TORONTO

KDD2025

RESEARCH TRACK

# GROOT: Effective Design of Biological Sequences with Limited Experimental Data

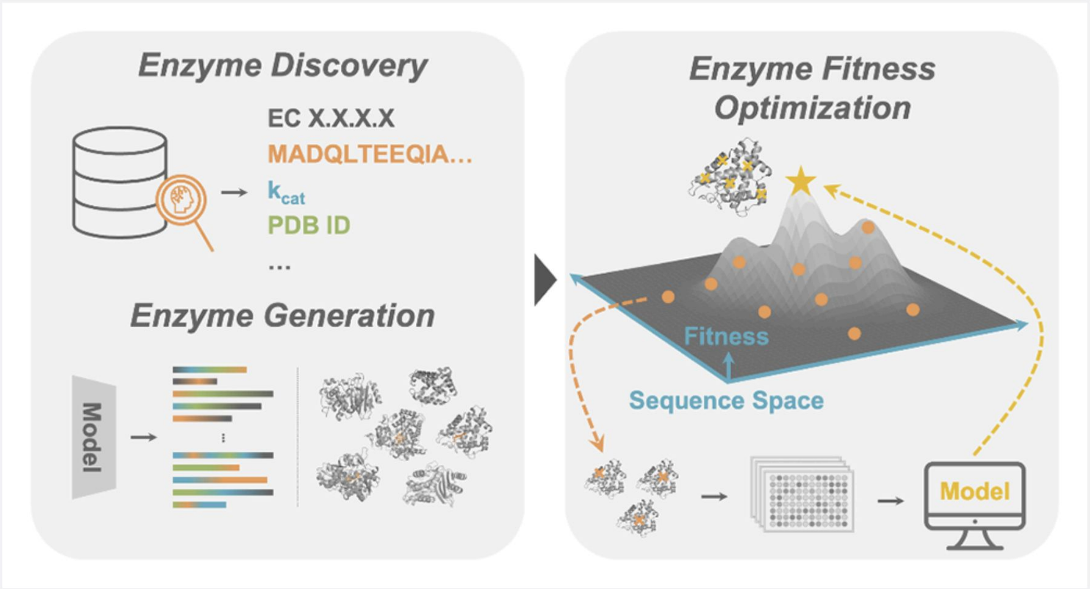
by:


**Thanh V. T. Tran** → FPT Software AI Center
**Nhat Khang Ngo** → FPT Software AI Center
**Viet Anh Nguyen** → FPT Software AI Center
**Truong Son Hy** → University of Alabama at Birmingham


August 3-7, 2025
KDD2025
TORONTO

# Motivation

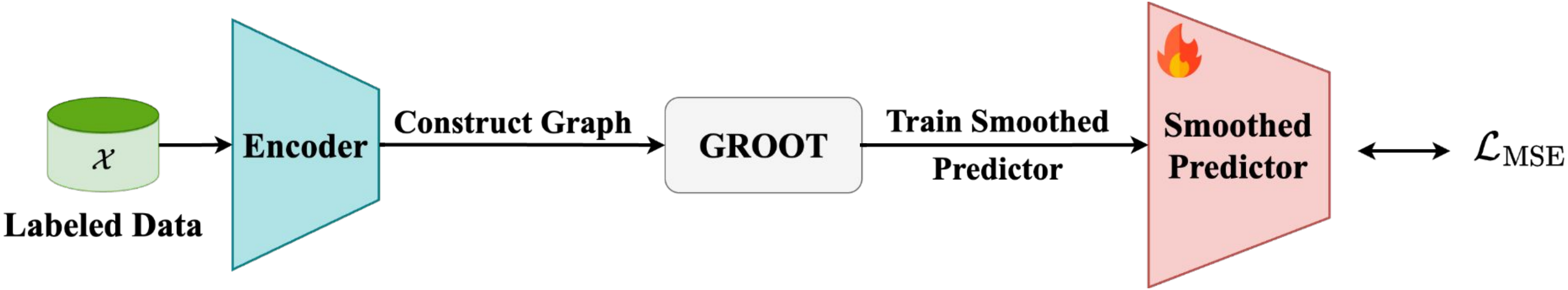
- **Rapid design via ML**: Modern protein optimization models can propose thousands of variants in silico, dramatically accelerating the design cycle.
- **Limited data scenarios**: Existing methods neglect the scenario of extremely limited labeled data and fail to utilize the abundant unlabeled data.
- **Label bottleneck**: Training such models **requires ground-truth fitness labels**, yet each measurement demands cloning, expression, purification, and functional assays—costly in time, money, and lab resources.
- **Need for sample-efficient methods**: To unlock ML-guided experimental design, we must develop algorithms that **learn from few labels**, reducing wet-lab burden without sacrificing performance.




[1] Yang, Jason, Francesca-Zhoufan Li, and Frances H. Arnold. "Opportunities and challenges for machine learning-assisted enzyme engineering." ACS Central Science 10.2 (2024): 226-241.

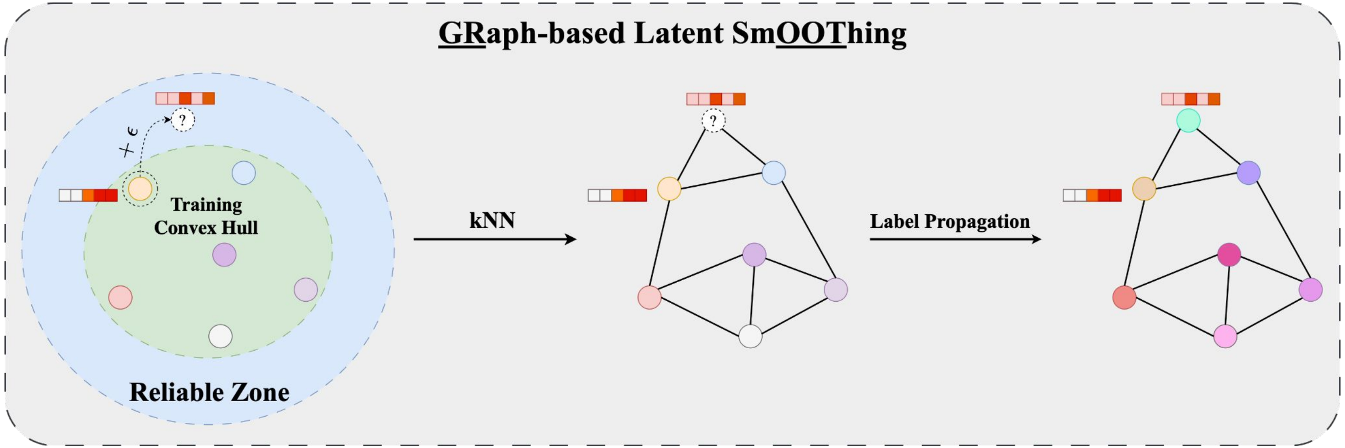
# Method: Overview






1. Train a VAE with labeled $\mathbb{X}$ and unlabeled data.
2. **Encode labeled sequences** → latent vectors.
3. **GROOT** synthesizes new samples $\mathbb{S}$ with pseudo experimental labels.
4. **Train a predictor** on these samples ($\mathbb{X} \cap \mathbb{S}$) with MSE loss.

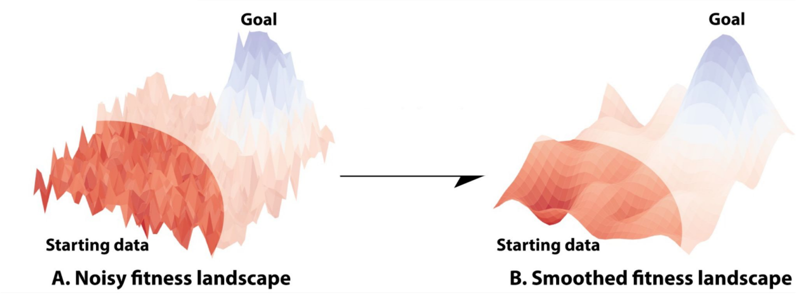
# Method: GROOT




1. **Introduce mutations** into latent sample.

2. **Construct kNN graph** in the latent space.

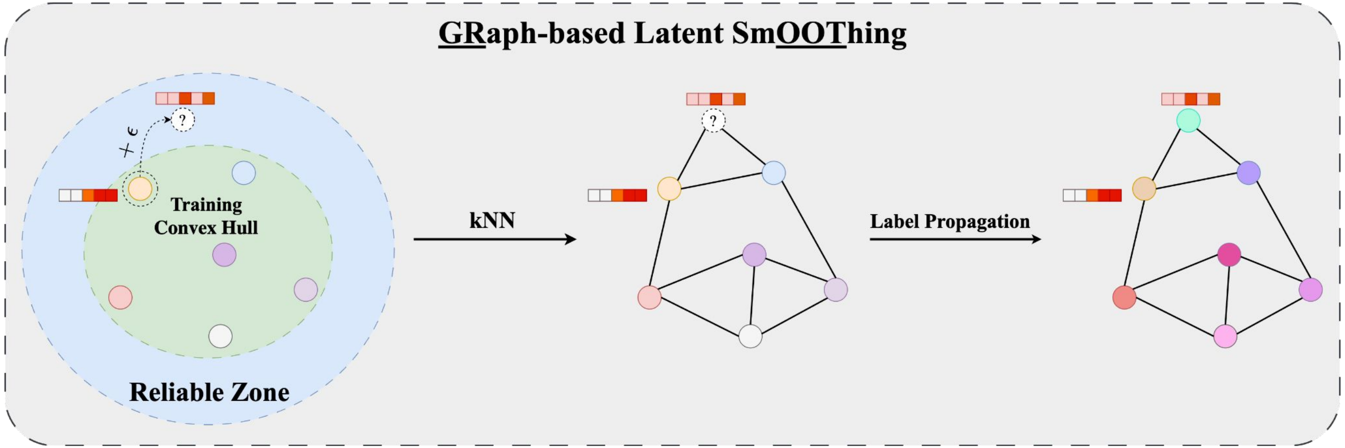
3. **Assign** pseudo labels for new samples and **smoothen** the landscape.




*[2] Kirjner, Andrew, et al. "Improving protein optimization with smoothed fitness landscapes." ICLR. 2024.*

# Why reliable?

August 3-7, 2025 KDD2025



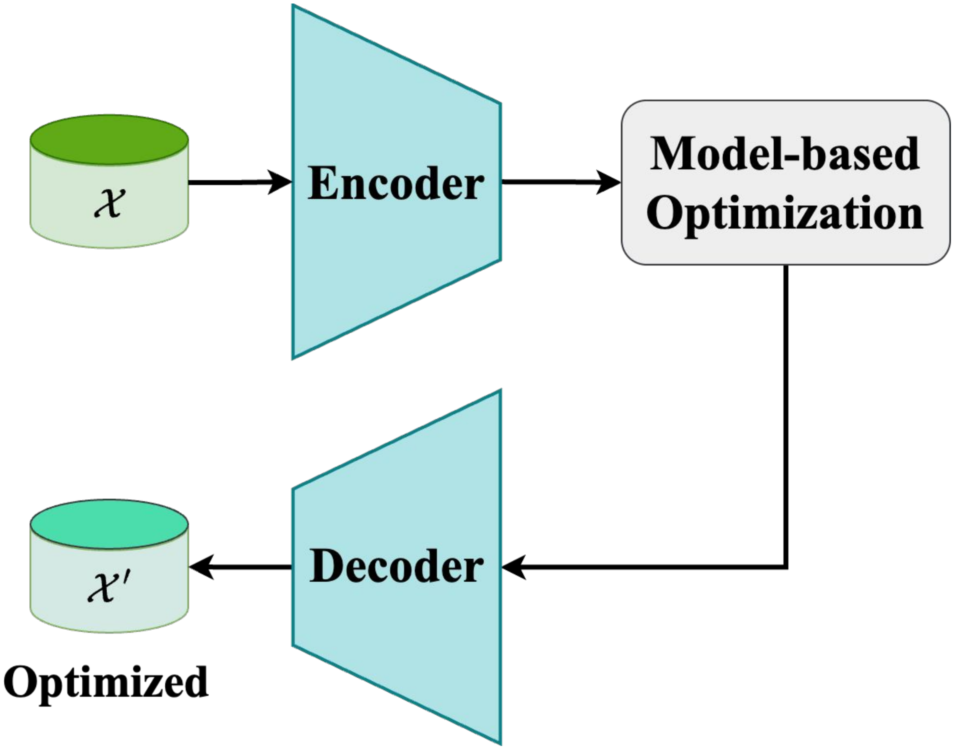

***Question:*** *Are assigned pseudo labels reliable?*

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

Distance from synthesized samples and training set are **<u>constrained</u>** and **<u>controllable</u>**

# Method: MBO






1. **Encode** initial sequences **X** → latent embeddings.
2. **MBO search**: surrogate explores the latent space for points with maximal predicted fitness.
3. **Decode** optimized embeddings → novel sequences **X′** ready for validation.

# Data Preparation

August 3-7, 2025 KDD2025

**Dataset and Task Definition**

| Task | Difficulty | Fitness Range (%) | Mutational Gap | Best Fitness | $\|\mathcal{D}\|$ |
|---|---|---|---|---|---|
| AAV | Harder1 | $< 30$th | 13 | 0.33 | 1157 |
| | Harder2 | $< 20$th | 13 | 0.29 | 920 |
| | Harder3 | $< 10$th | 13 | 0.24 | 476 |
| GFP | Harder1 | $< 30$th | 8 | 0.10 | 1129 |
| | Harder2 | $< 20$th | 8 | 0.01 | 792 |
| | Harder3 | $< 10$th | 8 | 0.01 | 397 |

*[2] Kirjner, Andrew, et al. "Improving protein optimization with smoothed fitness landscapes." ICLR. 2024.*

# Experimental Results

| Method | AAV *harder1* task | | | AAV *harder2* task | | | AAV *harder3* task | | |
|---|---|---|---|---|---|---|---|---|---|
| | Fitness ↑ | Diversity | Novelty | Fitness ↑ | Diversity | Novelty | 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) |
| BO | 0.00 (0.0) | 20.4 (0.3) | 21.8 (0.4) | 0.01 (0.0) | 20.4 (0.0) | 22.0 (0.0) | 0.01 (0.0) | 20.6 (0.3) | 22.0 (0.0) |
| 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 (GA)** | 0.37 (0.0) | 13.6 (0.9) | 10.0 (0.7) | 0.36 (0.0) | 13.7 (1.1) | 10.1 (0.9) | 0.34 (0.1) | 14.0 (2.2) | 10.0 (1.4) |
| **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) |

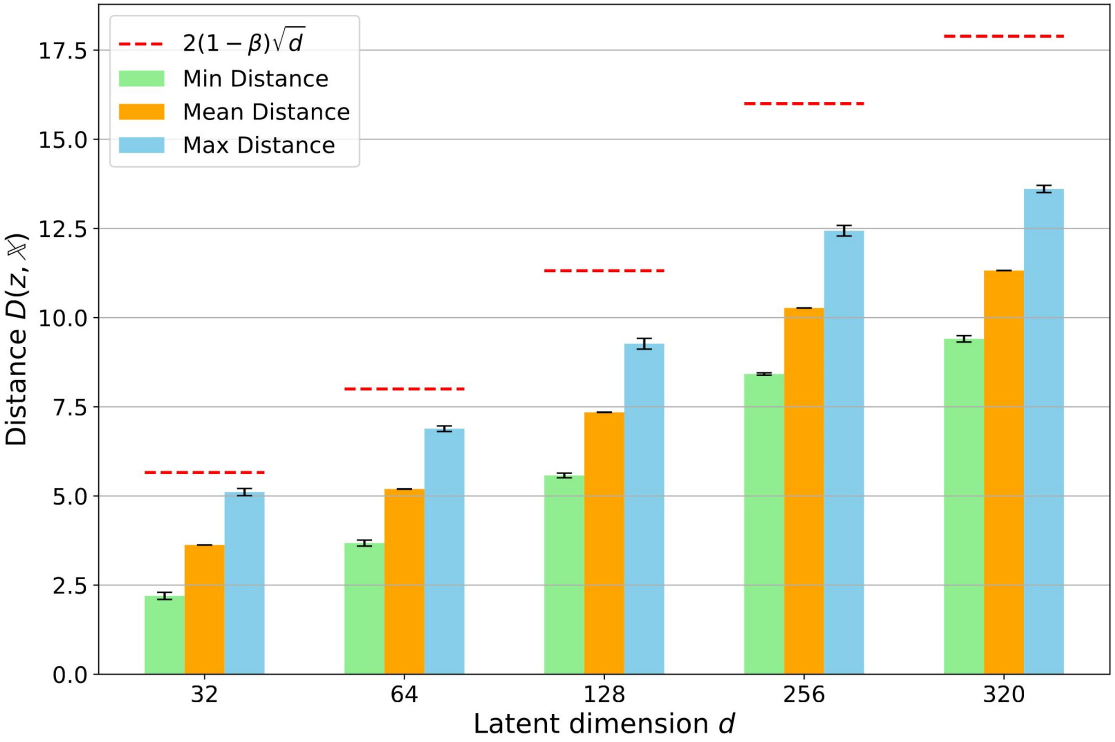
| Method | GFP *harder1* task | | | GFP *harder2* task | | | GFP *harder3* task | | |
|---|---|---|---|---|---|---|---|---|---|
| | Fitness ↑ | Diversity | Novelty | Fitness ↑ | Diversity | Novelty | Fitness ↑ | Diversity | Novelty |
| AdaLead | 0.39 (0.0) | 8.4 (3.2) | 9.0 (1.2) | 0.4 (0.0) | 7.3 (2.8) | 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).

**In-silico Evaluator**: An independent oracle whose checkpoint are taken from [2].

# Empirical Validation on Upper Bound

August 3–7, 2025
KDD2025



# *Effectiveness of GROOT*

August 3-7, 2025 KDD2025

| 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 |

# THANK YOU!

KDD2025
August 3–7, 2025