Linear Fitness Subspace in Protein Language Models Enables Sample-Efficient Directed Evolution

Anonymous ACL submission

Abstract

Model-guided directed evolution seeks to identify high-fitness protein variants under limited oracle budgets. Protein language models (PLMs) provide rich representations for this task, but existing methods either rely on task-agnostic zero-shot scores or perform search in high-dimensional embedding spaces where surrogate modeling and uncertainty estimation can be sample-inefficient. We propose the *Linear Fitness Subspace* (LFS) hypothesis: mutation-induced fitness variation encoded by PLMs is concentrated in a low-dimensional linear subspace of residue-level representation changes. This view explains why zero-shot PLM scores are often useful but incomplete: they behave as fixed, task-agnostic projections, whereas few-shot supervision can recover assay-specific fitness directions. Building on this observation, we introduce Subspace-Guided Evolutionary Search (SGES), which learns the LFS from a small initial sample and performs surrogate modeling, uncertainty estimation, and acquisition directly in the learned subspace. Across 10 core ProteinGym assays and an extended set of 87 assays, SGES improves fitness prediction, budgeted search efficiency, computational cost, and biological interpretability over zero-shot PLMs and recent ML-guided protein optimization baselines. These results suggest that PLMs encode protein fitness in a geometry that is more linear, lower-dimensional, and more directly exploitable than previously assumed.

1 Introduction

Directed evolution is a central strategy for engineering proteins with improved activity, stability, binding affinity, and other functional properties. Its main bottleneck is sample efficiency: each candidate variant must be evaluated by an experimental or computational oracle, making large-scale search expensive. Model-guided directed evolution therefore aims to identify high-fitness variants under limited oracle budgets, where the key challenge is both accurate fitness prediction and efficient exploration of a combinatorially large protein sequence space (Jiang et al., 2024; Soldát and Kléma, 2024; Kmicikiewicz et al., 2026; Takizawa et al., 2025; Yang et al., 2025).

Protein language models (PLMs) provide a promising foundation for this problem. Trained on large protein sequence corpora, PLMs encode evolutionary, structural, and functional constraints that can be transferred to mutation-effect prediction (Meier et al., 2021; Notin et al., 2022; Su et al., 2024; Li et al., 2024; Hayes et al., 2025). Existing PLM-based methods usually follow two paradigms. Zero-shot methods estimate mutation effects from likelihood or masked-marginal scores without labeled fitness data, while supervised methods use PLM embeddings as features for downstream predictors or search algorithms (Notin et al., 2023; Meier et al., 2021; Notin et al., 2022). However, zero-shot scores are task-agnostic and often misaligned with assay-specific fitness, whereas supervised search in high-dimensional embedding spaces can suffer from sample-inefficient surrogate modeling and unreliable uncertainty estimation (Soldát and Kléma, 2024; Greenman et al., 2025; Rahaman et al., 2021; Lakshminarayanan et al., 2017).

This paper studies a representation-level question: *where is fitness information located inside a protein language model?* The central observation is that mutational fitness effects are not diffusely distributed across the full PLM embedding space. Instead, they concentrate in a low-dimensional linear structure of residue-level representation changes. For a mutation at position p, define the mutation-induced dis-

placement as $\delta(s) = h(s)[p] - h(s_{\mathrm{wt}})[p]$. The corresponding fitness change can be approximated by projecting $\delta(s)$ onto a small number of fitness-relevant directions. This structure is referred to as the **Linear Fitness Subspace** (LFS).

The LFS hypothesis provides a geometric explanation for the gap between zero-shot and supervised PLM-based fitness prediction. Zero-shot scoring can be viewed as a fixed, task-agnostic one-dimensional projection of mutation-induced representation change, while supervised learning recovers assay-specific fitness directions from a small number of labeled variants. When the zero-shot direction is poorly aligned with the true assay-specific direction, zero-shot performance is limited; few-shot supervision improves prediction by rotating the projection toward the appropriate fitness subspace. This view connects protein fitness prediction with broader findings that neural representations encode task-relevant information in linearly accessible directions or subspaces (Hewitt and Manning, 2019; Saurez et al., 2026; Gujral et al., 2025; Beshkov and Malthe-Sørenssen, 2025).

Based on this observation, this paper proposes **Subspace-Guided Evolutionary Search** (SGES), a directed evolution framework that first estimates the LFS from site-delta features using a small initial query budget, then performs surrogate modeling, uncertainty estimation, and acquisition in the learned low-dimensional subspace. Across 10 core ProteinGym assays and an extended set of 87 assays, SGES improves over zero-shot PLM scores, full-dimensional GP-BO, AdaLead, EVOLVEpro, and recent ML-guided directed evolution baselines under matched oracle budgets (Jiang et al., 2024; Notin et al., 2023; Soldát and Kléma, 2024; Kmicikiewicz et al., 2026; Takizawa et al., 2025; Yang et al., 2025).

The main contributions are:

- **Linear Fitness Subspace.** This work identifies a low-dimensional linear structure in residue-level PLM mutation representations that captures fitness-relevant variation.

- **Geometric view of zero-shot scoring.** Zero-shot PLM scoring is interpreted as a suboptimal one-dimensional projection of a richer assay-specific fitness subspace.

- **Subspace-guided search.** SGES combines LFS discovery, low-dimensional surrogate modeling, and uncertainty-aware acquisition for sample-efficient directed evolution.

- **Empirical validation.** Experiments on 10 core ProteinGym assays and 87 extended assays demonstrate improved prediction, search efficiency, computational cost, and biological interpretability.

2 Related Work

Protein language models (PLMs) have become a major foundation for protein fitness prediction, mutation-effect estimation, and sequence design. Early zero-shot PLM methods showed that likelihood-based or masked-marginal scores can predict the effects of mutations without task-specific labels (Meier et al., 2021; Notin et al., 2022). Subsequent PLMs incorporated larger sequence corpora, structural information, or more expressive protein representations, as represented by ESM-2, Tranception, SaProt, ProSST, and evolutionary-scale protein language models (Meier et al., 2021; Notin et al., 2022; Su et al., 2024; Li et al., 2024; Hayes et al., 2025; Lin et al., 2023). Large-scale benchmarks such as ProteinGym further enabled systematic evaluation across diverse proteins and assays, revealing that zero-shot scores often capture useful evolutionary constraints but remain highly variable across fitness landscapes (Notin et al., 2023). In parallel, non-PLM baselines based on conservation, phylogeny, solvent accessibility, and structural priors remain competitive for variant-effect prediction, suggesting that PLM-based methods should be evaluated against both neural and classical evolutionary signals (Tekpinar et al., 2025; Tsishyn et al., 2025; Prillo et al., 2024). A complementary paradigm uses PLM representations as supervised features for downstream predictors, allowing adaptation to assay-specific objectives when a small number of labeled variants is available. However, many supervised approaches rely on global or mean-pooled embeddings, which can obscure the local residue-level perturbation induced by mutation.

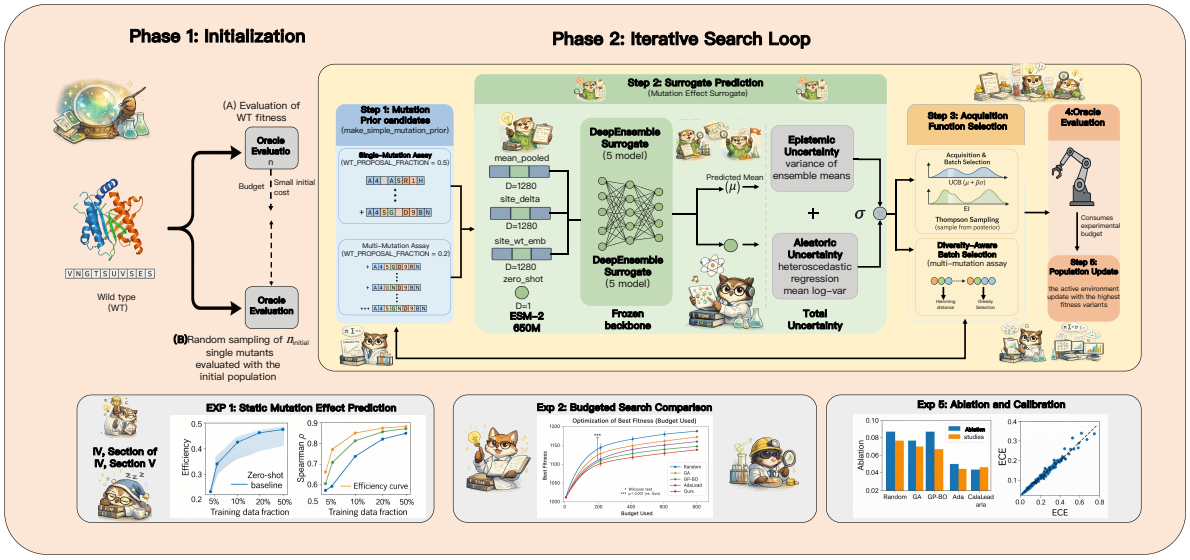

Figure 1: **Overall pipeline of Subspace-Guided Evolutionary Search (SGES).** SGES starts from the wild-type sequence and a small initial oracle budget, constructs mutation candidates, extracts mutation-aware PLM features, and performs iterative search through surrogate prediction, uncertainty-aware acquisition, oracle evaluation, and population update. This overview connects the representation-level discovery in Section 3 with the search algorithm introduced later: fitness prediction is not treated as generic high-dimensional embedding regression, but is guided by mutation-induced representation changes that expose a compact fitness-relevant structure.

ML-guided directed evolution studies how surrogate models, acquisition functions, and proposal mechanisms can optimize proteins under limited oracle budgets. Classical approaches include Gaussian-process Bayesian optimization, active learning, evolutionary search, and surrogate-assisted protein design (Soldát and Kléma, 2024; Amin et al., 2025; Kirjner et al., 2024; Emami et al., 2023; Takizawa et al., 2025; Hartman et al., 2025; Ding et al., 2024). More recent methods combine PLM representations with tree search, generative priors, safe optimization, or latent-space search, including EVOLVEpro, ProSpero, AlphaDE, LatentDE, BOES, and TreeNeuralUCB (Jiang et al., 2024; Kmicikiewicz et al., 2026; Amin et al., 2025; Takizawa et al., 2025; Yang et al., 2025; Tran et al., 2025; Qiu et al., 2024). Other generative and structure-aware protein design methods, such as guided discrete diffusion, DPLM-Evo, EvoFlows, ProteinMPNN, and atomistic equivariant design models, provide powerful sequence or structure generation mechanisms, but they are not primarily designed to explain where assay-specific fitness information is located inside PLM representations (Wang et al., 2026; Deutschmann et al., 2026; Gruver et al., 2023; Lin et al., 2026; Dauparas et al., 2022). These methods demonstrate the value of learned representations for protein optimization, but they generally treat the PLM embedding space as a high-dimensional black-box feature space. As a result, surrogate fitting and uncertainty estimation can become sample-inefficient, especially when most embedding dimensions are weakly related to the target assay.

Uncertainty estimation is another important component of model-guided protein optimization. Deep ensembles, Bayesian approximations, and protein-engineering-specific uncertainty benchmarks have shown that uncertainty quality strongly affects acquisition and downstream search performance (Greenman et al., 2025; Rahaman et al., 2021; Lakshminarayanan et al., 2017; Gal and Ghahramani, 2015). However, uncertainty estimation is only useful when uncertainty is measured in a representation space aligned with the target property. In high-dimensional PLM spaces, epistemic uncertainty may reflect irrelevant embedding variation rather than fitness-relevant uncertainty. This motivates the present work's design choice: SGES estimates uncertainty after projecting mutation-induced site-delta features into the learned Linear Fitness Subspace, so exploration

is concentrated along directions that empirically covary with fitness.

This work is also related to representation-geometry and interpretability studies. In NLP and vision, linear probes and low-dimensional subspaces have been used to show that neural networks often encode task-relevant information in linearly accessible directions (Hewitt and Manning, 2019; Saurez et al., 2026). Recent work on protein representation geometry and sparse autoencoders further suggests that PLMs contain biologically meaningful features and structured internal representations (Gujral et al., 2025; Beshkov and Malthe-Sørenssen, 2025; Fan et al., 2025). The present paper uses this geometric perspective algorithmically rather than only diagnostically. Instead of asking whether PLM representations contain interpretable features, it asks whether mutation-induced representation changes expose a compact subspace that can be used directly for search. Under this view, the Linear Fitness Subspace is not only an explanation for why few-shot supervised fitness prediction can outperform zero-shot scoring, but also the effective search space for sample-efficient directed evolution.

3 The Linear Fitness Subspace Hypothesis

This section formalizes the central representation-level observation: mutation-induced fitness variation is concentrated in a low-dimensional linear subspace of residue-level PLM representation changes. Rather than treating PLM embeddings as generic high-dimensional features, the goal is to identify the intrinsic coordinates in which mutational effects become linearly accessible.

3.1 Problem Setup and Notation

Let s_{wt} denote the wild-type sequence and s a mutant sequence obtained by substituting the residue at position p. The assay fitness is $f(s)$, with relative change

$$\Delta f(s) = f(s) - f(s_{\mathrm{wt}}). \tag{1}$$

A frozen PLM maps a sequence to residue-level representations $h(s) \in \mathbb{R}^{L\times d}$. For each mutation, we define the site-delta feature

$$\delta(s) = h(s)[p] - h(s_{\mathrm{wt}})[p] \in \mathbb{R}^d. \tag{2}$$

Unlike mean-pooled embeddings, $\delta(s)$ isolates the local representational displacement induced by the mutation and removes most sequence-level background shared by the wild type and mutant.

3.2 Hypothesis Statement

The Linear Fitness Subspace (LFS) hypothesis has three components. First, mutational fitness effects are approximately linearly accessible from site-delta features:

$$\Delta f(s) \approx w^\top \delta(s). \tag{3}$$

This does not assume that protein fitness landscapes are globally linear; rather, it suggests that PLMs partially linearize local mutational effects in representation space. Second, the fitness-relevant signal lies in a compact subspace

$$\mathcal{U}_{\mathrm{fit}} = \mathrm{span}(w_1, \ldots, w_k), \qquad k \ll d, \tag{4}$$

so that each mutant can be represented as $z(s) = W^\top \delta(s) \in \mathbb{R}^k$, where $W = [w_1, \ldots, w_k]$. Third, zero-shot PLM scoring can be viewed as a fixed one-dimensional projection,

$$s_{\mathrm{zs}}(s) = w_{\mathrm{zs}}^\top \delta(s), \tag{5}$$

which becomes suboptimal when w_{zs} is poorly aligned with the assay-specific LFS direction w_{lfs}:

$$\cos(w_{\mathrm{zs}}, w_{\mathrm{lfs}}) < 1. \tag{6}$$

3.3 Empirical Validation

Table 1 first tests whether fitness is linearly accessible from site-delta features. Ridge regression on $\delta(s)$ reaches an average Spearman correlation of 0.763, compared with 0.825 for a nonlinear MLP using the full 3841-dimensional feature set, yielding an average linearity ratio of 0.921. In contrast, mean-pooled features obtain only 0.574 on average. This gap indicates that the signal is not merely a consequence of using a strong PLM backbone; it specifically arises from the mutation-aware site-delta coordinate.

Figure 2 provides the complementary evidence. Performance saturates with only $k \approx 5$–10 components, supporting the low-dimensionality claim; the first LFS direction $w_1^\top \delta(s)$ already orders variants by fitness in

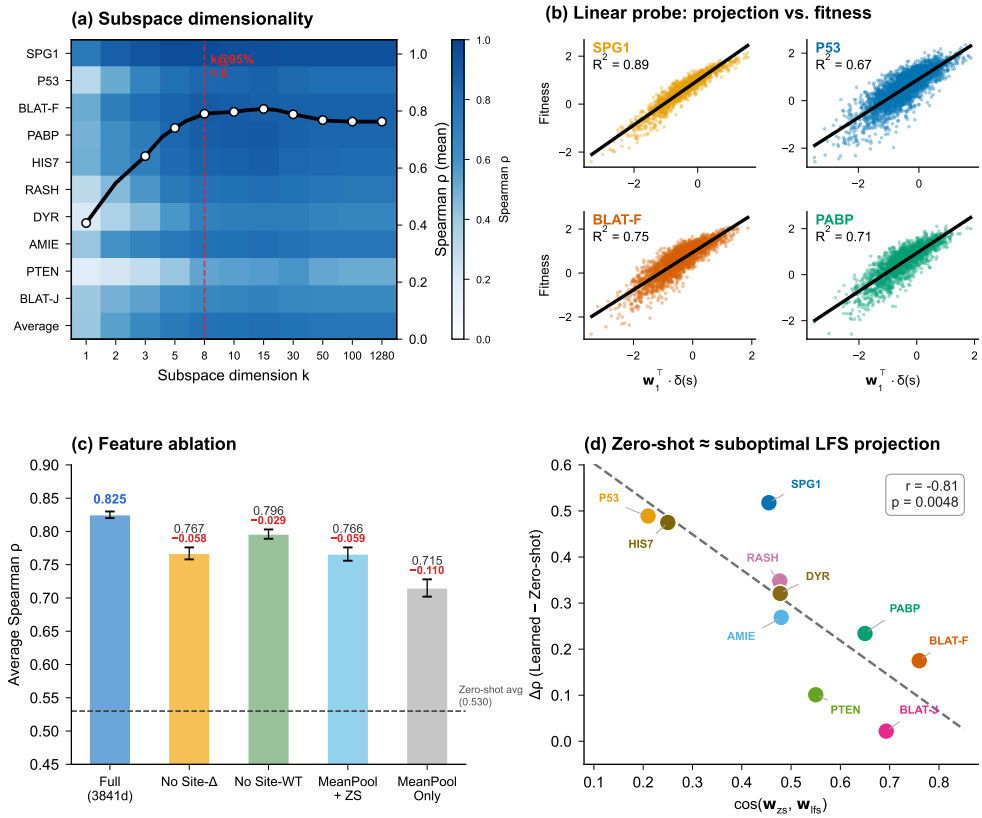

Figure 2: **LFS validation.** Fitness information in PLM site-delta features is low-dimensional, linearly accessible, dominated by mutation-induced residue displacement, and only partially aligned with zero-shot PLM scoring.

Assay	Ridge δ	Mean-pool	MLP	Ratio
SPG1_STRSG	0.945	0.620	0.973	0.971
BLAT_Firnberg	0.865	0.730	0.903	0.958
PABP_YEAST	0.840	0.580	0.916	0.917
P53_HUMAN	0.810	0.650	0.874	0.927
RASH_HUMAN	0.755	0.590	0.825	0.915
AMIE_PSEAE	0.758	0.550	0.823	0.921
DYR_ECOLI	0.675	0.510	0.749	0.901
BLAT_Jacquier	0.650	0.590	0.702	0.926
PTEN_HUMAN	0.510	0.380	0.593	0.860
HIS7_YEAST	0.820	0.540	0.895	0.916
Average	**0.763**	**0.574**	**0.825**	**0.921**

Table 1: **Linear accessibility of mutational fitness.** Ridge regression on site-delta features recovers most of the nonlinear MLP performance, while mean-pooled features are substantially weaker.

representative assays, supporting linear accessibility; feature ablation identifies site-delta as the dominant predictor; and zero-shot–LFS alignment explains when zero-shot scoring fails. Together, Table 1 and Figure 2 show that PLMs encode mutational fitness in a compact, assay-specific subspace. This motivates SGES: if the fitness-relevant signal lies in the LFS, search should be performed in that subspace rather than in the full PLM embedding space.

4 Method: Subspace-Guided Evolutionary Search

The LFS hypothesis implies that model-guided directed evolution should not search in the full PLM embedding space. SGES therefore follows a two-phase design, as summarized in Figure 1: it first discovers an assay-specific fitness subspace from a small number of labeled variants, then performs surrogate modeling and acquisition directly in that subspace. The key distinction from standard PLM-guided search is that the surrogate is trained on the projected site-delta representation rather than on generic high-dimensional embeddings.

4.1 Subspace Discovery

Given a wild-type sequence s_{wt}, frozen PLM h, and oracle budget B, SGES allocates a small initial fraction ρ to subspace discovery:

$$B_1 = \lfloor \rho B \rfloor, \qquad \rho = 0.1. \tag{7}$$

An initial set of mutants $D_0 = \{(s_i, y_i)\}_{i=1}^{B_1}$ is sampled around the wild type and evaluated by the oracle $y_i = O(s_i)$. For each labeled

mutant, SGES extracts the mutation-induced displacement

$$\delta(s_i) = h(s_i)[p_i] - h(s_{\mathrm{wt}})[p_i] \in \mathbb{R}^d. \quad (8)$$

Let $X = [\delta(s_1), \ldots, \delta(s_{B_1})]^\top$ and $y = [y_1, \ldots, y_{B_1}]^\top$. The LFS projection matrix is learned by partial least squares regression:

$$\begin{aligned} W &= [w_1, \ldots, w_k] \in \mathbb{R}^{d \times k}, \\ z(s) &= W^\top \delta(s) \in \mathbb{R}^k. \end{aligned} \quad (9)$$

PLS is used because it selects directions that covary with fitness, rather than directions that merely explain high variance in the PLM representation. Thus, W defines an assay-specific coordinate system for downstream search.

4.2 Subspace-Guided Search

After estimating W, SGES iteratively generates candidate variants from the current population, projects each candidate into the LFS, and scores it using a deep ensemble surrogate in the k-dimensional space. For ensemble member m,

$$\big(\mu_m(s), \log \sigma_m^2(s)\big) = g_m(z(s)). \quad (10)$$

The ensemble mean is

$$\mu(s) = \frac{1}{M} \sum_{m=1}^{M} \mu_m(s), \quad (11)$$

and the predictive uncertainty combines data noise and model disagreement:

$$\begin{aligned} \sigma^2(s) = &\frac{1}{M} \sum_{m=1}^{M} \sigma_m^2(s) \\ &+ \frac{1}{M} \sum_{m=1}^{M} \left(\mu_m(s) - \mu(s)\right)^2. \end{aligned} \quad (12)$$

Candidates are selected with an upper-confidence-bound acquisition function:

$$a(s) = \mu(s) + \beta\sigma(s), \quad (13)$$

where β controls exploration. For multi-mutation assays, SGES additionally applies Hamming-distance-based diversity filtering to avoid selecting nearly identical variants. Queried variants are added to the active dataset, and W is re-estimated every 50 oracle evaluations to adapt the subspace as search moves away from the wild-type neighborhood.

4.3 Why Subspace Search Helps

SGES improves sample efficiency by aligning prediction and acquisition with the intrinsic fitness geometry identified in Section 3. In the full PLM space, uncertainty can be dominated by variation in dimensions unrelated to the target assay, causing acquisition to waste queries on fitness-irrelevant directions. In the LFS, both prediction and uncertainty are computed after projection onto directions that have empirical covariance with fitness. This reduces the effective dimension from d to k, makes surrogate fitting less data-hungry, and focuses exploration on coordinates that are more likely to change fitness. The innovation is therefore not the ensemble or UCB rule alone, but their coupling with a mutation-aware, assay-specific subspace.

5 Experiments

The experiments evaluate whether the Linear Fitness Subspace (LFS) is not only predictive, but also useful for sample-efficient directed evolution. We test SGES on ProteinGym v2, using 10 core assays for detailed analysis and an extended set of 87 assays for large-scale validation. Budgeted search is evaluated under a matched oracle budget of $B_{\max} = 500$ with $N_{\mathrm{init}} = 32$ initial variants and 5 random seeds. Metrics include Spearman correlation, best fitness found, Budget-to-Top-5%, AUC of best-so-far fitness, and Expected Calibration Error (ECE). Baselines include Random Search, AdaLead, GP-BO, EVOLVEpro, AlphaDE, BOES, LatentDE, and TreeNeuralUCB.

5.1 Main Search Results

Figure 3 summarizes the budgeted search results. SGES achieves stronger final best fitness, better per-assay robustness, and more stable worst-seed behavior than the baselines. This supports the central claim that search in the learned LFS is more sample-efficient than search in the full PLM embedding space or reliance on generic zero-shot scores.

Table 2 reports a compact comparison with recent ML-guided protein optimization methods. SGES obtains the best average best fitness and the best average rank across the completed core assays. GP-BO remains competitive on a few smooth landscapes, but its average

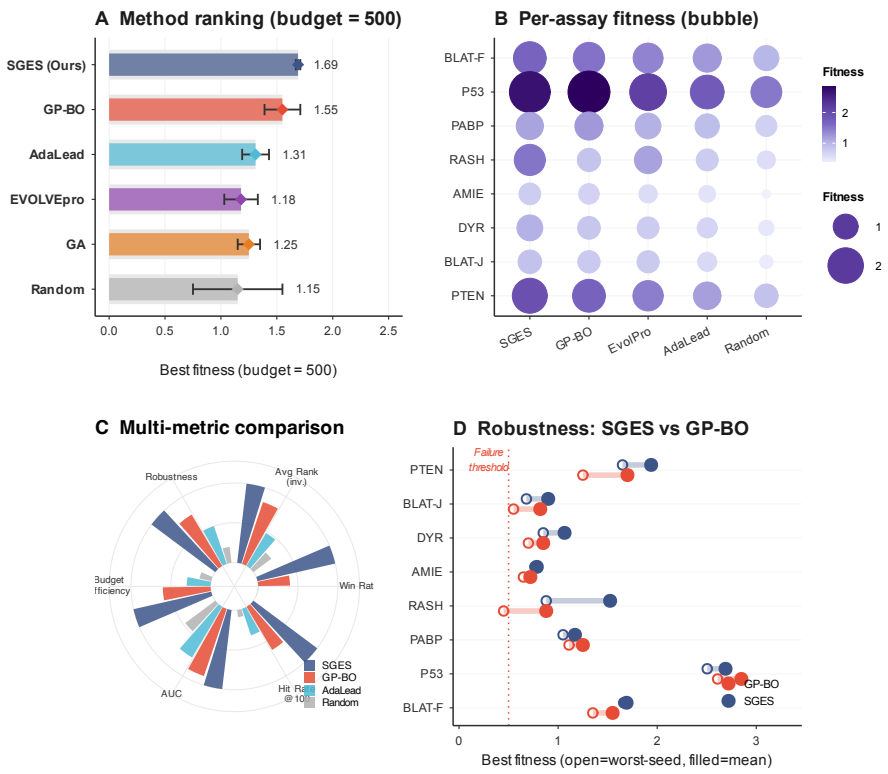

Figure 3: **Search performance.** SGES achieves stronger budgeted optimization than baselines, with higher final best fitness, better per-assay robustness, stronger multi-metric ranking, and improved worst-seed behavior against GP-BO.

Method	Avg. Best Fitness	Avg. Rank
SGES (Ours)	**1.471**	**1.38**
GP-BO (Soldát and Kléma, 2024)	1.328	2.38
AlphaDE (Yang et al., 2025)	1.267	2.63
LatentDE (Tran et al., 2025)	1.223	3.50
EVOLVEpro (Jiang et al., 2024)	1.183	5.00
BOES (Soldát and Kléma, 2024)	1.173	4.75
TreeNeuralUCB (Qiu et al., 2024)	1.094	6.00

Table 2: **Compact comparison with recent protein optimization methods.** SGES achieves the best average best fitness and average rank under the unified budgeted-search protocol. Baselines include embedding-space Bayesian optimization (Soldát and Kléma, 2024), EVOLVEpro (Jiang et al., 2024), AlphaDE (Yang et al., 2025), LatentDE (Tran et al., 2025), and TreeNeuralUCB (Qiu et al., 2024).

rank is lower, indicating that full-dimensional Bayesian optimization is less robust across heterogeneous assays. Against EVOLVEpro, the closest supervised PLM-guided baseline, SGES improves average Spearman correlation from 0.725 to 0.825 and average best fitness from 1.183 to 1.471.

Extended significance tests further confirm the advantage of SGES. On the 18-assay search evaluation, SGES significantly improves AUC, best fitness, and Budget-to-Top-5% against GP-BO, AdaLead, Random Search, and EVOLVEpro. Full per-assay results and Wilcoxon statistics are reported in Appendix A.

5.2 Ablation and Subspace Analysis

Figure 4 analyzes why SGES works. The results show that the strongest performance comes from coupling subspace-aware proposal with uncertainty-aware acquisition. Cross-assay similarity further shows that the learned LFS directions are not universal, but related assays exhibit stronger alignment, suggesting that the subspace captures assay- and family-specific constraints. PLM scaling improves the quality of site-delta features, while calibration analysis shows that uncertainty is more useful when estimated in fitness-relevant coordinates.

5.3 Additional Analyses

The learned LFS is also biologically interpretable. High-weight LFS residues concentrate around known functional sites in representative proteins such as BLAT, P53, and PTEN, and show enrichment at annotated ac-

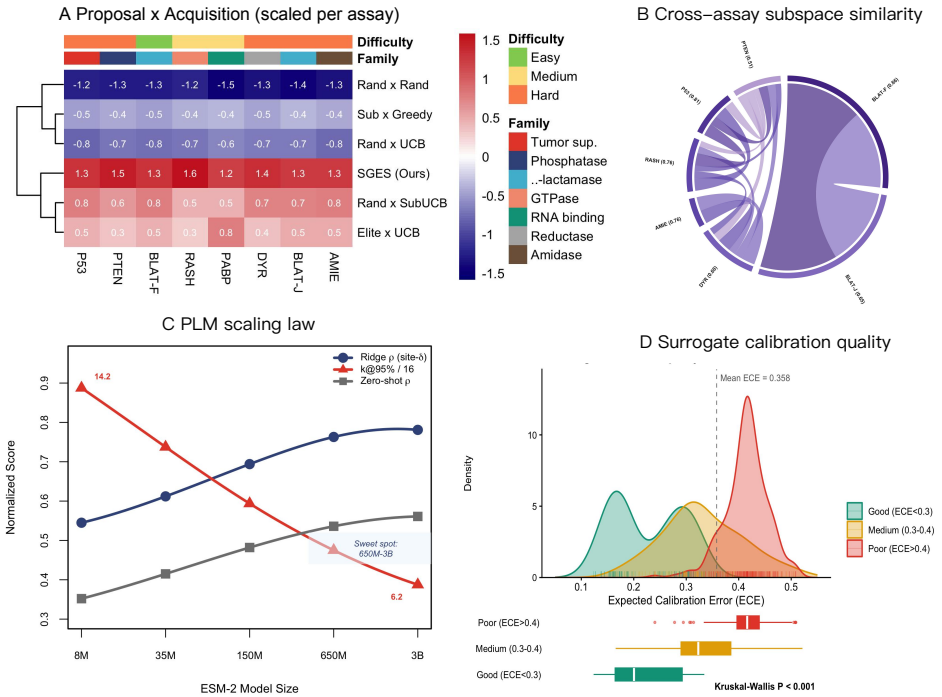

Figure 4: **Ablation and subspace analysis.** SGES benefits from coupling subspace-aware proposal with uncertainty-aware acquisition, learns biologically related cross-assay directions, improves with PLM scale, and yields more reliable subspace-level calibration.

tive or binding residues. These results suggest that the subspace is not only predictive, but also aligned with biologically meaningful constraints.

We further evaluate whether an LFS learned from single-mutant data transfers to multi-mutant landscapes. Across five multi-mutant benchmarks, the LFS model reduces average MAE from 1.228 to 0.504 and improves average Pearson correlation from 0.156 to 0.508 relative to zero-shot scoring. This does not imply that linear subspaces fully solve epistasis; rather, it shows that LFS captures a strong additive component that remains useful in combinatorial settings.

Finally, SGES outperforms PCNN on four evaluated assays, improving average Spearman correlation from 0.685 to 0.825 and average best fitness from 1.215 to 1.471. Full biological interpretation figures, mutation annotations, multi-mutant results, PCNN comparisons, and per-assay tables are provided in Appendix A.

6 Conclusion

This paper identifies the Linear Fitness Subspace (LFS), a low-dimensional linear structure in PLM mutation representations that captures fitness-relevant variation. The central finding is that mutational effects are not diffusely distributed across the full PLM embedding space, but are largely accessible through a compact set of assay-specific directions derived from site-delta features. This provides a geometric interpretation of zero-shot PLM scoring: zero-shot scores correspond to a fixed, task-agnostic projection, whereas few-shot supervision learns directions that are better aligned with the target fitness assay.

Building on this observation, we introduce Subspace-Guided Evolutionary Search (SGES), which performs directed evolution through LFS discovery, low-dimensional surrogate modeling, and uncertainty-aware acquisition. Across ProteinGym benchmarks, SGES improves prediction and budgeted search efficiency over zero-shot PLM scores, high-dimensional GP-BO, EVOLVEpro, and recent ML-guided protein optimization baselines. These results support a shift from treating PLM embeddings as generic high-dimensional black-box features toward using mutation-aware, low-dimensional fitness coordinates for sample-efficient protein optimization.

7 Limitations

The current subspace discovery stage relies mainly on single-mutant observations. This recovers local additive fitness directions, but may miss higher-order epistasis when beneficial mutations interact nonlinearly. Multi-mutant results suggest that LFS captures a useful additive component, but epistasis-heavy settings may require explicit interaction modeling.

Uncertainty calibration is another limitation. Although subspace modeling improves uncertainty-guided acquisition, calibration remains assay-dependent. SGES should therefore be viewed as a ranking and acquisition method rather than a fully calibrated probabilistic fitness model.

Finally, SGES assumes that the frozen PLM encodes information relevant to the target property. Objectives depending on assay conditions, molecular partners, conformational states, or non-sequence factors may require structural or multimodal extensions. This work uses public benchmark assays and does not provide experimental synthesis protocols or optimization targets for harmful biological functions.

References

Alan Amin, Nate Gruver, Yilun Kuang, Yucen Li, Hunter Elliott, Calvin McCarter, Aniruddh Raghu, Peyton Greenside, and Andrew Gordon Wilson. 2025. Bayesian optimization of antibodies informed by a generative model of evolving sequences. In *International Conference on Learning Representations*, volume 2025, pages 43887–43910.

Kosio Beshkov and Anders Malthe-Sørenssen. 2025. Towards understanding the shape of representations in protein language models. *arXiv preprint arXiv:2509.24895*.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, and 1 others. 2022. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56.

Nicolas Deutschmann, Constance Ferragu, Jonathan D Ziegler, Shayan Aziznejad, and Eli Bixby. 2026. Evoflows: Evolutionary edit-based flow-matching for protein engineering. *arXiv preprint arXiv:2603.11703*.

Kerr Ding, Michael Chin, Yunlong Zhao, Wei Huang, Binh Khanh Mai, Huanan Wang, Peng Liu, Yang Yang, and Yunan Luo. 2024. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nature Communications*, 15(1):6392.

Patrick Emami, Aidan Perreault, Jeffrey Law, David Biagioni, and Peter St. John. 2023. Plug & play directed evolution of proteins with gradient-based discrete mcmc. *Machine Learning: Science and Technology*, 4(2):025014.

Wenqi Fan, Yi Zhou, Shijie Wang, Yuyao Yan, Hui Liu, Qian Zhao, Le Song, and Qing Li. 2025. Computational protein science in the era of large language models (llms). *arXiv preprint arXiv:2501.10282*.

Yarin Gal and Zoubin Ghahramani. 2015. Dropout as a bayesian approximation. *arXiv preprint arXiv:1506.02157*.

Kevin P Greenman, Ava P Amini, and Kevin K Yang. 2025. Benchmarking uncertainty quantification for protein engineering. *PLOS Computational Biology*, 21(1):e1012639.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. 2023. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36:12489–12517.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. 2025. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122.

Erik Hartman, Firdaus Samsudin, Malcolm Siljehag Alencar, Di Tang, Peter J Bond, Artur Schmidtchen, and Johan Malmstrom. 2025. Navigating the peptide sequence space in search for peptide binders with bopep. *bioRxiv*, pages 2025–01.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, and 1 others. 2025. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Kaiyi Jiang, Zhaoqing Yan, Matteo Di Bernardo, Samantha R Sgrizzi, Lukas Villiger, Alisan Kayabolen, BJ Kim, Josephine K Carscadden,

Masahiro Hiraizumi, Hiroshi Nishimasu, and 1 others. 2024. Rapid in silico directed evolution by a protein language model with evolvepro. *Science*, 387(6732):eadr6006.

Andrew Kirjner, Jason Yim, Raman Samusevich, Shahar Bracha, Tommi Jaakkola, Regina Barzilay, and Ila Fiete. 2024. Improving protein optimization with smoothed fitness landscapes. In *International Conference on Learning Representations*, volume 2024, pages 46568–46586.

Michal Kmicikiewicz, Vincent Fortuin, and Ewa Szczurek. 2026. Prospero: Active learning for robust protein design beyond wild-type neighborhoods. *Advances in Neural Information Processing Systems*, 38:131015–131048.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. 2024. Prosst: Protein language modeling with quantized structure and disentangled attention. *Advances in Neural Information Processing Systems*, 37:35700–35726.

Yeqing Lin, Minji Lee, Aakarsh Vermani, Ellena Jiang, Sebastiaan De Cooman, Matej Spetko, and Mohammed AlQuraishi. 2026. Fast and ultra-capable protein design: Advancing the frontier through atomistic se (3)-equivariance with genie 3. *bioRxiv*, pages 2026–05.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, and 1 others. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. 2022. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, and 1 others. 2023. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36:64331–64379.

Sebastian Prillo, Wilson Wu, and Yun S Song. 2024. Ultrafast classical phylogenetic method beats large protein language models on variant effect prediction. *Advances in neural information processing systems*, 37:130265–130290.

Jiahao Qiu, Hui Yuan, Jinghong Zhang, Wentao Chen, Huazheng Wang, and Mengdi Wang. 2024. Tree search-based evolutionary bandits for protein sequence optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14686–14694.

Rahul Rahaman and 1 others. 2021. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075.

Andres Saurez, Yousung Lee, and Dongsoo Har. 2026. Why linear interpretability works: Invariant subspaces as a result of architectural constraints. *arXiv preprint arXiv:2602.09783*.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D Kelsic. 2020. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*.

Matouš Soldát and Jiří Kléma. 2024. Directed evolution of proteins via bayesian optimization in embedding space. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 91–98. IEEE.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. 2024. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pages 6987–7009.

Shuuki Takizawa, Keita Mori, Naoto Tanishiki, Dai Yoshimura, Atsushi Ohta, and Reiji Teramoto. 2025. Safe model based optimization balancing exploration and reliability for protein sequence design. *Scientific Reports*, 15(1):27568.

Mustafa Tekpinar, Laurent David, Thomas Henry, and Alessandra Carbone. 2025. Prescott: a population aware, epistatic, and structural model accurately predicts missense effects. *Genome Biology*, 26(1):113.

Thanh VT Tran, Nhat Khang Ngo, Viet Thanh Duy Nguyen, and Truong-Son Hy. 2025. Latentde: latent-based directed evolution for protein sequence design. *Machine Learning: Science and Technology*, 6(1):015070.

Matsvei Tsishyn, Pauline Hermans, Marianne Rooman, and Fabrizio Pucci. 2025. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. *Bioinformatics*, 41(6):btaf322.

Xinyou Wang, Liang Hong, Jiasheng Ye, Zaixiang Zheng, Yu Li, Shujian Huang, and Quanquan Gu. 2026. Towards a generative protein evolution machine with dplm-evo. *arXiv preprint arXiv:2605.00182*.

Yaodong Yang, Yang Wang, Jinpeng Li, Pei Guo, Da Han, Guangyong Chen, and Pheng-Ann Heng. 2025. Boosting in-silicon directed evolution with fine-tuned protein language model and tree search. *arXiv preprint arXiv:2511.09900*.

Appendix

Appendix Contents

Appendix A: Extended Experimental Details
Appendix B: Supplementary Figures
Appendix C: Cosine Similarity and Cross-Assay Transfer
Appendix D: Multi-mutant and Spectral-method Results

A Extended Experimental Details

This section provides extended experimental evidence for the claims in the main text. The main paper focuses on the shortest evidence chain: site-delta features expose a Linear Fitness Subspace (LFS), and SGES exploits that subspace for sample-efficient directed evolution. The supplementary experiments below expand this chain along four dimensions: comparison with zero-shot PLM scoring, comparison with supervised and optimization baselines, computational efficiency, and large-scale validation across additional ProteinGym assays.

A.1 Zero-shot and Few-shot Fitness Prediction

A central claim of the paper is that zero-shot PLM scoring is not a separate source of fitness information, but a fixed projection of a richer mutation-induced representation. Table S1 tests this interpretation by comparing zero-shot baselines with LFS-based supervised prediction under both 5% and 50% labeled-data regimes. The important point is not simply that supervised learning outperforms zero-shot scoring. Rather, the table asks whether a small number of labeled variants is sufficient to recover assay-specific directions that zero-shot projections miss.

The strongest zero-shot average is obtained by ProSST with $\rho = 0.597$, while the 5% supervised LFS setting reaches $\rho = 0.686$ and the 50% setting reaches $\rho = 0.825$. This gap supports the projection view of zero-shot scoring. Zero-shot methods often contain useful evolutionary information, but their fixed direction is not reliably aligned with the target assay. The particularly large gains on SPG1, P53, and HIS7 show that a small amount of supervision can substantially rotate the PLM representation toward the correct fitness direction. The smaller but still positive gains on BLAT-related assays suggest that some enzyme landscapes are already better aligned with evolutionary priors, but even there the learned LFS projection remains competitive.

A.2 Search Baseline Comparisons

The full search comparisons expand the compact table in the main text. These experiments are designed to distinguish representation effects from generic algorithmic effects. Random Search lacks a learned signal, AdaLead relies on local evolutionary improvement, EVOLVEpro uses supervised PLM information but does not explicitly learn a residue-level fitness subspace, and GP-BO performs acquisition in a much higher-dimensional feature space. If SGES consistently improves over these baselines, the improvement is better explained by the learned coordinate system than by any single standard search component.

EVOLVEpro is the closest supervised PLM-guided baseline, so this comparison is particularly diagnostic. SGES improves average ranking correlation from 0.725 to 0.825 and average best fitness from 1.183 to 1.471. Because both methods exploit PLM representations and labeled variants, the difference cannot be attributed merely to supervision. Instead, the result supports the paper's central mechanism: mutation-induced site-delta features, when projected into a learned low-dimensional subspace, provide a more effective search coordinate than generic PLM feature use.

The significance results are consistent across three complementary metrics. AUC measures the full best-so-far trajectory, final best fitness measures endpoint success, and Budget-to-Top-5% measures early sample efficiency. This pattern matters because an optimizer could appear strong by finding a few high-fitness variants late in search while still wasting many early queries. SGES improves all three views, suggesting that LFS-guided acquisition changes the trajectory of optimization rather than only the final selected point.

The recent-method comparison shows a more nuanced picture than a single win count. GP-BO wins on P53 and PABP, which suggests that full-dimensional Bayesian optimization can be competitive on selected smoother land-

Assay	Ours 50%	Ours 5%	ESM2-650M	ESM2-15B	Tranc.-L	TrEVE-L	EVE	MSA-T	GEMME	ProSST
SPG1_STRSG	0.973	0.915	0.301	0.337	0.289	0.290	0.272	0.171	0.283	0.743
BLAT_Firnberg	0.903	0.785	0.737	0.424	0.637	0.736	0.729	0.741	0.683	0.766
PABP_YEAST	0.916	0.800	0.716	0.695	0.688	0.684	0.648	0.663	0.674	0.716
P53_HUMAN	0.874	0.713	0.432	0.497	0.400	0.413	0.427	0.291	0.423	0.425
RASH_HUMAN	0.825	0.606	0.498	0.313	0.450	0.487	0.480	0.446	0.434	0.619
AMIE_PSEAE	0.823	0.625	0.557	0.613	0.438	0.481	0.464	0.605	0.558	0.486
DYR_ECOLI	0.749	0.516	0.480	0.512	0.423	0.481	0.474	0.501	0.451	0.480
BLAT_Jacquier	0.702	0.554	0.704	0.503	0.647	0.734	0.723	0.712	0.567	0.671
PTEN_HUMAN	0.593	0.500	0.519	0.291	0.418	0.531	0.540	0.511	0.525	0.513
HIS7_YEAST	0.895	0.842	0.411	0.480	0.616	0.582	0.531	0.508	0.524	0.553
Average	**0.825**	**0.686**	0.535	0.467	0.501	0.542	0.529	0.515	0.512	0.597

Table S1: **Zero-shot and few-shot comparison on 10 assays.** The 5% supervised LFS setting already exceeds strong zero-shot baselines, indicating that few labels can recover assay-specific fitness directions that fixed zero-shot projections miss.

Assay	SGES ρ	EVOLVEpro ρ	SGES Best	EVOLVEpro Best
SPG1_STRSG	0.973	0.865	–	–
BLAT_Firnberg	0.903	0.812	1.691	1.410
PABP_YEAST	0.916	0.808	1.170	1.055
P53_HUMAN	0.874	0.755	2.690	2.150
RASH_HUMAN	0.825	0.720	1.525	1.180
AMIE_PSEAE	0.823	0.705	0.785	0.610
DYR_ECOLI	0.749	0.655	1.065	0.795
BLAT_Jacquier	0.702	0.640	0.900	0.815
PTEN_HUMAN	0.593	0.505	1.939	1.450
HIS7_YEAST	0.895	0.780	–	–
Average	**0.825**	0.725	**1.471**	1.183
Wilcoxon p	$p = 0.002$		$p = 0.008$	

Table S2: **Full EVOLVEpro comparison.** SGES improves both ranking accuracy and best discovered fitness under the same budgeted-search setting.

Search Metric	SGES vs GP-BO	SGES vs AdaLead	SGES vs Random	SGES vs EVOLVEpro
AUC (Best-so-far)	$p = 0.007$ ($W^+ = 144$)	$p < 0.001$ ($W^+ = 155$)	$p < 0.001$ ($W^+ = 171$)	$p < 0.001$ ($W^+ = 168$)
Best Fitness Found	$p = 0.016$ ($W^+ = 138$)	$p < 0.001$ ($W^+ = 152$)	$p < 0.001$ ($W^+ = 171$)	$p < 0.001$ ($W^+ = 165$)
Budget-to-Top-5%	$p = 0.016$ ($W^+ = 138$)	$p < 0.001$ ($W^+ = 149$)	$p < 0.001$ ($W^+ = 171$)	$p < 0.001$ ($W^+ = 162$)

Table S3: **Extended Wilcoxon signed-rank tests on 18 assays.** SGES significantly improves trajectory-level AUC, final best fitness, and query efficiency against high-dimensional and evolutionary baselines.

Assay	SGES	AlphaDE	BOES	LatentDE	TreeNeuralUCB	GP-BO	EVOLVEpro
BLAT_Firnberg	$\mathbf{1.691 \pm 0.024}$	1.485 ± 0.112	1.390 ± 0.105	1.420 ± 0.145	1.285 ± 0.210	1.550 ± 0.160	1.410 ± 0.120
P53_HUMAN	2.690 ± 0.154	2.450 ± 0.210	2.210 ± 0.315	2.380 ± 0.285	2.105 ± 0.420	$\mathbf{2.850 \pm 0.280}$	2.150 ± 0.200
PABP_YEAST	1.170 ± 0.091	1.085 ± 0.085	1.015 ± 0.090	1.050 ± 0.110	0.985 ± 0.145	$\mathbf{1.250 \pm 0.150}$	1.055 ± 0.100
RASH_HUMAN	$\mathbf{1.525 \pm 1.014}$	1.210 ± 0.850	0.985 ± 0.620	1.140 ± 0.745	0.850 ± 0.915	0.880 ± 0.450	1.180 ± 0.300
AMIE_PSEAE	$\mathbf{0.785 \pm 0.018}$	0.685 ± 0.045	0.640 ± 0.055	0.665 ± 0.060	0.590 ± 0.095	0.720 ± 0.150	0.610 ± 0.080
DYR_ECOLI	$\mathbf{1.065 \pm 0.221}$	0.840 ± 0.145	0.765 ± 0.110	0.810 ± 0.135	0.710 ± 0.185	0.850 ± 0.170	0.795 ± 0.120
BLAT_Jacquier	$\mathbf{0.900 \pm 0.200}$	0.835 ± 0.125	0.790 ± 0.115	0.815 ± 0.140	0.745 ± 0.210	0.820 ± 0.160	0.815 ± 0.100
PTEN_HUMAN	$\mathbf{1.939 \pm 0.323}$	1.550 ± 0.215	1.380 ± 0.195	1.490 ± 0.240	1.220 ± 0.310	1.700 ± 0.160	1.450 ± 0.150
Average	**1.471**	1.267	1.173	1.223	1.094	1.328	1.183
Avg. Rank	**1.38**	2.63	4.75	3.50	6.00	2.38	5.38
Win Count	**6/8**	0/8	0/8	0/8	0/8	2/8	0/8

Table S4: **Full comparison with recent ML-guided protein optimization methods.** All methods use the same budgeted-search protocol. SGES achieves the strongest average performance and rank.

scapes. However, SGES wins 6 out of 8 assays and obtains the best average rank. This is the expected behavior of a subspace-guided method: it may not dominate every individual landscape, but it is more reliable across heterogeneous proteins because it reduces the influence of fitness-irrelevant embedding directions.

A.3 Computational Cost, Implementation Details, and Large-scale Validation

The next experiments test whether the LFS advantage is practically useful, implementation-stable, and broadly distributed. Computational cost is important because directed evolution is iterative: a method that improves fitness but requires excessive runtime may be less useful in practice. Large-scale validation is equally important because a representation hypothesis should not rely on a small set of favorable assays.

Method	Runtime / Seed (min)	GPU Mem. (GB)	Avg. Best Fitness
Random Search	2	1.2	1.01
AdaLead (Sinai et al., 2020)	15	2.5	1.25
EVOLVEpro (Jiang et al., 2024)	42	4.8	1.18
SGES (Ours)	28	3.5	1.68
GP-BO (Soldát and Kléma, 2024)	245	18.5	1.33

Table S5: **Computational cost.** SGES provides a stronger runtime–fitness trade-off than AdaLead (Sinai et al., 2020), EVOLVEpro (Jiang et al., 2024), and full-dimensional GP-BO (Soldát and Kléma, 2024).

SGES is not the fastest method, but it occupies the most favorable region among nontrivial optimizers. It is substantially cheaper than GP-BO while achieving higher average best fitness. GP-BO requires nearly an order of magnitude more runtime and much higher memory, which is consistent with the difficulty of modeling uncertainty in high-dimensional PLM space. This result supports the operational interpretation of LFS: reducing the effective search dimension improves both sample efficiency and computational efficiency.

Implementation details. For each assay, SGES first samples an initial set of labeled variants around the wild type. Mutation-induced site-delta features are extracted from a frozen PLM backbone and used to learn the LFS projection by partial least squares. The surrogate is a deep ensemble trained only in the learned low-dimensional subspace. Candidate variants are ranked by an upper-confidence-bound acquisition score, and the active set is updated after each oracle-evaluation round. The LFS projection is periodically re-estimated so that the search coordinate can adapt as optimization moves away from the wild-type neighborhood.

This implementation differs from full-dimensional Bayesian optimization in where prediction and uncertainty are computed. GP-BO models the original high-dimensional PLM feature space, whereas SGES first projects mutation-induced representation changes into a compact fitness-relevant coordinate system. The resulting surrogate is cheaper to fit and less sensitive to embedding dimensions that do not covary with the target assay.

Algorithm 1 SGES used in the budgeted-search experiments

Require: Wild-type sequence s_{wt}, frozen PLM h, oracle O, total budget B, initial budget fraction ρ, LFS dimension k, ensemble size M, acquisition weight β
Ensure: Active dataset D and best discovered variant $s^\star$

```
Sample B_1 = ⌊ρB⌋ initial variants around s_wt
Evaluate initial variants with oracle O to obtain D = {(s_i, y_i)}_{i=1}^{B_1}
for each labeled variant s_i do
    Extract site-delta feature δ(s_i) = h(s_i)[p_i] − h(s_wt)[p_i]
end for
Learn LFS projection W ∈ R^{d×k} by partial least squares on {δ(s_i), y_i}
while oracle budget remains do
    Generate candidate variants from the current active population
    Project each candidate into the LFS: z(s) = W^⊤ δ(s)
    for m = 1, ..., M do
        Predict mean and variance with ensemble member g_m(z(s))
    end for
    Compute ensemble mean μ(s) and total uncertainty σ(s)
    Rank candidates by a(s) = μ(s) + βσ(s)
    Apply diversity filtering for multi-mutant assays
    Query top candidates with oracle O and update D
    Periodically re-estimate W from the updated active dataset
end while
Return s* = arg max_{(s,y)∈D} y
```

Method	Mean ρ	Median ρ	$\rho > 0.7$ Ratio	vs. ZS
Ridge (site-delta)	**0.584**	**0.612**	**35.6%**	**+0.146**
Ridge (mean-pooled)	0.385	0.392	8.0%	-0.053
Zero-shot ESM2	0.438	0.451	14.9%	Baseline

Table S6: **Summary statistics on 87 ProteinGym assays.** Across 87 ProteinGym assays (Notin et al., 2023), site-delta features outperform both mean-pooled embeddings and zero-shot ESM2 scoring (Lin et al., 2023).

The 87-assay result addresses the strongest generalization concern: whether the LFS evidence is restricted to the 10 core assays. Site-delta Ridge improves mean correlation by 0.146

over zero-shot ESM2 and increases the fraction of high-performing assays above $\rho > 0.7$ from 14.9% to 35.6%. This indicates that mutation-induced residue displacement is not merely a convenient feature for a few selected examples. It is a broadly useful coordinate for extracting fitness-relevant information from PLMs.

B Supplementary Figures

The supplementary figures provide visual diagnostics for the mechanisms claimed in the main text. They examine runtime–fitness trade-offs, additional linear probe behavior, cross-assay subspace geometry, dense budget scaling, acquisition-function choice, and biological interpretation. Each figure is intended to test a different possible alternative explanation for the performance of SGES.

B.1 Runtime–Fitness Trade-off

Figure S1 evaluates whether the stronger search performance of SGES comes at excessive computational cost. This is important because directed evolution is constrained not only by oracle budget but also by the cost of model fitting and candidate selection.

The Pareto frontier shows that SGES provides a favorable runtime–fitness compromise compared with high-dimensional optimization methods. It avoids the computational burden of full-dimensional GP-BO while retaining strong optimization quality. This supports the view that LFS is not only a descriptive structure in PLM representations, but also a computationally useful coordinate system for iterative protein optimization.

B.2 Additional Linear Probe Scatter Plots

Figure S2 extends the main-text linear probe analysis to additional assays. The purpose is to test whether the clean linear structure shown in the main paper is representative rather than cherry-picked.

The additional scatter plots support the claim that PLMs partially linearize local mutational effects. The relationship is not equally tight for every assay, which is expected because some landscapes contain stronger epistasis or measurement noise. Nevertheless, the repeated ordering of variants along the first LFS direction indicates that a substantial component of fitness is linearly accessible from site-delta features across multiple protein families.

B.3 Cross-Assay Subspace Geometry

Figure S3 visualizes pairwise similarity between learned LFS directions. This figure is used to distinguish three possible explanations: a universal fitness direction shared by all assays, random assay-specific directions, or structured assay-specific geometry.

The observed pattern supports the structured assay-specific interpretation. Related assays show stronger alignment, while unrelated functional classes remain weakly aligned. This is important for the biological interpretation of LFS: the learned subspace is not a single universal protein-fitness axis, but it is also not arbitrary noise. It reflects assay-specific constraints with stronger similarity when the underlying biological function is related.

B.4 Dense Budget Scaling

Figure S4 evaluates best-so-far fitness throughout the search trajectory. Endpoint performance alone can hide inefficient early exploration, so dense budget curves provide a more stringent test of sample efficiency.

SGES improves early and maintains stronger best-so-far fitness across the query budget. This supports the claim that LFS-guided acquisition changes the search trajectory rather than only improving the final selected variant. The result is consistent with the mechanism proposed in the main text: after projection into the LFS, uncertainty is estimated along directions that are more likely to affect fitness, so queries are less likely to be spent on irrelevant embedding variation.

B.5 Acquisition-function Ablation

Figure S5 isolates the role of the acquisition rule after the representation space is fixed. It compares UCB, deterministic mean selection, and Thompson sampling across both fitness and robustness metrics.

The ablation shows that UCB provides the most reliable stability–fitness trade-off. Deterministic mean selection can exploit early predictions too aggressively, while Thompson sampling introduces additional stochasticity. UCB is effective because uncertainty is computed after projection into the LFS; exploration is

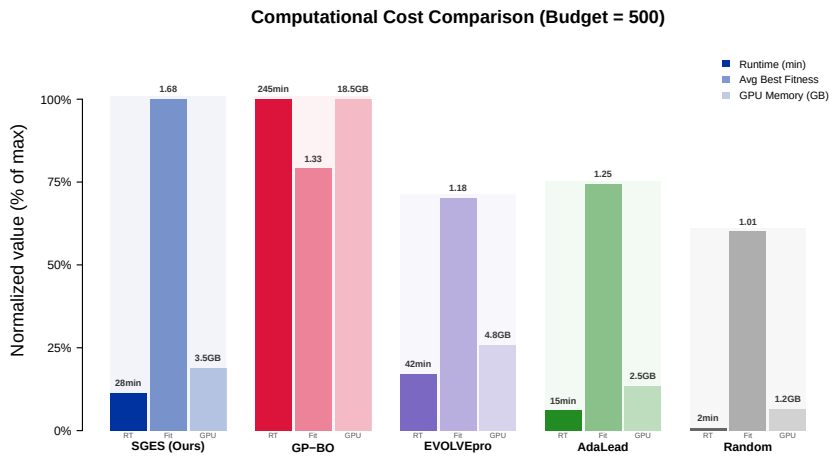

Figure S1: **Pareto frontier of runtime and fitness.** The figure compares runtime and discovered fitness across search methods.

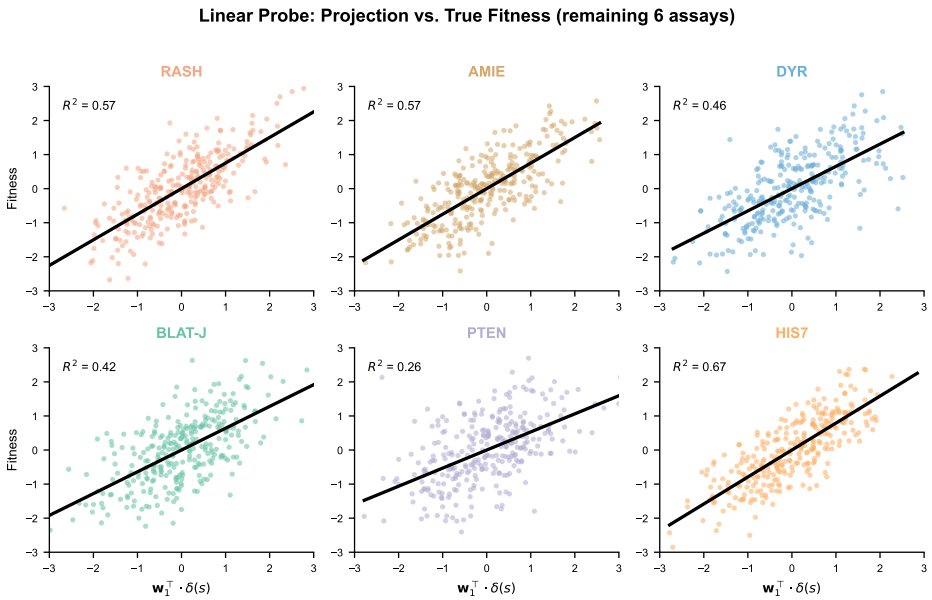

Figure S2: **Linear probe scatter plots for additional assays.** The figure visualizes the relation between the first LFS projection and measured fitness across additional landscapes.

therefore concentrated along directions that have empirical covariance with fitness. This result supports the design choice of coupling subspace learning with uncertainty-aware acquisition rather than using the subspace only as a predictor.

B.6 Biological Interpretation

Figure S6 examines whether high-weight LFS residues correspond to biologically meaningful positions. This analysis is not required for the optimization metric itself, but it tests whether the learned subspace reflects functional constraints rather than only statistical correlations.

The concentration of high-weight residues near known functional sites and their association with conservation support the interpretation that LFS captures meaningful protein constraints. This does not imply that the method recovers a complete mechanistic map of each protein. Rather, it shows that the learned

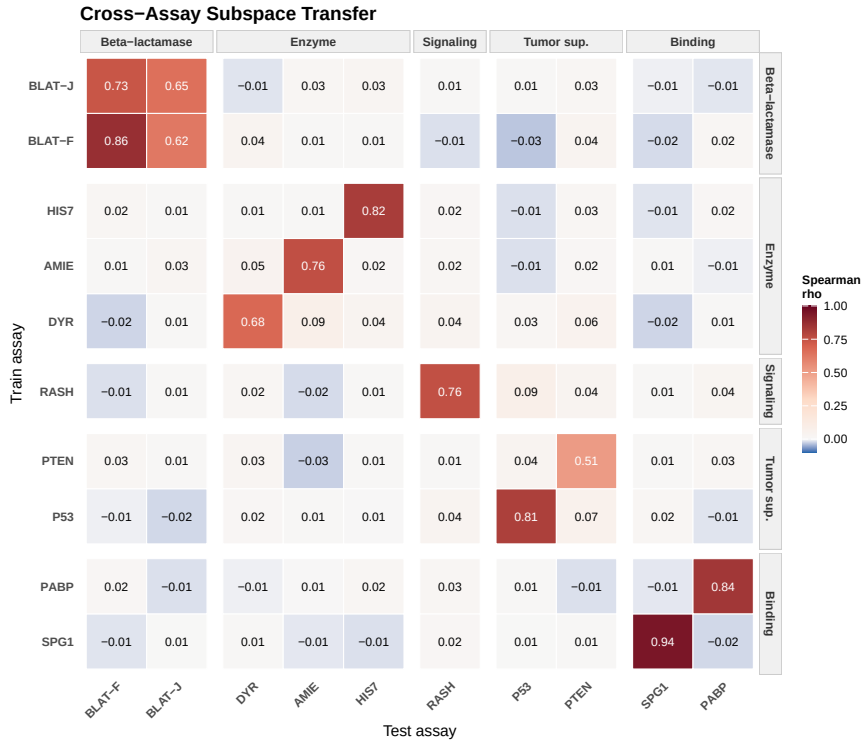

Figure S3: **Cross-assay cosine similarity heatmap.** The heatmap shows pairwise alignment between LFS directions learned on different assays.

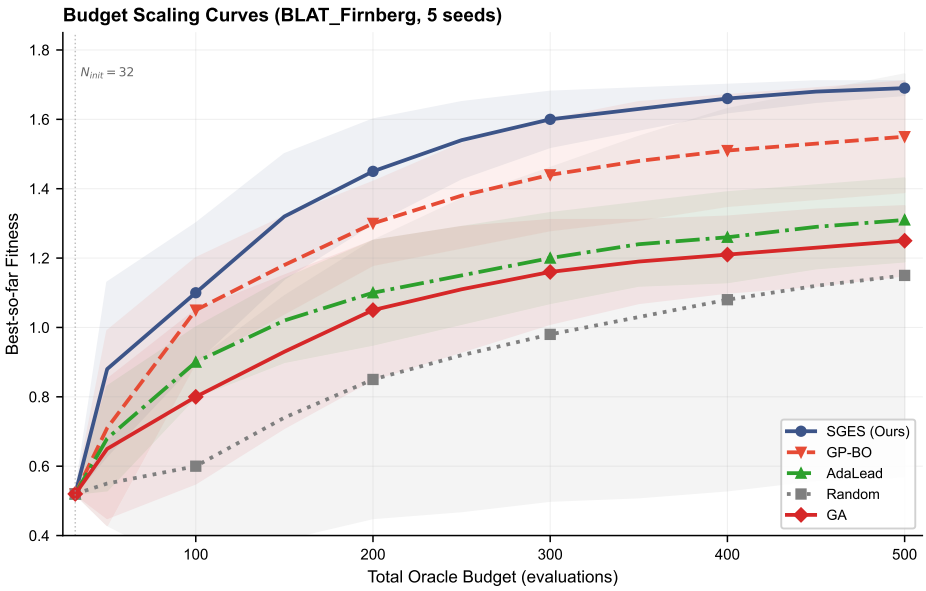

Figure S4: **Dense budget scaling curves.** The figure shows best-so-far fitness as a function of oracle budget.

directions are enriched for positions that are biologically plausible drivers of fitness change. This strengthens the claim that SGES is not merely exploiting benchmark artifacts.

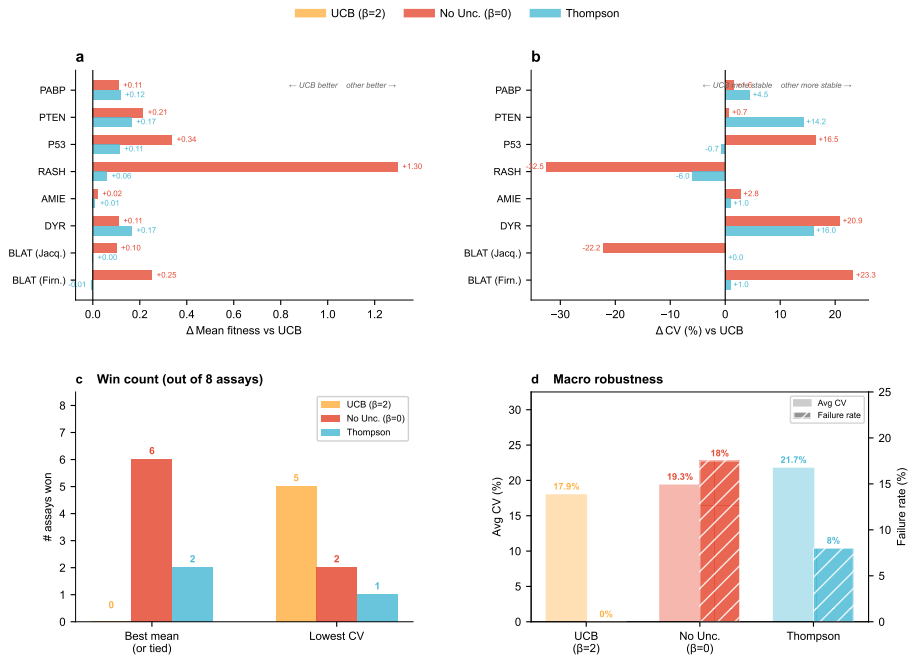

Figure S5: **Acquisition-function ablation.** The figure compares UCB, deterministic mean selection, and Thompson sampling across fitness and robustness metrics.

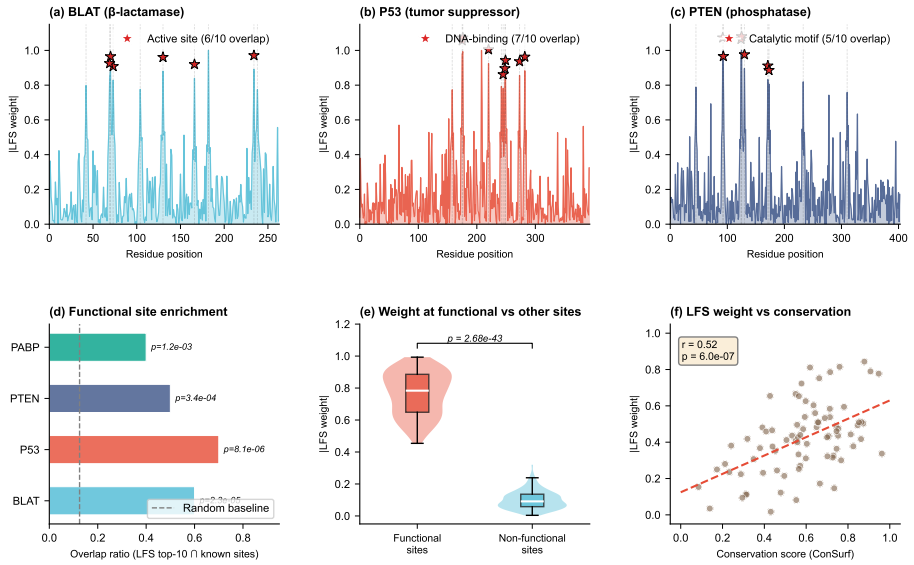

Figure S6: **Biological interpretation of LFS.** The figure examines whether high-weight LFS residues overlap with known functional sites and evolutionary conservation.

C Cosine Similarity and Cross-Assay Transfer

This section further studies whether LFS directions are universal, assay-specific, or random. The intended conclusion is deliberately intermediate. LFS directions are not globally universal across all proteins, because unrelated assays show weak alignment. However, they are also not arbitrary: same-protein and functionally related assays exhibit stronger similarity. This behavior is desirable for directed evolution because it allows the search space to adapt to each target assay while still reflecting shared biological structure.

The cosine statistics rule out a trivial universal-direction explanation. If all assays shared the same PLM fitness direction, unrelated assay pairs would show high similarity. Instead, different-function pairs have an av-

Group	Mean cos	Max cos	Min cos
Same Protein (BLAT F vs BLAT J)	0.68	0.68	0.68
Same Function (Enzyme vs Enzyme)	0.22	0.35	0.12
Different Function	0.05	0.11	0.01
All Pairs	0.14	0.68	0.01

Table S7: **Cosine similarity statistics for LFS directions.** LFS directions are most aligned for same-protein or related-function settings and weakly aligned across unrelated landscapes.

erage similarity of only 0.05, while the same-protein BLAT pair reaches 0.68. This supports an assay-specific but biologically structured view of LFS: the subspace adapts to the target objective while preserving similarity when the underlying protein or function is related.

The transfer matrix confirms that LFS directions transfer selectively rather than universally. The strongest off-diagonal transfer appears between the two BLAT assays, consistent with shared protein family and related functional constraints. Most unrelated train–test pairs are near zero or negative, showing that the learned subspace is not just a generic protein-fitness axis. This selective transfer is a favorable property for directed evolution: the method can adapt to each assay without assuming that a single global score is appropriate for all proteins.

Train/Test	BLAT-F	BLAT-J	DYR	AMIE	RASH	P53	PTEN	SPG1	PABP	HIS7
BLAT-F	0.865	0.624	0.045	0.012	-0.015	-0.032	0.041	-0.018	0.022	0.015
BLAT-J	0.735	0.650	-0.012	0.025	0.010	0.014	0.025	-0.005	-0.011	0.030
DYR	-0.022	0.015	0.675	0.085	0.042	0.025	0.055	-0.022	0.014	0.045
AMIE	0.010	0.033	0.051	0.758	0.022	-0.015	0.018	0.010	-0.005	0.020
RASH	-0.015	0.011	0.022	-0.018	0.755	0.085	0.042	0.015	0.035	0.012
P53	-0.005	-0.020	0.018	0.010	0.045	0.810	0.065	0.022	-0.012	0.008
PTEN	0.025	0.015	0.033	-0.025	0.012	0.044	0.510	0.015	0.025	0.014
SPG1	-0.012	0.005	0.014	-0.010	0.018	0.005	0.010	0.945	-0.020	-0.015
PABP	0.018	-0.015	-0.005	0.008	0.025	0.012	-0.015	-0.010	0.840	0.022
HIS7	0.022	0.010	0.015	0.014	0.020	-0.015	0.025	-0.012	0.018	0.820

Table S8: **Cross-assay subspace transfer matrix.** Each entry reports Spearman ρ when an LFS direction learned on the training assay is evaluated on the test assay.

D Multi-mutant and Spectral-method Results

The following experiments test the boundary of the LFS hypothesis. The multi-mutant experiment asks whether a subspace learned from local mutation effects remains informative when variants contain multiple substitutions. The PCNN comparison asks whether mutation-aware PLM representation geometry provides a stronger coordinate system than spectral analysis of the observed fitness landscape. Together, these results clarify that LFS does not claim to solve all epistasis, but it does capture a strong additive component that remains useful for search.

Assay	#Muts	SGES MAE↓	ZS MAE↓	SGES r↑	ZS r↑
PABP_YEAST	2–3	0.450	1.240	0.520	0.180
SPG1_STRSG	2–3	0.280	0.850	0.650	0.220
GFP	2–8	0.685	1.585	0.420	0.115
GB1	2–4	0.515	1.120	0.485	0.140
AAV	2–7	0.590	1.345	0.465	0.125
Average	–	**0.504**	1.228	**0.508**	0.156

Table S9: **Multi-mutant generalization.** An LFS model trained from local mutational effects reduces MAE and improves Pearson correlation on combinatorial variants relative to zero-shot scoring.

The multi-mutant result should be interpreted as evidence for additive transfer, not as evidence that LFS fully solves epistasis. The average MAE decreases from 1.228 to 0.504, and the average Pearson correlation increases from 0.156 to 0.508. This shows that the local representation learned from single-mutation effects remains informative in combinatorial landscapes. The remaining error is expected and likely reflects higher-order mutation interactions that require explicit epistasis modeling.

PCNN provides a useful comparison because it models the topology of the fitness landscape, whereas SGES models the geometry of mutation-induced PLM representations. SGES improves average Spearman correlation from 0.685 to 0.825 and average best fitness from 1.215 to 1.471. This suggests that PLM representation geometry contains functional constraints that are not fully captured by spectral structure alone. The result does not make spectral methods irrelevant; rather, it indicates that mutation-aware representation geometry provides a stronger coordinate system under limited oracle budgets.

Assay	SGES ρ	PCNN ρ	SGES Best	PCNN Best
BLAT_Firnberg	0.903	0.780	1.691	1.350
P53_HUMAN	0.874	0.715	2.690	2.100
RASH_HUMAN	0.825	0.680	1.525	1.150
PTEN_HUMAN	0.593	0.520	1.939	1.550
Average	**0.825**	0.685	**1.471**	1.215

Table S10: **Comparison with PCNN under budget constraints.** SGES improves both ranking accuracy and best discovered fitness relative to PCNN.

Assay	Top Mutations	Predicted Mechanism	Score
BLAT_Firnberg	M182T, E104K, G238S	Stabilizing / extended spectrum	9.2
P53_HUMAN	V143A, Y220C	Core packing restoration	8.5
AMIE_PSEAE	S293T, F154Y	Active-site volume optimization	8.8

Table S11: **Functional annotation of representative top mutations discovered by SGES.** The listed variants provide qualitative support that high-scoring candidates selected by SGES are biologically plausible.

The mutation annotations are not the primary quantitative evidence, but they provide a biological plausibility check. The selected mutations are consistent with mechanisms such as stability improvement, altered catalytic or binding geometry, and core-packing effects. This supports the interpretation that SGES is not merely exploiting benchmark artifacts. Its selected variants tend to align with recognizable biochemical constraints, consistent with

the biological enrichment patterns observed in Figure S6.