

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

In the format provided by the
authors and unedited

Supplementary Note 1

This note estimates the size of the sequence landscape that must be searched over to generate artificial guides.

The total number of possible 28-nt guides is $4^{28} \approx 7 \times 10^{16}$.

Optimally fit guide sequences would be nearby in sequence space to the complement of their target sequence. If M is the upper limit on the number of mismatches an optimal guide can have to this sequence, there are $\sum_{i=0}^M \binom{28}{i} \cdot 3^i$ in this target-adjacent guide set. This totals to $\sim 2.5 \times 10^7$ and $\sim 2.9 \times 10^9$ for $M = 5$ and $M = 7$, respectively. In many applications, we would explore this space not just once, but at every site in a targeted genome (tens to hundreds of thousands of sites for a typical viral genome); moreover, if a set of targets has high variation, we would explore the space around many observed alleles of the target. It would be inefficient to explore this vast space of potential guide sequences with an exhaustive, brute-force search.

Supplementary Note 2

Hyperparameter Search

We implemented a random search procedure to determine the optimal hyperparameters for our design algorithms. Because we developed two model-directed exploration algorithms (WGAN-AM and evolutionary) and we have two objective functions ($f_M(g|T)$ and $f_D(g|T_1, T_2)$), four sets of hyperparameters were determined by random search.

Multi-target detection objective function

For the multi-target detection objective function, $f_M(g|T)$, we constructed a random search dataset of genomes from NCBI's GenBank database¹. We randomly sampled ten viral families from a list of all viral families that have at least one vertebrate-infecting species. Then, within each of these ten families, we randomly sampled one vertebrate-infecting species that has at least 100 complete genomes. The ten selected viral species were Primate T-lymphotropic virus 1 (NCBI taxonomic identifier (taxid): 194440), avian orthoreovirus (taxid: 38170), alphapapillomavirus 10 (taxid: 333754), yellow fever virus (taxid: 11089), eastern equine encephalitis virus (taxid: 11021), pigeon circovirus (taxid: 1414603), SFTS phlebovirus (taxid: 1933190), human respirovirus 3 (taxid: 11216), human metapneumovirus (taxid: 162145), and Betacoronavirus 1 (taxid: 694003). For each of these ten viral species, we used ADAPT² to compile and curate complete genomes and used MAFFT³ create an alignment of the curated genomes. We randomly selected ten guide-length sites from each of the ten alignments, and extracted all of the sequences at each genomic site to create a target set T . In summary, we constructed 100 target sets from 100 genomic sites across ten different viral species—allowing our hyperparameter search process to reflect viral diversity.

We created 100 sets of hyperparameters for the algorithms by randomly sampling from the distributions listed [below](#). For each set of hyperparameters, we ran the WGAN-AM and evolutionary algorithms across all 100 target sets, and computed the mean fitness of the guides across the target sets ($\frac{1}{100} \sum_{i=1}^{100} f_M(g|T_i)$, for a set of targets T_i at each site). We ultimately selected the set of hyperparameters that the maximized mean guide fitness.

Variant identification objective function

For the variant identification objective function, $f_D(g|T_1, T_2)$, we created a synthetic random search dataset. We randomly generated 50 DNA sequences of length 150 to form the on-target sets $T_1^1, \dots, T_1^{50}$. For each of these DNA sequences, we mutated the base in the center of the sequence (position 75) to another randomly-selected base. These mutated DNA sequences comprised the off-target sets, $T_2^1, \dots, T_2^{50}$. These pairs of on-target and off-target sets $(T_1^1, T_2^1), \dots, (T_1^{50}, T_2^{50})$ represent two different sequences, one with a single nucleotide mutation and one without a single nucleotide mutation.

We created 100 sets of hyperparameters the algorithms by randomly sampling from the distributions listed [below](#). Furthermore, since the variant identification objective function $f_D(g|T_1, T_2)$ has hyperparameters as well, we created 100 sets of objective function hyperparameters by randomly sampling from the distributions also listed [below](#).

For each set of hyperparameters, we ran the algorithms across the 50 pairs of on-target and off-target sets $(T_1^1, T_2^1), \dots, (T_1^{50}, T_2^{50})$ in a sliding window fashion so that we would examine a guide placing the SNP at each position of the guide (window length = 28 nt, window stride = 1 nt; selecting the optimal of the 28 guides) to design 50 diagnostic guides. Because the objective function $f_D(g|T_1, T_2)$ has its own hyperparameters k , o , a , and r , it would not be appropriate to directly compare the fitness of the generated guides across the different hyperparameter sets. Since our goal is to design guides that have high on-target activity and low off-target activity, we selected the set of hyperparameters that maximized the mean difference between predicted on-target and off-target activity ($\frac{1}{50} \sum_{i=1}^{50} [A(g|T_1^i) - A(g|T_2^i)]$) while simultaneously generating guides with a high predicted on-target activity ($A(g|T_1^1), \dots, A(g|T_1^{50}) > -1.7$).

For both objectives, the evolutionary, AdaLead, and CbAS algorithms were provided a fixed number of model evaluations ($L = 2000$). Since the WGAN-AM algorithm’s hyperparameters influence L , a hyperparameter set was chosen such that the algorithm had L within 10% of 2000.

List of hyperparameters and search intervals

WGAN-AM algorithm parameters determined through random search:

- Number of starting points, $r_{outer} \sim$ uniform in $[5, 35]$ for multi-target detection objective, $r_{outer} \sim$ uniform in $[5, 15]$ for variant identification objective
- Number of steps in each search of the latent space, $r_{inner} \sim$ uniform in $[50, 325]$ for multi-target detection objective, $r_{inner} \sim$ uniform in $[50, 230]$ for variant identification objective
- optimizer learning rate, $\alpha \sim$ log-uniform in $[10^{-2}, 10^2]$

Evolutionary algorithm parameters determined through random search:

- Selection intensity, $\beta \sim$ log-uniform in $[10^{-1.2}, 10^{2.2}]$ for multi-target detection objective, $\beta \sim$ log-uniform in $[10^{-2}, 10^2]$ for variant identification objective
- Mutation rate, $\gamma \sim$ uniform in $[\frac{0}{28}, \frac{10}{28}]$ for multi-target detection objective, $\gamma \sim$ uniform in $[\frac{0}{28}, \frac{2}{28}]$ for variant identification objective
- Fraction of children in new generation, $j \sim$ uniform in $[0.1, 0.9]$
- Population size, $S \sim$ uniform in $[50, 300]$ for multi-target detection objective, $S \sim$ uniform in $[50, 250]$ for variant identification objective

AdaLead parameters determined through random search:

- Expected recombination partners per sequence, $\rho \sim$ discrete uniform in $[0, 2]$
- Recombination rate, $j \sim$ uniform in $[0, 0.5]$
- Expected mutations to sequence, $\omega \sim$ discrete uniform in $[1, 10]$ for multi-target detection objective, $\omega \sim$ discrete uniform in $[1, 2]$ for variant identification objective
- Fractional threshold $\kappa \sim$ uniform in $[0, 0.5]$

CbAS parameters determined through random search:

- Quantile threshold, $Q \sim$ uniform in $[0.5, 0.99]$

Variant identification objective function parameters determined through random search:

- $a \sim$ uniform in $[0.1, 9.1]$
- $k \sim$ uniform in $[-4, -1]$
- $r \sim$ uniform in $[0.1, 10.1]$
- $o \sim$ uniform in $[-4.5, -1]$

Supplementary Figures

Algorithm 1 WGAN-AM algorithm for guide design.

Input

T target set
 $f(g|T)$ objective function that returns the fitness of a guide g against T
 $G(z|T)$ pre-trained generator (from WGAN) function that generates a guide given the latent variable z and T
 r_{outer} number of randomly-selected starting points for the latent space search
 r_{inner} number of steps taken by the Adam optimizer in each latent space search
 $\text{Adam}(\nabla_z f(g|T), \alpha)$ Adam optimizer function that uses information about the gradient to ‘step’ the latent variable in the direction that increases fitness, where α is its learning rate

Output

set of generated guide sequences

```
1 function WGAN-DESIGN-GUIDE( $f, T, G, r_{outer}, r_{inner}, \alpha$ )
2    $Q \leftarrow \emptyset$ 
3   for  $i \leftarrow 1$  to  $r_{outer}$  do
4      $z_1 \sim \mathcal{N}(0, 1)$  ▷ Sample from normal dist. to obtain latent variable
5     for  $j \leftarrow 1$  to  $r_{inner}$  do
6        $g \leftarrow G(z_j|T)$  ▷ Use WGAN to generate guide  $g$ 
7        $z_{j+1} \leftarrow \text{Adam}(\nabla_{z_j} f(g|T), \alpha)$ 
8      $Q \leftarrow Q \cup \{G(z_{r_{inner}+1}|T)\}$ 
9   return  $Q$ 
```

Supplementary Figure 1 — Pseudocode of the WGAN-AM algorithm.

The WGAN-AM algorithm is further described in the Methods section.

Algorithm 2 Evolutionary algorithm for guide design.

Input

T	target set
$f(g T)$	objective function that returns the fitness of a guide g against T
β	selection intensity
I	initial set of diagnostic guides, generally the sequences extracted at the targeted genomic site in the multiple sequence alignment
γ	mutation rate
S	population size
L	limit on number of calls of the objective function, $f(g T)$
j	proportion of each generation composed of children

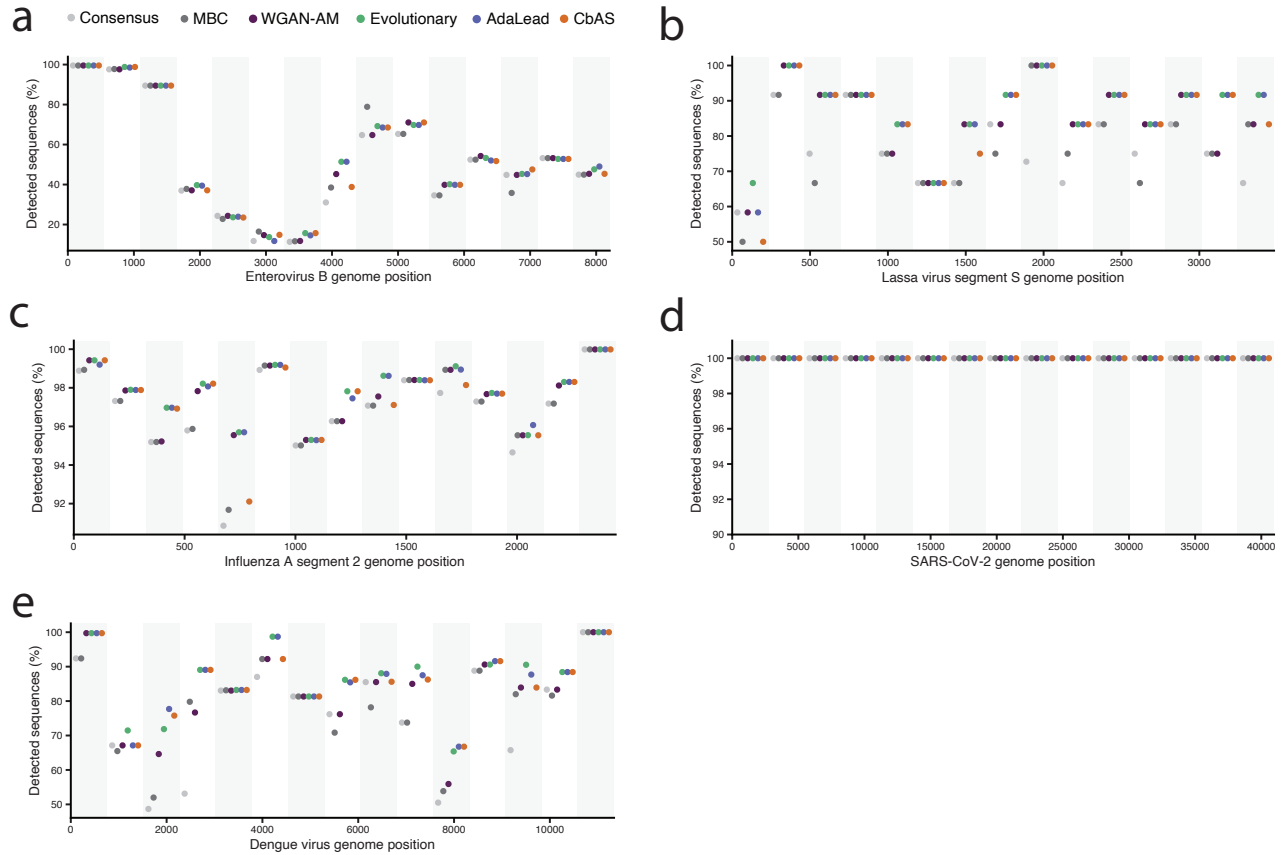
Output

set of generated guide sequences

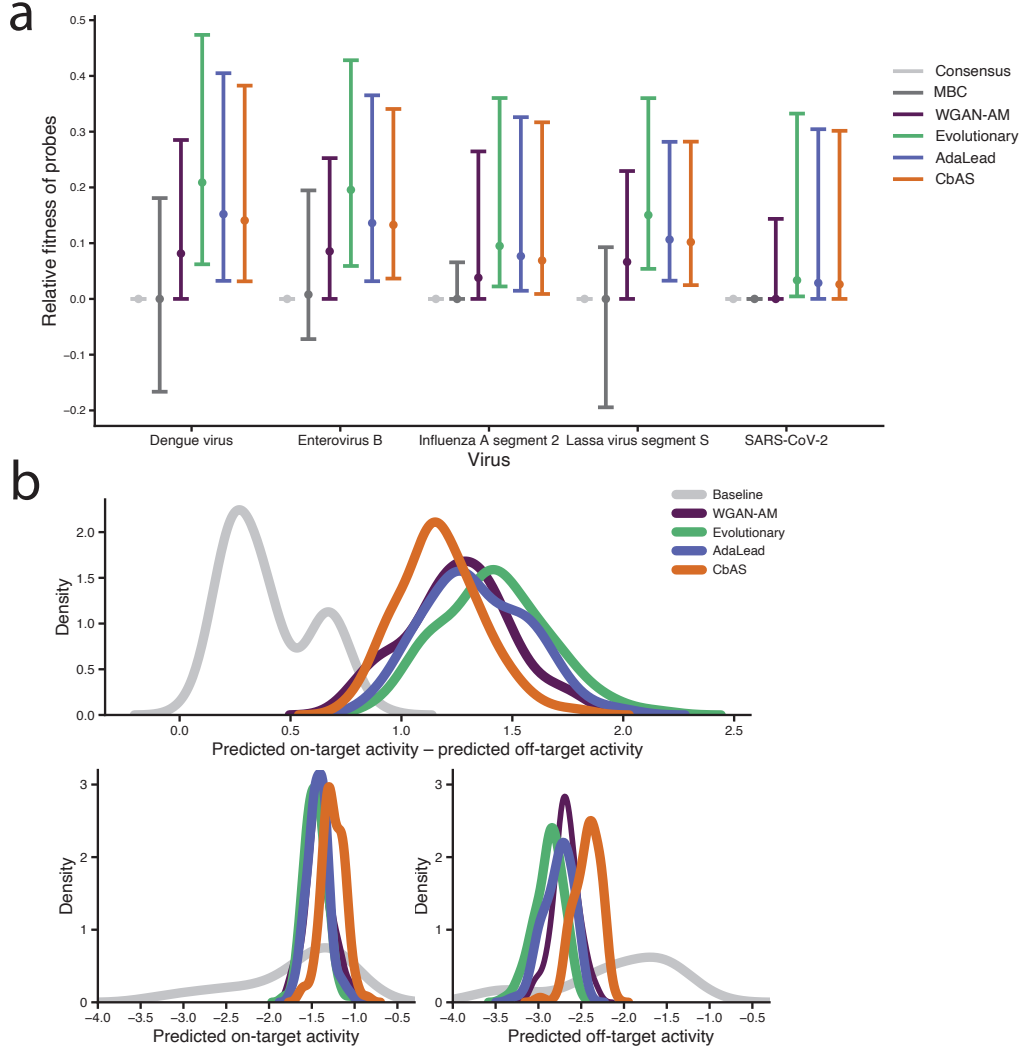
```
1 function EVOLUTIONARY-DESIGN-GUIDE( $f, \beta, \gamma, S, L, I, T, j$ )
2    $Q \leftarrow$  Sample  $S$  guides  $(a_1, \dots, a_S)$  from  $I$  with probability  $\frac{e^{\beta \cdot f(a_i|T)}}{\sum_{i=1}^{|I|} e^{\beta \cdot f(a_i|T)}}$ 
3    $z \leftarrow |Q| \cdot j$ 
4    $M \leftarrow 0$ 
5   while  $M < L$  do
6      $G \leftarrow$  Sample  $z$  guides  $(b_1, \dots, b_z)$  from  $Q$  with probability  $\frac{e^{\beta \cdot f(b_i|T)}}{\sum_{i=1}^{|Q|} e^{\beta \cdot f(b_i|T)}}$ 
7      $Q \leftarrow \text{Sort}(Q)$  ▷ Sort guides in  $Q$  by their fitness, highest to lowest
8      $Q \leftarrow Q[0 : S - z]$  ▷ Remove  $z$  lowest-fitness guide sequences from  $Q$ 
9     for  $g_i \in G$  do
10       $c_i \leftarrow \text{Mutate}(g_i, \gamma)$  ▷ Introduce random mutations to guide with frequency  $\gamma$ 
11       $Q \leftarrow Q \cup \{c_i\}$ 
12       $M \leftarrow M + 1$ 
13   return  $Q$ 
```

Supplementary Figure 2 — Pseudocode of the evolutionary algorithm.

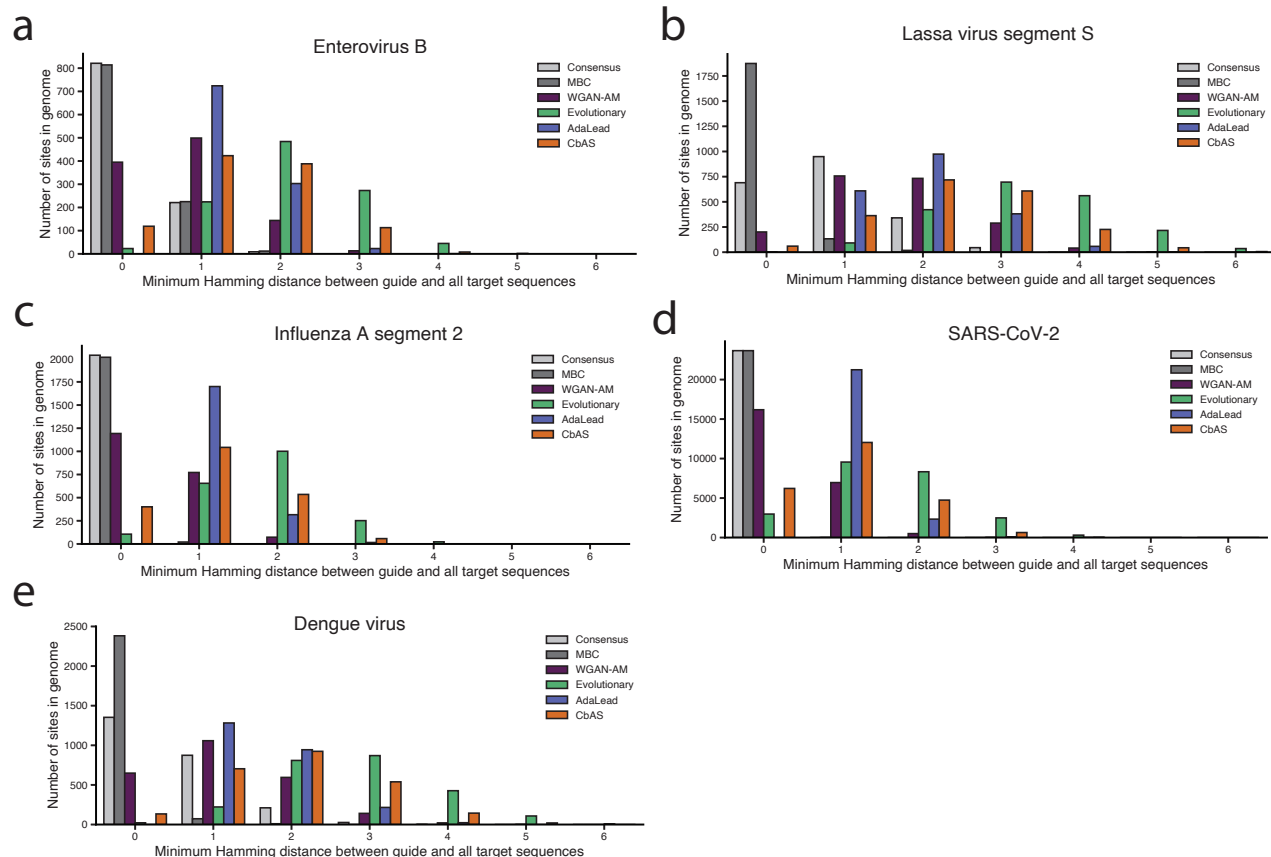
The evolutionary algorithm is further described in the Methods section.



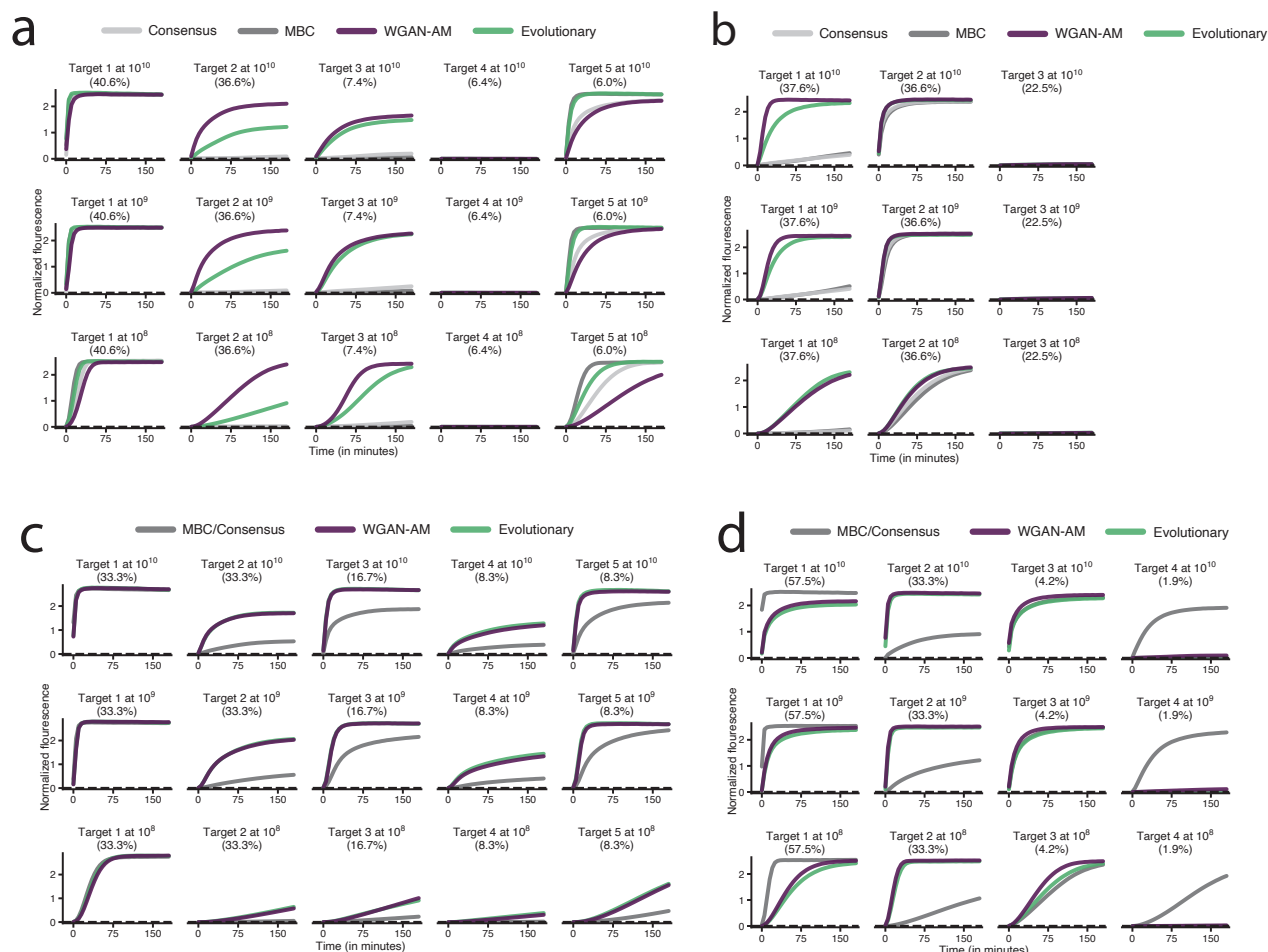
Supplementary Figure 3 — Coverage of guides designed by the generative design algorithms and baseline methods. Proportion of genomes predicted to be detected by guides designed by different methods. **(a)** enterovirus B, **(b)** Lassa virus segment S, **(c)** influenza A virus segment 2, **(d)** SARS-CoV-2, and **(e)** dengue virus. A guide is considered to detect a target if it meets the criteria described in the Methods.



Supplementary Figure 4 — Relative performance of generative design algorithms. (a) Relative fitness of the guides designed by the baseline methods and the generative design methods across the five viruses considered for the multi-target detection objective. For each distribution, the point represents the median, lines above and below the point represent the 10th percentile and 90th percentile, respectively, across targetable genomic sites (see Methods). The relative fitness is the difference between the fitness of the labeled algorithm's guide and the fitness of the consensus guide. All the generative design methods outperformed baseline approaches; across these 5 viruses, the design algorithms achieved an average relative fitness ($\frac{1}{\text{total \# of sites across all viruses}} \sum_{i=1}^5 \sum_{j=1}^{\# \text{ of sites in virus}_i} [\text{relative fitness of guide at site}_j \text{ in virus}_i]$) of: consensus, 0.000; MBC, 0.006; WGAN-AM, 0.090; evolutionary, 0.177; AdaLead, 0.140; CbAS, 0.134. **(b)** Distributions of predicted activities for the guides designed for the variant identification objective. Top, distributions of the difference in predicted on-target and off-target activities ($A(p|T_1) - A(p|T_2)$) across the 100 target pairs for the baseline and generative sequence design methods. Bottom, break-down of the difference into predicted on-target activity ($A(p|T_1)$; bottom left) and predicted off-target activity ($A(p|T_2)$; bottom right). The mean difference in predicted on-target and off-target activities was: baseline, 0.399; WGAN-AM, 1.26; evolutionary, 1.42; AdaLead, 1.34; CbAS: 1.17.



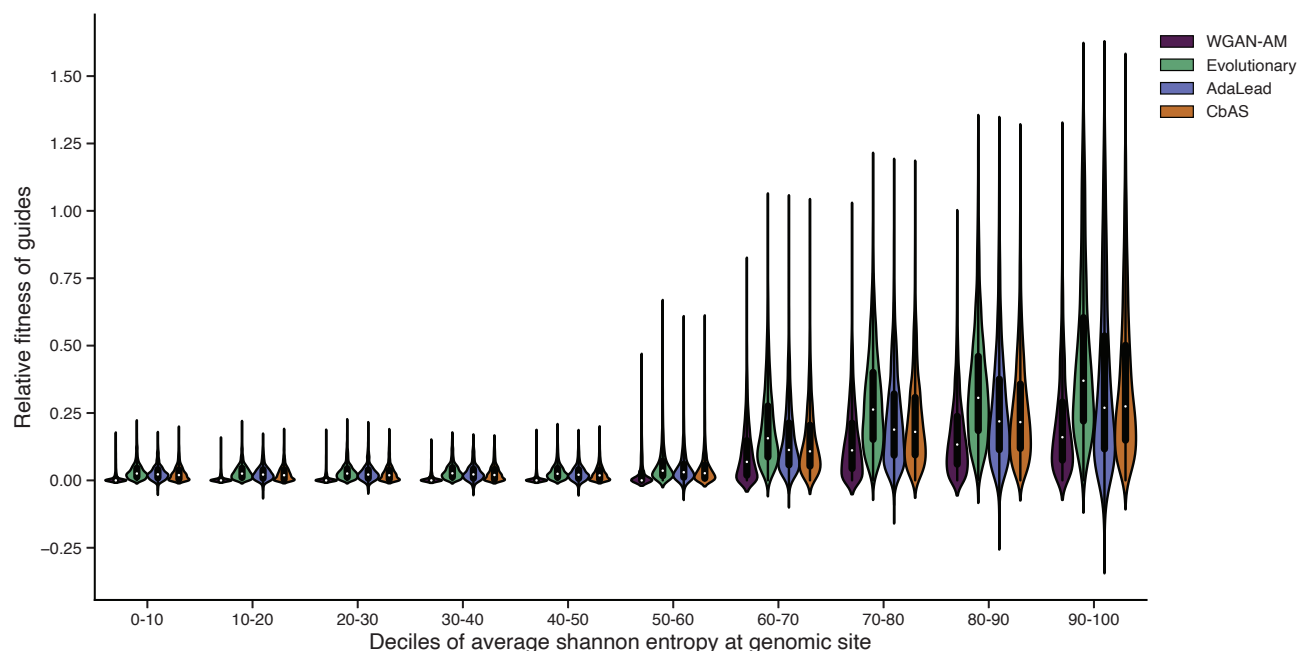
Supplementary Figure 5 — Divergence of guides from observed sequences. Histograms representing the minimum Hamming distance between the guides compared to all target sequences at a given genomic site—that is, the Hamming distance between a guide and the target sequence most similar to that guide. Shown for **(a)** enterovirus B, **(b)** Lassa virus segment S, **(c)** influenza A virus segment 2, **(d)** SARS-CoV-2, and **(e)** dengue virus. The generative sequence design methods (WGAN-AM, evolutionary, AdaLead, and CbAS) tend to produce guides that are more dissimilar to any of the target sequences (that is, more artificial) than the baseline algorithms (consensus and MBC).



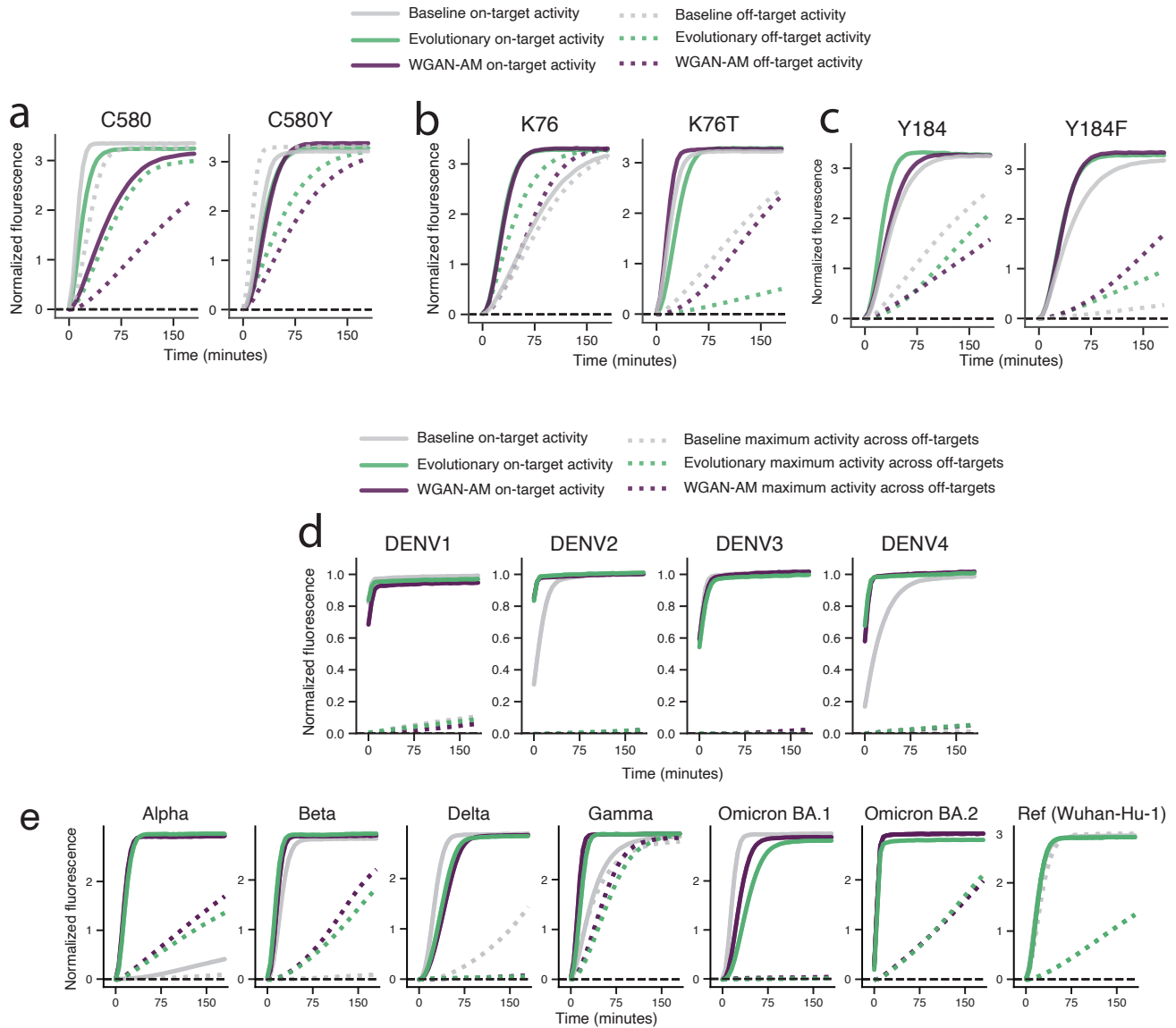
Supplementary Figure 6 — Kinetic curves for multi-target detection tasks. Normalized fluorescence over the course of the reaction for the guides targeting the genomic sites whose heatmaps are shown in Fig. 1. Parentheticals indicate the percentage of all genomes represented by the target and concentrations of targets are indicated above each plot with units of copies/ μ L. **(a)** Kinetic curves for a site in dengue virus (heatmap in Fig. 1f). **(b)** Kinetic curves for a site in enterovirus B (heatmap in Fig. 1g). **(c)** Kinetic curves for a site in Lassa virus segment S (heatmap in Fig. 1h). **(d)** Kinetic curves for a site in influenza A virus segment 2 (heatmap in Fig. 1e).



Supplementary Figure 7 — Experimental performance of multi-target detection guides. Each panel in this figure shows the experimental performance of guides designed for a specific genomic site. Panels (a–f) represent genomic sites in the top quartile of relative performance, while panels (g–k) represent genomic sites in the bottom quartile of relative performance (RP; see Methods for definition). The experimental data for the other four genomic sites in the top quartile of relative performance is shown in Fig. 1 and Supplementary Fig. 6. The left side of each panel is a heatmap representing fluorescence at 1 hour. Each column represents a target and has width proportional to the percentage of sequence diversity it represents, while each row is a concentration of the target sequence in copies/ μ L. The right side of each panel is a set of curves representing the normalized fluorescence of the guides against the specified targets. The color of these curves matches the color of the design method used to label each heatmap. Parentheticals indicate the percentage of all genomes represented by the target. **(a)** Site in dengue virus (top quartile of RP). **(b)** Site in enterovirus B (top quartile of RP). **(c)** Site in Lassa virus segment S (top quartile of RP). **(d)** Site in influenza A virus segment 2 (top quartile of RP). **(e)** Site in SARS-CoV-2 (top quartile of RP). **(f)** Site in SARS-CoV-2 (top quartile of RP). **(g)** Site in dengue virus (bottom quartile of RP). **(h)** Site in influenza A (bottom quartile of RP). **(i)** Site in enterovirus B (bottom quartile of RP). **(j)** Site in SARS-CoV-2 (bottom quartile of RP). **(k)** Site in Lassa virus (bottom quartile of RP).

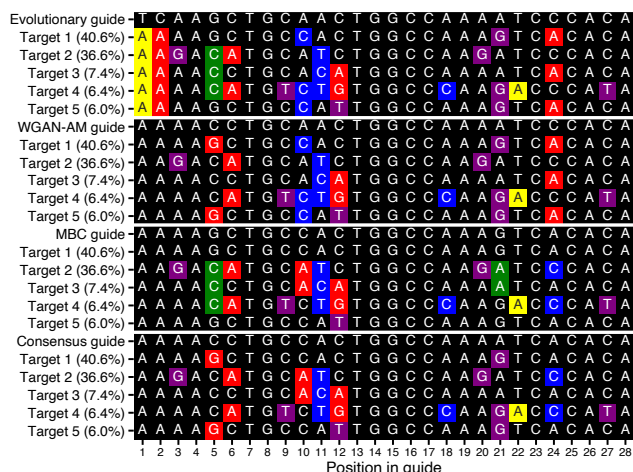


Supplementary Figure 8 — Improvement of the generative design algorithms over baseline methods when targeting variably diverse genomic sites. Relative improvement of generative algorithm-designed guides over the consensus guides grouped by the Shannon entropies of the genomic sites. The distribution is across all genomic sites in the 5 viral species considered for the multi-target detection objective (dengue virus, influenza A virus, enterovirus B, Lassa virus, and SARS-CoV-2) with a non-G PFS ($n = 33,939$). Relative fitness is defined as the fitness of the guide subtracted by the fitness of the consensus guide at that genomic site. By definition, if a guide has a relative fitness greater than zero, it is more fit than the consensus guide. The dots in the center of each violin represent the mean relative fitness for that decile.

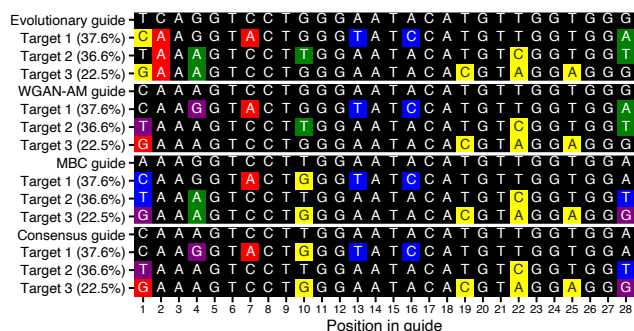


Supplementary Figure 9 — Experimental performance of guides designed for SNP and variant identification tasks. Normalized fluorescence over time is shown for (a) C580Y in Pfk13, (b) K76T in Pfcrt, (c) Y184F in Pfmdr1, (d) dengue virus serotypes 1-4, and SARS-CoV-2 lineages. The title of each plot indicates the target the guides were designed to detect as the on-target. On-target solid curves represent the fluorescence of the guide for its on-target and the dotted curves represent the fluorescence of the guide against its off-target. For panels (d) and (e), 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 DENV1 represent the maximal fluorescence across the DENV2, DENV3, and DENV4 targets at each time point in the reaction). Furthermore, in (e), for the Omicron BA.2 on-target, the WGAN-AM and baseline methods designed the same guide (purple). For the Ref on-target, the WGAN-AM and evolutionary methods designed the same guide (green). All targets were present at a concentration of 10^8 copies/ μ L.

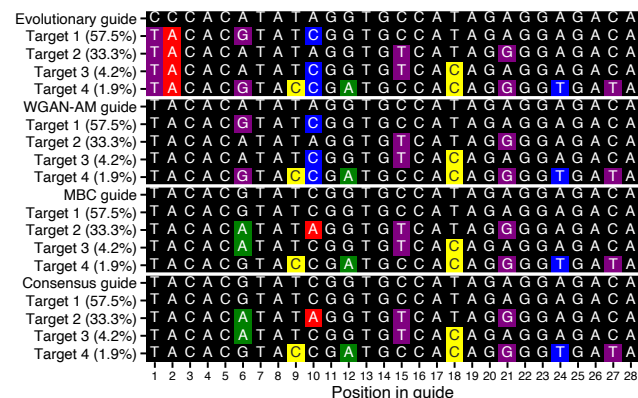
a



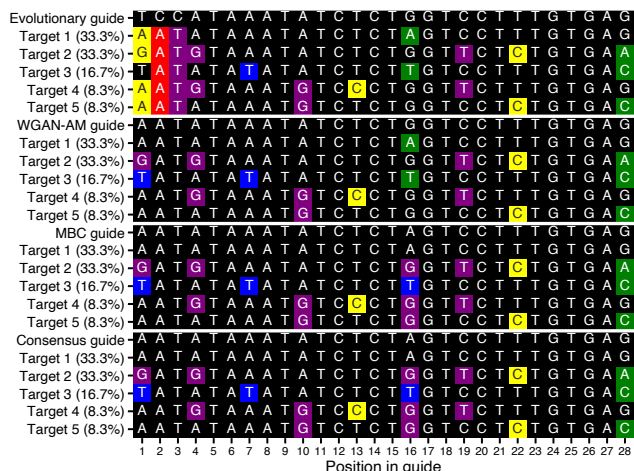
b



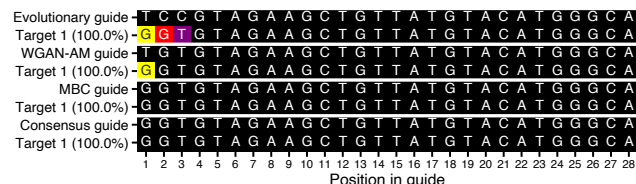
c



d

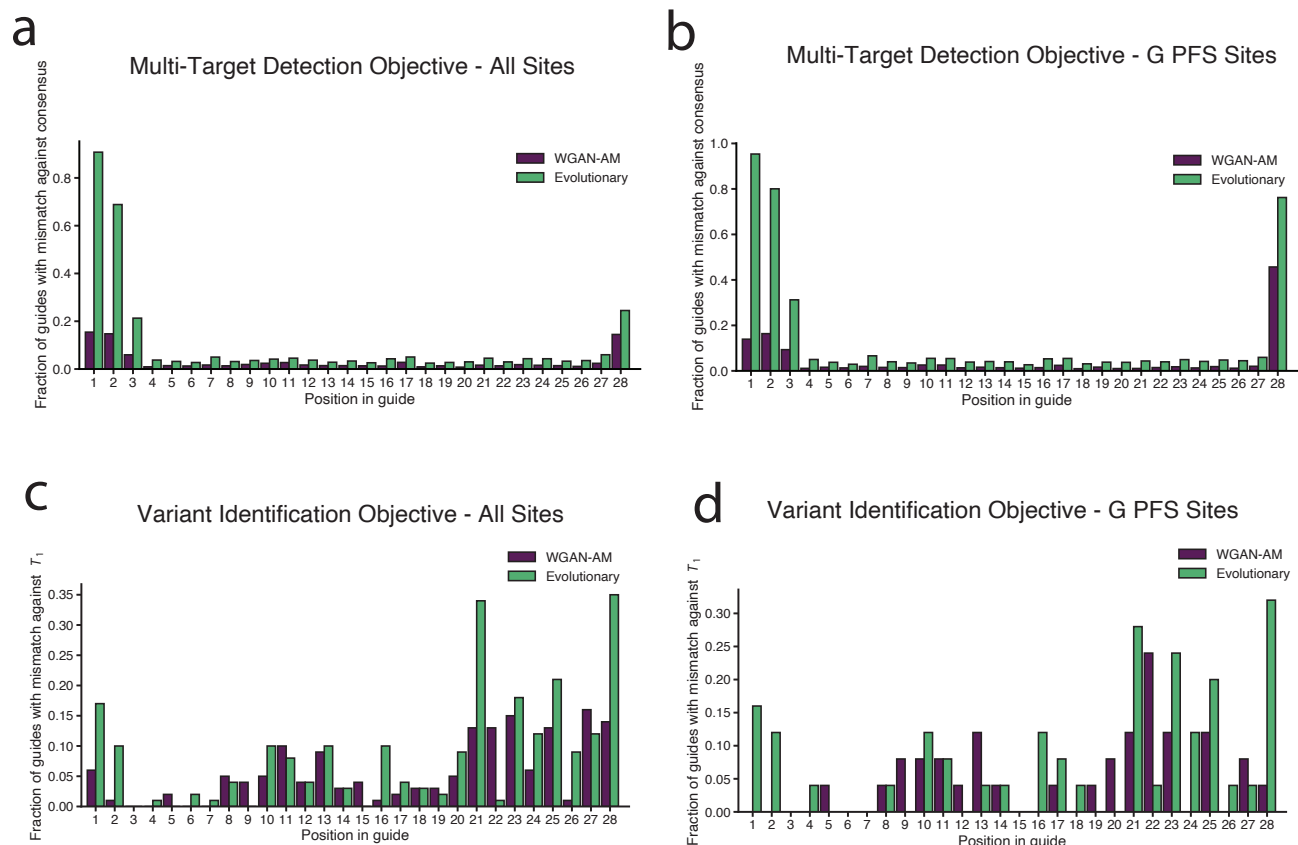


e

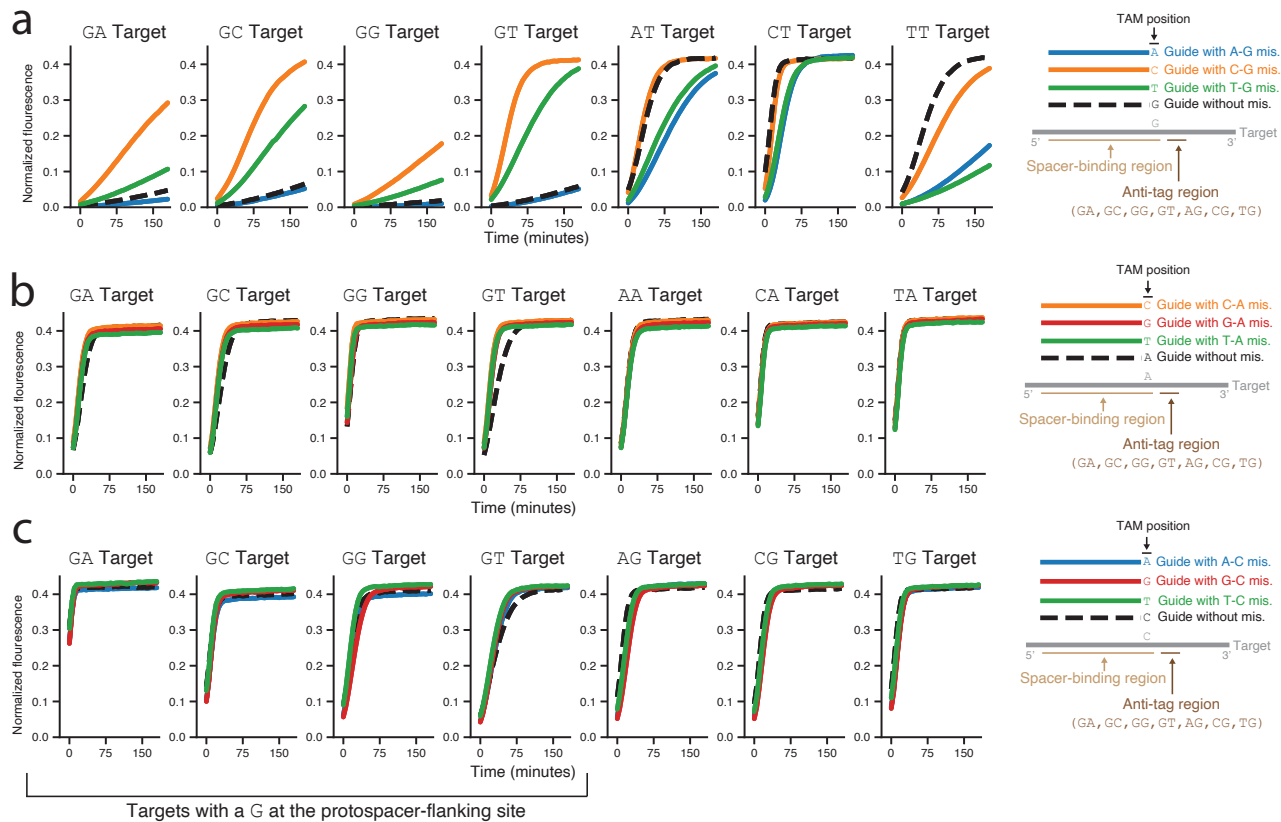


Supplementary Figure 10 — Visualization of guide-target mismatches in guides designed for the multi-target detection objective. Each panel shows guide sequences and representative target sequences for one experimentally-tested genomic site in each of the five viral pathogens considered. The guides are reverse-complemented, from their sequence in a crRNA, to be in the same frame as the target sequences, and both the guides and targets are shown as DNA sequences (with T replacing U). The black blocks indicate that there is no mismatch between the guide and target, while the blue, green, yellow, and red blocks represent A-B, G-H, T-V, and C-D mismatches respectively. The blocks in purple represent G-U wobble RNA base pairing. Parentheticals next to target names indicate the percent of sequence variation represented by that target. **(a)** A site in dengue virus. The baseline guides have mismatches with target 2 at positions 10 and 11 and also have mismatches with target 3 at positions 10, 11, and 12. The algorithms in BADGERS mutate the guide at position 10 to remove a mismatch at this position with targets 2 and 3, but introduce a mismatch at this position with targets 1, 4, and 5. Experimentally, the baseline guides have nearly no fluorescence on targets 2 and 3; in contrast, the BADGERS-designed guides achieve strong fluorescence on targets 2 and 3, while retaining similar performance on targets 1, 4, and 5 (Fig. 1f). Thus, the BADGERS-designed guides can detect nearly 44% more sequence diversity. (Legend continued on the subsequent page.)

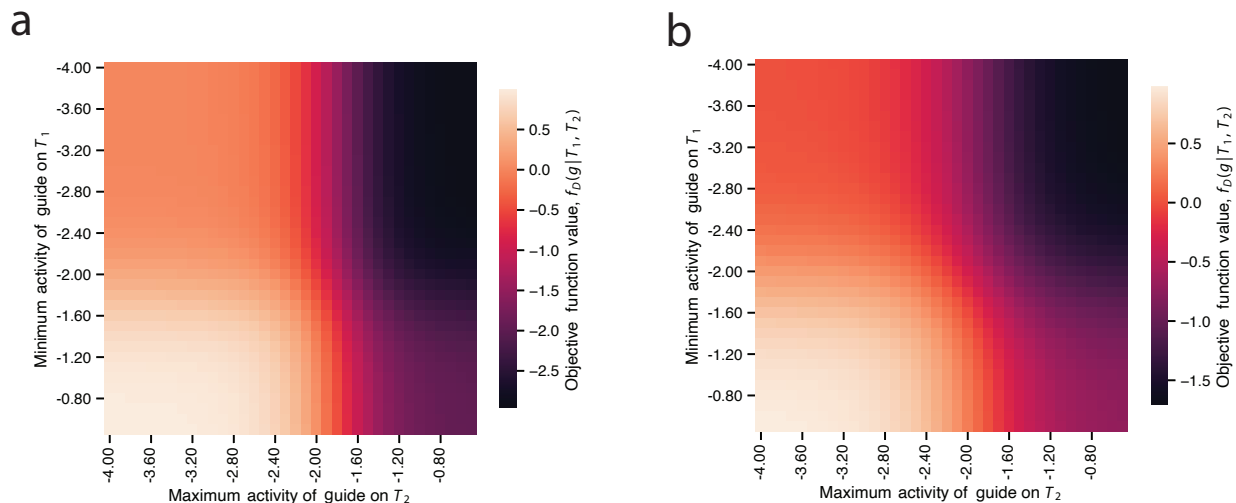
Supplementary Figure 10 — **(b)** A site in enterovirus B. Here, the guides designed by the baseline approaches have three mismatches in close proximity of one another (at positions 10, 13, and 16) with target 1, the target most representative of sequence variation. The algorithms in BADGERS mutate the nucleotide at position 10 and avoid the deleterious effect of this triple mismatch, but do introduce a mismatch with target 2 at this position. The baseline guides achieve nearly no fluorescence on target 1, while the BADGERS-designed guides are highly active on target 1, and still retain robust activity on target 2 (Fig. 1g). **(c)** A site in influenza A virus segment 2. The baseline guides have mismatches with target 2 at positions 6 and 10. The algorithms in BADGERS eliminate the position 6 mismatch against target 2 while not introducing mismatches to other targets (e.g., target 1) by employing G-U wobble base pairing between the RNAs. The methods implemented in BADGERS also mutate the nucleotide at position 10 to remove the mismatch against target 2, but by doing this, they introduce a mismatch at position 10 against target 1. These sequence changes enable the BADGERS-designed guides to nearly triple the fluorescence against target 2 as the baseline guides. However, because they do introduce a mismatch to target 1, they have slightly lower fluorescence at the initial timepoints against this target but still achieve saturating fluorescence by the later timepoints of the reaction (Fig. 1e). **(d)** A site in Lassa virus segment S. At this site, BADGERS-designed guides achieve substantially greater fluorescence on both targets 2 and 4 (Fig. 1h). From the pattern of guide-target mismatches, it is not clear why the BADGERS-designed guides perform better. As one hypothesis, the close proximity of the instances of G-U wobble base pairing at positions 16 and 19 of the baseline guides might be deleterious to activity; here, the algorithms in BADGERS might be exploiting features of Cas13a targeting that are not currently well-characterized. **(e)** A site in SARS-CoV-2. Previous work (Fig. 2e in ref. 2) has shown that a T-V guide-target mismatch at position 1 is strongly associated with increased guide activity. Both the WGAN-AM and evolutionary algorithms introduce this mismatch. Experimentally, the evolutionary and WGAN-AM guides both achieve slightly greater fluorescence than the baseline (MBC and consensus) guides, likely due to their position 1 mismatch (Supplementary Fig. 7e).



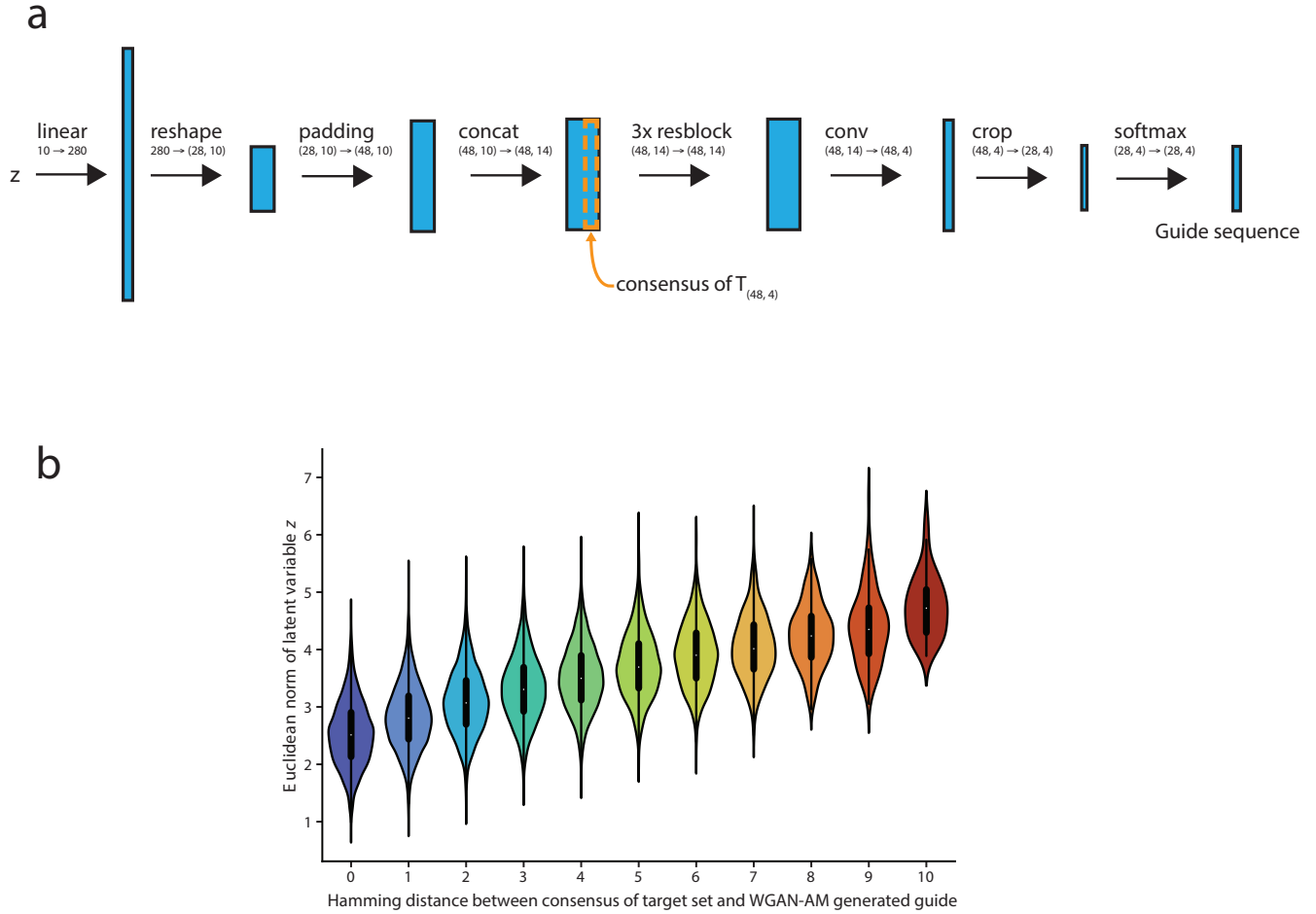
Supplementary Figure 11 — Divergence of BADGERS-designed guides from the consensus sequence. Fraction of BADGERS-designed guides that have a mismatch against the consensus genome sequence, at each position in the guide-target pairing. **(a)** Fractions taken across all genomic sites from all of the five viral pathogens considered in the multi-target detection objective (dengue virus, influenza A virus, enterovirus B, Lassa virus, and SARS-CoV-2). **(b)** Fractions taken across only the genomic sites with a G nucleotide at the PFS, again across all of the five viral pathogens considered in the multi-target detection objective. On this subset of sites, the BADGERS-designed guides are relatively likely to have a terminal mismatch in the protospacer, suggesting the benefit of such a mismatch when there is a G nucleotide at the PFS. **(c)** Fractions taken across all genomic sites considered in the variant identification detection objective. **(d)** Fractions taken across only the genomic sites with a G nucleotide at the PFS amongst the sites considered in the variant identification objective. In (c) and (d), T_1 is the on-target sequence.



Supplementary Figure 12 — Evaluation of the ability of the tag-adjacent mismatch to rescue guide activity for different target sets. Normalized fluorescence curves are shown for the target sets where the tag-adjacent position (TAM position) in the target is a (a) G nucleotide, (b) A nucleotide, or (c) C nucleotide. All guides are identical to the targets at positions 1-27 in the protospacer. The colored lines represent guides with a tag-adjacent mismatch (at position 28), while the black dashed lines represent guides without a mismatch to the target. The dinucleotide above each plot indicates the first two nucleotides of that target's anti-tag region, and the guide sequences represented in the schematic are reverse-complemented compared to the spacer sequence in the CRISPR RNA. The same curves for the target set with a T nucleotide at the tag-adjacent position are shown in Fig. 2h.



Supplementary Figure 13 — Visualization of the fitness function for the variant identification objective. Colors represent the value of the variant identification fitness function according to on-target (T_1) and off-target (T_2) activities (see Methods for the function). A maximally-fit guide g has a low maximum activity across the off-target set T_2 , $\max_{t \in T_2} A(g|t)$, and a high minimum activity across the on-target set T_1 , $\min_{t \in T_1} A(g|t)$. **(a)** Using hyperparameters determined by the WGAN-AM random search. **(b)** Using hyperparameters determined by the evolutionary algorithm random search.



Supplementary Figure 14 — Properties of WGAN-AM algorithm. **(a)** Architecture of generative model of Wasserstein generative adversarial network (WGAN) for generating guides, inspired by that of ref. 4. The input into the generative model, the latent vector z , is a 10-dimensional vector sampled from the standard normal distribution. z is upsampled through a linear layer and reshaped and padded to create a 48 by 10 matrix. In the concat operation, the one-hot-encoded consensus sequence of the target set T (with dimensions 48 by 4) is concatenated with the output of the previous layer to create a 48 by 14 matrix. This matrix is passed through three residual blocks. Each residual block (resblock) consists of two layers, with each layer containing a 1D convolutional layer with 14 filters of stride 1 and width 3. Next, in the conv operation, a convolutional layer with 4 filters of width 1 and stride 1 is applied to the output of the resblock to create 4-channel encoded guide sequence. Finally, the matrix is cropped to remove the 10 by 4 context on each side of the guide protospacer sequence, and the resulting 28 by 4 matrix is passed through a softmax layer so that each element in the matrix represents the probability of having a certain base at that position. The guide sequence contains the base at each position with the greatest probability. **(b)** Relationship between the Euclidean norm of the latent variable input to the WGAN's generator network and the number of mutations introduced into the guide relative to the consensus sequence for 500 generated guides. All guides with a Hamming distance ≤ 10 are shown ($n = 49,982$), and the computational experiment was performed as described in the Methods.

References

- [1] Federhen, S. The NCBI taxonomy database. *Nucleic Acids Research* **40**, D136–43 (2012).
- [2] Metsky, H. C. *et al.* Designing sensitive viral diagnostics with machine learning. *Nat Biotechnology* **40**, 1123–1131 (2022).
- [3] Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
- [4] Killoran, N., Lee, L. J., DeLong, A., Duvenaud, D. & Frey, B. J. Generating and designing dna with deep generative models. *arXiv* (2017). [1712.06148](#).