



Published in final edited form as:

Nat Biotechnol. 2025 August ; 43(8): 1266–1273. doi:10.1038/s41587-024-02422-w.

Model-directed generation of artificial CRISPR–Cas13a guide RNA sequences improves nucleic acid detection

Sreekar Mantena^{1,2,3}, Priya P. Pillai¹, Brittany A. Petros^{1,4,5,6}, Nicole L. Welch¹, Cameron Myhrvold^{7,8,9,10}, Pardis C. Sabeti^{1,11,12,13,14,∞}, Hayden C. Metsky^{1,14,∞}

¹Broad Institute of MIT and Harvard, Cambridge, MA, USA

²Department of Statistics, Harvard University, Cambridge, MA, USA

³Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

⁴Division of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, Cambridge, MA, USA

⁵MD–PhD Program, Harvard/Massachusetts Institute of Technology, Boston, MA, USA

⁶Department of Systems Biology, Harvard Medical School, Boston, MA, USA

⁷Department of Molecular Biology, Princeton University, Princeton, NJ, USA

⁸Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA

⁹Omenn–Darling Bioengineering Institute, Princeton University, Princeton, NJ, USA

¹⁰Department of Chemistry, Princeton University, Princeton, NJ, USA

¹¹Howard Hughes Medical Institute, Chevy Chase, MD, USA

¹²Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

¹³Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA

¹⁴These authors contributed equally: Pardis C. Sabeti, Hayden C. Metsky

Abstract

Reprints and permissions information is available at www.nature.com/reprints.

∞ Correspondence and requests for materials should be addressed to Pardis C. Sabeti or Hayden C. Metsky.

pardis@broadinstitute.org; hayden@broadinstitute.org.

Author contributions

S.M. implemented and evaluated model-directed exploration algorithms for Cas13-based viral detection. P.P.P., N.L.W., C.M. and B.A.P. provided critical insights. H.C.M. initiated the project and performed some of the early methods development and analysis. S.M. and H.C.M. wrote the manuscript with input from other authors. H.C.M. and P.C.S. supervised the project.

Competing interests

S.M., P.P.P., H.C.M. and P.C.S. are co-inventors on patent applications filed by the Broad Institute related to work in this study. H.C.M. is an employee of and holds equity in Inceptive Nucleics, Inc. P.C.S. is a co-founder of and consultant to Sherlock Biosciences and Delve Bio, and is a Board Member of Danaher Corporation, and holds equity in all three companies. C.M. is a co-founder of Carver Biosciences, a startup company developing Cas13-based antivirals, and holds equity in Carver Biosciences. The other authors declare no competing interests.

Extended data is available for this paper at <https://doi.org/10.1038/s41587-024-02422-w>.

CRISPR guide RNA sequences deriving exactly from natural sequences may not perform optimally in every application. Here we implement and evaluate algorithms for designing maximally fit, artificial CRISPR–Cas13a guides with multiple mismatches to natural sequences that are tailored for diagnostic applications. These guides offer more sensitive detection of diverse pathogens and discrimination of pathogen variants compared with guides derived directly from natural sequences and illuminate design principles that broaden Cas13a targeting.

Molecular diagnostic technologies have transformed the detection and surveillance of infectious pathogens. The performance of these assays depends critically on the diagnostic oligonucleotide sequences that they employ to bind to targeted nucleic acids. Current techniques for designing such diagnostic oligonucleotides, including PCR primers and clustered regularly interspaced short palindromic repeats (CRISPR) guide RNAs, select sequences directly from those in nature or from a simple function of natural sequences, such as a consensus, according to machine-learned discriminative models or thermodynamic heuristics^{1–6}. Thus, outputs of these techniques are identical (or nearly identical) to the input, natural sequences. However, natural diagnostic guide sequences may not perform the best in every application. Artificial diagnostic sequences that differ from any natural sequence by several mismatches may achieve superior diagnostic activity than their natural counterparts. Machine learning methods have been applied to generate new, artificial biological sequences such as proteins^{7–12} and DNA regulatory elements^{13–15}, but, to the best of our knowledge, such techniques have not been applied to design diagnostic sequences.

We hypothesized that harnessing a generative design approach to construct artificial diagnostic oligonucleotide sequences may advance performance on key detection tasks. For instance, if our goal were to detect all forms of a genetically diverse pathogen (‘multitarget detection’) and there are several polymorphisms in an otherwise efficacious target region, an artificial oligonucleotide sequence could more sensitively detect all combinations of polymorphisms than any sequence that fully matches one particular combination (Fig. 1a). This would enable us to achieve improved diagnostic sensitivity. Alternatively, if our goal were to distinguish between two highly similar targets (‘variant identification’), even those that differ by a single-nucleotide polymorphism (SNP), an artificial sequence with optimally positioned mismatches could achieve greater specificity than a sequence identical to one target (Fig. 2a). This would enable us to more accurately identify mutation(s) in a pathogen or discriminate between pathogen lineages. Previous work^{16,17} has described heuristic rules to introduce handcrafted specificity-enhancing mismatches into diagnostic oligonucleotide sequences, but these strategies are bespoke and are limited by the degree of variation they can distinguish.

In this study, we focus on designing guides for use with CRISPR–Cas13a-based detection assays; specifically, we design artificial 28 nt CRISPR guide RNA spacer sequences. Cas13a-based assays have increasingly received attention because of their minimal equipment requirements and ease of use in low-resource settings^{18,19}. We previously described Activity-informed Design with All-inclusive Patrolling of Targets (ADAPT)⁶, which designs diagnostic guides by using a convolutional neural network (CNN) as a discriminative model to select among natural sequences (or consensus of sequence

clusters). Since the approach selects among a predetermined list of sequences, ADAPT cannot design guides that are several mismatches or more from natural sequences, it cannot design guides for variant identification down to single-nucleotide resolution, and it is unable to uncover novel strategies that optimize guide design. By contrast, algorithms that create artificial guide sequences could achieve these goals while enabling superior diagnostic sensitivity and specificity, particularly for challenges in infectious diseases.

Designing artificial guide sequences necessitates exploring a vast high-dimensional space of RNA sequences. Each guide has a fitness according to the task at hand, such as multitarget detection (Fig. 1a) or variant identification (Fig. 2a). Highly fit artificial guide sequences should be nearby in sequence space to the complements of their targets, but this region, with $\sim 10^7$ – 10^9 sequences (Supplementary Note 1), is sufficiently large to motivate the development of algorithms that efficiently search the sequence space.

Here, we implement and evaluate a set of algorithms for designing artificial CRISPR–Cas13a guide RNAs. These algorithms combine a machine-learned model—trained to predict the enzymatic activity that results from a guide–target interaction—with search algorithms to explore a fitness landscape of guides. Our predictive model is the CNN we previously developed and implemented in ADAPT⁶. It predicts the enzymatic activity of Cas13a, a proxy of a given guide’s sensitivity for detecting a given target. We implement and evaluate two design algorithms: (1) a generative model that employs a conditional Wasserstein Generative Adversarial Network and activation maximization (WGAN-AM) to explore a continuous latent space of sequences^{13,20,21} (Extended Data Fig. 1a and Supplementary Fig. 1) and (2) an evolutionary algorithm^{22–24} that performs iterative rounds of random mutation, fitness evaluation and selection to explore the fitness landscape in discrete steps (Extended Data Fig. 1b and Supplementary Fig. 2). Our model-directed exploration process—including the predictive model, search algorithms and fitness functions—is conditioned on user-provided target sequences, making it broadly applicable to any input targets rather than being restricted to specific genomic sites. We built a software package to make these design approaches easily accessible for Cas13 guide RNA design. It is called Building Artificial Diagnostic Guides by Exploring Regions of Sequences (BADGERS) and is freely available at <https://github.com/broadinstitute/badgers-cas13> (ref. 25).

First, we applied the algorithms in BADGERS to design guides, for use with Cas13a-based detection, that are sensitive across a pathogen’s genomic variation, that is, for multitarget detection. The fitness function that we maximize is the average predicted activity across sequenced genomes (Fig. 1a and Methods). We focus on five RNA viruses, chosen for their extensive diversity (enterovirus B, Lassa and dengue) or public health relevance (influenza A and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)). We employed the design algorithms implemented in BADGERS to design guides and benchmarked them against algorithms representing current, state-of-the-art design approaches (Methods). These baseline methods include a simple, commonly used algorithm that complements the consensus sequence at the target site (‘consensus’) and a more advanced algorithm that we term ‘model-based choice’ (MBC), which reflects the method implemented in ADAPT, which we previously developed⁶. MBC employs clustering to compute a candidate set of

guides and selects the guide, from this predetermined list, with maximal fitness based on our model's predictions. Although the predictive model of guide activity used by MBC and the algorithms in BADGERS is the same, MBC is conceptually and algorithmically distinct; MBC does not use the predictive model to explore a sequence space and it cannot generate artificial guide sequences with several mismatches from the targets (Methods).

We computationally evaluated the multitarget detection performance of the different design methods across these five pathogens using the same predictive model of guide activity (Methods). BADGERS-designed guides were predicted to detect more diversity than guides designed by baseline methods across nearly all genomic regions of all five pathogens. In several regions within dengue, Lassa and influenza A viruses, the BADGERS-designed guides were predicted to detect 10–20% more sequence variation (Fig. 1b and Supplementary Fig. 3). The sequence and nucleotide composition of a viral pathogen vary across different genomic regions and influence the activity of diagnostic guides; thus, BADGERS-designed guides may achieve a coverage improvement in some regions but not others. The BADGERS-designed guides also achieved a higher mean fitness genome wide (Fig. 1c and Supplementary Fig. 4a). At most sites, the guides generated by the algorithms in BADGERS are several mismatches away from any sequenced genomes, suggesting that the WGAN-AM and evolutionary methods' ability to search fitness landscapes and generate artificial sequences contributes to their superior performance (Fig. 1d and Supplementary Fig. 5). For example, in Lassa virus segment S, the WGAN-AM and evolutionary algorithms designed guides that were distinct from any observed sequence at ~92% and ~99.9% of genomic sites, respectively.

Computational predictions are inherently limited so we set out to experimentally benchmark the multitarget detection performance of the BADGERS-designed guides. We used a random sampling strategy (Methods) to select three representative genomic sites from each of the five viruses we considered (15 sites in total) and employed both the algorithms in BADGERS and baseline methods to design guides for these sites. We tested these guides against multiple synthetic targets that represent the genomic diversity of each virus (Methods), measuring the fluorescence readout over time. Cas13a is activated through guide-mediated recognition of a target; once active, it engages in collateral cleavage of a fluorescent reporter to generate the fluorescent readout, with greater fluorescent signal indicating stronger guide–target affinity at a fixed target concentration. All guides were tested against the target concentrations 10^{10} copies per microlitre, 10^9 copies per microlitre and 10^8 copies per microlitre, which represent typical postamplification concentrations of viral pathogens^{6,26,27}.

For all but one of the 15 tested sites, the BADGERS-designed guides achieved higher activity across sequence diversity than the baseline guides (Fig. 1e–h, and Supplementary Figs. 6 and 7) and often enabled a lower limit of detection as well (Fig. 1i and Supplementary Fig. 6). The algorithms in BADGERS provide the greatest advantage in highly diverse viruses, for example, Lassa virus. This advantage is less pronounced at more conserved sites from less diverse viruses, for example, SARS-CoV-2 (Supplementary Figs. 3 and 8).

We next applied the algorithms in BADGERS to our second diagnostic objective: to design guides that optimally identify specific SNPs or distinguish between pathogen lineages, that is, variant identification (Fig. 2a). This objective can be conceptualized as designing a guide that achieves high activity on an on-target sequence and minimal activity on an off-target sequence; the fitness function is approximately the difference between these activities (Fig. 2a and Methods).

To computationally evaluate the performance of BADGERS's generative design approach in designing guides that identify SNPs, we considered a set of 100 pairs of randomly generated on-target and off-target sequences, each differing by one nucleotide (Methods). We applied the algorithms in BADGERS to these target sets and also designed guides using the canonical strategy¹⁷ for this problem: this strategy places the SNP at position 26 of the guide and introduces a synthetic mismatch, at position 24 of the guide, to both the on- and off-target sequences (Methods). MBC's design approach cannot create guides that can differentiate between lineages down to the single-nucleotide level, so it could not be applied to this task. In silico, the BADGERS-designed guides exhibited similar predicted on-target activity to those designed by this synthetic mismatch method, but achieved lower off-target activity (Fig. 2b–d).

We benchmarked the design algorithms' variant identification performance experimentally, first by designing guides for six clinically relevant SNPs and comparing them with the baseline canonical synthetic mismatch approach. The six SNPs include four associated with antimicrobial resistance in *Plasmodium falciparum* (Pfcrf K76T, Pfdhps A437G, Pfk13 C580Y and Pfmdr1 Y184F), one associated with microcephaly in Zika virus (PrM S139N) and one associated with immune evasion in SARS-CoV-2's spike protein (K417N/K417T)^{28–34}. Experimentally, across nearly all of these tasks, the BADGERS-designed guides exhibited lower off-target activities while maintaining similar on-target activities to baseline guides (Fig. 2e,f and Supplementary Fig. 9a–c). Several baseline guides (including those for the S139, A437G and K417T targets) had an off-target signal nearly identical to the on-target signal throughout the reaction, which complicates SNP identification. In contrast, the BADGERS-designed guides achieved an on-target signal two to three times above the off-target signal (Fig. 2e,f).

We further benchmarked the design algorithms' variant discrimination performance experimentally for designing guides that differentiate viral lineages, focusing on dengue virus (DENV) serotypes 1–4 and seven key SARS-CoV-2 lineages (Methods). In the DENV task, one serotype is the on-target set (for example, DENV-1) and the other three serotypes comprise the off-target set (for example, DENV-2, 3 and 4). Both the BADGERS-designed guides and baseline guides differentiated the DENV subtypes with high specificity at a target concentration of 10^8 copies per microlitre, probably because there are DENV genomic regions with sufficient dissimilarity across serotypes (Supplementary Fig. 9d). However, when tested at a lower target concentration of 10^6 copies per microlitre, the baseline guides exhibited low on-target fluorescence for serotypes 2 and 4, while the BADGERS-designed guides achieved, as desired, a high on-target fluorescence for those serotypes (Fig. 2g). In the SARS-CoV-2 task, we designed guides that differentiate the ancestral Wuhan lineage (that is, Ref), Alpha, Beta, Delta, Gamma, Omicron/BA.1 and Omicron/BA.2 (Methods).

The BADGERS-designed guides enabled robust lineage discrimination, while the baseline guides for Alpha, Gamma and Ref had off-target activities that were nearly as high as their on-target activities throughout the reaction (Supplementary Fig. 9e).

Overall, the WGAN-AM and evolutionary search algorithms performed similarly in our two tasks, despite employing distinct techniques and creating different guides. On average, the evolutionary algorithm designed guides that are more distinct from observed sequences than the WGAN-AM's guides (Figs. 1d and 2c and Supplementary Fig. 5). Fundamentally, the WGAN-AM and evolutionary methods represent two different classes of search algorithms: the WGAN-AM samples sequences from a generative network while the evolutionary algorithm simulates natural selection. To explore the utility of alternative search algorithms, we evaluated the performance of another method in each class. Specifically, we performed *in silico* benchmarking of conditioning by adaptive sampling (CbAS)³⁵, which employs a generative network, and AdaLead³⁶, which employs an evolutionary-inspired approach (Methods). All four search algorithms provide a substantial improvement over the baseline approaches. Different search algorithms do exhibit subtle differences in performance depending on the specific task; for example, the WGAN-AM-designed guides achieved a lower mean fitness on the multitarget detection objective than guides designed by the other three search algorithms, although this was not the case on the variant identification objective (Supplementary Fig. 4).

Next, we sought to understand whether the algorithms in BADGERS leverage known rules governing Cas13a activity to enhance guide performance. Remarkably, we found that the generative design algorithms implicitly learn Cas13a's protospacer and mismatch preferences without having them explicitly encoded and rationally apply this biological understanding among our two design objectives to generate optimally fit guides. For example, guide–target mismatches in close proximity to one another are known to decrease enzymatic activity^{37–39}. When applied to the first objective of multitarget detection, the BADGERS-designed artificial guide sequences reduced the number of adjacent guide–target mismatches at mismatch-sensitive positions relative to baseline guides, enabling superior sensitivity (Supplementary Fig. 10). In contrast, when applied to the variant identification objective, the algorithms in BADGERS introduced artificial mismatches within 4 nt of the SNP for the majority of target sets (WGAN-AM, 61% and evolutionary, 81%), reducing off-target activity (Fig. 2c). Additionally, for nearly 75% of simulated targets, the algorithms in BADGERS further optimized guide specificity by placing the SNP between positions 18 and 23, a region where mismatches are known⁶ to be highly deleterious to LwaCas13a activity (Fig. 2d).

Furthermore, we found that the algorithms in BADGERS take advantage of another biological property of Cas13a to design guides optimally suited for both objectives. Previous work has found that Cas13a guides exhibit reduced activity when targeting a genomic site with complementarity to Cas13a's tag sequence, particularly targets with a G nucleotide at the position immediately 3' to the protospacer (the protospacer-flanking site (PFS))^{39,40}. In the variant identification objective, when the targeted SNP is a mutation from a G to a non-G nucleotide (as it is in the experimentally tested S139N and C580Y targets), the algorithms in BADGERS positioned the G at the PFS to achieve minimal off-target activity (schematized

in Extended Data Fig. 2 and experimental results in Fig. 2e and Supplementary Fig. 9a). However, in the multitarget detection objective, the algorithms in BADGERS exhibited an unexpected design decision. For genomic sites with a G at the PFS, the algorithms often introduced a mismatch directly adjacent to the tag sequence at position 28 in the protospacer (the same as position 1 in the spacer), which we term the tag-adjacent mismatch (TAM). Intriguingly, among the genomic sites with a G at the PFS, the TAM was present in 45.7% of guides designed by the WGAN-AM algorithm and 76.2% of guides designed by the evolutionary algorithm (Supplementary Fig. 11).

We hypothesized that the TAM could represent a different design strategy, discovered by the algorithms in BADGERS, that enables improved guide activity on targets with a G at the PFS, so we explored the TAM's relevance further. Mechanistically, the TAM may disrupt base pairing between the G at the PFS of the target and the C in the CRISPR RNA (crRNA) tag sequence, promoting the separation of the Cas13–target complex and thereby increasing collateral cleavage activity and the resulting fluorescent signal. To test our hypothesis, we designed a library of guides against targets having unfavorable nucleotides (GN) and favorable nucleotides (AN, TN or CN) for the first two positions of the antitag region; the first nucleotide of the antitag region is the PFS (Methods and Fig. 2h). Guides with a TAM targeting a sequence with a G at the PFS exhibited substantially greater activity than guides without this TAM (Fig. 2h and Supplementary Fig. 12). Furthermore, the TAM conferred the greatest benefit when the targeted sequence was GT in the antitag region. This is mechanistically explainable: the first two nucleotides of Cas13a's tag sequence complement GT, so the TAM's disruption of this double base pairing may enhance activity more than for targets with a non-GT allele (Fig. 2h and Supplementary Fig. 12). As expected, introducing a TAM to guides targeting a sequence with a non-G at the PFS did not increase activity. To our knowledge, this design principle of introducing a mismatch in the spacer to rescue the activity of guides with extended tag:antitag complementarity has not been described before, suggesting that these algorithms are capable of elucidating new design principles.

In this study, we demonstrate that generative design algorithms can yield maximally fit guide sequences for pathogen detection and surveillance, and can illuminate guide design rules. These algorithms leverage a predictive model to explore a sequence landscape, only evaluating ~4,000 out of $\sim 10^{16}$ potential guides throughout their search process, yet are capable of designing artificial guide sequences that outperform those derived directly from nature (Supplementary Note 1). Crucially, our approach is not trained on nor restricted to a specific design task, but rather can generate guides for any conditioned target set. The machine-learned models of guide activity that steer our optimization process enable our method to exploit features of Cas13a's underlying biology (for example, antitag and mismatch preferences) to generate guides that optimize diagnostic performance. During the early 2022 surge in coronavirus disease 2019 (COVID-19) cases, we applied a prototype version of our method to design guides that distinguish SARS-CoV-2 variants, demonstrating that BADGERS-designed panels achieve high concordance with next-generation sequencing in a clinical context⁴¹. Furthermore, these methods can reveal interpretable guide design rules for CRISPR-based technologies. The TAM proposed by the generative design algorithms may enable Cas13a-based technologies to more efficiently

target a broader range of sequences and improve our understanding of the enzyme's nuclease activation mechanisms.

While false positive sequences, which score highly computationally yet are nonfunctional experimentally, are a challenge in generative sequence design, none of the BADGERS-designed guide sequences fell into this category. This may be partly due to the nature of our tasks as our focus on nucleotide base pairing entails substantially less structural complexity and a smaller sequence space relative to other tasks (for example, protein design). Nevertheless, the design algorithms that we evaluated still rely on predictive models of guide activity, which may return inaccurate predictions for guide–target pairs outside of the training domain. Certain design approaches—such as the WGAN-AM, which was trained to generate active guide–target pairs, or CbAS, which considers uncertainty in activity predictions—may be less susceptible to generating out-of-domain sequences than methods such as the evolutionary algorithm that are less constrained in their search. Additionally, these methods may not perform as well when applied to pathogens whose sequence composition is highly dissimilar from the viral sequences that comprised the predictive model's training dataset. Also, because different orthologs of Cas13 are known to have varying sequence preferences, the design algorithms that we evaluated may not be reliable in designing guides for non-Cas13a enzymes. Furthermore, although these algorithms do enhance diagnostic sensitivity, many regions of viral genomes are hypervariable and, in some cases, even the guides designed by our methods do not detect the full extent of sequence diversity (for example, DENV and enterovirus B; Fig. 1f,g).

In the realm of pathogen diagnostics, to our knowledge, all current design methods for nucleic acid technologies derive oligonucleotide sequences directly from pathogen genomes^{1–6}. Our results suggest that generative sequence design algorithms offer substantial benefits relative to approaches (like the one in ADAPT) that are not geared toward designing artificial guide sequences. Namely, the algorithms in BADGERS created guides that achieved superior performance to MBC (the technique employed by ADAPT) on the multitarget detection objective and enabled the discrimination of lineages down to single-nucleotide resolution, a task ADAPT cannot perform. To achieve this, BADGERS's methods engineered guide sequences by leveraging rational design principles (for example, G–U base pairing and nucleotide and position-specific effects of mismatches) and discovered strategies for guide design (including the terminal adjacent mismatch).

Though we focused on Cas13 guide RNAs in this study, artificial sequences stand to optimize the design of diagnostic sequences beyond CRISPR guide RNAs, including the primers and fluorescent probes employed in PCR, loop-mediated isothermal amplification and recombinase polymerase amplification. While the specific search algorithms employed in this work to generate artificial sequences achieved strong performance, the field of biological sequence design is rapidly advancing with the recent application of other classes of design methods (for example, autoregressive and diffusion-based generative methods) and future work should investigate how these approaches may further benefit diagnostic design. Beyond diagnostics, generative design methods could design rationally mismatched guide RNAs for base editors to minimize bystander editing rates¹⁶. For certain CRISPR screens, where the objective is to introduce maximal sequence diversity into protein-coding genes,

these methods could design guides that maximize the number of distinct edits made at the targeted sites⁴². Furthermore, while high-fidelity variants of CRISPR enzymes have lowered off-target editing rates⁴³, these class of algorithms could optimize guide RNA sequences to minimize these off-target effects further. Recent work has also highlighted the importance of accounting for human genetic diversity when designing CRISPR therapeutics^{43,44}, and generative design approaches are poised to generate guide sequences that optimally account for this diversity.

Generative methods are transforming the field of biological sequence design, with recent work focusing on generating long protein and DNA sequences^{10,12,45}. Here, we show that algorithms that explore sequence landscapes to design artificial sequences can not only optimize shorter RNA sequences, such as crRNAs, but can simultaneously elucidate mechanistic design rules and serve needs in pathogen surveillance.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-024-02422-w>.

Methods

The model-directed exploration algorithms developed in this work pair (1) a predictive model of diagnostic guide activity with (2) exploration algorithms that search through a landscape of guide sequences. Together, they design guides that optimize a well-defined objective function.

A guide sequence is a ~20–30 nt nucleic acid sequence that binds to a target. In our applications, the guide is the 28 nt crRNA spacer sequence for use with CRISPR–Cas13a-based detection. It allows detection of a target by leading to a readout when the target is present.

Predictive model of diagnostic guide activity

We previously developed the predictive models used in this work and they are described in detail in ref. 6. The input into these models is a 28 nt guide sequence g and a 48 nt target sequence t (having 10 nt of context on each site of the guide's binding region). The output of the classification model $C(g | t) \in [0,1]$ can be interpreted as the probability that a given guide–target pair is active. The output of the regression model $R(g | t) \in [-4,0]$ is the predicted activity of an active guide–target pair. Higher values of $C(g | t)$ may suggest that a guide is more likely to bind to a given target, and higher values of $R(g | t)$ may suggest that, after binding, a guide's signal on a given target will grow faster.

In this study, we combined the classifier and regression model's predictions to define a weighted activity metric, A , which follows from the law of total probability:

$$A(g | t) = P(g \text{ is active on } t) \times (\text{activity of active } g \text{ on } t) + (1 - P(g \text{ is active on } t)) \times (\text{activity of inactive } g \text{ on } t).$$

In our case, $C(g | t) = P(g \text{ is active on } t)$, $R(g | t) = \text{activity of active } g \text{ on } t$ and -4 is the activity of inactive g on t because guides that are inactive have a near-zero signal, which is represented in our dataset by an activity of -4 , the lower bound of measured signal.

Thus, for a given guide–target pair, we computed $A(g | t)$ as follows:

$$A(g | t) = C(g | t) \times R(g | t) + (1 - C(g | t)) \times -4$$

Objective functions for diagnostic and surveillance applications

Model-directed exploration algorithms can design a guide sequence that optimizes a well-defined objective function across a genome’s sequence variation.

We focus on designing guides that optimize two specific objective functions. These objectives have applications in diagnostic testing and pathogen genomic surveillance.

Multitarget detection. In the first application (‘multitarget detection’), our goal is to design guides that are maximally active across genome sequence diversity. As such, the objective function for this application is

$$f_M(g, |, T) = E_{t \in T} \left[A(g, |, t) \right] = \sum_{t \in T} w_t \times A(g, |, t),$$

where T is the set of genome sequence targets and $t \in T$ is a single target.

Genome sequence diversity varies spatially and temporally, so the w_t represents a prior probability of encountering a target $t \in T$. We set a uniform prior distribution (assuming that all w_t are equal to $\frac{1}{|T|}$) over the target genome sequences, following the choice in ref. 6, but one could modify w_t as certain lineages of a pathogen become more or less prevalent in a geographic region of interest or over time.

Additionally, one could also consider defining more complex objective functions (such as the tenth percentile of the distribution of activity across genomic diversity), although these functions may be nondifferentiable and more difficult for certain exploration algorithms to maximize.

Variant identification. In the second application (‘variant identification’), our goal is to design a diagnostic guide g that can optimally differentiate between closely related sequences, maximizing activity against one set of targeted sequences (on-target) while minimizing activity against a separate but homologous set of nontargeted sequences (off-target). In our notation, T_1 represents the on-target sequences and T_2 represents the off-target sequences.

Initially, we considered a simple objective function: the difference between the expected on-target activity and expected off-target activity, $E_{t \in T_1}[A(g | t)] - E_{t \in T_2}[A(g | t)]$. Although many guides designed using this objective had a large difference between the expected on-target activity and expected off-target activity, they were often inactive or only marginally active on the on-target sequences, making these guide designs unusable because the function does not enforce that the guide's activity on the on-target set of sequences be sufficiently high for detection.

Thus, we defined an objective function for the variant identification application that is maximized only when the designed guide has high activity on the on-target sequences and low activity on the off-target sequences. We based this objective function upon the logistic function so it is differentiable. The objective is defined below as

$$f_D(g, |, T_1, T_2) = \left(\frac{1}{1 + ae^{k(-\text{logsumexp}(-A(g, |, T_1)) - o)}} - 1 \right) + r \times \left(\frac{-1}{1 + ae^{k(\text{logsumexp}(A(g, |, T_2)) - o)}} \right),$$

where o , a , k and r are all parameters that modulate the slope and curvature of the objective function. The values of these parameters were determined through a random search procedure described in Supplementary Note 2. We use $\text{logsumexp}(A(g | T_i))$ as shorthand for $\text{logsumexp}(\{A(g | t) \forall t \in T_i\})$, where $\text{logsumexp}(x_1, \dots, x_n) = \log(\exp(x_1) + \dots + \exp(x_n))$ is a smooth approximation to the maximum.

The value of $f_D(g | T_1, T_2)$ increases as the minimum activity of the guide across the on-target set increases and the maximum activity of the guide across the off-target set decreases. A visualization of this objective function is included in Supplementary Fig. 13.

Exploration algorithms

The above objective functions define the fitness of a guide for two types of detection tasks. Our goal is to design maximally fit guide sequences—those that maximize the value of these objective functions. We developed two algorithms, a WGAN-AM approach and an evolutionary algorithm, that search over a landscape of guide sequences to generate these maximally fit guides.

Conditional WGAN-AM. A generative adversarial network (GAN) is an unsupervised deep learning framework that approximates an underlying data-generating distribution⁴⁶. A well-trained GAN can generate new samples of data (for example, photographs or genomic sequences) that have properties similar to those of the training set.

We use a Wasserstein GAN (WGAN) inspired by the application in ref. 20, but with one key difference. Our WGAN uses a conditional GAN, in particular conditional on a given target set⁴⁷. The generator network, $G(z | T)$, generates a guide sequence for the given target set. This difference allows our method to generalize well across different target sets, which contrasts to using a GAN that has been fit to data derived from a single target set. The WGAN uses a latent variable z as an input, and z can be modulated to explore the sequence

space and generate different guide sequences for a given target set. Supplementary Fig. 14a schematizes the architecture of the generative network of our WGAN.

We trained the WGAN on the same set of active guide–target pairs that our previously developed CNN regression model was trained on⁶. These guide–target pairs were experimentally demonstrated to be active via the Combinatorial Arrayed Reactions for Multiplexed Evaluation of Nucleic acids (CARMEN)–Cas13 system⁶. Thus, the generator outputs guides with measurable activity against the conditioned target. We trained the WGAN via the WGAN gradient penalty method in ref. 48. We used a batch size of 32 and trained the critic 5 times for each batch on which the generator was trained. We used the following hyperparameters for the Adam optimizer⁴⁸: $\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$.

The generator network $G(\mathbf{z} | T)$ captures the high-level features that make a guide active against a given target set and is capable of generating an artificial guide sequence given a target set. To generate maximally fit diagnostic guides, we employed a joint activation maximization framework²⁰. In this joint framework, the WGAN's generator, $G(\mathbf{z} | T)$, transforms the latent vector $\mathbf{z}$ into a candidate guide sequence g and the objective function, $f(g | T)$, computes the guide's fitness.

To generate optimal guides, we compute the gradient of the guide's fitness with respect to the latent vector $\mathbf{z}$, and use the Adam optimizer⁴⁹ to take steps in the direction of this gradient: $\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} + (\text{step size}) \times \nabla_{\mathbf{z}} f(g | T)$, where $\nabla_{\mathbf{z}} f(g | T) = \sum_d \frac{\partial f(g | T)}{\partial x_d} \times \frac{\partial x_d}{\partial \mathbf{z}}$ and d is one of the ten dimensions of the latent vector.

This procedure allows us to perform a continuous search over the latent space to generate guide sequences that have optimal fitness. To more thoroughly explore the fitness landscape and account for the fact that several local minima may exist in the fitness landscape, we perform multiple searches over the latent space with each search starting at a different random starting point (sampled from a normal distribution) in the latent space. A pseudocode overview of the WGAN-AM algorithm is available in Supplementary Fig. 1 and a high-level schematic is presented in Extended Data Fig. 1a.

The WGAN-AM algorithm's generated guide sequences are implicitly constrained. Since the WGAN's generator network approximates the conditional probability distribution $P(g | T)$ of the guide–target pairs in the training set, the guide sequences output by the WGAN's generator network have similar properties (for example, mismatch positions and mismatch types) to the guide sequences used to train the regression model. Thus, the WGAN-AM algorithm searches regions of the fitness landscape in which the regression model is likely to accurately predict guide activity. Owing to the generator network's structure, it is unlikely that the WGAN-AM algorithm will design guides that are so different than the guide sequences in the training set that the regression model's predictions are no longer accurate and the model behaves pathologically. For example, we would not expect the WGAN-AM algorithm to generate guides that have a much higher number of mismatches against their target sets than those in the training data for the predictive models.

Evolutionary algorithm. We also tested an unconstrained search algorithm that could reach any region of the fitness landscape—an evolutionary algorithm. Evolutionary algorithms are biologically inspired optimization algorithms that are widely used in computer science to search across a domain of candidate solutions. They have been applied to design biological sequences that exhibit a desired property, such as peptide sequences that have maximal antimicrobial activity^{23,50,51}.

We build upon this work to develop an evolutionary algorithm that searches over our fitness landscape of guide sequences by performing iterative rounds of fitness evaluation, selection and mutation⁵¹.

First, we initialize a population of guide sequences G_0 by extracting all of the guide-length sequences at the genomic site of interest. Then, we use the objective function $f(g | T)$ to evaluate the fitness of each of these candidate guides. We sample S parent guides from this population with a probability $q_i = \frac{e^{\beta \times (g_i | T)}}{\sum_{j=1}^{(G_i)} e^{\beta \times (g_j | T)}}$, where β represents the selection intensity⁵¹. We then mutate the randomly sampled guides with a mutation frequency γ . Specifically, for each position in the guide, we randomly sample a value from the standard uniform distribution, $x \sim \text{Uniform}(0,1)$. If $x < \gamma$, we mutate the nucleotide at that position (for example, ‘A’) to one of the other nucleotides (for example, ‘C’, ‘G’ or ‘U’) with uniform probability. Finally, we add the resulting child guides to the guide population. We repeat these rounds of fitness evaluation, selection and mutation for several generations until the limit on the number of calls to the objective function, L , is exceeded. We return the guide sequences in the final population.

Throughout the search process, the algorithms in BADGERS are capable of efficiently recognizing the positions at which fitness increases if the guide diverges from the consensus sequence, but they do not explicitly evaluate each possible nucleotide at those positions. Thus, after the last generation (of the evolutionary algorithm) and after the last outer round (of the WGAN-AM algorithm), the algorithms perform a local search, where they identify positions at which the generated guide differs from the consensus, mutate the guide to the other possible nucleotides at each of these positions and then select the resulting guide with the best fitness.

A pseudocode overview of the evolutionary algorithm is available in Supplementary Fig. 2 and a high-level schematic is presented in Extended Data Fig. 1b.

CbAS and AdaLead search algorithms. To computationally evaluate the performance of other search algorithms used in the field of biological sequence design in this domain, we applied CbAS³⁵ and AdaLead³⁶ to design guides for both objectives. The random search process used to determine the hyperparameters of the search algorithms is described in Supplementary Note 2. We implemented and tested all models using TensorFlow 2.8.0⁵² and FLEXS 0.2.8³⁶ in Python 3.7.10.

Computational benchmarking of model-directed exploration algorithms

To characterize the ability of the WGAN's generator network to introduce mismatches at different positions in the guide and between different alleles, we ran a simulation. We computed the consensus sequence at each of the 100 genomic sites used in the multitarget detection random search. For each consensus sequence $C_1, \dots, C_{100}$, we randomly sampled 500 latent variables $z \sim N(0,1)$ and generated 500 guides conditioned on that consensus sequence using the generator network $G(z | C_i)$. We computed the Euclidean norm of the latent variable as well as the Hamming distance between the WGAN-AM generated guide and the consensus sequence, which is the number of positions at which the two guides have different nucleotides.

We found that that, as the Euclidean norm of the latent variable z increases, the guide generated by the WGAN's generator network has a higher Hamming distance from the sequence that it was conditioned upon. In other words, as the latent variable z gets farther from the origin, the generator introduces more mutations into the generated guide. This property is illustrated in Supplementary Fig. 14b. Thus, the latent space z has an interpretable biological meaning.

Additionally, to benchmark the algorithms' ability to design guides for the variant identification objective, we created a synthetic benchmarking dataset. To avoid biasing the sequence composition toward a limited set of pathogen genomes and evaluate the performance of the methods across a large distribution of sequences, we randomly generated 100 pairs of DNA sequences, each pair consisting of targets that differ by one nucleotide. The algorithms in BADGERS were run on this dataset and the results are shown in Fig. 2b–d.

Designing diagnostic guides for experiments

We applied our model-directed exploration algorithms to design diagnostic guides for both the multitarget detection and variant identification applications.

Multitarget detection experiments. For the multitarget detection application, we selected five RNA viruses that are of public health interest and cause substantial morbidity and mortality globally. We used ADAPT⁶ to download and curate complete genomes for DENV (taxid: 12637), influenza A virus (segment 2, taxid: 11320) and enterovirus B (taxid: 138949) on 30 August 2021 from the National Center for Biotechnology Information (NCBI) GenBank⁵³. For SARS-CoV-2, we downloaded all complete genomes available on the Global Initiative on Sharing All Influenza Data (GISAID)⁵⁴ on 28 June 2021, removed the sequences flagged as 'low quality' and randomly sampled 10,000 of the genomes. For Lassa virus (segment S, taxid: 11620), we directly downloaded complete genomes from GenBank⁵³ using the following search criteria: 'txid11620[Organism:exp]' AND 'segment S' AND 'complete sequence' on 31 August 2021. We used MAFFT⁵⁵ to generate alignments for these five viral species. The predictive models ($R(g | T)$ and $C(g | T)$) employed in this work require an input of a 48 nt target sequence (a 28 nt guide-binding target sequence plus 10 nt of context on each side). Thus, we extracted 48 nt sliding windows along the genome and considered the target set T at that genomic site. We removed

the genomic sites with less than ten valid sequences (sequences that are a contiguous 48 nt and have no ambiguity or gaps). Both of the model-based exploration algorithms were run on all of these target sets using the multitarget objective function, $f_M(g | T)$, to design guides for these genomic sites.

To benchmark the performance of our model-based exploration algorithms against current methods, we used ADAPT to design guides at each of these genomic sites. ADAPT is software suite that employs machine learning and combinatorial optimization to design diagnostic guides and is state-of-the-art in the field⁶. When running ADAPT, we used the arguments ‘-w 28 -gl 28 --predict-cas13a-activity-model --obj maximize-activity -hgc 1 --cluster-threshold 0.3’ with SEARCH-TYPE set to sliding window, such that the CNN-based predictive model was used and ADAPT was restricted to designing one guide at each site. We refer to this usage of ADAPT as MBC since the predictive model is being used to select a maximally fit sequence present in the ground set.

Furthermore, we computed the consensus of the sequence alignment at each of these genomic sites to represent a simple baseline approach to guide sequence design. These guides are called ‘consensus guides’ throughout the text.

We sought to appropriately compare the performance of the guides designed by the baseline methods (MBC and the consensus) with the performance of the guides designed by our model-directed exploration algorithms (WGAN-AM and evolutionary). It is not tractable to experimentally test the tens of thousands of diagnostic guides designed. Thus, for every genomic site, we computed a relative performance metric, RP_i , the summed fitness of the guides designed by the model-directed exploration algorithms subtracted by the summed fitness of the guides designed by the baseline methods:

$$RP_i = (f(g_{\text{WGAN}} | T_i) + f(g_{\text{evolutionary}} | T_i)) - (f(g_{\text{consensus}} | T_i) + f(g_{\text{MBC}} | T_i)).$$

Previous studies^{6,17} have shown that LwaCas13a exhibits reduced activity when targeting a genomic sites with a G at the PFS, so we only considered genomic sites with a non-G at the PFS for experimental testing. For each viral species, all genomic sites with a non-G at the PFS were sorted into quantiles by their respective RP_i , so genomic sites with the highest RP_i were in the fourth quartile. We randomly sampled two sites from the fourth quartile and randomly sampled one site from the first quartile for each viral species. This method for selecting genomic sites allows us to experimentally test sites at which the guides designed by the model-directed exploration algorithms were likely to have superior performance than the guides designed by previous methods, as well as sites at which guides designed by previous methods were likely to have similar or better performance than the guides designed by the model-directed exploration algorithms. Each panel with experimental multitarget detection results shows one such site.

To computationally characterize guide performance, we computed the coverage of the guides designed by the algorithms in BADGERS and the baseline methods across different genomic windows in the viral pathogens, as shown in Fig. 1b and Supplementary Fig. 3.

A guide g was considered to detect a target if it was predicted to be active by the classification model ($C(g | t) > 0.577467$) and was predicted to be in the top quartile of activity by the regression model ($R(g | t) > -1.2801363$), based on the ‘highly active’ criterion used in ref. 6. The coverage was computed by determining the percentage of targets a guide detects in the given target set. To visualize these results in Fig. 1b and Supplementary Fig. 3, the guide designed by each method in each genomic window with the maximum predicted coverage was selected. The percentage of targets detected by this guide is shown in the graphs.

We computed the predicted guide activity across all targetable genomic sites in these pathogens, as shown in Fig. 1c and Supplementary Fig. 4a. Targetable genomic sites are defined as sites at which the baseline methods had meaningful activity ($\frac{A(g_{\text{consensus}} | T) + A(g_{\text{MBC}} | T)}{2} > -2.5$; in practice, it would not make sense to target sites that fall below this minimal threshold). In these figures and in all box plots throughout the manuscript, the bounds of the box represent the first and third quartiles (Q1 and Q3), the center represents the median and the whiskers extend from $Q1 - 1.5 \times \text{interquartile range}$ to $Q3 + 1.5 \times \text{interquartile range}$.

Variant identification experiments. For the variant identification application, we sought to design diagnostic guides that could differentiate between DENV serotypes 1–4 and the SARS-CoV-2 World Health Organization-designated variants of concern. For DENV, we downloaded all available complete genomes for serotypes 1–4 from the Virus Pathogen Database and Analysis Resource⁵⁶ on 1 February 2022. We used MAFFT⁵⁵ to generate an alignment of the genomes.

For SARS-CoV-2, we downloaded the ‘full length variant alignment’ from GISAID on 22 January 2022. This alignment contains representative sequences for the targeted SARS-CoV-2 lineages: Wuhan reference, Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Omicron (BA.1) and Omicron (BA.2). The GISAID accessions for these reference genomes are EPI_ISL_402124, EPI_ISL_674612, EPI_ISL_940877, EPI_ISL_2777382, EPI_ISL_1758376, EPI_ISL_6914029 and EPI_ISL_7190366, respectively. We referenced analyses on the CoVariants⁵⁷ and outbreak.info⁵⁸ websites to confirm that the insertions, deletions and mutations targeted by our guides were highly conserved and were present in at least 85% of genomic sequences of each variant.

The model-directed exploration algorithms were run on the DENV and SARS-CoV-2 alignments described above to design guides that were specific to each serotype/genotype/lineage.

To benchmark the performance of the model-directed exploration algorithms for variant identification of lineages (beyond a single SNP), we implemented a baseline approach. This baseline approach designs guide sequences by finding the 28 nt subsequence of the on-target consensus that has the highest Hamming distance from the off-target consensus and has an average nucleotide identity of at least 90% across the on-target genomes. This approach

mimics the strategy (usually performed by hand) of identifying regions that are conserved among the on-target genomes but divergent from the off-targets.

For the antimalarial resistance panel, we downloaded reference sequences for the *P. falciparum* genes *Pfmdr1*, *Pfcr1*, *Pfk13* and *Pfdhps* from GenBank accessions [HQ215532.1](#), [LC498250.1](#), [KT328114.1](#) and [KE123491.1](#), respectively.

For the K417N/T identification task, we used the same SARS-CoV-2 sequences that we previously downloaded from GISAID, as described above. For the Zika virus S139N identification task, we downloaded the reference sequence from GenBank accession [MT483911.1](#). We extracted the 250 nt genomic regions encapsulating the sites of interest in each gene and used them as the target sequences.

The model-directed exploration algorithms were run to design guides for each of the four SNPs of interest in the *P. falciparum* genome, the K417N/T SNP in SARS-CoV-2 and the S139N SNP in Zika virus. We also sought to benchmark our algorithms against a baseline. The ‘synthetic mismatch’ strategy (where the SNP is placed at position 26 of the protospacer and a mismatch at position 24 is introduced against both the derived and ancestral targets) is currently the standard approach for designing Cas13a guides that identify SNPs^{17,59}. The synthetic mismatch creates a double mismatch against the off-target sequence (positions 24 and 26), but only one mismatch against the on-target (position 24). We employed the synthetic mismatch strategy to design guides that target these four SNPs of interest, enabling us to benchmark the performance of our methods against the current standard.

Tag-adjacent mismatch experiments

As discussed in the main text, the model-directed exploration algorithms often introduce a mismatch at position 28 in the guide when targeting a genomic site with G nucleotide at the PFS (Fig. 2h). We refer to this mismatch the TAM and hypothesized that it may enhance guide–target activity.

To determine whether this was the case, we designed an experimental library consisting of targets that were representative of viral sequence diversity. We randomly sampled four viral families from a list of all viral families that have at least one vertebrate-infecting species. Then, within each of these four families, we randomly sampled one vertebrate-infecting species that has at least 100 complete genomes. The four selected viral species were primate tick-borne encephalitis virus (NCBI taxid: 11084), human mastadenovirus D (taxid: 130310), hepatitis B virus (taxid: 10407) and Zaire ebolavirus (taxid: 186538). For each of these viral species, we downloaded all available complete genomes from GenBank⁵³ on 15 February 2022 and used MAFFT⁵⁵ to create an alignment. We randomly selected a genomic site with a G nucleotide at the PFS in each of the four alignments and used this as our target sequence. We mutated the G nucleotide at the PFS in the original target sequence to create three additional targets with a non-G nucleotide at the PFS. Furthermore, since the nucleotide directly 3′ of a site with a G at the PFS also impacts activity⁶, we additionally mutated the original target sequence to create targets that have an antitag region of GC, GG, GA and GT.

For each of the four species, we designed the guide sequence without a TAM by simply extracting the 28 nt protospacer-binding region from the original target sequence. We also designed three guide sequences with the TAM by mutating the nucleotide at position 28 of the guide's protospacer to the other three possible bases at that position.

Preparing target sequences for experiments

To experimentally test the performance of our guides, we first designed target sequences against which we would test guides. We designed these targets to be representative of viral genomic diversity. We ran ADAPT's `pick_test_targets.py` program developed in ref. 6, with the arguments '`--num-representative-targets 5 --min-target-len 250`'. This script extracts a region of the alignment that encapsulates the guide-binding sites and is at least 250 nt long. Then, it clusters the resulting sequences and determines the medoid of each cluster, which are used as the representative target sequences. At all of the genomic sites we experimentally tested, the targets represented at least 95% of the total genomic diversity.

We added a T7 promoter sequence (5'-GAAATTAATACGACTCACTATAGGG-3') followed by a positive control sequence (5'-CACTATAGGGGCTCTAGCGACTTCTTTAAATAGTGGCTTAAAATAAC-3') to the 5' end of every target sequence. In each experiment, we included a crRNA with a protospacer that was complementary to this positive control sequence (5'-GCTCTAGCGACTTCTTTAAATAGTGGCT-3'). This positive control enabled us to verify that the target sequences were synthesized correctly. *P. falciparum*'s genome may pose DNA synthesis challenges because it is GC poor, so we added a second positive control sequence (5'-GAATGGAAGCACCGAGAGTATATGAAGATCTTCATGTGTGCAAAAGAATGGTAAAGCAGAGAAGGAGC-3') to the 3' end of all the *P. falciparum* target sequences. In each experiment that had *P. falciparum* targets, we included crRNA with a protospacer that was complementary to this positive control (5'-TTCTTTTGCACACATGAAGATCTTCATAT-3').

Preparing guide sequences for experiments

At each genomic site, the WGAN-AM algorithm outputs r_{outer} guides and the evolutionary algorithm outputs S guides; the WGAN-AM guide sequence and the evolutionary guide sequence with the highest predicted fitness were used for experimental testing. All of the methods employed in this work (including the model-based exploration algorithms and the baseline methods) output guide sequences in the frame of Cas13's protospacer. Thus, to prepare the guide sequences for experiments, we took the reverse complement of them to transform them to the frame of the CRISPR-Cas13 spacer and added the LwaCas13 direct repeat (5'-GAUUUAGACUACCCCAAAAACGAAGGGGACUAAAAC-3') to the 5' end of the crRNA sequence.

All oligonucleotide sequences employed are available at <https://github.com/broadinstitute/badgers-cas13-analysis>.

Experimentally evaluating diagnostic guide designs

Experimental methods. We used the mCARMEN⁴¹ platform to experimentally test the performance of diagnostic guides. Briefly, mCARMEN is a CRISPR-based diagnostic technology that uses a Fluidigm microfluidic chip to enable highly multiplexed testing of dozens of diagnostic guides against dozens of RNA targets. We followed the protocol described in the ‘General mCARMEN procedures’ section of ref. 41, with the following modifications:

All DNA targets were ordered as gBlocks from Integrated DNA Technologies. All crRNAs and the quenched synthetic fluorescent RNA reporter (FAM/rUrUrUrUrUrU/3IABkFQ) were also ordered from Integrated DNA Technologies.

The DNA targets were serially diluted to the concentrations of 10^{10} , 10^9 and 10^8 copies per microlitre and 1.43 μ l of each target served as input to each sample mix. The quenched synthetic fluorescent reporter was included in the sample mix at a concentration of 500 nM. The crRNAs were included in the assay mixes at a concentration of 212.5 nM, and LwaCas13 from GenScript was included in the assay mix at a concentration of 42.5 nM.

After chip loading, the Fluidigm Biomark HD was set to a constant temperature of 37 °C and was used to image the integrated fluidic circuit on the FAM and ROX channels every 5 min for 3 h.

Analysis of experimental data. To robustly characterize the performance of each of the guide–target pairs, we computed reference-normalized background-subtracted fluorescence values, as was done in refs. 6,41: reference-normalized fluorescence = $\frac{S_t - S_0}{R_t - R_0}$, where S_t is the guide–target pair’s FAM signal at time point t , S_0 is the guide–target pair’s FAM signal at time point 0, R_t is the guide–target pair’s reference ROX signal at time point t and R_0 is the guide–target pair’s reference ROX signal at time point 0.

To compute the reference-normalized background-subtracted fluorescence of each guide–target pair at time t , we subtracted the reference-normalized fluorescence of the no-template control at time t from the reference-normalized fluorescence of the guide–target pair at time t . Thus, if a guide–target pair has a reference-normalized background-subtracted fluorescence greater than 0, it has a fluorescence that is higher than the no-template control signal.

In the heat maps of fluorescence, we plotted the reference-normalized background-subtracted fluorescence at the 1 h time point. In the kinetic curves of fluorescence, we plotted all the reference-normalized background-subtracted fluorescence values collected throughout the course of the reaction ($t = 0$ –180 min).

Reporting summary

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

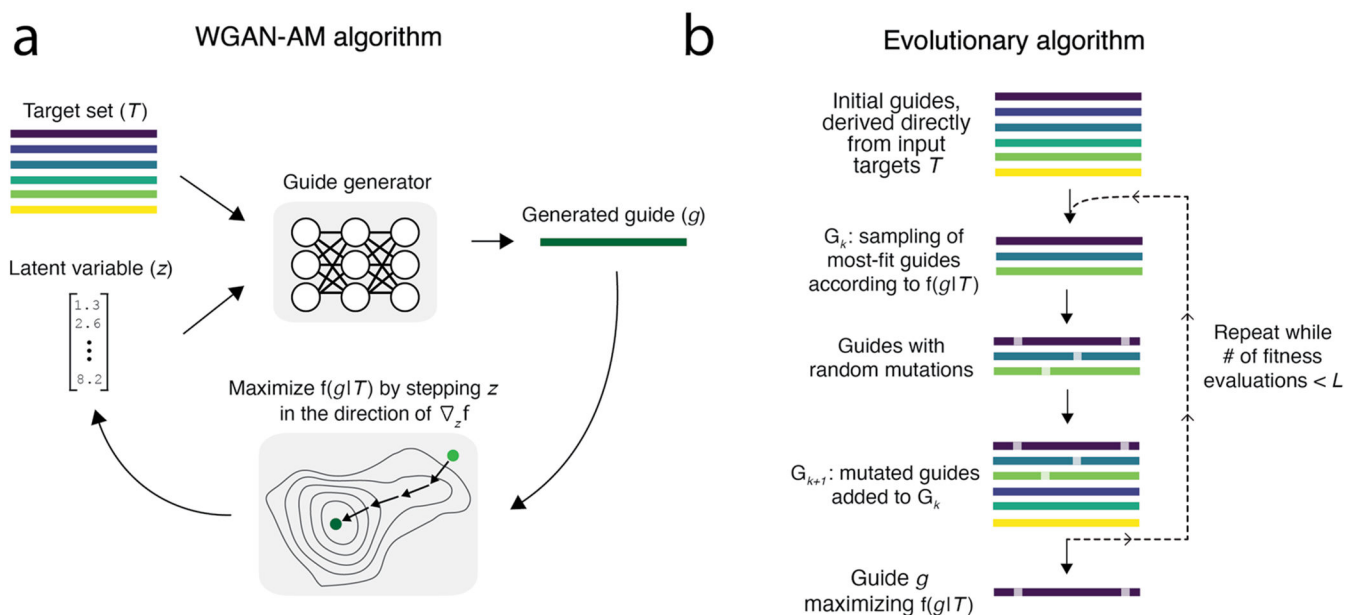
Data availability

The viral genomes used in this manuscript were downloaded from the publicly accessible NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and GISAID (<https://gisaid.org/>) databases. The data generated and analyzed in the manuscript is available via GitHub at <https://github.com/broadinstitute/badgers-cas13-analysis> (ref. 25).

Code availability

The BADGERS software package is available under the MIT license via GitHub at <https://github.com/broadinstitute/badgers-cas13> (ref. 25). The repository contains detailed instructions on how to run the algorithms on any input genomic sequences. The code used to run the analyses in this manuscript is available via GitHub at <https://github.com/broadinstitute/badgers-cas13-analysis>.

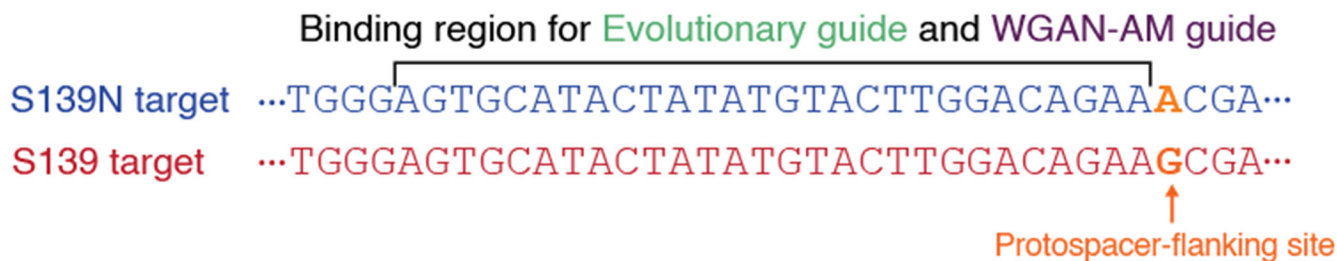
Extended Data



Extended Data Fig. 1 | BADGERS designs artificial diagnostic guides using the WGAN-AM and evolutionary search algorithms.

(a) WGAN-AM algorithm for guide design. The generative model, $G(z|T)$, generates active guide sequences conditional on a given target set T . A latent variable z modulates the generator's output, allowing for different guide sequences. To generate optimal guides, the algorithm starts at a random value of z , computes the fitness of the generated guide $f(g|T)$, and adjusts z in the direction of $\nabla_z f(g|T)$. **(b)** Evolutionary algorithm for guide design. The algorithm initializes a population of candidate guides, G_0 , by extracting the guide-length subsequences at the targeted genomic site. To form the population G_k in each generation k , n parent guides are sampled from the population G_{k-1} with probability proportional to their fitness $f(g|T)$. The n guides are randomly mutated, and the resulting mutated child guides are

added to the population. This process is repeated until L , the limit on the number of guides whose fitness has been evaluated, is reached.



Extended Data Fig. 2 |. Algorithms in BADGERS exploit Cas13a's protospacer context preferences to design optimal variant identification guides.

When targeting a SNP from a G nucleotide to a non-G nucleotide, the algorithms in BADGERS often place the G nucleotide in the off-target at the PFS for optimal SNP discrimination. In this schematized example, the guides were designed to achieve high activity on the S139N target and low activity on the S139 target. Both the WGAN-AM and evolutionary algorithm positioned the guides such that the G nucleotide in the S139 off-target sequence is located at the PFS, thus lowering off-target activity, as is experimentally observed (Fig. 2e).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank T. Nguyen, J. Arizti Sanz and L. Pinello for valuable discussions and advice. This project was made possible by the Amazon Web Services Diagnostic Development Initiative and Defense Advanced Research Projects Agency (DARPA) grant no. D18AC00006. H.C.M. was supported by National Institutes of Health (NIH)/National Institute of Allergy and Infectious Diseases (NIAID) grant no. K01AI163498 and U.S. Agency for International Development (USAID) Prime grant no. 58-3022-2-031. S.M. and B.A.P. were supported by NIH/National Institute of General Medical Sciences (NIGMS) grant no. T32GM144273. B.A.P. was also supported by NIH/NIGMS grant no. T32GM007753 and the Broad Institute MD-PhD Fellowship. C.M. was supported by INV-034761 from the Bill and Melinda Gates Foundation and the Princeton Catalysis Initiative. P.C.S. was supported by Howard Hughes Medical Institute (HHMI), NIH/NIAID grant no. U19AI110818 and U01AI151812, Centers for Disease Control and Prevention (CDC) grant no. NU50CK000629 and 75D30122C15113, Flu Lab, and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation and Open Philanthropy.

References

1. Jabado OJ et al. Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Res.* 34, 6605–6611 (2006). [PubMed: 17135211]
2. Kreer C et al. openPrimeR for multiplex amplification of highly diverse templates. *J. Immunol. Methods* 480, 112752 (2020). [PubMed: 31991148]
3. Duitama J et al. PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Res.* 37, 2483–2492 (2009). [PubMed: 19264805]
4. Brodin J et al. A multiple-alignment based primer design algorithm for genetically highly variable DNA targets. *BMC Bioinform.* 14, 255 (2013).

5. Varliero G, Wray J, Malandain C & Barker G PhyloPrimer: a taxon-specific oligonucleotide design platform. *PeerJ* 9, e11120 (2021). [PubMed: 33986979]
6. Metsky HC et al. Designing sensitive viral diagnostics with machine learning. *Nat. Biotechnol* 40, 1123–1131 (2022). [PubMed: 35241837]
7. Gupta A & Zou J Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intell* 1, 105–111 (2019).
8. Repecka D et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell* 3, 324–333 (2021).
9. Shin J-E et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun* 12, 2403 (2021). [PubMed: 33893299]
10. Watson JL et al. De novo design of protein structure and function with RFdiffusion. *Nature* 620, 1089–1100 (2023). [PubMed: 37433327]
11. Strokach A & Kim PM Deep generative modeling for protein design. *Curr. Opin. Struct. Biol* 72, 226–236 (2022). [PubMed: 34963082]
12. Madani A et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol* 41, 1099–1106 (2023). [PubMed: 36702895]
13. Taskiran II et al. Cell type directed design of synthetic enhancers. *Nature* 626, 212–220 (2024). [PubMed: 38086419]
14. de Almeida BP et al. Targeted design of synthetic enhancers for selected tissues in the *Drosophila* embryo. *Nature* 626, 207–211 (2024). [PubMed: 38086418]
15. Gosai SJ et al. Machine-guided design of synthetic cell type-specific cis-regulatory elements. Preprint at <https://www.biorxiv.org/content/10.1101/2023.08.08.552077v1> (2023).
16. Zhao D et al. Imperfect guide-RNA (igRNA) enables CRISPR single-base editing with ABE and CBE. *Nucleic Acids Res.* 50, 4161–4170 (2022). [PubMed: 35349689]
17. Gootenberg JS et al. Nucleic acid detection with CRISPR–Cas13a/C2c2. *Science* 356, 438–442 (2017). [PubMed: 28408723]
18. Arizti-Sanz J et al. Simplified Cas13-based assays for the fast identification of SARS-CoV-2 and its variants. *Nat. Biomed. Eng* 6, 932–943 (2022). [PubMed: 35637389]
19. Kaminski MM, Abudayyeh OO, Gootenberg JS, Zhang F & Collins JJ CRISPR-based diagnostics. *Nat. Biomed. Eng* 5, 643–656 (2021). [PubMed: 34272525]
20. Killoran N, Lee LJ, Delong A, Duvenaud D & Frey BJ Generating and designing DNA with deep generative models. Preprint at <https://arxiv.org/abs/1712.06148> (2017).
21. Zimec J et al. Controlling gene expression with deep generative design of regulatory DNA. *Nat. Commun* 13, 5099 (2022). [PubMed: 36042233]
22. Sample PJ et al. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol* 37, 803–809 (2019). [PubMed: 31267113]
23. Porto WF et al. In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat. Commun* 9, 1490 (2018). [PubMed: 29662055]
24. Fox R Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J. Theor. Biol* 234, 187–199 (2005). [PubMed: 15757678]
25. Mantena S et al. BADGERS: a package for designing artificial Cas13a diagnostic guides. *GitHub* <https://github.com/broadinstitute/badgers-cas13> (2024).
26. Granados A, Peci A, McGeer A & Gubbay JB Influenza and rhinovirus viral load and disease severity in upper respiratory tract infections. *J. Clin. Virol* 86, 14–19 (2017). [PubMed: 27893998]
27. Dayarathna S et al. Are viral loads in the febrile phase a predictive factor of dengue disease severity? Preprint at <https://www.medrxiv.org/content/10.1101/2023.07.31.23293412v1> (2023).
28. Yuan L et al. A single mutation in the prM protein of Zika virus contributes to fetal microcephaly. *Science* 358, 933–936 (2017). [PubMed: 28971967]
29. Apinjoh TO, Ouattara A, Titanji VPK, Djimde A & Amambua-Ngwa A Genetic diversity and drug resistance surveillance of plasmodium falciparum for malaria elimination: is there an ideal tool for resource-limited sub-Saharan Africa? *Malar. J* 18, 217 (2019). [PubMed: 31242921]

30. Carter TE et al. Evaluation of dihydrofolate reductase and dihydropteroate synthetase genotypes that confer resistance to sulphadoxine-pyrimethamine in *Plasmodium falciparum* in Haiti. *Malar. J* 11, 275 (2012). [PubMed: 22889367]
31. Quan H et al. High multiple mutations of *Plasmodium falciparum*-resistant genotypes to sulphadoxine-pyrimethamine in Lagos, Nigeria. *Infect. Dis. Poverty* 9, 91 (2020). [PubMed: 32653033]
32. Yoshida N, Yamauchi M, Morikawa R, Hombhanje F & Mita T Increase in the proportion of *Plasmodium falciparum* with kelch13 C580Y mutation and decline in *pfcrt* and *pfmdr1* mutant alleles in Papua New Guinea. *Malar. J* 20, 410 (2021). [PubMed: 34666779]
33. Hyde JE Drug-resistant malaria. *Trends Parasitol.* 21, 494–498 (2005). [PubMed: 16140578]
34. Harvey WT et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol* 19, 409–424 (2021). [PubMed: 34075212]
35. Brookes D, Park H & Listgarten J Conditioning by adaptive sampling for robust design. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri K & Salakhutdinov R) vol. 97, 773–782 (PMLR, 2019).
36. Sinai S et al. AdaLead: a simple and robust adaptive greedy search algorithm for sequence design. Preprint at <https://arxiv.org/abs/2010.02141> (2020).
37. Tambe A, East-Seletsky A, Knott GJ, Doudna JA & O'Connell MR RNA binding and HEPN-nuclease activation are decoupled in CRISPR–Cas13a. *Cell Rep.* 24, 1025–1036 (2018). [PubMed: 30044970]
38. Abudayyeh OO et al. RNA targeting with CRISPR–Cas13. *Nature* 550, 280–284 (2017). [PubMed: 28976959]
39. Abudayyeh OO et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353, aaf5573 (2016). [PubMed: 27256883]
40. Meeske AJ & Marraffini LA RNA guide complementarity prevents self-targeting in type VI CRISPR systems. *Mol. Cell* 71, 791–801.e3 (2018). [PubMed: 30122537]
41. Welch NL et al. Multiplexed CRISPR-based microfluidic platform for clinical testing of respiratory viruses and identification of SARS-CoV-2 variants. *Nat. Med* 28, 1083–1094 (2022). [PubMed: 35130561]
42. Bock C et al. High-content CRISPR screening. *Nat. Rev. Methods Primers* 2, 9 (2022). [PubMed: 37214176]
43. Huang X, Yang D, Zhang J, Xu J & Chen YE Recent advances in improving gene-editing specificity through CRISPR–Cas9 nuclease engineering. *Cells* 11, 2186 (2022). [PubMed: 35883629]
44. Saha K Accounting for diversity in the design of CRISPR-based therapeutic genome editing. *Nat Genet.* 55, 6–7 (2023). [PubMed: 36593304]
45. Shanehsazzadeh A et al. Unlocking de novo antibody design with generative artificial intelligence. Preprint at <https://www.biorxiv.org/content/10.1101/2023.01.08.523187v1> (2023).
46. Goodfellow IJ et al. Generative adversarial networks. Preprint at <https://arxiv.org/abs/1406.2661> (2014).
47. Mirza M & Osindero S Conditional generative adversarial nets. Preprint at <https://arxiv.org/abs/1411.1784> (2014).
48. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V & Courville A Improved training of Wasserstein GANs. Preprint at <https://arxiv.org/abs/1704.00028> (2017).
49. Kingma DP & Ba J Adam: a method for stochastic optimization. Preprint at <https://arxiv.org/abs/1412.6980> (2014).
50. Yoshida M et al. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* 4, 533–543 (2018).
51. Sinai S & Kelsic ED A primer on model-guided exploration of fitness landscapes for biological sequence design. Preprint at <https://arxiv.org/abs/2010.10614> (2020).
52. Abadi M, et al. TensorFlow: large-scale machine learning on heterogeneous systems (2015).
53. Federhen S The NCBI taxonomy database. *Nucleic Acids Res.* 40, D136–D143 (2012). [PubMed: 22139910]

54. Shu Y & McCauley J GISAID: global initiative on sharing all influenza data—from vision to reality. *Euro Surveil.* 22, 30494 (2017).
55. Katoh K & Standley DM MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol* 30, 772–780 (2013). [PubMed: 23329690]
56. Pickett BE et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 40, D593–D598 (2011). [PubMed: 22006842]
57. Hodcroft EB SARS-CoV-2 mutations and variants of interest. *CoVariants* <https://covariants.org/> (2020).
58. Gangavarapu K et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat. Methods* 20, 512–522 (2023). [PubMed: 36823332]
59. Kellner MJ, Koob JG, Gootenberg JS, Abudayyeh OO & Zhang F SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nat. Protoc* 14, 2986–3012 (2019). [PubMed: 31548639]

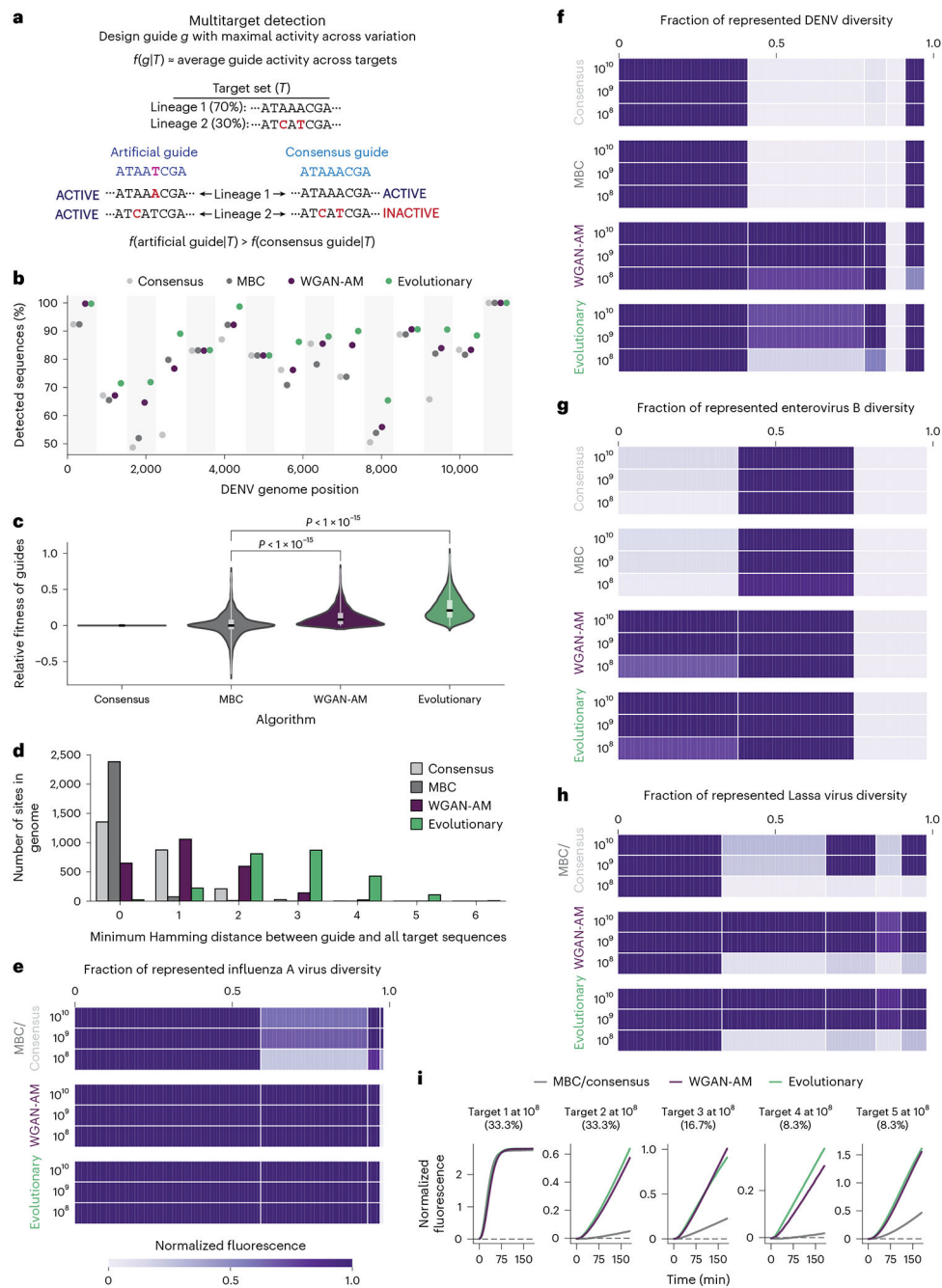


Fig. 1 |. Algorithms in BADGERS design guides that are maximally sensitive across genomic variation.

a, Artificial guide sequences—that is, those that differ from any naturally observed sequence—can increase detection across sequence variation. In the example target set shown, there are two lineages. The consensus of the targets has two mismatches with lineage 2 (in red), which could reduce enzymatic activity and thus sensitivity for that lineage. The artificial guide sequence, which has only one mismatch with lineage 1 and one with lineage 2, could yield superior performance since Cas13a better tolerates a single mismatch than

two mismatches in close proximity. **b**, The proportion of dengue virus (DENV) genomes detected by the diagnostic guides designed with methods implemented in BADGERS (WGAN-AM and evolutionary) and baseline methods (MBC and consensus) within fixed-size windows along the genome. A guide is considered to detect a target if it meets the criteria described in Methods. **c**, The relative fitness of the guides designed by MBC, WGAN-AM and evolutionary algorithms at sites in the DENV genome. The relative fitness is the difference between the fitness of the labeled algorithm's guide and the fitness of the consensus guide. A predictive model of guide–target activity was used to compute these fitnesses, and the distribution is across targetable genomic sites ($n = 2,465$; Methods). *P* values were computed using one-sided Wilcoxon signed-rank tests. **d**, Minimum Hamming distance between the guides designed by each algorithm and all target sequences at a given genomic site. The distribution is shown across targetable genomic sites. **e–h**, Normalized fluorescence for guides detecting representative targets of genomic sites in influenza A virus (**e**), DENV (**f**), enterovirus B (**g**) and Lassa virus (**h**) measured at the 1 h time point. The color bar below **e** and its scale apply to all of **e–h**. Each column represents a target and has width proportional to the percentage of sequence diversity it represents. Each row is a concentration of the target sequence in copies per microlitre, representing postamplification target concentrations. **i**, Normalized fluorescence, over the course of the reaction, when detecting the five representative targets of Lassa virus at 10^8 copies per microlitre. The parentheses indicate the percentage of all genomes represented by the target. The MBC-designed guide and consensus guide were identical at this site and are represented by 'MBC/consensus'.

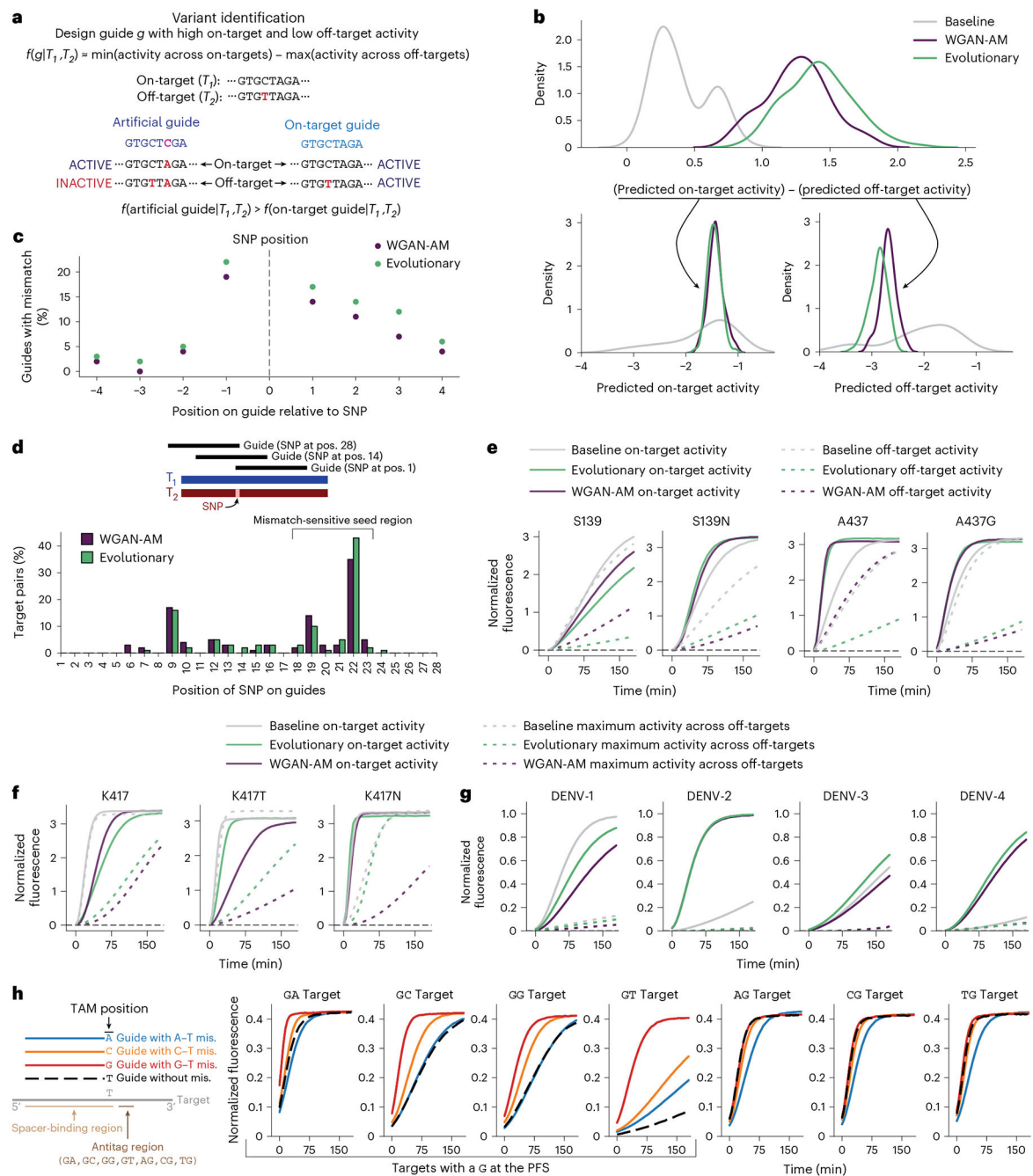


Fig. 2 | Algorithms in BADGERS design optimal guides for variant identification and propose a mechanistic guide design principle.

a, Artificial guide sequences can increase the specificity of a guide designed to discriminate one target from another. A baseline approach is to design a guide whose sequence is directly derived from the on-target sequence. Here, the baseline guide has only one mismatch against the off-target sequence, and thus has substantial off-target activity and achieves poor specificity. An artificial guide sequence that has an additional mismatch introduced could have a double mismatch to the off-target sequence, enabling lower off-target activity and

more robust specificity. **b–d**, Design features of 28 nt guides that differentiate between 100 pairs of synthetic targets such that, in each pair, T_2 has a single-nucleotide difference (1 SNP) with T_1 . For each pair of targets, we applied the algorithms in BADGERS to design guides for each of the 28 possible SNP placements along the guide, and chose the guide with the highest fitness among the 28 as best able to differentiate the two targets. **b**, Benchmarking predicted activities against a widely used design approach that introduces a synthetic mismatch between the guide and its targets (baseline; Methods). Top: distributions of the difference in predicted on-target and off-target activities ($A(g|T_1) - A(g|T_2)$) across the 100 target pairs. Bottom: breakdown of the difference into predicted on-target activity ($A(g|T_1)$; bottom left) and predicted off-target activity ($A(g|T_2)$; bottom right). **c**, Additional guide mismatches were introduced by BADGERS's design algorithms relative to the targets. Namely, the plotted value is the percent of BADGERS-designed guides that have a mismatch against both T_1 and T_2 at positions around the SNP (positive positions are 3' to the SNP and negative are 5'). **d**, The position (pos.) of the SNP within the BADGERS-designed guides. For the majority of target pairs, BADGERS's design algorithms place the SNP within Cas13a's mismatch-sensitive seed region. **e**, Experimental results of guides designed with different approaches for identifying the S139N SNP in Zika virus and A437G antimalarial resistance SNP in *P. falciparum* at a target concentration of 10^8 copies per microlitre. The title of each plot indicates the target the guides were designed to detect as the on-target. **f**, Normalized fluorescence over time of guides designed to identify the K417N/T SNP in SARS-CoV-2 at 10^8 copies per microlitre. **g**, Normalized fluorescence over time of guides designed to identify each of the four DENV serotypes at 10^6 copies per microlitre. In **f** and **g**, there are multiple off-targets for each discrimination task, so the dotted off-target curve shows the maximum fluorescence of the guide across the off-targets computed at each time point (for example, the off-target curves for K417 represent the fluorescence for the higher of the K417N and K417T targets at each time point in the reaction). **h**, Evaluation of the effect of the tag-adjacent mismatch (mis.) on rescuing guide activity when there are unfavorable (GN) nucleotides at the antitag region. The curves show the normalized fluorescence of guides detecting targets that have different nucleotide pairs at the antitag region, at 10^7 copies per microlitre. All guides are identical to the targets at positions 1–27 in the protospacer. The colored lines represent guides with a terminal adjacent mismatch (at position 28), while the dashed black lines represent guides without a mismatch to the target. The dinucleotide above each plot indicates the first two nucleotides of that target's antitag region, and the guide sequences represented in the schematic are reverse complemented.