

Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design

Supplemental Material

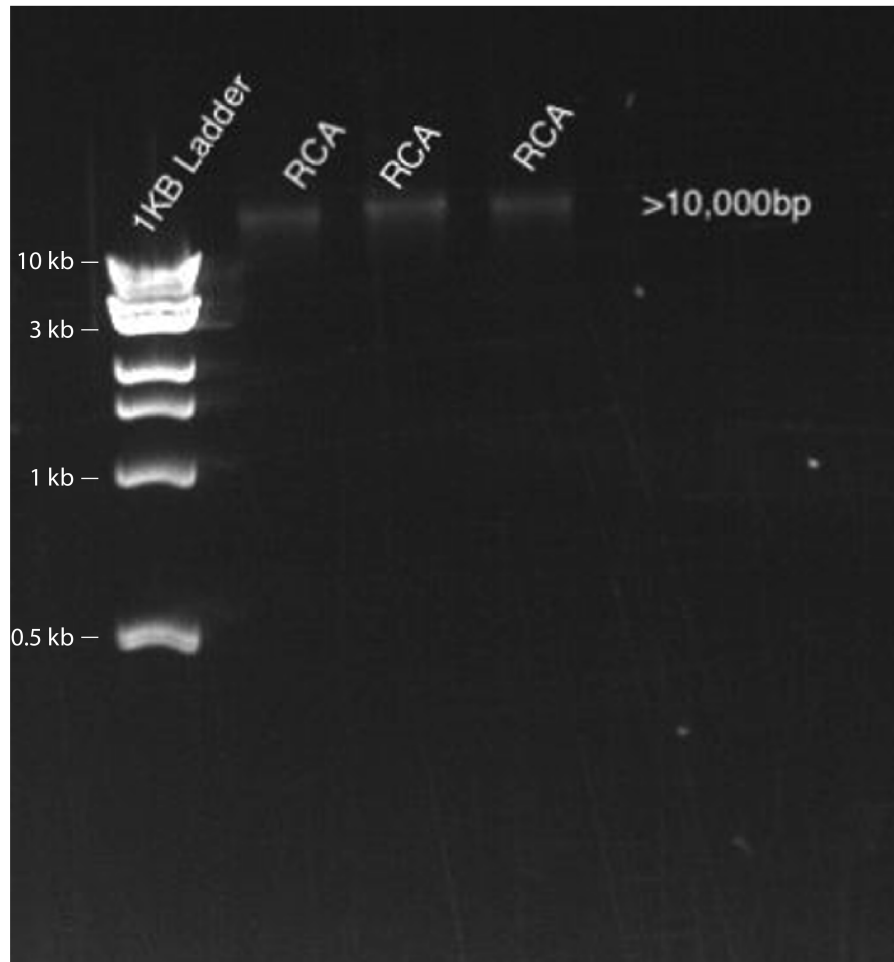
Imburgia, Organick, Zhang, Cardozo *et al.*

Contents

1	Size of RCA Products	3
2	Calculating Enrichment Score (ES)	4
3	Random Access Cas9 gRNA and PAM Sequences	5
4	Number of Payloads per Read	6
5	Examining Sequences After Similarity Search Encoding	7
6	Sequencing and Filtering Summary	8
7	The Impact of Edit Distance on Activation	10
8	Effect of High Specificity on Encoding	11
9	Similarity Search Primer Sequences	12

1 Size of RCA Products

A 1% Agarose Gel Electrophoresis Results for RCA Products shows that each amplified library is roughly the same size and has a length greater than 10,000 bp.



Supplementary Figure 1: 1% Agarose Gel Electrophoresis Results for RCA Products

2 Calculating Enrichment Score (ES)

The calculation for the Enrichment Score (ES) is the ratio of the concentration of the target file in the enriched library to the concentration of target file in the original (or baseline), un-enriched library.

$$ES = \frac{\frac{R_f}{R}}{\frac{r_f}{r}} \quad (1)$$

Where:

- ES is the Enrichment Score
- R_f is the number of file-specific reads in the enriched library experiment.
- R is the total number of reads in the enriched library experiment.
- r_f is the number of file-specific reads in the un-enriched library experiment.
- r is the total number of reads in the un-enriched library experiment.

3 Random Access Cas9 gRNA and PAM Sequences

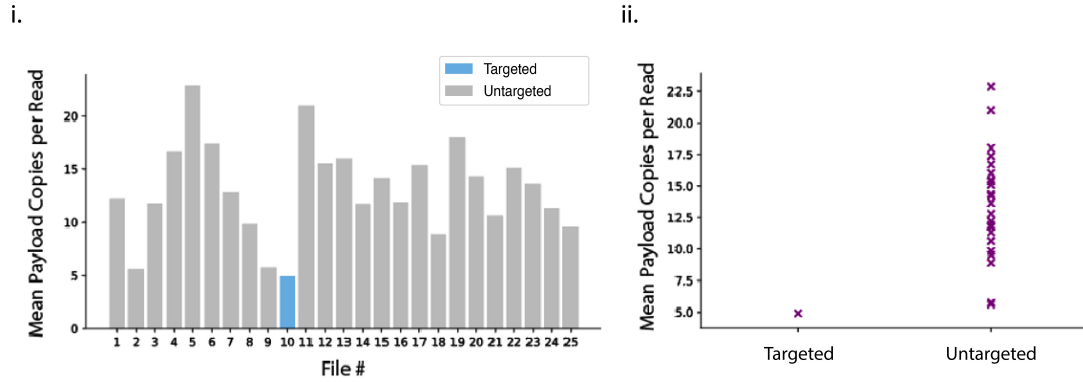
The Cas9 gRNA and PAM sequences for all 25 files in our random access library.

Guide Name	Guide Sequence	PAM
Guide 1	CGCAGAGGTGGCGTCGACCA	AGG
Guide 2	CGCAGAGGTGGCGCGATTGT	TGG
Guide 3	CGCAGAGGTGGCGATAGCAT	AGG
Guide 4	CGCAGAGGTGGCGTCGCCTT	TGG
Guide 5	CGCAGAGGTGGCGTGCCAAC	AGG
Guide 6	CGCAGAGGTGGCGATCGCTC	TGG
Guide 7	CGCAGAGGTGGCGAGAACCG	AGG
Guide 8	CGCAGAGGTGGCGGACGAAC	TGG
Guide 9	CGCAGAGGTGGCGCACAGAG	AGG
Guide 10	CGCAGAGGTGGCGCATTAAA	TGG
Guide 11	CGCAGAGGTGGCGAGAATCT	AGG
Guide 12	CGCAGAGGTGGCGCTAGCTG	TGG
Guide 13	CGCAGAGGTGGCGCTTTACA	AGG
Guide 14	CGCAGAGGTGGCGAAGTGCT	TGG
Guide 15	CGCAGAGGTGGCGGGTTCCT	AGG
Guide 16	CGCAGAGGTGGCGTTTCCAG	TGG
Guide 17	CGCAGAGGTGGCGTGGGATG	AGG
Guide 18	CGCAGAGGTGGCGTAAACGC	TGG
Guide 19	CGCAGAGGTGGCGGATGAGA	AGG
Guide 20	CGCAGAGGTGGCGTCCTCAG	TGG
Guide 21	CGCAGAGGTGGCGGCCACAT	AGG
Guide 22	CGCAGAGGTGGCGCCCATCC	TGG
Guide 23	CGCAGAGGTGGCGCCTGGTT	AGG
Guide 24	CGCAGAGGTGGCGTCCATCG	TGG
Guide 25	CGCAGAGGTGGCGGACCTTA	AGG

Supplementary Table 1: All Cas9 gRNA sequences in the random access library.

4 Number of Payloads per Read

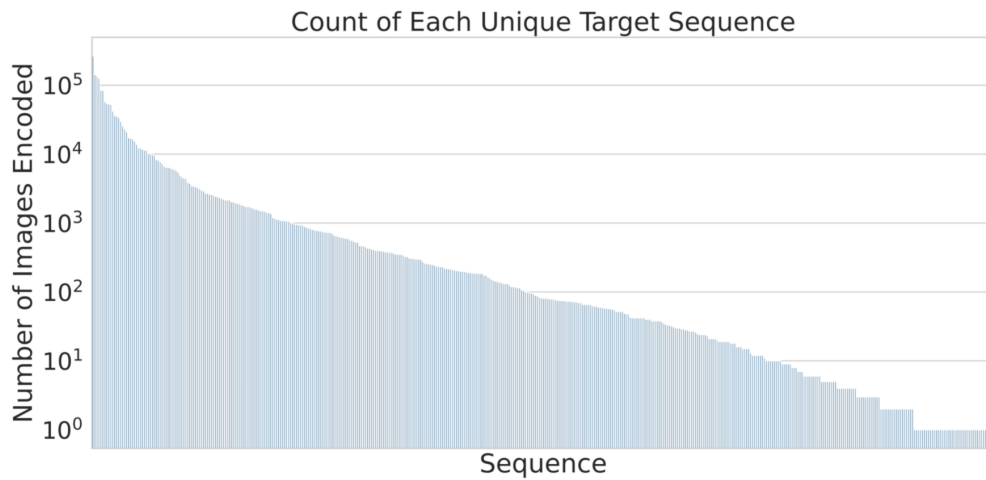
In the random access protocol there are many copies of the same payload after RCA. The mean number of payload copies (concatemers) per read was calculated by dividing the average total read lengths for each file by the circularized file length (580 bp).



Supplementary Figure 2: **(i, ii)** The mean number of payload copies of file 10 per read in the first enrichment experiment (1 μ M gRNAs and 1 minute reaction incubation) was about 5. Though we expect file 10 to have fewer payload copies per read due to Cas9 cleavage necessary for adapter ligation, longer reads with more payload copies are more conducive for nanopore sequencing and downstream alignment.

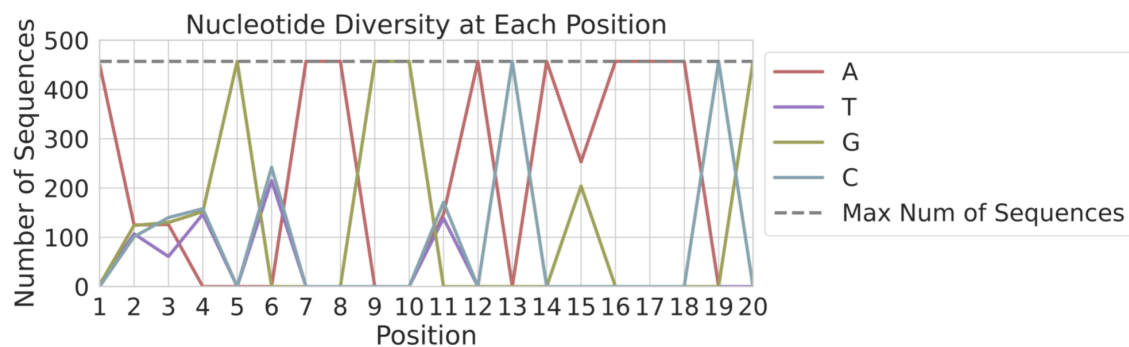
5 Examining Sequences After Similarity Search Encoding

After encoding with the best performing similarity search model, a 1.74 million image database is represented by 457 unique sequences. In **Supplementary Figure 3**, we show how many images are represented by each unique DNA sequence.



Supplementary Figure 3: **Examining the Encoded Sequence Relationship to Number of Represented Images.** Each column in the x-axis is a unique sequence encoded by our best performing encoder. The y-axis shows the number of images that one sequence represents.

To understand the composition of each sequence, we examine all unique sequences and report the rate each base is present at each position as shown in **Supplementary Figure 4**. We see the sequences are highly conserved, with only positions 2, 3, 4, 6 and 11 having high diversity.

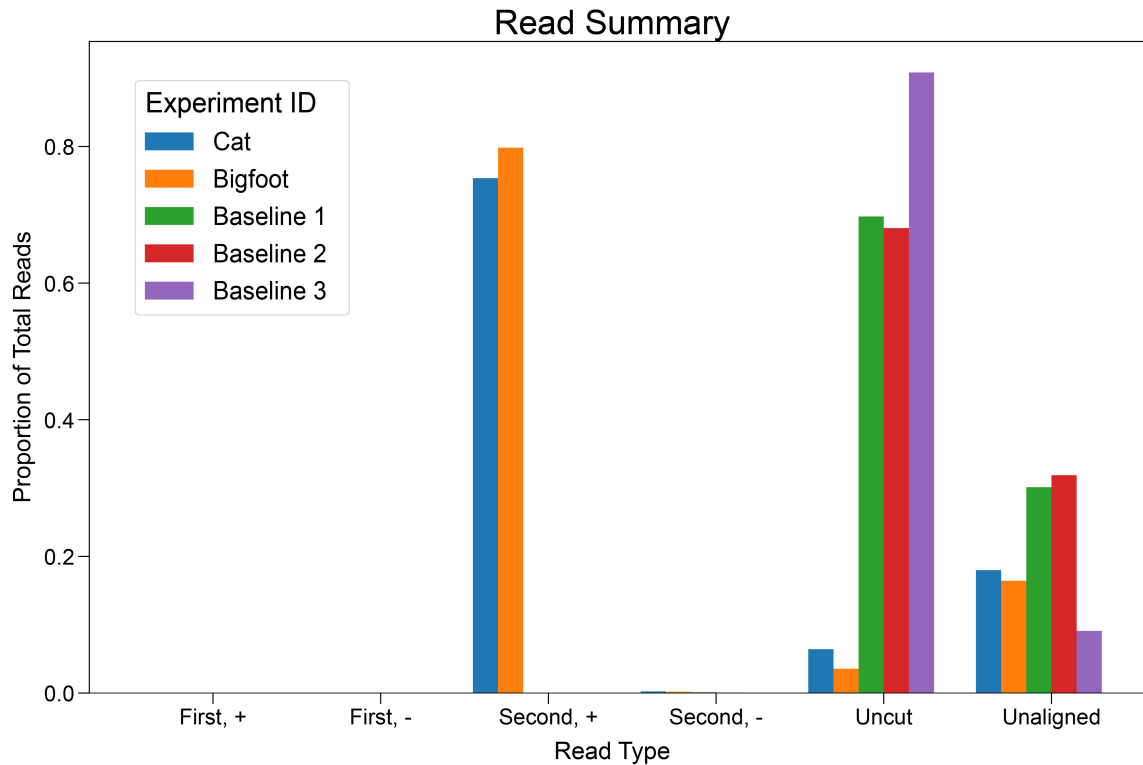


Supplementary Figure 4: **Examining the Encoded Sequence Diversity.** The PAM distal region is to the left and the PAM proximal region is to the right.

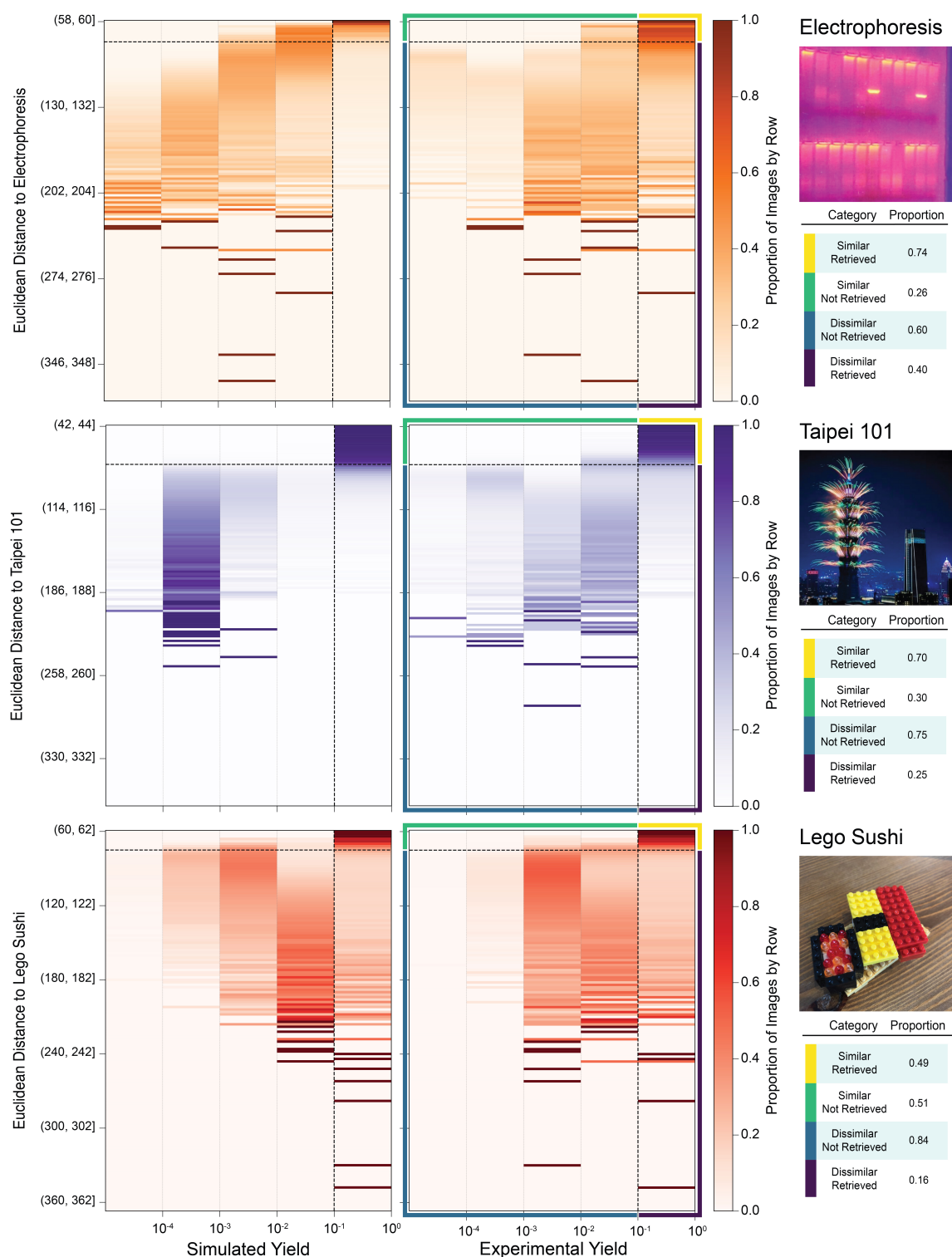
6 Sequencing and Filtering Summary

Experiment ID	First, +	First, -	Second, +	Second, -	Uncut	Unaligned	Total Reads
Cat	9	196	875,839	2,644	74,532	209,001	1,162,221
Bigfoot	11	772	2,365,731	5,320	105,317	486,675	2,963,826
Baseline 1	17	1	417	1,696	1,071,839	462,935	1,536,905
Baseline 2	0	0	75	136	212,323	99,466	312,000
Baseline 3	1	0	23	28	56,312	5,636	62,000

Supplementary Table 2: Raw read counts categorized by alignment start and end positions. First and second refer to the two halves of the sequences after Cas9 cleavage where 'first' contains the universal forward primer and partial Cas9 file address (~43 nt) and 'second' contains the PAM, payload, and universal back primer (~120 nt). The direction the read sequence was to the reference is distinguished as '+' for the exact match and '-' for the reverse complement.



Supplementary Figure 5: **Read Summary by Sequencing Run** The proportion of total read sequences that fall into each category of read alignment to reference (+ for exact, - for reverse complement), cleaved strand half (first/second), and alignment success.

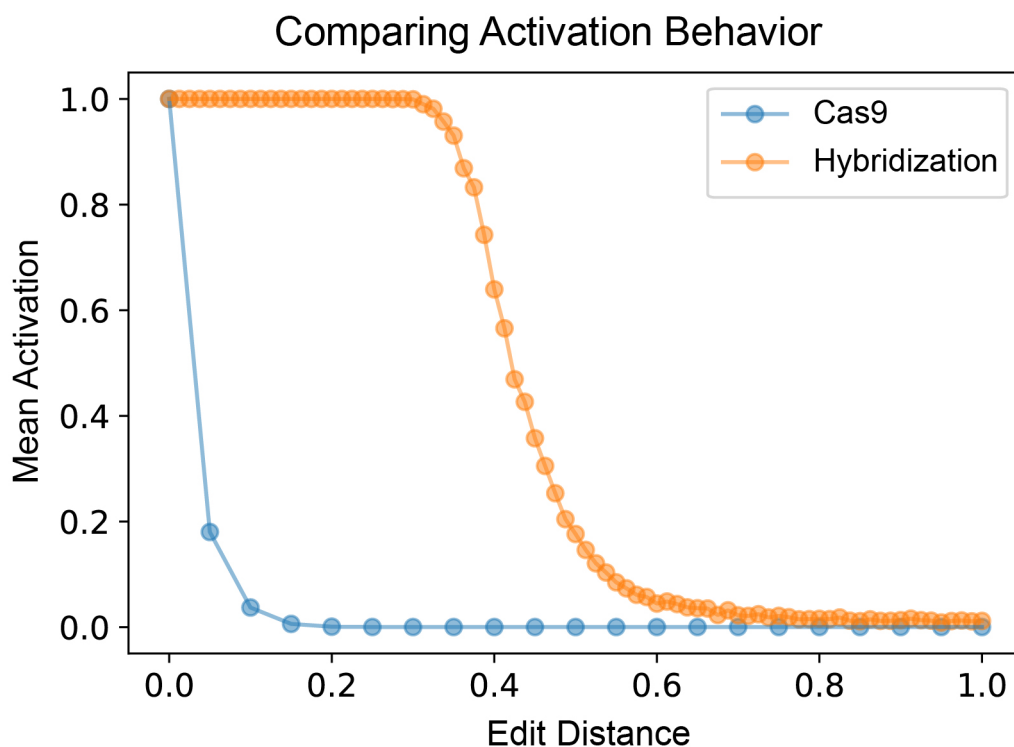


Supplementary Figure 6: **Results from Additional Queries.** Heat maps of simulated and experimental retrieval results and a summary of the proportion of similar and dissimilar images are displayed for each query. For the heatmaps, images are binned by Euclidean distance and a normalized simulated yield or enrichment score. Color indicates the proportion of total images in a row that fall into each yield bin. Images are partitioned into four categories (similar or dissimilar and retrieved or not retrieved).

7 The Impact of Edit Distance on Activation

We examine the activation behavior of Cas9 as a function of edit distance and compare it to its predecessor method, hybridization in **Supplementary Figure 7**. 100,000 randomly generated sequence pairs were run through the appropriate predictor model (hybridization or Cas9 cleavage) to obtain activation scores. The activation scores and edit distances between each pair were recorded.

A model's encoding space refers to the all the possible changes that sequences can undergo to achieve different activation scores. A larger encoding space enables more potential activation scores, which helps preserve similarity relationships (discussed in detail in the main text). The hybridization method spans a greater edit distance range than Cas9 before activation drops below a useful range. Therefore its encoding space is much larger. This supports the hypothesis presented in the main text that Cas9 sequence recognition is too stringent. Any increase in edit distance drastically reduces Cas9 activation behavior. Consequently, encoding space is reduced and similarity search using the Cas9 architecture proves difficult for complex databases.

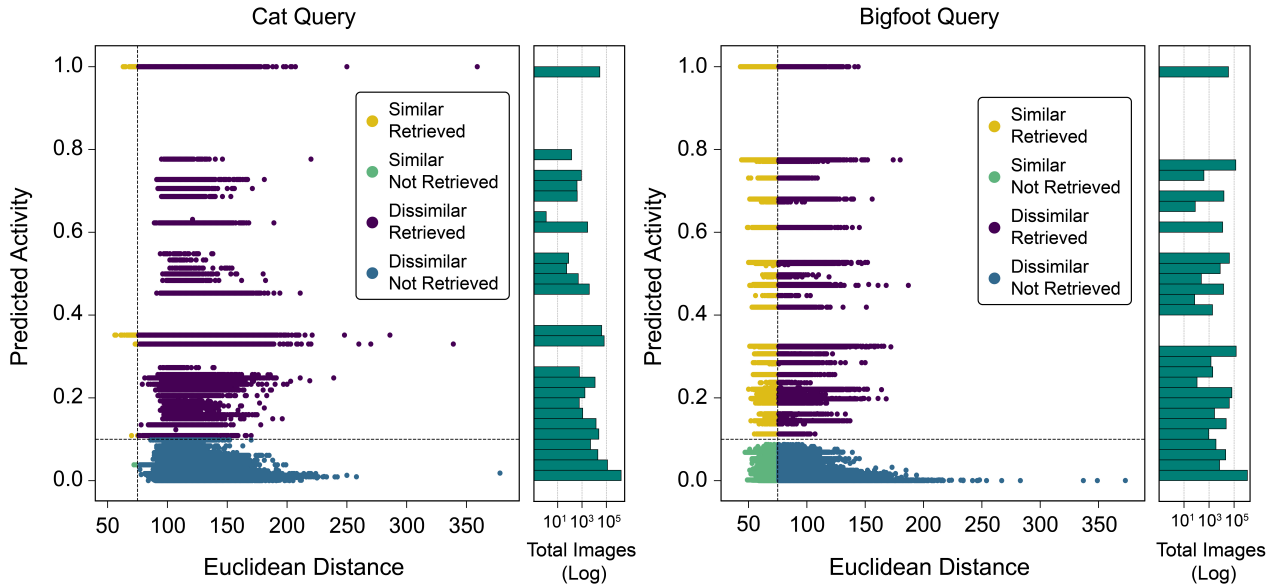


Supplementary Figure 7: **Examining Activation Behavior.** An edit distance of 0 indicates there are no differences between two sequences, an edit distance of 1 indicates all positions in a sequence are different between two sequences. An activation of 0 indicates no activity (Cas9 cleavage of a sequence, or the hybridization of two sequences, respectively) while a score of 1 indicates maximum activity.

8 Effect of High Specificity on Encoding

Observing the distribution of Euclidean distances of each image encoded to a specific sequence helps understand the consequence of reduced encoding space in our model. **Supplementary Figure 8** shows that retrieval of similar images is impossible without the retrieval of dissimilar images that are represented by the same encoded sequence. The retrieval accuracy of the two queries is impacted in part by the distribution of similar images within the dataset. The Bigfoot query has many ambiguous features and therefore has similarity to a wide range of images across the dataset. In contrast, the cat query has features that distinguish it from the image database. As a result, there are fewer images that are classified as "similar" to it, and retrieval of most of these images doesn't increase the retrieval of false negatives beyond what we see in other queries.

Predicted Activity and Distribution of Euclidean Distances to Query by Sequence Encoding



Supplementary Figure 8: **Euclidean Distance Distribution by Sequence.** Each point represents an image's predicted activity (based on sequence encoding) and Euclidean distance to the specified query. All 1.74 million images are included, but many overlap on the graph. To resolve skew from this, we include a histogram of the total number of images contained in predicted activity ranges.

9 Similarity Search Primer Sequences

All images are encoded, then flanked with a universal primer pair to allow PCR amplification of the entire pool. Supplementary Table 3 contains the primer sequences.

Forward Primer Sequence	Reverse Primer Sequence
TACTCGCTGCGTGCAATTTA	TTGACACGTTCCGGACACTTT

Supplementary Table 3: Universal primer sequences for similarity search work.