Closed-Form Test Functions for Biophysical Sequence Optimization Algorithms

Samuel Stanton, Robert Alberstein, Nathan Frey, Andrew Watkins, and Kyunghyun Cho
Genentech

Prescient Design
A Genentech Accelerator

Summary

Testing AI molecule design systems end-to-end is hard.

Experimental feedback is slow, we need something faster for development.

Simulating experimental feedback is very hard, if you've solved that you've almost solved the whole problem.

Instead of simulating experiments, what if we defined a closed-form problem with the same structure?

Paper **Code**

@samuel_stanton_

Background

What Makes A Good Benchmark?

A. **Low barriers to entry**
Should be inexpensive and easy to use.

B. **Well-characterized solutions**
Should be very clear when the benchmark is "solved".

C. **Non-trivial difficulty**
Should be hard to solve with a naïve baseline.

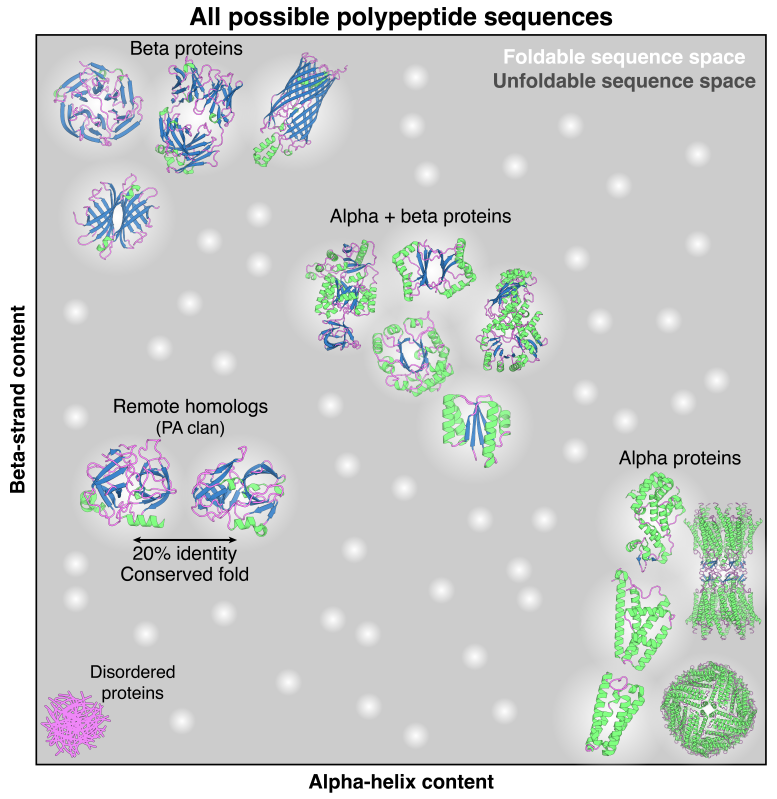
D. **Similarity to real applications**
Should have similar problem structure.

Types of Test Functions for Sequence Optimization

- **Database lookups**
  Requires brute force enumeration, $$ $, hard to verify
  Example: TF DNA Binding [1]
- **Empirical function approximation**
  Inaccurate far from data, hard to define a good trust region
  Example: TAPE fluorescence [2]
- **Physics-based simulations**
  Difficult to use, slow, and poorly characterized solutions
  Example: ΔΔG simulations [3]
- **Closed-form functions**
  Tend to be too easy and simplistic
  Example: beta sheet motif count [4]

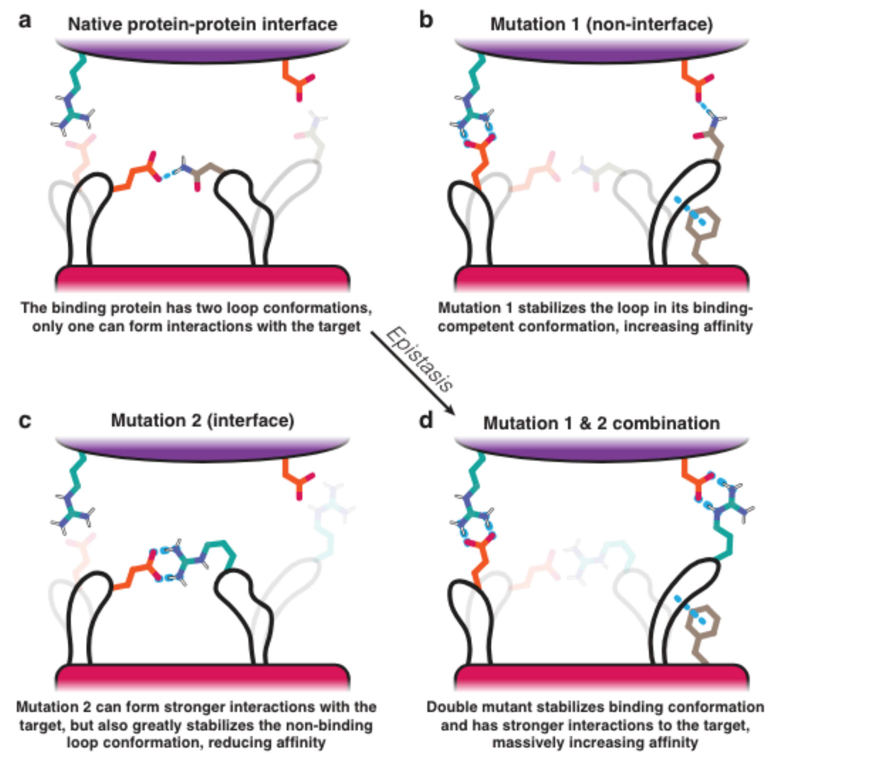
What Is The Essential Geometry of Sequence Optimization?

Feasible sequence space is sparse



Feasibility constraint: $\mathcal{F} = \{\mathbf{x} \in \mathcal{X} \mid A[x_{\ell-1}, x_\ell] > 0 \; \forall \ell \geq 2\},$

Non-additive, position-dependent sensitivity


a
Native protein-protein interface
b
Mutation 1 (non-interface)
Epistasis
c
Mutation 2 (interface)
d
Mutation 1 & 2 combination

The binding protein has two loop conformations, only one can form interactions with the target

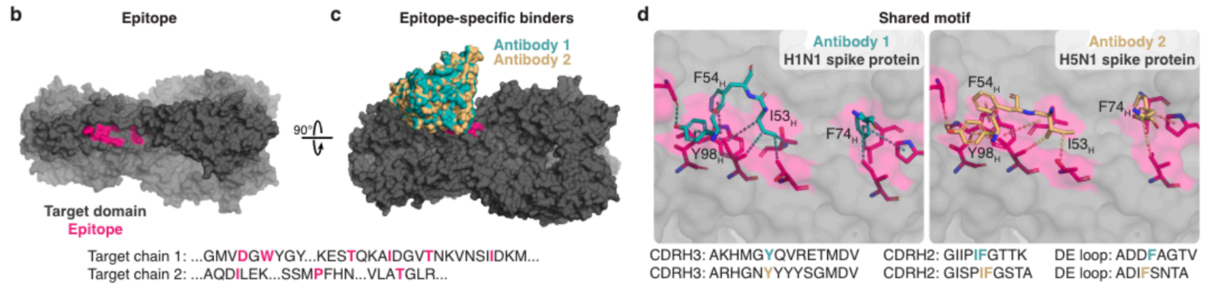
Mutation 1 stabilizes the loop in its binding-competent conformation, increasing affinity

Mutation 2 can form stronger interactions with the target, but also greatly stabilizes the non-binding loop conformation, reducing affinity

Double mutant stabilizes binding conformation and has stronger interactions to the target, massively increasing affinity

Motif satisfaction: $h_q(\mathbf{x}, \mathbf{m}^{(i)}, \mathbf{s}^{(i)}) = \max_{\ell < L} \left(\sum_{j=1}^{k} \mathbb{1}\{x_{\ell + s_j^{(i)}} = m_j^{(i)}\} \right) // \frac{k}{q} \, / q$

Many Local Optima with Similar Latent Structure



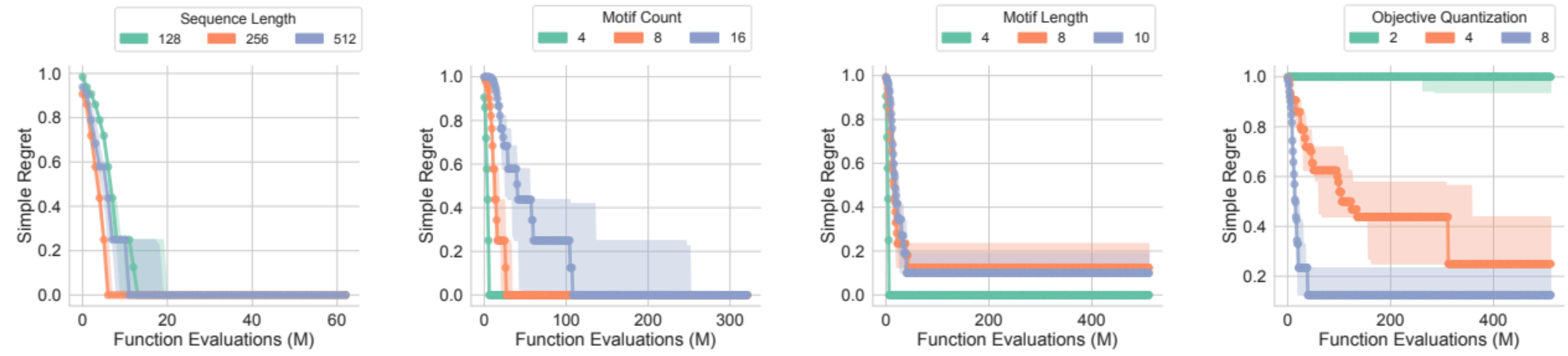
Target chain 1: ...GMVDGWYGY...KESTQKAIDGVTNKVNSIIDKM...
Target chain 2: ...AQDILEK...SSMPFHN...VLATGLR...

CDRH3: AKHMGYQVRETMDV CDRH2: GIIPIFGTTK DE loop: ADDFAGTV
CDRH3: ARHGNYYYYSGMDV CDRH2: GISPIFGSTA DE loop: ADIFSNTA

Ehrlich Function:

$$f(\mathbf{x}) = \begin{cases} \prod_{i=1}^{c} h_q(\mathbf{x}, \mathbf{m}^{(i)}, \mathbf{s}^{(i)}) & \text{if } \mathbf{x} \in \mathcal{F} \\ -\infty & \text{else} \end{cases}.$$

Ehrlich Functions Have Tunable Difficulty



Test function parameters like sequence length, # motifs, # of motif elements, and signal quantization can be varied to change the problem difficulty and reveal weaknesses in algorithms.

Easy to Install, Fast to Evaluate

Installation

```
python -m pip install pytorch-holo
```

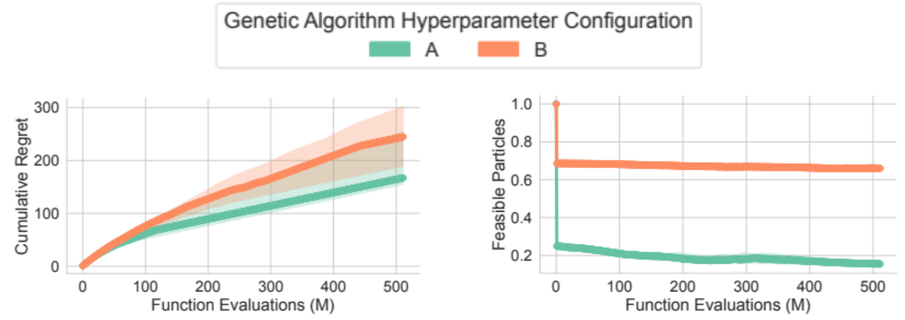
Minimal Code Example

```
import torch
from holo.test_functions import closed_form
from holo.optim import DiscreteEvolution

test_fn = closed_form.Ehrlich(negate=True)
params = [
    torch.nn.Parameter(
        test_fn.initial_solution().float(),
    )
]
optimizer = DiscreteEvolution(
    params,
    vocab=list(range(test_fn.num_states)),
    mutation_prob=1/test_fn.dim,
    recombine_prob=1/test_fn.dim,
    num_particles=1024,
    survival_quantile=0.01
)

for _ in range(4):
    loss = optimizer.step(
        lambda x: test_fn(x[0])
    )
```

200B test function calls? No problem!



Explore hyperparameter tradeoffs more systematically. Faster test functions -> more experiments.

References

1. Barrera, L. A., Vedenko, A., Kurland, J. V., Rogers, J. M., ... & Bulyk, M. L. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, *351*(6280), 1450-1454.
2. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., ... & Song, Y. (2019). Evaluating protein transfer learning with TAPE. *Advances in neural information processing systems*, *32*.
3. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic acids research*, *33*(suppl_2), W382-W388.
4. Gligorijevic, V., Berenberg, D., Ra, S., Watkins, A., Kelow, S., Cho, K., & Bonneau, R. (2021). Function-guided protein design by deep manifold sampling. bioRxiv. *preprint*.