Enhancing active learning in protein design with self-supervised representations

Anonymous submission

Abstract

We study sample-efficient protein fitness optimization under a fixed evaluation budget K. Let $x \in \mathcal{X}$ be a sequence, $f(x)$ its unknown fitness, and $\phi(x) \in \mathbb{R}^d$ a pretrained embedding. We examine whether there exists a representation $\psi(x)$ such that $f(x) \approx w^\top \psi(x)$ can be estimated from $n \ll d$ samples and effectively guide search.

We revisit the ProSpero pipeline and show that ridge regression matches or exceeds neural surrogates, implying that fitness is often well-approximated by low-complexity linear models. However, replacing one-hot encodings with pretrained embeddings does not improve performance at high n, and degrades it at low n, indicating a mismatch between representation geometry and the target function.

To investigate the linear accessibility of latent variables, we evaluate linear probes on sparse autoencoder (SAE) features and observe gains in extreme low-data regimes on specific tasks, suggesting that relevant latent variables are linearly accessible only after partial disentanglement. Finally, epistasis analysis reveals that some fitness landscapes are strongly additive, explaining why simple sequence-local models are near-optimal.

Overall, representation learning improves optimization only when it yields a basis aligned with the functional decomposition of f.

Problem Formulation We consider the estimation of $f : \mathcal{X} \to \mathbb{R}$ under a sample constraint $n \ll |\mathcal{X}|$. Given a representation $\psi(x)$, we fit a linear surrogate $\hat{f}(x) = w^\top \psi(x)$ from n labeled samples and use it to guide sequence optimization. The objective is to identify representations ψ that minimize generalization error at fixed n and maximize achieved fitness under a fixed search budget K. Under the superposition hypothesis, pretrained embeddings encode latent factors as linear directions, but these may not be directly usable without disentanglement.

Linear Surrogates Match Neural Networks in the ProSpero benchmark We replace the neural surrogate with ridge regression over sequence features. Across tasks, linear models match or outperform neural baselines, particularly at low n. This demonstrates that the effective complexity of f is low and that linear estimation suffices to guide search, reducing variance and improving computational efficiency.

Pretrained Embeddings Do Not Outperform One-Hot Encodings in the Low-Data Regime We compare $\psi(x)$ given by one-hot encodings and flattened ESM-2 embeddings. Ridge regression on embeddings does not improve and often degrades performance at low n. This indicates that pretrained representations do not provide a basis aligned with f, despite encoding rich sequence statistics.

Disentangled Representations Enable Linear Readout in Extreme Low-Data Regimes We evaluate linear probes across datasets with $n \in \{8, \ldots, 512\}$ (5 splits), using test Spearman correlation. SAE features yield strong gains at $n = 8$ on specific tasks (e.g., LGK). This shows thùat latent variables relevant to f are linearly accessible after partial disentanglement, but obscured in raw embeddings due to superposition.

Epistasis Analysis Reveals Task-Dependent Additivity of Fitness Landscapes We quantify epistasis via deviations from additive effects of mutations. Certain tasks (e.g., AAV) are near-additive, implying $f(x)$ is well-approximated by a linear model in the one-hot basis. Other tasks exhibit non-additive structure, where performance depends on whether $\psi(x)$ aligns with the underlying factorization. This might explain when representations help.