Quality Diversity for One-Shot Biological Sequence Design

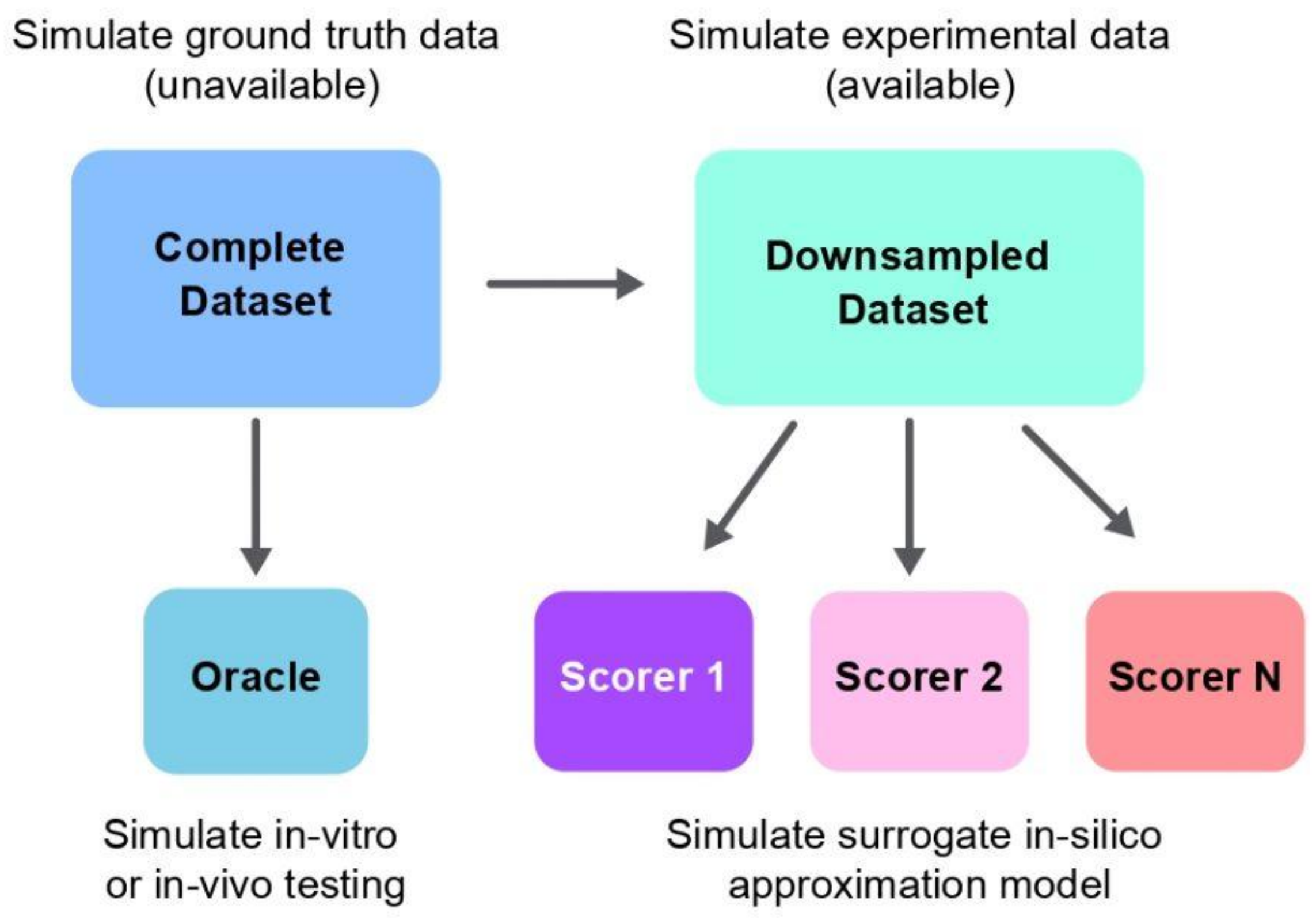
Jérémie Donà[1], Arthur Flajolet[1], Andrei Marginean[1], Antoine Cully[2], Thomas Pierrot[1]

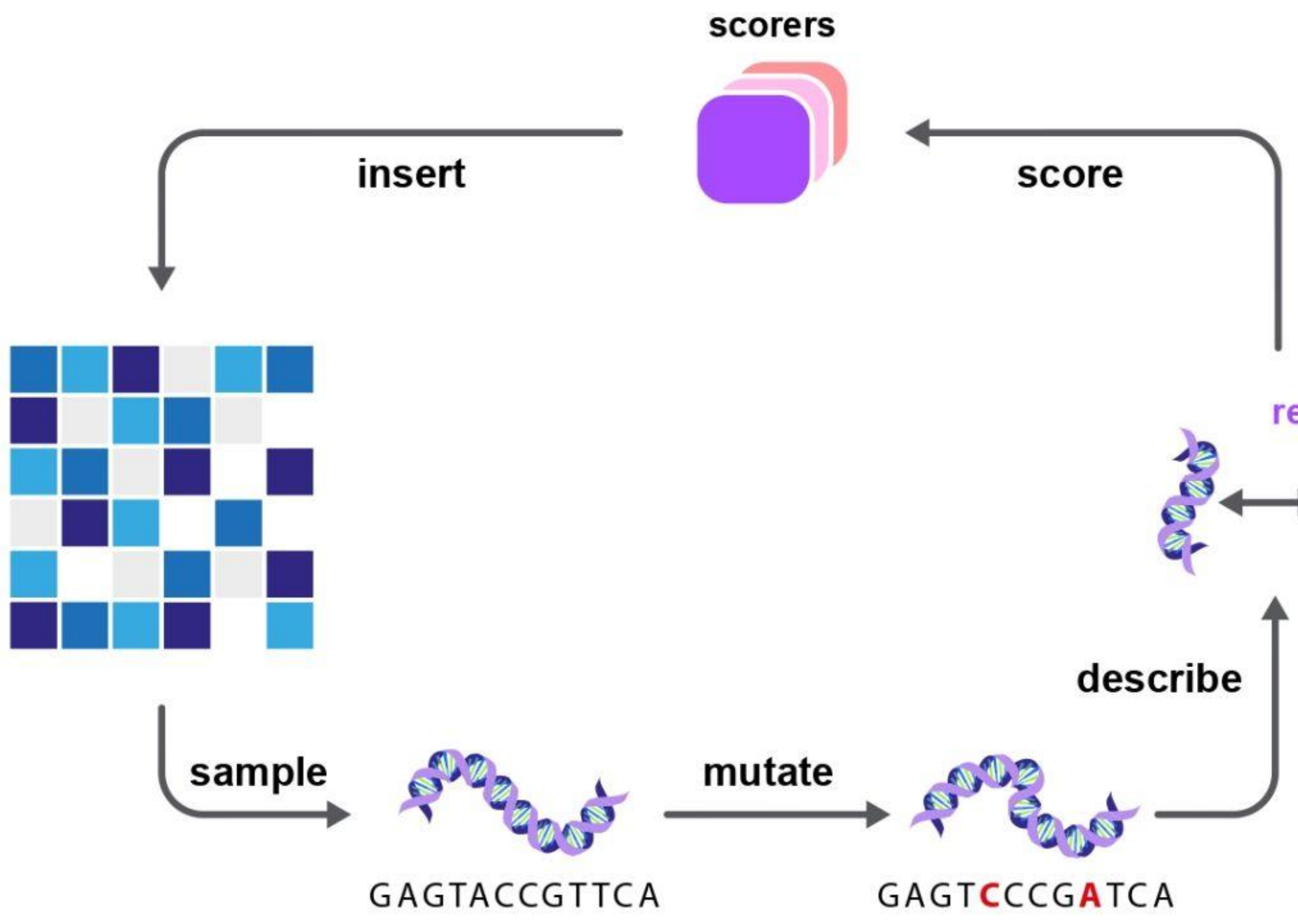
[1]InstaDeep, Paris, France [2]Imperial College, London, United Kingdom

Objectives and Task

We are given a paired dataset of sequences and structures : $\{(x_i, y_i)\}_{i \leq N}$

We aim at designing new sequences that optimize the property But in vitro / vivo experiments are expensive and in silico evaluation may be inaccurate: **Propose a batch of diverse sequences to maximize the odds to get at least one working sequence !**





Example for the design of 5 UTR (nucleotide sequence).

We resort to Quality-Diversity (QD) Formalism!

Fitness and Description

QD Algorithms need two main functions:

1. *Score*: estimating the property we optimize
2. *Descriptors*: characterizing the designed sequences

Score: Be Conservative

Learn **robust** scorers that minimize the error and penalize high prediction of adversarially constructed inputs:

$$L(\theta) = \mathbb{E}_{p_{data}} \|y - f_\theta(x)\|_2^2 + \alpha(\mathbb{E}_{\mu(x)}[f_\theta] - \mathbb{E}_{p_{data}}[f_\theta])$$

Optimize a lower bound of robust scorers:

$$s(x) = L^{-1} \sum \hat{f}_{\theta_l}(x) - \beta.\hat{\sigma}\left((\hat{f}_{\theta_l}(x))_{1 \leq l \leq L}\right)$$

Description: Leverage your Datasets

Get a subsample $\mathcal{X}^{ref}$ of the sequence dataset. Compare any proposed sequence to your reference set:

$$\phi(x) = \left(\mathrm{H}(x, z)\right)_{z \in \mathcal{X}^{\mathrm{ref}}} \in \mathbb{R}^{N_{ref}}$$

Normalize this similarity vector, push sequences distant from $\mathcal{X}^{ref}$ far from the origin of the behaviour description space:

$$b(x) = e^{d_{nn}(x)} W \phi_n(x) \in \mathbb{R}^d$$

Experiments

From N=128 sequences output 128 new sequences optimized for the property at stake. We test on 3 datasets: 1) Antibody Design. 2) 5'UTR. 3) Green Fluorescence Proteins.

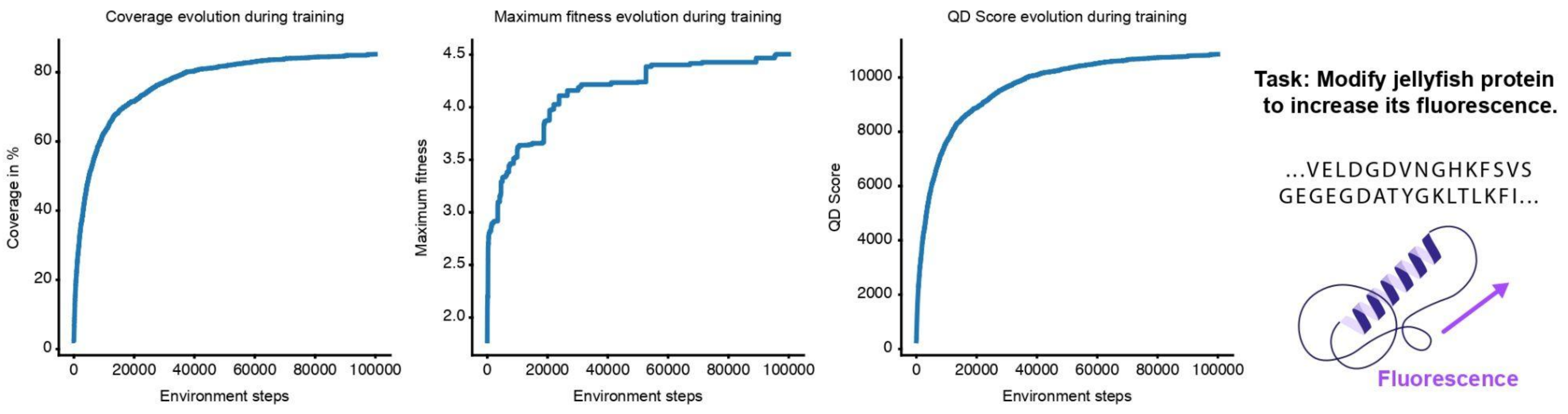


Fig.1: Evolution of QD metrics during the optimization for the GFP dataset

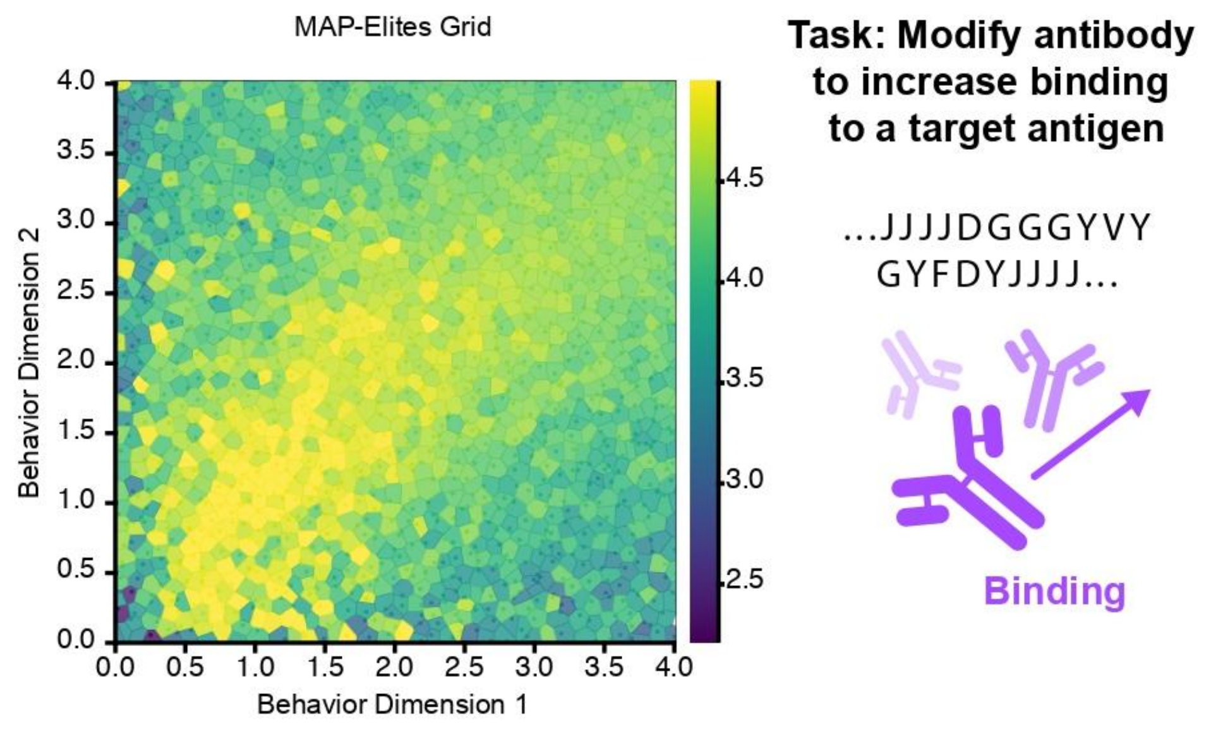


Fig.2: Illuminated QD-repertoire for the AB dataset

Method	Max	Mean	Diversity	Novelty
CBAS	0.69 ± 0.00	0.55 ± 0.00	37.1 ± 0.0	22.5 ± 0.0
CMA-ES	0.71 ± 0.00	0.66 ± 0.00	$\mathbf{39.6 \pm 0.1}$	$\mathbf{48.4 \pm 0.1}$
COMS	0.70 ± 0.00	0.68 ± 0.00	36.5 ± 0.0	1.0 ± 0.1
GA	0.69 ± 0.00	0.68 ± 0.00	25.5 ± 1.9	11.8 ± 5.4
GFLOWNET	0.70 ± 0.05	0.56 ± 0.06	27.3 ± 4.2	21.6 ± 1.0
GRAD	0.70 ± 0.00	0.65 ± 0.01	33.9 ± 0.6	22.4 ± 0.2
REINFORCE	0.68 ± 0.00	0.53 ± 0.01	36.9 ± 0.0	22.6 ± 0.0
OURS	0.71 ± 0.01	$\mathbf{0.69 \pm 0.00}$	14.7 ± 2.5	22.0 ± 0.6
OURS-BIO	$\mathbf{0.72 \pm 0.00}$	0.68 ± 0.00	35.1 ± 1.6	21.7 ± 0.3

Method	Max	Mean	Diversity	Novelty
CBAS	0.55 ± 0.02	0.34 ± 0.01	12.9 ± 0.1	6.38 ± 0.1
CMA-ES	0.53 ± 0.00	0.43 ± 0.01	$\mathbf{19.0 \pm 0.0}$	$\mathbf{19.8 \pm 0.1}$
COMS	$\mathbf{0.67 \pm 0.03}$	0.52 ± 0.03	11.3 ± 0.5	12.0 ± 0.7
GA	0.55 ± 0.02	0.40 ± 0.00	13.3 ± 0.2	6.2 ± 0.1
GFLOWNET	0.41 ± 0.01	0.28 ± 0.00	12.6 ± 0.2	5.7 ± 0.2
GRAD	0.64 ± 0.02	0.55 ± 0.02	3.2 ± 1.2	16.8 ± 0.3
REINFORCE	0.44 ± 0.03	0.32 ± 0.02	12.6 ± 0.8	7.1 ± 0.6
OURS	0.66 ± 0.02	$\mathbf{0.56 \pm 0.01}$	9.7 ± 0.4	7.1 ± 0.3
OURS-BIO	0.64 ± 0.02	0.50 ± 0.01	12.6 ± 0.3	8.0 ± 0.6

Method	Max	Mean	Diversity	Novelty
CBAS	0.84 ± 0.05	0.81 ± 0.06	2.7 ± 1.3	0.9 ± 0.4
CMA-ES	0.00 ± 0.09	-0.19 ± 0.04	$\mathbf{232 \pm 0.4}$	$\mathbf{200 \pm 1.5}$
COMS	0.86 ± 0.00	0.75 ± 0.00	5.9 ± 0.0	0.0 ± 0.1
GA	0.86 ± 0.01	0.80 ± 0.00	7.9 ± 0.2	3.7 ± 0.1
GFLOWNET	0.86 ± 0.02	0.42 ± 0.16	86.9 ± 9.9	110.3 ± 12
GRAD	0.86 ± 0.00	0.75 ± 0.00	5.9 ± 0.0	1.1 ± 0.7
REINFORCE	0.83 ± 0.06	0.71 ± 0.03	5.9 ± 0.0	2.1 ± 0.0
OURS	0.86 ± 0.00	$\mathbf{0.82 \pm 0.01}$	8.5 ± 0.4	4.3 ± 0.3
OURS-BIO	$\mathbf{0.87 \pm 0.00}$	0.44 ± 0.08	8.2 ± 0.9	8.2 ± 0.2

Tables: Max and mean fitness, diversity and novelty of the generated sequences on respectively 5' UTR, Antibody design and GFP dataset

ICML
International Conference
On Machine Learning