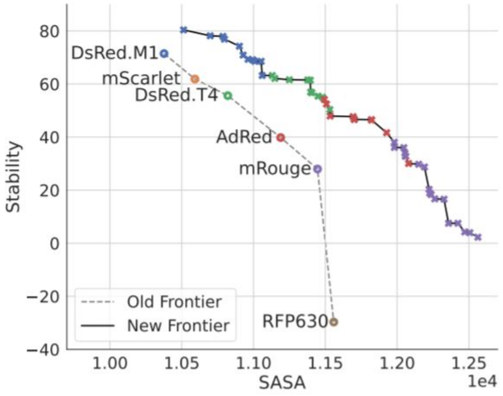
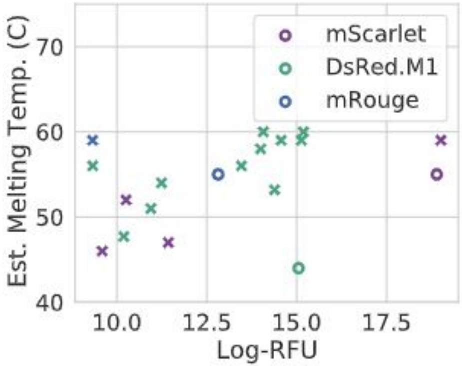
Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders

Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, Andrew Gordon Wilson

in silico



in vitro



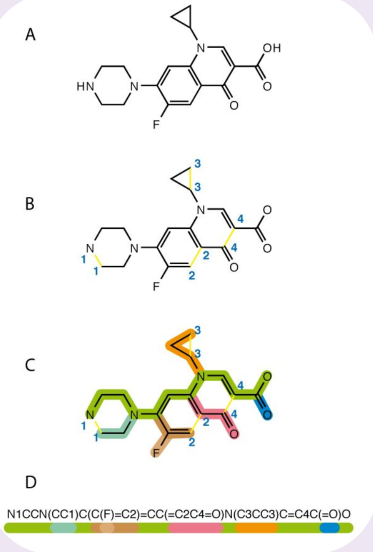
Biological Sequence Design

intractable combinatorial optimization problem

slow/impossible to compute

$$\max_{\mathbf{x}\in\mathcal{X}}(f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))$$

- Discrete, high-dim. inputs
- Multiple black-box objectives
- Batched experiments
- Noisy labels



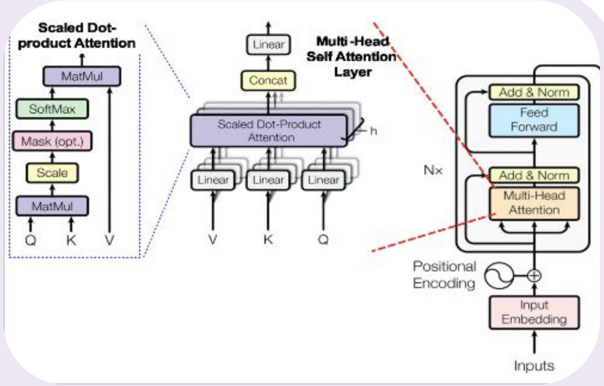
```
        10         20         30         40         50
MVSKGEELFT GVVPILVELD GDVNGHKFSV SGEGEGDATY GKLTLKFICT
        60         70         80         90        100
TGKLPVPWPT LVTTLTYGVQ CFSRYPDHMK QHDFFKSAMP EGYVQERTIF
       110        120        130        140        150
FKDDGNYKTR AEVKFEGDTL VNRIELKGID FKEDGNILGH KLEYNYNSHN
       160        170        180        190        200
VYIMADKQKN GIKVNFKIRH NIEDGSVQLA DHYQQNTPIG DGPVLLPDNH
       210        220        230
YLSTQSALSK DPNEKRDHMV LLEFVTAAGI TLGMDELYK
```

Deep Generative Models (DGMs)

optimize over small DGM-generated subset

ad-hoc proxy objective

$$\max_{\mathbf{x}\in\mathcal{X}'} ???$$



- How do we rank generated sequences, accounting for multiple objectives, explore-exploit, etc.?

- How do we generate good subsets for ranking?

ICML International Conference On Machine Learning

Figure reproduced from Shu, Raphael, ..K. Cho *Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior.* Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.

Bayesian Optimization (BayesOpt)

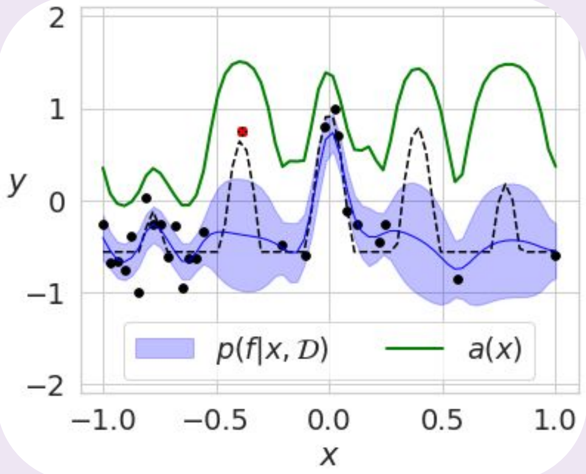
intractable combinatorial optimization subproblem

principled proxy objective

$$\max_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[u(\mathbf{x})]$$

- Strategy 1: solve subproblem with genetic algorithms (inefficient).

- Strategy 2: use a frozen pretrained generative model, solve in latent space (data-hungry).



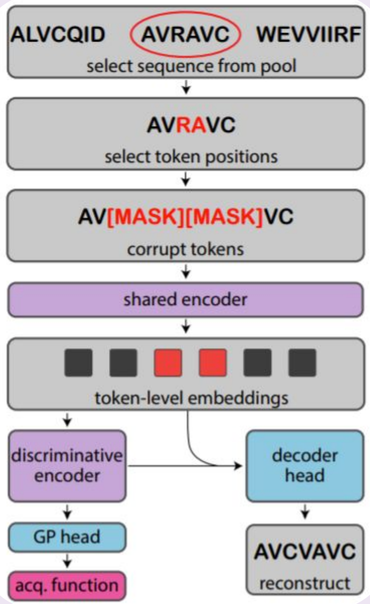
Latent Multi-Objective BayesOpt (LaMBO)

LaMBO is designed from the ground up to combine the best attributes of DGMs and BayesOpt.

ICML
International Conference On Machine Learning

The LaMBO architecture

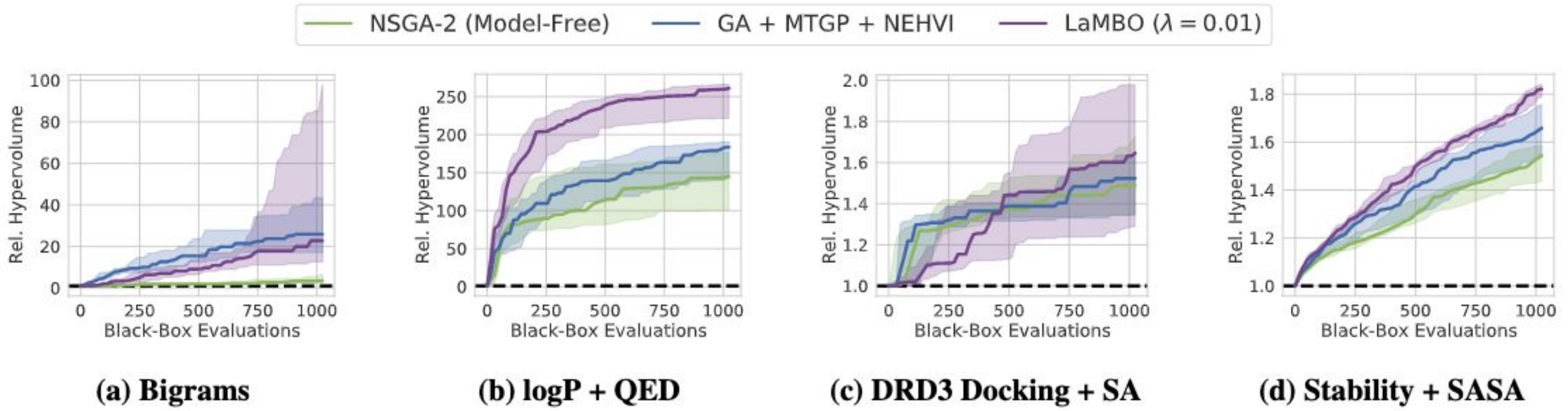
- Jointly train a denoising autoencoder and a discriminative deep kernel GP.
- Rank samples with the NEHVI acquisition function, optimize in latent space.
- Pretraining is optional!



ICML International Conference On Machine Learning

Previewing the results

Comparing **LaMBO** to model-free and model-based genetic algorithms



small molecule

large molecule

NYU

BigHat BIOSCIENCES

Collaborators

NYU

Samuel Stanton

Wesley Maddox

Nate Gruver

Andrew Wilson

BigHat BIOSCIENCES

Phil Maffetone

Emily Delaney

Peyton Greenside

Thanks, come check out the poster!

Thursday, July 21, 6PM - 8PM
Poster Session 1, Hall E #533

@samuel_stanton_

ss13641@nyu.edu

Paper

SCAN ME

Code

SCAN ME

ICML
International Conference On Machine Learning