ACS Synthetic Biology

pubs.acs.org/synthbio
Review

Engineering a New Generation of Gene Editors: Integrating Synthetic Biology and AI Innovations

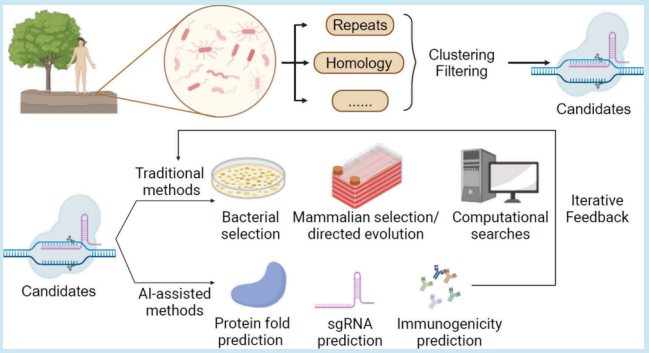
Bing Shao Chia,[#] Yu Fen Samantha Seah,[#] Bolun Wang, Kimberle Shen, Diya Srivastava, and Wei Leong Chew*

Cite This: *ACS Synth. Biol.* 2025, 14, 636–647

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: CRISPR-Cas technology has revolutionized biology by enabling precise DNA and RNA edits with ease. However, significant challenges remain for translating this technology into clinical applications. Traditional protein engineering methods, such as rational design, mutagenesis screens, and directed evolution, have been used to address issues like low efficacy, specificity, and high immunogenicity. These methods are labor-intensive, time-consuming, and resource-intensive and often require detailed structural knowledge. Recently, computational strategies have emerged as powerful solutions to these limitations. Using artificial intelligence (AI) and machine learning (ML), the discovery and design of novel gene-editing enzymes can be streamlined. AI/ML models predict activity, specificity, and immunogenicity while also enhancing mutagenesis screens and directed evolution. These approaches not only accelerate rational design but also create new opportunities for developing safer and more efficient genome-editing tools, which could eventually be translated into the clinic.



KEYWORDS: *Synthetic Biology, Genome Editing, Protein Design, Artificial Intelligence*

INTRODUCTION

CRISPR-Cas technology has impacted biology by enabling precise DNA/RNA edits with accuracy and ease. Despite these advances, current genome editing tools still face significant hurdles that limit their clinical translation and real-world impact. Key challenges include improving editing efficacy and specificity, as well as reducing protein size and immunogenicity. Traditionally, protein engineering has relied on labor-intensive techniques like rational protein design, directed evolution, and mutagenesis screens, but these approaches are often time-consuming and constrained by the need for pre-existing, detailed experimental and structural knowledge.

It is here that the Synthetic Biology toolkit in enzyme discovery and engineering presents an enticing opportunity in addressing key limitations of DNA/RNA editors. Here, we review the recent advances that develop novel and improved DNA/RNA editors through computational strategies. For broader discussions of CRISPR-Cas discovery, engineering and applications, we refer readers to previous reviews.[1−3]

Computational biology has been employed to tackle many challenges faced in Synthetic Biology, by enabling the design, modeling, simulation and optimization of biological systems in a speedy and scalable manner. Moreover, the recent trajectory of artificial intelligence (AI) and machine learning (ML) tools opens new avenues for designing more efficient and precise gene editors. AI/ML models are particularly useful for predicting on-target and off-target activity of guide RNAs (gRNAs), a critical factor in ensuring the specificity (and safety) of gene editing. AI/ML-driven approaches can also enhance mutagenesis screens and support directed evolution by identifying promising variants for testing, accelerating the discovery of optimized editors by reducing the vast search space of all possible mutants that would otherwise have to be tested.

We also address novel modalities, such as base editing and prime editing, which offer greater precision and versatility over the base CRISPR-Cas endonucleases. Increased precision could refer to fewer off-target effects outside the target locus, and/or reduced bystander edits, which are undesired changes within the target locus. New modalities that enable larger segments of the genome to be rewritten are also emerging, including the Insertion Sequences IS110/IS1111, recombinases, and transposases. Data-informed computational strat-

Received: October 4, 2024
Revised: January 6, 2025
Accepted: January 16, 2025
Published: February 25, 2025

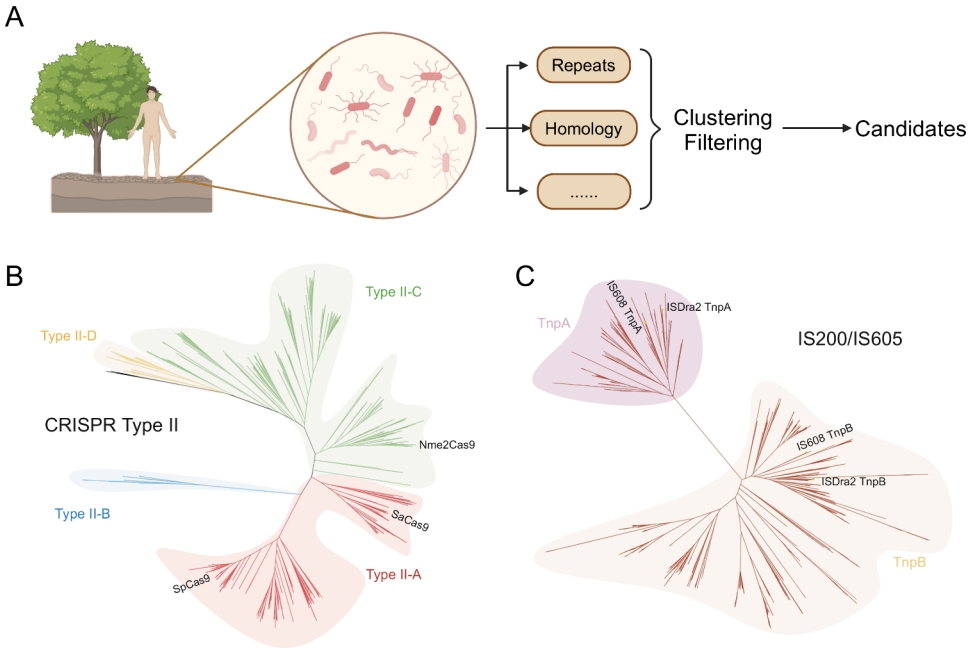


Figure 1. Finding genome editors. (A) Schematic representation of the key features of finding new genome editors. This figure was created with BioRender.com. (B) Phylogenetic tree illustrating CRISPR Type II proteins, with data sourced from CasPedia.[16] (C) Phylogenetic tree of IS200/IS605 family proteins, constructed from sequences retrieved from the ISfinder database,[17] aligned using MAFFT, and analyzed with IQ-TREE2.[18] Both phylogenetic trees were visualized using iTOL with representative proteins annotated.[19]

egies have immense potential for the enhancement and refinement of these novel modalities.

FINDING NEW EDITORS

Clustered regularly interspaced short palindromic repeats (CRISPR) was independently discovered by several groups in the 1980s and 1990s,[4] but its biological function had remained elusive until the 2000s when it was characterized as a prokaryotic defense mechanism.[5] CRISPR was subsequently repurposed for gene editing in bacterial and mammalian cells,[6−8] opening the door to potentially efficient gene editing therapies. Despite its high programmability and relative ease of use, the translation and extensive adoption of CRISPR as genetic medicines have been limited by efficiency of targeting, specificity, size and immunogenicity of CRISPR-associated (Cas) proteins, as well as the sequence contexts that Cas complexes could target. Therefore, there is continued interest in discovering new CRISPR systems with desirable features.

A suite of bioinformatics tools has been used to identify CRISPR arrays and their associated Cas protein genes within bacterial and archaeal genomes. Some key characteristics of CRISPR-Cas systems are the presence of repetitive sequences (direct repeats), which constitute the RNAs that program CRISPR-Cas complexes toward the target sites, and highly conserved genes such as Cas1 and Cas2, both of which are involved in spacer acquisition. Commonly utilized tools include CRISPR Recognition Tool (CRT),[9] PILER-CR,[10] CRISPRfinder[11] and its upgraded version CRISPRCasfinder.[12] CRT directly scans DNA sequences for exact k-mer matches to detect repeats without additional preprocessing. PILER-CR builds upon a repeat analysis algorithm of PILER targeting CRISPR repeats specifically, while CRISPRfinder uses Vmatch, Fuzznuc, and CLUSTALW to identify CRISPR repeats and perform BLAST searches against the GenBank database. A schematic of this process is presented in Figure 1A.

ML has been utilized to augment the search for new editors, particularly through the enhanced detection and clustering of genes encoding these enzymes. CRISPRCasFinder excels in Cas gene detection and shows improved CRISPR specificity by integrating CasFinder from MacSyFinder,[13] utilizing protein similarity searches via Hidden Markov Models. Another tool, CRISPRCasTyper,[14] employs 680 Hidden Markov Models to identify Cas genes and uses BLAST and a k-mer-based ML approach within its RepeatType module to detect CRISPR arrays. More recently, a fast locality-sensitive hashing−based clustering (FLSHclust) algorithm was reported.[15] Using this approach of clustering similar, but not identical, protein and DNA sequences, Altae-Tran et al. were able to significantly reduce the time needed to segregate the vast data consisting of 8 billion proteins and 10.2 million CRISPR arrays from months to weeks without significantly sacrificing clustering performance. When compared to Linclust and MMSeqs2 on a large data set containing 51 million proteins, FLSHclust was found to cluster 58% more proteins than Linclust and only 12% fewer than MMSeqs2, while FLSHclust exhibits linearithmic scaling in practice, which allows it to run faster than Linclust and MMSeqs2, two quadratic-scaling algorithms. Over 100 novel CRISPR-associated gene modules were subsequently identified, several of which (DinG-HNH, Cas5-HNH, Cas8-HNH, and a candidate type VII system) were experimentally characterized and/or engineered.

New systems incorporating potential gene editors have also been recently discovered through computational analyses of evolutionary genomic data. For example, IscB from the IS200/IS605 transposon family has been identified as a putative ancestor to Cas9 via clustering and phylogenetic analysis of the RuvC endonuclease, bridge helix (BH), and HNH endonuclease domains[20] (Figure 1B, C). At approximately 400 amino acids in size, IscB is smaller than Cas9 and thus more easily packaged into delivery vectors, making it promising for

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

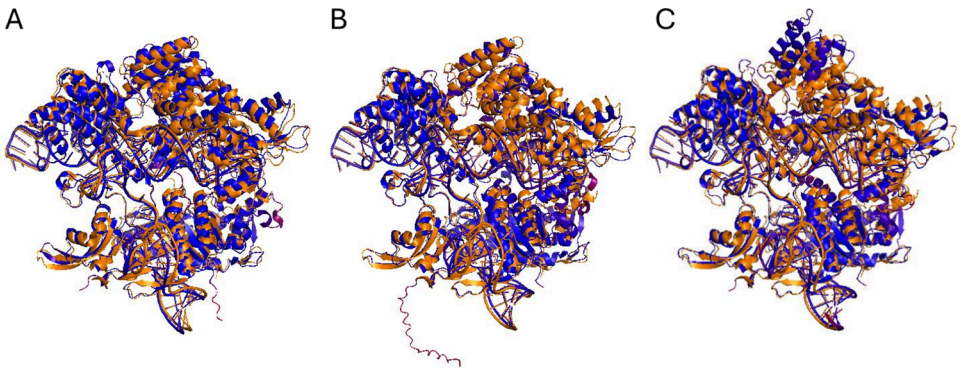


Figure 2. Alignment of experimentally determined *Streptococcus pyogenes* Cas9 structure in complex with sgRNA and target DNA (orange) with Alphafold3 (AF3) generated structures[44] (blue, red). The predicted structure is colored by pLDDT values, with residues with high pLDDT values shown in blue, residues with moderate values in purple and residues with low values in red. (A) Alignment of experimentally determined *Streptococcus pyogenes* Cas9 structure in complex with sgRNA and target DNA[45] (orange) with the AF3-generated structure using the same amino acid and nucleic acid sequences (blue, red). The AF3 prediction is of high confidence, as most residues have pLDDT > 90, pTM = 0.92 and iPTM = 0.94. The experimentally determined and AF3-predicted structures align well (RMSD = 2.004). (B) Alignment of experimentally determined *Streptococcus pyogenes* Cas9 structure in complex with sgRNA and target DNA (orange) with the AF3-generated structure of OpenCRISPR-1[35] (blue, red). The AF3 prediction is of high confidence, as most residues have pLDDT > 90, pTM = 0.82 and iPTM = 0.92. The experimentally determined and AF3-predicted structures align well (RMSD = 2.114). (C) Alignment of experimentally determined *Streptococcus pyogenes* Cas9 structure in complex with sgRNA and target DNA (orange) with the AF3-generated structure of EvoCas9-1[36] (blue, red). The AF3 prediction is of high confidence, as most residues have pLDDT > 90, pTM = 0.84 and iPTM = 0.91. The experimentally determined and AF3-predicted structures align well (RMSD = 2.028).

therapy. Other IscB homologs are even smaller − IsrB, which lacks the HNH domain, is approximately 340 amino acids, while IshB, which lacks the RuvC domain, is approximately 180 amino acids. Although the original IscB displays low editing activity in human cells, protein engineering that enhances its DNA-binding affinity resulted in >30-fold increase in gene editing frequency.[21] TnpB, another member in the IS200/IS605 family that is widespread in nature, has been posited to be the ancestor of Cas12 based on bioinformatic prediction of the conserved RuvC-like active site within the TnpB sequence. Notably, the TnpB from *Deinococcus radiodurans* has been shown to work as a reprogrammable DNase within human cells.[22] Subsequently, a screen of 78 TnpB systems (64 of which were curated from all 107 members of the IS605 group annotated in the ISfinder database, along with additional *de novo* annotated systems) unveiled 33 TnpB proteins that could efficiently cleave plasmids in *E. coli*, among which five also demonstrated robust editing capabilities in human cells.[23] Interestingly, a group of larger eukaryotic proteins, Fanzors, which are evolutionarily related to TnpBs, were found during a systematic screening of transposable elements (TEs).[24] These proteins are encoded diversely in eukaryotic transposons, and can also be reprogrammed for human genome engineering applications.[25]

In addition to endonucleases like IscB, TnpB, and Fanzors, programmable editor systems capable of inserting, deleting or flipping large segments of DNA have also been discovered, although such activity has yet to be demonstrated in mammalian cells.[26−28] Unlike other systems that depend on endogenous DNA repair processes and/or the fusion of effector proteins, these RNA-guided recombinases of the IS110/IS1111 family are smaller, easier to deliver and potentially less immunogenic. More importantly, these recombinases could facilitate large DNA edits, such as inversions, duplications or translocations, which are challenging with current endonucleases. Engineering these systems for programmable, site-specific activity in mammalian cells would expand the toolkit for treating a wider range of genetic diseases that cannot be remedied through simple SNP editing or gene knockouts.

DESIGNING NEW EDITORS

Beyond mining sequencing data for naturally occurring gene editors, researchers are now envisioning the design of *de novo* editors with desired properties. This design objective is rooted in the correlation between protein structure and function, which is bolstered by the backdrop of recent advances in deep learning (DL) that has yielded improvements in protein structure prediction and design. Examples of new protein prediction tools include ESMFold,[29] OmegaFold[30] and Alphafold.[31,32] ESMFold utilizes transformer-based architectures and is adept at handling sequence data and capturing long-range dependencies in amino acid sequences. Meanwhile, OmegaFold and AlphaFold have achieved remarkable accuracy in protein structure prediction, with AlphaFold particularly noted for its use of attention mechanisms and DL techniques to model inter-residue distances and angles (Figure 2). However, these models are limited in their ability to predict disordered and flexible regions of proteins. Nonetheless, these advancements enable high-accuracy structure predictions, which facilitate structural evaluation of the designed proteins and lay a foundation for generative *de novo* protein designs. Protein design tools can be used to design proteins that have never been generated by nature (unconditional), or to improve existing proteins (conditional). Conditional design imposes specific constraints such as protein function, stability, molecular interactions, or environmental suitability, making it useful when the desired engineering outcome is well-defined. Unconditional design, however, imposes no such constraints and is used when the desired characteristics of the protein are not fully known, or when one is creating entirely new structures. There are two main strategies for protein design: structure-based and sequence-based, which can both be used for conditional and unconditional design. Structure-based protein design typically uses diffusion models, such as RoseTTAFold Diffusion (RFdiffusion)[33] and its derivatives

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

(e.g. RF diffusion All-Atom[34]), while sequence-based design typically uses large language models (LLMs).[35,36] For more detailed discussions pertaining to protein design with generative AI, we refer readers to previous reviews.[37,38] It is not difficult to envision the use of these models for the creation of gene-editing proteins with tailored functions, whether by structure- or sequence-based approaches.

Diffusion models have already been used to generate novel sequence-specific nucleic-acid binding proteins. Glasscock et al. optimized existing RFdiffusion methodology for designing sequence-specific DNA-binders,[39] creating specific helix-turn-helix (HTH) DNA-binding domains and using RIFdock to identify optimal variants. They employed LigandMPNN,[40] a variation of ProteinMPNN[41] to design high-affinity DNA binding proteins (DBPs) that could bind 7-mer DNA sequences, demonstrating functionality in bacterial and mammalian cells. In contrast, Zhou et al. designed DBPs with enzymatic activity,[42] creating 27 artificial Argonaute proteins, 24 of which display DNA cleavage activity, with 74% of these outperforming the template protein. Although these two studies have designed DBPs successfully, designing an RNA-guided *de novo* editor would be significantly more challenging due to the need for ternary interactions with both guide RNA and target DNA. However, with the development of newer tools like RoseTTAFoldNA,[43] which can predict protein-nucleic acid complexes more accurately, we anticipate that improved design of *de novo* gene editors using diffusion models will emerge.

Several recent examples utilize a different approach to generate *de novo* editors. LLMs leverage 2D sequence-based knowledge, and thus require less computational resources than 3D structure-based strategies (allowing for greater scalability and speed) and are less reliant on accurate structural data that has been traditionally onerous to obtain. An example of a LLM tool is Evo, a deep signal processing model that enables prediction and generation of tasks over multiple modalities,[36] including the design of CRISPR-Cas complexes. While these designs have not yet been empirically shown for activity in cells, the protein and RNA sequences have been predicted to form structures resembling their canonical counterparts in key enzymatic domains. Nguyen et al. also used Evo to design IS200/IS605 elements, including the programmable nuclease TnpB. Only 25.5% of the designed TnpBs were predicted to fold well when assessed by ESMFold pLDDT, highlighting both the potential and challenges of using sequence-based models for protein design. Separately, Ruffolo et al. employed LLMs trained on biological diversity at scale to design programmable gene editors.[35] Of the 209 Cas-like proteins selected for validation, their top hit, PF-CAS-182 (a.k.a. OpenCRISPR-1), displayed both lower off-target editing and higher editing activity than SpCas9, with indel rates of 55.7% versus 48.3% at on-target sites.

In conclusion, the development of AI tools for protein design and modeling holds great promise. While only a limited number of publicly available tools exist for designing nucleic-acid binding enzymes, recent breakthroughs herald an exponential trajectory toward radical design of effective gene editors. The ability to generate new proteins and enzymes raises new concerns about biosafety, as there is now the potential to generate pathogenic proteins. While this issue is beyond the scope of our review, we refer readers to other articles in which these are discussed in more detail.[36,46]

TRADITIONAL MEANS OF ENGINEERING EDITORS

Native gene editors often have low activity, especially in mammalian cells. This highlights the need to improve on-target efficiencies, while minimizing unintended off-target effects, before these editing systems could be deployed as research tools and therapeutics. Both rational design and random mutagenesis approaches have been employed to engineer the RNA and protein components.

In terms of engineering gRNA, several features have emerged as important: the protospacer adjacent motif (PAM),[47−50] gRNA sequence motifs,[49,51,52] overall nucleotide usage,[47,49−51] nucleotide composition in the seed region,[47,49,50,53] overall secondary structure,[49] GC content,[47] and overall length.[50,51] Such rules were determined by creating a pool of gRNAs that tile across multiple target sites, and then assessing corresponding cleavage activity of each gRNA, via downstream assays,[47] positive/negative screening, and/or high throughput sequencing.[48] Alternatively, these rules could also be investigated through the analysis of publicly available data sets to identify novel features.[49,50]

These efforts eventually led to the development of various sgRNA design tools, many of which employ ML to analyze the complex interplay of parameters described above. One of the few hypothesis-driven tools, which relies on experimentally informed, human-interpreted rules to design gRNAs, is CHOPCHOP.[54] Its first version was based on two parameters: GC content and whether the gRNA contains a G at position 20. Later versions incorporated more sophisticated rules, updated with the growing literature on parameters governing CRISPR activity, and added functionalities such as CRISPR activation/repression and RNA cleavage.[55,56] Other hypothesis-driven tools include E-CRISP[57] and GuideScan,[58] but ML tools are far more common.

Additionally, off-target prediction can be integrated into these computational tools. Experimentally, off-target activity is measured by performing targeted or whole genome sequencing, and examples include GUIDE-seq,[59] CIRCLE-seq,[60] SITE-seq[61] and DISCOVER-seq.[62] These data sets provide valuable input for training and validating off-target prediction models, which improves the utility of gRNA design tools.

Besides engineering gRNA, extensive efforts have also been directed at protein engineering, chiefly to increase specificity, on-target activity, and targeting range. To increase the specificity of Cas9, groups have utilized structure-guided rational design to minimize Cas9 activity on off-target sequences, by reducing nonspecific Cas9 interactions with DNA,[63] attenuating the helicase activity of Cas9,[64] and by raising the threshold for HNH nuclease conformational activation.[65] Besides rational design, others have utilized bacterial selection,[66,67] *in vivo* screening in yeast,[68] phage-assisted continuous evolution (PACE),[69] and directed evolution in *E. coli*[70] to identify Cas9 variants that have higher fidelity.

In contrast, to improve on-target efficiency of Cas9, strategies have sought to maximize Cas9 activity through the fusion of multiple nuclear localization signals,[71] optimizing codon usage,[72] increasing accessibility through chromatin remodelling,[73−75] and enhancing association between Cas9 and its DNA substrate.[76] Knowledge of structure, molecular interactions, and dynamics within the editor complexes together has informed the tuning of editor activity.

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

To increase targeting range, engineering efforts include developing SpCas9 that recognize non-NGG PAMs. The PAM refers to a 2−6 base-pair sequence located immediately adjacent to the target sequence that is complementary to the RNA guide, and is essential for the recognition and cleavage of DNA sequences by the Cas protein. While multiple Cas proteins may recognize a particular PAM sequence, each Cas protein only recognizes one PAM. For example, SpCas9 only recognizes and cleaves next to the NGG sequence. Hence, editing sequences that are not adjacent to NGG would require the usage of other Cas proteins that can recognize non-NGG PAMs. Examples of expanded-PAM Cas9 variants include SpCas9-VQR (recognizes NAG and NGAG PAMs[66]), SpCas9-EQR (recognizes NGCG PAMs[66]), xCas9s (recognize NG, GAA and GAT PAMs[69]), variants that recognize NRNH PAMs,[77] and a nearly PAMless SpRY variant (recognize NGN PAMs[78]). These variants were identified through various approaches, such as error-prone mutagenesis,[66] PACE,[69,77] phage-assisted non-continuous evolution (PANCE)[77] and structure-guided mutagenesis.[78] These studies have provided an expanded targeting range for flexibility in genome editing.

NOVEL MODALITIES

Most genetic diseases cannot be addressed with simple gene knockouts, requiring more advanced genome editing techniques. Novel modalities like base and prime editing have emerged, offering greater precision in genetic modifications.

Base editors allow the targeted conversion of specific base pairs without the induction of double-strand breaks. For instance, cytosine base editors (CBE) can convert C to T or G to A,[79] while adenine base editors (ABE)[80] convert A to I (which polymerases interpret as G). Recent base editors can also mediate C-to-G conversions.[81,82] Base editors-focused reviews provide a more thorough description of their structure and functions.[83,84]

Prime editing is another significant advancement. It uses a Cas9 nickase fused to an engineered reverse transcriptase, and a pegRNA that both locates the target site and provides the template for the desired edit. This method allows all 12 base substitutions, alleviates the less controllable and unpredictable mutagenic outcomes generated from the cell endogenous repair of Cas9-mediated DSBs, and is less confined by PAM requirements.[85] Strategies to optimize prime editing, such as using PACE to enhance the reverse transcriptase domain[86] have produced more efficient prime editor variants.

Other gene editing modalities have been developed and improved by combining different proteins and optimizing each component through screens, protein engineering and directed evolution. For example, find and cut-and-transfer (FiCAT) allows the targeted insertion of multikilobase DNA fragments into the genome.[87] The use of DNA-dependent DNA polymerases has also been used to improve the efficiency and versatility of prime editing. Researchers have replaced the reverse transcriptase (RT) in prime editors with the DNA-dependent DNA polymerase phi29[88] to enable increased editing efficiency for long sequences. Similarly, the click editor system uses a DNA polymerase fused to a gRNA-programmable nickase and an HUH endonuclease to enable substitutions, insertions and deletions.[89]

INCORPORATING AI/ML TO ENGINEER EDITORS

Despite the above-mentioned successes, the sheer number of possible gRNA variants and engineered Cas proteins creates an enormous search and design space that is impossible to interrogate through experimental means alone. ML and DL tools, by capturing the underlying patterns without explicitly identifying the relationship between features (target characteristics) and labels (experimental outcomes), can efficiently shortlist promising variants to be experimentally tested, thus minimizing resource wastage. Figure 3 presents a schematic of the relevant methods.

Predicting gRNA On-Target and Off-Target Activity. Given the many tools available, we will focus on the main ways the field has changed over time and some overarching themes and challenges that remain. For more comprehensive reviews that evaluate algorithmic aspects of machine and deep learning in predicting on- and off-target activity, we refer readers to previous reviews.[90−92]

ML tools, such as regression[93] and classification models,[47,94] were trained using early large sgRNA screening data sets. Over time, DL models, particularly convolutional neural networks (CNNs),[95−97] were introduced. Later studies expanded into hybrid models[98,99] and included more sophisticated strategies from natural language processing, such as attention-based architectures. DL also enabled the discovery of features through representation learning. However, different models sometimes yield contradictory results, highlighting the evergreen importance of experimental validation. For instance, DeepCRISPR[96] and CNN-SVR[100] found that uracils are disfavored at the four positions closest to the PAM, which is consistent with experimental data,[47] but C-RNNCrispr[101] favored the presence of uracil at the similar protospacer position 20.

Early on-target prediction models primarily use the nucleotide sequence of the protospacer sequence, the PAM and some flanking sequences, while off-target prediction models consider the sequence similarity between the gRNA and its target, with the general rule that off-target potential decreases with increasing base-pairing mismatch between the gRNA and off-target DNA. Various encoding methods, such as one-hot encoding and the more advanced loss-free encoding, impact model performance (with the latter outperforming previous methods in off-target prediction by 35%).[102] Additional features such as chromatin accessibility and epigenetic information[99] have also been incorporated. However, the choice of data set may introduce certain cell-type-specific biases into these models, given that epigenetic factors and chromatin accessibility can be cell-type specific. For instance, only partial correlations (0.462−0.752) were obtained when knockout efficiencies of gRNAs across different cell lines in the same species were compared.[91,103] More recent developments include incorporating physically informed features such as sequence context around potential off-target sites, the GC ratio, nucleosome positioning,[104] and even molecular dynamics of gRNA-DNA interactions,[105] showing a trend toward predictions that go beyond simple sequence context.

The underlying data set (and by corollary, the experimental design and labels) can affect the performance of these ML/DL tools. For on-target predictions, functional screens measure the phenotypic changes after gene editing, such as changes in cell surface markers[47] or drug resistance,[94,106] while other screens

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

quantify editing outcomes at a nucleotide level after introducing synthetic exogenous libraries of sgRNAs and target pairs.[107] While the latter approach yields more detailed information, such as indel rate and editing patterns, such predictions may not necessarily translate to phenotypic outcomes. Ultimately, models are more accurate for use cases that closely resemble their training data but may not be transferable to different experimental conditions. This is exemplified by the differences noted by Haeussler et al. when evaluating 20 data sets, specifically that the performance of the on-target efficiency model strongly depends on the promoter used for guide RNA transcription.[103] Despite these challenges, ML and DL models are increasingly being applied to other genome editing modalities like base editors,[108−110] prime editors[111] and RNA-targeting Cas proteins,[112] and one expects these conceptual frameworks to be extrapolated to the newer generation of nucleases and editors, such as IscB, TnpB, Fanzor, and insertion sequences (IS).

DL models are generally more powerful than traditional ML models given their ability to automatically learn features. However, their potential is currently limited by the relatively small sizes of sgRNA training data sets. Many public data sets contain only tens of thousands of guide sequences, insufficient for robust DL training. Some studies have tried to combine multiple existing data sets,[103] but the intrinsic differences in data collection lead to conventional ML models performing comparably or even better than DL models. For example, conventional ML models, including Azimuth 2.0,[48] perform well against DL models,[91] and CRISPR-GNL, based on Bayesian ridge regression, outperforms DeepCas9,[113] partly due to the use of feature selection. In addition, CRISPR sequences have fewer features (20−30nt) with which models can be trained on, compared to other deep learning applications, such as image or speech analysis, where millions of data points are used as features, possibly limiting the utility of DL for gene editing.

Researchers therefore elect to build multiple models in parallel before identifying the best model for their application.[114,115] Studies have also shown that using multiple ML models in combination can outperform individual predictive tools for off-target activity.[116]

One challenge with existing data sets, especially for off-target prediction, is the imbalance between positive (off-target) and negative (nonoff-target) samples. This imbalance can lead models to overlearn from the majority class, reducing accuracy for minority samples. To address this, Chuai et al. pretrained their model with a vast data set of unlabeled sgRNAs (0.68 billion) through deep unsupervised representation learning,[96] before refining it with a supervised neural network using labeled sgRNAs. Their platform, DeepCRISPR, also uses a bootstrapping sampling to balance the data and incorporates epigenetic features from 13 cell types. Similarly, Zhang et al. improved their DL-CRISPR model by quadrupling the number of positive samples through a technique borrowed from image classification, rotating the original feature matrix to rebalance their training data.[117] These methods proved crucial for improving DL model performance, despite the challenges of small and imbalanced data sets.

Supporting Mutagenesis Screens and Directed Evolution. Conventional protein engineering relies on rational design and random mutagenesis. Rational design is traditionally limited by the availability of protein structures and data on beneficial mutations, while random mutagenesis is laborious for large proteins like Cas nucleases. Epistatic enhancement to the proteins, which is conferred by combinations of multiple mutations, is also difficult to access with random mutagenesis approaches. Directed evolution circumvents some of these challenges by using multiple iterative rounds of mutation and selection to optimize protein function. However, in cases where mutations interact in nonadditive ways (epistatic effects), DE may lead to the selection of mutants trapped in local, rather than global, optima.[118]

ML offers a solution to reduce experimental burden in protein engineering. In a machine learning-assisted directed evolution (MLDE) framework,[118] information from activity screens can guide the selection of new variants through iterative test-learn-design cycles, narrowing the search space more efficiently and minimizing the need for exhaustive experimental screening.

In protein variant libraries, most variants have no or low activity, limiting the overall yield from random sampling. ML models will benefit from selection strategies (zero-shot predictions) that prioritise functional variants.[119] The choice of ML model will also influence prediction reliability, and should be selected based on factors such as the size of the training data set and desired accuracy.[118] Similarly, model accuracy can be affected by encoding methods, which refer to how sequences are mapped to numerical representations used to train, test and run ML models. For example, when compared to traditional one-hot encoding, which indiscriminately assigns numbers to amino acids, the use of Georgiev representation[120] and learned embeddings from the MSA Transformer,[121] which consider physicochemical and evolutionary properties, improved MLDE outcomes.[119] In the context of RNA-guided nucleases, the choice of sgRNAs also plays a role in the reliability of predictions by the ML model.[122]

For Cas nucleases, MLDE has been successfully used to identify KKH-SaCas9 variants with increased activity.[122] After validating the MLDE model[119,123] for predicting SpCas9 activity using a previous high-throughput data set,[124] leveraging only 20% of the input data can accurately nominate top-performing variants in downstream predictions, which reduces the screening burden by 80%.[122] This shows that ML can enhance mutagenesis screens and directed evolution, enabling the exploration of larger search spaces with fewer experiments. Although the field is currently partially limited by the lack of large, consistent and unbiased data sets, we are optimistic that newer tools with improved predictions could significantly advance genome editing and gene therapy by prioritising high-activity proteins and gRNAs with fewer off-target effects.

REDUCING IMMUNOGENICITY

In view of clinical applications, ML and DL tools are also increasingly employed to reduce the immunogenicity of genome-editing proteins like Cas nucleases. Immunogenicity can pose safety risks and reduce the efficacy of gene editing therapies. Despite initial assumptions that transient expression of these nucleases within cells could limit immune reactions, studies have shown otherwise. For instance, anti-SpCas9 antibodies were observed in mice 14 days after adenoviral delivery,[125] and Cas9 expression itself, regardless of delivery method, induces an immune response.[126] Pre-existing antibodies and T-cells reactive to SaCas9 and SpCas9 have also

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

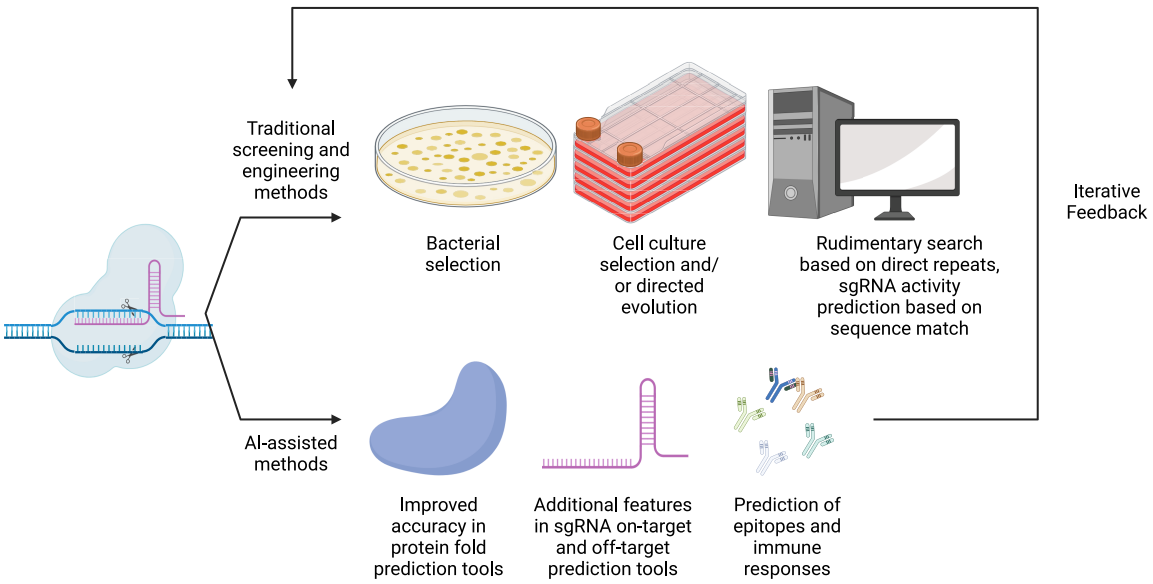


Figure 3. Traditional screening and engineering methods and AI-assisted methods both provide valuable information for improving editor function, including via the improvement of sgRNA activity, protein function, and the reduction of immunogenicity. Created with BioRender.com.

been found in human blood samples at nontrivial prevalences,[127−130] which suggests that the majority of the human population might mount potent adverse reactions to these editors. There is hence a pressing need to reduce the immunogenicity of these gene editors for safer therapeutic applications. Here we will only discuss Synthetic Biology-informed engineering of Cas proteins to mask immunogenic epitopes. For an in-depth understanding of pre-existing immunity to Cas9 and the other strategies to evade such immunity, we refer readers to other reviews.[131−133]

One approach to minimize immunogenicity is through epitope masking, which involves identifying immunogenic peptide sequences and modifying them to prevent recognition by antibodies or T-cells. Experimental methods, such as phage-immunoprecipitation sequencing (PhIP-seq),[134] can map epitopes, but these processes are often costly. Therefore, a plethora of different epitope prediction software, such as ElliPro,[135] SEPPA 3.0,[136] epitope3D,[137] DiscoTope-3.0,[138] many utilizing ML and DL, have been developed. The next generation of prediction tools take advantage of advances in protein language models to outperform previous models. For example, the integration of ESM-2 numerical representations[29] into BepiPred-3.0[139] improved linear and conformational epitope prediction. These traditional and newer AI/ML computational tools have been used to predict binding to Cas9 epitopes. For example, Shen et al. predicted human B-cell and $CD8^+$ T-cell epitopes in SaCas9, using DiscoTope2.0[140] and the Immune Epitope Database (IEDB) consensus MHC-binding prediction algorithm,[48] and validated these candidates through immunological assays. In contrast, Ferdosi et al. identified two T-cell epitopes for SpCas9[141] using ANN-Hydro, a T-cell epitope prediction model based on a feed-forward artificial neural network,[142] and showed that mutations at these sites reduced Cas9 immunogenicity without compromising Cas9 activity. These demonstrations suggest emerging opportunities in engineering gene editors in properties beyond on-target efficiency and off-target avoidance (specificity). As these properties, such as immunogenicity, tend to be more onerous to examine functionally, the resultant sparseness of data sets limits the training of high-performing ML/DL models. As we harness the capability and scalability of Synthetic Biology toward engineering and evaluating gene editors, one can envision that the wealth of generated data will fuel more sophisticated computational models suited for this task.

CONCLUSION

The rapidly advancing field of synthetic biology presents transformative opportunities for discovering, designing and engineering next-generation gene-editing tools to overcome key challenges such as efficiency, specificity, and immunogenicity. By mining naturally occurring systems and employing conventional means of protein engineering, such as rational design, saturation mutagenesis and directed evolution, significant progress has been made in the field of gene editing. A prime example of such success is the discovery and optimization of CRISPR-Cas systems, which has been successfully adapted for efficient gene-editing in mammalian cells. Subsequently, various higher-activity, high-specificity, and/or expanded-PAM variants have been generated, along with various inventive modalities such as base editors and prime editors. These innovations highlight the dramatic advancements made in enhancing the precision and versatility of gene-editing tools.

Moving forward, the integration of AI/ML will revolutionize future gene editor discovery and engineering by reducing the cost and time required for such efforts. AI/ML also has the potential to allow the design of *de novo* editors with unprecedented efficiency, specificity and safety, while introducing novel editing capabilities tailored to specific therapeutic applications. The convergence of synthetic biology and AI/ML, underpinned by training on expansive and high-quality data sets, could unlock the full potential of gene-editing technologies, redefining the landscape of gene therapy and personalized medicine and transforming healthcare on a global scale.

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

AUTHOR INFORMATION

Corresponding Author

Wei Leong Chew − *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore; Synthetic Biology Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117596, Singapore;* orcid.org/0000-0002-4774-7959; Email: chewwl@gis.a-star.edu.sg

Authors

Bing Shao Chia − *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore;* orcid.org/0000-0003-0101-5805

Yu Fen Samantha Seah − *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore;* orcid.org/0000-0002-8589-8459

Bolun Wang − *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore*

Kimberle Shen − *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore*

Diya Srivastava − *Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.4c00686

Author Contributions

#B.S.C. and Y.F.S.S. contributed equally. B.S.C.: Investigation, Writing−original draft, review and editing. Y.F.S.S.: Investigation, Writing−original draft, review and editing. B.W.: Investigation, Writing−original draft, review and editing. K.S.: Investigation, Writing−original draft, review and editing. D.S.: Investigation, Writing−original draft, review and editing. W.L.C.: Funding acquisition, Conceptualization, Project administration, Supervision, Writing−review and editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

W.L.C. was supported by Agency for Science, Technology and Research (A*STAR), the A*STAR MTC Individual Research Grant 242 K2115, the National Medical Research Council (NMRC) OFIRG24jul-0096, and the Ministry of Health Programme for Research in Epidemic Preparedness and REsponse (PREPARE) award PREPARE-OC-VT-2024-008.

REFERENCES

(1) Liu, G.; Lin, Q.; Jin, S.; Gao, C. The CRISPR-Cas Toolbox and Gene Editing Technologies. *Mol. Cell* **2022**, *82* (2), 333−347.

(2) Wang, J. Y.; Doudna, J. A. CRISPR Technology: A Decade of Genome Editing Is Only the Beginning. *Science* **2023**, *379* (6629), No. eadd8643.

(3) Villiger, L.; Joung, J.; Koblan, L.; Weissman, J.; Abudayyeh, O. O.; Gootenberg, J. S. CRISPR Technologies for Genome, Epigenome and Transcriptome Editing. *Nat. Rev. Mol. Cell Biol.* **2024**, *25* (6), 464−487.

(4) Ishino, Y.; Krupovic, M.; Forterre, P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J. Bacteriol.* **2018**, *200* (7), No. e00580-17.

(5) Makarova, K. S.; Grishin, N. V.; Shabalina, S. A.; Wolf, Y. I.; Koonin, E. V. A Putative RNA-Interference-Based Immune System in Prokaryotes: Computational Analysis of the Predicted Enzymatic Machinery, Functional Analogies with Eukaryotic RNAi, and Hypothetical Mechanisms of Action. *Biol. Direct* **2006**, *1* (1), 7.

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

(7) Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; Zhang, F. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **2013**, *339* (6121), 819−823.

(8) Mali, P.; Yang, L.; Esvelt, K. M.; Aach, J.; Guell, M.; DiCarlo, J. E.; Norville, J. E.; Church, G. M. RNA-Guided Human Genome Engineering via Cas9. *Science* **2013**, *339* (6121), 823−826.

(9) Bland, C.; Ramsey, T. L.; Sabree, F.; Lowe, M.; Brown, K.; Kyrpides, N. C.; Hugenholtz, P. CRISPR Recognition Tool (CRT): A Tool for Automatic Detection of Clustered Regularly Interspaced Palindromic Repeats. *BMC Bioinformatics* **2007**, *8* (1), 209.

(10) Edgar, R. C. PILER-CR: Fast and Accurate Identification of CRISPR Repeats. *BMC Bioinformatics* **2007**, *8* (1), 18.

(11) Grissa, I.; Vergnaud, G.; Pourcel, C. CRISPRFinder: A Web Tool to Identify Clustered Regularly Interspaced Short Palindromic Repeats. *Nucleic Acids Res.* **2007**, *35* (Web Server), W52−W57.

(12) Couvin, D.; Bernheim, A.; Toffano-Nioche, C.; Touchon, M.; Michalik, J.; Néron, B.; Rocha, E. P. C.; Vergnaud, G.; Gautheret, D.; Pourcel, C. CRISPRCasFinder, an Update of CRISRFinder, Includes a Portable Version, Enhanced Performance and Integrates Search for Cas Proteins. *Nucleic Acids Res.* **2018**, *46* (W1), W246−W251.

(13) Abby, S. S.; Néron, B.; Ménager, H.; Touchon, M.; Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS One* **2014**, *9* (10), No. e110726.

(14) Russel, J.; Pinilla-Redondo, R.; Mayo-Muñoz, D.; Shah, S. A.; Sørensen, S. J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J.* **2020**, *3* (6), 462−469.

(15) Altae-Tran, H.; Kannan, S.; Suberski, A. J.; Mears, K. S.; Demircioglu, F. E.; Moeller, L.; Kocalar, S.; Oshiro, R.; Makarova, K. S.; Macrae, R. K.; Koonin, E. V.; Zhang, F. Uncovering the Functional Diversity of Rare CRISPR-Cas Systems with Deep Terascale Clustering. *Science* **2023**, *382* (6673), No. eadi1910.

(16) Adler, B. A.; Trinidad, M. I.; Bellieny-Rabelo, D.; Zhang, E.; Karp, H. M.; Skopintsev, P.; Thornton, B. W.; Weissman, R. F.; Yoon, P. H.; Chen, L.; Hessler, T.; Eggers, A. R.; Colognori, D.; Boger, R.; Doherty, E. E.; Tsuchida, C. A.; Tran, R. V.; Hofman, L.; Shi, H.; Wasko, K. M.; Zhou, Z.; Xia, C.; Al-Shimary, M. J.; Patel, J. R.; Thomas, V. C. J. X.; Pattali, R.; Kan, M. J.; Vardapetyan, A.; Yang, A.; Lahiri, A.; Maxwell, M. F.; Murdock, A. G.; Ramit, G. C.; Henderson, H. R.; Calvert, R. W.; Bamert, R. S.; Knott, G. J.; Lapinaite, A.; Pausch, P.; Cofsky, J. C.; Sontheimer, E. J.; Wiedenheft, B.; Fineran, P. C.; Brouns, S. J. J.; Sashital, D. G.; Thomas, B. C.; Brown, C. T.; Goltsman, D. S. A.; Barrangou, R.; Siksnys, V.; Banfield, J. F.; Savage, D. F.; Doudna, J. A. CasPEDIA Database: A Functional Classification System for Class 2 CRISPR-Cas Enzymes. *Nucleic Acids Res.* **2024**, *52* (D1), D590−D596.

(17) Siguier, P.; Perochon, J.; Lestrade, L.; Mahillon, J.; Chandler, M. ISfinder: The Reference Centre for Bacterial Insertion Sequences. *Nucleic Acids Res.* **2006**, *34* (90001), D32−D36.

(18) Minh, B. Q.; Schmidt, H. A.; Chernomor, O.; Schrempf, D.; Woodhams, M. D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37* (5), 1530−1534.

(19) Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v6: Recent Updates to the Phylogenetic Tree Display and Annotation Tool. *Nucleic Acids Res.* **2024**, *52* (W1), W78−W82.

(20) Altae-Tran, H.; Kannan, S.; Demircioglu, F. E.; Oshiro, R.; Nety, S. P.; McKay, L. J.; Dlakić, M.; Inskeep, W. P.; Makarova, K. S.;

Macrae, R. K.; Koonin, E. V.; Zhang, F. The Widespread IS200/IS605 Transposon Family Encodes Diverse Programmable RNA-Guided Endonucleases. *Science* **2021**, *374* (6563), 57−65.
(21) Yan, H.; Tan, X.; Zou, S.; Sun, Y.; Ke, A.; Tang, W. Assessing and Engineering the IscB-ωRNA System for Programmed Genome Editing. *Nat. Chem. Biol.* **2024**, *20*, 1617.
(22) Karvelis, T.; Druteika, G.; Bigelyte, G.; Budre, K.; Zedaveinyte, R.; Silanskas, A.; Kazlauskas, D.; Venclovas, Č.; Siksnys, V. Transposon-Associated TnpB Is a Programmable RNA-Guided DNA Endonuclease. *Nature* **2021**, *599* (7886), 692−696.
(23) Xiang, G.; Li, Y.; Sun, J.; Huo, Y.; Cao, S.; Cao, Y.; Guo, Y.; Yang, L.; Cai, Y.; Zhang, Y. E.; Wang, H. Evolutionary Mining and Functional Characterization of TnpB Nucleases Identify Efficient Miniature Genome Editors. *Nat. Biotechnol.* **2024**, *42* (5), 745−757.
(24) Bao, W.; Kojima, K. K.; Kohany, O. Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mob. DNA* **2015**, *6* (1), 11.
(25) Saito, M.; Xu, P.; Faure, G.; Maguire, S.; Kannan, S.; Altae-Tran, H.; Vo, S.; Desimone, A.; Macrae, R. K.; Zhang, F. Fanzor Is a Eukaryotic Programmable RNA-Guided Endonuclease. *Nature* **2023**, *620* (7974), 660−668.
(26) Durrant, M. G.; Perry, N. T.; Pai, J. J.; Jangid, A. R.; Athukoralage, J. S.; Hiraizumi, M.; McSpedon, J. P.; Pawluk, A.; Nishimasu, H.; Konermann, S.; Hsu, P. D. Bridge RNAs Direct Programmable Recombination of Target and Donor DNA. *Nature* **2024**, *630* (8018), 984−993.
(27) Siddiquee, R.; Pong, C. H.; Hall, R. M.; Ataide, S. F. A Programmable seekRNA Guides Target Selection by IS1111 and IS110 Type Insertion Sequences. *Nat. Commun.* **2024**, *15* (1), 5235.
(28) Hiraizumi, M.; Perry, N. T.; Durrant, M. G.; Soma, T.; Nagahata, N.; Okazaki, S.; Athukoralage, J. S.; Isayama, Y.; Pai, J. J.; Pawluk, A.; Konermann, S.; Yamashita, K.; Hsu, P. D.; Nishimasu, H. Structural Mechanism of Bridge RNA-Guided Recombination. *Nature* **2024**, *630* (8018), 994−1002.
(29) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.
(30) Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; Ma, J.; Peng, J. High-Resolution *de Novo* Structure Prediction from Primary Sequence. *bioRxiv* **2022**. DOI: 10.1101/2022.07.21.500999.
(31) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577* (7792), 706−710.
(32) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(33) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089−1100.
(34) Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R.; McHugh, R.; Vafeados, D.; Li, X.; Sutherland, G. A.; Hitchcock, A.; Hunter, C. N.; Kang, A.; Brackenbrough, E.; Bera, A. K.; Baek, M.; DiMaio, F.; Baker, D. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *Science* **2024**, *384* (6693), No. eadl2528.
(35) Ruffolo, J. A.; Nayfach, S.; Gallagher, J.; Bhatnagar, A.; Beazer, J.; Hussain, R.; Russ, J.; Yip, J.; Hill, E.; Pacesa, M.; Meeske, A. J.; Cameron, P.; Madani, A. Design of Highly Functional Genome Editors by Modeling the Universe of CRISPR-Cas Sequences. *bioRxiv* **2024**. DOI: 10.1101/2024.04.22.590591.
(36) Nguyen, E.; Poli, M.; Durrant, M. G.; Kang, B.; Katrekar, D.; Li, D. B.; Bartie, L. J.; Thomas, A. W.; King, S. H.; Brixi, G.; Sullivan, J.; Ng, M. Y.; Lewis, A.; Lou, A.; Ermon, S.; Baccus, S. A.; Hernandez-Boussard, T.; Ré, C.; Hsu, P. D.; Hie, B. L. Sequence Modeling and Design from Molecular to Genome Scale with Evo. *Science* **2024**, *386* (6723), No. eado9336.
(37) Winnifrith, A.; Outeiral, C.; Hie, B. L. Generative Artificial Intelligence for de Novo Protein Design. *Curr. Opin. Struct. Biol.* **2024**, *86*, 102794.
(38) Kortemme, T. De Novo Protein Design—From New Structures to Programmable Functions. *Cell* **2024**, *187* (3), 526−544.
(39) Glasscock, C. J.; Pecoraro, R.; McHugh, R.; Doyle, L. A.; Chen, W.; Boivin, O.; Lonnquist, B.; Na, E.; Politanska, Y.; Haddox, H. K.; Cox, D.; Norn, C.; Coventry, B.; Goreshnik, I.; Vafeados, D.; Lee, G. R.; Gordan, R.; Stoddard, B. L.; DiMaio, F.; Baker, D. Computational Design of Sequence-Specific DNA-Binding Proteins. *bioRxiv* **2023**. DOI: 10.1101/2023.09.20.558720.
(40) Dauparas, J.; Lee, G. R.; Pecoraro, R.; An, L.; Anishchenko, I.; Glasscock, C.; Baker, D. Atomic Context-Conditioned Protein Sequence Design Using LigandMPNN. *bioRxiv* **2023**. DOI: 10.1101/2023.12.22.573103.
(41) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; de Haas, R. J.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King, N. P.; Baker, D. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **2022**, *378*, 49−56.
(42) Zhou, B.; Zheng, L.; Wu, B.; Yi, K.; Zhong, B.; Tan, Y.; Liu, Q.; Liò, P.; Hong, L. A Conditional Protein Diffusion Model Generates Artificial Programmable Endonuclease Sequences with Enhanced Activity. *bioRxiv* **2024**. DOI: 10.1101/2023.08.10.552783.
(43) Baek, M.; McHugh, R.; Anishchenko, I.; Jiang, H.; Baker, D.; DiMaio, F. Accurate Prediction of Protein-Nucleic Acid Complexes Using RoseTTAFoldNA. *Nat. Methods* **2024**, *21* (1), 117−121.
(44) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493−500.
(45) Nishimasu, H.; Ran, F. A.; Hsu, P. D.; Konermann, S.; Shehata, S. I.; Dohmae, N.; Ishitani, R.; Zhang, F.; Nureki, O. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **2014**, *156* (5), 935−949.
(46) Baker, D.; Church, G. Protein Design Meets Biosecurity. *Science* **2024**, *383* (6681), 349−349.
(47) 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. Rational Design of Highly Active sgRNAs for CRISPR-Cas9-Mediated Gene Inactivation. *Nat. Biotechnol.* **2014**, *32* (12), 1262−1267.
(48) 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. Optimized sgRNA Design to

Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34* (2), 184−191.
(49) Wong, N.; Liu, W.; Wang, X. WU-CRISPR: Characteristics of Functional Guide RNAs for the CRISPR/Cas9 System. *Genome Biol.* **2015**, *16* (1), 218.
(50) 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. Sequence Determinants of Improved CRISPR sgRNA Design. *Genome Res.* **2015**, *25* (8), 1147−1157.
(51) Moreno-Mateos, M. A.; Vejnar, C. E.; Beaudoin, J.-D.; Fernandez, J. P.; Mis, E. K.; Khokha, M. K.; Giraldez, A. J. CRISPRscan: Designing Highly Efficient sgRNAs for CRISPR-Cas9 Targeting in Vivo. *Nat. Methods* **2015**, *12* (10), 982−988.
(52) Graf, R.; Li, X.; Chu, V. T.; Rajewsky, K. sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. *Cell Rep.* **2019**, *26* (5), 1098−1103.
(53) Chari, R.; Mali, P.; Moosburner, M.; Church, G. M. Unraveling CRISPR-Cas9 Genome Engineering Parameters via a Library-on-Library Approach. *Nat. Methods* **2015**, *12* (9), 823−826.
(54) Montague, T. G.; Cruz, J. M.; Gagnon, J. A.; Church, G. M.; Valen, E. CHOPCHOP: A CRISPR/Cas9 and TALEN Web Tool for Genome Editing. *Nucleic Acids Res.* **2014**, *42* (W1), W401.
(55) Labun, K.; Montague, T. G.; Gagnon, J. A.; Thyme, S. B.; Valen, E. CHOPCHOP v2: A Web Tool for the next Generation of CRISPR Genome Engineering. *Nucleic Acids Res.* **2016**, *44* (W1), W272−W276.
(56) Labun, K.; Montague, T. G.; Krause, M.; Torres Cleuren, Y. N.; Tjeldnes, H.; Valen, E. CHOPCHOP v3: Expanding the CRISPR Web Toolbox beyond Genome Editing. *Nucleic Acids Res.* **2019**, *47* (W1), W171−W174.
(57) Heigwer, F.; Kerr, G.; Boutros, M. E-CRISP: Fast CRISPR Target Site Identification. *Nat. Methods* **2014**, *11* (2), 122−123.
(58) Perez, A. R.; Pritykin, Y.; Vidigal, J. A.; Chhangawala, S.; Zamparo, L.; Leslie, C. S.; Ventura, A. GuideScan Software for Improved Single and Paired CRISPR Guide RNA Design. *Nat. Biotechnol.* **2017**, *35* (4), 347−349.
(59) 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. GUIDE-Seq Enables Genome-Wide Profiling of off-Target Cleavage by CRISPR-Cas Nucleases. *Nat. Biotechnol.* **2015**, *33* (2), 187−197.
(60) Tsai, S. Q.; Nguyen, N. T.; Malagon-Lopez, J.; Topkar, V. V.; Aryee, M. J.; Joung, J. K. CIRCLE-Seq: A Highly Sensitive in Vitro Screen for Genome-Wide CRISPR-Cas9 Nuclease off-Targets. *Nat. Methods* **2017**, *14* (6), 607−614.
(61) 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.; Young, J. K.; May, A. P. Mapping the Genomic Landscape of CRISPR-Cas9 Cleavage. *Nat. Methods* **2017**, *14* (6), 600−606.
(62) Wienert, B.; Wyman, S. K.; Yeh, C. D.; Conklin, B. R.; Corn, J. E. CRISPR Off-Target Detection with DISCOVER-Seq. *Nat. Protoc.* **2020**, *15* (5), 1775−1799.
(63) Kleinstiver, B. P.; Pattanayak, V.; Prew, M. S.; Tsai, S. Q.; Nguyen, N. T.; Zheng, Z.; Joung, J. K. High-Fidelity CRISPR-Cas9 Nucleases with No Detectable Genome-Wide off-Target Effects. *Nature* **2016**, *529* (7587), 490−495.
(64) Slaymaker, I. M.; Gao, L.; Zetsche, B.; Scott, D. A.; Yan, W. X.; Zhang, F. Rationally Engineered Cas9 Nucleases with Improved Specificity. *Science* **2016**, *351* (6268), 84−88.
(65) Chen, J. S.; Dagdas, Y. S.; Kleinstiver, B. P.; Welch, M. M.; Sousa, A. A.; Harrington, L. B.; Sternberg, S. H.; Joung, J. K.; Yildiz, A.; Doudna, J. A. Enhanced Proofreading Governs CRISPR-Cas9 Targeting Accuracy. *Nature* **2017**, *550* (7676), 407−410.
(66) Kleinstiver, B. P.; Prew, M. S.; Tsai, S. Q.; Topkar, V. V.; Nguyen, N. T.; Zheng, Z.; Gonzales, A. P. W.; Li, Z.; Peterson, R. T.; Yeh, J. R. J.; Aryee, M. J.; Joung, J. K. Engineered CRISPR-Cas9 Nucleases with Altered PAM Specificities. *Nature* **2015**, *523* (7561), 481−485.
(67) 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. A High-Fidelity Cas9Mutant Delivered as a Ribonucleoprotein Complex Enables Efficient Gene Editing in Human Hematopoietic Stem and Progenitor Cells. *Nat. Med.* **2018**, *24* (8), 1216−1224.
(68) Casini, A.; Olivieri, M.; Petris, G.; Montagna, C.; Reginato, G.; Maule, G.; Lorenzin, F.; Prandi, D.; Romanel, A.; Demichelis, F.; Inga, A.; Cereseto, A. A Highly Specific SpCas9 Variant Is Identified by in Vivo Screening in Yeast. *Nat. Biotechnol.* **2018**, *36* (3), 265−271.
(69) 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. Evolved Cas9 Variants with Broad PAM Compatibility and High DNA Specificity. *Nature* **2018**, *556* (7699), 57−63.
(70) Lee, J. K.; Jeong, E.; Lee, J.; Jung, M.; Shin, E.; Kim, Y.; Lee, K.; Jung, I.; Kim, D.; Kim, S.; Kim, J.-S. Directed Evolution of CRISPR-Cas9 to Increase Its Specificity. *Nat. Commun.* **2018**, *9* (1), 3048.
(71) Wu, Y.; Zeng, J.; Roscoe, B. P.; Liu, P.; Yao, Q.; Lazzarotto, C. R.; Clement, K.; Cole, M. A.; Luk, K.; Baricordi, C.; Shen, A. H.; Ren, C.; Esrick, E. B.; Manis, J. P.; Dorfman, D. M.; Williams, D. A.; Biffi, A.; Brugnara, C.; Biasco, L.; Brendel, C.; Pinello, L.; Tsai, S. Q.; Wolfe, S. A.; Bauer, D. E. Highly Efficient Therapeutic Gene Editing of Human Hematopoietic Stem Cells. *Nat. Med.* **2019**, *25* (5), 776−783.
(72) Cheng, B.; Groshong, T.; Madejski, I.; Sweeney, R.; De, L. Codon Optimization of Gene Editing CRISPR-SaCas9 Augments Protein Expression in Human Liver Cells to Boost *in Vivo* Therapeutic Application. *FASEB J.* **2019**, *33* (S1), 495.3−495.3.
(73) Horlbeck, M. A.; Gilbert, L. A.; Villalta, J. E.; Adamson, B.; Pak, R. A.; Chen, Y.; Fields, A. P.; Park, C. Y.; Corn, J. E.; Kampmann, M.; Weissman, J. S. Compact and Highly Active Next-Generation Libraries for CRISPR-Mediated Gene Repression and Activation. *eLife* **2016**, *5*, No. e19760.
(74) Ding, X.; Seebeck, T.; Feng, Y.; Jiang, Y.; Davis, G. D.; Chen, F. Improving CRISPR-Cas9 Genome Editing Efficiency by Fusion with Chromatin-Modulating Peptides. *CRISPR J.* **2019**, *2* (1), 51−63.
(75) Liu, G.; Yin, K.; Zhang, Q.; Gao, C.; Qiu, J.-L. Modulating Chromatin Accessibility by Transactivation and Targeting Proximal dsgRNAs Enhances Cas9 Editing Efficiency in Vivo. *Genome Biol.* **2019**, *20* (1), 145.
(76) Yin, S.; Zhang, M.; Liu, Y.; Sun, X.; Guan, Y.; Chen, X.; Yang, L.; Huo, Y.; Yang, J.; Zhang, X.; Han, H.; Zhang, J.; Xiao, M.-M.; Liu, M.; Hu, J.; Wang, L.; Li, D. Engineering of Efficiency-Enhanced Cas9 and Base Editors with Improved Gene Therapy Efficacies. *Mol. Ther.* **2023**, *31* (3), 744−759.
(77) Miller, S. M.; Wang, T.; Randolph, P. B.; Arbab, M.; Shen, M. W.; Huang, T. P.; Matuszek, Z.; Newby, G. A.; Rees, H. A.; Liu, D. R. Continuous Evolution of SpCas9 Variants Compatible with Non-G PAMs. *Nat. Biotechnol.* **2020**, *38* (4), 471−481.
(78) Walton, R. T.; Christie, K. A.; Whittaker, M. N.; Kleinstiver, B. P. Unconstrained Genome Targeting with Near-PAMless Engineered CRISPR-Cas9 Variants. *Science* **2020**, *368* (6488), 290−296.
(79) Komor, A. C.; Kim, Y. B.; Packer, M. S.; Zuris, J. A.; Liu, D. R. Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage. *Nature* **2016**, *533* (7603), 420−424.
(80) Gaudelli, N. M.; Komor, A. C.; Rees, H. A.; Packer, M. S.; Badran, A. H.; Bryson, D. I.; Liu, D. R. Programmable Base Editing of A•T to G•C in Genomic DNA without DNA Cleavage. *Nature* **2017**, *551* (7681), 464−471.
(81) 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. Efficient C•G-to-G•C Base Editors Developed Using

https://doi.org/10.1021/acssynbio.4c00686
ACS Synth. Biol. 2025, 14, 636−647

CRISPRi Screens, Target-Library Analysis, and Machine Learning. *Nat. Biotechnol.* **2021**, *39* (11), 1414−1425.
(82) Chen, L.; Park, J. E.; Paa, P.; Rajakumar, P. D.; Prekop, H.-T.; Chew, Y. T.; Manivannan, S. N.; Chew, W. L. Programmable C:G to G:C Genome Editing with CRISPR-Cas9-Directed Base Excision Repair Proteins. *Nat. Commun.* **2021**, *12* (1), 1384.
(83) Rees, H. A.; Liu, D. R. Base Editing: Precision Chemistry on the Genome and Transcriptome of Living Cells. *Nat. Rev. Genet.* **2018**, *19* (12), 770−788.
(84) Porto, E. M.; Komor, A. C.; Slaymaker, I. M.; Yeo, G. W. Base Editing: Advances and Therapeutic Opportunities. *Nat. Rev. Drug Discovery* **2020**, *19* (12), 839−859.
(85) 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. Search-and-Replace Genome Editing without Double-Strand Breaks or Donor DNA. *Nature* **2019**, *576* (7785), 149−157.
(86) Doman, J. L.; Pandey, S.; Neugebauer, M. E.; An, M.; Davis, J. R.; Randolph, P. B.; McElroy, A.; Gao, X. D.; Raguram, A.; Richter, M. F.; Everette, K. A.; Banskota, S.; Tian, K.; Tao, Y. A.; Tolar, J.; Osborn, M. J.; Liu, D. R. Phage-Assisted Evolution and Protein Engineering Yield Compact, Efficient Prime Editors. *Cell* **2023**, *186* (18), 3983−4002.
(87) Pallarès-Masmitjà, M.; Ivančić, D.; Mir-Pedrol, J.; Jaraba-Wallace, J.; Tagliani, T.; Oliva, B.; Rahmeh, A.; Sánchez-Mejías, A.; Güell, M. Find and Cut-and-Transfer (FiCAT) Mammalian Genome Engineering. *Nat. Commun.* **2021**, *12* (1), 7071.
(88) Liu, B.; Dong, X.; Zheng, C.; Keener, D.; Chen, Z.; Cheng, H.; Watts, J. K.; Xue, W.; Sontheimer, E. J. Targeted Genome Editing with a DNA-Dependent DNA Polymerase and Exogenous DNA-Containing Templates. *Nat. Biotechnol.* **2024**, *42* (7), 1039−1045.
(89) Ferreira Da Silva, J.; Tou, C. J.; King, E. M.; Eller, M. L.; Rufino-Ramos, D.; Ma, L.; Cromwell, C. R.; Metovic, J.; Benning, F. M. C.; Chao, L. H.; Eichler, F. S.; Kleinstiver, B. P. Click Editing Enables Programmable Genome Writing Using DNA Polymerases and HUH Endonucleases. *Nat. Biotechnol.* **2024**, DOI: 10.1038/s41587-024-02324-x.
(90) Lee, M. Deep Learning in CRISPR-Cas Systems: A Review of Recent Studies. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1226182.
(91) Konstantakos, V.; Nentidis, A.; Krithara, A.; Paliouras, G. CRISPR-Cas9 gRNA Efficiency Prediction: An Overview of Predictive Tools and the Role of Deep Learning. *Nucleic Acids Res.* **2022**, *50* (7), 3616−3637.
(92) Sherkatghanad, Z.; Abdar, M.; Charlier, J.; Makarenkov, V. Using Traditional Machine Learning and Deep Learning Methods for On- and off-Target Prediction in CRISPR/Cas9: A Review. *Brief. Bioinform.* **2023**, *24* (3), bbad131.
(93) Fusi, N.; Smith, I.; Doench, J.; Listgarten, J. *In Silico* Predictive Modeling of CRISPR/Cas9 Guide Efficiency. *bioRxiv* **2015**. DOI: 10.1101/021568.
(94) Wang, T.; Wei, J. J.; Sabatini, D. M.; Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **2014**, *343* (6166), 80−84.
(95) Lin, J.; Wong, K.-C. Off-Target Predictions in CRISPR-Cas9 Gene Editing Using Deep Learning. *Bioinformatics* **2018**, *34* (17), i656−i663.
(96) 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. DeepCRISPR: Optimized CRISPR Guide RNA Design by Deep Learning. *Genome Biol.* **2018**, *19* (1), 80.
(97) Xue, L.; Tang, B.; Chen, W.; Luo, J. Prediction of CRISPR sgRNA Activity Using a Deep Convolutional Neural Network. *J. Chem. Inf. Model.* **2019**, *59* (1), 615−624.
(98) Liu, Q.; He, D.; Xie, L. Prediction of Off-Target Specificity and Cell-Specific Fitness of CRISPR-Cas System Using Attention Boosted Deep Learning and Network-Based Gene Feature. *PLOS Comput. Biol.* **2019**, *15* (10), No. e1007480.
(99) Liu, Q.; He, D.; Xie, L. Identifying Context-Specific Network Features for CRISPR-Cas9 Targeting Efficiency Using Accurate and Interpretable Deep Neural Network. *bioRxiv* **2018**. DOI: 10.1101/505602.
(100) Zhang, G.; Dai, Z.; Dai, X. A Novel Hybrid CNN-SVR for CRISPR/Cas9 Guide RNA Activity Prediction. *Front. Genet.* **2020**, *10*, 1303.
(101) Zhang, G.; Dai, Z.; Dai, X. C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA Activity Using Convolutional and Recurrent Neural Networks. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 344−354.
(102) Charlier, J.; Nadon, R.; Makarenkov, V. Accurate Deep Learning Off-Target Prediction with Novel sgRNA-DNA Sequence Encoding in CRISPR-Cas9 Gene Editing. *Bioinformatics* **2021**, *37* (16), 2299−2307.
(103) 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. Evaluation of Off-Target and on-Target Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR. *Genome Biol.* **2016**, *17* (1), 148.
(104) Störtz, F.; Mak, J. K.; Minary, P. piCRISPR: Physically Informed Deep Learning Models for CRISPR/Cas9 off-Target Cleavage Prediction. *Artif. Intell. Life Sci.* **2023**, *3*, 100075.
(105) Chen, Q.; Chuai, G.; Zhang, H.; Tang, J.; Duan, L.; Guan, H.; Li, W.; Li, W.; Wen, J.; Zuo, E.; Zhang, Q.; Liu, Q. Genome-Wide CRISPR off-Target Prediction and Optimization Using RNA-DNA Interaction Fingerprints. *Nat. Commun.* **2023**, *14* (1), 7521.
(106) Koike-Yusa, H.; Li, Y.; Tan, E.-P.; Velasco-Herrera, M. D. C.; Yusa, K. Genome-Wide Recessive Genetic Screening in Mammalian Cells with a Lentiviral CRISPR-Guide RNA Library. *Nat. Biotechnol.* **2014**, *32* (3), 267−273.
(107) Kim, H. K.; Song, M.; Lee, J.; Menon, A. V.; Jung, S.; Kang, Y.-M.; Choi, J. W.; Woo, E.; Koh, H. C.; Nam, J.-W.; Kim, H. In Vivo High-Throughput Profiling of CRISPR-Cpf1 Activity. *Nat. Methods* **2017**, *14* (2), 153−159.
(108) Marquart, K. F.; Allam, A.; Janjuha, S.; Sintsova, A.; Villiger, L.; Frey, N.; Krauthammer, M.; Schwank, G. Predicting Base Editing Outcomes with an Attention-Based Deep Learning Algorithm Trained on High-Throughput Target Library Screens. *Nat. Commun.* **2021**, *12* (1), 5114.
(109) Park, J.; Kim, H. K. Prediction of Base Editing Efficiencies and Outcomes Using DeepABE and DeepCBE. In *Base Editors*; Bae, S., Song, B., Eds.; Methods in Molecular Biology; Springer US: New York, 2023; Vol. *2606*, pp 23−32. DOI: 10.1007/978-1-0716-2879-9_3.
(110) Arbab, M.; Shen, M. W.; Mok, B.; Wilson, C.; Matuszek, Ż.; Cassa, C. A.; Liu, D. R. Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. *Cell* **2020**, *182* (2), 463−480.
(111) Mathis, N.; Allam, A.; Tálas, A.; Benvenuto, E.; Schep, R.; Damodharan, T.; Balázs, Z.; Janjuha, S.; Schmidheini, L.; Böck, D.; Van Steensel, B.; Krauthammer, M.; Schwank, G. Predicting Prime Editing Efficiency across Diverse Edit Types and Chromatin Contexts with Machine Learning. *bioRxiv* **2023**. DOI: 10.1101/2023.10.09.561414.
(112) Wessels, H.-H.; Stirn, A.; Méndez-Mancilla, A.; Kim, E. J.; Hart, S. K.; Knowles, D. A.; Sanjana, N. E. Prediction of On-Target and off-Target Activity of CRISPR-Cas13d Guide RNAs Using Deep Learning. *Nat. Biotechnol.* **2024**, *42* (4), 628−637.
(113) Wang, J.; Xiang, X.; Cheng, L.; Zhang, X.; Luo, Y. CRISPR-GNL: An Improved Model for Predicting CRISPR Activity by Machine Learning and Featurization. *bioRxiv* **2019**. DOI: 10.1101/605790.
(114) Wang, D.; Zhang, C.; Wang, B.; Li, B.; Wang, Q.; Liu, D.; Wang, H.; Zhou, Y.; Shi, L.; Lan, F.; Wang, Y. Optimized CRISPR Guide RNA Design for Two High-Fidelity Cas9 Variants by Deep Learning. *Nat. Commun.* **2019**, *10* (1), 4284.
(115) Kim, H. K.; Yu, G.; Park, J.; Min, S.; Lee, S.; Yoon, S.; Kim, H. H. Predicting the Efficiency of Prime Editing Guide RNAs in Human Cells. *Nat. Biotechnol.* **2021**, *39* (2), 198−206.

(116) Zhang, S.; Li, X.; Lin, Q.; Wong, K.-C. Synergizing CRISPR/Cas9 off-Target Predictions for Ensemble Insights and Practical Applications. *Bioinformatics* **2019**, *35* (7), 1108−1115.
(117) Zhang, Y.; Long, Y.; Yin, R.; Kwoh, C. K. DL-CRISPR: A Deep Learning Method for Off-Target Activity Prediction in CRISPR/Cas9 With Data Augmentation. *IEEE Access* **2020**, *8*, 76610−76617.
(118) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16* (8), 687−694.
(119) Wittmann, B. J.; Yue, Y.; Arnold, F. H. Informed Training Set Design Enables Efficient Machine Learning-Assisted Directed Protein Evolution. *Cell Syst.* **2021**, *12* (11), 1026−1045.
(120) Georgiev, A. G. Interpretable Numerical Descriptors of Amino Acid Space. *J. Comput. Biol.* **2009**, *16* (5), 703−723.
(121) Rao, R.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J. F.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. *bioRxiv* **2021**. DOI: 10.1101/2021.02.12.430858.
(122) Thean, D. G. L.; Chu, H. Y.; Fong, J. H. C.; Chan, B. K. C.; Zhou, P.; Kwok, C. C. S.; Chan, Y. M.; Mak, S. Y. L.; Choi, G. C. G.; Ho, J. W. K.; Zheng, Z.; Wong, A. S. L. Machine Learning-Coupled Combinatorial Mutagenesis Enables Resource-Efficient Engineering of CRISPR-Cas9 Genome Editor Activities. *Nat. Commun.* **2022**, *13* (1), 2219.
(123) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (18), 8852−8858.
(124) Choi, G. C. G.; Zhou, P.; Yuen, C. T. L.; Chan, B. K. C.; Xu, F.; Bao, S.; Chu, H. Y.; Thean, D.; Tan, K.; Wong, K. H.; Zheng, Z.; Wong, A. S. L. Combinatorial Mutagenesis En Masse Optimizes the Genome Editing Activities of SpCas9. *Nat. Methods* **2019**, *16* (8), 722−730.
(125) Wang, D.; Mou, H.; Li, S.; Li, Y.; Hough, S.; Tran, K.; Li, J.; Yin, H.; Anderson, D. G.; Sontheimer, E. J.; Weng, Z.; Gao, G.; Xue, W. Adenovirus-Mediated Somatic Genome Editing of *Pten* by CRISPR/Cas9 in Mouse Liver in Spite of Cas9-Specific Immune Responses. *Hum. Gene Ther.* **2015**, *26* (7), 432−442.
(126) Chew, W. L.; Tabebordbar, M.; Cheng, J. K. W.; Mali, P.; Wu, E. Y.; Ng, A. H. M.; Zhu, K.; Wagers, A. J.; Church, G. M. A Multifunctional AAV-CRISPR-Cas9 and Its Host Response. *Nat. Methods* **2016**, *13* (10), 868−874.
(127) Charlesworth, C. T.; Deshpande, P. S.; Dever, D. P.; Camarena, J.; Lemgart, V. T.; Cromer, M. K.; Vakulskas, C. A.; Collingwood, M. A.; Zhang, L.; Bode, N. M.; Behlke, M. A.; Dejene, B.; Cieniewicz, B.; Romano, R.; Lesch, B. J.; Gomez-Ospina, N.; Mantri, S.; Pavel-Dinu, M.; Weinberg, K. I.; Porteus, M. H. Identification of Preexisting Adaptive Immunity to Cas9 Proteins in Humans. *Nat. Med.* **2019**, *25* (2), 249−254.
(128) Simhadri, V. L.; McGill, J.; McMahon, S.; Wang, J.; Jiang, H.; Sauna, Z. E. Prevalence of Pre-Existing Antibodies to CRISPR-Associated Nuclease Cas9 in the USA Population. *Mol. Ther. - Methods Clin. Dev.* **2018**, *10*, 105−112.
(129) Wagner, D. L.; Amini, L.; Wendering, D. J.; Burkhardt, L.-M.; Akyüz, L.; Reinke, P.; Volk, H.-D.; Schmueck-Henneresse, M. High Prevalence of Streptococcus Pyogenes Cas9-Reactive T Cells within the Adult Human Population. *Nat. Med.* **2019**, *25* (2), 242−248.
(130) Shen, X.; Lin, Q.; Liang, Z.; Wang, J.; Yang, X.; Liang, Y.; Liang, H.; Pan, H.; Yang, J.; Zhu, Y.; Li, M.; Xiang, W.; Zhu, H. Reduction of Pre-Existing Adaptive Immune Responses Against SaCas9 in Humans Using Epitope Mapping and Identification. *CRISPR J.* **2022**, *5* (3), 445−456.
(131) Ewaisha, R.; Anderson, K. S. Immunogenicity of CRISPR Therapeutics—Critical Considerations for Clinical Translation. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1138596.
(132) Mehta, A.; Merkel, O. M. Immunogenicity of Cas9 Protein. *J. Pharm. Sci.* **2020**, *109* (1), 62−67.
(133) Wignakumar, T.; Fairchild, P. J. Evasion of Pre-Existing Immunity to Cas9: A Prerequisite for Successful Genome Editing In Vivo? *Curr. Transplant. Rep.* **2019**, *6* (2), 127−133.
(134) Mohan, D.; Wansley, D. L.; Sie, B. M.; Noon, M. S.; Baer, A. N.; Laserson, U.; Larman, H. B. PhIP-Seq Characterization of Serum Antibodies Using Oligonucleotide-Encoded Peptidomes. *Nat. Protoc.* **2018**, *13* (9), 1958−1978.
(135) Ponomarenko, J.; Bui, H.-H.; Li, W.; Fusseder, N.; Bourne, P. E.; Sette, A.; Peters, B. ElliPro: A New Structure-Based Tool for the Prediction of Antibody Epitopes. *BMC Bioinformatics* **2008**, *9* (1), 514.
(136) Zhou, C.; Chen, Z.; Zhang, L.; Yan, D.; Mao, T.; Tang, K.; Qiu, T.; Cao, Z. SEPPA 3.0—Enhanced Spatial Epitope Prediction Enabling Glycoprotein Antigens. *Nucleic Acids Res.* **2019**, *47* (W1), W388−W394.
(137) Da Silva, B. M.; Myung, Y.; Ascher, D. B.; Pires, D. E. V. epitope3D: A Machine Learning Method for Conformational B-Cell Epitope Prediction. *Brief. Bioinform.* **2022**, *23* (1), bbab423.
(138) Høie, M. H.; Gade, F. S.; Johansen, J. M.; Würtzen, C.; Winther, O.; Nielsen, M.; Marcatili, P. DiscoTope-3.0: Improved B-Cell Epitope Prediction Using Inverse Folding Latent Representations. *Front. Immunol.* **2024**, *15*, 1322712.
(139) Clifford, J. N.; Høie, M. H.; Deleuran, S.; Peters, B.; Nielsen, M.; Marcatili, P. BepiPred -3.0: Improved B-cell Epitope Prediction Using Protein Language Models. *Protein Sci.* **2022**, *31* (12), No. e4497.
(140) Kringelum, J. V.; Lundegaard, C.; Lund, O.; Nielsen, M. Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Comput. Biol.* **2012**, *8* (12), No. e1002829.
(141) Ferdosi, S. R.; Ewaisha, R.; Moghadam, F.; Krishna, S.; Park, J. G.; Ebrahimkhani, M. R.; Kiani, S.; Anderson, K. S. Multifunctional CRISPR-Cas9 with Engineered Immunosilenced Human T Cell Epitopes. *Nat. Commun.* **2019**, *10* (1), 1842.
(142) Chowell, D.; Krishna, S.; Becker, P. D.; Cocita, C.; Shu, J.; Tan, X.; Greenberg, P. D.; Klavinskis, L. S.; Blattman, J. N.; Anderson, K. S. TCR Contact Residue Hydrophobicity Is a Hallmark of Immunogenic CD8 $^{+}$ T Cell Epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (14), E1754−E1762.