ProSpero: Active Learning for Robust Protein Design Beyond Wild-Type Neighborhoods

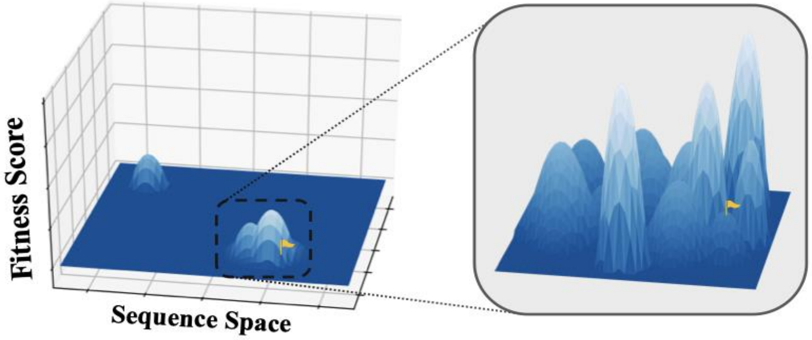
Michal Kmicikiewicz Vincent Fortuin Ewa Szczurek

HELMHOLTZ MUNICH

Introduction & Motivation

Introduction

Main goal – Designing novel protein sequences with desired properties
Challenges:
- rugged and sparse ”fitness landscape”
- combinatorial search space
- expensive black-box evaluations



Z. Ren et al., Proximal Exploration for Model-Guided Protein Sequence Design, ICML 2022

HELMHOLTZ MUNICH

How to reliably explore further away from the wild-type?

Wild-type – A naturally occurring sequence serving as a reference/starting point in the optimization

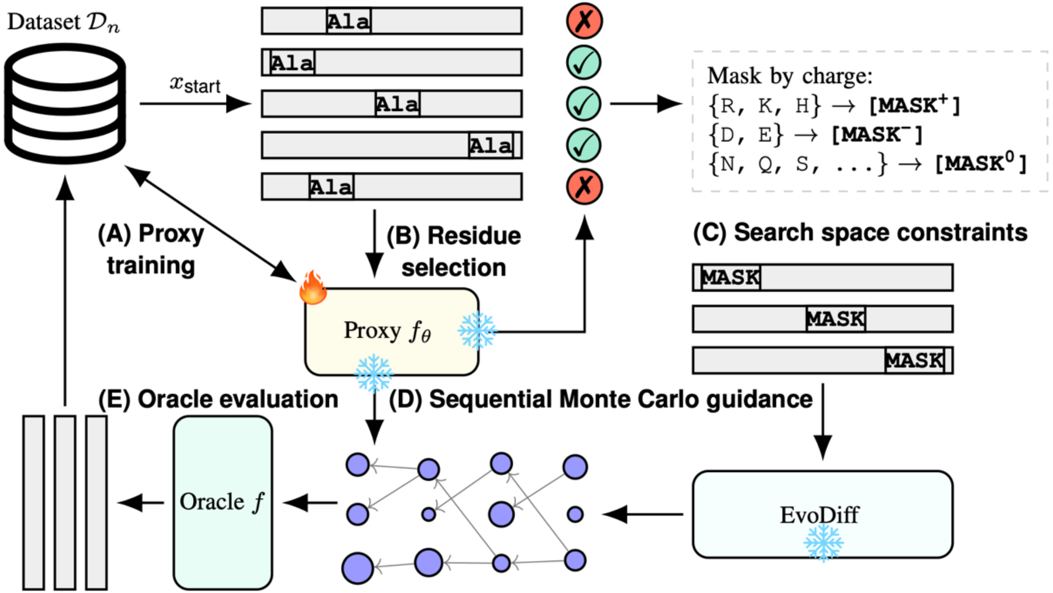
Potential solutions:

- Active learning: iterative re-training of the surrogate to progressively expand its support

- Biological priors: leverage prior biological knowledge to ensure plausibility even with a potentially misspecified surrogate

Proposed Framework

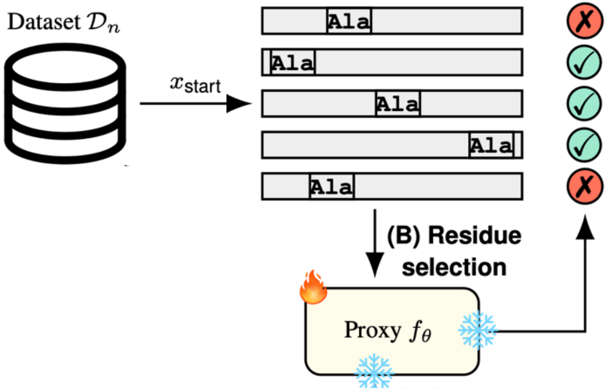
ProSpero



Inference-time guidance of a pre-trained pLM with a surrogate updated in an active learning loop – seamless integration of biological priors into online optimization, regardless of the target protein family

HELMHOLTZ MUNICH

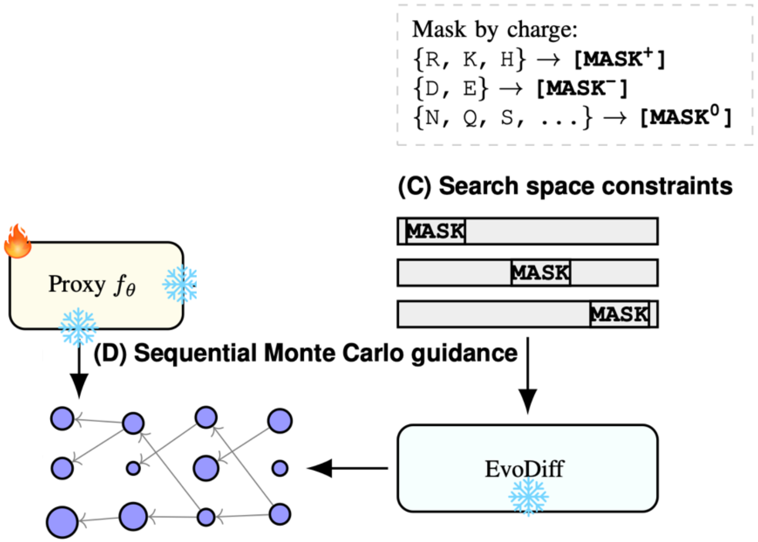
Targeted Masking



We focus edits on fitness-relevant residues, while preserving structurally and functionally important sites

HELMHOLTZ MUNICH

Biologically-constrained Sequential Monte Carlo



Biologically-constrained Sequential Monte Carlo

Constrained proposal

$$\tilde{x}^{(i)}_{\pi(t)} \sim \mathcal{P}_{RAA}(\tilde{x}^{(i)}_{\pi(t)} \mid \tilde{x}^{(i)}_{\pi(<t)})$$

Weighting

$$x^{(i)}_{\text{unroll}} \sim \prod_{s=t+1}^{T} \mathcal{P}_{RAA}(\tilde{x}^{(i)}_{\pi(s)} \mid \tilde{x}^{(i)}_{\pi(<s)})$$

$$\hat{y}^{(i)} = \mu_\theta(x^{(i)}_{\text{unroll}}) + k \cdot \sigma_\theta(x^{(i)}_{\text{unroll}})$$

$$\text{Perp}(x^{(i)}_{\text{unroll}}) = \exp\left(-\frac{1}{|\mathcal{I}|} \sum_{s=T-|\mathcal{I}|+1}^{T} \log \mathcal{P}(x^{(i)}_{\text{unroll}_{\pi(s)}} \mid x^{(i)}_{\text{unroll}_{\pi(<s)}})\right)$$

$$w^{(i)}_t = \hat{y}^{(i)} / \text{Perp}(x^{(i)}_{\text{unroll}})$$

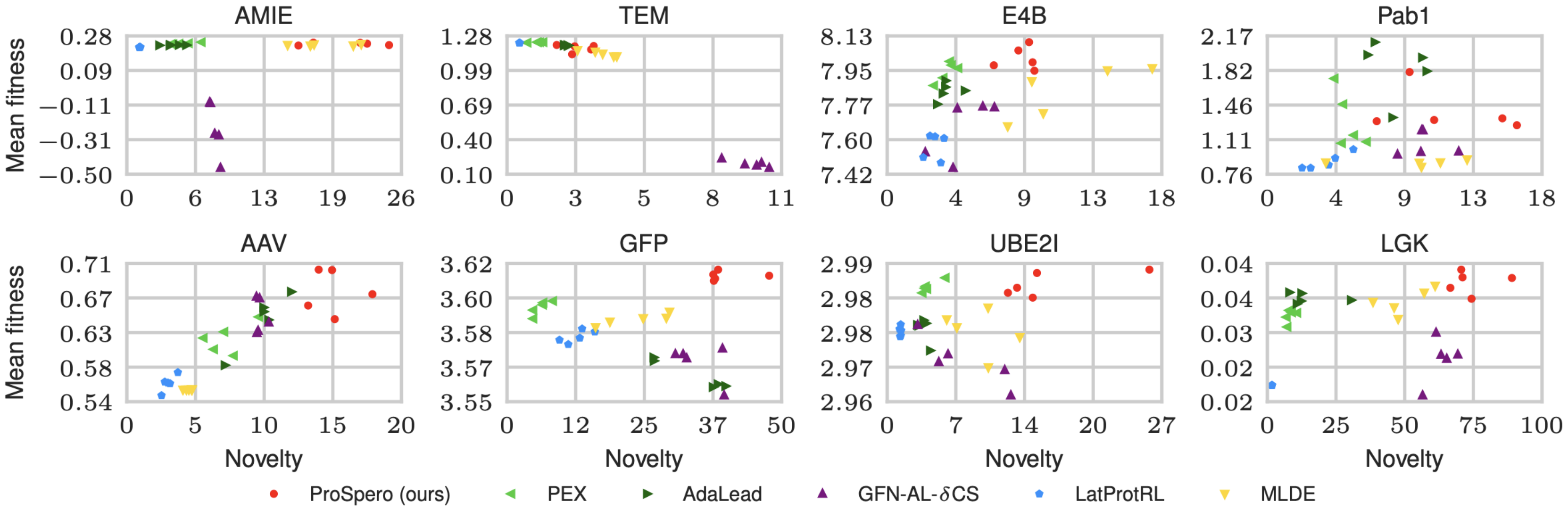
Resampling

$$\tilde{x}^{(i)} \sim \text{Cat}\left(\{\tilde{x}^{(i)}\}^B_{i=1}, \left\{\frac{w^{(i)}_t}{\sum_{j=1}^{B} w^{(j)}_t}\right\}^B_{i=1}\right)$$

Biologically-constrained SMC restricts proposals to residues with similar phys-chem properties to their wild-type counterparts, improving the likelihood of finding high fitness sequences under surrogate misspecification

Results

Breaking the fitness–novelty Pareto front



HELMHOLTZ MUNICH

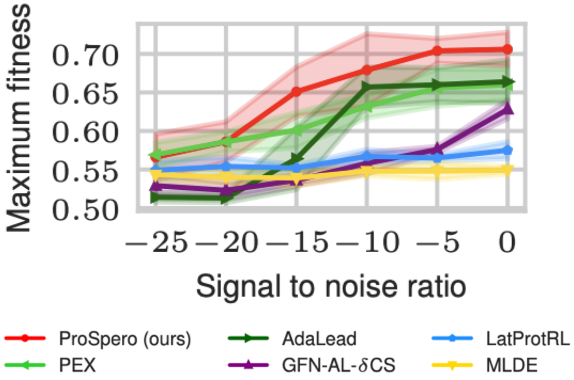
Robustness to surrogate misspecification

Method	Maximum pTM	Mean pTM	Diversity	Novelty
AdaLead	0.796 ± 0.013	0.755 ± 0.011	8.83 ± 2.54	8.36 ± 2.97
PEX	0.807 ± 0.023	$\underline{0.760 \pm 0.012}$	6.14 ± 0.89	4.45 ± 0.38
GFN-AL-δCS	0.791 ± 0.010	0.729 ± 0.005	$\mathbf{16.92 \pm 0.88}$	9.56 ± 0.60
MLDE	$\underline{0.810 \pm 0.020}$	0.752 ± 0.004	9.89 ± 1.11	$\mathbf{20.88 \pm 2.98}$
LatProtRL	0.787 ± 0.013	0.743 ± 0.003	6.32 ± 0.32	5.90 ± 0.53
ProSpero	$\mathbf{0.822 \pm 0.027}$	$\mathbf{0.777 \pm 0.020}$	$\underline{11.50 \pm 1.62}$	$\underline{17.74 \pm 3.20}$

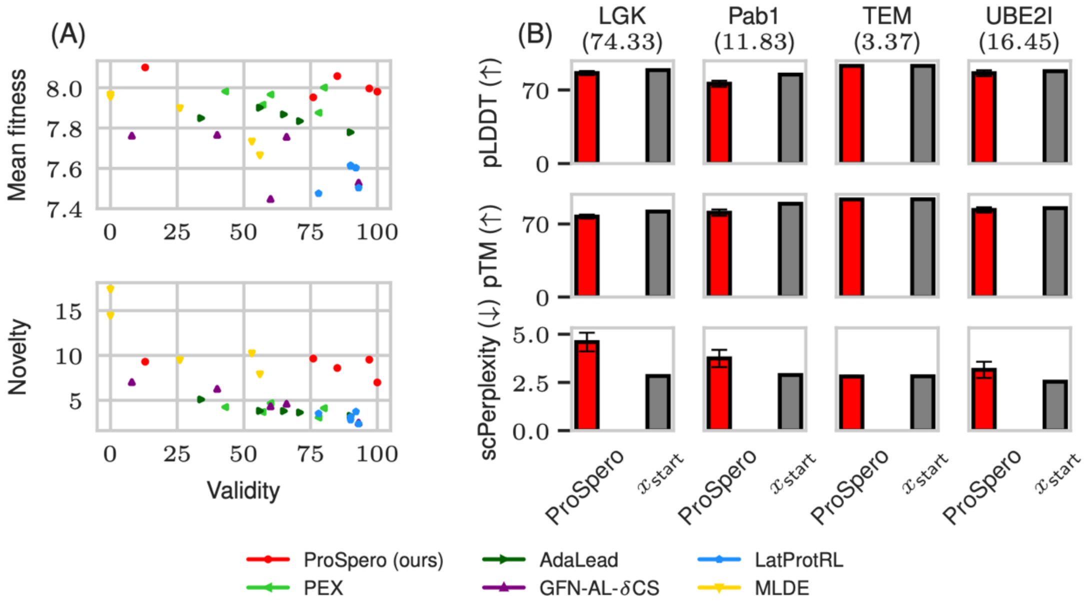
Starting sequence different from the wild-type by 35 residues

Method	Maximum pTM	Mean pTM	Diversity	Novelty
AdaLead	0.593 ± 0.028	0.526 ± 0.007	14.26 ± 1.91	7.66 ± 1.08
PEX	0.578 ± 0.014	0.518 ± 0.003	3.40 ± 0.07	1.72 ± 0.04
GFN-AL-δCS	0.630 ± 0.024	0.542 ± 0.006	$\mathbf{24.13 \pm 1.47}$	14.63 ± 1.16
MLDE	$\underline{0.652 \pm 0.059}$	$\underline{0.572 \pm 0.035}$	13.10 ± 1.18	$\underline{21.68 \pm 3.85}$
LatProtRL	0.560 ± 0.000	0.508 ± 0.003	2.24 ± 0.14	1.78 ± 0.16
ProSpero	$\mathbf{0.672 \pm 0.031}$	$\mathbf{0.599 \pm 0.014}$	$\underline{14.51 \pm 1.99}$	$\mathbf{22.03 \pm 1.69}$

Starting sequence different from the wild-type by 75 residues



Biologically plausible sequences



Conclusion

Conclusion

ProSpero facilitates protein design beyond wild-type neighborhoods by incorporating biological priors through:

- Inference-time guidance of a pre-trained pLM with a surrogate updated in an active learning loop
- Targeted masking of fitness-relevant residues while preserving key structural sites
- Biologically-constrained SMC sampling that restricts proposals to wild-type-like residues

We demonstrated robustness of ProSpero across diverse *in silico* protein engineering tasks

Thank you!

Ewa Szczurek

Vincent Fortuin