Similarity Search with DNA

Using DNA to do things silicon computers have traditionally done

Lee Organick 5-24-21

Overview

What is similarity search?

Why would you want to use DNA?

How do you use DNA?

How could we make this better?

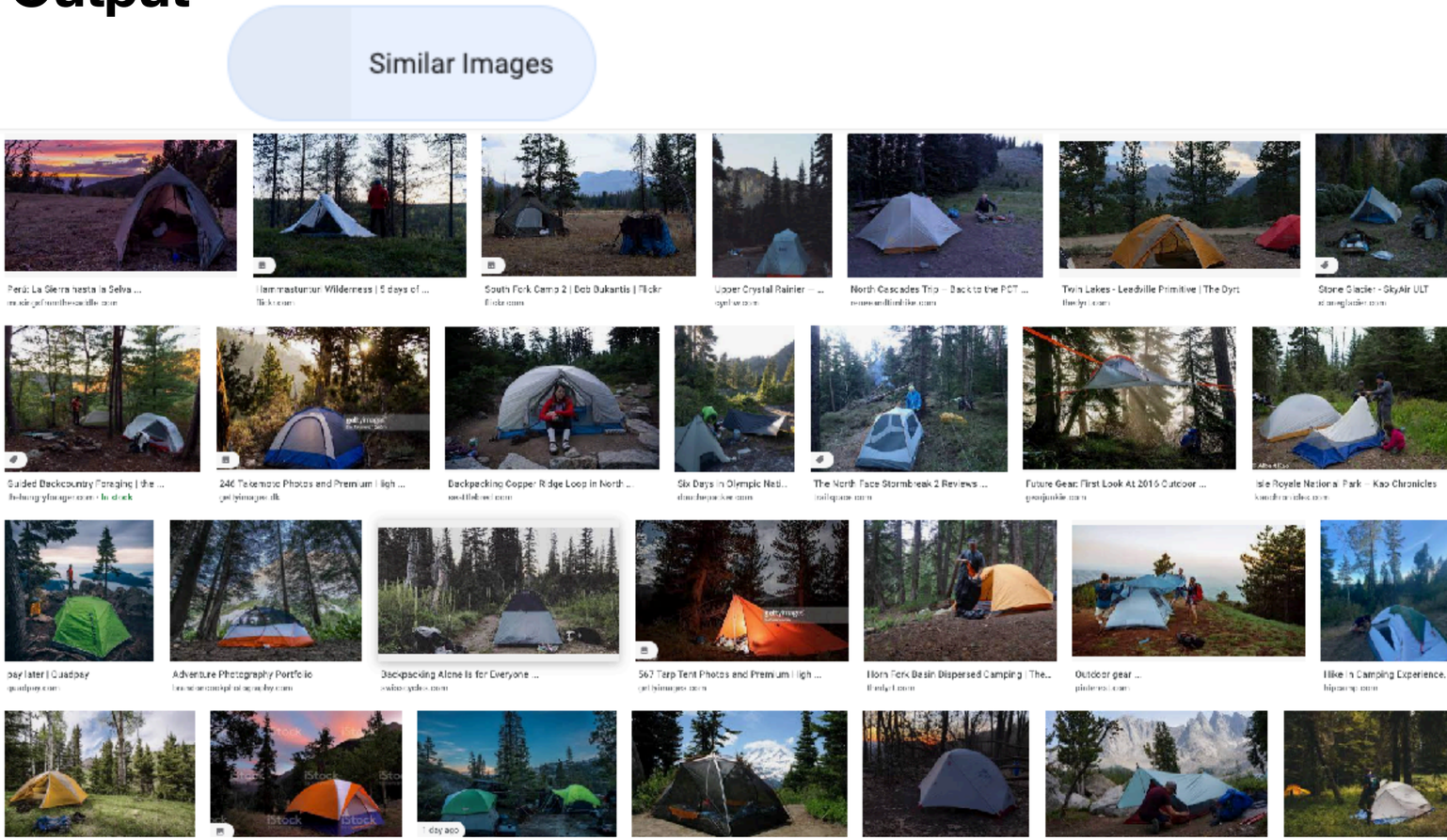
How do you use Cas9 to perform similarity search?

Musings on the intersection of Comp Bio, Molecular Computing, SynBio

What is similarity search?

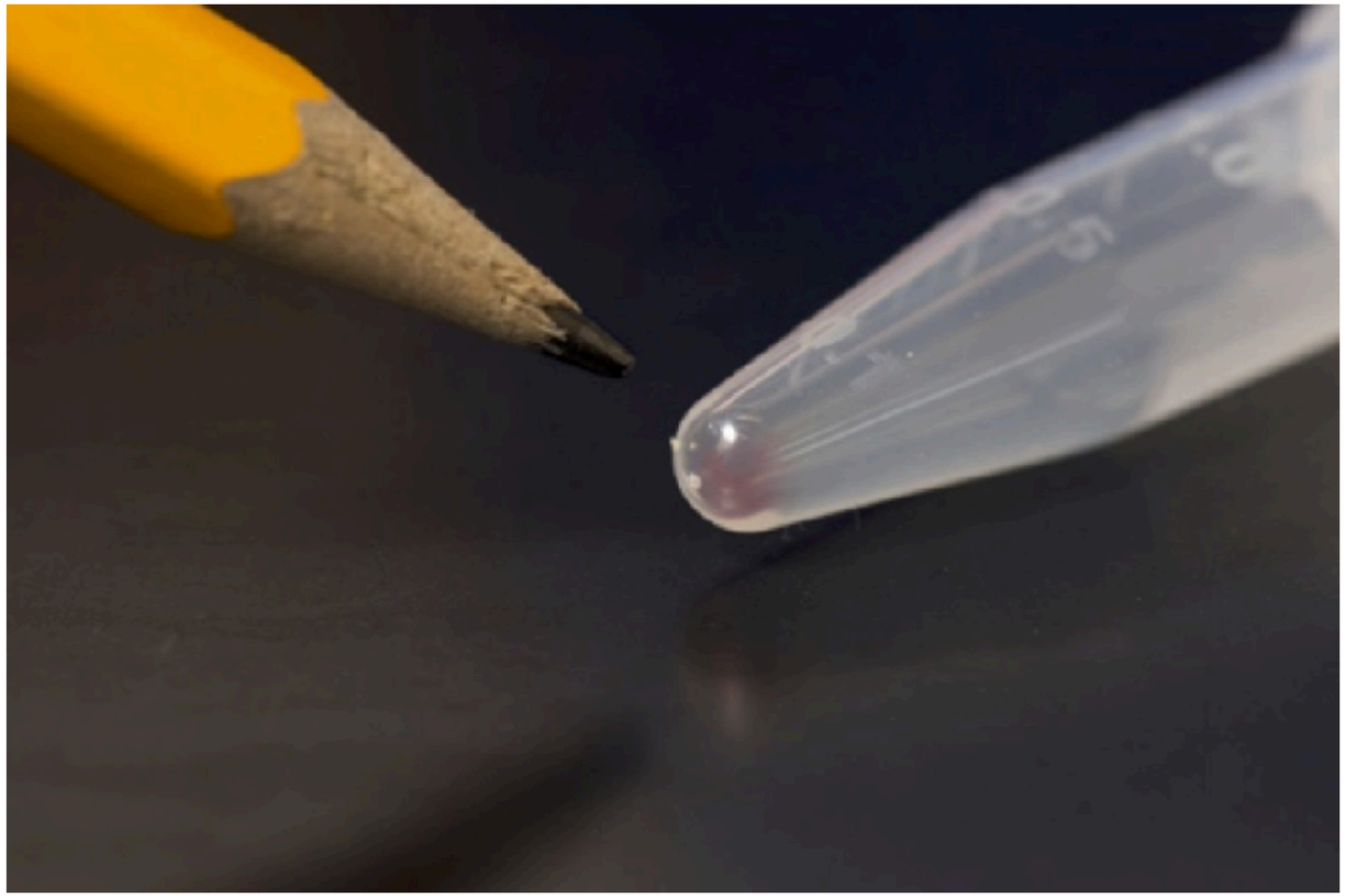
Input

Output



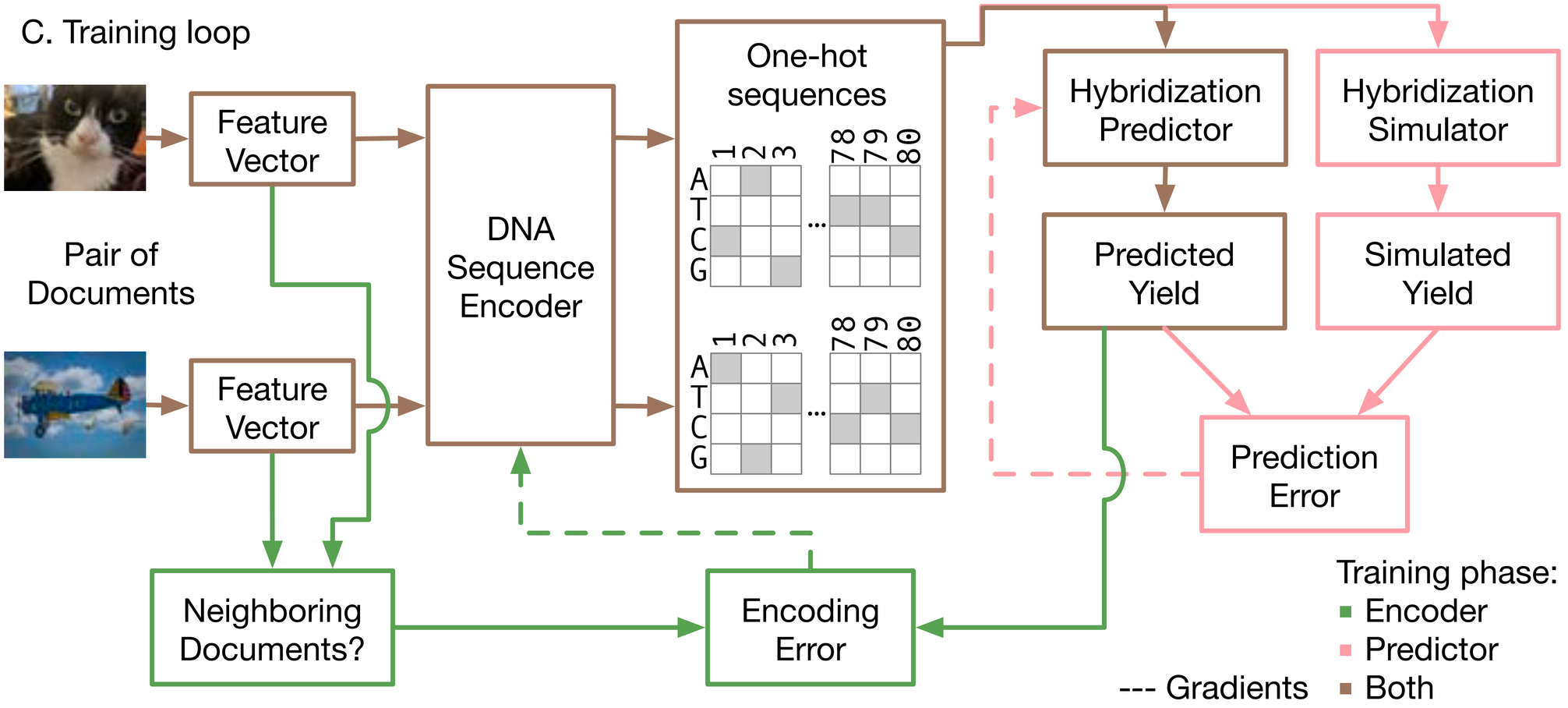
Why would you want to use DNA?

- Parallelism
- DNA information density
- DNA longevity
- Ease of distributing DNA databases
- Sometimes faster
- Sometimes more energy efficient



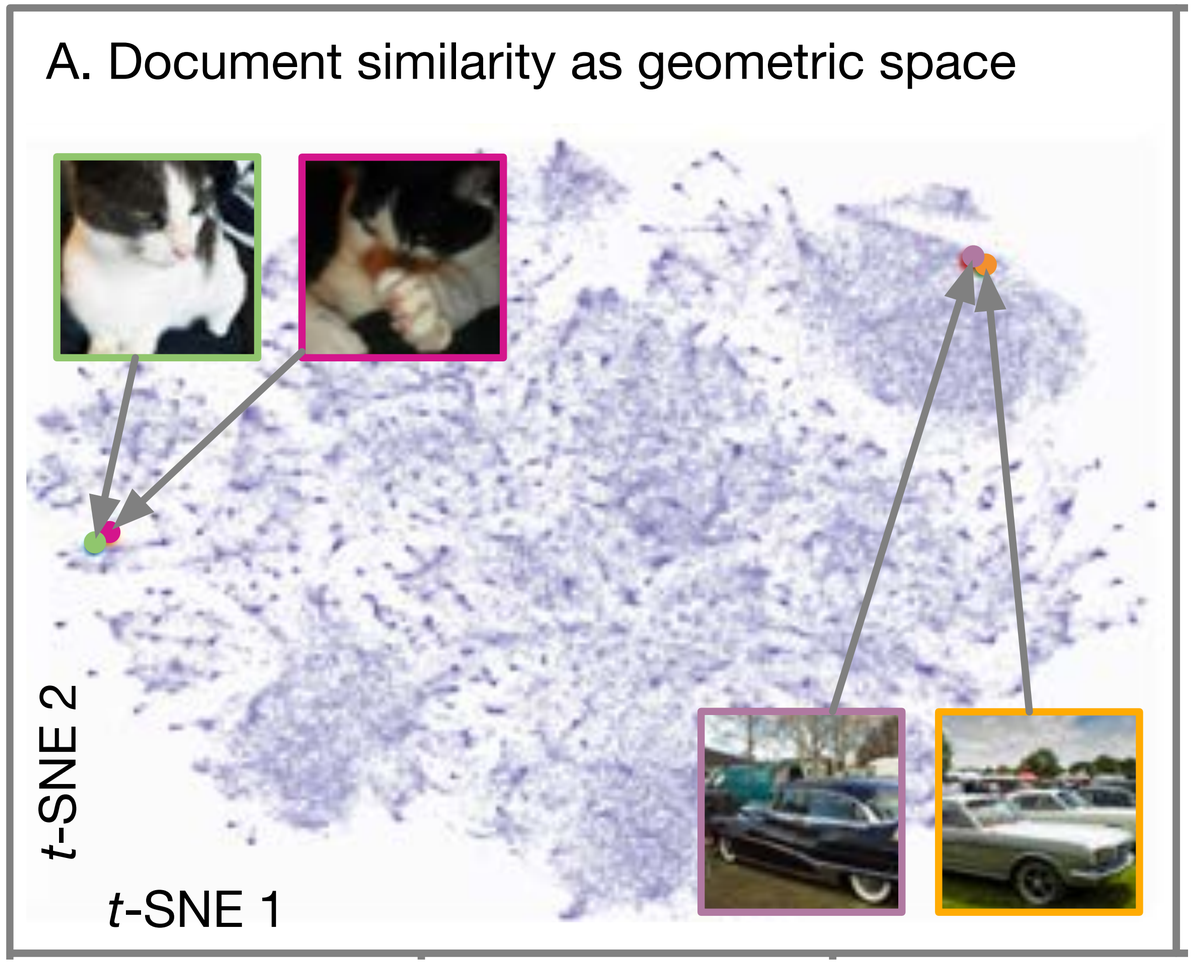
In that faint pink smear is ~10TB of data

Similarity Search



**Content-Based Similarity Search in Large-Scale DNA Data Storage Systems.* Bee et al. *BioRXiv 2020*

Similarity Search



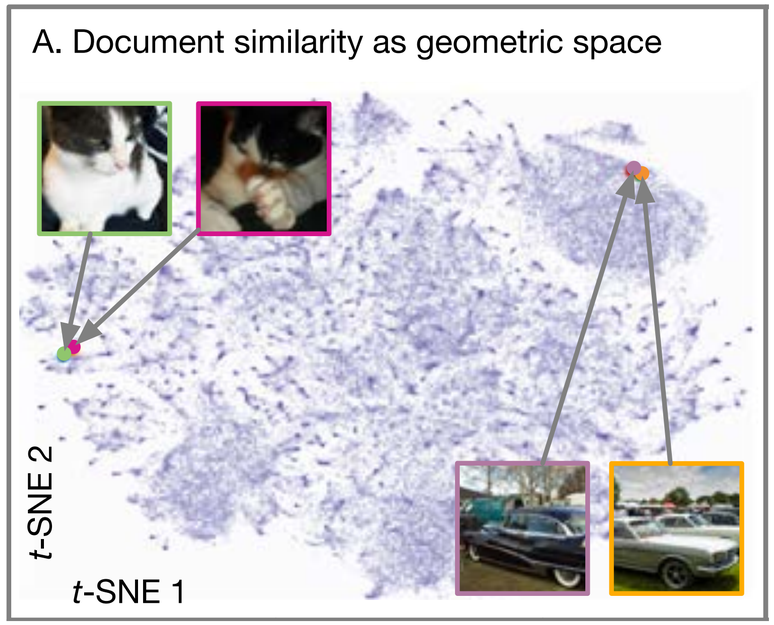
Curse of high dimensionality:

Exact indexing schemes in high-dimensionality spaces are no better than a costly linear or “brute force” search, which is infeasible for large databases

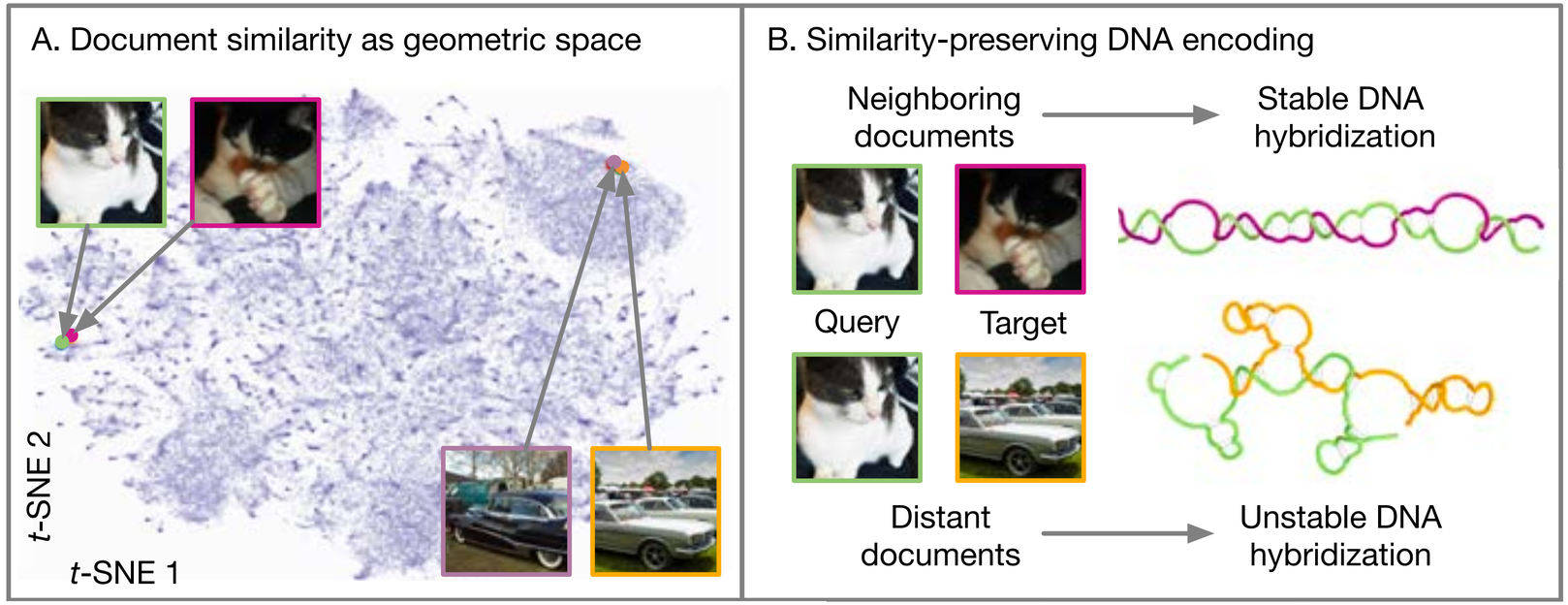
Tradeoff:

Rather than finding exact nearest neighbors, our goal is to maximize the number of near neighbors retrieved while minimizing the number of irrelevant results

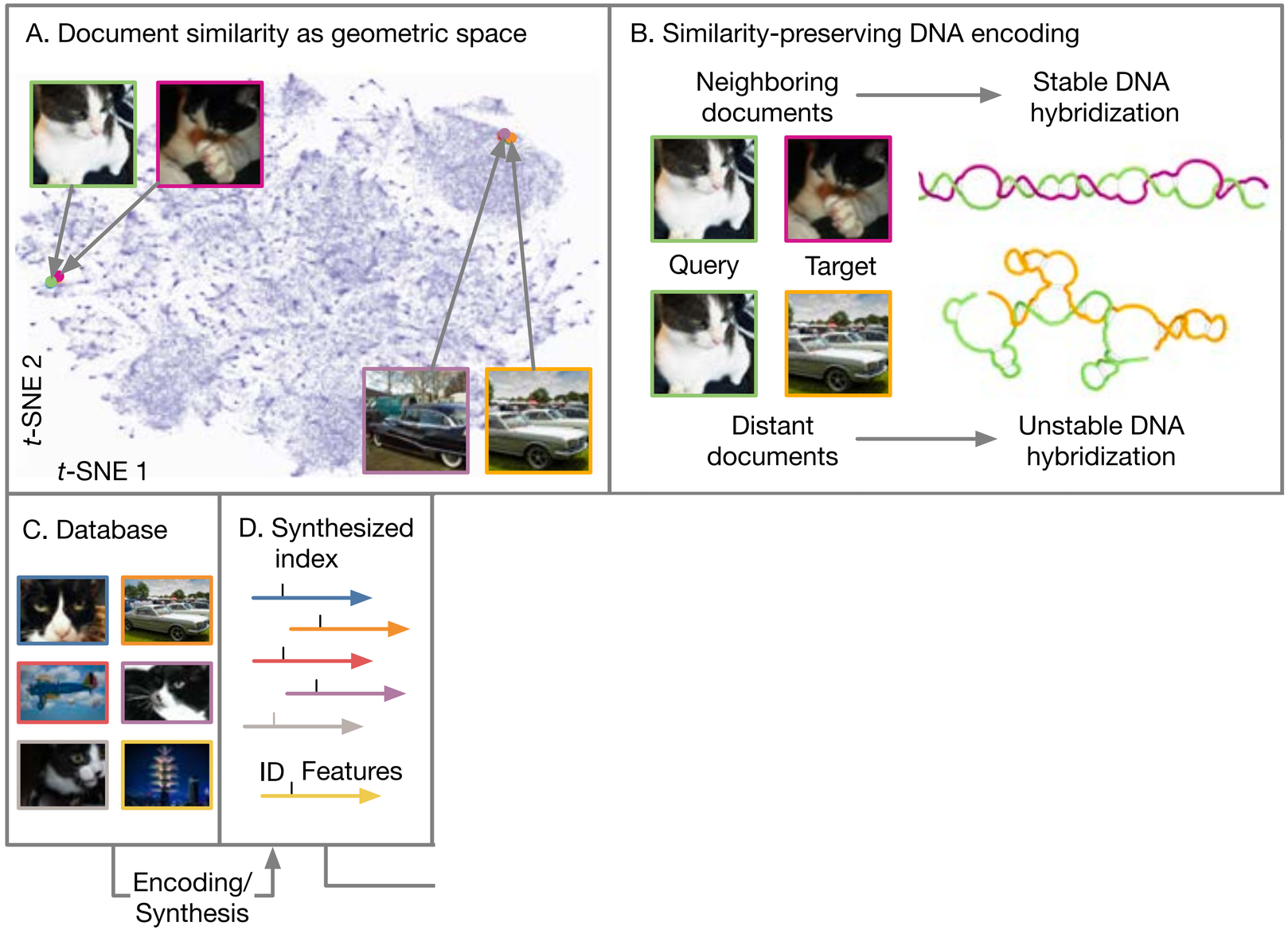
How do we do similarity search with DNA ?



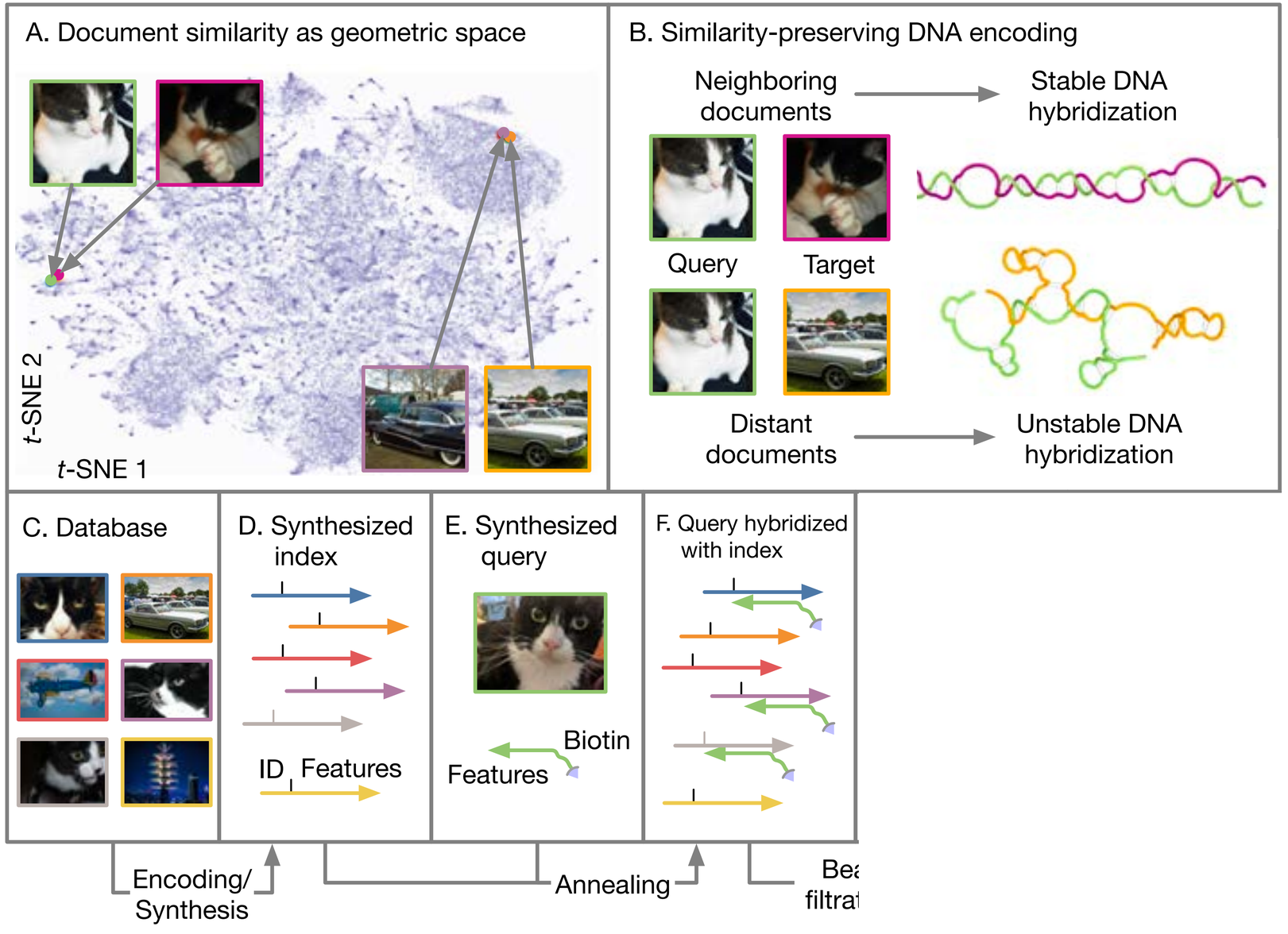
How do we do similarity search with DNA ?



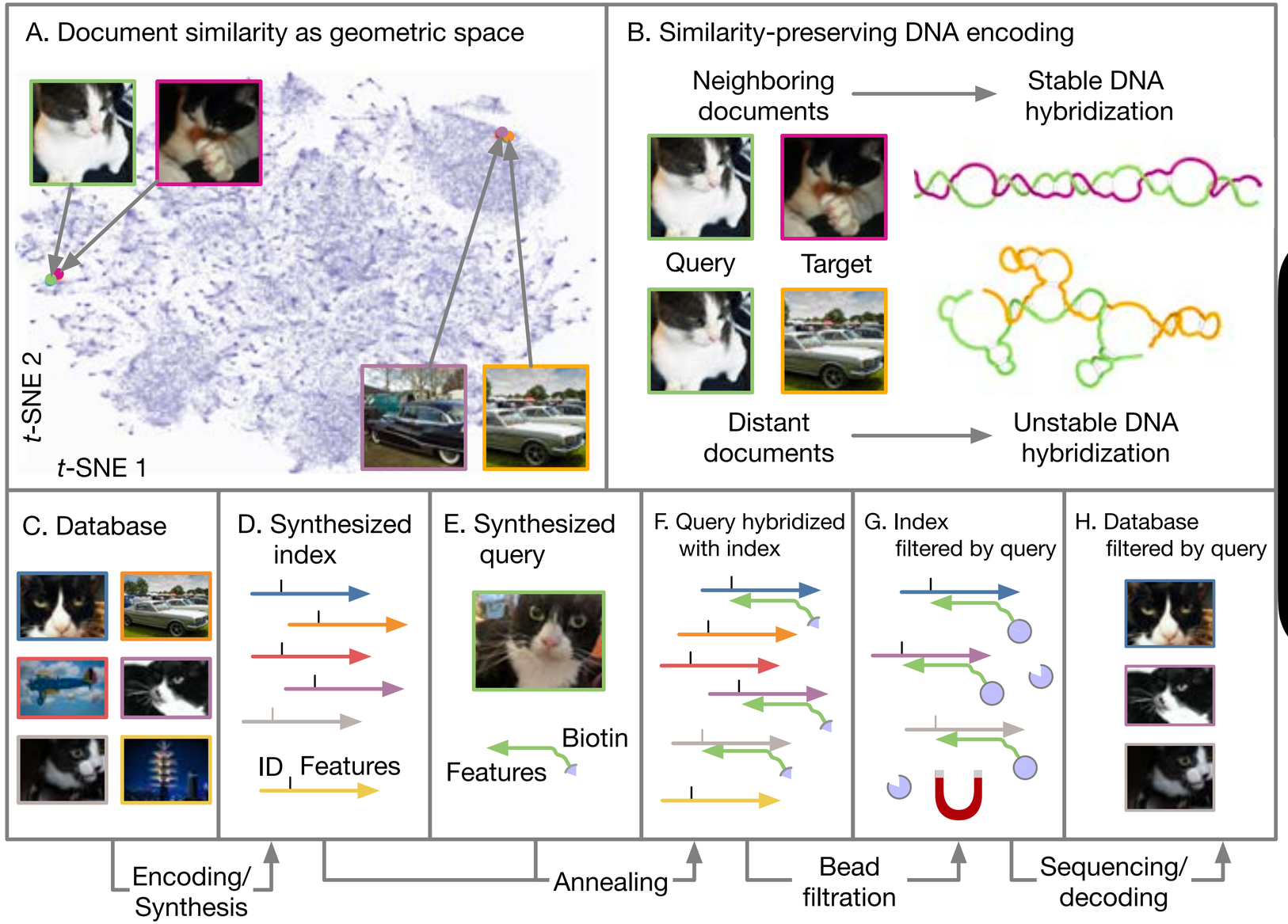
How do we do similarity search with DNA ?

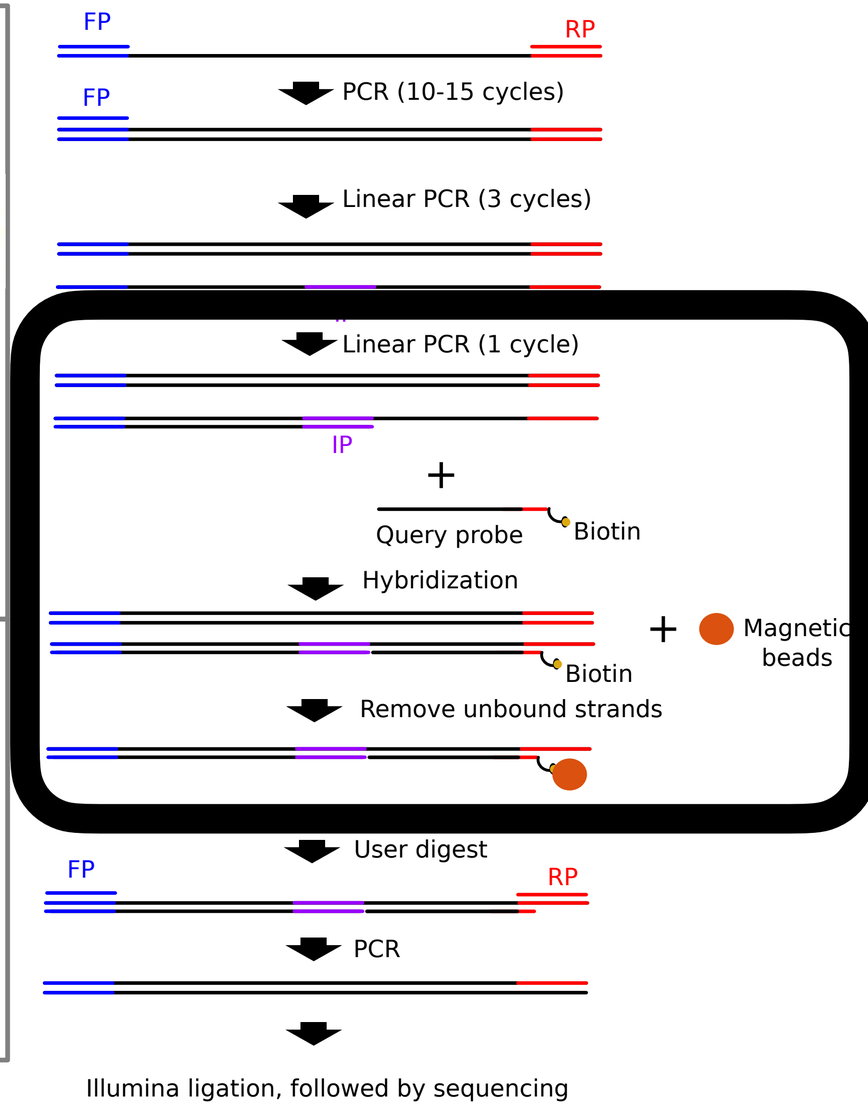


How do we do similarity search with DNA ?



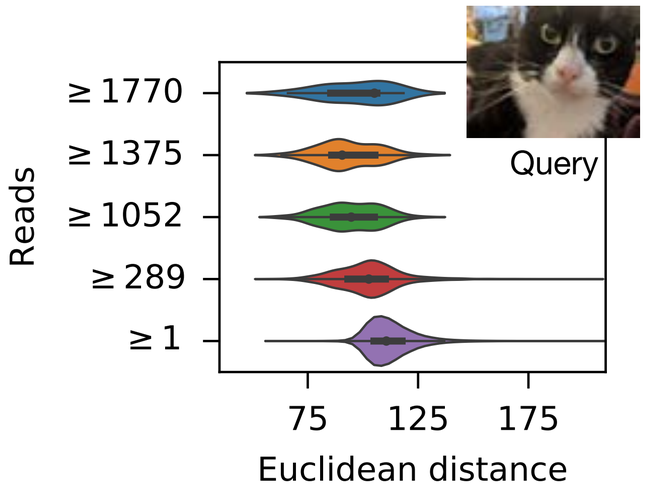
How do we do similarity search with DNA ?



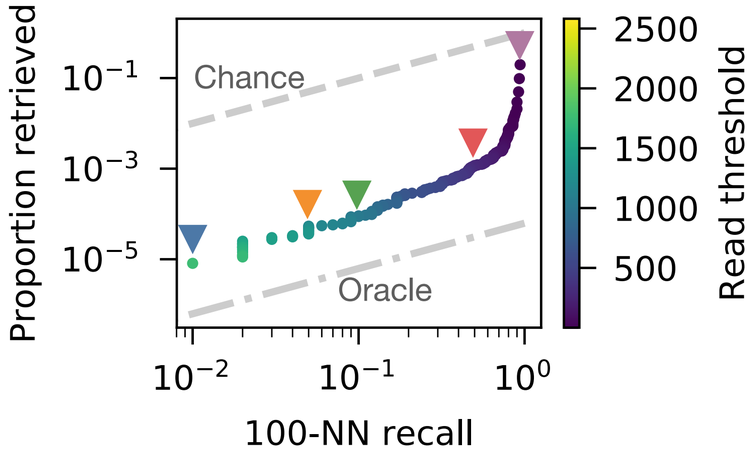


It works!

A. Distribution of similarity across read depths



B. Retrieval as a function of read depth



The proportion of the entire dataset that must be retrieved (y-axis) to retrieve a certain proportion of the 100 most similar images (x-axis)

C. Sets of retrieved images for select read depth thresholds

Number retrieved	100-NN recall	Top 5 results (sorted by similarity)
13	0.01	
58	0.05	
141	0.10	
1,831	0.50	
315,736	0.94	

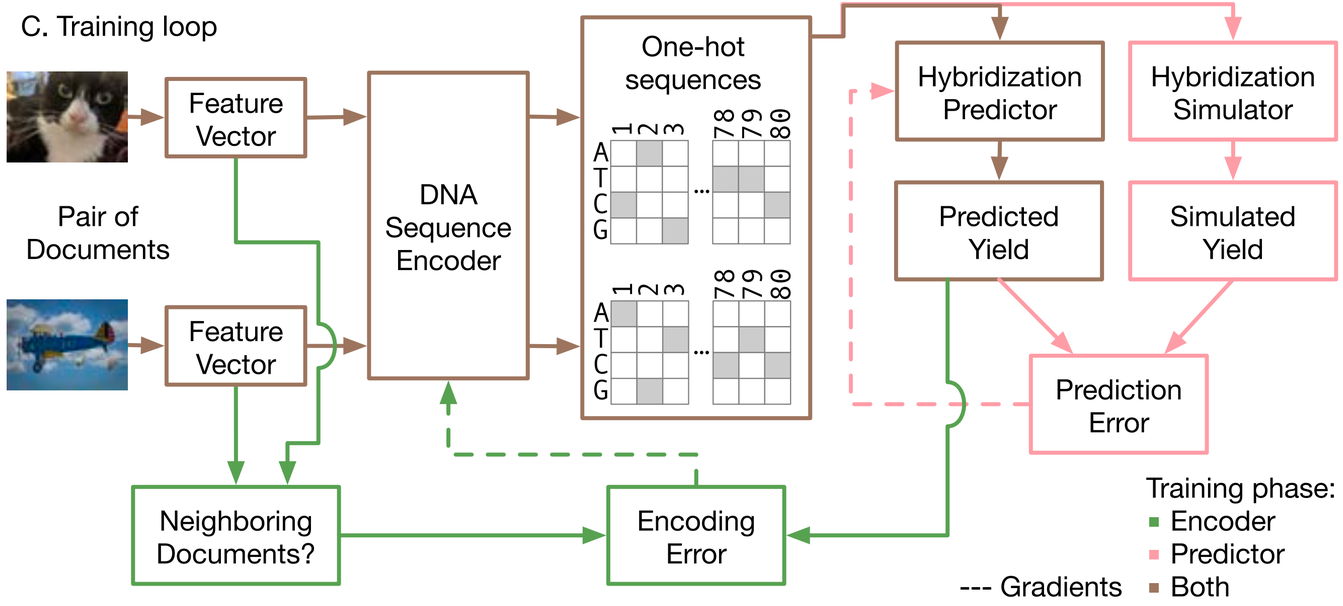
How can we improve?

- More energy efficient
- Faster

…what if we used Cas9?

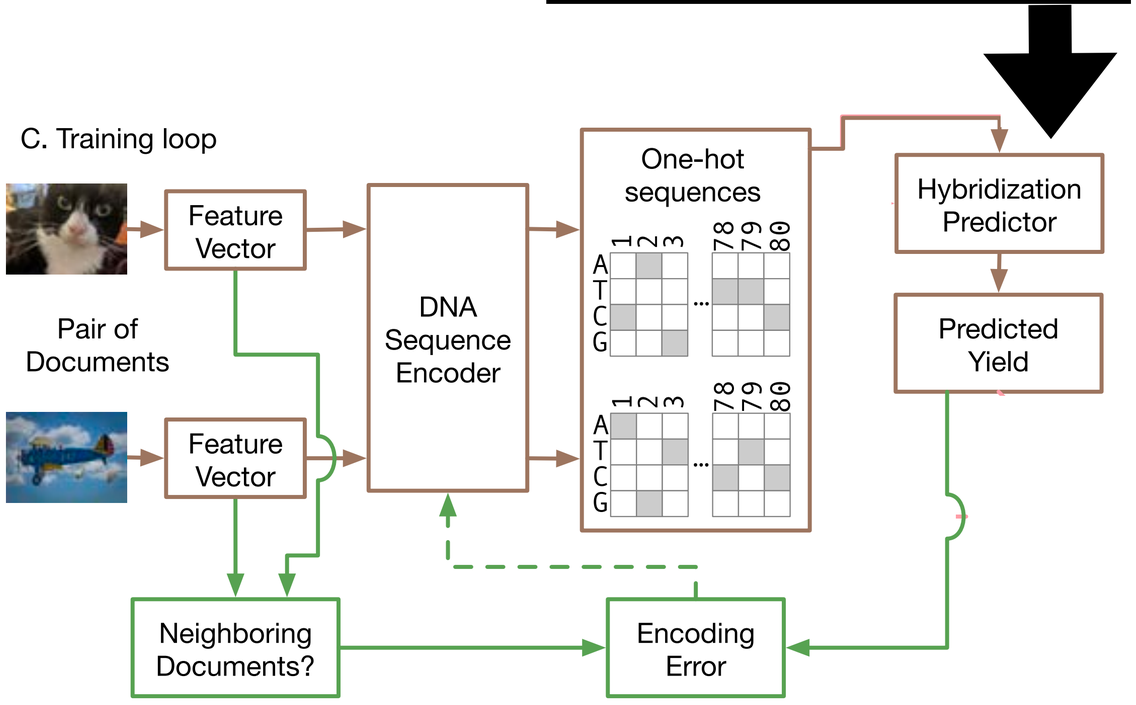
Using Cas9

Hybridization and Bead Extraction



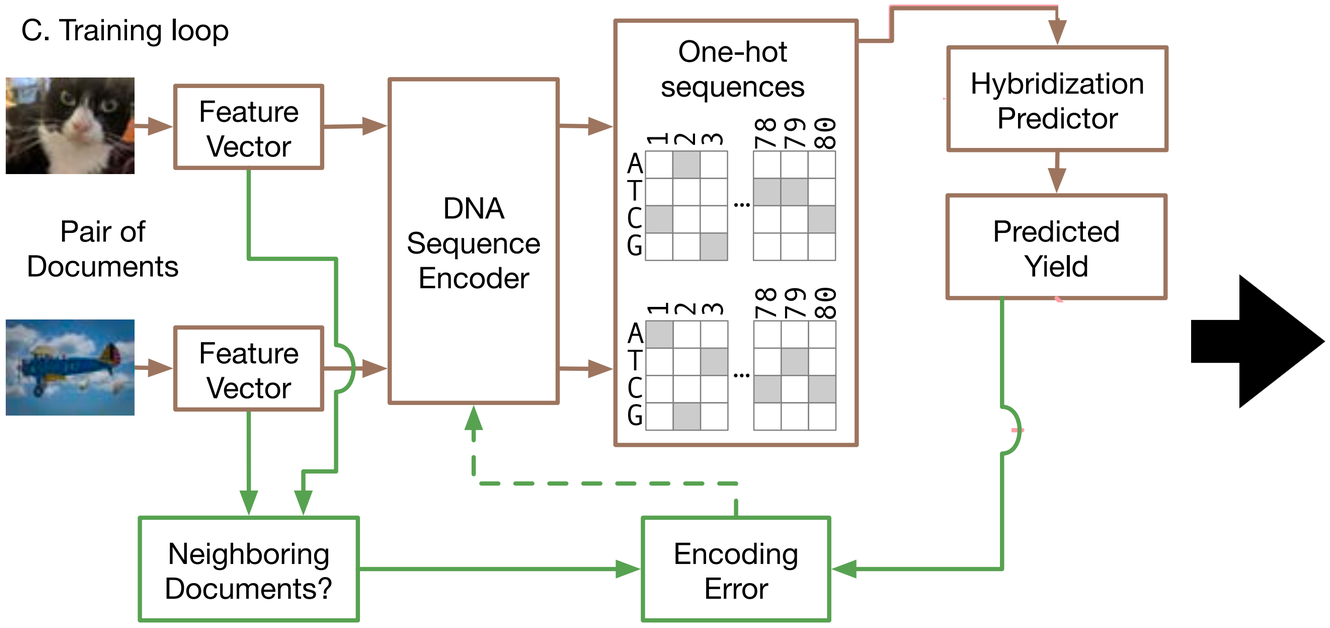
Cas9

*Cas9 Binding Predictor



Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. Jones et al. *Nature Biotechnology Sept. 2020*

Using Cas9

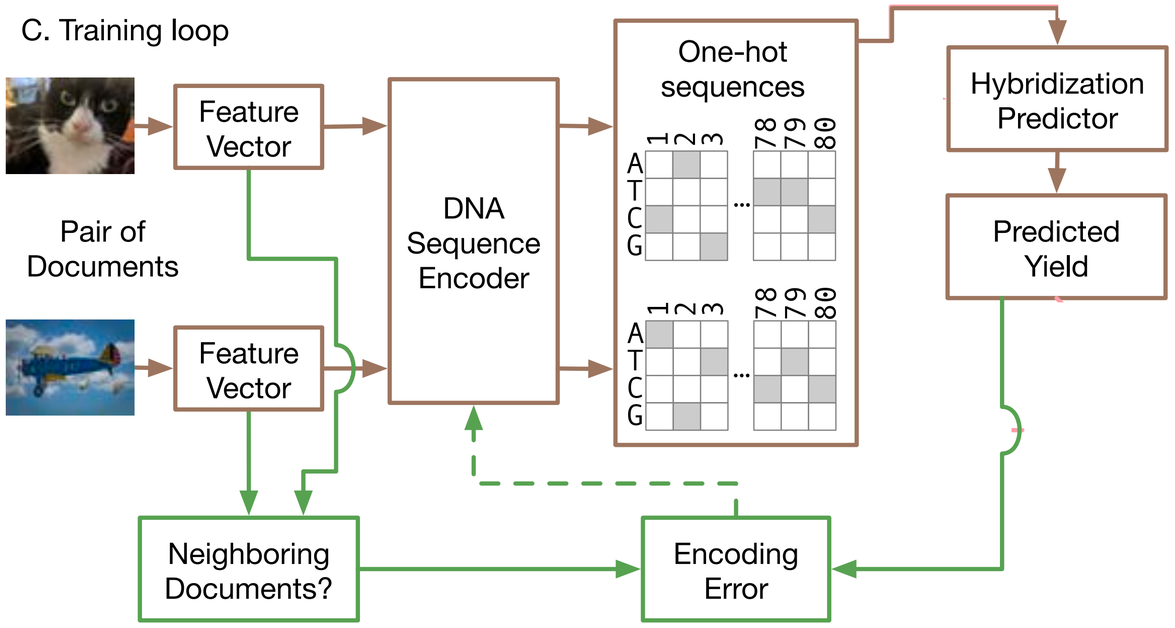


Cas9

20 features Data ID

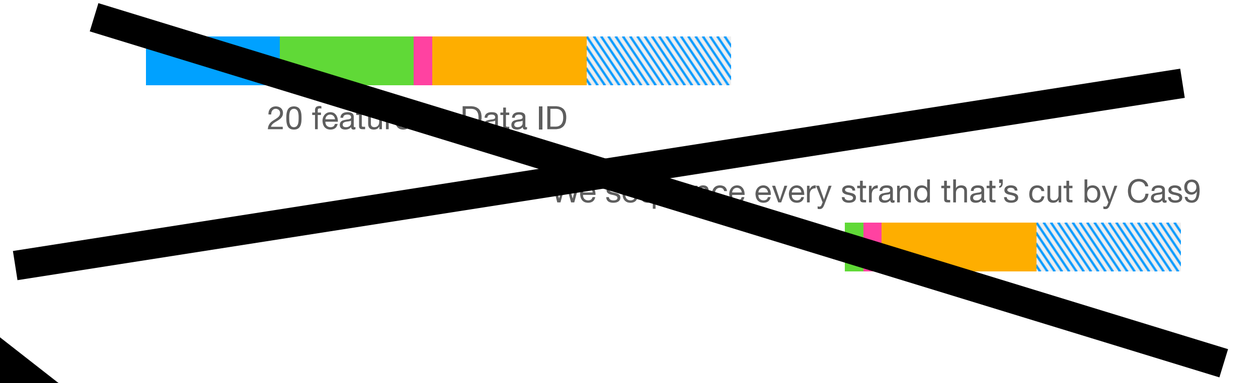
We sequence every strand that's cut by Cas9

Using Cas9



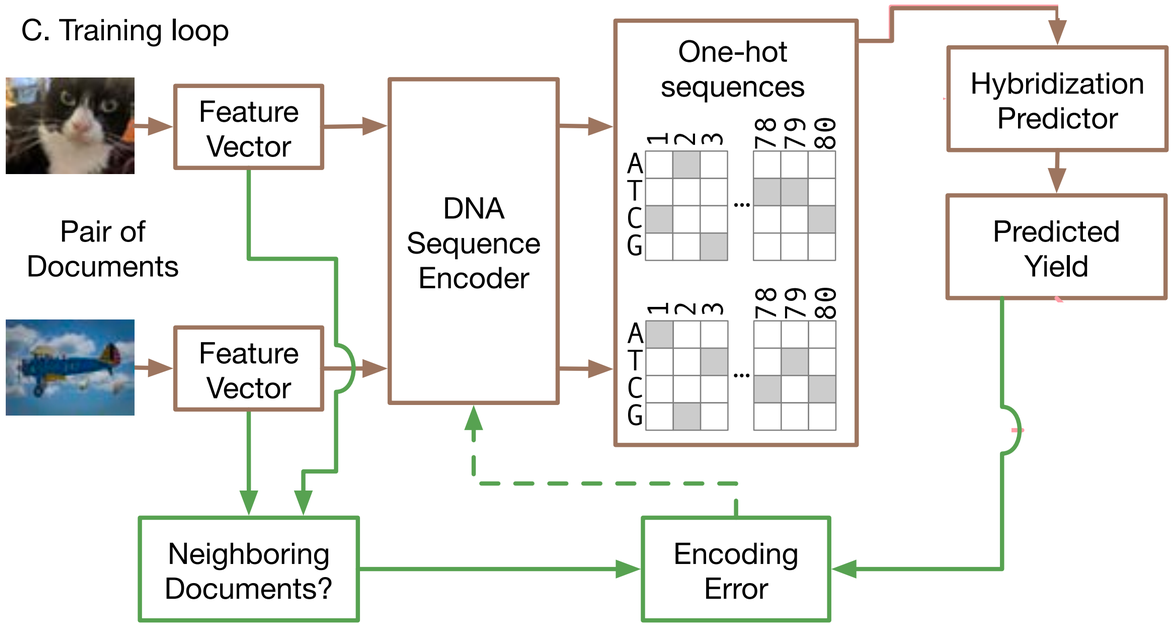
Cas9

Hard to do in wet lab



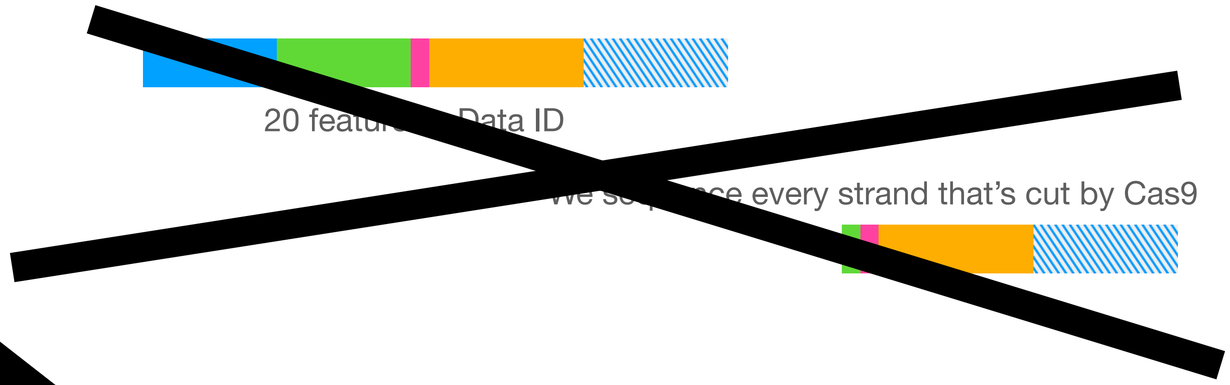
We sequence every strand that's cut at BOTH Cas9 regions…AND logic!

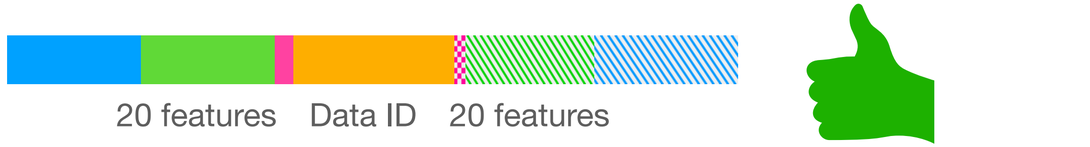
Using Cas9



Cas9

Hard to do in wet lab





We sequence every strand that's cut at BOTH Cas9 regions…AND logic!

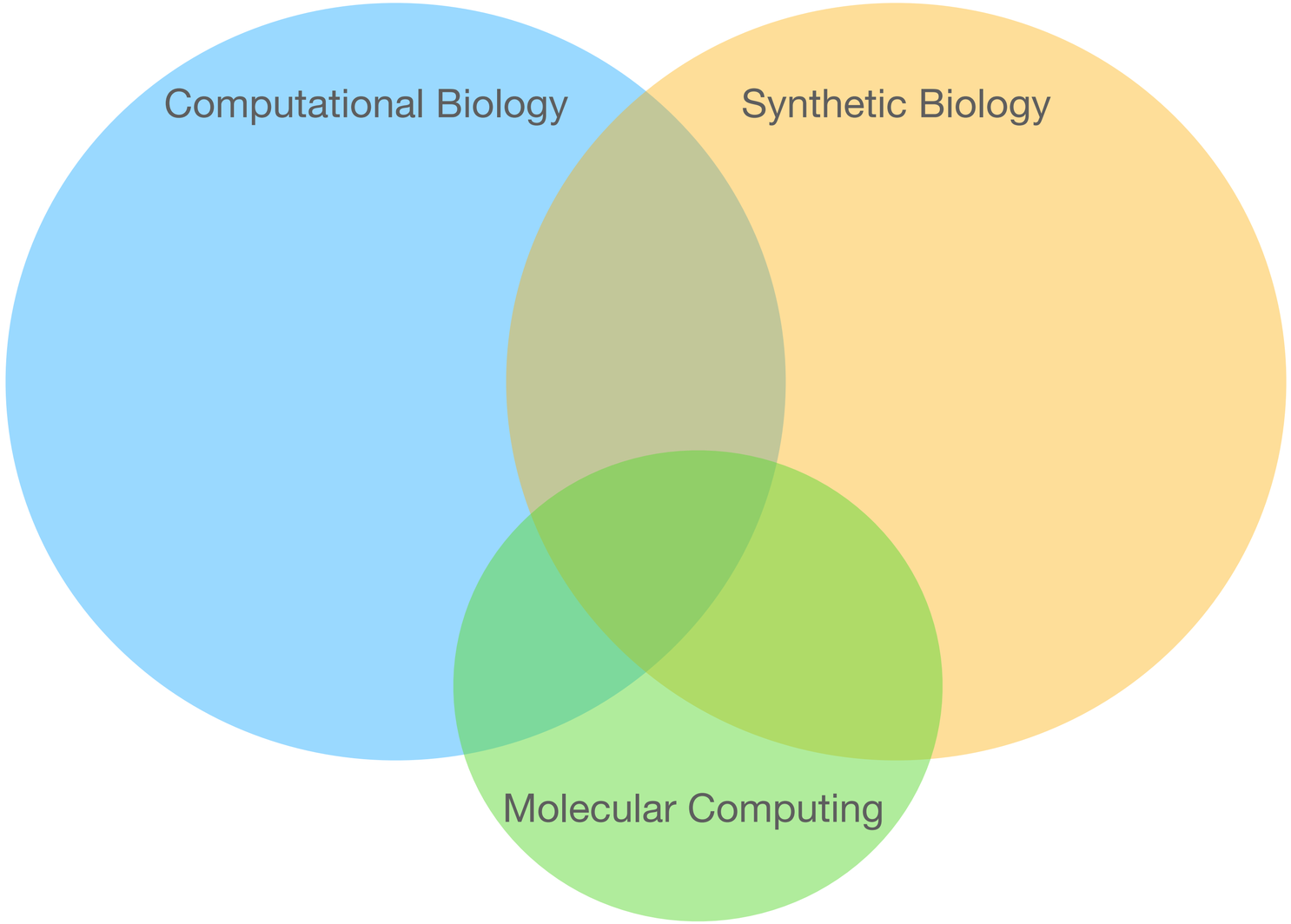
What if we allow for nested queries??

Comp Bio ∩ Molecular Computing ∩ SynBio

The lines are still being drawn

My two cents for people in any of these three fields:

- If you're a computation-centered person, get comfortable talking to wet lab-centered people
- vice-versa
- Bigger computational AND molecular toolboxes tend to make it easier to design experiments
- We need more tools

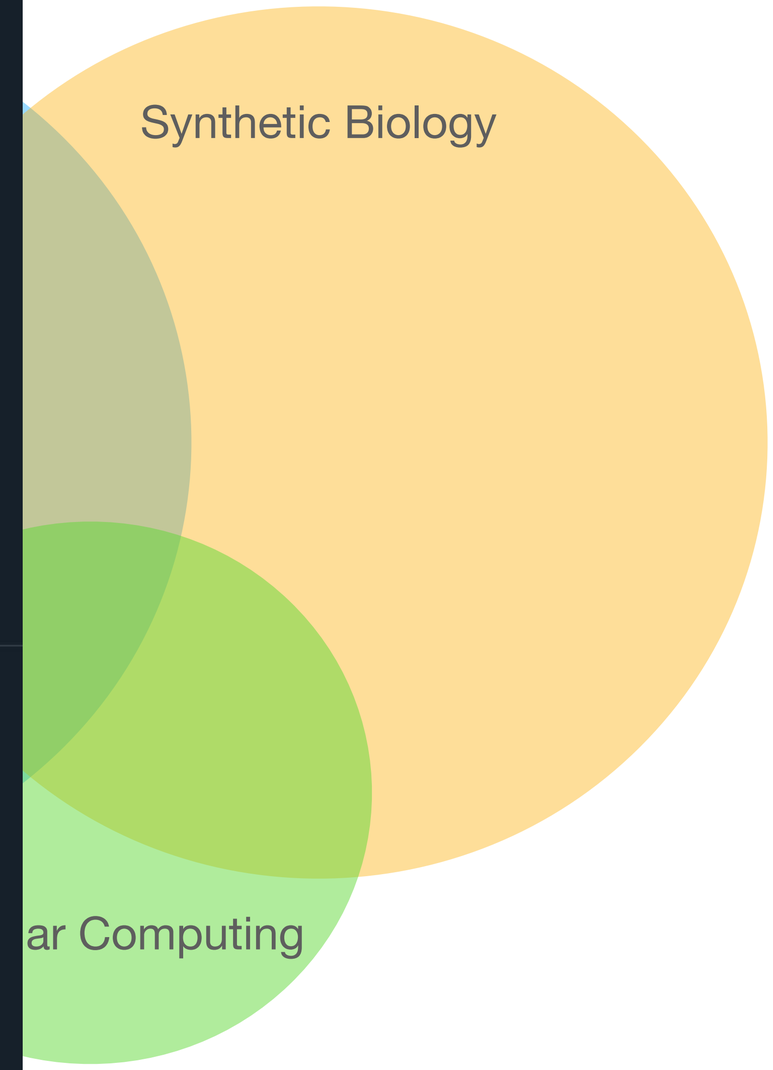


Comp Bio ∩ ing ∩ SynBio

The lines are still being draw

My two cents for people in a
these three fields:

- If you're a computation-c
person, get comfortable t
wet lab-centered people
- vice-versa
- Bigger computational AN
molecular toolboxes tend
make it easier to design
experiments
- We need more tools



Things to think about

- What are some tools you'd like to see developed?
- What are things you'd like to see standardized?
- Are there times when having a deeper background in a different field (i.e., biology) would have helped you?
- Anything other wishes for the future?