPR (SSC) (H. Xu et al., 2015), which incorporates the sequence context of the gRNAs, was significantly higher than that of the non-functional gRNAs, suggesting the gRNA sequence features still play a role in on-target efficiency for CRISPRi.

Comparing design rules for SpCas9 and CRISPRi/a, despite the fact that Rule Set 2 concluded no preference in a specific region in the protein-coding sequences, it is recommended to target near transcription start sites for both CRISPRi and CRISPRa. Additionally, sequence context is of less importance in CRISPRi/a. Nevertheless, the optimal targeting windows for CRISPRi/a need to be tailored according to annotation, the Cas variants, and target organisms.

Given that available data from large-scale screens are accumulating, more comprehensive design rules have been extracted from published data with alternative analyzing methods. Here I include two other examples to describe how the publicly available data have been reanalyzed to extract novel design rules, aside from the reannotation mentioned above (Radzisheuskaya et al., 2016), The first example is the reanalysis of the data from Doench *et al.* (Doench et al., 2014) by comparing the top and bottom 20% of gene-wise ranked sgRNAs, highlighting the importance of secondary structure and sequence characteristics for SpCas9 (Wong et al., 2015). The higher accessibility of the PAM-proximal 3nt in gRNA, lower gRNA self-folding stability, and lower GC content distinguished the 20% most efficient gRNAs. Homopolymers, especially four consecutive guanines, were associated with poor activity. Adenine was surprisingly enriched in functional gRNAs despite the GC content effect, while GG and GGG were depleted. Preference for guanine and dispreference for cysteine at the 20th position of gRNAs were again detected.

Another example is from Xu and collaborators (H. Xu et al., 2015), who combined large-scale functional screen data in human and mouse cells from other two studies (Koike-Yusa et al., 2014; T. Wang et al., 2014) to identify reproducible sequence features across species and experimental designs.In the study, the authors first identified and classified essential genes into three sets: ribosomal, nonribosomal, and mouse embryonic stem cells (mESC). Secondly, gRNAs were split into functional and nonfunctional groups with an individual threshold in each set, followed by calculating the log odds ratio of the functional group for each sequence feature. Thirdly, gRNAs were randomly permuted to construct a null distribution, and the 95% confidence interval was used as the threshold to define whether the feature was reproducible across gene sets in addition to the concordant effects across gene sets. Consistent with previous studies, guanine was favored at the 3' end of gRNAs, and thymine was disfavored at the last four nucleotides of gRNAs. Novel features such as a preference for cytosine at the cleavage site of Cas9 were identified. In the same study, CRISPRi genome-wide data from Gilbert *et al.* were reanalyzed to reinforce the importance of sequence features. gRNAs targeting the 500 most essential genes in the CRISPRi screen were selected followed by grouping into efficient and inefficient gRNAs according to the activity scores of non-targeting gRNAs. This reanalysis landed on novel design rules for CRISPRi: purine was preferred in most positions of gRNAs; sequence preference was dominated by the PAM-proximal end of gRNAs.

In summary, sequence composition, GC content, targeting position, and secondary structure are proven to play a role in on-target efficiency although the feature importance varies due to cell lines, organisms, promoters, and analytic methods. This variation was also observed in the decreased prediction accuracy of design tools in other data sets (Haeussler et al., 2016). To improve the prediction accuracy of the on-target efficiency, numerous studies leveraged the power of machine learning models, which I will describe in more detail in section 1.5.

1.4.2 Off-target specificity

Quantitative measurement of the tolerance of mismatches, detection of unintended mutations, and characterizing the functional PAM sequences laid the foundation for scoring the off-target specificity. Quantitating the tolerance of mismatches depends on the change in sgRNA activity with a variable number and/or position of deletion, insertion, and mismatches (Hsu et al., 2013; Fu et al., 2013; S. W. Cho et al., 2014; Gilbert et al., 2014; Wu et al., 2014; Doench et al., 2016; X. Xu et al., 2017). Direct detection of unintended mutations is performed using *in vitro* or *in vivo* genome-wide assays, such as GUIDE-seq (Tsai et al., 2015), Digenome–seq (Richardson et al., 2016), SITE-seq (Cameron et al., 2017), CIRCLE–seq (Cameron et al., 2017), DISCOVER-Seq (Wienert et al., 2019), GOTI (Zuo et al., 2019), EndoV-seq (Liang et al., 2019), and SURRO-seq (Pan et al., 2022). These approaches depicted the landscapes of off-targets for Cas9, dCas9, and base editors in human cell lines, mouse cell lines, and plants, and observations were incorporated into scoring tools for off-target specificity. For example, up to 5 mismatches were observed in the off-targets (Fu et al., 2013); one mismatch in the PAM-distal region was better tolerated than that in the PAM-proximal region and gRNAs with off-targets harboring fewer than three mismatches should be avoided (Hsu et al., 2013); one mismatch in the 3' end had a negative effect on gRNA activity in CRISPRi while combinations of mismatches could abolish the activity (Gilbert et al. 2014). The identification of functional PAM was described in section 1.1.

Tools for scoring off-target specificity are either alignment-based or scoring-based. Alignment-based tools rely on the number and the position of the mismatches in the aligned sequences, such as CasOT (Xiao et al., 2014), CHOPCHOP (Montague et al., 2014), CRISPR-OFFinder (Bae, Park, et al., 2014), sgRNAcas9 (S. Xie et al., 2014), FlashFry (McKenna & Shendure, 2018) and Crisflash (Jacquin et al., 2019). Scoring-based tools are based on calculations from experimental data or machine learning model predictions, including MIT (Hsu et al., 2013), Cutting Frequency Determination (CFD) (Doench et al., 2016), CCTop (Stemmer et al., 2015), CRISPRoff (Alkan et al., 2018), CRISTA (Abadi et al., 2017), Elevation (Listgarten et al., 2018), and DeepCRISPR (Chuai et al., 2018). The MIT score and CFD score were included in most of the gRNA design platforms. While the MIT score was derived from more than 700 gRNAs, over 27,000 gRNAs were investigated for the CFD score. The CFD score was validated with GUIDE-Seq and was shown to outperform the MIT score. The comprehensive screen of the CFD score provided the percent activity values for each mismatch, deletion, or insertion at any position of 20 nt gRNA, allowing scoring potential off-targets with a single or combination of mutations. But the aggregation of multiple mutations was simply calculated by multiplying the individual percent activity values. CRISTA, Elevation, and DeepCRISPR are more recent machine learning-based tools to predict gRNA activity with mismatches, and all three exhibited higher accuracy than the MIT score and CFD score.

The combinatorial effect of the number and location of mutations in potential off-target sites dominates the guide sequence selection for high specificity. User-friendly tools are available for both off-target prediction and scoring, greatly simplifying the gRNA design to limit off-targets. In combination with other experimental settings, described in the following section, minimizing the off-target effects have been shown possible.

1.4.3 Other considerations in experimental design

Besides selecting the guide sequence carefully, on-target efficiency can be impacted by target genes, gRNA expression machinery, gRNA length, gRNA stability, and Cas9 loading. For target genes, in the study of Rule Set 1, targeting the N terminus of one gene exhibited noticeably weaker activity. The gene-specific low activity and gene-associated off-target effects might both be related to chromatin accessibility (Wu et al., 2014), given that low chromatin accessibility was shown to decrease the efficiency (Chari et al. 2015; Smith et al. 2016; Chuai et al. 2018; Labuhn et al. 2018; Kim et al. 2019). Combining the Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) and sgRNA activity, the chromatin accessibility and nucleosome occupancy explained the asymmetric target window relative to TSS in CRISPRi (Smith et al., 2016). Regarding gRNA expression machinery, it was shown that the dispreference of thymine towards the 3’ end of sgRNAs resulted from the recognition of the transcription termination signal by RNA polymerase III (Doench et al., 2014; Gagnon et al., 2014). The length of gRNA was first modified to increase the targetable positions by truncating the sgRNAs, but the truncated sgRNAs resulted in comparable or lower on-target efficiency depending on the target organisms (Gilbert et al., 2014; Moreno-Mateos et al., 2015; Smith et al., 2016). For gRNA stability, Gagnon and collaborators (Doench et al., 2014; Gagnon et al., 2014) improved the sgRNA indel efficiency by the direct microinjection of the Cas9-gRNA complex into the zebrafish zygotes. In line with this, Moreno-Mateos and collaborators (Moreno-Mateos et al., 2015) disentangled the sgRNA stability from expression level by direct injection of transcribed sgRNA and Cas9 encoding RNA and demonstrated that the sgRNA stability positively correlated with on-target efficiency, and the enrichment of guanine in stable sgRNAs suggested that G-quadruplex structure enhanced the sgRNA stability. Another study also showed the engineered sgRNA with a highly stable hairpin structure exhibited higher cleavage efficiency (Riesenberg et al., 2022). Although Moreno-Mateos and collaborators pointed out that loading sgRNA to Cas9 had a limited effect on the on-target efficiency, Wang and collaborators (T. Wang et al., 2014) found sequence composition near the 3’ end of sgRNA, a preference for guanine and a dispreference for uracil, played a determinant role in the affinity of sgRNA-Cas9 binding and showed that the affinity for Cas9 positively correlated with sgRNA efficiency.

Moreover, other experimental settings can be taken into consideration to limit undesired off-targets, such as the type of Cas protein, the ratio between the amount of Cas protein and gRNAs, and the incubation duration. For the selection of Cas protein, using nCas9 to induce SSB instead of DSB avoids the error-prone NHEJ (Ran, Hsu, Lin, et al., 2013; B. Shen et al., 2014) and can reduce off-target effects by 1,500 times. Engineered high-fidelity SpCas9 variants, such as SpCas9–HFI, HiFi Cas9, and Sniper–Cas9, are able to cut down the off-target effects to a negligible level despite some showing lower activity at certain loci (Kleinstiver et al., 2016; J. K. Lee et al., 2018; Vakulskas et al., 2018). Further, the duration of the execution can be fine-tuned to minimize the off-target effects using reversible CRISPR-Cas methods, i.e. CRISPRoff (Carlson-Stevermer et al., 2020), and anti-CRISPR (Acr) proteins (Marino et al., 2020). CRISPRoff incorporated a light-inducible switch by replacing two positions in the sgRNA with photocleavable residues. Turning on the switch can degrade the sgRNA to prevent new DSB formation, thus avoiding further undesired editing after the target is edited. The delivery of Acr protein AcrIIA4 to inhibit the Cas9 protein after 6 hours maintained the on-target efficiency while significantly abridging the off-target editing (Shin et al., 2017). Dedicated experimental methods like dual nCas9s using multiple gRNAs to locate the target position also aid the leaking off-target editing.

Overall, many attempts were made to optimize the recipe for gRNA design in various cell lines, organisms, and CRISPR-Cas systems, despite the generalization of this recipe being far from perfect. Given that off-target specificity is more controllable in combination with other experimental considerations, optimization of on-target efficiency prediction requires more effort to meet the needs of the expanding applications of CRISPR-Cas tools.

1.5 Machine learning methods in CRISPR-Cas gRNA efficiency prediction

After machine learning was first termed in 1959 (Samuel, 1959), the term “machine learning” now can be understood as a self-learning machine, a pattern-recognition model, or a model that predicts unseen data after training on appropriate data. The growing size and complexity of biological data quickly attracted the involvement of machine learning techniques, as was also the case for CRISPR-Cas gRNA efficiency prediction (Konstantakos et al., 2022). A fair amount of on-target efficiency prediction tools are machine learning model-based. These tools not only score the on-target efficiency but also shed light on the design rules and mechanisms of CRISPR-Cas systems. The majority of the established tools are designed for CRISPR-Cas9 system in eukaryotes, despite that the dedicated tools for other CRISPR-Cas systems are emerging, including CRISPRi/a (Calvo-Villamañán et al., 2020; Hawkins et al., 2020; H. Xu et al., 2015), SaCas9 (Najm et al., 2018), Cas12a (H. K. Kim et al., 2018; Luo et al., 2019), base editors (Song et al., 2020; Koblan et al., 2021), and Cas13 (Wessels et al., 2020; Krohannon et al., 2022; Metsky et al., 2022).

The flood of machine learning-related research raised concerns about how the machine learning models were trained (Whalen et al., 2021; Greener et al., 2022), given that its performance relies heavily on training data quality, feature set, model type, model evaluation metrics, and validation methods. As follows, I discuss these topics in the context of on-target guide efficiency.

1.5.1 Data collection and integration

The training data set plays a determinant role in the machine learning model development, considering that the model learns from the training data and its competence is consequently restrained by the data (Jain et al., 2020). Correctly predicting on-target efficiency demands informative data to understand the cause of variability of gRNAs targeting the same gene. In line with this, a better resolution of the variation landscape can be obtained by a large number of gRNAs per gene. Therefore, similar to design rule extraction, large-scale screen data were often exploited for machine learning model development, both genome-wide and tiling (targeting only a set of genes), given that more than five gRNAs per gene are commonly included. As described in section 1.4.1, diverse experimental and analytic methods offer different but equally effective read-out for quantitative gRNA efficiency. The screening data from, for example, Doench *et al*. (Doench et al., 2014, 2016), Koike-Yusa *et al*. (Koike-Yusa et al., 2014), Wang *et al*. (T. Wang et al., 2014), Moreno-Mateos *et al*. (Moreno-Mateos et al., 2015), Hart *et al*. (Hart et al., 2015) and Kim *et al*. (H. K. Kim et al., 2019) were used in one or more models.

However, due to the diversity in experimental setups, models trained on data from one experimental method exhibited poor generalization in data from another experimental method (Haeussler et al., 2016). For example, the model trained on U6-derived gRNAs resulted in undesirable accuracy in T7-derived data. Similarly, the choice of cell lines and organisms has an impact on the model performance. Therefore, multiple predictive models are often implemented in gRNA design platforms to suit users' needs. For instance, one can choose a model according to the cell line, organism, and promoter in the experimental design. But this does not guarantee a model available to any experimental setup. In contrast, improving the generalizability of the model can improves its applicability in other, even novel, experimental setups.

One way to improve the generalizability of the model is data integration. By including data from different sources, the model is opt to learn the concordance across datasets, alleviating the batch effects and the influence of methodological variety. For instance, Xu and collaborators (H. Xu et al., 2015) combined screen data from human and mouse cell lines (Koike-Yusa et al., 2014; T. Wang et al., 2014), whereas Doench *et al*. (Doench et al., 2016), Chuai *et al*. (Chuai et al., 2018) and Xiang *et al*. (Xiang et al., 2021) enlarged the training data from the human cell line with previously published data from another human cell line. Furthermore, Chari *et al*. (Chari et al., 2017) combined data from two Cas9 orthologs, SpCas9 and Sth1Cas9. In these studies, models trained on integrated data showed higher accuracy than that of individual datasets. Notably, the numeric differences due to experimental setups and batch effects need to be considered when integrating the datasets. For instance, Chuai *et al*. (Chari et al., 2017) applied a weighted sum of mean values of measured efficiency per experiment, gRNA, and all gRNAs to normalize the on-target efficiency, whereas Xiang *et al*. calculated the scaling factor using linear regression on overlapping gRNAs between datasets to rescale one dataset before training. Alternatively, Doench *et al*. transformed the quantitative gRNA efficiency to rankings in each target gene, which bypassed the measurement difference but sacrificed the comparability across genes and numeric resolution, given that the same ranking difference might indicate different efficiency variations across genes.

A problem- or CRISPR-Cas system-oriented collection of data is of great importance to developing a predictive and accurate model for gRNA efficiency, given the sensitivity of model performance in experimental choices. But the models can achieve better generalization by taking advantage of data integration.

1.5.2 Feature selection and engineering

There are two main types of models in machine learning, supervised and unsupervised learning models. Unlike unsupervised learning, which requires no label for the input data, supervised learning entails labeled examples to supervise the model. The labeled examples are structured as columns with specific feature names in tabular data. Hence, for supervised models, the construction of a feature set demands domain knowledge, and the selection of features can drastically change the model performance. Here, I describe the common features included in CRISPR-Cas gRNA efficiency prediction tools.

Sequence, thermodynamic, genetic, and epigenetic features are the primary features investigated in the gRNA efficiency prediction tools, considering the mechanism of CRISPR-Cas systems and their proven relevance to the on-target efficiency. But the exact feature sets vary in the prediction models (G. Liu et al., 2020), and the definitions or calculations of a specific feature might differ. Given that sequence context is essential for gRNA design, sequence features are contained –to the best of my knowledge– in each model without exception. Sequence features include position-dependent and position-independent features. Position-dependent features record the presence or absence of sequence patterns in specific positions. Position-independent features count the occurrence of sequence patterns of varied lengths. The length of 23, 30, or 35 mer for position-dependent features are often considered, where 23 mer includes only 20 nt gRNA and 3 nt PAM sequences, 30 mer includes an extra 4 nt upstream of gRNA and 3 downstream of PAM, and 35 mer includes an extra 6 nt upstream and downstream. Although most frequently only single nucleotides and di-nucleotides are enlisted, up to 4 adjacent nucleotide patterns were tested (Rahman & Rahman, 2017; Muhammad Rafid et al., 2020). In comparison, position-independent features are less often incorporated (Doench et al., 2016; Rahman & Rahman, 2017; Muhammad Rafid et al., 2020), except for the count of guanine and cytosine due to its connection to thermodynamic features (Doench et al., 2014, 2016; Wilson et al., 2018; D. Wang et al., 2019). Apart from GC content and GC counts, melting temperature, self-folding energy, and gRNA-target-DNA binding energy allow the estimation of the accessibility of gRNA and the gRNA-target-DNA binding stability (Wong et al., 2015; Doench et al., 2016; D. Wang et al., 2019; Xiang et al., 2021). Genetic features such as target position in protein-coding sequence have been considered in a few tools (Doench et al., 2016; Wilson et al., 2018), but in the form of amino acid cut position and percentage peptide. Considering that in-frame mutations might retain the protein function, tools such as Microhomology-Predictor (Bae, Kweon, et al., 2014) and Vienna Bioactivity CRISPR score (Michlits et al., 2020) integrated protein function prediction. Despite the fact that chromosome accessibility has been linked to on-target efficiency, limited tools have attempted to include this information (Kuan et al., 2017), probably due to the lack of available data.

In addition to the selection of features, feature preprocessing helps to increase the effectiveness of the learning process and improve model performance, which includes but is not limited to transformation, imputation, and feature reduction (Kotsiantis et al., 2006). Transformation methods are specific to feature types, for example, standard scaling for numeric features to remove the mean and one-hot encoding for categorical features. Without proper encodings, like one-hot encoding, sequence features cannot be used as input to the training given that only numeric values are accepted in machine learning algorithms. Imputation involves handling missing data by assigning i.e. average values. Feature reduction aims to remove less important features using either bottom-up or top-down methods (Abadi et al., 2017; Rahman & Rahman, 2017). One way to determine the reduced feature set is to train models with various reduced sets and evaluate the respective performance. Of note, the model used for feature reduction is independent of the final predictive model. Alternatively, after training a model with the full set, one can reduce the features based on the model interpretation, i.e. based on the feature importance (Calvo-Villamañán et al., 2020). Feature reduction also curtails the risk of overfitting, especially when the sample size is relatively small, thereby improving the generalizability of the model. For instance, Xu and colleagues (H. Xu et al., 2015) constructed a reduced, reproducible feature set from the full feature set incorporated in the linear regression models trained with three individual datasets. This reproducible feature set was chosen based on the feature concordance across models, which was measured by comparing the coefficients of the features from each model. Using the reproducible feature set, the model performance was improved.

Together, a careful selection of features prior to the model training deserves more attention. In particular, patterns in the gRNA efficiency differ across CRISPR-Cas systems. One simple example is that epigenetic features are less significant when applied in prokaryotic systems. While choosing the right features for the corresponding data, one should be aware of the necessity of feature preprocessing.

1.5.3 Model type and hyperparameter tuning

Another impactful factor in machine learning model training is model type (training algorithm). While the model types can be classified as supervised or unsupervised, as mentioned in the previous section, another way to classify them is as shallow or deep learning models (Chauhan & Singh, 2018). Shallow models, also known as traditional models, are models that are not neural network-based. Deep learning models are principally neural networks, which attempt to mimic the neural network in the human brain. Numerous machine learning algorithms have been applied to the on-target efficiency prediction problem. While the early-developed models for efficiency predictions are shallow model-based, deep learning models are increasingly popular in recent years.

Even though diverse algorithms have been shown applicable in on-target efficiency prediction, the algorithms are fundamentally different, with no consensus having yet been reached. The applied shallow models include linear regression (Moreno-Mateos et al., 2015), logistic regression (Doench et al., 2014), least absolute shrinkage and selection operator (LASSO) (Calvo-Villamañán et al., 2020), Elastic Net (H. Xu et al., 2015), support vector machines (SVM) (Wong et al., 2015; Chari et al., 2015; Rahman & Rahman, 2017), gradient-boosted decision trees (Doench et al., 2016), and random forest (Wilson et al., 2018). Linear regression models (Su et al., 2012) essentially attempt to fit a line to the data, thus relying on the linear relationship between the features and the training target, which is on-target efficiency in this case. In regression models, weights are assigned to each feature and the prediction is the weighted sum of feature values. Logistic regression is similar to linear regression but applies a logistic function to transform the values between 0 and 1 to predict the probability for classification problems. To avoid overfitting the training data, regularization was introduced to add penalties to the feature weights (Ying, 2019). Two major types of regularization are L1 and L2, penalizing the sum of absolute values of weights and the sum of the squared values respectively. While LASSO only incorporates L1 regularization, Elastic Net considers both. Instead of a line, SVM (Noble, 2006) fits a two-dimensional hyperplane, and gradient-boosted decision trees and random forests build decision trees (Kotsiantis, 2013). While a hyperplane is similar to lines, decision trees split the nodes and grow the branches with if-then rules. The end nodes are intuitively called leaves, the same for a tree. Each decision tree is considered one separate model, which is at risk of overfitting. Hence, methods like gradient-boosted decision trees and random forests combine the predictions from numerous decision trees to avoid overfitting and improve accuracy. Such techniques are termed ensemble methods (Dietterich, 2000). As for deep learning models, convolutional neural networks (CNNs) (Chuai et al., 2018; H. K. Kim et al., 2019; Xiang et al., 2021) and recurrent neural networks (RNNs) (D. Wang et al., 2019) have been applied. Both CNNs and RNNs are popular in genomics problems (Eraslan et al., 2019) because their assumptions can represent patterns and orders in the nucleic acid sequences respectively.

Although no particular model has been favoured, tree-based models have been shown to outperform linear regression models (Doench et al., 2016; Muhammad Rafid et al., 2020), given that the dominant features are sequence features and the linear relation between sequence features and on-target efficiency is weak, although the opposite has been argued due to the similar performance between tree-based models and linear regression models (Calvo-Villamañán et al., 2020). In contrast to shallow models, deep learning algorithms often demand a larger size of samples. The increasing availability of data in CRISPR-Cas systems explains the growing popularity of deep learning models, although Muhammad Rafid *et al.* (Muhammad Rafid et al., 2020) have argued that deep learning models did not outperform shallow models.

For each machine learning algorithm, hyperparameters are parameters that cannot be inferred from the training data, thus requiring prior knowledge to assign functional values (L. Yang & Shami, 2020; Bischl et al., 2023). Hyperparameters are crucial to the learning process and faulty hyperparameters can fail the learning process, leading to disastrous model performance. Different model types bear different hyperparameters. For example, a LASSO model requires the penalty factor and a deep learning model requires the number of hidden layers. For a random forest model, typical hyperparameters include the number of trees, the number of features to consider in splitting branches, the minimum number of samples in the leaves.

To select a satisfactory set of hyperparameters, grid search, random search, bayesian optimization, or tree-structured parzen estimators (TPE) can free one from the laborious manual process. Grid search looks for the exhaustive combinations of values in the search space, while random search (Bergstra & Bengio, 2012) looks for random combinations till it hits the set number of trials. Bayesian optimization (Shahriari et al., 2016) leverages the power of bayesian inference and sampling methods and examines the probability of the model performance given the hyperparameters. Similarly, TPE models (Bergstra et al., 2011) estimate the probability of hyperparameters given the model performance. Search space is however required for each method, in which a range or list of values is predefined for each hyperparameter to be tuned. Dedicated tools for hyperparameter tuning have been developed. One example is hyperopt (Bergstra et al., 2013), which uses both random search and TPE methods and is simple to use.

The decision on the model type and the corresponding hyperparameters is laboriously intensive but crucial. It is worth exploring the data structure and the feature values before an exhaustive search and comparison. Considering the difficulty, automated machine-learning techniques were developed to simplify this process, which will be described in section 1.5.5.

1.5.4 Model performance evaluation and validation

Feature and model selection rely on the resulting model performance (Raschka, 2018). Hence, how to evaluate the model performance is a critical decision. Ample metrics for either classification or regression models allow the evaluation from different perspectives, and metrics for classification and regression models are different.

On-target efficiency prediction has been shaped into both classification and regression problems. Therefore, I describe metrics for each type of problem here. In the classification problem, gRNAs are classified as either efficient or inefficient based on the ranking of each target gene. Often top quantile gRNAs are considered efficient and the bottom 20% or the rest were the opposite (Doench et al., 2014; Wong et al., 2015; Chari et al., 2017; Rahman & Rahman, 2017), or a cut-off in logFC values was applied (Chuai et al., 2018; Muhammad Rafid et al., 2020). The classification models, also called classifiers, output the prediction probability between 0 and 1. If the prediction probability is higher than 0.5, it is considered as predicted positive, otherwise, it is predicted negative. The accuracy score, the simplest metric for classifiers, is calculated based on the percentage of correctly predicted classes. Other metrics that take the class balance into account include the F1 score, ROC-AUC score, and Matthews correlation coefficient (MCC). The ROC-AUC score (Bradley, 1997) is the most commonly used in on-target efficiency classifiers. ROC stands for receiver operating characteristic and AUC for the area under the ROC curve. The ROC curve is constructed on all possible thresholds in the prediction probability to classify the predicted true or negative. From each threshold, the true positive rate and false positive rate are calculated and plotted as values on the x- and y-axis respectively. The connected dots from all thresholds are the ROC curve, and thus the AUC score provides a balanced and robust evaluation of the classification accuracy in each class. In the regression problem, the continuous numeric efficiency values were used as the training target. Despite that mean squared error (MSE) and R squared are routinely used to evaluate the regressors, the Spearman correlation between the true and predicted values is the most applied, given that the order of the prediction weighs more to advise the selection of gRNAs for each target gene.

In addition, the generalizability of the predictive model is worthy of attention, which measures how accurately the model predicts unseen data, possibly data from remote sources. While data integration amends the lack of generalizability, cross-validation allows the evaluation of model generalizability by training and testing on different subsets of the data and is also used in hyperparameter optimization (Raschka, 2018). The basic form of cross-validation is k-fold cross-validation (Rodriguez et al., 2010), the procedure of which is as follows: the samples are firstly divided equally into k folds; the model is trained on k-1 folds and tested on the one held-out fold; the training is repeated k times to train and test on different folds. Leave-one-out cross-validation is another cross-validation method, in which one sample is held-out in each iteration for testing.

How the samples are divided into folds also impacts the evaluation. A simple train-test split on all training samples provides the evaluation of accuracy in the overall landscape of gRNA efficiency (Moreno-Mateos et al., 2015). In comparison, splitting based on the target gene might be more practical in the case of gRNA efficiency prediction, given that the gRNAs are designed to target different genes and the unseen data mean novel genes. Therefore, the leave-one-gene-out method has been adopted to evaluate the models (Wong et al., 2015; Doench et al., 2016; Rahman & Rahman, 2017), in which gRNAs targeting one gene were held out as a test set while the remaining were used as the training set in each iteration. Further, leave-one-sgRNA-out, leave-one-cell-out, and leave-one-dataset-out have been applied (Doench et al., 2014; H. Xu et al., 2015; Chuai et al., 2018; Muhammad Rafid et al., 2020). However, these leave-one-out methods are still at risk of high similarity between train and test sets. To tackle this, Xiang and colleagues (Xiang et al., 2021) split the samples into partitions based on the hamming distance of the 30 mer extended sequences to keep the samples sharing high similarity in one partition, followed by cross-validation on these partitions.

In summary, choosing a suitable and practical metric for evaluation steers the model to a better performance on unseen data, and correctly splitting the data into train and test sets ensures precise assessment.

1.5.5 Automated machine learning (AutoML) technique

AutoML automates the model development by converting every step from raw data to the final model into an optimization problem (Waring et al., 2020). It is dedicated to developing a ready-to-use predictive model while reducing the dependency on knowledge of machine learning. Compared to a laborious search for an optimal model in combinations of methods in each step (such as feature preprocessing, model selection, etc.), one can save tremendous time by focusing on the best-performing models from autoML. Available autoML frameworks (Thornton et al., 2013; Feurer et al., 2015; Olson & Moore, 2016; Ledell & Poirier, 2020) however only include partial essential steps in the automation, such as preprocessing, model selection, and hyperparameter tuning. Among these frameworks, I will focus on auto-sklearn here, given that it won the first (Feurer et al., 2015) and second (Feurer et al., 2020) international AutoML challenge.

Auto-sklearn combines algorithm selection and hyperparameter optimization problems with Bayesian optimization, ensemble construction, and meta-learning. Bayesian optimization is used to select and optimize the algorithm, similar to solely hyperparameter tuning. Auto-sklearn includes configurations, which include 14 preprocessors, 15 estimators (classifiers or regressors), and over 150 hyperparameters, in the search space and constructs a probabilistic model to map the configurations to their performance. For each well-performed configuration, auto-sklearn creates a pipeline and evaluates it with a predefined metric from the user, such as MSE for regression models. In ensemble construction, pipelines are added to an ensemble with weight based on the performance of the pipeline. Combining the predictions from different pipelines in the ensemble can further minimize the prediction errors. Meta-learning (Vanschoren, 2018) is a means to leverage the knowledge of the existing datasets and their optimized configurations to speed up the learning process. Considering the limitation in time and computational resources, meta-learning was implemented in auto-sklearn to prioritize the configurations from existing datasets that are similar to the training data as a warm-start for Bayesian optimization. The similarity was calculated based on the distance of 38 meta-features learned from over 100 datasets. The meta-features include, for example, the number of samples, the number of features, and the distribution of feature values. Auto-sklearn is built around the most popular machine learning package scikit-learn (Pedregosa et al., 2011) and offers great flexibility. One can include only certain steps or model types for optimization within specific execution durations. More than one metric can be recorded, allowing exploration in suboptimal configurations. Further, the tool continues to be improved, and the second generation of auto-sklearn with faster speed and better performance on limited computational resources is already available for classification problems.

AutoML arises as an alternative approach to effectively develop machine learning models with less labor and likely better performance. However, more attention should be drawn to its implementation in the analysis of CRISPR-Cas systems considering the need for continuous model optimization for emerging CRISPR-Cas techniques.

1.5.6 Model interpretability

A machine learning model is more than a predictive tool to accept or refuse candidates. Interpretation of the model can unlock the machine learning black box and retrieve the predictive features that contribute to the model prediction (Lipton, 2018; Molnar, 2022). As mentioned in section 1.5.2, model interpretation advises an optimal feature set for better performance and higher training efficiency.

Whether to make high stakes decisions based on model interpretation is debatable (Rudin, 2019). It is clear that the interpretation of gRNA efficiency prediction models is less likely to involve high-stakes decisions and potentiates a deeper understanding of the CRISPR-Cas mechanisms. Commonly, design rule extraction, based on model interpretation, is dependent on the performance of the efficiency prediction tool, given that only a well-performing model is assumed to be built on correctly recognized patterns in the data. This dependence again underlines the importance of model performance evaluation.

Aside from the dependence on model performance, the interpretability varies across model types. For linear regression models, such as LASSO and elastic-net models, the higher absolute values of weights of each feature, also known as coefficients, can indicate the higher importance of features. And the sign of the coefficients, positive or negative, indicates how the feature impacts the prediction. In a previous study, the sign of the coefficients was used to determine whether a feature was concordant across models (H. Xu et al., 2015). But the direct interpretation of coefficients requires mean-centered numeric features and properly coded categorical features, otherwise, the difference in the scales of feature values confounds the comparability. For tree-based models, one can measure how many splits across all trees depend on each feature. The higher count leads to higher Gini importance, suggesting the higher importance of the feature (Doench et al., 2016). Gini importance is also the most common method to interpret a tree-based model.

Beyond these model-derived characteristics to explain the feature importance, model-agnostic methods (Molnar et al., 2022) spare more flexibility, not limited by model types. Model-agnostic methods can also provide a description of the average behavior of the model, such as the average contribution of a feature in all predictions, and an explanation of individual prediction, such as how the prediction is derived for a sample. One example is SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017), which is a game theory-derived optimal Shapley value. In game theory, the contribution of each player can be measured by their presence or absence in a game. When adopted to interpret a model, the feature is the player, and when the feature value is known it is present. TreeSHAP (Lundberg et al., 2020) was developed for tree-based models. Xiang and colleagues (Xiang et al., 2021) built a gradient-boosting regression model and used SHAP to analyze the feature importance, even though the proposed predictive tool was a CNN-based model. By treating the pixels in the images as a group, which is similar to a feature in supervised models, the application of SHAP extends to deep learning models. Wang and colleagues (D. Wang et al., 2019) incorporated DeepSHAP, a function in the shap package (Lundberg & Lee, 2017) specific for deep learning models, to interpret the position-specific nucleotide contributions in guide efficiency prediction and exploited the sum of SHAP values to understand the effects of repetitive nucleotides. For deep learning models, besides SHAP, saliency maps, also known as pixel attribution, were applied to highlight the pixels that support a certain image class. By treating four bases as channels in pixels, Chuai and colleagues (Chuai et al., 2018) managed to adopt the saliency map for CNN model interpretation.

Aside from facilitating model optimization, model interpretability enhances our confidence to make decisions on the model predictions, enabling the use of machine learning techniques to their full potential.

Together, applying machine learning gradually became the regular practice in predicting on-target efficiency. Previous models presented the diversity and complexity of the model development, albeit focusing mostly on the CRISPR-Cas9 system in eukaryotes. The field of machine learning is continuously growing, and novel methods, such as autoML, lead to a promising future of effective model optimization for CRISPR-Cas9 and other CRISPR-Cas systems and techniques.

1.6 Aim of study

The past ten years experienced an explosion of CRISPR-related studies. The continuous discovery of CRISPR-Cas systems inspires the active development of CRISPR-based tools, which revolutionized research in many fields, such as cancer and infectious disease. The application of CRISPR-based tools in bacteria started more slowly than in eukaryotic systems but CRISPRi quickly attracted attention in functional interrogation and synthetic gene circuit design. Efficient transcriptional regulation through CRISPRi however demands rational gRNA design, for which effective on-target efficiency prediction tools are of significance, especially considering that thousands of gRNAs are routinely employed in large-scale screens. Unfortunately, despite dozens of tools available for gRNA on-target efficiency prediction, the majority have been developed for the CRISPR-Cas9 system in eukaryotes and their lack of robustness has been reported (Haeussler et al., 2016). To date, only one LASSO model (Calvo-Villamañán et al., 2020) has been devised for CRISPRi guide design in bacteria while another model Mismatch-CRISPRi (Hawkins et al., 2020) focused on predicting titratable gene expression from mismatch-harboring gRNAs instead of selecting efficient gRNAs. Efficiency prediction tools for other CRISPR-Cas systems in bacteria are also under-investigated. Therefore, more work needs to address the need for effective guide design in bacteria for CRISPRi and other CRISPR-Cas systems or techniques.

The work in this thesis can be divided into two parts. In the first part, I describe a machine learning approach exploiting multiple CRISPRi genome-wide essentiality screens for CRISPRi guide efficiency prediction in bacteria. The aim is to improve the predictive accuracy of the machine learning model by leveraging the advantages of autoML, data integration, and model interpretability. As the model interpretation suggested dominant effects of gene-specific features, I subsequently made further attempts to segregate the gene and guide effects to build a more reliable model. In the second part, I describe the applications of the machine learning approach developed in the first part in the analysis of three distinct CRISPR-Cas genome-wide screens to probe the robustness and applicability of the approach across CRISPR-Cas systems and bacteria. The analysis of these screens also aims at understanding the mechanism of CRISPR-Cas systems, extracting design rules, and improving the CRISPR-based tools.

2. Methods

All codes necessary to reproduce these results in the thesis are available at: https://github.com/yanyyyy3/PhD_thesis.

2.1 Methods for Result section 3.1

2.1.1 Training datasets

I collected the data from three previous CRISPRi genome-wide essentiality screens in *Escherichia coli* (*E. coli*) K12 MG1655 (Cui et al., 2018; Rousset et al., 2018; T. Wang et al., 2018). The sequence, targeted gene, gene position, and fitness effect of each gRNA were retrieved from the supplementary information of each study. Gene sequences and positions were updated to be consistent with the latest reference genome version (NCBI: NC_000913.3). I discarded gRNAs from the Wang data set previously removed as having insufficient read counts (T. Wang et al., 2018) or sequences from the Rousset and Cui datasets that differed from the reference sequence due to differences in the genome versions. 8099 gRNAs targeting the coding-strand within the coding regions of essential genes were extracted in total from all three datasets. I removed genes with less than 5 gRNAs in each dataset to stabilize estimates of median gRNA activity scores, resulting in 7400 gRNAs in total. This included 1618 gRNAs targeting 171 genes in E75 Rousset/E18 Cui and 4164 targeting 300 genes in Wang.

2.1.2 Feature engineering

A Python script (feature_engineering.py on GitHub) was used to compute 137 sequence, thermodynamic, genomic, and transcriptomic features (**Table S1**). 30 mer extended sequence, 4 bp upstream of gRNA to 3 bp downstream of PAM, were one-hot encoded to 120 (30 × 4) features. Thermodynamic features including minimum free energy for different interactions (the hybridization of gRNA and target DNA, the hybridization of the seed region of gRNA and target DNA, the homodimer of gRNA, and the monomer of gRNA) were computed using the ViennaRNA Package (Lorenz et al., 2011): RNAduplex (version 2.4.12) for RNA:RNA hybrids; RNAduplex (version 2.1.9h) for DNA:RNA hybrids (Lorenz et al., 2012); RNAfold (version 2.4.12) for single RNA folding. The seed region was defined as the 8 nt PAM-proximal region in the gRNA. The CRISPRoff score (deltaGB) was calculated using the energy function in the CRISPRoff pipeline v1.1.2 (Alkan et al., 2018) with ViennaRNA Package version 2.2.5 (deltaGB_calculation.py on GitHub). Homopolymer was defined as the number of consecutive nucleotides in gRNA sequences. Genomic features including gene and operon organizations were based on the reference genome, essential genes as determined in the Keio collection (Baba et al., 2006), and transcriptional unit definitions from RegulonDB (Tierrafría et al., 2022). Transcriptomic data including gene expression levels across growth at ten different ODs were obtained from a previous study (Conway et al., 2014). Minimal or maximal expression levels were calculated across the range of ODs until the growth phase when cells were collected in each CRISPRi screen: OD 1.4 for the Wang dataset, and all ODs for the Rousset and Cui datasets. The codon adaptation index (CAI) for each gene was calculated using CAIcal (http://genomes.urv.es/CAIcal/) (Puigbò et al., 2008). The resulting feature sets are available on GitHub.

2.1.3 Cross-validation for machine learning methods

Paired-wise hamming distance of 5602 unique 30 mer sequences was calculated with hamming function from SciPy (version 1.10.0) (Virtanen et al., 2020). The median hamming distance is 21 while the minimum is 6.

To evaluate the models for depletion prediction, training and test sets were split guide-wise based on unique gRNA sequences. 10-fold cross-validation was used to evaluate model performance. 10 test sets, with the number of samples ranging from 786 to 855 targeting 245 to 258 genes, were kept identical regardless of training data. The minimum hamming distance between the train and test sets in each iteration or among test sets was 6. For each iteration, the Spearman correlation between measured depletion values and predicted values for all test samples was calculated.

To evaluate the models for gene-effect segregation (mix-effect random forest (MERF) and median subtracting (MS) models), training and test sets were split gene-wise based on gene identifier. 10-fold cross-validation was used to evaluate model performance. 10 test sets, with the number of samples ranging from 624 to 855 targeting 30 or 31 genes, were kept identical regardless of training data. The minimum hamming distance between the train and test sets in each iteration or among test sets was 7. For each iteration, the Spearman correlation between measured depletion values and predicted values for each held-out gene was calculated. For MERF, the predicted values were from the fixed-effect model.

The gRNAs used in each train-test split are described in **Table S2**.

2.1.4 Predictive models for depletion prediction

The automated machine learning toolkits auto-sklearn (version 0.15.0) (Feurer et al., 2015) and H2O (version 3.38.0.4) (Ledell & Poirier, 2020) were used to develop optimized machine learning regression models. For auto-sklearn, the AutoSklearnRegressor function was used and all possible estimators were included. The following parameters were used: “time_left_for_this_task” = 3600, “per_run_time_limit” = 360, “resampling_strategy” = ‘cv’, ensemble_kwargs = {"ensemble_size": 1}, “resampling_strategy_arguments” = {“fold”: 10}, “metric” = autosklearn.metrics.mean_squared_error, include = {'feature_preprocessor' : ["no_preprocessing"]}. To optimize hyperparameters for the random forest model, "regressor":["random_forest"] was added to the dictionary for parameter include. Feature types for each feature were listed in **Table S1**. The selected models were saved and used with scikit-learn (version 0.24.0) for downstream analysis. For H2O, the "StackedEnsemble" algorithm was excluded and parameters "max_runtime_secs = 0" and "seed = 1" were used. If not otherwise specified, parameters were left as default.

Simple linear regression, LASSO, elastic net, SVR, random forest, and histogram-based gradient boosting models were trained using scikit-learn.

2.1.5 Segregation of guide and gene effects with MERF and MS models

MERF models were trained using package merf (version 1.0) (Hajjem et al., 2014). Hyperparameters for the final fixed-effect random forest model were optimized using hyperopt (version 0.2.5) (Bergstra et al., 2013). Search space included: 'bootstrap' either True or False, 'n_estimators' from range 50 to 1000 with a step of 10, 'max_features' from range 0 to 1, 'max_depth' from range 2 to 30 with a step of 1, 'min_samples_leaf' from range 1 to 20 with a step of 1, 'min_samples_split' from range 2 to 20 with a step of 1. The objective function was the highest median gene-wise Spearman correlation of a 5-fold cross-validation split on the target gene (for more details, see MERF_crispri.py on GitHub). 128 guide-specific features were assigned as fixed effects, while 9 gene-specific features were assigned as random effects. The random effect matrix was standardized with the default StandardScaler function from scikit-learn. 301 unique gene IDs were used as cluster IDs. To train simplified models excluding transcriptomic measurements (**Figure S1.8E**), CAI value, gene length, gene GC content, and dataset were included for the random-effect model.

For MS models, logFC values were scaled to integrate the datasets, as an adaptation of a previously applied data fusion method (Xiang et al., 2021). First, the mean of logFCs of E75 Rousset and E18 Cui were calculated and used as the scaled logFC (**Figure S1.4B-D**). Then linear regression was performed between the logFCs in Wang and scaled logFCs in E75 Rousset for 378 overlapping gRNAs. All of the logFCs from Wang were then scaled by the fitted slope and intercept (**Figure S1.4B**). The 378 overlapping gRNAs in Wang were excluded in the subsequent training for MS models. Activity scores were calculated by subtracting the scaled logFC of each gRNA from the median scaled logFC for each gene across all 3 datasets (**Figure S1.4A**).

MS models were trained with guide-specific features to predict activity scores for each gRNA (**Figure S1.4A**). The hyperparameters of the MS random forest model were the same as for MERF, while those of the LASSO model were optimized using hyperopt with a search space for alpha ranging from 0 to 0.1 and 100 trials. Deep learning models were trained using pytorch (version 1.8.1) (Paszke et al.,

2019) and pytorch-lightning (version 1.5.10). For our custom 1D CNN model, sequence features were processed using 1D convolutional layers and later concatenated with other guide features (**Figure S1.5**). Concatenated features were further processed with fully connected layers. Three 1D convolutional layers were implemented sequentially with input channels 4, 64, and 64, output channels 64, 64, and 32, kernel size 5, 3, and 1, and stride 2, 2, and 1 respectively. For fully connected layers, output dimensions are 128, 64, 32, and 1 (which is predicted gRNA efficiency). The first three fully connected layers are accompanied by batch normalization (Ioffe & Szegedy, 2015), ReLU, and dropout (Srivastava et al., 2014) (p=0.5). I trained the model using AdamW (Loshchilov & Hutter, 2017) optimizer with learning rate of 0.001 and batch size of 32. For CRISPRon, CGx was implemented with eight non-sequential guide features (distance features, thermodynamic features, etc.) concatenated to processed sequential features. To test the effect of incorporating the deltaGB score (Alkan et al. 2018), the four thermodynamic features were replaced with deltaGB, resulting in concatenating five non-sequential guide features to the processed sequential features. I trained the model using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and batch size of 32. I additionally tested a learning rate of 0.0001 and batch size of 500 as used in the original CRISPRon implementation (Xiang et al. 2021), but saw only minor differences in performance.

The trained fixed-effect models from MERF were used to predict gRNA efficiency (i.e. in cross-validation and validation with independent data), while the trained MS models were directly used, requiring only features associated with the guide sequence.

2.1.6 Model interpretation

Tree-based models, including depletion prediction models, the fixed-effect models from MERF, and the MS random forest model, were interpreted using TreeExplainer from the python shap package (version 0.39.0) (Lundberg et al., 2020).

SHAP values were calculated using the 'shap_values' function in TreeExplainer with 80% of the samples for depletion prediction models and all samples for the rest. SHAP value plots were generated with the 'summary_plot' function in shap.

SHAP interaction values were calculated using the 'shap_interaction_values' function in TreeExplainer with 1000 guides. Absolute SHAP interaction values were averaged over 1000 samples. The rank of interaction was obtained based on the sorted mean absolute SHAP interaction values across all unique feature pairs. To compare interaction effects to expectations based on single-feature SHAP values, four feature combinations were considered: both absent (-/-), only the first feature present (+/-), only the second feature present (-/+), and both present (+/+). For the top 5,000 interacting feature pairs, the SHAP values for each feature in samples with each combination of features were extracted. For each feature pair (F1 and F2), the expected value for +/+ was calculated as the sum of the median F1 SHAP values for +/- samples with the median of F2 SHAP values for -/+ samples, while the expected value for -/- was calculated as the sum of the median F1 SHAP values for -/+ samples and the median of F2 SHAP values for +/- samples.

2.1.7 Strains and growth conditions

All strains, plasmids, and primers are listed in Supplementary **Table S15, S16, and S17**. *E. coli* cells were grown in Lysogeny Broth (LB) (10 g/L NaCl, 5 g/L yeast extract, 10 g/L tryptone) at 37 °C with shaking at 250 rpm. To maintain plasmids, the antibiotics ampicillin, chloramphenicol, and/or kanamycin were added at 50 µg/mL, 34 µg/mL, and 50 µg/mL, respectively as necessary. For screening experiments, *E. coli* MG1655 was grown in M9 minimal medium (1x M9 salts, 1 mM thiamine hydrochloride, 0.4% glucose, 0.2% casamino acids, 2 mM MgSO4, 0.1 mM CaCl2) supplemented with the appropriate antibiotics.

2.1.8 Validation of GFP silencing by flow cytometry

To investigate gene repression efficiency, 19 sgRNAs were selected to target the coding strand of a degfp reporter gene at different positions in *E. coli* BL21(DE3) (**Table S15**). Cells were initially transformed with three compatible plasmids encoding dCas9, a degfp-targeting sgRNA, and a deGFP reporter (**Table S17**). For normalization purposes, a positive control strain harboring a non-targeting sgRNA and a negative control strain lacking the degfp encoding reporter plasmid was included. Overnight cultures of cells harboring the above-mentioned plasmids were back-diluted to optical density OD_{600} ~0.01 in LB medium with ampicillin, chloramphenicol and/or kanamycin and incubated with shaking at 250 rpm at 37 °C, until reaching an OD_{600} of 1. Cultures were then diluted 1:25 in 1x phosphate-buffered saline (PBS) and analyzed on an Accuri C6 flow cytometer with C6 sampler plate loader (Becton Dickinson) equipped with CFlow plate sampler, a 488-nm laser, and a 530+/− 15-nm bandpass filter. Forward scatter (cutoff of 11,500) and side scatter (cutoff of 600) were used to eliminate non-cellular events. The mean green fluorescence value (measured by the FL1-H channel) across 30,000 events within a gate set for *E. coli* was used for further analysis. The log fold repression of each gRNA was calculated as the ratio between the difference in fluorescence values between the gRNA and negative control, and the difference between the positive and the negative control, followed by log transformation. The mean log fold repression across three replicates was compared to predicted values from the machine learning models (**Table S10 and S12**).

For experiments with *Salmonella* Typhimurium, the procedure was similar, but cells were grown until an OD_{600} of ~0.8 before analysis on an Accuri C6 flow cytometer. To eliminate non-cellular events, the forward scatter (cutoff of 10,000) was used and the mean green fluorescence value (FL1-H) across 30,000 events within a gate set for *S.* Typhimurium was used for data analysis as described above across four replicates (**Table S10 and S12**).

2.1.9 Generation of the sgRNA library

9 genes (purA, purC, purD, purE, purF, purH, purK, purL, and purM) in purine biosynthesis pathway in *E. coli* MG1655 (NCBI: NC_000913.3) were selected for the saturating screen in M9 minimal medium. All possible 20 nt gRNAs with NGG PAM, GC content between 30% to 85%, and without BbsI restriction enzyme cut site were included in the library, resulting in 750 gRNAs (**Table S13**). The minimum and median hamming distance of 30 mer sequences between the 750 gRNAs and the training data from three essentiality screens were 7 and 21. To include 50 randomized non-targeting gRNAs as the negative control, sequences were firstly randomly generalized with the same length as the target gRNAs in the library, followed by aligning to all possible NGG gRNAs in the target genome. Sequences with more than 11 bp matches were removed. After filtering with 5 consecutive nucleotides and BbsI restriction sites, 50 of the remaining sequences were included.

For the sgRNA library, plasmid DC512 served as a backbone, following a previously established protocol (Liao, Slotkowski, et al., 2019). To generate a library with 800 sgRNAs (including 50 non-targeting sgRNAs; **Table S13**), 800 forward and reverse oligonucleotides each encoding one spacer and a 4-nt junction, were synthesized as an oPool (1,600 oligos at 10pmol/oligo) by Integrated Device Technology (IDT). The same 5′ and 3′ assembly junction sequences were used for all spacer pairs leading to the same integration site within the backbone (5′ TAGT overhang at the 5′ end and a 3′ AAAC overhang at the 3′ end). Supplementary **Table S17** contains the specific oligonucleotides and assembly junctions used for the library generation. The oligos were phosphorylated and annealed to form dsDNA with a 5′ and 3′ overhang. The steps of phosphorylation and annealing were combined and conducted in one pot, by adding 8,000 fmol of the oPool and 1 µl T4 polynucleotide kinase (10 units) to 5 µl 10 × T4 ligation buffer and then, adding water until reaching a final volume of 50 µl. After mixing briefly by pipetting the mix was incubated at 37°C for 30 minutes in a thermocycler and then incubated at 65°C for 20 minutes in a thermocycler to heat-inactivate the kinase. For the annealing of the forward and reverse oligo pairs, the following thermocycler steps were added: 95°C for 5 min, 94°C for 15 s, decrease by 1°C, and hold for 30 seconds for 79 cycles. For integrating the dsDNA inserts into DC512, 400 fmol of the dsDNA, 20 fmol of backbone plasmid, 0.5 µL of T4 ligase (1,000 units), and 1.5 µL of BbsI (15 units) were added to 2 µL of 10x T4 ligation buffer, then water was added to reach a total volume of 20 µl. A thermocycler was used to perform 35 cycles of digestion and ligation (37 °C for 2 min, 16 °C for 5 min) followed by a final digestion step (60 °C for 10 min) and a heat inactivation step (80 °C for 10 min). After

NdeI digestion (37°C, 1h) of the ligation mix to remove any remaining original backbone plasmids and subsequent ethanol precipitation, 10 µl of the ligation mix was transformed into electrocompetent *E. coli* NEB10 beta (NEB, Ipswich, MA, USA), following the manufacturer's instructions. After transformation and recovery in 1 ml SOC for 1 h at 37 °C with shaking at 250 rpm, different dilutions of the recovered cells were plated on LB agar containing the appropriate antibiotic and incubated for 16 h to check the number and color of the resulting colonies (ensuring a ~58X coverage). The rest of the recovered culture was added to 100 mL LB media containing the appropriate antibiotic and incubated at 37 °C with shaking at 250 rpm to $OD_{600} \approx 1$. Cells were harvested by centrifugation and subjected to plasmid extraction. Sanger sequencing was used to validate the library plasmid DNA.

2.1.10 Screening experiment

E. coli strain MG1655 was initially transformed with a dCas9 encoding plasmid (2.0 kV, 200 Omega, and 25 µF). The resulting strain SG332 was then transformed with the sgRNA library by electroporation and recovered in 900 µl SOC for 1.5h at 37 °C with shaking at 250 rpm. Different dilutions of the recovered cells were plated on LB agar containing the appropriate antibiotics and incubated for 16 h to check the number of the resulting colonies (~56^5 colonies). The recovered culture was back-diluted to OD_{600} 0.01 in LB medium with appropriate antibiotics and incubated at 37 °C with shaking for 13 hours. Subsequently, 5 mL of the culture was sampled and the library was extracted by miniprep (Nucleospin Plasmid, Macherey-Nagel) to obtain the initial sgRNA distribution. The calculated amount of culture to reach OD_{600} 0.01 in 50 ml M9 minimal medium, was sampled and washed twice with M9 minimal medium to remove traces of the LB medium. The culture was incubated at 37°C with shaking until it reached OD_{600} 1, allowing ~6 replications. 5 ml of the culture was sampled at OD_{600} 0.2, OD_{600} 0.6, and at OD_{600} 1, and the library was extracted by miniprep. The experiment was performed in duplicate starting from two independent transformations of MG1655 with the plasmid library.

2.1.11 Library sequencing

The sequencing library was generated using the KAPA HiFi HotStart Library Amplification Kit for Illumina® platforms (Roche) and the primers listed in Supplementary **Table S17**. The first PCR adds the first index. The second PCR adds the second index and flow cell-binding sequence. The amplicons of the first and second PCR reactions were purified using solid-phase reversible immobilization beads (AMPure XP, Beckman Coulter) following the manufacturer's instructions to remove excess primers and possible primer dimers. The sequencing library samples, with the required DNA concentrations ranging from 100 pg - 200 ng in a total volume of 10 µL, were submitted to the HZI NGS sequencing facility

(Braunschweig, Germany) for paired-end 2 × 50 bp deep sequencing with 800,000 reads per sample on a NovaSeq 6000 sequencer.

The resulting raw sequencing data have been deposited in GEO under accession GSE196911.

2.1.12 Sequencing data processing

Paired-end reads were merged using BBMerge (version 38.69) (Bushnell et al., 2017) with parameters “qtrim2=t, ecco, trimq=20, -Xmx1g”. Merged reads with perfect matches were assigned to the gRNA library using a Python script. After filtering guides for at least 1 count per million in at least 4 samples, read counts of each gRNA were normalized by factors derived from non-targeting guides using the trimmed mean of m-values method in edgeR (version 3.28.0) (Robinson et al., 2009). An extra column was added to the design matrix to capture batch effects between the two replicate experiments. Differential abundance (log fold change, logFC) of gRNAs between time points and the input library was estimated using edgeR, and a quasi-likelihood F test was used to test for significance after fitting in a generalized linear model. Spearman correlation between the logFC and the predicted score was calculated for each of the nine genes in the purine biosynthesis pathway. For positive predictive value calculation, gRNAs with fold change values within N folds of the maximum fold change value in each gene were classified as positive, while the best 5 predicted gRNAs were defined as predicted positive. The positive predictive values were calculated with TP/(TP+FP) for all gRNAs at each time point for each fold (N=1.5 - 5 with a step of 0.5, TP = True positive, FP = False positive).

2.1.13 Score functions from previous studies

Adapted Python scripts (gRNADesigner.py and DeepSpCas9.py on GitHub) from the source codes of gRNA Designer (Doench et al., 2016) and DeepSpCas9 (Kim et al. 2019) were used to calculate the predicted scores. The source code of TUSCAN (https://github.com/BauerLab/TUSCAN) (Wilson et al., 2018) was directly used. For SSC (H. Xu et al., 2015), I used the web-based application at http://crispr.dfci.harvard.edu/SSC/ with option CRISPR inhibition or activation and 20nt gRNA length. For the Pasteur model (Calvo-Villamañán et al., 2020), the trained LASSO model was saved from the adapted Python script (Pasteur_model.py on GitHub) based on the jupyter notebook on their GitLab (https://gitlab.pasteur.fr/dbikard/ecowg1).

2.2 Methods for Result section 3.2.1

2.2.1 Base-editing in *E. coli* with ScBE3

The rAPOBEC1 deaminase was joined to the n-terminus of *Streptococcus canis* (ScCas9) D10A by a 16AA linker. This construct, SPC914, was co-transformed by electroporation into *E. coli* MG1655 alongside a constitutively expressed guide RNA targeted to a codon which could produce a stop codon or disrupt a start codon. The targets were flanked by the minimal NNG PAM required by ScCas9. These transformants were plated on LB with glucose and the appropriate antibiotics. Colonies were picked and inoculated into 2mL of LB media shaking at 37°C and 250 rpm overnight with the appropriate antibiotics. The next day, cultures were diluted 1:500 into LB supplemented with antibiotics and 1mM IPTG, and 0.2% L-arabinose for induction of the base editor, and cultured for 8h. An aliquot of these cultures was plated for analysis, and another aliquot was diluted 1:500 for 16h of further induction and culturing before plating. Base editing frequency was measured by the fraction of colonies having a white or blue color on X-gal indicator plates, and/or sanger sequencing.

2.2.2 Library design

After taking over this project from Bozena Mika-Gospodorz, I further optimized an in-house Python script for library design. The reference genome and annotation of *E. coli* K12 MG1655 (NCBI: NC_000913.3) were used for gRNA library design. For the essentiality screen, I first located the targeting codons-start codon ATG, TGG, CAA, CAG, and CGA-in the first half of the coding region and searched for NNG PAM given the defined activity window from the position 5 to 8 from the 5' end of 20nt gRNA. gRNAs (with flanking primer sequences) containing longer than 4 consecutive nucleotides or BsmBI restriction sites were removed. For genes with more than 15 gRNAs, a maximum of 15 gRNAs per gene was included. Consequently, excessive gRNAs were filtered step by step based on off-targets and GC contents until no more than 15 gRNAs per gene were included. In the first 4 steps, gRNAs with off-targets harboring 0 to 3 mismatches were removed. In the last step, gRNAs with GC content outside of the range of 30 to 85 were removed. If all gRNAs were removed in one step, the first 15 gRNAs based on the target position in the coding sequence from the previous step were kept. It resulted in a library with 37,762 gRNAs, including 37,362 targeting guides covering 4,086 genes and 400 randomized non-targeting guides. Potential off-targets were evaluated using SeqMap (H. Jiang & Wong, 2008)(version 1.0.12).

For randomized non-targeting gRNAs, sequences were firstly randomly generalized with the same length as the target gRNAs in the library, followed by aligning them to the target genome. Sequences with more than 11 bp matches were removed. After filtering with 5 consecutive nucleotides and BsmBI restriction sites, 400 of the remaining sequences were included.

2.2.3 Guide library cloning and verification

The library of guides was assembled by golden-gate BsmBI assembly. After assembly, the library was transformed into *E. coli* to confirm >50-fold library coverage. The library was also transformed by 12 separate electroporation reactions, recovered for 1 hour in 50 mL of LB media, and then supplemented with ampicillin to select for library transformants. This 50 mL culture was grown until the early stationary phase, shaking at 37 °C and 250 rpm. The 50 mL cultured was then maxi-prepped.

2.2.4 Guide library screen

The base editor ScBE3 (SPC914) was transformed into *E. coli* MG1655, followed by electroporation of 1 µg of library DNA to produce the library with the base editor. To screen for gene essentiality, induction of base editing and selection of non-essential genes occurred in tandem. All culturing occurred shaking at 37 °C and 250 rpm, in LB media with antibiotics, 1 mM IPTG, and 0.2% L-arabinose. The dilution and culturing series were as follows: 1:100 with 4h culturing, 1:100 with 4h culturing, and 1:500 with 16h culturing. Cultures were sampled and plasmids were isolated for sequencing.

2.2.5 Library sequencing

Sequencing libraries were prepared by amplifying the guide RNA sequence with primers and adding the Illumina sequence adaptors. Five forward primers were used, with the terminal end staggered by one base pair so as to add to the sequencing library complexity. The resulting amplicons were then further amplified to add Illumina indices for sample identification. Sequencing was performed on an Illumina NextSeq system, 400M 75bp single-ended reads.

2.2.6 Sequencing data processing

Sequence reads with perfect match were mapped to the gRNA library using an in-house Python script. gRNAs were first filtered by 1 count per million in minimal 2 samples in the essentiality screen. Read counts of each gRNA were normalized with non-targeting guides using the trimmed Mean of M method in edgeR (version 3.28.0). RUVs in RUVSeq (version 1.20.0) (Risso et al., 2014) was used to capture batch effects between the two replicate experiments. Differential abundance (log2FC) of gRNAs between time points was calculated using the quasi-likelihood F test after fitting in a generalized linear model.

2.2.7 Applying Machine learning

The machine learning model was developed with 556 sequence features and 2813 gRNAs targeting 307 essential genes using auto-sklearn (version 0.14.6). Sequence features include one-hot encoded 20 nt gRNA and 3 nt PAM, and dinucleotide features of the 30 mer extended sequences from 4 nt upstream of gRNA to 3 nt downstream of PAM (23×4+29×16). An essential gene list in the LB Lennox medium was obtained from EcoCyc (Keseler et al., 2017). The logFC values between the last and initial time points for ScBE3 were used as training targets. For auto-sklearn, the AutoSklearnRegressor function was used and all possible estimators were included. The following parameters were used: “time_left_for_this_task” = 3600, “per_run_time_limit” = 360, “resampling_strategy” = ‘cv’, ensemble_kwargs = {"ensemble_size": 1}, “resampling_strategy_arguments” = {“fold”: 5}, “metric” = autosklearn.metrics.r2, include = {'feature_preprocessor' : ["no_preprocessing"]}. Feature types are all categorical. The selected histogram-based gradient boosting model was saved and used with scikit-learn (version 0.24.2) for downstream analysis. 10-fold cross-validation was used to evaluate model performance. The training and test sets were split guide-wise based on unique gRNA sequences. For each iteration in 10-fold cross-validation, the Spearman correlation between logFC values and predicted values for all test samples was calculated. The histogram-based gradient boosting model was interpreted using the ‘shap_values’ function in TreeSHAP (version 0.39.0) and all samples.

2.3 Methods for Result section 3.2.2

2.3.1 Library design and validation

The reference genome and annotation of *E. coli* K12 MG1655 (NCBI: NC_000913.3) were used for crRNA library design. First, all potential 32 nt guides were designed for protein-coding genes (limited to the coding sequence) and rRNAs with non-GU PFS and GC content between 40% and 60%, resulting in an average of 484 guides per gene. To reduce the size of the library, and considering the unknown effect of the targeting location within a gene on guide efficiency, each gene was divided into a maximum of 10 sections with equal length. Within each section, guides were filtered based on the strength of the local secondary structure, defined as ΔG, in both repeat-crRNA sequence and the mRNA targeting region (including a region of 2 times the length of gRNA before and after the target). ΔG was calculated as the energy difference between the unconstrained minimum free energy (MFE) structure and the constrained MFE structure with no base pairs, estimated using RNAfold from the Vienna RNA Package (version 2.4.12). Sequences (either guide or flanking primer sequences) containing BsmBI restriction sites or homopolymer stretches of more than four consecutive nucleotides were excluded to facilitate synthesis and cloning. The guide with the lowest secondary structure strength in each section was selected, resulting in a library of 25,997 guides, including 25,470 guides targeting protein-coding genes, 127 guides targeting rRNAs, and 400 randomized non-targeting guides. For randomized non-targeting gRNAs, sequences were firstly randomly generalized with the same length as the target gRNAs in the library, followed by aligning them to the target genome. Sequences with more than 12 bp matches were removed. After filtering with 5 consecutive nucleotides and BsmBI restriction sites, 400 of the remaining sequences were included. A sequence containing a universal primer binding site and the BsmBI restriction site was added to the guides to amplify the oligo library and digest it before ligating it into the backbone. The library was synthesized by Twist Bioscience.

The base backbone pFT50 was slightly modified to insert the direct repeat before the GFP dropout site to be able to limit the insertion size to the spacer itself. A BsmBI restriction site present in pFT50 was also eliminated through Q5 mutagenesis.

2.3.2 Guide library cloning and verification

The library was amplified with Kapa Hifi polymerase (Roche Diagnostics, 7958935001) (20 ng DNA) for 10 cycles following the manufacturer's instructions (Ta = 64°C; 30 s denaturation, 20 s annealing, 15 s extension) using primers SPCpr 349/350. 5 µL library (150 nM) and 5 µL of backbone (50 nM) were mixed in 25 µL of total reaction volume. The mixture was subjected to 50 cycles of BsmBI digestion (3 min at 42°C) and ligation (T4 ligase - 5 min at 16°C) with final digestion at 55°C for 60 min to ensure complete removal of the backbone, followed by a 10-minute heat inactivation at 80°C. The sample was then ethanol precipitated, and 5 µg was transformed into fresh electrocompetent Top10 cells (90 µL). The transformation was conducted with two separate batches of electrocompetent cells to ensure enough transformants were obtained. After recovering the two cultures in 500 µL of SOC medium (SOB medium: 20 g Tryptone, 5 g Yeast Extract, 0.5 g NaCl, 800 mL dH2O, and 10 mL 250 mM KCl adjusted to pH 7. To SOB medium added 5ml 2M MgCl2, 20 ml of 1 M Glucose) shaking at 37°C for 1 h, the recovered cultures were back-diluted into 150 mL LB (10 g tryptone, 5 g yeast extract, and 10 g NaCl in 1 L of dH2O) with ampicillin (Amp, 100 µg/mL) and cultured with shaking at 37°C for 12 h. The next day, plasmid DNA from the culture was isolated using the ZymoPURE II Maxiprep Kit (Zymo Research, D4203) and further purified by ethanol precipitation.

2.3.3 Guide library screen

Two replicates of *E. coli* MG1655 cells with or without the nuclease plasmid were inoculated overnight, then the next day the ABS_{600} was normalized and the cells back-diluted to $ABS_{600} \approx 0.1$ in fresh LB with or without chloramphenicol (Cm, 34 µg/mL). Once each culture reached $ABS_{600} \approx 0.8$, cells were made electrocompetent by washing twice with 10% glycerol and finally resuspended in 480 μL 10% glycerol. For each sample, six separate transformations were conducted each with 1 μg of library DNA (40 μL/transformation). Transformed cells were then recovered in 500 μL SOC medium for 1 h with shaking at 37°C. The six reactions were combined to yield 3 mL of culture per condition. Serial dilutions of this culture were made, and 100 μL of 1:10,000 dilutions were plated for the targeting and no-Cas13a samples with the appropriate antibiotics (Cm and Amp, or Amp only), yielding a theoretical library coverage of ~9,500. The remaining culture was diluted 1:100 in LB with Amp and Cm to ABS_{600} = 0.06 and cultured for 12 h with shaking at 37°C. Finally, the library was isolated with the ZymoPure II Plasmid Midiprep Kit (Zymo Research, D4200).

2.3.4 Library sequencing

The guides sequences from the purified library DNA were amplified with Kapa Hifi polymerase using primers oEV-315/316 (NT1), oEV-317/318 (NT2), oEV-319/320 (T1), oEV-321/322 (T2). 10 ng of DNA were included in a 50-μL PCR reaction for 15 amplification cycles (15 s at 98°C, 30 s at 64°C, and 30 s extension). The amplification products were purified using AMPure beads (Beckman Coulter, A63881) and further amplified with primers oEV-323/324 (NT1), oEV-325/326 (NT2), oEV-327/328 (T1), oEV-329/330 (T2) to add the appropriate indices and Illumina adaptors. For this reaction, the same settings were used with the only difference being the amount of input DNA (25 ng) and the number of cycles (10). The resulting amplification products were purified with AMPure beads and resolved on a gel to verify the presence of the correct amplicon. The samples were submitted for Sanger sequencing and Bioanalyzer 2100 analysis (Agilent Technologies) as a quality check. Finally, samples were submitted for next-generation sequencing at the NextSeq 500 sequencer (Illumina) with a 150 bp paired-ends kit (130 million reads, Illumina, 20024905) to obtain 1,000-fold coverage. To increase the library diversity, 20% of phiX phage was spiked-in.

To correlate transcript expression levels with guide depletion, we measured transcript levels in *E. coli* MG1655 under conditions paralleling the library screen. Briefly, two replicates of cells containing the plasmid cBAD33 (empty backbone for nuclease plasmid) were cultured overnight and then normalized to ABS_{600} = 0.06 in LB with Cm and cultured to $ABS_{600} \approx 0.5$ or $ABS_{600} \approx 0.8$. At those growth points, cells were pelleted and snap-frozen for RNA extraction with the Direct-zol RNA MiniPrep Plus kit. The samples were also DNase-treated with TURBO DNase (Thermo Fisher Scientific, AM2239) and quality verified using a 5200 Fragment Analyzer System (Agilent Technologies). Finally, rRNA was removed with the Rybo-off rRNA depletion kit (Vazyme Biotech, N407-01), and the samples were prepared with NEBNext® Ultra™ II Directional RNA Library Prep Kit (New England BioLabs) and sequenced on a NovaSeq 6000 (Illumina) with 50-bp paired-end reads.

The resulting NGS data were deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE179913 for the genome-wide screen (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179913) and GSE179914 for transcriptomic analysis (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179914).

2.3.5 Sequencing data processing

After merging using BBMerge (version 38.69) with parameters "qtrim2=t, ecco, trimq=20, -Xmx1g, mix=f", paired-end sequence reads with a perfect match were assigned to gRNA sequences. After filtering guides for at least 1 count per million reads in at least 2 samples, the library sizes were normalized using the read counts for non-targeting guides with the trimmed mean of M-values method in edgeR (version 3.28.0). Differential abundance (logFC) of gRNAs between targeting samples and control samples lacking the Cas13a nuclease was assessed using the edgeR quasi-likelihood F test after fitting a generalized linear model. The translation initiation rate of each gene was predicted using the RBS calculator (version 1.0) (Salis, 2011). For RNA seq analysis, sequencing reads were aligned to the *E. coli* K12 MG1655 genome (NCBI: NC_000913.3) using STAR (Dobin et al., 2013) (version 2.7.4a) with parameters "–alignIntronMax 1 –genomeSAindexNbases 10 –outSAMtype BAM SortedByCoordinate" and the count of reads mapping to each gene was obtained using HTSeq (Anders et al., 2015) (version 0.9.1) with parameters "-i locus_tag -r pos –stranded reverse –nonunique none -t gene", followed by calculating transcripts per million (TPM).

2.3.6 Applying Machine learning

The machine learning regression model was developed with 144 features and the logFC values of gRNAs from the genome-wide screen as targets using auto-sklearn (version 0.10.0) with all possible estimators and preprocessors included and parameters 'ensemble_size': 1, 'resampling_strategy': 'cv', 'resampling_strategy_arguments': {'folds': 5}, 'per_run_time_limit': 360, and 'time_left_for_this_task': 3600. Features included gene expression level (log2 transformed TPM at ABS_{600} 0.5), gene essentiality, gene id, gene length, (percent) targeting position in gene, delta G of repeat-crRNA and mRNA targeting region, and the one-hot-encoded PFS sequence and gRNA sequence. Gene essentiality information in the LB Lennox medium was obtained from EcoCyc (Keseler et al., 2017). The optimal histogram-based gradient boosting model was evaluated using 10-fold cross-validation and interpreted using test samples and the 'shap_values' function in TreeSHAP (version 0.36.0). In order to further explore the contribution of guide features to depletion, a MERF was applied to remove the effects of gene expression level, gene length, and gene essentiality using a random-effect linear regression model, with guide features used to predict the residual depletion using a fixed-effect random forest model. The contribution of the guide features to predictions of depletion by the fixed-effect random forest was investigated using the 'shap_values' function in TreeSHAP and test samples.

2.4 Methods for Result section 3.2.3

2.4.1 Library design and cloning

The library was designed to target a non-redundant collection of complete genomes of *Klebsiella pneumoniae*. Guides matching at least 15 bp with the *E. coli* MG1655 genome were removed. A guide RNA library containing 11,900 targeting and 100 non-targeting guides was constructed by Golden Gate cloning and transformed into *E. coli* MG1655. At least 10,67 clones were recovered, pooled in 10 mL LB (10 g tryptone, 5 g yeast extract, and 10 g NaCl in 1 L of dH2O), and stored as 1 mL aliquots in DMSO 10% at -80°C. The library was mini-prepped and electroporated into *E. coli* MFDpir, plated in LB with kanamycin (Kan, 50 µg/mL) and incubated at 37°C, then pooled in 10 mL LB and stored as 1 mL aliquots in DMSO 10% at -80°C.

2.4.2 Guide library screen

The plasmid library was delivered by conjugation to two *K. pneumoniae* receptor strains (NTUH-K2044 and KPPR1). Briefly, 108 cells were used to inoculate 50 mL of LB with Kan and diaminopimelic acid (DAP) 0.3 uM, and cultivated at 37°C and 190 rpm until they reached late-exponential phase (OD_{600} ≈1), and 1:1 ratio of donor and receptor were mixed. The cells were plated and collected after 4 hours incubation at 42°C in LB agar with Kan, then pooled in 10 mL LB and stored as 1 mL aliquots in DMSO 10% at -80°C. To perform the assay, 1mL of cells from -80 °C was inoculated in 100 mL LB with Kan and grown to an ABS_{600} of ~ 0.2 at 42°C. Then, 1 nM anhydrotetracycline (aTc) was added and cells were recovered at time zero, after three hours and 16 hours.

2.4.3 Library sequencing

To measure the relative abundance of guide RNAs in each sample, two nested PCR reactions were used to generate the sequencing library. The first set amplifies the sgRNA region, adding a variable region next to the sequencing primer, that prevents the sequencing from starting all at the same position in all the clusters and at the same time can be used as index. The second PCR adds the index and the flow cells attachment sequences. Sequencing is then performed with Illumina sequencing primers using a NextSeq 500 benchtop sequencer.

2.4.4 Sequencing data processing

After filtering guides for at least 1 count per million reads in at least 2 samples from the count table of gRNAs, the library sizes were normalized using the read counts for non-targeting guides with the trimmed mean of M-values method in edgeR (version 3.38.1). Differential abundance (logFC) of gRNAs between different time points (3h - 0h and overnight - 0h) was assessed using the edgeR quasi-likelihood F test after fitting a generalized linear model.

2.4.5 Applying Machine learning

The machine learning regression model was developed with 738 features as predictors and the logFC values between overnight and 0 hours of gRNAs from the genome-wide screens in individual or both strains as targets using auto-sklearn (version 0.14.6) with all possible estimators included, all feature preprocessors excluded and parameters 'ensemble_size': 1, 'resampling_strategy': 'cv', 'resampling_strategy_arguments': {'folds': 10}, 'per_run_time_limit': 360, and 'time_left_for_this_task': 3600. Features included dataset, gRNA GC content, three thermodynamic features (delta G of the hybridization of repeat-gRNA and target DNA, delta G of the homodimer of repeat-gRNA, and delta G of the monomer of repeat-gRNA), the 732 one-hot-encoded single-nucleotide and dinucleotide features. Thermodynamic features were computed using the ViennaRNA Package: RNAduplex (version 2.4.14) for RNA:RNA hybrids; RNAduplex (version 2.1.9h) for DNA:RNA hybrids; RNAfold (version 2.4.14) for single RNA folding. 732 sequence features included one-hot encoded 4th position in PAM and 26 positions of gRNA, and dinucleotide features from 5nt upstream of PAM to 5nt downstream of gRNA. When the data from only one screen were used to train, the dataset feature was subsequently removed. The optimal histogram-based gradient boosting model was evaluated using 10-fold cross-validation based on unique gRNA sequences and interpreted using TreeSHAP (version 0.39.0) and all samples.

3. Result

Section 3.1 is the modified version of the manuscript for "Automated interference of CRISPRi guide efficiency in bacteria from genome-wide essentiality screens" (Yu et al., 2022). The work is a result of collaboration with Sandra Gawlitt (Helmholtz Institute for RNA-based Infection Research), who performed the validation experiments, and Lisa Barros de Andrade e Sousa, Erinc Merdivan, and Marie Piraud (Helmholtz AI), who constructed the deep learning models. Section 3.2 includes the other three side projects, in which the machine learning approach developed in section 3.1 was applied to analyze other independent CRISPR-Cas genome-wide screens.

3.1 Automated interference of CRISPRi guide efficiency in bacteria from genome-wide essentiality screens

3.1.1 Automated machine learning and feature engineering identifies gene-specific effects in CRISPRi depletion screens

I set out to devise design rules for CRISPRi in bacteria by combining machine learning with large experimental datasets. Compared to large-scale tiling screens, which have been routinely adopted for gRNA on-target efficiency prediction in the CRISPR-Cas9 system, genome-wide screens offer higher gene coverage. Higher gene variation in the training will potentially benefit the generalizability of the resulting machine learning model. CRISPRi genome-wide screens allow the investigation of the gene essentiality based on the gRNA depletion log2 fold-changes (logFCs) compared to the initial time point (**Figure 3.1.1A**). Silencing essential genes should result in growth defect or cell death thus strong guide depletion, while the depletion of guides targeting non-essential genes would be modest. But the degree of silencing relies on the efficiency of gRNAs. Hence, I reasoned that understanding the variation in the guide depletion of targeting essential genes, defined by the Keio collection (Baba et al., 2006), in CRISPRi genome-wide screens would hint at the factors attributed to guide efficiency.

I first began my investigation by applying machine learning to predict the guide depletion in a published *E. coli* CRISPRi essentiality screen using dCas9 performed in rich media (Rousset et al., 2018), which included 1,618 guides targeting 171 essential genes. Given the potential and simplicity of automated machine learning (autoML) techniques, which attempt to automate the often labor-intensive process of model selection and optimization, I applied the leading autoML tool auto-sklearn (Feurer et al., 2015) to develop a model to predict the guide depletion. Beyond the exhaustive search of best-performed configurations with model types and their hyperparameters, auto-sklearn turns model selection itself into a Bayesian optimization problem, mapping the performance of each configuration to a probabilistic model. Auto-sklearn further takes advantage of prior knowledge in over a hundred datasets to prioritize the configurations for time-efficient searching.

Features are required to describe the training target (guide depletion) to apply auto-sklearn. Hence, I next asked what features would be required for accurate prediction. In combination with optimizing models using auto-sklearn, I engineered a series of feature sets of increasing complexity (**Table S1; Method 2.1.2**). I started with the one-hot-encoded extended 30 mer sequence including four bases upstream of the gRNA sequence and three bases following the NGG PAM (**Figure S1.1)**, which resulted in a poorly performing model with a median Spearman's ρ of ~0.19 in 10-fold cross-validation splitted based on unique gRNA sequences (**Figure 3.1.1B; Table S2; Table S3; Method 2.1.3**). I therefore iteratively added a set of additional features while monitoring changes in model performance. As targeting efficiency for CRISPRi has been suggested to depend on the distance to the transcriptional start site (Qi et al., 2013; Doench et al., 2014; Gilbert et al., 2014; Radzisheuskaya et al., 2016; T. Wang et al., 2018), the set included absolute and relative distance to the start codon. I also included a suite of thermodynamic features describing gRNA:target interactions predicted using the ViennaRNA package (Lorenz et al., 2011): minimum free energy of the gRNA self-folding, hybridization of two identical gRNAs, and hybridization of the targeted DNA and gRNA (Lorenz et al., 2012). These additional feature sets resulted in only moderate improvement in Spearman correlation (ρ ~ 0.22) for the predictions (**Figure 3.1.1B)**.

Given that guide depletion mainly reports the fitness defect of the target genes, I constructed eight gene-specific features describing each target gene that I reasoned may have some explanatory power (**Table S1; Method 2.1.2**). I collected publicly available RNA expression data over growth in minimal media stretching from OD_{600} 0.1 to 180 minutes after the stationary phase (Conway et al., 2014). The Rousset dataset was collected at OD_{600} 2.2-2.5, so I include data from all time points to compute the maximal and minimal gene expression level. I collected transcription unit (TU) information from RegulonDB (Tierrafría et al., 2022) and calculated the distance from the target site to the start of the TU, the number of downstream genes in each TU, and the presence of downstream essential genes in the TU. Finally, I also included gene GC content and gene length. Incorporating these additional gene features led to a major improvement in prediction accuracy, with cross-validation performance jumping to a Spearman's ρ of ~0.67 (**Figure 3.1.1B**).

To understand the contribution of these features to the prediction of gRNA depletion, I used SHapley Additive exPlanation values (SHAP values) computed with TreeExplainer (Lundberg et al., 2020) on the best-performing histogram-based gradient boosting model produced by auto-sklearn (**Figure 3.1.1C; Table S4**). SHAP values are a game-theoretic approach to feature importance that capture the marginal contribution of a given feature to a prediction. Looking at average absolute SHAP values provides a measure of feature importance while plotting individual SHAP values shows how each feature affects each individual prediction. Of all considered features, maximal RNA expression had the single largest average effect on depletion, making an average of a ~1.6-fold difference to the predictions. Unexpectedly, high target gene expression tended to be associated with higher depletion. There was also evidence for polar effects from CRISPRi, as the number of downstream essential genes was highly predictive of increased depletion. The most predictive effects that could actually be modified by guide design were guanine at the 20th position of gRNAs and cytosine at the N position of PAM, followed by guide distance to the transcriptional start site, but on average these had fairly small effects compared to features associated with the target gene. In summary, I found that autoML can rapidly produce predictive models of CRISPRi guide depletion, and the predictions made by these models are dominated by the effects of gene features that cannot be modified in guide design. These findings outline key features that need to be accounted for to accurately infer guide efficiency from genome-wide depletion screens.

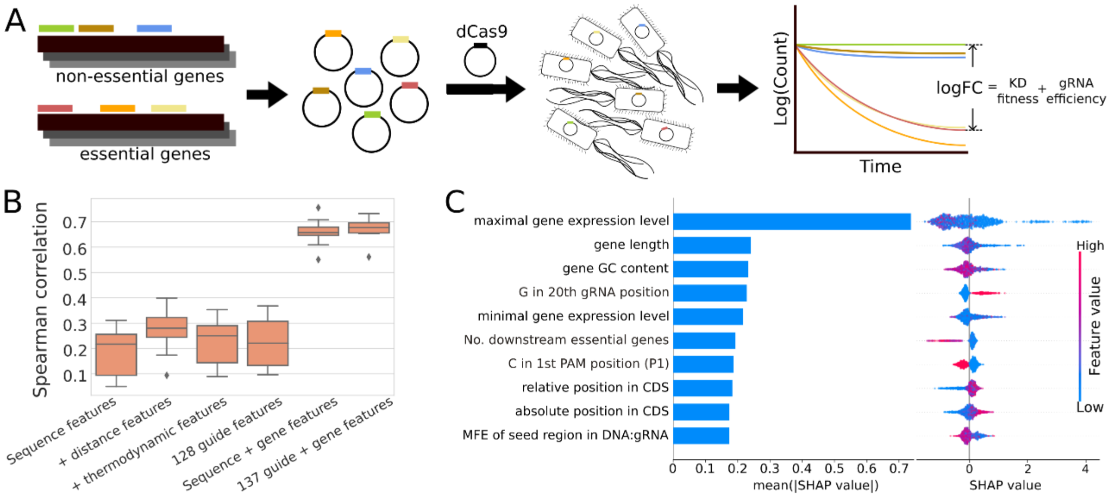

Figure 3.1.1: Automated machine learning predicts depletion in CRISPRi essentiality screens. (**A**) An overview of CRISPRi essentiality screens. gRNAs are designed targeting every gene in the genome and cloned into an appropriate plasmid for expression. This plasmid collection is then transformed into the target bacteria, and depletion is measured as the change in guide frequency over growth determined by sequencing relative to a set of non-targeting gRNAs. The measured depletion (logFC) is then a mixture of the fitness effect of gene knockdown with the efficiency of silencing itself. (**B**) Comparison of Spearman correlation between actual and predicted guide depletion in 10-fold cross-validation (CV) of the best model trained with auto-sklearn with different feature combinations, using data from (Rousset et al., 2018). (**C**) The ten most predictive features determined using TreeExplainer on the optimal histogram-based gradient boosting tree model trained with auto-sklearn and 137 guide and gene features. Mean absolute SHAP value (left) provides a global measure of feature importance, while the beeswarm plot (right) shows the effect of each feature on each individual gRNA prediction. CDS: coding sequence.

3.1.2 Data fusion improves prediction performance

Considering the decent performance with the set of 137 features, I next asked whether the sample was limiting the accuracy of the predictions. To this end, I collected data from two additional CRISPRi screens of *E. coli* in rich media. First, I included data from an additional screen using the same gRNA library and experimental setup but with dCas9 expressed from a stronger promoter, which I refer to here as E18 Cui (Cui et al., 2018). Second, I included data from a completely independent screen using a higher density library containing twice as many guides targeting essential genes (4164; 378 are identical to gRNAs contained in Cui/Rousset), which I refer to as Wang (T. Wang et al., 2018). I refer to the original data set as E75 Rousset. It is also worth noting that while the E18 Cui and E75 Rousset libraries were grown repeatedly to the stationary phase, the Wang screen was collected in the log phase. The level of depletion in each dataset exhibited qualitative differences, with Wang showing a clearer bimodal separation between depleted and non-depleted guides (**Figure 3.1.2A**). There was a reasonable correlation of depletion between datasets, with E18 Cui and E75 Rousset exhibiting a Spearman's ρ of ~0.9. The correlation between Wang and the other two datasets was lower (ρ~0.80-0.82), but this seemed mostly attributable to a saturation effect in Wang, possibly due to the shorter growth period (**Figure S1.2**).

To investigate the impact of fusing these datasets on model performance, I trained a series of models using auto-sklearn with each dataset individually or in combination including a dataset indicator as a potential predictor. For comparable and generalizable evaluation, I tested them on sets of guides held out from each dataset as well as a mixed test set (**Figure 3.1.2B**; **Table S5**) and applied 10-fold cross-validation splitted based on unique gRNA sequences (**Table S2; Method 2.1.3**). Unsurprisingly, models trained on single datasets tended to perform best on their cognate test set. Similarly, models trained on E18 Cui and E75 Rousset appeared to generalize better to each other than to the Wang dataset and vice versa. Combined three training datasets exhibited generalized better across datasets without degrading performance relative to models trained on individual datasets. Training on only two datasets followed the same trend, particularly those mixing at least one of the Cui/Rousset sets with Wang (**Figure S1.3A**). In some cases, particularly with the Cui dataset, fused training sets actually improved performance on a test set drawn from a single dataset. In each case, the best-performing model chosen by auto-sklearn was constantly gradient-boosted decision tree models.

To illustrate that the performance increases I saw when combining datasets were not an artifact of my autoML workflow, I tested data fusion with both an alternative autoML package, H2O (Ledell & Poirier, 2020) (**Figure S1.3A; Table S5**) as well as a suite of individual model types (**Figure S1.3; Table S6**). H2O independently selected the histogram-based gradient boosting model as the best-performed model. Other tested model types responded differently to the fused data, with linear regression-based models showing little improvement (e.g. linear regression, LASSO linear regression, elastic net linear regression; **Figure S1.3B-D**), while tree-based methods (e.g. random forest regression, histogram-based gradient boosted trees; **Figure S1.3F-G**) showed clear improvement. Importantly, none of the tested models appeared to degrade in performance when trained with fused data. These findings suggest that both increased generalizability and accuracy can be achieved by integrating multiple data sources for training tree-based models for CRISPRi depletion.

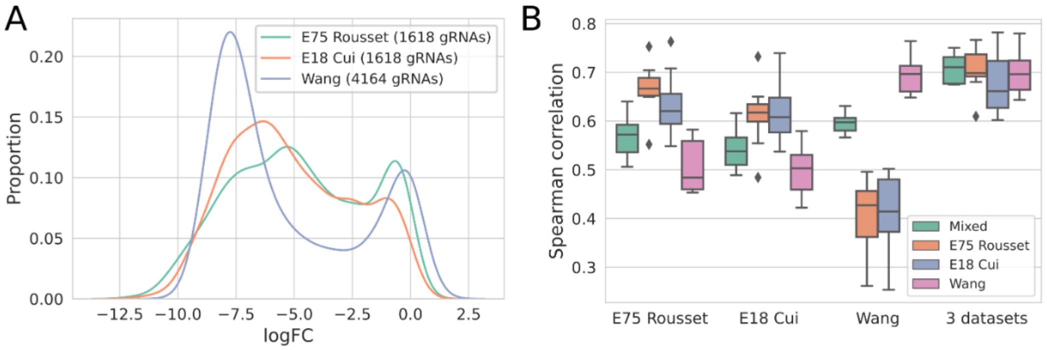

Figure 3.1.2: Data fusion improves prediction of depletion in genome-wide CRISPRi screens. (**A**) Distribution of logFCs of gRNAs targeting essential genes from three CRISPRi genome-wide essentiality screens in *E. coli.* (**B**) Comparison of Spearman correlation from 10-fold CV of the best auto-sklearn trained model on one dataset or the integrated three datasets.

3.1.3 Segregating guide and gene effects produces a predictive model for CRISPRi guide efficiency

My exploration of the features that are most predictive of gRNA depletion in competitive screens highlighted that features describing the targeted gene often made much larger contributions to the prediction than features describing the guide sequence, strengthening that the guide depletion describes mainly the gene-to-gene variation. However, this dominant gene-to-gene variation masks the guide-to-guide variation, thereby obscuring the underlying factors for guide efficiency. I reasoned that removing gene-specific effects would allow us to extract the contribution of guide efficiency properly.

I took two distinct approaches to separate guide and gene effects. The first was to explicitly model both effects jointly using Mixed-Effect Random Forest (MERF) (Hajjem et al., 2014). The MERF model handles data with an underlying cluster structure by defining two separate models: a linear model that captures random effects associated with the cluster, and a random forest (or other tree-based models) that captures fixed effects associated with each individual measurement. These models are then jointly optimized in an iterative process using the expectation-maximization algorithm (**Figure 3.1.3**). In my case, random effects correspond to features associated with each gene (e.g. gene GC content, expression level) as well as dataset, while fixed effects correspond to features that could be manipulated in gRNA design (e.g. PAM and guide sequence, thermodynamic properties). I used the gene identifier as a cluster ID for the random-effect model.

I refer to the second approach as median subtracting (MS), where I subtract the gene-wise median logFC from each gRNA depletion value to calculate relative "activity scores" following previous work (Calvo-Villamañán et al., 2020) (**Figure S1.4A**). The published LASSO model from the same work was referred to as "Pasteur". However, this leads to problems integrating multiple datasets, as the range of depletion values varies across datasets (**Figure 3.1.2A**). I adapted a previously described approach used for fusing CRISPR gene deletion datasets (Xiang et al., 2021). First, I averaged the logFCs between E75 Rousset and E18 Cui which share all guides in common. I then calculated a linear scale factor for guides shared between Wang and the averaged Rousset/Cui data set to make logFCs for the unshared guides in Wang comparable to logFCs derived from Rousset/Cui (**Figure S1.4B-D**). For cross-validation, scaling was performed within each test fold to avoid possible leakage of information between test and training sets. The scaled activity scores from integrated three datasets are referred to as MS data.

Both the fixed effect model from the MERF and activity scores in the MS method remove gene-specific effects to estimate guide efficiency, making guide-wise cross-validation difficult as the true guide efficiency is unknown. As an alternative to guide-wise cross-validation, I implemented a gene-wise cross-validation scheme. I trained new models using 10-fold cross-validation (**Method 2.1.3**), this time holding out all guides targeting a set of held-out genes, evaluating the Spearman correlation between predictions and measured depletion within each gene under the assumption that rank order should reflect guide efficiency within a gene.

To assess the MERF model, I first optimized the hyperparameters for the fixed-effect random forest model using hyperopt, which outperformed the random forest and histogram-based gradient boosting models optimized to predict depletion using auto-sklearn (**Table S7**), thereby adopted for the following evaluations. To evaluate the MS method, I began with the Pasteur model, which is to date the only model for guide efficiency prediction for CRISPRi in bacteria. Considering the Pasteur model was trained on the E75 Rousset dataset alone, I trained it on the MS data, which is hereafter referred to as "Pasteur (retrained)". Besides, the Pasteur model incorporated a reduced sequence feature set describing the 6 nt upstream of the di-guanine in the PAM to 16 nt downstream of PAM. Therefore, the discrepancy between MERF and Pasteur (retrained) would indicate the contribution of both the extended feature set and the gene-effect segregation methods, while the comparison between models trained on individual or integrated datasets allows the exploration of the effect of data fusion.

As I had previously observed in my evaluation of depletion predictions (**Figure 3.1.2B**), data fusion between multiple CRISPRi screens consistently improved performance in both MERF and Pasteur models (**Table 1**). In aggregate, the MERF model performed better than the LASSO-based models (median ρ=0.393 (MERF) vs. 0.366 (Pasteur (retrained)) and 0.357 (Pasteur)). When I broke this down into performance on held-out genes in individual datasets, Pasteur performed the best on the E75 Rousset data (ρ=0.429 vs. 0.406 (MERF) and 0.394 (Pasteur (retrained))), probably due to the overlapping in my test sets and its training data. But MERF performed the best in the E18 Cui data from the same lab (ρ=0.429 vs. 0.400 (Pasteur (retrained)) and 0.411 (Pasteur)). The improvement in the Wang dataset was more distinguishable (ρ=0.371 (MERF) vs. 0.327 (Pasteur (retrained)) and 0.298 (Pasteur)).

Even though MERF outperformed the retrained Pasteur model, it is not clear whether it is due to the feature set or the gene-effect segregation method. Therefore, I optimized a LASSO model using hyperopt with the extended feature set on MS data. The performances were slightly better than Pasteur (retrained) models (ρ=0.371) and the trend in individual datasets was similar (**Table S7**), indicating my extended feature set could provide better estimates for guide efficiency. When I trained a random forest model on MS data with the same feature set and hyperparameters as MERF, the performance however degraded (ρ=0.372), despite the high correlation observed between median gene-wise logFC and the MERF-predicted random effects across the datasets (**Figure S1.4E**), suggesting MERF could better segregate the gene effects than MS method.

In sum, I found that MERF trained on multiple datasets and extended feature set outperforms MS models in predicting guide efficiency for held-out genes and that the MERF approach provides a straight-forward means of integrating datasets while isolating effects important for guide efficiency.

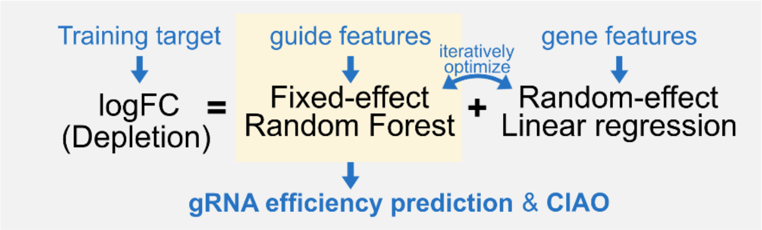
Training target
guide features
iteratively optimize
gene features
logFC (Depletion)
=
Fixed-effect Random Forest
+
Random-effect Linear regression
gRNA efficiency prediction & CIAO

Figure 3.1.3: Segregation of gene effects using Mixed-Effect Random Forest (MERF). An overview of the training process of the MERF model. A MERF model segregates the training target into predictions from a fixed-effect random forest model and a random-effect linear regression model, where fixed effects are associated with each individual measurement and random effects are associated with the cluster. These two models are then jointly optimized in an iterative process using the expectation-maximization algorithm. In my case, to predict the measured depletion (logFC), 9 gene-specific features are enclosed for the random-effect model, while 128 guide-specific features that could be manipulated in gRNA design (e.g. PAM and guide sequence, thermodynamic properties) are assigned to train the fixed-effect models. The trained fixed-effect random forest model was then used for gRNA efficiency prediction and the web-based tool CIAO (ciao.helmholtz-hiri.de).

Table 1: Evaluating predictions of guide efficiency after removing gene effects. Spearman correlations between predictions and measured logFC for held-out genes. Genes were held out in 10-fold cross-validation, and the reported median Spearman correlation was calculated across all held-out genes.

Model	Training data	Median Spearman Correlation Across held-out genes			
		E75 Rousset	E18 Cui	Wang	Mixed
MERF	E75 Rousset	0.331	0.333	0.257	0.301
	E18 Cui	0.371	0.400	0.283	0.329
	Wang	0.357	0.400	0.336	0.357
	3 datasets	**0.406**	**0.429**	**0.371**	**0.393**
Pasteur (retrained)	E75 Rousset	0.333	0.333	0.256	0.310
	E18 Cui	0.363	0.352	0.286	0.322
	Wang	0.367	0.377	0.307	0.339
	3 datasets	0.394	0.400	0.327	0.366
Pasteur	-	0.429	0.411	0.298	0.357

3.1.4 Model interpretation with explainable AI illustrates rational design rules for CRISPRi

To understand the features underlying model performance, I again examined SHAP values for my random forest models using TreeExplainer (Lundberg et al., 2020). I observed similar features with large impacts on predictions from both MERF and MS random forest models (**Figure 3.1.4A; Figure S1.4F; Table S8**). In the MERF model, the strongest average effects were seen for distances from the start codon, followed by a cytosine at the +1 position following the PAM. In particular, I found that targeting positions further from the start codon led to reduced guide efficiency, as has been inferred previously (Qi et al., 2013; T. Wang et al., 2018). Other predictive features involved the nucleotide at position 20 of the guide, directly adjacent to the PAM sequence (**Figure 3.1.4A-B**). Guanine and particularly adenine at this position negatively impacted silencing efficiency, while cytosine and thymine increased efficiency — almost the exact inverse of previous reports for Cas9 efficiency in eukaryotic genome editing applications (Doench et al., 2014; Michlits et al., 2020). Within and following the PAM sequence, the SHAP values were qualitatively similar to previous observations in Cas9 genome editing (Doench et al., 2014; Moreno-Mateos et al., 2015; H. Xu et al., 2015; Corsi et al., 2022). Cytosine was again favored at the variable position of the NGG PAM, and a guanine residue immediately following the PAM had a negative impact on the silencing, though I additionally observed a positive impact of cytosine at this position. Together, these effects displayed a preference for cytosine and a disfavoring of guanine and thymine within and around the PAM sequence.

To investigate potential interactions between features, I estimated SHAP interaction values that quantify situations in which the presence of one feature changes the impact of another so that the combined SHAP value for both features together is not the simple sum of each feature's SHAP value. To provide a visualization of these interactions, I calculated expected effects using the median SHAP value for each feature from guides containing only one of the interacting features and compared the expected sum to the actual SHAP values for guides containing both features (**Figure S1.5; Table S9**).

The majority of these interactions involved distance features or bases in the vicinity of the PAM. For instance, the interactions between the favored cytosine and disfavored guanine lead to guides with either increased (**Figure S1.5 I**) or reduced efficiency (**Figure S1.5 III**) compared to expectations based on single feature SHAP values, whereas multiple cytosines bring an enhanced efficiency (**Figure S1.5 II**). Similarly, consecutive thymines can lead to a stronger reduction in efficiency (**Figure S1.5 IV**). The existence of such interactions between features in the guide sequence may provide one explanation for the superior performance of tree-based methods over linear regression, as tree regressors are particularly well suited to capture interaction effects.

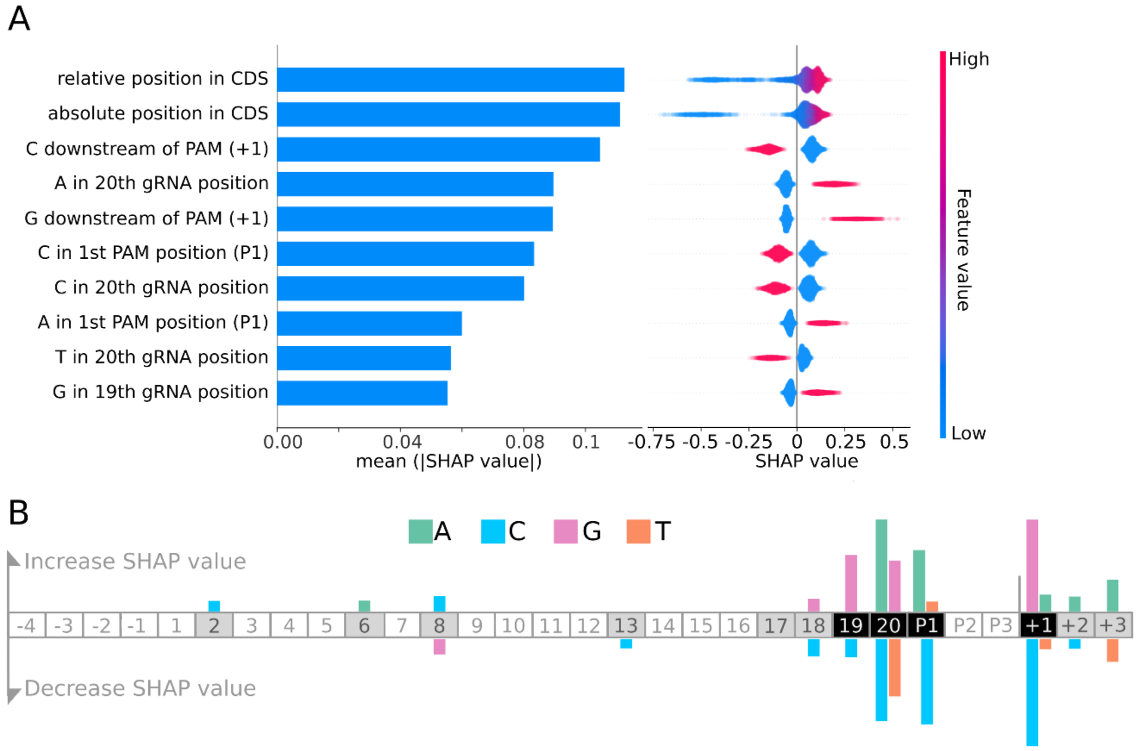

Figure 3.1.4: Important features for CRISPRi guide efficiency illustrate sequence preferences. (**A**) SHAP values for the top 10 features from MERF optimized random forest model. Global feature importance is given by the mean absolute SHAP value (left), while the beeswarm plot (right) illustrates feature importance for each guide prediction. (**B**) A summary of the effects of sequence features. Increased SHAP values indicate features that lead to reduced guide efficacy, while decreased SHAP values indicate increased guide efficacy. The guide sequence is numbered G1 to G20 and the three positions of the PAM sequence are labeled P1, P2, and P3. Negative and positive numbers refer to positions preceding the guide sequence and following the PAM, respectively.

3.1.5 Deep learning approaches do not improve prediction performance

Given the increasing popularity of deep learning approaches in CRISPR guide efficiency prediction (Chuai et al., 2018; H. K. Kim et al., 2018, 2019; D. Wang et al., 2019; Xiang et al., 2021), I next asked whether the combination of MS method and deep learning model would improve performance in predicting gRNA efficiency for CRISPRi in bacteria. I tested two convolutional neural networks (CNNs), which run a series of kernel filters across the sequence to extract local features. One is a custom CNN, where the convolutional layers were used to extract sequence features before concatenating them to the eight non-sequential guide features (**Figure S1.6**). This concatenated feature set was then fed through a fully connected 4-layer multilayer perceptron (MLP) for regression using activity scores for guide efficiency. In addition to a custom CNN architecture, I reimplemented and tested the state-of-the-art deep learning architecture used for predicting Cas9 gene editing efficiency by CRISPRon (Xiang et al., 2021) for Cas9 genome editing in eukaryotes, only trained using the CRISPRi MS data and guide feature set.

Both the custom CNN and adapted CRISPRon models exhibited lower Spearman correlations as compared to my previously trained random forest models when tested on held-out gene sets (**Table S7**; CNN ρ=0.326, CRISPRon ρ=0.333, vs. MERF ρ=0.396). These results show that conventional machine learning approaches can outperform deep learning models, which might be caused by the limitation of data size in applying more complex models.

3.1.6 A saturating screen of purine biosynthesis genes independently validates performance of tree-based models and data fusion

The cross-validation within the training datasets demonstrated that MERF trained on multiple datasets outperformed MS methods in predicting guide efficiency. To further validate the model performance by producing a truly independent test set, a plasmid-encoded GFP construct was targeted with 19 gRNAs across a range of predicted guide efficiencies, and the reduction in cell fluorescence by flow cytometry was approximated as the measured guide efficiency (**Figure S1.7A; Table S12**). Using Spearman correlation to measure the ranked order accuracy, I found MERF performed best (ρ=0.64), while the Pasteur models performed comparatively poorly (ρ=0.33 retrained, 0.26 original). Replicating this study in *Salmonella* Typhimurium gave qualitatively similar results, though with lower Spearman correlations (**Figure S1.7B; Table S12**). However, when I reanalyzed the data from a Miller assay (measuring β-Galactosidase activity) previously used to validate the Pasteur model (Calvo-Villamañán et al., 2020) (**Figure S1.7C; Table S12**), I found that the Pasteur models performed modestly better (ρ=0.71 original, 0.65 retrained) than the MERF (ρ=0.63). I also tested three tools designed for predicting Cas9 guide efficiency for genome editing in eukaryotes (H. Xu et al., 2015; Doench et al., 2016; Wilson et al., 2018; H. K. Kim et al., 2019), and all performed universally poorly on both data sets.

While the exact reasons for the discrepancies in performance between the GFP measurements and the Miller assay are unclear, one plausible explanation is that these data sets simply have sample sizes too small to discriminate between prediction methods. To resolve this, I performed a high-throughput screen targeting nine genes from the purine biosynthesis pathway of *E. coli* known to be essential in minimal media, spread across seven independent transcriptional units (**Figure 3.1.5A**). To avoid any bias in guide selection, I saturated all potential target sites in each gene, ending with a total of 750 gRNAs, including between 35 and 223 guides per gene. Duplicate samples were then collected at three time points during growth in M9 minimal medium, and gRNA depletion was measured with reference to input samples, normalized using a set of 50 non-targeting guides not complementary to any potential target sequences in *E. coli*.

Comparing the experimentally determined depletion values to predictions from my tested models confirmed the results of my previous cross-validation in the training set (**Figure 3.1.5B; Table S14**): the MERF performed best (median ρ~0.57 across all time points vs. 0.53 Pasteur (retrained) and 0.45 Pasteur). While Spearman correlation overviews the ranked orders of the gRNA efficiency, commonly only three to five guides are chosen for each target gene in practice. Therefore, I calculated positive predictive values (PPV), which indicate the number of true positives relative to the sum of both true and false positives. To classify the measured logFC into positive and negative classes, I defined gRNAs

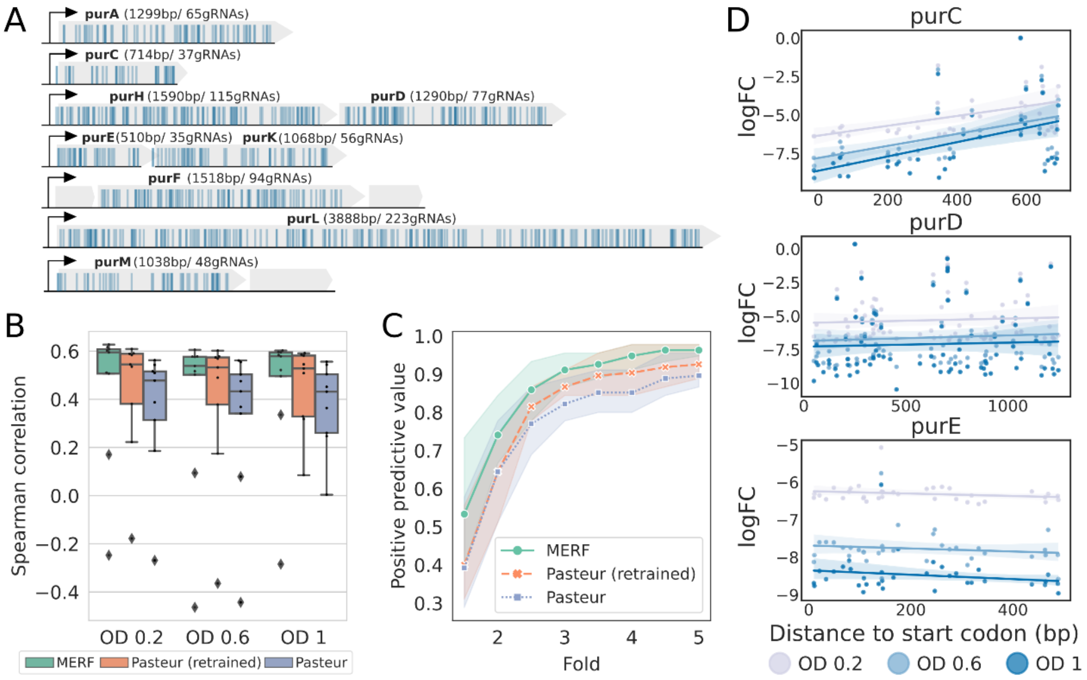

Figure 3.1.5: Independent validation of model performance using a saturating screen of purine biosynthesis genes. High-throughput screening of 750 gRNAs targeting 9 purine biosynthesis genes in *E. coli* K12 MG1655. (**A**) Transcriptional architecture of the targeted genes. All possible gRNAs were designed for each gene. Each blue vertical line represents a gRNA. Grey boxes represent genes, and black arrows transcriptional start sites. (**B**) Spearman correlations between the predicted scores and measured logFC for each gene across collected time points. (**C**) Positive predictive values of all gRNAs for each time point. The predicted positives are defined as the top 5 predicted gRNAs in each gene, while the positive class includes gRNAs with fold changes values within N fold of the fold change value of the strongest depleted gRNA in each gene (N= 1.5 - 5 with a step of 0.5). (**D**) Measured logFCs for each guide as a function of distance to the start codon for purC, purD, and purE. The plots for the other 6 genes are in **Figure S1.9**.

within the N-fold depletion fold change of the most strongly depleted gRNA for each gene as the positive class and tested a set of thresholds with N ranging from 1.5 to 5. I observed that MERF performed constantly the best when the top three to five guides are defined as the predicted positive (**Figure 3.1.5C; Figure S1.8A-B**), supporting the utility of MERF as a predictive tool for gRNA selection in practice.

Comparing the MERF trained on the three fused data sets to the MERF trained on any single data set also showed improved performance, similar to the retrained and original Pasteur models (**Figure S1.8C; Table S14**). The choice of a tree-based model also made a clear difference in performance, as the LASSO and deep learning models trained on the MS data showed worse performance (ρ~0.48—0.50) than either random forest (ρ~0.53) (**Figure S1.8D; Table S14**). The performance with PPV follows the same trend.

Beyond validating the performance of my models, my saturating screen of purine biosynthesis genes also revealed previously unobserved features of CRISPRi depletion screens. First, there were two genes, purE and purK, on which all methods performed poorly as measured by Spearman correlation. Upon inspection of the depletion values, it became clear that this was because there was surprisingly little variation in guide efficiency along these transcripts (**Figure 3.1.5D; Figure S1.9A-B**). This meant that for these genes, differences in ranking reflected very small differences in depletion, likely within the error of the experimental measurements. I examined my initial training set to see if this might be a more widespread phenomenon, finding a substantial number of genes with low variation in their guide depletion values (**Figure S1.9C**). This may be a factor in the overall low average Spearman correlations I report in the cross-validation.

A second unexpected feature was the overall lack of a clear relationship between guide efficiency and distance to the transcriptional start site. Of the nine tested genes, only two, purC and purM, showed a clear linear dependency of depletion on position within the gene sequence. This was particularly surprising, as distance features were clearly important to the model predictions. I attempted to train a model excluding distance features, but this substantially degraded performance on predicting depletion in the high-throughput screen (**Figure S1.8E**). Whether this is an artifact of the training data, based on screens that used small collections of guides biased towards the 5′ ends of genes, or if other guide features compensate for positional differences in guide efficiency remains unclear. In support of the latter, my analysis of feature interactions found many of the strongest effects came from interactions between distance features and sequence features in the vicinity of the PAM (**Table S9**), suggesting that sequence features have larger effects on efficacy as the distance from the transcription start site increases.

In sum, the screen of guides targeting purine biosynthesis genes independently validated the better performance of my MERF model compared to the state-of-the-art, while also highlighting some unexpected features of CRISPRi.

3.2 Analysis of CRISPR-Cas genome-wide screens

3.2.1 Analysis of CRISPR base editor genome-wide screen in *Escherichia coli*

*The section is part of the results for "**Enhanced genome-wide knockout screens in bacteria with CRISPR base editors**". The work is a result of collaboration with Sandra Gawlitt (Helmholtz Institute for RNA-based Infection Research) and Scott P Collins (North Carolina State University), who performed all laboratory experiments and finalized the figures.*

CRISPR base editors enable persistent point mutation in the genome, not only minimizing the errors from double-strand break repairing but also showing promising future treatments for point mutation diseases. Whereas the first-generation cytosine base editors (CBEs) were fusions of dCas9 and a cytidine deaminase, the third generation replaced the dCas9 with Cas9 nickase to achieve higher editing efficiency (Komor et al., 2016). CBEs alter cytidine to thymine, thus consequently altering guanine to adenine on the opposite strand. By specifically editing CAA, CAG, CGA, and TGG codons, base editors can introduce premature stop codons thus becoming an alternative gene silencing technique to Cas9 or CRISPRi (**Figure 3.2.1A**) (Billon et al., 2017; Kuscu et al., 2017). Despite its proven applicability in bacteria (Banno et al., 2018), insufficient attention was paid to optimizing the base editing in bacterial systems and assessing its capability for genome-wide screenings. Further, it remains under-investigated whether a gene can be disrupted by editing the start codon (ATG). To fill the gap, Sandra and Scott engineered a *Streptococcus canis* Cas9 (ScCas9) nickase-derived base editor (ScBE3) that disrupts transcription efficiently in *E. coli*. ScBE3 also exhibits flexible PAM recognition (NNG), increasing the number of targetable genes.

To assess the gene silencing efficiency of ScBE3 by introducing premature stop codon on a genome-wide scale, a library of 37,362 gRNAs was designed to target ATG (start codon), CAA, CAG, CGA, or TGG codons for 4086 protein-coding genes in *E. coli* (**Method 2.2.2**). gRNAs were selected to target the first half of the coding sequences to ensure silencing efficacy. The number of gRNAs per gene was balanced by removing the excessive guides with potential off-target effects and extreme GC content. The library further included 400 randomized non-targeting guides lacking complementarity to the target genome. The sgRNA library was then transformed into *E. coli* cells together with ScBE3-carrying plasmid. After culturing for 16 hours, the transformed cells were induced to express ScBE3 to silence the target genes. Short-read sequencing was finally applied to sgRNA-carrying plasmids collected at 0, 4, 8, and 24 hours after induction (**Figure 3.2.1B**). Under this setup, guides targeting essential guides

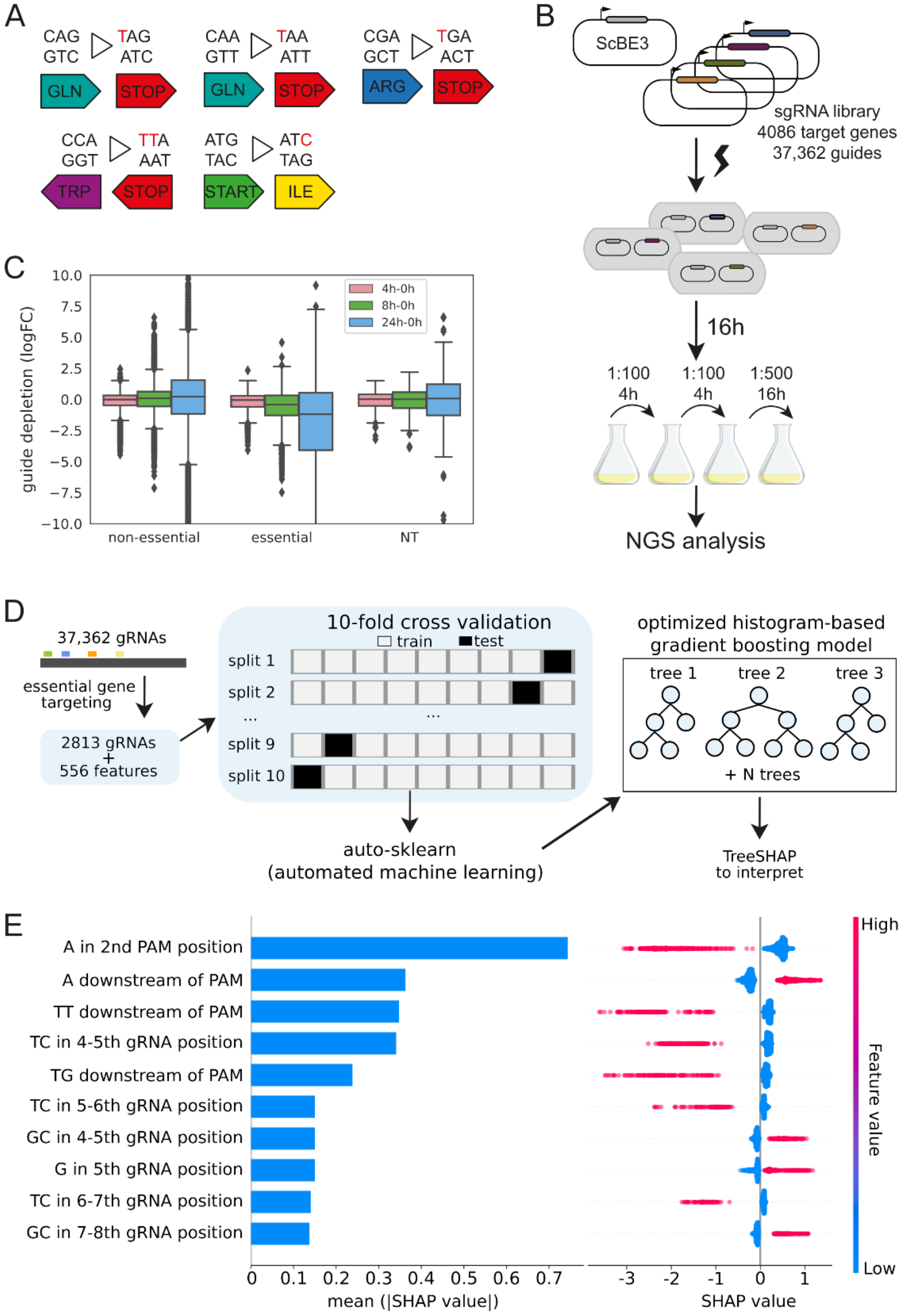

Figure 3.2.1: A genome-wide base-editor screen of essential genes in *E. coli* reveals hidden target preferences of ScBE3. (**A**) Nucleobase context of ScBE3 targets cytosines that can be converted into premature stop codons or lead to a mutated start codon. (**B**) Overview of the genome-wide gene essentiality screen in *E. coli* using ScBE3. Plasmids carrying sgRNA and ScBE3 were co-transformed in *E. coli*. Samples were collected at 0, 4, 8, and 24 h after induction of ScBE3 expression, followed by next-generation sequencing (NGS) of the extracted sgRNA-carrying plasmids. Guide depletion is determined in comparison to the initial time point (0 h). (**C**) log2 fold-changes (logFC) of sgRNAs targeting essential, non-essential genes or a non-targeting control for ScBE3 between time points. Boxes and whiskers indicate 50 and 99% intervals, respectively. (**D**) Depiction of method used for developing the machine learning algorithm. Screen data between 0 and 24 h after induction were used to determine the depletion of each of the 37,362 guides. The derived log2 fold-changes of 2,813 essential-gene-targeting guides were used as prediction targets, with 556 sequence features as predictors. Auto-sklearn was used to train and optimize an assortment of candidate model types. A histogram-based gradient boosting model was the best-performing model. This model was then interpreted using TreeSHAP to explain the contribution of each predictor to each prediction. (**E**) SHAP values for the top 10 features from optimized histogram-based gradient boosting model. Global feature importance is given by the mean absolute SHAP value (left), while the beeswarm plot (right) illustrates feature importance for each guide prediction.

would be more strongly depleted compared to guides targeting non-essential genes and the depletion would be stronger over time. Indeed, the depletion of guides targeting essential genes, defined by Keio collection (Baba et al., 2006), was enhanced as the induction time was prolonged, with median depletion of 0.04, 0.42, and 1.19-fold comparing 4, 8, and 24 hours after induction compared to the initial time point, while the non-essential gene targeting and non-targeting guides were enriched by median 0.22-fold and 0.07-fold at the 24 hours after induction (**Figure 3.2.1C**).

Next, I applied auto-sklearn to optimize machine learning models to extract design rules for effective silencing (**Figure 3.2.1D**). Similar to the model for the CRISPRi, only guides targeting essential genes were subjected to the training. Interpretation of the resulting model using SHAP values (Lundberg et al., 2020) highlighted the predictive sequence features near the PAM region and activity window, including A at the 2nd position of PAM, A/TT/TG downstream of PAM, and TC at the 4th to 5th position from the 5' end of gRNA (**Figure 3.2.1E**).

Here, even though the depletion of guides targeting essential genes was modest compared to the previous CRISPRi screens (**Figure 3.1.2**), likely due to low editing in the experimental setup, the machine learning model was able to extract important features for ScBE3. The preferences in sequence context revealed by the machine learning model can guide us towards maximization of the efficiency of the CRISPR base editor in bacteria, thereby facilitating its application in a broader range of functional interrogations and bacterial species.

3.2.2 Analysis of CRISPR-Cas13a genome-wide screen in *Escherichia coli*

*This section is part of the results for the publication "**A target expression threshold dictates invader defense and autoimmunity by CRISPR-Cas13**" (Vialetto et al., 2022). The work is a result of collaboration with Elena Vialetto (Helmholtz Institute for RNA-based Infection Research), who performed all laboratory experiments and finalized the figures.*

CRISPR-Cas13a is a class 2 type VI RNA cleaving single-effector system. Upon target recognition it cleaves non-specifically cellular RNA, inducing cellular dormancy or cell death to inhibit reproduction and dissemination of the invader as an immune response (Abudayyeh et al., 2016; Meeske et al., 2019). Instead of PAM, the invader-identification of Cas13a relies on the protospacer flanking site (PFS). The G PFS base-pairing with the repeat-derived portion of the crRNA aborts the Cas13a activation (Meeske & Marraffini, 2018). The known mechanism of Cas13-induced immunity, including the necessity of target RNA and crRNA pairing and non-G PFS, was principally assessed through targeting transcripts expressed from plasmids or phages or from the genome with an inducible promoter (Abudayyeh et al., 2016, 2017; Meeske & Marraffini, 2018; Kiga et al., 2020; Meeske et al., 2020). The impact of targeting native transcripts by Cas13a was however under-investigated despite the fact that CRISPR-Cas systems infrequently require spacers from the endogenous sequences (Stern et al., 2010). Even though the Cas13-mediated immune response is known to induce cell death or dormancy, it has been largely associated with gene silencing instead in eukaryotes (Abudayyeh et al., 2017; Cox et al., 2017). To gain a deeper understanding of the mechanism of Cas13-induced immunity, Elena set out to investigate the impact of targeting endogenous transcripts by employing the *Leptotrichia shahii* Cas13a (LshCas13a) (Abudayyeh et al., 2016; L. Liu et al., 2017) in *E. coli* as a simple model and using the reduction in plasmid transformation as the readout of the Cas13a activation, with a high reduction indicating successful of Cas13a activation. Surprisingly, the majority of 12 endogenous transcripts from both essential and non-essential genes failed to induce a measurable reduction in transformation efficiency but the reduction fold surged to 280-fold to 1000-fold when the targets were expressed at a higher level from a plasmid. Elena observed a sharp transition between the low and high reduction folds when the target transcripts were expressed under eight different constitutive promoters exhibiting varying expression strengths, suggesting the transcript expression level had an impact on Cas13a activation. Considering the observations were limited to a small number of transcripts, we aim to achieve a global assessment of the causal effects of expression level and the role of other factors (such as target position and sequence context) with a genome-wide screen in *E. coli*.

A library of 25,597 crRNAs targeting 4228 mRNAs and rRNAs in *E. coli* was designed (**Figure 3.2.2A; Methods 2.3.1**). The crRNA candidates were selected with non-GU PFS, GC content, and secondary structure to increase the binding efficiency and avoid decoupling of activation and one crRNA targeting each of ten evenly-spaced sections in the gene was maintained to explore the effect of the targeting locations. The library further included 400 randomized non-targeting guides lacking complementarity to any endogenous transcripts as the negative control. The crRNA library was then transformed into *E. coli* cells with or without LshCas13a. Cell death or dormancy caused by the successful activation of Cas13a could deplete the crRNAs in the transformed cells. Short-read sequencing was finally applied to extracted plasmids after culturing to measure the depletion of each guide compared to the no-LshCash13a control cultured under the same conditions (**Figure 3.2.2B**). Under this setup, highly active guides would be heavily depleted within the library, while poorly active guides would be minimally depleted.

Given the stronger fitness defect of silencing essential genes, I first compared the depletion of guides targeting essential or non-essential genes using the 0.01 quantile in the logFC values of non-targeting guides as the threshold to define depletion. 85% of guides targeting essential genes exhibiting logFC below the threshold were depleted, substantially higher than 43% for guides targeting non-essential genes. The median depletion was also higher for guides targeting essential (2.9-fold) versus non-essential (0.3-fold) genes (**Figure 3.2.2C**). The screen was validated by testing individual highly depleted guides using the transformation assay (**Figure S2.1A**). This discrepancy in the extent of crRNA depletion indicated a gene-dependent effect while targeting non-essential genes led to less frequent collateral cleavage and dormancy despite following the same crRNA design rules.

In comparison to essential and non-essential genes, guides targeting rRNAs, the highest expressed RNAs in the cell, were strongly and consistently depleted in the library (**Figure 3.2.2C**). In line with the previous result, I investigated how transcript expression levels were associated with guide depletion across the library. A strong correlation was found between guide depletion and the levels of the targeted transcript with Spearman coefficients 0.72 and 0.68 in middle and late exponential growth, respectively (**Figure 3.2.2D; Figure S2.1B**). The stronger correlation in middle exponential growth might be due to the dominance of middle exponential growth in the screen. Translational strength predicted with the ribosome-binding site (RBS) calculator showed minimal correlation (Salis, 2011) (**Figure S2.1C**), indicating that protection of the mRNA by translating ribosomes does not account for differences in guide depletion. The observed correlation applied to both essential and non-essential genes (**Figure S2.1D**). The stronger depletion of guides targeting essential genes with similar expression levels to non-essential genes, further supporting the added growth defect when silencing essential genes (**Figure S2.1D-E**).

To determine the factors that play a role in the successful induction of cytotoxic autoimmunity, I applied auto-sklearn to optimize a machine learning model to predict the guide depletion in the screen (**Figure S2.2A**). The model interpretation using SHAP values (Lundberg et al., 2020) revealed that transcript levels were the strongest predictor of guide depletion. In contrast, the effect of gene essentiality was modest, which could be a result of the small number of essential genes in *E. coli*. I also identified nucleotide preferences within the crRNA guide or the PFS with the predictive value, such as a C in the 1st or 2nd PFS position or a U in the 19th crRNA guide position (**Figure 3.2.2E; Figure S2.2B-D**). The existing Cas13d design tool based on experiments in human cell culture (Wessels et al., 2020) showed poor accuracy in predicting depletion scores for our bacterial experiments with Cas13a (**Figure S2.2E**), indicating the differences in the captured predictive patterns. The results from the machine learning

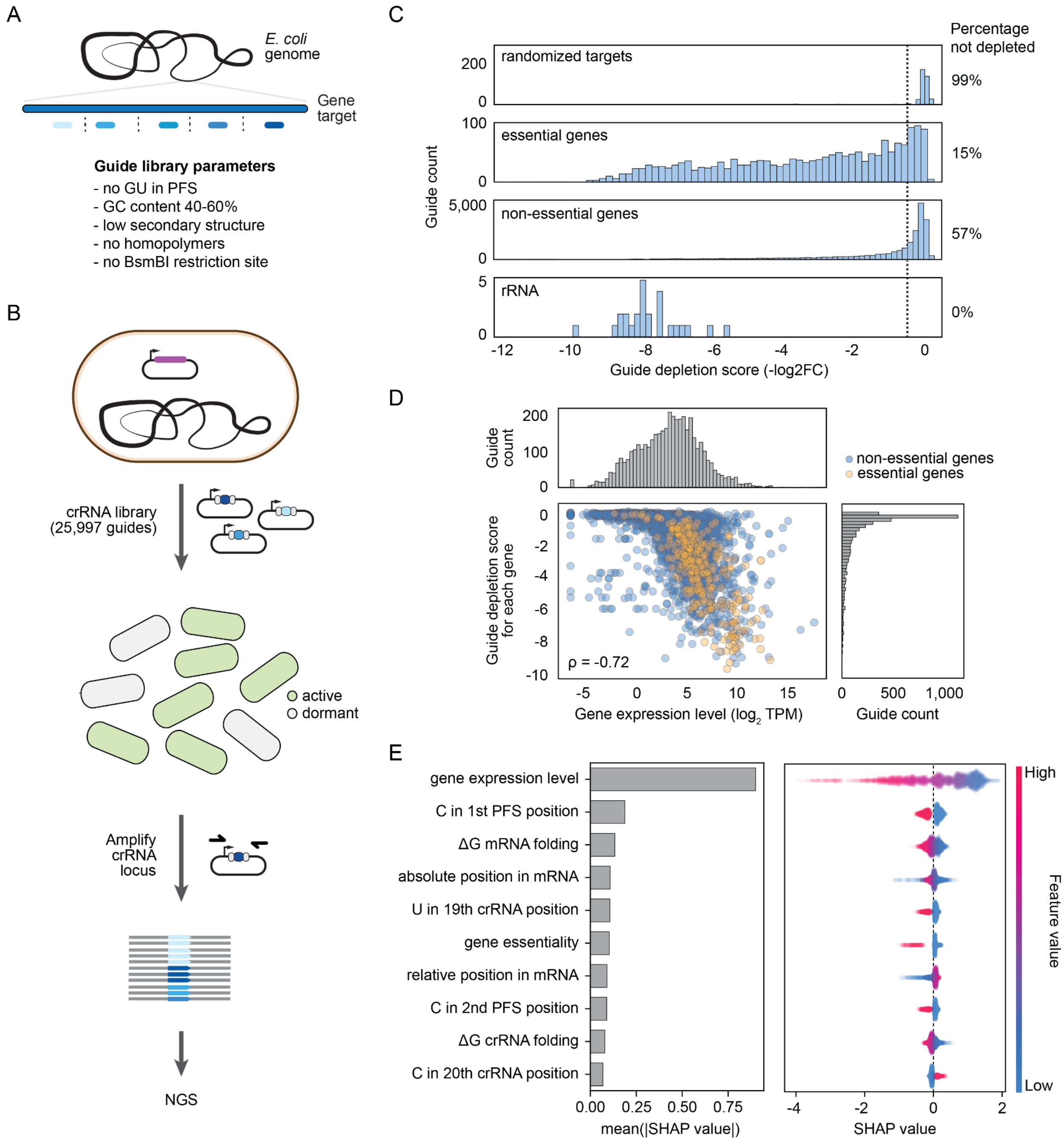

Figure 3.2.2: A genome-wide CRISPR-Cas13a screen reveals target expression levels as the main determinant of cytotoxic self-targeting. (**A**) Design of crRNA guide library. Guide selection accounted for standard rules lending to efficient targeting and spanning the entire coding region of each target gene. Other parameters (i.e., homopolymers, BsmBI sites) were included to facilitate library synthesis and cloning. The resulting library included 25,470 guides targeting protein-coding genes, 127 guides targeting rRNAs, and 400 randomized guides as negative controls. (**B**) Workflow for library screening. As part of the screen, cells with or without a LshCas13a plasmid (purple) are transformed with the crRNA plasmid library and cultured while selecting all present plasmids. Guide depletion is determined in comparison to the same workflow with the no-LshCas13a control. (**C**) Distribution of depletion scores for different groups of guides within the library. The cutoff for no fitness defect is based on the range of depletion scores for a set of randomized guides. (**D**) Correlation between guide depletion score and the expression levels of the target gene. Expression levels were measured by RNA-seq analysis with *E. coli* cells harboring the no-LshCas13a control and subjected to the library workflow to turbidity of $ABS_{600} \approx 0.5$. Values for transcript levels and guide depletion are the average of duplicate independent experiments and screens, respectively. ρ: Spearman coefficient. See **Figure S2.1B** for the correlation for turbidity of $ABS_{600} \approx 0.8$. (**E**) SHAP values for the strongest predictors of guide depletion from the library. The left barplot indicates the average absolute contribution of each feature to the predicted depletion values, while the right beeswarm plot shows the impact of each feature on each individual prediction.

approach together support the role of the target expression level as the third criterion to activate the Cas13a-induced immunity and that most self-targeting guides do not activate cytotoxic autoimmunity due to insufficient target expression. This additional criterion proposes an explanation for the often observed gene silencing instead of cell death or dormancy by Cas13 in eukaryotes: given the larger cell size and distribution of transcripts, the expression threshold in eukaryotes is likely to be higher, leading to the uncoupling between target binding and collateral cleavage. It was further demonstrated in the publication (Vialetto et al., 2022) that the expression level threshold also allows the Cas13a system to tolerate a lowly expressed invader such as a plasmid yet display a robust immune response to a highly expressed invader as an actively replicating lytic phage.

The application of the machine learning approach in this study provides independent support other than the experimental observation to understand the role of transcription level in the mechanism of the CRISPR-Cas13 system.

3.2.3 Analysis of CRISPR-Cas12a genome-wide screen in *Klebsiella pneumoniae*

*The section is part of the results for "**Target-dependent features dictate Cas12a antimicrobial activity against multidrug resistant and hypervirulent strains of Klebsiella pneumoniae**". The work is a result of collaboration with Elena Vialetto (Helmholtz Institute for RNA-based Infection Research), Solange Miele, and Moran Goren (Pasteur Institute), who performed library design, initial sequencing data processing, all laboratory experiments, and finalized the figures.*

Klebsiella pneumoniae has been a threat to thousands of lives due to its increasing multi-drug resistance (MDR) and hypervirulent (HV) genes (Santajit & Indrawattana, 2016). The decreasing efficacy of antibiotics stresses an alternative antimicrobial strategy to prevent the death of millions (Murray et al., 2022). To address the pressing need, CRISPR-Cas systems emerged as a promising means to achieve infection clearance by specifically eliminating pathogens from a microbial community (Uribe et al., 2021; Rubin et al., 2022). To extend the application of CRISPR-Cas systems as antimicrobials, Elena identified *Acidaminococcus sp.* Cas12a (AsCas12a) as one of the best candidates in *K. pneumoniae* strain ATCC 10031 after comparing the activity of several Cas nucleases. However, the antimicrobial activity of AsCas12a in different MDR and HV *Klebsiella* strains was not always consistent. Elena showed that it was associated with the secondary structure of the gRNAs.

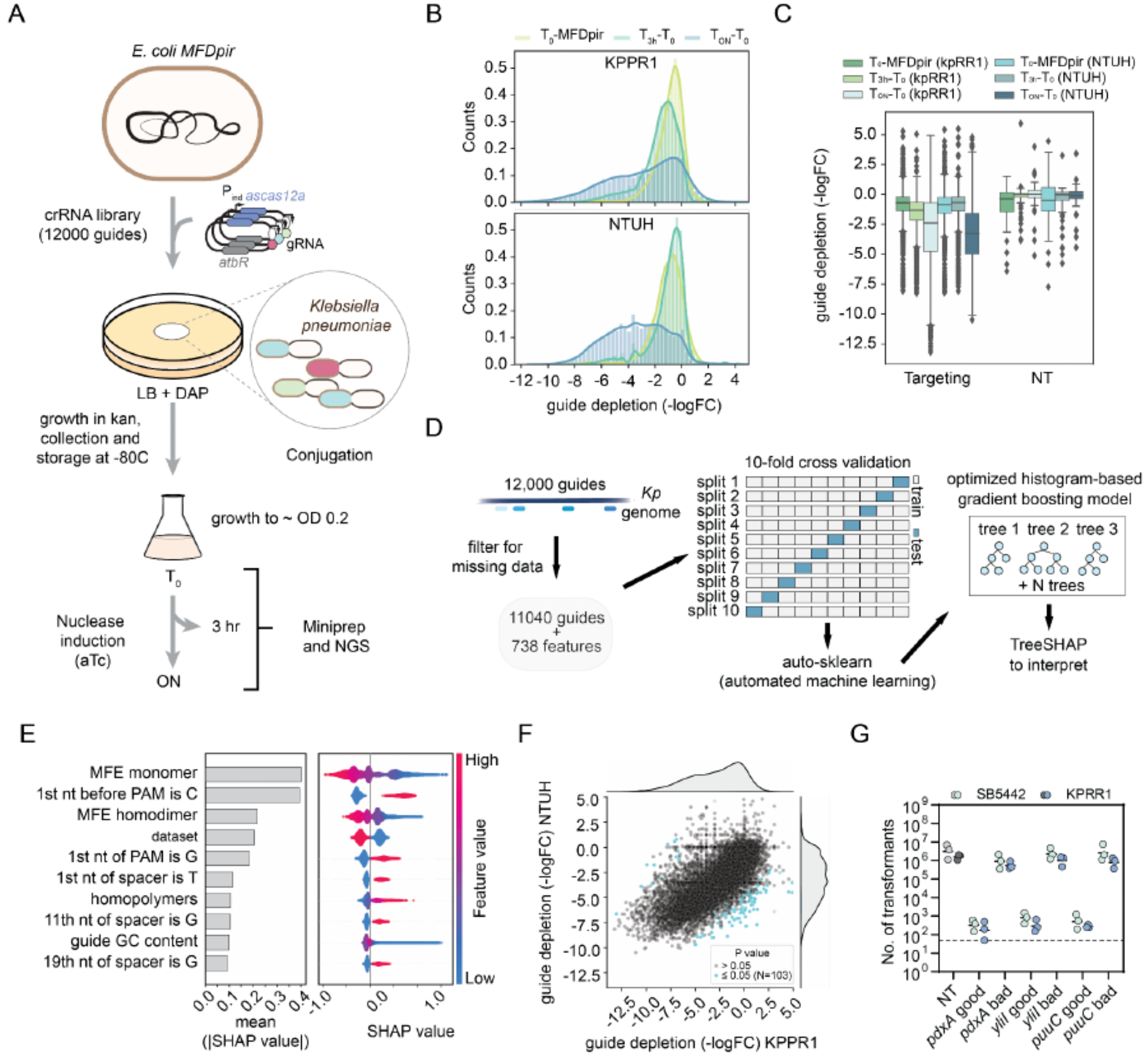

Figure 3.2.3: A genome-wide AsCas12a screen confirms guide differential depletion across strains and advises guide design rules for efficient antimicrobial activity. (**A**) Schematic of library screen setup. Samples were collected at 0, 3 h, and overnight (ON) after induction. Guide depletion is determined by comparing the abundance of guides between time points. DAP: diaminopimelic acid. aTc: anhydrotetracycline. (**B**) Depletion of guides prior to induction (T_0 - MFDpir) and following nuclease induction for 3 hours or overnight ($T_{3h/ON}$ - T_0). MFDpir: the *E. coli* strain used for guide library preparation. (**C**) Box plots of guide depletion distribution before and after aTc nuclease induction for the targeting and NT guides. Boxes and whiskers indicate 50 and 99% intervals, respectively. (**D**) Depiction of method used for developing the machine learning algorithm. The derived log2 fold-changes between 0 h and overnight after induction from the screen data were used as prediction targets, with 738 sequence and thermodynamic features, as well as dataset, as predictors. Auto-sklearn was used to train and optimize an assortment of candidate model types. A histogram-based gradient boosting model was the best-performing model. This model was then interpreted using TreeSHAP to explain the contribution of each predictor to each prediction. (**E**) Contribution of different features to guide depletion from the library using the mixed dataset for training the algorithm. The left barplot depicts the absolute weight of each feature to the depletion values, while the right beeswarm plot shows how each datapoint impacts depletion according to each feature. (**F**) Correlation between guide depletion in the kPPR1 and NTUH strains. Blue dots indicate the 103 gRNAs with the differential difference between strains. (**G**) Validation of algorithm prediction of six high and low-efficiency guides in KPPR1 and SB5442. The dashed line represents the limit of detection of the assay.

To improve the consistency of antimicrobial activity across strains, genome-wide screens in two HV *Klebsiella* strains (NTUH-2044, KPPR1) were employed to determine across-strain transferable guide design rules. A library of 12,000 gRNAs including 11,900 guides targeting the coding and non-coding chromosomal regions in both strains and 100 non-targeting guides was designed, followed by transforming the cells with a conjugative plasmid encoding the constitutively expressed gRNA library and inducibly expressed AsCas12a (**Figure 3.2.3A**). Next-generation sequencing was applied to extracted gRNA-carrying plasmids after inducing the expression of AsCas12a for three hours and overnight. The variation in the gRNA abundance compared to the initial time point prior to induction would inform the gRNA activity. While I observed significant guide depletion after overnight incubation, the depletion was moderate after a three-hour induction (**Figure 3.2.3B-C**).

I next applied the machine learning approach to extract rules for guide design using the depletion scores after overnight induction from both screens (**Figure 3.2.3D**). The model interpretation using SHAP values emphasized the detrimental effect of a more stable self-folding and homodimer structure of gRNAs and the disfavoring of a cytosine before the PAM sequence (**Figure 3.2.3E**), consistent with the previous observation of the importance of gRNA folding secondary structure and uncovering a previously unnoticed PAM preference: guanine at the last position in PAM immediately upstream of the guide or thymine at the first position of the guide was neither favored for efficient antimicrobial. The data source was also listed as a strongly predictive feature, strengthening the idea that guide activity can vary between strains. However, when comparing guide depletion between the two strains, I observed a strong correlation (Pearson correlation coefficient 0.705) with only 0.9% of the guides having a significant differential activity (**Figure 3.2.3F**). This higher consistency across strains compared to the previous observation in MDR and HV strains suggests strain-specific responses or that the genome-wide screens yielded robust antimicrobial activity. Training the models on individual strains also delivered almost identical predictive factors with comparable prediction performance (**Figure S2.3A-C**), indicating the decent generalization of the captured guide design rules. These design rules were further validated using six independent guides with either high or low activity according to the model prediction. The test was performed in the KPPR1 strain and in the SB5442 strain which was not used to train the model. For both strains, the model prediction was accurate with predicted efficient guides exhibiting more than 103-fold reduction in transformation while the predicted inefficient guides showed an undetectable decrease (**Figure 3.2.3G**), strengthening the fundamental role of secondary structure and sequence compositions in antimicrobial activity and suggesting the good generalizability of the predictive model.

In conclusion, the machine learning approach explicated the rational design rules for AsCas12 in *K. pneumoniae* strains and the predictive model could benefit the guide design for the pathogen clearance in other strains, fueling the application of CRISPR-Cas systems as an alternative strategy to combat the infectious diseases.

4. Discussion

In this thesis, I developed a random forest model from three genome-wide essentiality screens in *E. coli* to predict guide efficiency for CRISPRi in bacteria after segregating the confounding gene-relevant effects using mixed-effect models. I extensively explored the effects of feature engineering, data integration, model selection, and hyperparameter tuning on depletion prediction and demonstrated the rigorousness of autoML in model development. In light of the dominant gene effects from model interpretation, I showed that MERF provided a more reliable prediction of guide efficiency and outperformed existing tools for both Cas9 and CRISPRi in independent validation experiments, including a large-scale saturating screen of purine biosynthesis genes in minimal media. Further, I applied the derived machine learning approach to analyze genome-wide screens of CRISPR base editor, Cas13a, and Cas12a in either *E. coli* or *K. pneumoniae*, reinforcing its robustness and adaptiveness while gaining insights into CRISPR-Cas mechanisms.

I addressed the need for a predictive machine learning model specifically to predict gRNA efficiency for CRISPRi in bacteria, given the rapid development of the CRISPRi techniques and the crucial role of gRNA design. Beyond a simple alternative to gene knockout, the CRISPRi techniques possess the potential for genome-wide screening, multiplexed targeting, and gene expression tuning. Genome-wide CRISPRi screens enable exhaustive functional interrogation of thousands of genes in one sitting, avoiding the laborious collection of single knockout mutants (Baba et al., 2006). Similarly, CRISPRi multiplexed targeting would largely simplify the investigation of genetic interactions (Kuzmin et al., 2018) and the construction of complex synthetic circuits (Lian et al., 2017; Santos-Moreno et al., 2020), in comparison to the state-of-art double mutant construction method by mating (Butland et al., 2008; Typas et al., 2008). A genetic interaction network can additionally reveal synergetic targets for combinations of antibiotics or antibiotic-antibody to obtain more efficient pathogen clearance and to avoid antimicrobial resistance (Doern, 2014; X. Xu et al., 2018; Leshchiner et al., 2022), while synthetic circuits can be devised to optimize production of particular small molecules or proteins, such as biofuel, pharmaceuticals, and biomaterials (Khalil & Collins, 2010; M. Xie & Fussenegger, 2018). The CRISPR array technologies allow the co-expression of up to 22 gRNAs in one array and a large-scale screen of arrays (Liao, Ttofali, et al., 2019; A. C. Reis et al., 2019), enabling comprehensive analysis of even higher-order genetic interactions. The inducible and titratable CRISPRi systems shed light on the complex relationship between gene expression and fitness (Hawkins et al., 2020; Bosch et al., 2021). Optimal gRNA efficiency is however indispensable to realize the full potential of CRISPRi techniques. While large-scale experiments demand the selection of gRNAs for thousands of genes, my predictive model and web-based tool are able to design and select efficient gRNAs from input sequences or specified genes, providing a straightforward solution to simplify and improve the CRISPRi-based experiments.

Additionally, I successfully applied auto-sklearn to optimize the machine learning model and demonstrated its robust performance in predicting depletion. Despite the fact that machine learning models have already been exploited for guide efficiency prediction in CRISPR-Cas systems, the application of the emerging autoML techniques (Waring et al., 2020) is under-investigated. My exploration presented in this thesis is a proof-of-concept, showing the great potential for model improvement using the continuously developing autoML tools. While CRISPRi genome-wide screens have been performed on a wide range of microbes and under various conditions (Rock et al., 2017; Rousset et al., 2018; T. Wang et al., 2018; H. H. Lee et al., 2019; Hawkins et al., 2020; McNeil et al., 2021; Spoto et al., 2022), increasing CRISPR-based tools have been applied to prokaryotes (Vento et al., 2019; Vigouroux & Bikard, 2020; Z. Liu et al., 2020; Meliawati et al., 2021). In this thesis, I also showed that the model development workflow leveraging the power of autoML led to meaningful model interpretations in three distinct genome-wide screens with the base editor, Cas13a, and Cas12a, again confirming the robustness of auto-sklearn. Surprisingly, histogram-based gradient boosting models were uniformly selected for all the screens. Tree-based models showed superior performance in some previous studies (Doench et al., 2016; Muhammad Rafid et al., 2020), probably due to the non-linear relationship between sequence features and guide efficiency and the fact that tree-based models can better capture the interactions between features. Indeed, the performance of other model types failed to outperform the selected histogram-based gradient boosting model (**Figure S1.3B-G**), suggesting the selection is not a technical bias from auto-sklearn and tree-based models can better untangle the connections between gRNA features and efficiency. Another autoML tool, H2O, arrived at the same best-performing model selection, albeit with worse performance (**Figure S1.3A**). The simplicity of implementing the autoML tool and the minimum demand of relevant experience play a central role in the machine learning approach that I have developed and the subsequent successful application of the approach has proven to have the potential to accelerate the development of machine learning models and lead to a deeper understanding in other CRISPR-Cas systems and organisms.

Despite the simplification in model development using autoML, data collection and feature selection and engineering to refine the models remain challenging and limitations persist in autoML tools. In this work, I vastly adopted genome-wide screens for model training. While genome-wide screens guarantee a considerable number of data points and minimal batch effects, simultaneously providing a complete landscape of gRNAs, the high number of target genes exhibit substantial gene-specific effects. I found that these gene-specific effects confounded the interpretation of guide efficiency when predicting depletion scores from CRISPRi essentiality screens (**Figure 3.1.1C**). My observations in the saturating screen for purine biosynthesis genes also indicated that the gene-specific effects lower the prediction accuracy (**Figure 3.1.5B-D; Figure S1.9A-B**), leading to an open question of whether the fine-tuned selection of target genes will improve the performance and the generalization, such as selecting target genes based on the standard deviation of depletion scores of guides (**Figure S1.9C**). Moreover, the minimal number of gRNAs per gene or target genes required for comparable performance needs future investigation. I included only the gRNAs targeting previously proven essential genes from the CRISPRi screens for the model training, considering those genes are expected to deplete during the screening thus enabling measurable variation of gRNA efficiency. Surprisingly, these roughly 10% essential-gene-targeting gRNAs among all available guides were sufficient for developing a moderately accurate model for guide efficiency (**Figure 3.1.1B**). But I demonstrated that data integration of multiple screens improved the model performance both in direct depletion prediction and in gene-segregation-based methods (**Figure 3.1.2B; Table 1; Figure S1.8C**). Therefore, it remains possible to significantly improve the model accuracy by including more data, such as other screens and gRNAs targeting non-essential but strongly depleted genes. Besides accuracy, data integration also gives rise to better generalization (**Figure 3.1.2B**) probably due to the reduction of batch effects, suggesting training the model with screens from different conditions and organisms can generate a better-performing model for the application in understudied bacteria, such as the non-model bacteria and novel pathogenic isolates. Together with the issue of the gene-specific effects, whether there is a balance or it would be a trade-off between expanding the sample size and fine-tuning the target gene set leaves ample space for further model improvement for wider applications.

In addition, I showed that my rich, biologically relevant feature set is important for accurate prediction of CRISPRi depletion and it can be easily adapted for other CRISPR-Cas systems. I found that only guide-specific features are insufficient to predict CRISPRi depletion. Starting with sequence features, I included single nucleotide 30 mer extended gRNA sequences, considering that the 30 mer covers the potential relevant regions suggested by the previous study (Calvo-Villamañán et al., 2020). The model interpretation confirmed that only the nucleotide immediately downstream of PAM has a significant contribution among flanking nucleotides around gRNA and PAM. Despite both single and di-nucleotide features being commonly enlisted, additionally including di-nucleotides worsened the performance (Yu et al., 2022), probably due to the excessive number of features compared to the limited sample size. Inspired by previous findings in relation to the importance of targeting position relative to transcription start site for CRISPRi (Qi et al., 2013; Gilbert et al., 2014; Smith et al., 2016), I included absolute and relative distances in CDS. Interestingly, the depletion model did not perform better but distance features are the most predictive predictors in the fixed-effect model from MERF (**Figure 3.1.1C; Figure 3.1.4A**). Despite the gene-specific effects in the screen of purine biosynthesis genes, which exhibited a weak correlation between distance to start codon and guide efficiency for a few genes (**Figure 3.1.5D; Figure S1.9A**), removing distance features degraded the model performance (**Figure S1.8E**). The seemingly inconsistent importance of target position might stem from two perspectives: 1) dCas9 can more effectively stop the transcription by preventing the RNA polymerase (RNAP) from escaping the promoter when targeting shortly downstream of the start codon but the efficacy drops when dCas9 is required to stop the actively transcribing RNAP when targeting farther downstream positions; 2) the function of the truncated protein caused by CRISPRi depends on the location of the protein domains and it is gene-specific. Compared to sequence and distance features, thermodynamic features are less predictive in either the depletion (**Figure 3.1.1C**) or MERF model (**Figure 3.1.4A**), which is inconsistent with the previous finding for the Cas9 system (Xiang et al., 2021) where the thermodynamic feature ($\triangle G_B$) was the most predictive. $\triangle G_B$ is calculated using the CRISPRoff pipeline (Alkan et al., 2018), which considers DNA opening, gRNA self-folding, the hybridization of gRNA and target DNA, and the PAM sequence. Replacing these four thermodynamic features with $\triangle G_B$ however did not improve the model performance for depletion prediction or the purine screen (**Figure S1.10A-D**) and $\triangle G_B$ is not among the most predictive features (**Figure S1.10E**), suggesting that the low importance is independent of the choice of thermodynamic features. The minor role of the thermodynamic characteristic in CRISPRi in bacteria compared to Cas9 in eukaryotes might be explained by the more compacted DNA structure in eukaryotes, where binding to the target DNA becomes a chokepoint for efficient editing by Cas proteins. One of the most surprising findings, when I investigated the features, is that the gene-specific features contributed more to the prediction, measured by SHAP values (**Figure 3.1.1C**), indicating that the variation in the depletion scores is mostly driven by the target genes. The gene-specific fitness effect that is indicated by varying continuous values suggests a non-binary character of gene essentiality. Among the gene features, the expression level is the strongest predictor, supporting the previous findings that gene essentiality can be varied by transcriptional activity (Bauer et al., 2015; Keren et al., 2016; Hawkins et al., 2020; Bosch et al., 2021). Including the number of downstream genes in the operon asserts the potential polar effect in the transcription units (Cui et al., 2018), while the higher contribution of the number of essential genes proposes a more specific explanation. The rich gene features improved the model performance for depletion prediction and contributed to segregating the gene-specific effects using MERF. Considering the lack of operon information and transcriptomic data for other organisms, I also approximated expression level with the codon adaptive index (CAI) (M. dos Reis et al., 2003) and removed operon-related information, with promising results (**Figure S1.8E**).

The importance of biologically relevant predictors in predicting the depletion from CRISPRi essentiality screens also emphasizes the necessity to adapt the features according to the data type and source. For example, when developing prediction algorithms for a CRISPR base editor, attention should be paid to the editing window which differs across base editors and poses importance in editing efficiency. For the Cas13a system, given it targets RNAs, thermodynamic features to describe the secondary structure of the target RNA and the hybridization energy between the gRNA and target RNA should be considered. Moreover, the definition of seed region, in which one mismatch might abort the target binding, might be unclear or different in different CRISPR-Cas systems.

Whereas previous guide efficiency models were often trained on direct measurement of guide efficiency (i.e. indel rate), CRISPRi essentiality screens only provide indirect measurement, which includes both gene- and guide-specific effects. Therefore, gene-specific features, such as the previously unreported gene expression, significantly improved the model performance in predicting depletion, strengthening the importance of the data-oriented feature set. Beyond this, I adopted two distinct methods to remove gene-specific effects in order to extract guide efficiency from CRISPRi essentiality screens: modeling gene-specific effects as random effects in a mixed-effect model and using the median logFC value as a proxy for gene-specific effects. While both methods accomplished the task, the former method using mixed-effect random forest (MERF) (Hajjem et al., 2014) outperformed the latter one which subtracts the median gene-wise logFC values from depletion scores as described previously (Calvo-Villamañán et al., 2020) (**Figure 3.1.5B-C; Figure S1.8D; Table 1; Table S7; Table S14**), which might be explained by the more accurate segregation enabled by the rich gene-specific features for the random-effect model. The superior performance of MERF can be enhanced when more data from various sources are integrated, considering the potentially larger batch effect. Moreover, it might become more important to describe the gene-specific effects using biologically relevant features when incorporating data of, for example, non-essential genes and other strains or organisms.

As I optimized the random forest fixed-effect model for MERF, I noticed the technical difficulty of implementing auto-sklearn directly. Hence, I tested the auto-sklearn-selected best-performing histogram-based gradient boosting and random forest models for depletion prediction, in addition to the optimized random forest model using a hyperparameter tuning tool hyperopt. While the models selected by auto-sklearn resulted in a decent performance, the hyperopt-optimized model performed the best (**Figure S1.8F; Table S7; Table S14**), suggesting the robust performance of auto-sklearn might not necessarily be optimal. Actually, applying hyperparameter tuning tools like hyperopt requires more domain knowledge, indicating a trade-off between optimal performance and runtime.

The state-of-art tool for CRISPRi in bacteria, which I refer to as 'Pasteur', implemented a LASSO model and pointed out that a more complex model would not improve the performance (Calvo-Villamañán et al., 2020), which is inconsistent with the better performance of the median subtracting (MS) random forest model compared to the MS LASSO model (**Figure S1.8D; Table S14**). But the retrained Pasteur on our integrated data performed similarly to the MS random forest model. On one hand, given the same feature set in both MS random forest and MS LASSO models, a tree-based model showed potential for performance improvement. On the other hand, the reduced sequence feature set used by the Pasteur model might be favored by simple models like LASSO, suggesting the feature set is crucial to the choice of the model type.

The MS method is however more flexible in the choice of models, whereas MERF relies on tree-based models. I tested shallow models, such as random forest and LASSO, and two CNN-based deep learning models using the MS method. Even though I did not see a significant improvement in the MS deep learning models (**Table S7; Figure S1.8D**), they possess the potential for improvement when more data is available. Moreover, transfer learning in combination with deep learning, which applies a pre-trained model to a related problem to reduce the required sample size and training time, can take advantage of the tremendous data generated in laboratory strains, such as strains of *E. coli*, to understand other bacteria. The transfer learning approach has been found promising, for instance, to translate the findings in the mouse model to human (Brubaker et al., 2019) and the information from the reference atlas to new single-cell data (Lotfollahi et al., 2021). Of note, integrating data for the MS method demands a non-trivial scaling normalization for the depletion scores to avoid batch effects given that the readout from screens under different conditions exhibits qualitative discrepancy (**Figure 3.1.2A**). Whether the integration process can be simplified using more advanced methods, i.e. conditional variational autoencoder (Simidjievski et al., 2019), requires further investigation.

To evaluate the models, I applied various validation methods and showed MERF outperformed previous existing tools in both training data and independent data. Cross-validation is common for machine learning model hyperparameter tuning and evaluation, providing better generalization by using non-overlapping test sets in each iteration. For MERF, I optimized the hyperparameters using gene-wise cross-validation. The better performance of MERF in individual datasets besides training data demonstrated satisfactory generalizability and the practicality of cross-validation. While cross-validation focuses on the train-test split, Spearman correlation is often applied as the metric to evaluate the prediction accuracy. But Spearman correlation also suffered from the high sensitivity to the sample size, as the median Spearman correlation in the training data is below 0.4 where less than 10 gRNAs were designed for most of the genes (**Table 1**). In contrast, the median Spearman correlation fluctuates from 0.5 to 0.7 in the independent validation data (**Table S14**), where as high as 223 guides were designed for one gene. While evaluating with a few dozen of guides in fluorescent-based and miller assays is less convincing, the saturating screen of nine purine biosynthesis genes provided a more reliable estimate. Aside from the Spearman correlation, I further included positive predictive value to evaluate the accuracy of selecting the efficient guides given that only three to five guides are commonly selected for one gene in practice. These two metrics together provided an in-depth evaluation and suggested that caution should be taken in choosing evaluation metrics.

Beyond the predictive power of the machine learning models, I leveraged the interpretability of the optimized models to extract rational design rules for various CRISPR-Cas systems. For CRISPRi in bacteria, I observed highly consistent interpretation across models (**Figure 3.1.4A; Figure S1.4F; Figure S1.10E**), indicating the proposed design rules are not model-dependent. To my surprise, despite most of the highlighted sequence features being concordant with the CRISPR-Cas9 system, the effect of the 20th position in gRNA is the opposite. The reversed effect might indicate the structural differences between the target binding and cleavage considering that dCas9 only binds to DNA without introducing DSB. Strikingly, the interpretation using TreeSHAP captured the specific effects for different CRISPR-Cas systems. For the CRISPR base editor, the sequence context in the editing window (position 5 to 8 in the gRNA) has a major effect (**Figure 3.2.1E**), corresponding to the importance of successful editing to introduce premature stop codons for gene silencing. Since the NGG PAM of *Streptococcus pyogenes* Cas9 (SpCas9) limits the number of targetable positions, *Streptococcus canis* Cas9 with a more flexible NNG PAM was utilized in the base editor screen. In spite of the proposed PAM flexibility, SHAP values suggested that there is a distinct preference in the PAM sequence and the downstream sequence. Such a rapid means to investigate the PAM preferences of Cas proteins with the combination of large-scale screening and machine learning is valuable for adapting CRISPR-based tools to more microbes, given that the adaptation of alternative Cas proteins is inevitable (Rock et al., 2017) due to the low efficiency and toxicity of the widely used SpCas9 in some bacteria (S. Cho et al., 2018).

For Cas13a, the target expression level exhibited the highest contribution, and the effects of this predictor correspond to the observed strong correlation between the target expression level and depletion scores (**Figure 3.2.2D-E**). Despite gene expression level also impacting the depletion score in the CRISPRi essentiality screen, the effect was only observed in essential genes. Whereas gene essentiality plays a pivotal role in the extent of depletion in the CRISPRi essentiality screen, gene essentiality has a much weaker impact in the Cas13a system (**Figure 3.2.2E**), strengthening the determinant role of transcriptional level in endogenous targeting by Cas13a. In addition, the model interpretation also pointed out the PFS sequence preference. Given that the PFS sequence in the Cas13a system plays an equal role as the PAM sequence in the Cas9 system, the highlighted attributions of both PFS and PAM sequences in various systems suggest the importance of non-self recognition in guide efficiency. Interestingly, in both Cas13a and Cas12a systems, thermodynamic features have more significant contributions than that in CRISPRi (**Figure 3.2.2E, Figure 3.2.3E**), more concordant with the Cas9 system (Xiang et al., 2021). The complexity in the secondary structure of the target sequence has a higher contribution in the Cas13a system, whereas the gRNA self-folding is the strongest predictor in the Cas12a system, which can be explained by the fundamental difference in the target sequence type - Cas13a targets RNA while Cas12a targets DNA. Despite the uniformity of predictive features for Cas12a in different strains of *K. pneumoniae* (**Figure S2.3A-B**), it is unclear whether the guidelines for CRISPR-Cas techniques generalize well across bacterial species. The degraded prediction accuracy of the CRISPRi model trained with *E. coli* data in *Salmonella* (**Figure S1.7B**) however hinted at the lack of model generability across bacteria, but the model was trained on data from the same strain. Nevertheless, the discrepancy in predictive features for different CRISPR-Cas systems underlines the need for independent design rules. As CRISPR genome-wide screens are more routinely performed to characterize microbes (Todor et al., 2021), the interpretation of a predictive algorithm trained on the screen data with the proposed machine learning approach can provide insights into the CRISPR systems, facilitating the further improvement and application of the CRISPR-based tools.

Even though I demonstrated that my MERF model outperformed existing tools, there is still room for improvement. Firstly, expanding the datasets by integrating screens under various conditions or from different strains can enhance the generalizability of the model, and including gRNAs targeting non-essential genes that are significantly depleted can enlarge the sample size, potentially improving the accuracy. Secondly, the feature set can be fine-tuned. While an exhaustive search of an optimal reduced feature set can be performed to remove correlated or redundant features, features can also be optimized using other calculation methods or the feature set can be extended to describe other gRNA properties. For example, one can alternatively incorporate melting temperature as a thermodynamic feature, implement transcriptional data from the same experimental condition, and include di- or tri-nucleotide features. Thirdly, the searching space for the hyperparameters of the model can be optimized and other tools for hyperparameter tuning, such as Optuna (Akiba et al., 2019) and Tune (Liaw et al., 2018), can be tested. Lastly, multiple models can be trained on a portion of data, followed by selecting one of these models for prediction based on the similarity between the training and the query data to obtain a better performance while avoiding overfitting the training data.

In combination with autoML, my exploration of data integration, feature engineering, model selection, and interpretation provides a blueprint for applying machine learning in the analysis of CRISPR systems. After demonstrating the robustness of my adaptive machine learning approach in characterizing various CRISPR systems in a variety of bacteria, I expect the approach can be easily applied to any bacterium and CRISPR-Cas system of interest. Therefore, my adaptive machine learning approach can not only address the growing demand for rational guide design with the development of novel CRISPR-based tools and their increasing practice, but also expand the application of CRISPR-Cas systems in other bacteria. While much more efforts have been invested in the application of CRISPR in eukaryotes (Komor, Badran, et al., 2017), the wealth of knowledge in immune responses in humans enables host-directed personalized therapies for infectious disease based on immune profiling (Sundaresh et al., 2021). Advancing the application of CRISPR-Cas systems in prokaryotes can uncover novel drug targets (Bosch et al., 2021; McNeil et al., 2021; Feng et al., 2022) and inform host-pathogen interactions (Sidik et al., 2016; Ferrand et al., 2018; Casadevall Arturo & Pirofski Liise-anne, 2000), strengthening both host- and pathogen-directed treatments (Abreu et al., 2020). Moreover, the availability of CRISPR-based tools in more bacteria reinforces the strain-specific targeting using CRISPR in a complex microbial population (Uribe et al., 2021; Rubin et al., 2022; Rottinghaus et al., 2023) as an alternative therapeutic strategy for infectious diseases. For instance, the machine learning approach can be applied to develop a predictive algorithm with strong generalizability from CRISPR screens in multiple laboratory culturable strains. Given that whole genome sequencing is inexpensive, efficient gRNAs can be designed to eliminate unculturable pathogenic strains that are close to the laboratory strains using the developed predictive algorithm. While antimicrobial resistance has been a growing threat to millions of lives (Murray et al., 2022), expanding the application of CRISPR-Cas systems in bacteria paves a path to advanced therapeutics for bacterial infection.

References

Abadi, S., Yan, W. X., Amar, D., & Mayrose, I. (2017). A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Computational Biology*, *13*(10), e1005807.

Abreu, R., Giri, P., & Quinn, F. (2020). Host-pathogen interaction as a novel target for host-directed therapies in tuberculosis. *Frontiers in Immunology*. https://www.frontiersin.org/articles/10.3389/fimmu.2020.01553/full

Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., Verdine, V., Cox, D. B. T., Kellner, M. J., Regev, A., Lander, E. S., Voytas, D. F., Ting, A. Y., & Zhang, F. (2017). RNA targeting with CRISPR-Cas13. *Nature*, *550*(7675), 280–284.

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B. T., Shmakov, S., Makarova, K. S., Semenova, E., Minakhin, L., Severinov, K., Regev, A., Lander, E. S., Koonin, E. V., & Zhang, F. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, *353*(6299), aaf5573.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H., & Gorodkin, J. (2018). CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology*, *19*(1), 177.

Allen, F., Behan, F., Khodak, A., Iorio, F., Yusa, K., Garnett, M., & Parts, L. (2019). JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Research*, *29*(3), 464–471.

Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L., & Lombardo, A. (2016). Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. *Cell*, *167*(1), 219–232.e14.

Amitai, G., & Sorek, R. (2016). CRISPR-Cas adaptation: insights into the mechanism of action. *Nature Reviews. Microbiology*, *14*(2), 67–76.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* , *31*(2), 166–169.

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), 149–157.

Arnoult, N., Correia, A., Ma, J., Merlo, A., Garcia-Gomez, S., Maric, M., Tognetti, M., Benner, C. W., Boulton, S. J., Saghatelian, A., & Karlseder, J. (2017). Regulation of DNA repair pathway choice in

S and G2 phases by the NHEJ inhibitor CYREN. *Nature*, *549*(7673), 548–552.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, *2*, 2006.0008.

Bae, S., Kweon, J., Kim, H. S., & Kim, J.-S. (2014). Microhomology-based choice of Cas9 nuclease target sites. *Nature Methods*, *11*(7), 705–706.

Bae, S., Park, J., & Kim, J.-S. (2014). Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* , *30*(10), 1473–1475.

Banno, S., Nishida, K., Arazoe, T., Mitsunobu, H., & Kondo, A. (2018). Deaminase-mediated multiplex genome editing in Escherichia coli. *Nature Microbiology*, *3*(4), 423–429.

Bauer, C. R., Li, S., & Siegal, M. L. (2015). Essential gene disruptions reveal complex relationships between phenotypic robustness, pleiotropy, and fitness. *Molecular Systems Biology*, *11*(1), 773.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, *24*. https://proceedings.neurips.cc/paper/4443-algorithms-for-hyper-parameter-optimization

Bergstra, J., & Bengio, Y. (2012). *Random search for hyper-parameter optimization*. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf?source=post_page---------------------------

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (Vol. 28, pp. 115–123). PMLR.

Bétermier, M., Bertrand, P., & Lopez, B. S. (2014). Is non-homologous end-joining really an inherently error-prone process? *PLoS Genetics*, *10*(1), e1004086.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., & Marraffini, L. A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Research*, *41*(15), 7429–7437.

Billon, P., Bryant, E. E., Joseph, S. A., Nambiar, T. S., Hayward, S. B., Rothstein, R., & Ciccia, A. (2017). CRISPR-Mediated Base Editing Enables Efficient Disruption of Eukaryotic Genes through Induction of STOP Codons. *Molecular Cell*, *67*(6), 1068–1079.e4.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1484

Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C., & Brown, C. M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, *17*, 356.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, *8*, 209.

Bodapati, S., Daley, T. P., Lin, X., Zou, J., & Qi, L. S. (2020). A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biology*, *21*(1), 62.

Bosch, B., DeJesus, M. A., Poulton, N. C., Zhang, W., Engelhart, C. A., Zaveri, A., Lavalette, S., Ruecker, N., Trujillo, C., Wallach, J. B., Li, S., Ehrt, S., Chait, B. T., Schnappinger, D., & Rock, J. M. (2021). Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. *Cell*, *184*(17), 4579–4592.e24.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Braun, S. M. G., Kirkland, J. G., Chory, E. J., Husmann, D., Calarco, J. P., & Crabtree, G. R. (2017). Rapid and reversible epigenome editing by endogenous chromatin regulators. *Nature Communications*, *8*(1), 560.

Bravo, J. P. K., Liu, M.-S., Hibshman, G. N., Dangerfield, T. L., Jung, K., McCool, R. S., Johnson, K. A., & Taylor, D. W. (2022). Structural basis for mismatch surveillance by CRISPR–Cas9. *Nature*, *603*(7900), 343–347.

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V., & van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, *321*(5891), 960–964.

Brubaker, D. K., Proctor, E. A., Haigis, K. M., & Lauffenburger, D. A. (2019). Computational translation of genomic responses from experimental model systems to humans. *PLoS Computational Biology*, *15*(1), e1006286.

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PloS One*, *12*(10), e0185056.

Butland, G., Babu, M., Díaz-Mejía, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G., Pogoutse, O., Mori, H., Wanner, B. L., Lo, H., Wasniewski, J., Christopolous, C., Ali, M., Venn, P., Safavi-Naini, A., Sourour, N., … Emili, A. (2008). eSGA: E. coli synthetic genetic array analysis. *Nature Methods*, *5*(9), 789–795.

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. *Nature Reviews. Genetics*, *21*(9), 526–540.

Calvo-Villamañán, A., Ng, J. W., Planel, R., Ménager, H., Chen, A., Cui, L., & Bikard, D. (2020).

On-target activity predictions enable improved CRISPR-dCas9 screens in bacteria. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa294

Cameron, P., Fuller, C. K., Donohoue, P. D., Jones, B. N., Thompson, M. S., Carter, M. M., Gradia, S., Vidal, B., Garner, E., Slorach, E. M., Lau, E., Banh, L. M., Lied, A. M., Edwards, L. S., Settle, A. H., Capurso, D., Llaca, V., Deschamps, S., Cigan, M., … May, A. P. (2017). Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nature Methods*, *14*(6), 600–606.

Campa, C. C., Weisbach, N. R., Santinha, A. J., Incarnato, D., & Platt, R. J. (2019). Multiplexed genome engineering by Cas12a and CRISPR arrays encoded on single transcripts. *Nature Methods*, *16*(9), 887–893.

Carlson-Stevermer, J., Kelso, R., Kadina, A., Joshi, S., Rossi, N., Walker, J., Stoner, R., & Maures, T. (2020). CRISPRoff enables spatio-temporal control of CRISPR editing. In *Nature Communications* (Vol. 11, Issue 1). https://doi.org/10.1038/s41467-020-18853-3

Carroll, D. (2011). Genome engineering with zinc-finger nucleases. *Genetics*, *188*(4), 773–782.

Casadevall Arturo, & Pirofski Liise-anne. (2000). Host-Pathogen Interactions: Basic Concepts of Microbial Commensalism, Colonization, Infection, and Disease. *Infection and Immunity*, *68*(12), 6511–6518.

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., & Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews. Molecular Cell Biology*, *18*(8), 495–506.

Chao, M. C., Abel, S., Davis, B. M., & Waldor, M. K. (2016). The design and analysis of transposon insertion sequencing experiments. *Nature Reviews. Microbiology*, *14*(2), 119–128.

Chari, R., Mali, P., Moosburner, M., & Church, G. M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods*, *12*(9), 823–826.

Chari, R., Yeo, N. C., Chavez, A., & Church, G. M. (2017). sgRNA Scorer 2.0: A Species-Independent Model To Predict CRISPR/Cas9 Activity. *ACS Synthetic Biology*, *6*(5), 902–904.

Charpentier, E., & Marraffini, L. A. (2014). Harnessing CRISPR-Cas9 immunity for genetic engineering. *Current Opinion in Microbiology*, *19*, 114–119.

Chauhan, N. K., & Singh, K. (2018). A Review on Conventional Machine Learning vs Deep Learning. *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 347–352.

Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C. D., Wiegand, D. J., Ter-Ovanesyan, D., Braff, J. L., Davidsohn, N., Housden, B. E., Perrimon, N., Weiss, R., Aach, J., Collins, J. J., & Church, G. M. (2015). Highly efficient Cas9-mediated transcriptional programming. *Nature Methods*, *12*(4), 326–328.

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S., & Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, *155*(7), 1479–1491.

Chen, J. S., Ma, E., Harrington, L. B., Da Costa, M., Tian, X., Palefsky, J. M., & Doudna, J. A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science*, *360*(6387), 436–439.

Cho, S., Shin, J., & Cho, B.-K. (2018). Applications of CRISPR/Cas System to Bacterial Metabolic Engineering. *International Journal of Molecular Sciences*, *19*(4). https://doi.org/10.3390/ijms19041089

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S., & Kim, J.-S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Research*, *24*(1), 132–141.

Chow, R. D., Wang, G., Ye, L., Codina, A., Kim, H. R., Shen, L., Dong, M. B., Errami, Y., & Chen, S. (2019). In vivo profiling of metastatic double knockouts through CRISPR–Cpf1 screens. *Nature Methods*, *16*(5), 405–408.

Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B., Gu, F., Qu, S., Huang, D., Wei, J., & Liu, Q. (2018). DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology*, *19*(1), 80.

Collias, D., Leenay, R. T., Slotkowski, R. A., Zuo, Z., Collins, S. P., McGirr, B. A., Liu, J., & Beisel, C. L. (2020). A positive, growth-based PAM screen identifies noncanonical motifs recognized by the S. pyogenes Cas9. *Science Advances*, *6*(29), eabb4054.

Conway, T., Creecy, J. P., Maddox, S. M., Grissom, J. E., Conkle, T. L., Shadid, T. M., Teramoto, J., San Miguel, P., Shimada, T., Ishihama, A., Mori, H., & Wanner, B. L. (2014). Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *mBio*, *5*(4), e01442–14.

Corsi, G. I., Qu, K., Alkan, F., Pan, X., Luo, Y., & Gorodkin, J. (2022). CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context. *Nature Communications*, *13*(1), 3006.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, *46*(W1), W246–W251.

Cox, D. B. T., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J., Joung, J., & Zhang, F. (2017). RNA editing with CRISPR-Cas13. *Science*, *358*(6366), 1019–1027.

Cui, L., Vigouroux, A., Rousset, F., Varet, H., Khanna, V., & Bikard, D. (2018). A CRISPRi screen in E. coli reveals sequence-specific toxicity of dCas9. *Nature Communications*, *9*(1), 1912.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, *14*(3), 297–301.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602–607.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., & Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of Bacteriology*, *190*(4), 1390–1400.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15.

Dion, M. B., Plante, P.-L., Zufferey, E., Shah, S. A., Corbeil, J., & Moineau, S. (2021). Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Research*, *49*(6), 3127–3138.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853–1866.e17.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, *34*(2), 184–191.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., & Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature Biotechnology*, *32*(12), 1262–1267.

Doern, C. D. (2014). When does 2 plus 2 equal 5? A review of antimicrobial synergy testing. *Journal of Clinical Microbiology*, *52*(12), 4124–4128.

dos Reis, M., Wernisch, L., & Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Research*, *31*(23), 6976–6985.

Edgar, R. C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, *8*, 18.

Ellis, N. A., Kim, B., Tung, J., & Machner, M. P. (2021). A multiplex CRISPR interference tool for virulence gene interrogation in Legionella pneumophila. *Communications Biology*, *4*(1), 157.

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews. Genetics*, *20*(7), 389–403.

Feng, X., Tang, M., Dede, M., Su, D., Pei, G., Jiang, D., Wang, C., Chen, Z., Li, M., Nie, L., Xiong, Y., Li, S., Park, J.-M., Zhang, H., Huang, M., Szymonowicz, K., Zhao, Z., Hart, T., & Chen, J. (2022). Genome-wide CRISPR screens using isogenic cells reveal vulnerabilities conferred by loss of tumor suppressors. *Science Advances*, *8*(19), eabm6638.

Ferrand, J., Croft, N. P., Pépin, G., Diener, K. R., Wu, D., Mangan, N. E., Pedersen, J., Behlke, M. A., Hayball, J. D., Purcell, A. W., Ferrero, R. L., & Gantier, M. P. (2018). The Use of CRISPR/Cas9 Gene Editing to Confirm Congenic Contaminations in Host-Pathogen Interaction Studies. *Frontiers in Cellular and Infection Microbiology*, *8*, 87.

Ferreira, R., Skrekas, C., Nielsen, J., & David, F. (2018). Multiplexed CRISPR/Cas9 Genome Editing and Gene Regulation Using Csy4 in Saccharomyces cerevisiae. *ACS Synthetic Biology*, *7*(1), 10–15.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2020). Auto-Sklearn 2.0: The Next Generation. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2007.04074

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 2962–2970). Curran Associates, Inc.

Fontana, J., Dong, C., Ham, J. Y., Zalatan, J. G., & Carothers, J. M. (2018). Regulated Expression of sgRNAs Tunes CRISPRi in E. coli. *Biotechnology Journal*, *13*(9), e1800069.

Friedman, J. R., Fredericks, W. J., Jensen, D. E., Speicher, D. W., Huang, X. P., Neilson, E. G., & Rauscher, F. J., 3rd. (1996). KAP-1, a novel corepressor for the highly conserved KRAB repression domain. *Genes & Development*, *10*(16), 2067–2078.

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, *31*(9), 822–826.

Gagnon, J. A., Valen, E., Thyme, S. B., Huang, P., Akhmetova, L., Pauli, A., Montague, T. G., Zimmerman, S., Richter, C., & Schier, A. F. (2014). Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PloS One*, *9*(5), e98186.

Gasiunas, G., Barrangou, R., Horvath, P., & Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National*

Academy of Sciences, *109*(39), E2579–E2586.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, *551*(7681), 464–471.

Gehrke, J. M., Cervantes, O., Clement, M. K., Wu, Y., Zeng, J., Bauer, D. E., Pinello, L., & Joung, J. K. (2018). An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nature Biotechnology*, *36*(10), 977–982.

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M., & Weissman, J. S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, *159*(3), 647–661.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, *154*(2), 442–451.

Gootenberg, J. S., Abudayyeh, O. O., Lee, J. W., Essletzbichler, P., Dy, A. J., Joung, J., Verdine, V., Donghia, N., Daringer, N. M., Freije, C. A., Myhrvold, C., Bhattacharyya, R. P., Livny, J., Regev, A., Koonin, E. V., Hung, D. T., Sabeti, P. C., Collins, J. J., & Zhang, F. (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science*, *356*(6336), 438–442.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews. Molecular Cell Biology*, *23*(1), 40–55.

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(Web Server issue), W52–W57.

Gu, T., Zhao, S., Pi, Y., Chen, W., Chen, C., Liu, Q., Li, M., Han, D., & Ji, Q. (2018). Highly efficient base editing in using an engineered CRISPR RNA-guided cytidine deaminase. *Chemical Science* , *9*(12), 3248–3253.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J.-S., & Concordet, J.-P. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*, *17*(1), 148.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A., & Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nature*

Biotechnology, *35*(5), 463–474.

Hanna, R. E., & Doench, J. G. (2020). Design and analysis of CRISPR–Cas experiments. *Nature Biotechnology*, *38*(7), 813–823.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., & Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, *163*(6), 1515–1526.

Hawkins, J. S., Silvis, M. R., Koo, B.-M., Peters, J. M., Osadnik, H., Jost, M., Hearne, C. C., Weissman, J. S., Todor, H., & Gross, C. A. (2020). Mismatch-CRISPRi Reveals the Co-varying Expression-Fitness Relationships of Essential Genes in Escherichia coli and Bacillus subtilis. *Cell Systems*, *11*(5), 523–535.e9.

Heler, R., Marraffini, L. A., & Bikard, D. (2014). Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. *Molecular Microbiology*, *93*(1), 1–9.

He, W., Zhang, L., Villarreal, O. D., Fu, R., Bedford, E., Dou, J., Patel, A. Y., Bedford, M. T., Shi, X., Chen, T., Bartholomew, B., & Xu, H. (2019). De novo identification of essential protein domains from CRISPR-Cas9 tiling-sgRNA knockout screens. *Nature Communications*, *10*(1), 4541.

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, *33*(5), 510–517.

Höck, J., & Meister, G. (2008). The Argonaute protein family. *Genome Biology*, *9*(2), 210.

Horvath, P., Romero, D. A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., & Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *Journal of Bacteriology*, *190*(4), 1401–1412.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, *31*(9), 827–832.

Hu, J. H., Miller, S. M., Geurts, M. H., Tang, W., Chen, L., Sun, N., Zeina, C. M., Gao, X., Rees, H. A., Lin, Z., & Liu, D. R. (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*, *556*(7699), 57–63.

Hwang, G.-H., Park, J., Lim, K., Kim, S., Yu, J., Yu, E., Kim, S.-T., Eils, R., Kim, J.-S., & Bae, S. (2018). Web-based design and analysis tools for CRISPR base editing. *BMC Bioinformatics*, *19*(1), 542.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1502.03167

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., & Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of Bacteriology*, *169*(12), 5429–5433.

Jacquin, A. L. S., Odom, D. T., & Lukk, M. (2019). Crisflash: open-source software to generate CRISPR guide RNAs against genomes annotated with individual variation. *Bioinformatics* , *35*(17), 3146–3147.

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562.

Jayavaradhan, R., Pillis, D. M., Goodman, M., Zhang, F., Zhang, Y., Andreassen, P. R., & Malik, P. (2019). CRISPR-Cas9 fusion to dominant-negative 53BP1 enhances HDR and inhibits NHEJ specifically at Cas9 target sites. *Nature Communications*, *10*(1), 2866.

Jiang, F., & Doudna, J. A. (2017). CRISPR-Cas9 Structures and Mechanisms. *Annual Review of Biophysics*, *46*, 505–529.

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E., & Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, *351*(6275), 867–871.

Jiang, F., Zhou, K., Ma, L., Gressel, S., & Doudna, J. A. (2015). A Cas9–guide RNA complex preorganized for target DNA recognition. *Science*, *348*(6242), 1477–1481.

Jiang, H., & Wong, W. H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* , *24*(20), 2395–2396.

Jiao, C., Sharma, S., Dugar, G., Peeck, N. L., Bischler, T., Wimmer, F., Yu, Y., Barquist, L., Schoen, C., Kurzai, O., Sharma, C. M., & Beisel, C. L. (2021). Noncanonical crRNAs derived from host transcripts enable multiplexable RNA detection by Cas9. *Science*, *372*(6545), 941–948.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821.

Jusiak, B., Cleto, S., Perez-Piñera, P., & Lu, T. K. (2016). Engineering Synthetic Gene Circuits in Living Cells with CRISPR Technology. *Trends in Biotechnology*, *34*(7), 535–547.

Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F., & Collins, J. J. (2021). CRISPR-based diagnostics. *Nature Biomedical Engineering*, *5*(7), 643–656.

Kantor, A., McClements, M. E., & MacLaren, R. E. (2020). CRISPR-Cas9 DNA Base-Editing and Prime-Editing. *International Journal of Molecular Sciences*, *21*(17). https://doi.org/10.3390/ijms21176240

Karvelis, T., Gasiunas, G., Young, J., Bigelyte, G., Silanskas, A., Cigan, M., & Siksnys, V. (2015). Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biology*, *16*, 253.

Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M., & Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nature Methods*, *12*(5), 401–403.

Kellner, M. J., Koob, J. G., Gootenberg, J. S., Abudayyeh, O. O., & Zhang, F. (2019). SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nature Protocols*, *14*(10), 2986–3012.

Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., Weinberger, A., Alon, U., Milo, R., & Segal, E. (2016). Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell*, *166*(5), 1282–1294.e18.

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., … Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, *45*(D1), D543–D550.

Khalil, A. S., & Collins, J. J. (2010). Synthetic biology: applications come of age. *Nature Reviews. Genetics*, *11*(5), 367–379.

Kiga, K., Tan, X.-E., Ibarra-Chávez, R., Watanabe, S., Aiba, Y., Sato'o, Y., Li, F.-Y., Sasahara, T., Cui, B., Kawauchi, M., Boonsiri, T., Thitiananpakorn, K., Taki, Y., Azam, A. H., Suzuki, M., Penadés, J. R., & Cui, L. (2020). Development of CRISPR-Cas13a-based antimicrobials capable of sequence-specific killing of target bacteria. *Nature Communications*, *11*(1), 2934.

Kim, H. K., Kim, Y., Lee, S., Min, S., Bae, J. Y., Choi, J. W., Park, J., Jung, D., Yoon, S., & Kim, H. H. (2019). SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Science Advances*, *5*(11), eaax9249.

Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S., & Kim, H. H. (2018). Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nature Biotechnology*, *36*(3), 239–241.

Kim, Y. B., Komor, A. C., Levy, J. M., Packer, M. S., Zhao, K. T., & Liu, D. R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nature Biotechnology*, *35*(4), 371–376.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., & Keith Joung, J.

(2016). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. In *Nature* (Vol. 529, Issue 7587, pp. 490–495). https://doi.org/10.1038/nature16526

Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S., & Sternberg, S. H. (2019). Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature*, *571*(7764), 219–225.

Koblan, L. W., Arbab, M., Shen, M. W., Hussmann, J. A., Anzalone, A. V., Doman, J. L., Newby, G. A., Yang, D., Mok, B., Replogle, J. M., Xu, A., Sisley, T. A., Weissman, J. S., Adamson, B., & Liu, D. R. (2021). Efficient C•G-to-G•C base editors developed using CRISPRi screens, target-library analysis, and machine learning. *Nature Biotechnology*, *39*(11), 1414–1425.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C., & Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, *32*(3), 267–273.

Komor, A. C., Badran, A. H., & Liu, D. R. (2017). CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell*, *168*(1-2), 20–36.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), 420–424.

Komor, A. C., Zhao, K. T., Packer, M. S., Gaudelli, N. M., Waterbury, A. L., Koblan, L. W., Kim, Y. B., Badran, A. H., & Liu, D. R. (2017). Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Science Advances*, *3*(8), eaao4774.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., & Zhang, F. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588.

Konstantakos, V., Nentidis, A., Krithara, A., & Paliouras, G. (2022). CRISPR-Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research*, *50*(7), 3616–3637.

Koonin, E. V., Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology*, *37*, 67–78.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, *1*(2), 111–117.

Krohannon, A., Srivastava, M., Rauch, S., Srivastava, R., Dickinson, B. C., & Janga, S. C. (2022). CASowary: CRISPR-Cas13 guide RNA predictor for transcript depletion. *BMC Genomics*, *23*(1), 172.

Kuan, P. F., Powers, S., He, S., Li, K., Zhao, X., & Huang, B. (2017). A systematic evaluation of nucleotide properties for CRISPR sgRNA design. *BMC Bioinformatics*, *18*(1), 297.

Kuscu, C., Parlak, M., Tufan, T., Yang, J., Szlachta, K., Wei, X., Mammadov, R., & Adli, M. (2017). CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nature Methods*, *14*(7), 710–712.

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Mattiazzi Usaj, M., van Leeuwen, J., Koch, E. N., Pons, C., Dagilis, A. J., Pryszlak, M., Wang, J. Z. Y., Hanchard, J., Riggi, M., Xu, K., Heydari, H., … Myers, C. L. (2018). Systematic analysis of complex genetic interactions. *Science*, *360*(6386). https://doi.org/10.1126/science.aao1729

Kwon, D. Y., Zhao, Y.-T., Lamonica, J. M., & Zhou, Z. (2017). Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC. *Nature Communications*, *8*, 15315.

Labuhn, M., Adams, F. F., Ng, M., Knoess, S., Schambach, A., Charpentier, E. M., Schwarzer, A., Mateo, J. L., Klusmann, J.-H., & Heckl, D. (2018). Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Research*, *46*(3), 1375–1385.

Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., & Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, *19*(12), 2308–2316.

Ledell, E., & Poirier, S. (2020). H2O AutoML: Scalable Automatic Machine Learning. *H2O AutoML: Scalable Automatic Machine Learning*, 61.

Lee, H. H., Ostrov, N., Wong, B. G., Gold, M. A., Khalil, A. S., & Church, G. M. (2019). Functional genomics of the rapidly replicating bacterium Vibrio natriegens by CRISPRi. *Nature Microbiology*, *4*(7), 1105–1113.

Lee, J. K., Jeong, E., Lee, J., Jung, M., Shin, E., Kim, Y.-H., Lee, K., Jung, I., Kim, D., Kim, S., & Kim, J.-S. (2018). Directed evolution of CRISPR-Cas9 to increase its specificity. *Nature Communications*, *9*(1), 3048.

Leenay, R. T., Maksimchuk, K. R., Slotkowski, R. A., Agrawal, R. N., Gomaa, A. A., Briner, A. E., Barrangou, R., & Beisel, C. L. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell*, *62*(1), 137–147.

Leshchiner, D., Rosconi, F., Sundaresh, B., Rudmann, E., Ramirez, L. M. N., Nishimoto, A. T., Wood, S. J., Jana, B., Buján, N., Li, K., Gao, J., Frank, M., Reeve, S. M., Lee, R. E., Rock, C. O., Rosch, J. W., & van Opijnen, T. (2022). A genome-wide atlas of antibiotic susceptibility targets and pathways to tolerance. *Nature Communications*, *13*(1), 3165.

Liang, P., Xie, X., Zhi, S., Sun, H., Zhang, X., Chen, Y., Chen, Y., Xiong, Y., Ma, W., Liu, D., Huang, J., & Songyang, Z. (2019). Genome-wide profiling of adenine base editor specificity by EndoV-seq.

Nature Communications, *10*(1), 67.

Lian, J., HamediRad, M., Hu, S., & Zhao, H. (2017). Combinatorial metabolic engineering using an orthogonal tri-functional CRISPR system. *Nature Communications*, *8*(1), 1688.

Liao, C., Slotkowski, R. A., & Beisel, C. L. (2019). CRATES: A one-step assembly method for Class 2 CRISPR arrays. *Methods in Enzymology*, *629*, 493–511.

Liao, C., Ttofali, F., Slotkowski, R. A., Denny, S. R., Cecil, T. D., Leenay, R. T., Keung, A. J., & Beisel, C. L. (2019). Modular one-pot assembly of CRISPR arrays enables library generation and reveals factors influencing crRNA biogenesis. *Nature Communications*, *10*(1), 2948.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A Research Platform for Distributed Model Selection and Training. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1807.05118

Li, J., Wang, Y., Wang, B., Lou, J., Ni, P., Jin, Y., Chen, S., Duan, G., & Zhang, R. (2022). Application of CRISPR/Cas Systems in the Nucleic Acid Detection of Infectious Diseases. *Diagnostics (Basel, Switzerland)*, *12*(10). https://doi.org/10.3390/diagnostics12102455

Lin, S., Staahl, B. T., Alla, R. K., & Doudna, J. A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife*, *3*, e04766.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queueing Systems. Theory and Applications*, *16*(3), 31–57.

Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., Gao, K., Hoang, L., Elibol, M., Doench, J. G., & Fusi, N. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering*, *2*(1), 38–47.

Liu, G., Zhang, Y., & Zhang, T. (2020). Computational approaches for effective CRISPR guide RNA design and evaluation. *Computational and Structural Biotechnology Journal*, *18*, 35–44.

Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., Wang, M., Zhang, X., & Wang, Y. (2017). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell*, *170*(4), 714–726.e10.

Liu, X. S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R. A., & Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. *Cell*, *167*(1), 233–247.e17.

Liu, Y., Mao, S., Huang, S., Li, Y., Chen, Y., Di, M., Huang, X., Lv, J., Wang, X., Ge, J., Shen, S., Zhang, X., Liu, D., Huang, X., & Chi, T. (2020). REPAIRx, a specific yet highly efficient programmable A > I RNA base editor. *The EMBO Journal*, *39*(22), e104748.

Liu, Y., Wang, R., Liu, J., Lu, H., Li, H., Wang, Y., Ni, X., Li, J., Guo, Y., Ma, H., Liao, X., & Wang, M. (2022). Base editor enables rational genome-scale functional screening for enhanced industrial phenotypes in *Corynebacterium glutamicum*. *Science Advances*, *8*(35), eabq2157.

Liu, Z., Dong, H., Cui, Y., Cong, L., & Zhang, D. (2020). Application of different types of CRISPR/Cas-based systems in bacteria. *Microbial Cell Factories*, *19*(1), 172.

Li, W., Köster, J., Xu, H., Chen, C.-H., Xiao, T., Liu, J. S., Brown, M., & Liu, X. S. (2015). Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biology*, *16*, 281.

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology*, *15*(12), 554.

Li, X., Wang, Y., Liu, Y., Yang, B., Wang, X., Wei, J., Lu, Z., Zhang, Y., Wu, J., Huang, X., Yang, L., & Chen, J. (2018). Base editing with a Cpf1-cytidine deaminase fusion. *Nature Biotechnology*, *36*(4), 324–327.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology: AMB*, *6*, 26.

Lorenz, R., Hofacker, I. L., & Bernhart, S. H. (2012). Folding RNA/DNA hybrid duplexes. *Bioinformatics* , *28*(19), 2530–2531.

Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1711.05101

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., Rybakov, S., Misharin, A. V., & Theis, F. J. (2021). Mapping single-cell data to reference atlases by transfer learning. In *Nature Biotechnology*. https://doi.org/10.1038/s41587-021-01001-7

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

Luo, J., Chen, W., Xue, L., & Tang, B. (2019). Prediction of activity and specificity of CRISPR-Cpf1 using convolutional deep learning neural networks. *BMC Bioinformatics*, *20*(1), 332.

Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., & Joung, J. K. (2013). CRISPR RNA–guided activation of endogenous human genes. *Nature Methods*, *10*(10), 977–979.

Ma, H., Naseri, A., Reyes-Gutierrez, P., Wolfe, S. A., Zhang, S., & Pederson, T. (2015). Multicolor CRISPR labeling of chromosomal loci in human cells. *Proceedings of the National Academy of*

Sciences of the United States of America, *112*(10), 3002–3007.

Ma, H., Tu, L.-C., Naseri, A., Chung, Y.-C., Grunwald, D., Zhang, S., & Pederson, T. (2018). CRISPR-Sirius: RNA scaffolds for signal amplification in genome imaging. *Nature Methods*, *15*(11), 928–931.

Ma, H., Tu, L.-C., Naseri, A., Huisman, M., Zhang, S., Grunwald, D., & Pederson, T. (2016a). Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nature Biotechnology*, *34*(5), 528–530.

Ma, H., Tu, L.-C., Naseri, A., Huisman, M., Zhang, S., Grunwald, D., & Pederson, T. (2016b). CRISPR-Cas9 nuclear dynamics and target recognition in living cells. *The Journal of Cell Biology*, *214*(5), 529–537.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Scott, D., Shah, S. A., Siksnys, V., Terns, M. P., Venclovas, Č., White, M. F., Yakunin, A. F., … Koonin, E. V. (2019). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews. Microbiology*, *18*(2), 67–83.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826.

Marino, N. D., Pinilla-Redondo, R., Csörgő, B., & Bondy-Denomy, J. (2020). Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. *Nature Methods*. https://doi.org/10.1038/s41592-020-0771-6

Marraffini, L. A., & Sontheimer, E. J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*, *463*(7280), 568–571.

Marshall, R., Maxwell, C. S., Collins, S. P., Jacobsen, T., Luo, M. L., Begemann, M. B., Gray, B. N., January, E., Singer, A., He, Y., Beisel, C. L., & Noireaux, V. (2018). Rapid and Scalable Characterization of CRISPR Technologies Using an E. coli Cell-Free Transcription-Translation System. *Molecular Cell*, *69*(1), 146–157.e3.

McCarty, N. S., Graham, A. E., Studená, L., & Ledesma-Amaro, R. (2020). Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nature Communications*, *11*(1), 1281.

McKenna, A., & Shendure, J. (2018). FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biology*, *16*(1), 74.

McNeil, M. B., Keighley, L. M., Cook, J. R., Cheung, C.-Y., & Cook, G. M. (2021). CRISPR interference identifies vulnerable cellular pathways with bactericidal phenotypes in Mycobacterium tuberculosis. *Molecular Microbiology*, *116*(4), 1033–1043.

Meeske, A. J., Jia, N., Cassel, A. K., Kozlova, A., Liao, J., Wiedmann, M., Patel, D. J., & Marraffini, L.

A. (2020). A phage-encoded anti-CRISPR enables complete evasion of type VI-A CRISPR-Cas immunity. *Science*, *369*(6499), 54–59.

Meeske, A. J., & Marraffini, L. A. (2018). RNA Guide Complementarity Prevents Self-Targeting in Type VI CRISPR Systems. *Molecular Cell*, *71*(5), 791–801.e3.

Meeske, A. J., Nakandakari-Higa, S., & Marraffini, L. A. (2019). Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature*, *570*(7760), 241–245.

Meliawati, M., Schilling, C., & Schmid, J. (2021). Recent advances of Cas12a applications in bacteria. *Applied Microbiology and Biotechnology*, *105*(8), 2981–2990.

Metsky, H. C., Welch, N. L., Pillai, P. P., Haradhvala, N. J., Rumker, L., Mantena, S., Zhang, Y. B., Yang, D. K., Ackerman, C. M., Weller, J., Blainey, P. C., Myhrvold, C., Mitzenmacher, M., & Sabeti, P. C. (2022). Designing sensitive viral diagnostics with machine learning. *Nature Biotechnology*, *40*(7), 1123–1131.

Michlits, G., Jude, J., Hinterndorfer, M., de Almeida, M., Vainorius, G., Hubmann, M., Neumann, T., Schleiffer, A., Burkard, T. R., Fellner, M., Gijsbertsen, M., Traunbauer, A., Zuber, J., & Elling, U. (2020). Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. *Nature Methods*. https://doi.org/10.1038/s41592-020-0850-8

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable . christophm. github. io/interpretable-ml-book.*

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 39–68). Springer International Publishing.

Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., & Valen, E. (2014). CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Research*, *42*(Web Server issue), W401–W407.

Moreno-Mateos, M. A., Vejnar, C. E., Beaudoin, J.-D., Fernandez, J. P., Mis, E. K., Khokha, M. K., & Giraldez, A. J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods*, *12*(10), 982–988.

Mougiakos, I., Bosma, E. F., Ganguly, J., van der Oost, J., & van Kranenburg, R. (2018). Hijacking CRISPR-Cas for high-throughput bacterial metabolic engineering: advances and prospects. *Current Opinion in Biotechnology*, *50*, 146–157.

Muhammad Rafid, A. H., Toufikuzzaman, M., Rahman, M. S., & Rahman, M. S. (2020).

CRISPRpred(SEQ): a sequence-based method for sgRNA on target activity prediction using traditional machine learning. *BMC Bioinformatics*, *21*(1), 223.

Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, *399*(10325), 629–655.

Najm, F. J., Strand, C., Donovan, K. F., Hegde, M., Sanson, K. R., Vaimberg, E. W., Sullender, M. E., Hartenian, E., Kalani, Z., Fusi, N., Listgarten, J., Younger, S. T., Bernstein, B. E., Root, D. E., & Doench, J. G. (2018). Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. *Nature Biotechnology*, *36*(2), 179–189.

Nambiar, T. S., Billon, P., Diedenhofen, G., Hayward, S. B., Taglialatela, A., Cai, K., Huang, J.-W., Leuzzi, G., Cuella-Martin, R., Palacios, A., Gupta, A., Egli, D., & Ciccia, A. (2019). Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. *Nature Communications*, *10*(1), 3395.

Niaz, S. (2018). The AGO proteins: an overview. *Biological Chemistry*, *399*(6), 525–547.

Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O. O., Gootenberg, J. S., Mori, H., Oura, S., Holmes, B., Tanaka, M., Seki, M., Hirano, H., Aburatani, H., Ishitani, R., Ikawa, M., Yachie, N., … Nureki, O. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science*, *361*(6408), 1259–1262.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567.

Olson, R. S., & Moore, J. H. (2016). TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Proceedings of the Workshop on Automatic Machine Learning* (Vol. 64, pp. 66–74). PMLR.

Paez-Espino, D., Morovic, W., Sun, C. L., Thomas, B. C., Ueda, K.-I., Stahl, B., Barrangou, R., & Banfield, J. F. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature Communications*, *4*, 1430.

Pan, X., Qu, K., Yuan, H., Xiang, X., Anthon, C., Pashkova, L., Liang, X., Han, P., Corsi, G. I., Xu, F., Liu, P., Zhong, J., Zhou, Y., Ma, T., Jiang, H., Liu, J., Wang, J., Jessen, N., Bolund, L., … Luo, Y. (2022). Massively targeted evaluation of therapeutic CRISPR off-targets in cells. *Nature Communications*, *13*(1), 4049.

Pardee, K., Green, A. A., Takahashi, M. K., Braff, D., Lambert, G., Lee, J. W., Ferrante, T., Ma, D., Donghia, N., Fan, M., Daringer, N. M., Bosch, I., Dudley, D. M., O'Connor, D. H., Gehrke, L., & Collins, J. J. (2016). Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular

Components. *Cell*, *165*(5), 1255–1266.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1912.01703

Pattanayak, V., Lin, S., Guilinger, J. P., Ma, E., Doudna, J. A., & Liu, D. R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature Biotechnology*, *31*(9), 839–843.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830.

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nature Methods*, *10*(10), 973–976.

Peters, J. M., Koo, B.-M., Patino, R., Heussler, G. E., Hearne, C. C., Qu, J., Inclan, Y. F., Hawkins, J. S., Lu, C. H. S., Silvis, M. R., Harden, M. M., Osadnik, H., Peters, J. E., Engel, J. N., Dutton, R. J., Grossman, A. D., Gross, C. A., & Rosenberg, O. S. (2019). Enabling genetic analysis of diverse bacteria with Mobile-CRISPRi. *Nature Microbiology*, *4*(2), 244–250.

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.-P., Couvin, D., Toffano-Nioche, C., & Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Research*, *48*(D1), D535–D544.

Puigbò, P., Bravo, I. G., & Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct*, *3*, 38.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152*(5), 1173–1183.

Radzisheuskaya, A., Shlyueva, D., Müller, I., & Helin, K. (2016). Optimizing sgRNA position markedly improves the efficiency of CRISPR/dCas9-mediated transcriptional repression. *Nucleic Acids Research*, *44*(18), e141.

Rahman, M. K., & Rahman, M. S. (2017). CRISPRpred: A flexible and efficient tool for sgRNAs on-target activity prediction in CRISPR/Cas9 systems. *PloS One*, *12*(8), e0181943.

Raleigh, E. A., & Brooks, J. E. (1998). Restriction Modification Systems: Where They Are and What They Do. In F. J. de Bruijn, J. R. Lupski, & G. M. Weinstock (Eds.), *Bacterial Genomes: Physical Structure and Analysis* (pp. 78–92). Springer US.

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y., & Zhang, F. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, *154*(6), 1380–1389.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308.

Raschka, S. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1811.12808

Rauscher, B., Heigwer, F., Henkel, L., Hielscher, T., Voloshanenko, O., & Boutros, M. (2018). Toward an integrated map of genetic interactions in cancer cells. *Molecular Systems Biology*, *14*(2), e7656.

Ray, U., Vartak, S. V., & Raghavan, S. C. (2020). NHEJ inhibitor SCR7 and its different forms: Promising CRISPR tools for genome engineering. *Gene*, *763*, 144997.

Rees, H. A., Yeh, W.-H., & Liu, D. R. (2019). Development of hRad51–Cas9 nickase fusions that mediate HDR without double-stranded breaks. *Nature Communications*, *10*(1), 1–12.

Reis, A. C., Halper, S. M., Vezeau, G. E., Cetnar, D. P., Hossain, A., Clauer, P. R., & Salis, H. M. (2019). Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0286-9

Richardson, C. D., Ray, G. J., Bray, N. L., & Corn, J. E. (2016). Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes. *Nature Communications*, *7*, 12463.

Riesenberg, S., Helmbrecht, N., Kanis, P., Maricic, T., & Pääbo, S. (2022). Improved gRNA secondary structures allow editing of target sites resistant to CRISPR-Cas9 cleavage. *Nature Communications*, *13*(1), 489.

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, *32*(9), 896–902.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Rock, J. M., Hopkins, F. F., Chavez, A., Diallo, M., Chase, M. R., Gerrick, E. R., Pritchard, J. R., Church, G. M., Rubin, E. J., Sassetti, C. M., Schnappinger, D., & Fortune, S. M. (2017). Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. *Nature Microbiology*, *2*, 16274.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575.

Rottinghaus, A. G., Vo, S., & Moon, T. S. (2023). Computational design of CRISPR guide RNAs to enable strain-specific control of microbial consortia. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(1), e2213154120.

Rousset, F., Cui, L., Siouve, E., Becavin, C., Depardieu, F., & Bikard, D. (2018). Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLoS Genetics*, *14*(11), e1007749.

Rubin, B. E., Diamond, S., Cress, B. F., Crits-Christoph, A., Lou, Y. C., Borges, A. L., Shivram, H., He, C., Xu, M., Zhou, Z., Smith, S. J., Rovinsky, R., Smock, D. C. J., Tang, K., Owens, T. K., Krishnappa, N., Sachdeva, R., Barrangou, R., Deutschbauer, A. M., … Doudna, J. A. (2022). Species- and site-specific genome editing in complex bacterial communities. *Nature Microbiology*, *7*(1), 34–47.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Salis, H. M. (2011). The ribosome binding site calculator. *Methods in Enzymology*, *498*, 19–42.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Sanjana, N. E., Cong, L., Zhou, Y., Cunniff, M. M., Feng, G., & Zhang, F. (2012). A transcription activator-like effector toolbox for genome engineering. *Nature Protocols*, *7*(1), 171–192.

Santajit, S., & Indrawattana, N. (2016). Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens. *BioMed Research International*, *2016*, 2475067.

Santos-Moreno, J., Tasiudi, E., Stelling, J., & Schaerli, Y. (2020). Multistable and dynamic CRISPRi-based synthetic circuits. *Nature Communications*, *11*(1), 2746.

Scholefield, J., & Harrison, P. T. (2021). Prime editing – an update on the field. *Gene Therapy*, *28*(7), 396–401.

Schuster, A., Erasimus, H., Fritah, S., Nazarov, P. V., van Dyck, E., Niclou, S. P., & Golebiewska, A. (2019). RNAi/CRISPR Screens: from a Pool to a Valid Hit. *Trends in Biotechnology*, *37*(1), 38–55.

Semenova, E., Jore, M. M., Datsenko, K. A., Semenova, A., Westra, E. R., Wanner, B., van der Oost, J., Brouns, S. J. J., & Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10098–10103.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, *104*(1), 148–175.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., & Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, *343*(6166), 84–87.

Shapiro, R. S., Chavez, A., Porter, C. B. M., Hamblin, M., Kaas, C. S., DiCarlo, J. E., Zeng, G., Xu, X., Revtovich, A. V., Kirienko, N. V., Wang, Y., Church, G. M., & Collins, J. J. (2017). A CRISPR–Cas9-based gene drive platform for genetic interaction analysis in Candida albicans. *Nature Microbiology*, *3*(1), 73–82.

Shen, B., Zhang, W., Zhang, J., Zhou, J., Wang, J., Chen, L., Wang, L., Hodgkins, A., Iyer, V., Huang, X., & Skarnes, W. C. (2014). Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nature Methods*, *11*(4), 399–402.

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C.-C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., … Mali, P. (2017). Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, *14*(6), 573–576.

Shin, J., Jiang, F., Liu, J.-J., Bray, N. L., Rauch, B. J., Baik, S. H., Nogales, E., Bondy-Denomy, J., Corn, J. E., & Doudna, J. A. (2017). Disabling Cas9 by an anti-CRISPR DNA mimic. *Science Advances*, *3*(7), e1701620.

Sidik, S. M., Huet, D., Ganesan, S. M., Huynh, M.-H., Wang, T., Nasamu, A. S., Thiru, P., Saeij, J. P. J., Carruthers, V. B., Niles, J. C., & Lourido, S. (2016). A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. *Cell*, *166*(6), 1423–1435.e12.

Simidjievski, N., Bodnar, C., Tariq, I., & Scherer, P. (2019). Variational autoencoders for cancer data integration: design principles and computational practice. *Frontiers in*. https://www.frontiersin.org/articles/10.3389/fgene.2019.01205/full

Smith, J. D., Suresh, S., Schlecht, U., Wu, M., Wagih, O., Peltz, G., Davis, R. W., Steinmetz, L. M., Parts, L., & St Onge, R. P. (2016). Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. *Genome Biology*, *17*, 45.

Song, M., Kim, H. K., Lee, S., Kim, Y., Seo, S.-Y., Park, J., Choi, J. W., Jang, H., Shin, J. H., Min, S., Quan, Z., Kim, J. H., Kang, H. C., Yoon, S., & Kim, H. H. (2020). Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nature Biotechnology*, *38*(9), 1037–1043.

Spoto, M., Riera Puma, J. P., Fleming, E., Guan, C., Ondouah Nzutchi, Y., Kim, D., & Oh, J. (2022). Large-Scale CRISPRi and Transcriptomics of Staphylococcus epidermidis Identify Genetic Factors Implicated in Lifestyle Versatility. *mBio*, e0263222.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research: JMLR*, *15*(1), 1929–1958.

Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J., & Mateo, J. L. (2015). CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PloS One*, *10*(4), e0124633.

Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in Genetics: TIG*, *26*(8), 335–340.

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., & Doudna, J. A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, *507*(7490), 62–67.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J. L., Makarova, K. S., Koonin, E. V., & Zhang, F. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. *Science*. https://doi.org/10.1126/science.aax9181

Sundaresh, B., Xu, S., Noonan, B., Mansour, M. K., Leong, J. M., & van Opijnen, T. (2021). Host-informed therapies for the treatment of pneumococcal pneumonia. *Trends in Molecular Medicine*, *27*(10), 971–989.

Su, X., Yan, X., & Tsai, C.-L. (2012). Linear regression. *Wiley Interdisciplinary Reviews. Computational Statistics*, *4*(3), 275–294.

Tambe, A., East-Seletsky, A., Knott, G. J., Doudna, J. A., & O'Connell, M. R. (2018). RNA Binding and HEPN-Nuclease Activation Are Decoupled in CRISPR-Cas13a. *Cell Reports*, *24*(4), 1025–1036.

Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S., & Vale, R. D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell*, *159*(3), 635–646.

Tan, J., Zhang, F., Karcher, D., & Bock, R. (2019). Engineering of high-precision base editors for site-specific single nucleotide replacement. *Nature Communications*, *10*(1), 439.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 847–855.

Tierrafría, V. H., Rioualen, C., Salgado, H., Lara, P., Gama-Castro, S., Lally, P., Gómez-Romero, L., Peña-Loredo, P., López-Almazo, A. G., Alarcón-Carranza, G., Betancourt-Figueroa, F., Alquicira-Hernández, S., Polanco-Morelos, J. E., García-Sotelo, J., Gaytan-Nuñez, E., Méndez-Cruz, C.-F., Muñiz, L. J., Bonavides-Martínez, C., Moreno-Hagelsieb, G., … Collado-Vides, J. (2022). RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microbial Genomics*, *8*(5). https://doi.org/10.1099/mgen.0.000833

Todor, H., Silvis, M. R., Osadnik, H., & Gross, C. A. (2021). Bacterial CRISPR screens for gene function. *Current Opinion in Microbiology*, *59*, 102–109.

Tong, H., Huang, J., Xiao, Q., He, B., Dong, X., Liu, Y., Yang, X., Han, D., Wang, Z., Wang, X., Ying, W., Zhang, R., Wei, Y., Xu, C., Zhou, Y., Li, Y., Cai, M., Wang, Q., Xue, M., … Yang, H. (2022). High-fidelity Cas13 variants for targeted RNA degradation with minimal collateral effects. *Nature*

Biotechnology. https://doi.org/10.1038/s41587-022-01419-7

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., Aryee, M. J., & Joung, J. K. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, *33*(2), 187–197.

Tu, M., Lin, L., Cheng, Y., He, X., Sun, H., Xie, H., Fu, J., Liu, C., Li, J., Chen, D., Xi, H., Xue, D., Liu, Q., Zhao, J., Gao, C., Song, Z., Qu, J., & Gu, F. (2017). A "new lease of life": FnCpf1 possesses DNA cleavage activity for genome editing in human cells. *Nucleic Acids Research*, *45*(19), 11295–11304.

Typas, A., Nichols, R. J., Siegele, D. A., Shales, M., Collins, S. R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B. L., Mori, H., Weissman, J. S., Krogan, N. J., & Gross, C. A. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. *Nature Methods*, *5*(9), 781–787.

Unterholzner, S. J., Poppenberger, B., & Rozhon, W. (2013). Toxin-antitoxin systems: Biology, identification, and application. *Mobile Genetic Elements*, *3*(5), e26219.

Uribe, R. V., Rathmer, C., Jahn, L. J., Ellabaan, M. M. H., Li, S. S., & Sommer, M. O. A. (2021). Bacterial resistance to CRISPR-Cas antimicrobials. *Scientific Reports*, *11*(1), 17267.

Vakulskas, C. A., Dever, D. P., Rettig, G. R., Turk, R., Jacobi, A. M., Collingwood, M. A., Bode, N. M., McNeill, M. S., Yan, S., Camarena, J., Lee, C. M., Park, S. H., Wiebking, V., Bak, R. O., Gomez-Ospina, N., Pavel-Dinu, M., Sun, W., Bao, G., Porteus, M. H., & Behlke, M. A. (2018). A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature Medicine*, *24*(8), 1216–1224.

Van Melderen, L., & De Bast, M. S. (2009). Bacterial Toxin–Antitoxin Systems: More Than Selfish Entities? *PLoS Genetics*, *5*(3), e1000437.

van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, *6*(10), 767–772.

Vanschoren, J. (2018). Meta-Learning: A Survey. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1810.03548

Vento, J. M., Crook, N., & Beisel, C. L. (2019). Barriers to genome editing with CRISPR in bacteria. *Journal of Industrial Microbiology & Biotechnology*, *46*(9-10), 1327–1341.

Vialetto, E., Yu, Y., Collins, S. P., Wandera, K. G., Barquist, L., & Beisel, C. L. (2022). A target expression threshold dictates invader defense and prevents autoimmunity by CRISPR-Cas13. *Cell Host & Microbe*. https://doi.org/10.1016/j.chom.2022.05.013

Vigouroux, A., & Bikard, D. (2020). CRISPR Tools To Control Gene Expression in Bacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *84*(2). https://doi.org/10.1128/MMBR.00077-19

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Vo, P. L. H., Ronda, C., Klompe, S. E., Chen, E. E., Acree, C., Wang, H. H., & Sternberg, S. H. (2020). CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nature Biotechnology*, *39*(4), 480–489.

Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., & Wang, Y. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nature Communications*, *10*(1), 4284.

Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, *85*, 227–264.

Wang, T., Guan, C., Guo, J., Liu, B., Wu, Y., Xie, Z., Zhang, C., & Xing, X.-H. (2018). Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nature Communications*, *9*(1), 2475.

Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, *343*(6166), 80–84.

Wang, X., Li, J., Wang, Y., Yang, B., Wei, J., Wu, J., Wang, R., Huang, X., Chen, J., & Yang, L. (2018). Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion. *Nature Biotechnology*, *36*(10), 946–949.

Ward, H. N., Aregger, M., Gonatopoulos-Pournatzis, T., Billmann, M., Ohsumi, T. K., Brown, K. R., Blencowe, B. J., Moffat, J., & Myers, C. L. (2021). Analysis of combinatorial CRISPR screens with the Orthrus scoring pipeline. *Nature Protocols*, *16*(10), 4766–4798.

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, *104*, 101822.

Wessels, H.-H., Méndez-Mancilla, A., Guo, X., Legut, M., Daniloski, Z., & Sanjana, N. E. (2020). Massively parallel Cas13 screens reveal principles for guide RNA design. *Nature Biotechnology*. https://doi.org/10.1038/s41587-020-0456-9

Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2021). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews. Genetics*. https://doi.org/10.1038/s41576-021-00434-9

Wiedenheft, B., van Duijn, E., Bultema, J. B., Waghmare, S. P., Zhou, K., Barendregt, A., Westphal, W., Heck, A. J. R., Boekema, E. J., Dickman, M. J., & Doudna, J. A. (2011). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10092–10097.

Wienert, B., Wyman, S. K., Richardson, C. D., Yeh, C. D., Akcakaya, P., Porritt, M. J., Morlock, M., Vu, J. T., Kazane, K. R., Watry, H. L., Judge, L. M., Conklin, B. R., Maresca, M., & Corn, J. E. (2019). Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science*, *364*(6437), 286–289.

Wilson, L. O. W., Reti, D., O'Brien, A. R., Dunne, R. A., & Bauer, D. C. (2018). High Activity Target-Site Identification Using Phenotypic Independent CRISPR-Cas9 Core Functionality. *The CRISPR Journal*, *1*, 182–190.

Wimmer, F., & Beisel, C. L. (2019). CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. *Frontiers in Microbiology*, *10*, 3078.

Wimmer, F., Mougiakos, I., Englert, F., & Beisel, C. L. (2022). Rapid cell-free characterization of multi-subunit CRISPR effectors and transposons. *Molecular Cell*, *82*(6), 1210–1224.e6.

Wong, N., Liu, W., & Wang, X. (2015). WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biology*, *16*, 218.

Wu, X., Scott, D. A., Kriz, A. J., Chiu, A. C., Hsu, P. D., Dadon, D. B., Cheng, A. W., Trevino, A. E., Konermann, S., Chen, S., Jaenisch, R., Zhang, F., & Sharp, P. A. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology*, *32*(7), 670–676.

Xiang, X., Corsi, G. I., Anthon, C., Qu, K., Pan, X., Liang, X., Han, P., Dong, Z., Liu, L., Zhong, J., Ma, T., Wang, J., Zhang, X., Jiang, H., Xu, F., Liu, X., Xu, X., Wang, J., Yang, H., … Luo, Y. (2021). Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nature Communications*, *12*(1), 3238.

Xiao, A., Cheng, Z., Kong, L., Zhu, Z., Lin, S., Gao, G., & Zhang, B. (2014). CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics* , *30*(8), 1180–1182.

Xie, K., Minkenberg, B., & Yang, Y. (2015). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3570–3575.

Xie, M., & Fussenegger, M. (2018). Designing cell function: assembly of synthetic gene circuits for cell biology applications. *Nature Reviews. Molecular Cell Biology*, *19*(8), 507–525.

Xie, S., Shen, B., Zhang, C., Huang, X., & Zhang, Y. (2014). sgRNAcas9: a software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. *PloS One*, *9*(6), e100448.

Xiong, T., Meister, G. E., Workman, R. E., Kato, N. C., Spellberg, M. J., Turker, F., Timp, W., Ostermeier, M., & Novina, C. D. (2017). Targeted DNA methylation in human cells using engineered dCas9-methyltransferases. *Scientific Reports*, *7*(1), 6732.

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., Brown, M., & Liu, X. S. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Research*, *25*(8), 1147–1157.

Xu, L., Zhao, L., Gao, Y., Xu, J., & Han, R. (2017). Empower multiplex cell and tissue-specific CRISPR-mediated gene manipulation with self-cleaving ribozymes and tRNA. *Nucleic Acids Research*, *45*(5), e28.

Xu, P., Liu, Z., Liu, Y., Ma, H., Xu, Y., Bao, Y., Zhu, S., Cao, Z., Zhou, Z., & Wei, W. (2021). Genome-wide interrogation of gene functions through base editor screens empowered by barcoded sgRNAs. *Nature Biotechnology*, 1403–1413.

Xu, X., Duan, D., & Chen, S.-J. (2017). CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. *Scientific Reports*, *7*(1), 143.

Xu, X., Xu, L., Yuan, G., Wang, Y., Qu, Y., & Zhou, M. (2018). Synergistic combination of two antimicrobial agents closing each other's mutant selection windows to prevent antimicrobial resistance. *Scientific Reports*, *8*(1), 7237.

Yang, B., Yang, L., & Chen, J. (2019). Development and Application of Base Editors. *The CRISPR Journal*, *2*(2), 91–104.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316.

Yilmaz, A., Peretz, M., Aharony, A., Sagi, I., & Benvenisty, N. (2018). Defining essential genes for human pluripotent stem cells by CRISPR–Cas9 screening in haploid cells. *Nature Cell Biology*, *20*(5), 610–619.

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics. Conference Series*, *1168*(2), 022022.

Yuan, Q., & Gao, X. (2022). Multiplex base- and prime-editing with drive-and-process CRISPR arrays. *Nature Communications*, *13*(1), 2771.

Yu, Y., Gawlitt, S., de Andrade e Sousa, L. B., Merdivan, E., Piraud, M., Beisel, C., & Barquist, L. (2022). Improved prediction of bacterial CRISPRi guide efficiency through data integration and automated machine learning. In *bioRxiv* (p. 2022.05.27.493707). https://doi.org/10.1101/2022.05.27.493707

Zamanighomi, M., Jain, S. S., Ito, T., Pal, D., Daley, T. P., & Sellers, W. R. (2019). GEMINI: a variational Bayesian approach to identify genetic interactions from combinatorial CRISPR screens. *Genome*

Biology, *20*(1), 137.

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V., & Zhang, F. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, *163*(3), 759–771.

Zhang, Y., Wang, J., Wang, Z., Zhang, Y., Shi, S., Nielsen, J., & Liu, Z. (2019). A gRNA-tRNA array for CRISPR-Cas9 based rapid multiplexed genome editing in Saccharomyces cerevisiae. *Nature Communications*, *10*(1), 1053.

Zheng, K., Wang, Y., Li, N., Jiang, F.-F., Wu, C.-X., Liu, F., Chen, H.-C., & Liu, Z.-F. (2018). Highly efficient base editing in bacteria using a Cas9-cytidine deaminase fusion. *Communications Biology*, *1*, 32.

Zuo, E., Sun, Y., Wei, W., Yuan, T., Ying, W., Sun, H., Yuan, L., Steinmetz, L. M., Li, Y., & Yang, H. (2019). Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science*, *364*(6437), 289–292.

Appendix

The description of supplement tables S1 to S17 are as follows, and the file containing the tables is accessible on GitHub:
https://github.com/yanyyyy3/PhD_thesis/blob/main/CRISPRi/Yu_CRISPRi_supplementary_tables.xlsx

Table S1	Description of the features used in model training
Table S2	Description of train-test splits in the study
Table S3	Model evaluation of feature engineering
Table S4	SHAP values from model predicting the gRNA depletion
Table S5	Model evaluation of data fusion
Table S6	Model evaluation of data fusion with other model types
Table S7	Model evaluation of segregating gene and guide effects
Table S8	SHAP values from segregation models
Table S9	SHAP interaction values from MERF
Table S10	Description of gRNAs targeting deGFP
Table S11	Description of gRNAs in miller assay
Table S12	Model evaluation for targeting deGFP and miller assay
Table S13	Description of gRNAs for screen of purine biosynthesis genes
Table S14	Model evaluation of saturating screen of purine biosynthesis genes
Table S15	List of strains
Table S16	List of plasmids
Table S17	List of oligos for plasmid construction, library generation, and sequencing

Figure S1.1

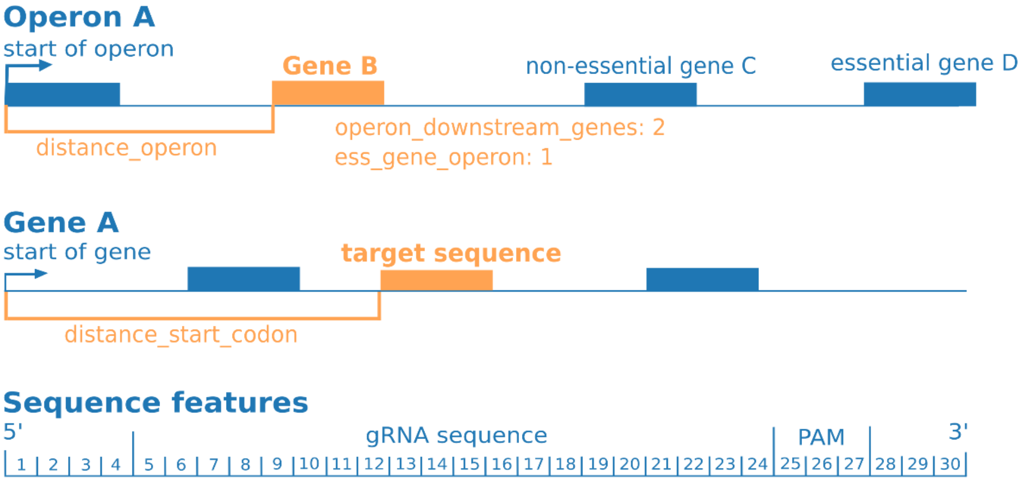

Figure S1.1: Illustration of the genomic and sequence features used, see also Table S1.

Figure S1.2

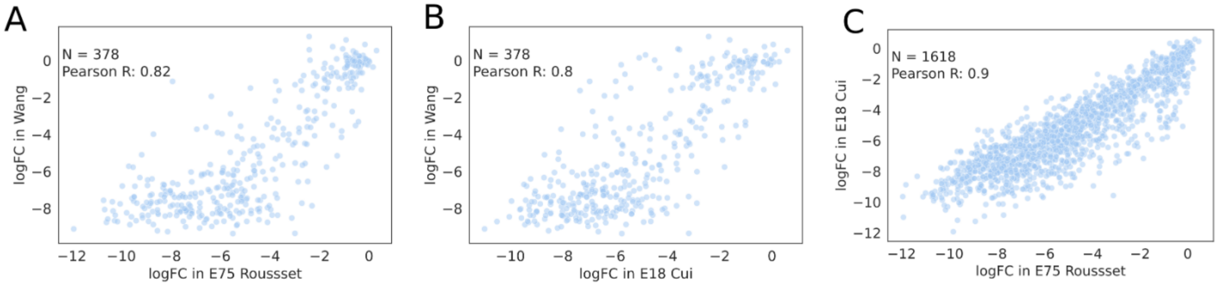

Figure S1.2: Comparison of guide depletion across datasets. (**A**) The logFC of gRNAs in E75 Rousset plotted against that in Wang for shared gRNAs. (**B**) The logFC of gRNAs in E18 Cui was plotted against that in Wang for shared gRNAs. (**C**) The logFC of gRNAs in E18 Cui was plotted against that in E75 Rousset for overlapping gRNAs.

Figure S1.3

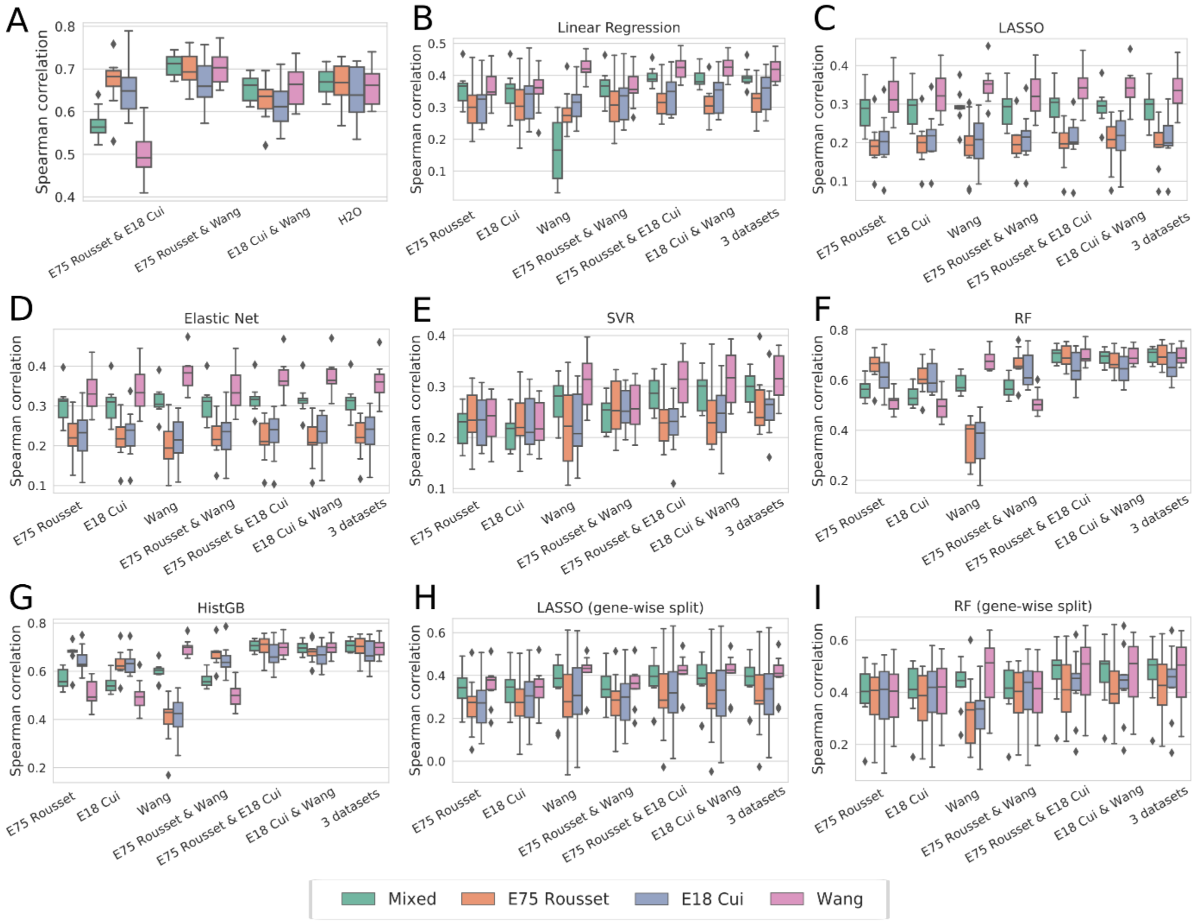

Figure S1.3: Spearman correlation of 10-fold cross-validation of models trained with one or mixed datasets. (**A**) Auto-sklearn optimized models trained with the indicated combination of datasets and H2O optimized model trained with integrated three datasets. (**B**) linear regression, (**C**) LASSO, (**D**) Elastic net, (**E**) support vector regression (SVR), (**F**) Random forest (RF) regression, (**G**) Histogram-based gradient boosting regression (HistGB), (**H**) LASSO (same hyperparameters as the MS LASSO model, gene-wise split). (**I**) Random forest (same hyperparameters as the MS random forest model, gene-wise split).

Figure S1.4

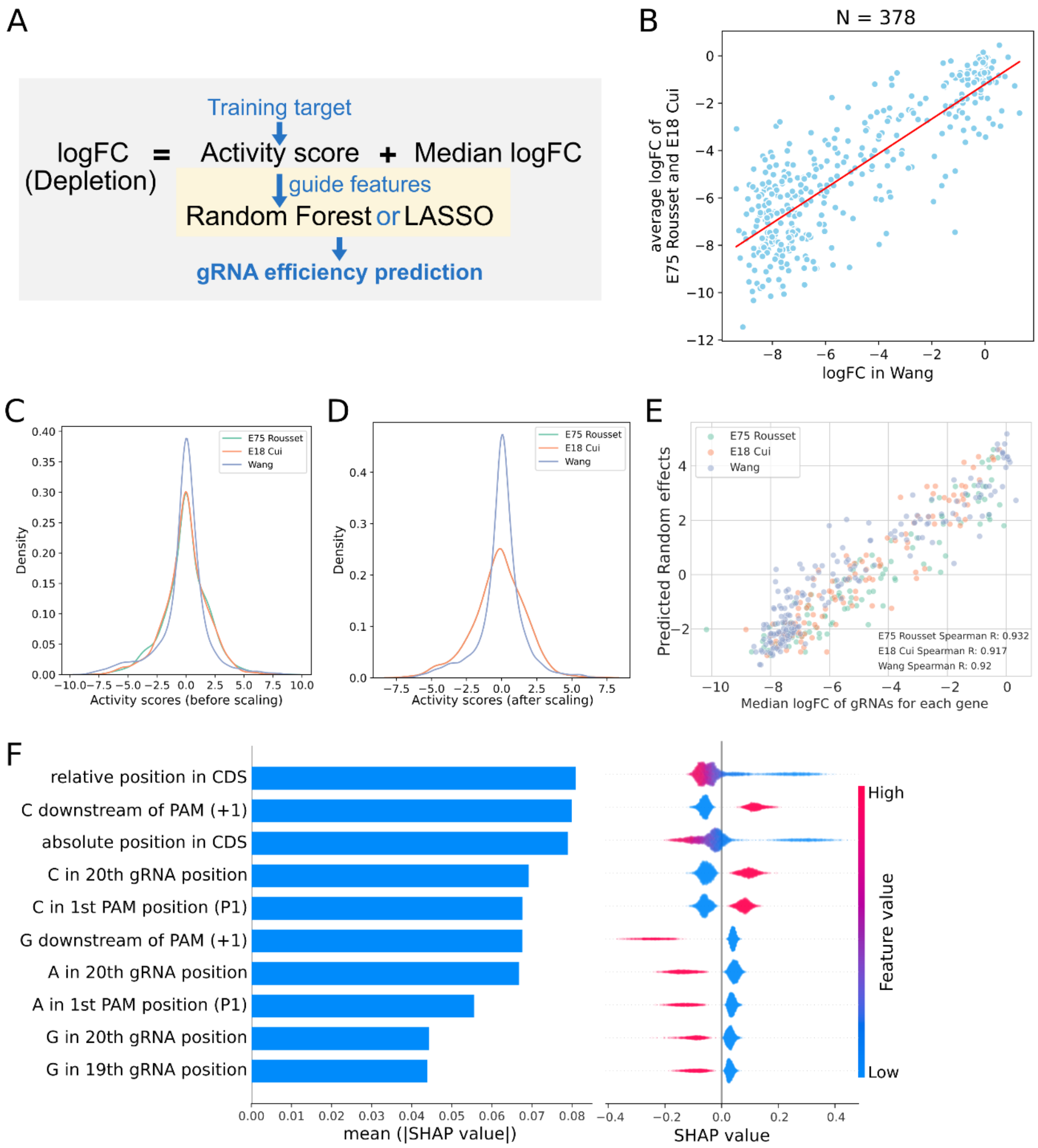

Figure S1.4: MS models for segregation of gene and guide effects. (**A**) The overview of training an MS model. We subtract the gene-wise median logFC from each gRNA depletion value upon data fusion to obtain the activity scores of each gRNA. The activity scores were used as training targets and the random forest or LASSO models were trained with 128 guide-specific features. (**B**) The logFC values in Wang were scaled based on the linear regression between the original logFC of Wang and the average logFC of E75 Rousset and E18 Cui for the 378 overlapping gRNAs. The distributions of activity scores (**C**) with and (**D**) without scaling are shown. (**E**) Predicted scores of the random effect model from MERF (y-axis) compared to the median logFC across gRNAs (x-axis) for each gene in each dataset. (**F**) SHAP values for the top ten features in the MS random forest model. Global feature importance is given by the mean absolute SHAP value (left), while the beeswarm plot (right) illustrates feature importance for each guide prediction.

Figure S1.5

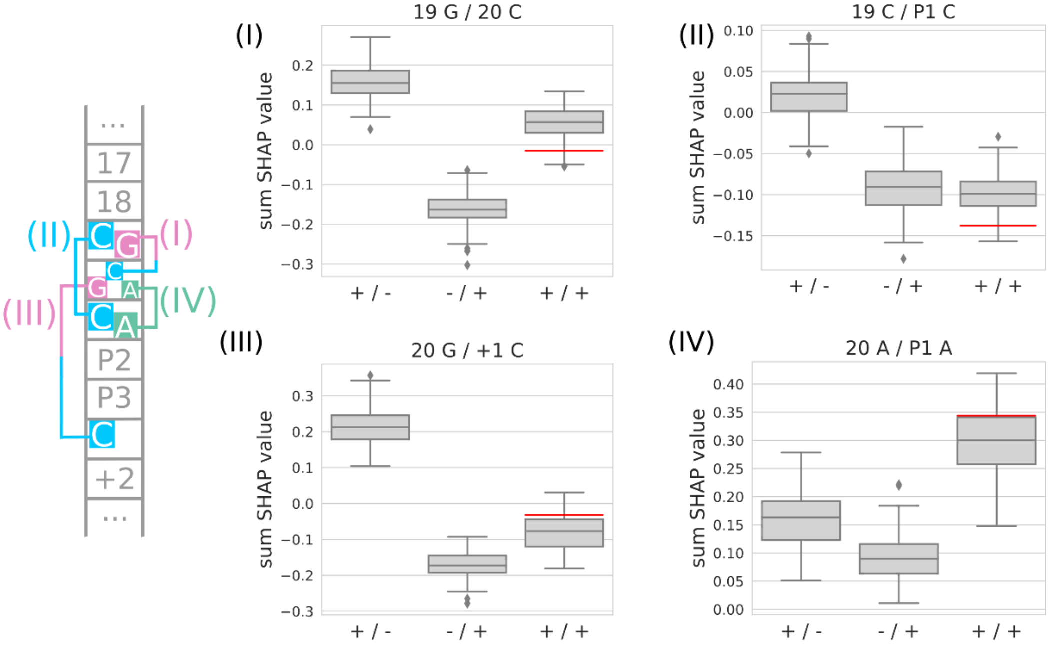

Figure S1.5: An illustration of feature interactions. The schematic on the left illustrates the positions of three representative interacting positions in the vicinity of the PAM sequence. (I-IV) show SHAP values for features in guides containing one (+/-) or the other (-/+) feature, or both (+/+). The expected SHAP value (red line) is calculated as the sum of the median SHAP values observed for each feature when occurring independently. 20 C/G/A: C/G/A in 20th gRNA position, +1 C: C downstream of PAM, P1 A/C: A/C in the variable position of the NGG PAM.

Figure S1.6

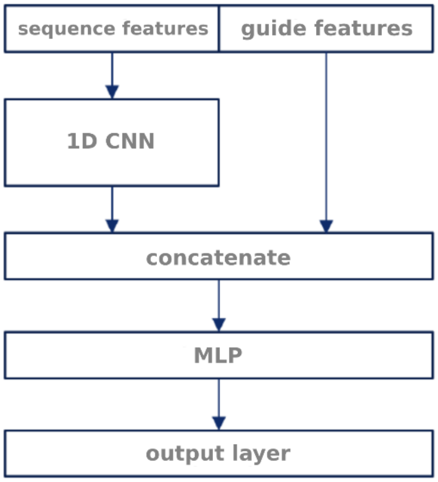

Figure S1.6: Deep learning approaches do not improve prediction performance. Architectures of the applied deep learning models for MS method. Sequence features were processed using 1D convolutional layers and later concatenated with other guide features. Guide features refer to guide-specific features apart from sequence features. MLP: multilayer perceptron.

Figure S1.7

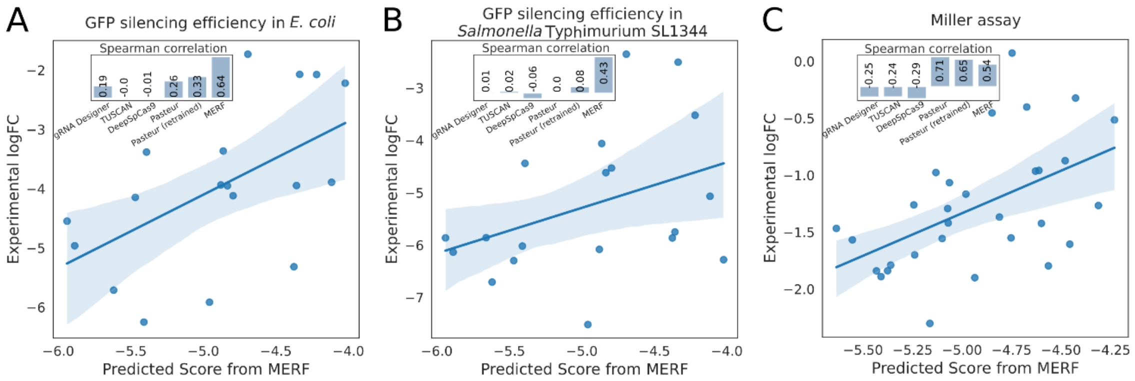

Figure S1.7: Independent validation of model performance. The activity of 19 gRNAs targeting a plasmid-expressed deGFP gene was measured in (**A**) *E. coli* and in (**B**) *Salmonella* Typhimurium SL1344 using a flow cytometry-based assay. The measured activity compared to the control gRNA is plotted against the score predicted by the MERF model. The inset barplot illustrates Spearman correlations for six methods for predicting guide efficiency. (**C**) The activity of 30 gRNAs targeting lacZ was measured with a Miller assay by Calvo-Villamañán et al., plotted as in A and B.

Figure S1.8

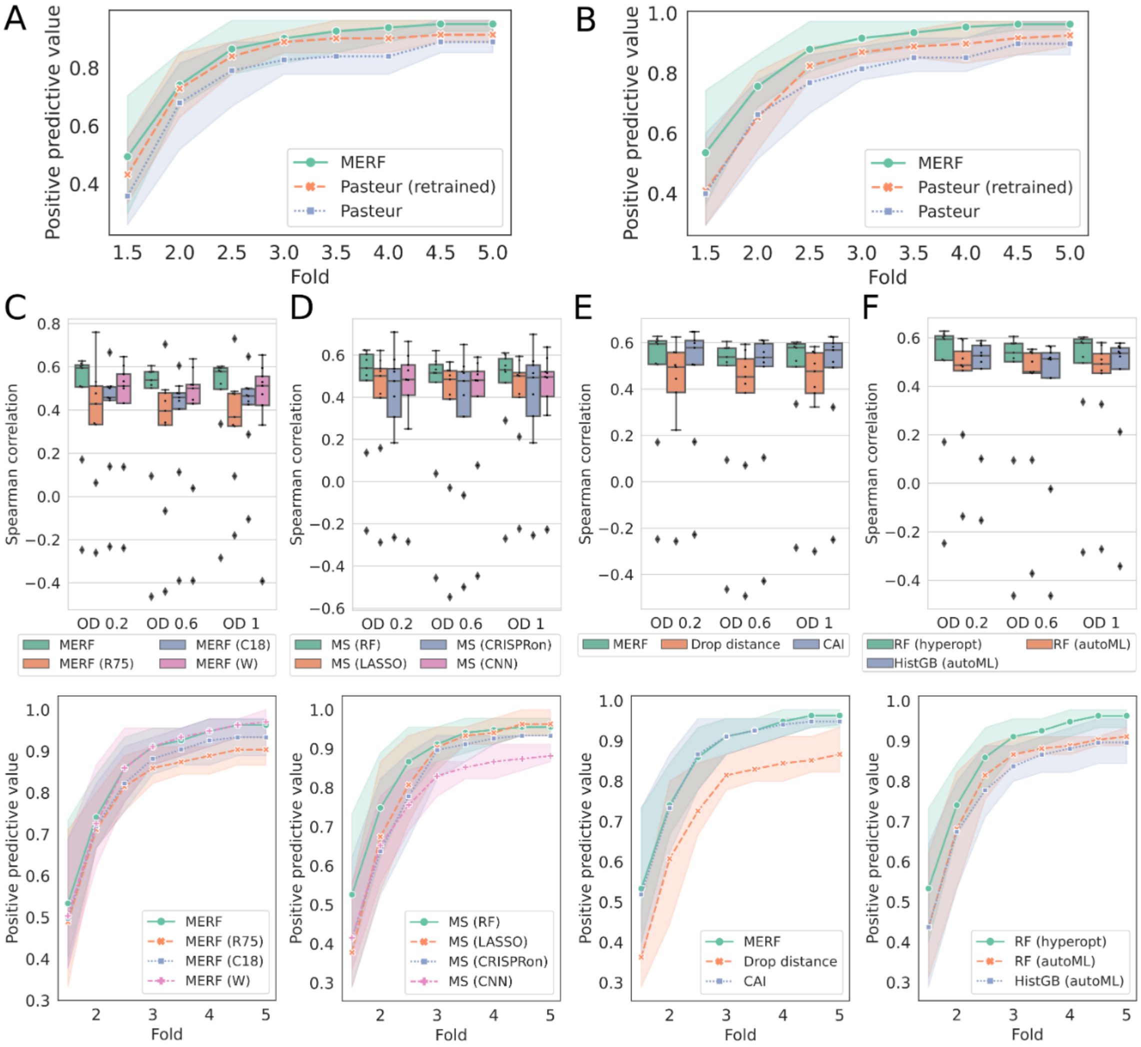

Figure S1.8: Additional figures related to model validation using a saturating screen of purine biosynthesis genes. (**A, B**) Positive predictive values of all gRNAs for each time point. The predicted positives are defined as (**A**) the top 3 and (**B**) the top 4 predicted gRNAs in each gene, while true positives are gRNAs with fold change within the N fold of the strongest depleted gRNA in each gene (N= 1.5 - 5 with a step of 0.5). (**C-F**) Performance on the purine screen of different models. The calculation of Spearman correlation and positive predictive value is the same as **Figure 3.1.5B-C**. (**C**) MERF random forest model trained with individual or integrated three datasets, (**D**) MS random forest, LASSO, and deep learning models trained with integrated three datasets, (**E**) MERF random forest model trained with or without distance features (Drop distance) and with 4 instead of 9 gene features for the random-effect model (CAI value, gene length, gene GC content, and dataset), (**F**) MERF fixed-effect model with random forest optimized using hyperopt (same in **Figure 3.1.5**), random forest optimized using auto-sklearn, and histogram-based gradient boosting model optimized using auto-sklearn. The hyperparameters in the auto-sklearn models were optimized for depletion prediction.

Figure S1.9

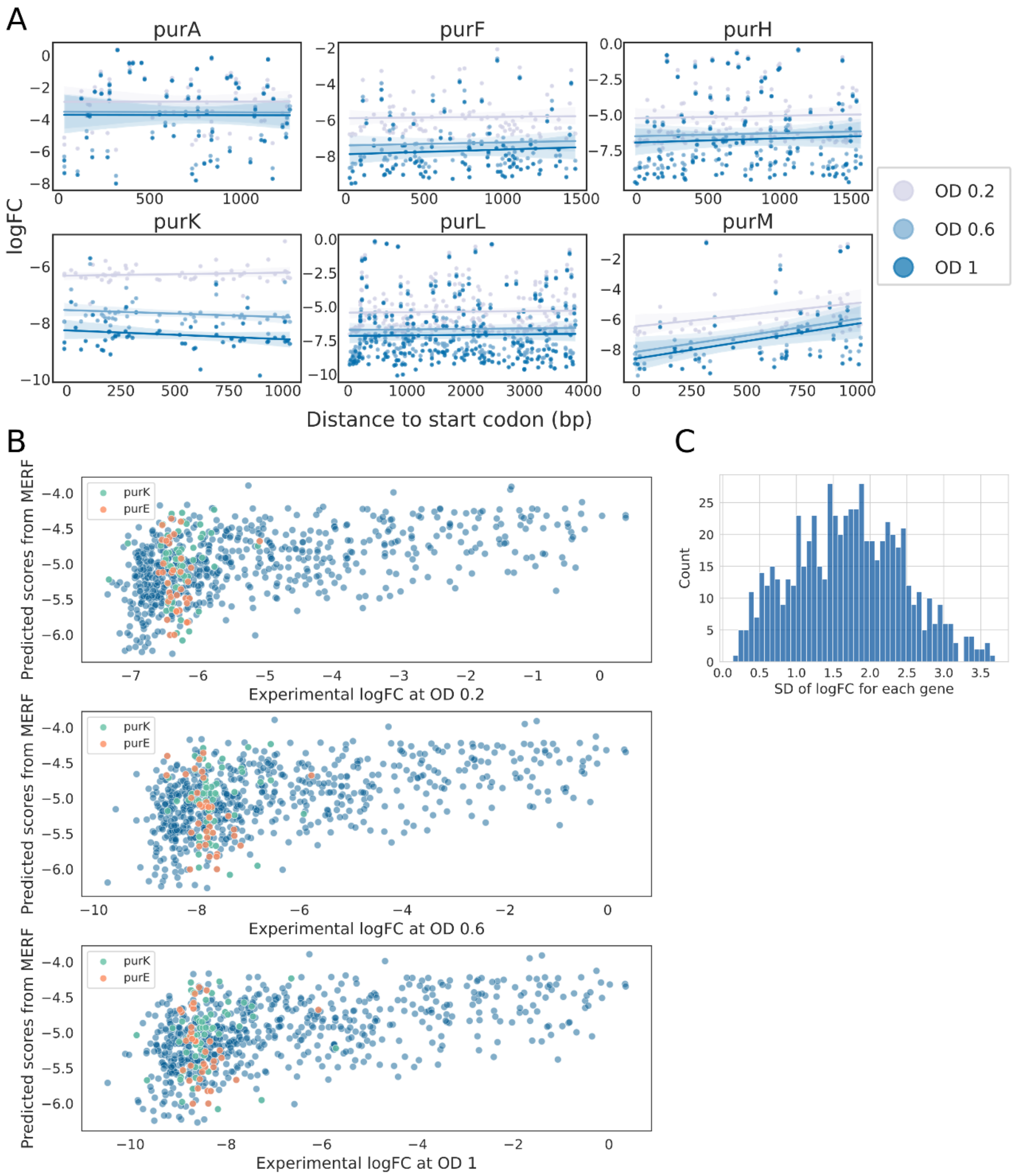

Figure S1.9: Additional figures related to measured logFC in the saturating screen of purine biosynthesis genes. (**A**) Measured logFCs for each guide as a function of distance to the start codon for the other 6 genes not shown in Figure 3.1.5. (**B**) The predicted scores from the MERF random forest model were plotted against experimental logFC at different time points. Guides targeting purE and purK were marked with orange and green respectively. (**C**) The distribution of the standard deviation of the logFC for guides targeting each gene in the training data.

Figure S1.10

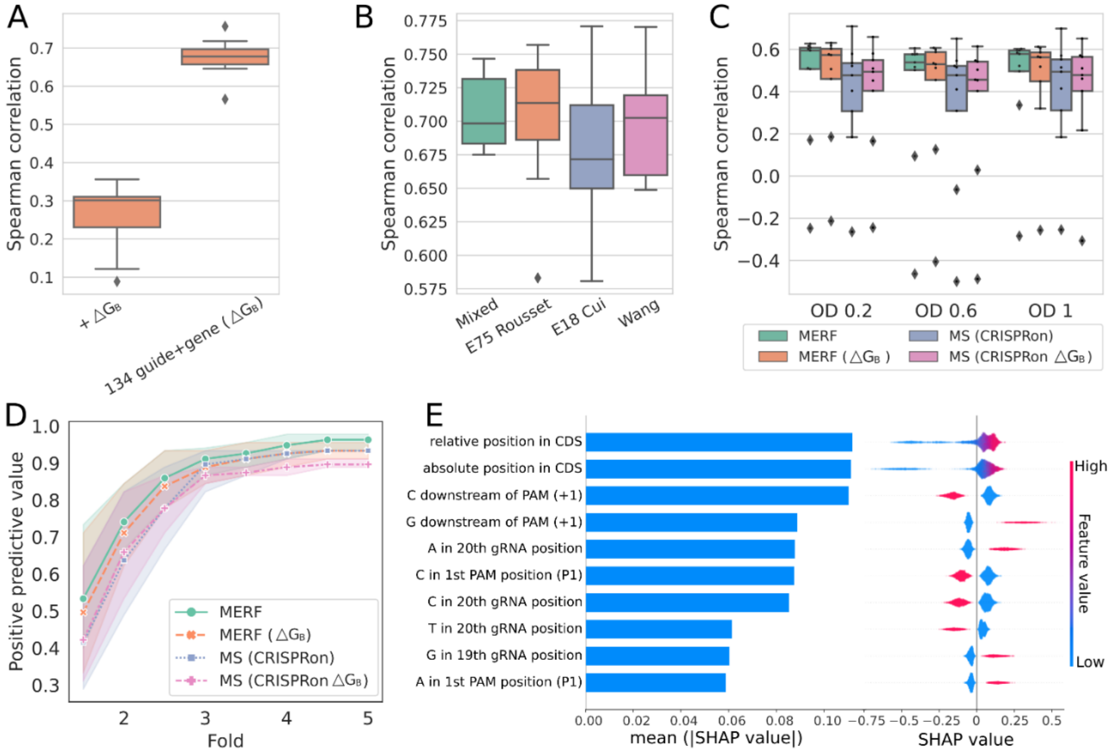

Figure S1.10: The model performance replacing 4 thermodynamic minimum free energy (MFE) features with $\triangle G_B$. (**A**) Comparison of Spearman correlation between actual and predicted guide depletion in 10-fold cross-validation (CV) of the best model trained with Auto-Sklearn with different feature combinations, using data from (Rousset et al., 2018). + $\triangle G_B$: only sequence, distance, and $\triangle G_B$ features. 134 guide + gene ($\triangle G_B$): all features. (**B**) Comparison of Spearman correlation from the 10-fold CV of the best auto-sklearn trained model on the integrated three datasets to predict depletion. The x-axis indicates the test sets. (**C**) Spearman correlations between the predicted scores and measured logFC for each gene across collected time points in the saturating screen of purine biosynthesis genes. (**D**) Positive predictive values of all gRNAs for each time point in the saturating screen of purine biosynthesis genes. The predicted positives are defined as the top 5 predicted gRNAs in each gene, while true positives are gRNAs with logFC within the N fold of the strongest depleted gRNA in each gene (N= 1.5 - 5 with a step of 0.5). When $\triangle G_B$ is not specified, the MERF or MS CRISPRon models were trained with 4 MFE features. (**E**) SHAP values for the top 10 features from MERF model with $\triangle G_B$. Global feature importance is given by the mean absolute SHAP value (left), while the beeswarm plot (right) illustrates feature importance for each guide prediction.

Figure S2.1

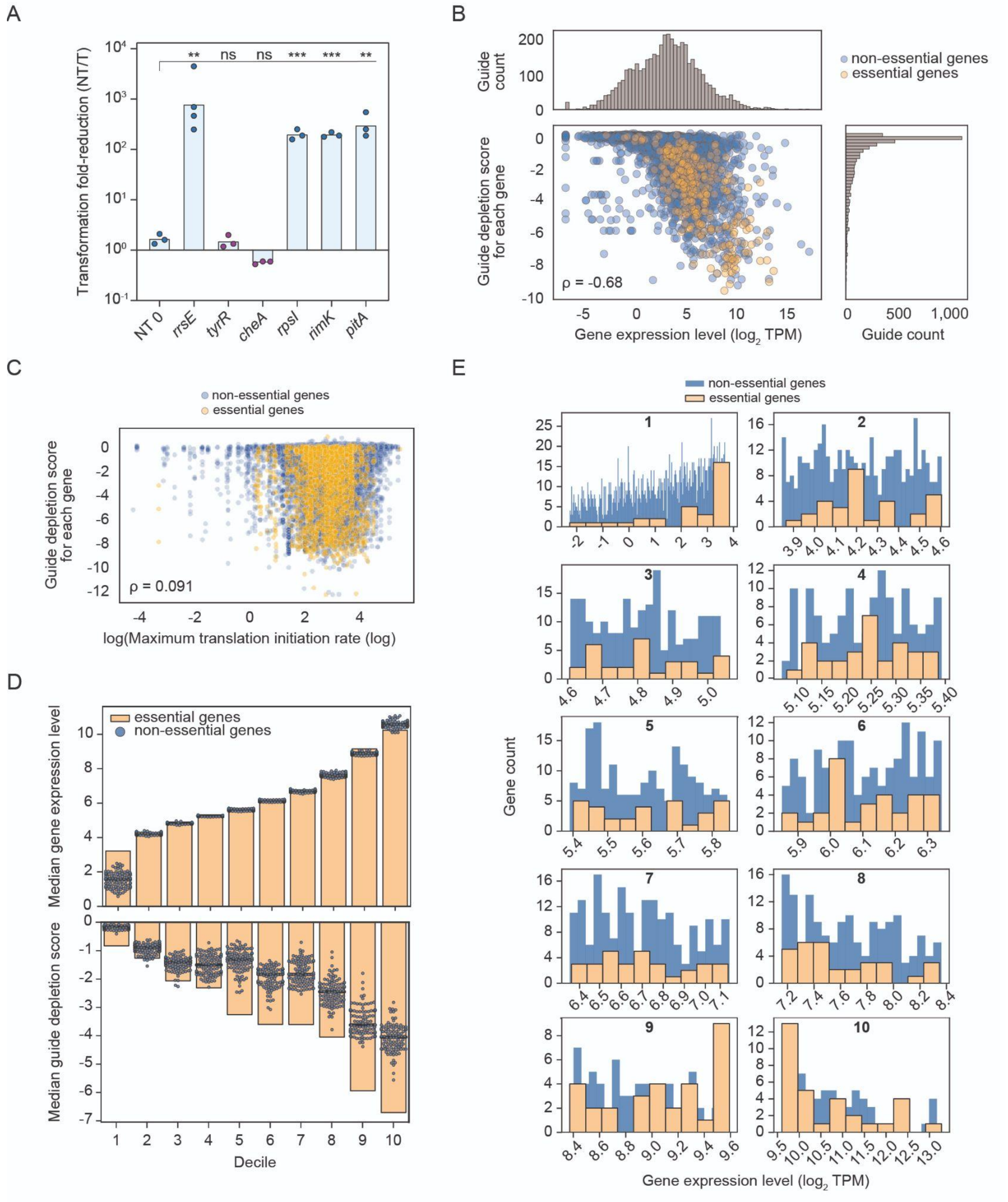

Figure S2.1: Additional results related to the guide depletion screen. Related to Figure 3.2.2. (**A**) Selected highly-depleted guides from the library yield reduced transformation or growth. A randomized crRNA from the library (NT 0) served as the negative control, while a crRNA targeting the rRNA encoded by rrsE served as a positive control. Purple data points represent small colonies indicative of reduced growth. Bars represent the mean

of at least triplicate independent experiments. Statistical significance was calculated by comparing the transformation fold-reduction to that of NT 0. ***: $p < 0.001$. **: $p < 0.01$. *: $p < 0.05$. ns: not significant or average below the reference. Tested guides resulted in significant fold-reduction in transformation or small colonies. (**B**) Correlation between guide depletion score and the expression levels of the target gene for cells cultured to a turbidity of $ABS_{600} \approx 0.8$. ρ: Spearman correlation coefficient. Values for transcript levels are the average of duplicate independent experiments, while values for the depletion score are the average of duplicate independent screens. (**C**) Predicted translation rate poorly correlates with guide depletion score in the library screen. Translation rates were predicted using the RBS calculator. (**D**) Median depletion scores of guides targeting non-essential genes are lower compared to essential genes having the same range of transcript expression. The data are derived from the experiment at $OD_{600} \approx 0.5$. (**E**) Distribution of expression level for the essential and non-essential genes in different deciles. Essential genes were evenly distributed into ten deciles while non-essential genes were split based on the expression values of essential genes of each decile. In the first decile, there is a large number of non-essential genes and few essential genes which distribute on the higher end of the range. This explains the lower median expression levels of guide targeting non-essential genes compared to essential genes for that decile.

Figure S2.2

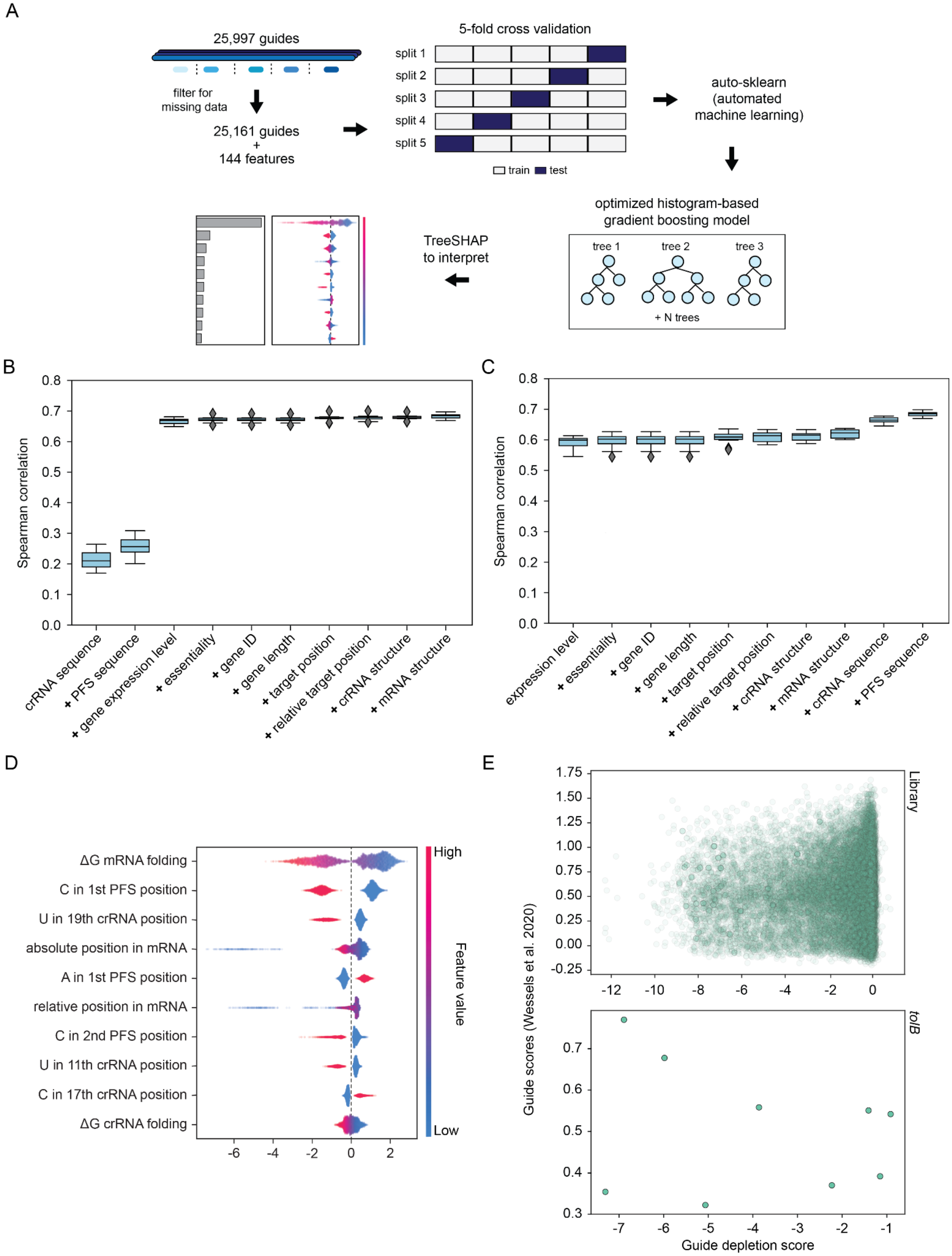
A
25,997 guides
filter for missing data
25,161 guides + 144 features
5-fold cross validation
split 1
split 2
split 3
split 4
split 5
train
test
auto-sklearn (automated machine learning)
optimized histogram-based gradient boosting model
tree 1
tree 2
tree 3
+ N trees
TreeSHAP to interpret
B
Spearman correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
crRNA sequence
+ PFS sequence
+ gene expression level
+ essentiality
+ gene ID
+ gene length
+ target position
+ relative target position
+ crRNA structure
+ mRNA structure
C
Spearman correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
expression level
+ essentiality
+ gene ID
+ gene length
+ target position
+ relative target position
+ crRNA structure
+ mRNA structure
+ crRNA sequence
+ PFS sequence
D
ΔG mRNA folding
C in 1st PFS position
U in 19th crRNA position
absolute position in mRNA
A in 1st PFS position
relative position in mRNA
C in 2nd PFS position
U in 11th crRNA position
C in 17th crRNA position
ΔG crRNA folding
-6
-4
-2
0
2
High
Feature value
Low
E
Library
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-12
-10
-8
-6
-4
-2
0
tolB
0.7
0.6
0.5
0.4
0.3
-7
-6
-5
-4
-3
-2
-1
Guide scores (Wessels et al. 2020)
Guide depletion score

Figure S2.2: Machine learning model and feature importance. Related to Figure 3.2.2. (**A**) Overview of the machine learning model development protocol used. Screen data was used to determine the depletion of each of 25,161 guides as part of growth in liquid culture. The derived log2 fold-changes were used as prediction targets, with 144 sequence, gene, thermodynamic, and expression features as predictors. Auto-sklearn was used to train and optimize an assortment of candidate model types. A histogram-based gradient boosting model was the best performing model. This model was then interpreted using TreeSHAP to explain the contribution of each predictor to each prediction. (**B**) Alternative evaluation of feature importance using cross-validation. The predictive power of each feature was evaluated by first training a simple histogram-based gradient boosting model with a single feature or feature set, and then iteratively adding additional features and comparing the Spearman correlation in 10-fold cross-validation of the model with the expanded feature set to previously trained models. For B, the base feature was the crRNA sequence. Adding PFS sequence features slightly increases the Spearman correlation, but a large jump in correlation is observed when adding expression level features. Other features lead to only minor changes in the Spearman correlation. (**C**) As in B, but using expression level as the base feature. This again shows that expression level alone has high predictive utility; features other than the crRNA and PFS sequence features have relatively minor effects on overall prediction accuracy. (**D**) SHAP values for the strongest predictors contributing to guide depletion independently from gene features. Guide features were assigned to a random forest model, while gene features were assigned to a random effect model using a mixed effect random forest (MERF). Most features are similar to the model with both guide and gene features, such as the presence of a C in 1st and 2nd PFT position, a U in 19th crRNA position and low mRNA and crRNA secondary structure. Additionally, there are three new parameters, one favoring depletion, which is the presence of a U in 11th crRNA position and two worsening guide activity, an A in 1st PFS position and a C in 17th crRNA position. (**E**) Correlation between guide scores from Wessels et al. algorithm and guide depletion scores in this study. The graph above depicts this correlation for all crRNA guides in the library (top) and crRNA guides targeting tolB (below).

Figure S2.3

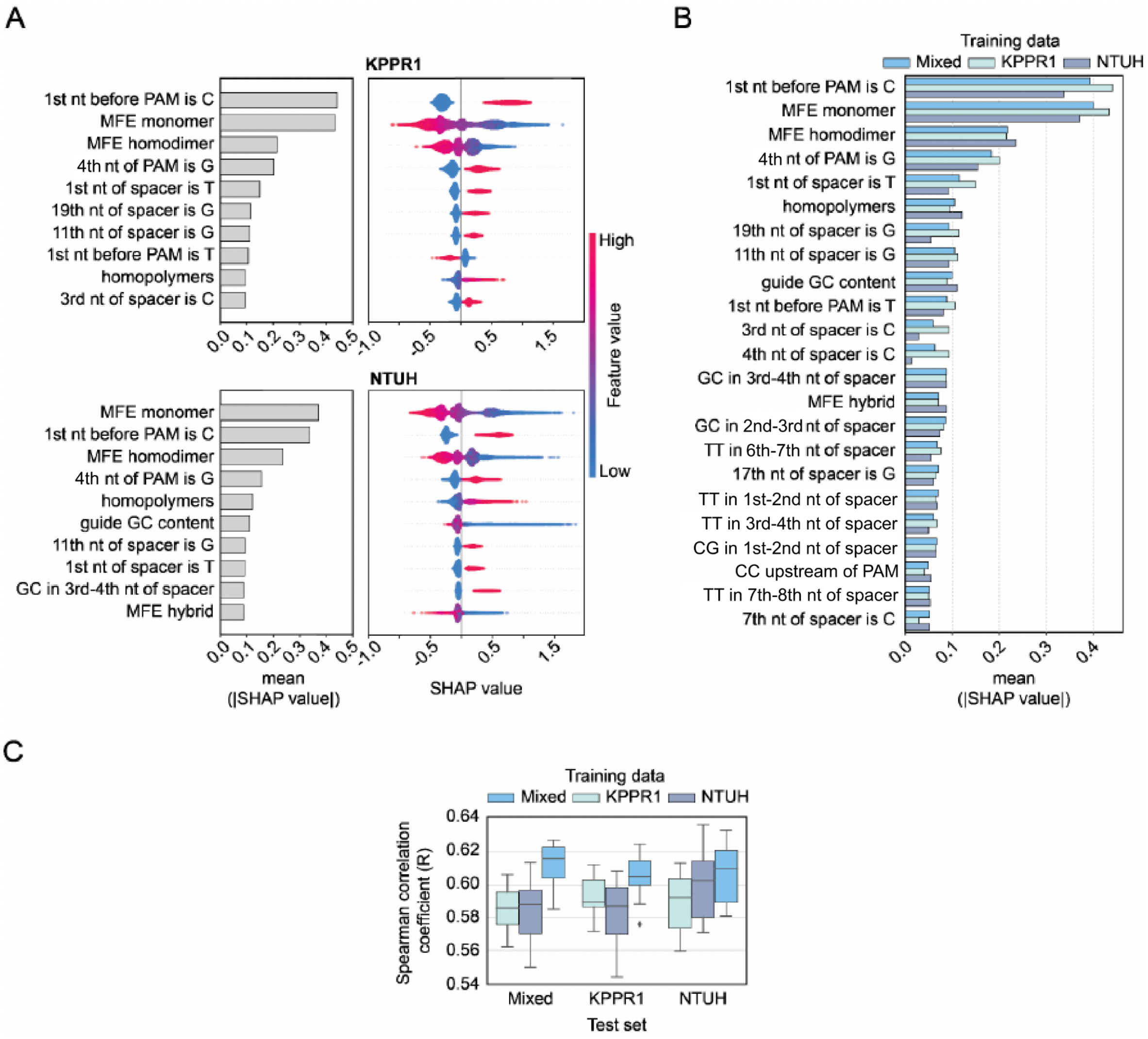

Figure S2.3: Features contributing to guide depletion in the single strains and in the mixed model. Related to Figure 3.2.3. (**A**) SHAP values for the ten strongest features determining guide depletion in the KPPR1 and NTUH strains when using the individual screen data for training the algorithm. The left barplot depicts the absolute weight of each feature to the depletion values, while the right beeswarm plot shows how each datapoint impacts depletion according to each feature. (**B**) Features determining guide depletion in the library screen by using the single strains or the mixed data to train the algorithm. (**C**) Comparison of Spearman correlation from 10-fold CV of the best auto-sklearn trained model on one dataset or the integrated three datasets.

List of Figures

List of Tables

Abbreviation index

Roman symbols

Adenine, Cytosine, Guanine, Thymine, Uracil	A, C, G, T, U
A/C	Y
A/C/T/G	N
A/G/C	V
A/G/T	D
A/T	W
T/G	K

Acronyms and Abbreviations

Abortive initiation	Abi
Acidaminococcus sp. Cas12a	AsCas12a
Adenine base editor	ABE
Ampicillin	Amp
Anhydrotetracycline	aTc
Anti-CRISPR	Acr
Argonaute	Ago
Automated machine learning	AutoML
Base pair	bp
CRISPR activation	CRISPRa
CRISPR interference	CRISPRi
CRISPR interference and activation	CRISPRi/a
CRISPR-associated	Cas
Cap Analysis of Gene Expression	CAGE
Cas9 nickase	nCas9
Chloramphenicol	Cm

Coding sequence	CDS
Codon adaptation index	CAI
Convolutional neural network	CNN
Cutting Frequency Determination	CFD
Cytosine base editor	CBE
Diaminopimelic acid	DAP
DNA endonuclease-targeted CRISPR trans reporter	DETECTR
Dead Cas9	dCas9
Double-strand break	DSB
False positive	FP
Fluorescence-activated cell sorting	FACS
Fluorescent in-situ hybridization	FISH
Francisella tularensis Cas12a	FnCas12a
Guide RNA	gRNA
Histogram-based gradient boosting	HistGB
Homology-directed repair	HDR
Hypervirulent	HV
Integrated Device Technology	IDT
KRAB-box-associated protein-1	KAP-1
Kanamycin	Kan
Kruppel-associated Box	KRAB
Lachnospiraceae bacterium Cas12a	LbCas12a
Least absolute shrinkage and selection operator	LASSO
Leptotrichia shahii Cas13a	LshCas13a
Leveraging engineered tracrRNAs and on-target DNAs for parallel RNA detection	LEOPARD
Log fold change	logFC
Lysogeny Broth	LB

Matthews correlation coefficient	MCC
Maximum likelihood estimation	MLE
Mean squared error	MSE
Median subtracting	MS
Minimum free energy	MFE
Mix-effect random forest	MERF
Mouse embryonic stem cells	mESC
Multi-drug resistance	MDR
Multilayer perceptron	MLP
Non-canonical crRNA	ncrRNA
Non-homologous end joining	NHEJ
Nucleic acid sequence-based amplification	NASBA
Nucleotide	nt
Optical density	OD
Overnight	ON
Phosphate-buffered saline	PBS
Polymerase chain reaction	PCR
Positive predictive values	PPV
Premature CRISPR RNA	pre-crRNA
Prime editing guide RNA	pegRNA
Protospacer adjacent motif	PAM
Protospacer flanking site	PFS
RNA binding protein	RBP
RNA polymerase	RNAP
Random forest	RF
Receiver operating characteristic	ROC
Recombinase polymerase amplification	RPA

Recurrent neural network	RNN
Reprogrammed tracrRNA	Rptr
Reverse transcription RPA	RT-RPA
SHapley Additive exPlanations	SHAP
ScCas9 nickase derived base editor	ScBE3
Sequence Scan for CRISPR	SSC
Single guide RNA	sgRNA
Single-cell CRISPR	scCRISPR
Single-cell RNA sequencing	scRNA-seq
Streptococcus canis Cas9	ScCas9
Streptococcus pyogenes Cas9	SpCas9
Streptococcus thermophilus Cas9 from CRISPR1 locus	Sth1Cas9
Support vector machines	SVM
Synergistic activation mediator	SAM
The area under the ROC curve	AUC
Trans-activating CRISPR RNA	tracrRNA
Transcription activator-like effector	TALE
Transcription start site	TSS
Transcription unit	TU
Transcripts per million	TPM
Transposon Directed Insertion Sequencing	TraDIS
Transposon sequencing	Tn-Seq
Tree-structured parzen estimators	TPE
True positive	TP
Untranslated regions	UTRs
VP64-p65-Rta	VPR
Zinc finger nucleases	ZFNs

Publications list

Preprint:

1. **Yu, Y.**, Gawlitt, S., e Sousa, L.B.D.A., Merdivan, E., Piraud, M., Beisel, C. and Barquist, L., 2022. Improved prediction of bacterial CRISPRi guide efficiency through data integration and automated machine learning. BioRxiv.

Published:

2. Vialetto, E., **Yu, Y.**, Collins, S. P., Wandera, K. G., Barquist, L., & Beisel, C. L. (2022). A target expression threshold dictates invader defense and prevents autoimmunity by CRISPR-Cas13. **Cell Host & Microbe**.
3. Jiao, C., Sharma, S., Dugar, G., Peeck, N.L., Bischler, T., Wimmer, F., **Yu, Y.**, Barquist, L., Schoen, C., Kurzai, O. and Sharma, C.M., (2021). Noncanonical crRNAs derived from host transcripts enable multiplexable RNA detection by Cas9. **Science**.
4. Shahraki, A., **Yu, Y**., Gul, Z.M., Liang, C. and Iyison, N.B., (2020). Whole genome sequencing of Thaumetopoea pityocampa revealed putative pesticide targets. **Genomics**.
5. Hamprecht, A., Barber, A.E., Mellinghoff, S.C., Thelen, P., Walther, G., **Yu, Y.**, Neurgaonkar, P., Dandekar, T., Cornely, O.A., Martin, R. and Kurzai, O., (2019). Candida auris in Germany and previous exposure to foreign healthcare. **Emerging infectious diseases**.

Curriculum vitae

Affidavit

I hereby confirm that my thesis entitled “Applied machine learning for the analysis of CRISPR-Cas systems” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date

Yanying Yu

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation “Angewandtes maschinelles Lernen für die Analyse von CRISPR-Cas-Systemen" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ort, Datum

Yanying Yu