GROOT : Effective Design of Biological Sequences with Limited Experimental Data

Thanh V. T. Tran*, Nhat Khang Ngo*, Viet Anh Nguyen, Truong Son Hy

Motivation

- MBO requires large amount of labeled data, which is limited.

➡ Sub-optimal performance ☹

- Rich unlabeled data.

➡ Utilize to generate better designs 😊

Overview

1. Train a robust VAE with unlabeled data

➡ Rich representations in the latent space

2. Process labeled data in the latent space

- Interpolate to create new data point

$$z = \beta * \overline{x} + (1 - \beta) * \epsilon$$

- Construct graph in the latent space

$$\mathcal{V} \leftarrow \mathcal{V} \cup \{z\}$$
$$\mathcal{E} \leftarrow \cup_{x \in \mathcal{V}} \mathrm{kNN}(x, \mathcal{V})$$

- Apply smoothing to generate pseudo-labels

$$\tilde{Y} \leftarrow \mathrm{LabelPropagation}(\mathcal{E}, \mathcal{V}, Y)$$

3. GROOT-augmented MBO

- Train surrogate model based on newly generated data

➡ Versatile for optimization across domains.

Theoretical Justification

Definition 1. *Let $\mathbb{X} = \{x_1, \ldots, x_N\}$ be a set of N points in $\mathbb{R}^d$, a convex hull of $\mathbb{X}$ is defined as:*

$$\mathrm{Conv}(\mathbb{X}) \triangleq \left\{ \sum_{i=1}^{N} \lambda_i x_i | \lambda_i \geq 0 \text{ and } \sum_{i=1}^{N} \lambda_i = 1 \right\}$$

Definition 2. *Interpolation occurs for a sample whenever this sample belongs to $\mathrm{Conv}(\mathbb{X})$, if not, extrapolation occurs*

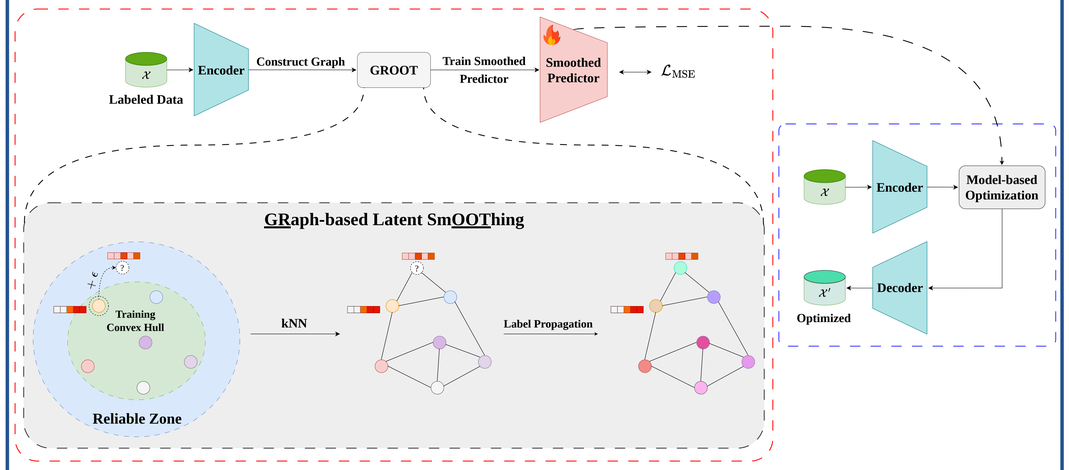
Proposition 1. *In limited data scenario:*

$$\lim_{d \to \infty} \mathbf{P}(z \notin \mathrm{Conv}(\mathbb{X})) = 1$$

Proposition 2. $\mathbb{E}[D(z, \mathrm{Conv}(\mathbb{X}))] < 2(1 - \beta)\sqrt{d}$

Method

GROOT-augmented Model-based Optimization



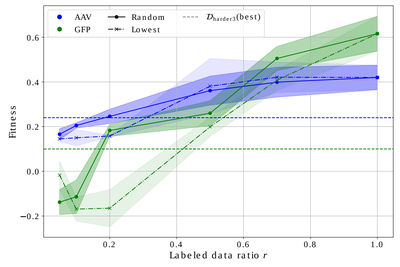
Experimental Results

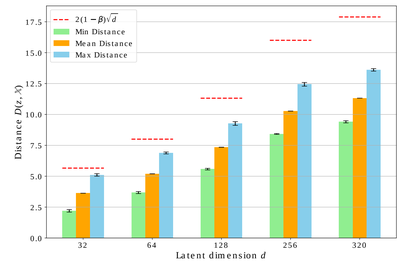
Main results. GROOT outperform SOTA methods in label-scarcity scenarios.

Method	AAV *harder1* task Fitness ↑	Diversity	Novelty	AAV *harder2* task Fitness ↑	Diversity	Novelty	AAV *harder3* task Fitness ↑	Diversity	Novelty
AdaLead	0.38 (0.0)	5.5 (0.5)	7.0 (0.7)	0.43 (0.0)	4.2 (0.7)	7.8 (0.8)	0.37 (0.0)	6.22 (0.9)	8.0 (1.2)
CbAS	0.02 (0.0)	22.9 (0.1)	18.5 (0.5)	0.01 (0.0)	23.2 (0.1)	19.3 (0.4)	0.01 (0.0)	23.2 (0.1)	19.3 (0.4)
GFN-AL	0.00 (0.0)	15.4 (6.2)	21.6 (0.5)	0.00 (0.0)	8.1 (3.5)	21.6 (1.0)	0.00 (0.0)	7.6 (0.8)	22.6 (1.4)
PEX	0.23 (0.0)	6.4 (0.5)	3.8 (0.7)	0.30 (0.0)	7.8 (0.4)	5.0 (0.0)	0.26 (0.0)	7.3 (0.7)	4.4 (0.5)
GGS	0.30 (0.0)	13.6 (0.2)	14.5 (0.3)	0.27 (0.0)	16.0 (0.0)	19.4 (0.0)	0.38 (0.0)	7.0 (0.1)	9.6 (0.1)
ReLSO	0.15 (0.0)	20.9 (0.0)	13.0 (0.0)	0.17 (0.0)	20.3 (0.0)	13.0 (0.0)	0.22 (0.0)	17.8 (0.0)	11.0 (0.0)
S-ReLSO	0.24 (0.0)	11.5 (0.0)	13.0 (0.0)	0.28 (0.0)	16.4 (0.0)	6.5 (0.0)	0.27 (0.0)	17.7 (0.0)	11.0 (0.0)
GROOT	**0.46 (0.1)**	9.8 (1.6)	12.2 (0.5)	**0.45 (0.0)**	9.9 (0.8)	13.0 (0.0)	**0.42 (0.1)**	11.0 (2.0)	13.0 (0.0)
	GFP *harder1* task			GFP *harder2* task			GFP *harder3* task		
Method	Fitness ↑	Diversity	Novelty	Fitness ↑	Diversity	Novelty	Fitness ↑	Diversity	Novelty
AdaLead	0.39 (0.0)	8.4 (3.2)	9.0 (1.2)	0.4 (0.0)	7.3 (2.8)	9.8 (0.4)	0.42 (0.0)	6.4 (2.3)	9.0 (1.2)
CbAS	-0.08 (0.0)	172.2 (35.7)	201.5 (1.5)	-0.09 (0.0)	158.4 (34.8)	202.0 (0.7)	-0.08 (0.0)	186.4 (33.4)	201.5 (0.9)
GFN-AL	0.21 (0.1)	74.3 (55.3)	219.2 (3.3)	0.14 (0.2)	27.0 (9.5)	223.5 (2.4)	0.21 (0.0)	37.5 (21.7)	219.8 (4.3)
PEX	0.13 (0.0)	12.6 (1.2)	7.1 (1.1)	0.17 (0.0)	12.6 (1.2)	7.1 (1.1)	0.19 (0.0)	12.2 (1.1)	7.8 (1.7)
GGS	0.67 (0.0)	4.7 (0.2)	9.1 (0.1)	0.60 (0.0)	5.4 (0.2)	9.8 (0.1)	0.00 (0.0)	15.7 (0.4)	19.0 (2.2)
ReLSO	*0.94 (0.0)*†	0.0 (0.0)	8.0 (0.0)	*0.94 (0.0)*†	0.0 (0.0)	8.0 (0.0)	*0.94 (0.0)*†	0.0 (0.0)	8.0 (0.0)
S-ReLSO	*0.94 (0.0)*†	0.0 (0.0)	8.0 (0.0)	*0.94 (0.0)*†	0.0 (0.0)	8.0 (0.0)	*0.94 (0.0)*†	0.0 (0.0)	8.0 (0.0)
GROOT	**0.88 (0.0)**	3.0 (0.2)	7.0 (0.0)	**0.87 (0.0)**	3.0 (0.1)	7.5 (0.5)	**0.62 (0.2)**	7.6 (1.5)	8.6 (1.5)

† indicates that the generated population has collapsed (i.e., producing only a single sequence).

Ablation Studies. Verify the propositions and test with extreme cases (under 100 labeled data points)





Ablation Studies. Demonstrate the effect of smoothing technique

Task	Difficulty	Smoothed	Train MAE ↓	Holdout MAE ↓
AAV	Harder1	No	4.94	8.93
		Yes	**1.02**	**5.78**
	Harder2	No	4.67	8.91
		Yes	**1.10**	**6.11**
	Harder3	No	4.13	8.87
		Yes	**1.44**	**7.30**
GFP	Harder1	No	1.39	2.81
		Yes	**0.22**	**1.71**
	Harder2	No	1.34	2.81
		Yes	**0.31**	**1.85**
	Harder3	No	1.33	2.80
		Yes	**0.51**	**2.09**

Task	Difficulty	Smoothed	Fitness ↑	Diversity	Novelty
AAV	Harder1	No	0.12	20.0	10.0
		Yes	**0.46**	9.8	12.2
	Harder2	No	0.11	20.0	9.6
		Yes	**0.45**	9.9	13.0
	Harder3	No	0.12	20.1	10.0
		Yes	**0.42**	11.0	13.0
GFP	Harder1	No	-0.12	71.0	42.2
		Yes	**0.88**	3.0	7.0
	Harder2	No	-0.18	69.5	41.1
		Yes	**0.87**	3.0	7.5
	Harder3	No	-0.17	64.0	37.0
		Yes	**0.62**	7.6	8.6