LAVEP:LANGUAGE GUIDED PROTEIN LANGUAGE MODELS FOR VARIANT EFFECT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how sequence variants alter protein function is central to protein engineering and variant interpretation. However, mutation effects are inherently tied to the phenotype being measured. Protein language models (PLMs) provide transferable sequence priors, but typically evaluate mutations with a fixed scoring function independent of whether the experiment measures binding, activity, stability, expression, or other phenotypes. Supervised methods can capture assay specificity, but rely on target-assay labels that are experimentally costly and time-consuming to obtain and often scarce in practice, limiting their use for new assays. We introduce a natural-language-guided framework that uses assay descriptions to tell a frozen PLM what "fitness" means in the experiment at hand. We treat language as an *open-vocabulary task specification*: assay context is grounded in the protein representation and used to steer the PLM's mutation prior toward the measured phenotype, without fine-tuning the PLM or requiring target-assay labels. We further establish a fundamental limit of assay-agnostic prediction. When the same variants are ranked differently across assays, any shared predictor must compromise between distinct fitness landscapes, imposing an intrinsic ceiling on attainable performance. Conditioning on assay language allows the model to represent these experimental objectives separately and escape this shared-ranking constraint. Across homology-separated ProteinGym assays, our approach consistently improves zero-shot mutation-effect prediction on *unseen assays* across ranking and classification objectives. It benefits further from limited labeled support variants, showing the value of assay language when new experimental labels are unavailable or scarce.

1 INTRODUCTION

Proteins can acquire new properties when their amino-acid sequences change. Even a single amino-acid substitution can alter protein folding, stability, interactions, or activity. Predicting these *variant effects* is therefore central to protein engineering and variant interpretation (Notin et al., 2023a; Marquet et al., 2022). Deep mutational scanning (DMS) measures thousands of variants (Fowler et al., 2010; Notin et al., 2023a), but such measurements are costly, leaving many assays with scarce labels (Teufel et al., 2025). Zero-shot prediction can help prioritize variants before those labels exist.

Variant effects are inherently dependent on the phenotype being measured: the same mutation may have different functional consequences when a protein is evaluated for binding, activity, stability, expression, or growth under different conditions (Notin et al., 2023a; Melnikov et al., 2014; Chiasson et al., 2020). Pretrained protein language models (PLMs) provide transferable sequence representations for zero-shot mutation effect prediction without target-assay labels (Lin et al., 2023; Meier et al., 2021; Marquet et al., 2022; Notin et al., 2022; Dinh et al., 2026). However, relying on a fixed sequence prior makes these predictions largely *assay-agnostic*: they do not account for the experimental phenotype being measured. Supervised approaches can learn assay-specific effects but require labeled measurements from the assay of interest (Notin et al., 2023b; Hsu et al., 2022). This leaves an important gap: enabling **zero-shot assay-aware** prediction for unseen assays without requiring assay-specific labels.

We explore a simple source of such information: the experiment can be specified in natural language. An assay description can specify the measured phenotype, molecular context, interaction partner,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

experimental conditions, and meaning of the reported score. We therefore view language not merely as annotation, but as an *open-vocabulary* **task specification** that tells a protein model what "fitness" means in the experiment at hand. This perspective complements recent protein–text models that use language for functional annotation and retrieval (Xu et al., 2023; Su et al., 2025), mutation explanation (Luo et al., 2024), and protein design (Zhuang et al., 2025). Here, the PLM provides a general **sequence prior**, while assay language specifies the **experimental objective** for interpreting that prior.

Based on this insight, we introduce **LAVEP** (**L**anguage-guided, **A**ssay-aware **V**ariant **E**ffect **P**rediction), a framework that leverages assay language as a task specification to adapt transferable PLM priors to different experimental objectives. Rather than learning separate predictors for individual assays, LAVEP enables zero-shot assay-aware prediction on unseen assays by conditioning mutation effect predictions on assay context. When measurements become available, LAVEP further supports lightweight adaptation through label-efficient residual correction. This perspective also motivates a fundamental analysis of assay-agnostic predictors: when two assays induce different mutation effects for the same variants, an assay-agnostic predictor must produce identical predictions across assays, creating an intrinsic limit on achievable accuracy. We formalize this limitation as an *assay-agnostic ranking ceiling* and characterize when assay-conditioned models can surpass it.

We evaluate our framework on ProteinGym (Notin et al., 2023a), using a homology-separated split in which target assays are held out during training. On these **unseen assays**, conditioning on assay language improves aggregate zero-shot mutation-effect prediction over assay-agnostic baselines across ranking and classification metrics. Together, these results support natural language as a flexible interface for adapting pretrained protein sequence priors to new experimental objectives without first collecting fitness measurements from those experiments.

Our main contributions are:

- **Problem formulation:** We formulate zero-shot assay-aware variant effect prediction: given an unseen assay described in natural language, infer assay-specific variant effects without target-assay fitness labels. This setting addresses a gap left by existing zero-shot methods, which typically do not condition predictions on the experimental objective.
- **Language-guided conditioning:** We introduce a framework that treats natural-language assay descriptions as an open-vocabulary task specification for a frozen PLM, adapting a transferable sequence prior to the phenotype being measured without learning a separate predictor for each assay.
- **Theory and empirical validation:** We formalize an *assay-agnostic ranking ceiling*, showing that methods producing the same predictions across assays face an intrinsic limit when assays induce different variant effects, and derive sufficient conditions under which assay conditioning can surpass this ceiling. On homology-separated ProteinGym assays, LAVEP improves zero-shot performance on held-out assays across complementary evaluation metrics.

2 Related Work

2.1 Variant Effect Prediction

Variant effect prediction estimates how sequence variants alter experimentally measured phenotypes such as activity, binding, stability, or expression. Zero-shot methods draw on several sources of biological prior. **Sequence-based PLMs** learn statistical regularities from natural protein sequences and score mutations by their compatibility with the learned sequence distribution (Meier et al., 2021; Marquet et al., 2022; Dinh et al., 2026). **Homology-aware methods** use homologous sequences or multiple-sequence alignments to capture conservation and co-variation (Laine et al., 2019; Rao et al., 2021; Truong Jr. & Bepler, 2023; Notin et al., 2022; Weitzman et al., 2025). **Structure-aware methods** incorporate three-dimensional context through structure-conditioned likelihoods or structural representations (Li et al., 2024; Guo et al., 2026). Despite these different priors, their mutation scores are generally *assay-agnostic*: the same protein–variant pair receives the same prediction regardless of what the experiment measures.

Other methods introduce increasing amounts of assay information. ProteinNPT (Notin et al., 2023b) learns assay-specific fitness from labeled variants in the target assay. METALIC (Beck et al., 2025) meta-learns across fitness landscapes and supports zero-shot transfer to unseen tasks, but its zero-shot predictions remain *assay-agnostic* in the sense relevant here: the target assay itself is not semantically specified to the model. PRIMO (Teufel et al., 2025) adds an explicit assay signal through a categorical property type such as binding, stability, or activity. This provides task information but remains coarse and closed-vocabulary, since different experiments within the same category share the same representation. We instead use the natural-language assay description as a **fine-grained, open-vocabulary** specification of the experimental objective, enabling assay-aware prediction on previously unseen assays without target-assay fitness labels.

2.2 Protein–Text Multimodal Learning

Protein–text models use natural language to complement information available from protein sequences. Prior work has aligned protein and text representations for retrieval and zero-shot functional prediction (Xu et al., 2023; Su et al., 2025), generated protein descriptions or mutation explanations (Jararweh et al., 2025; Fei et al., 2025; Luo et al., 2024), and used functional descriptions or instructions to guide protein design (Liu et al., 2025b; Guo et al., 2025; Liu et al., 2025a; Zhuang et al., 2025).

Our use of language is different: language specifies *the experiment under which a mutation should be evaluated*. The assay description defines the measured phenotype and experimental context that distinguish one fitness landscape from another. Language therefore serves as an task specification for variant-effect prediction, instead of only describing, explaining, or specifying the protein itself.

3 Method

3.1 Overview

Method workflow. As illustrated in Figure 1, LAVEP proceeds in four steps. **First**, a frozen protein language model encodes the sequence and candidate mutation into mutation-centric protein representations. **Second**, a frozen text encoder represents each assay description using five semantic fields, which can be contrasted at inference for proteins with multiple assays. **Third**, bidirectional cross-attention refines each text field with the mutation-specific protein representation. The resulting field-wise updates are then integrated with the original sequence prior to yield the zero-shot prediction. **Fourth**, when labeled support variants are available, kernel ridge regression (KRR) models the residual discrepancy between support-set label ranks and scaled zero-shot scores, providing lightweight nonparametric adaptation without updating the neural model.

3.2 Assay-Agnostic PLM Representation

Encoding the masked mutation context. We first extract the assay-independent protein representation to be steered by assay language. Let $x_a \equiv x_a^{\mathrm{WT}}$ be the wild-type sequence and consider a substitution $x_{a,i} \to m$. A frozen ESM-2 model encodes both x_a and its masked counterpart $x_a^{\setminus i}$:

$$H_a^{\mathrm{WT}} = E_{\mathrm{ESM}}(x_a), \qquad H_a^{\mathrm{mask}(i)} = E_{\mathrm{ESM}}(x_a^{\setminus i}), \qquad h_i^0 = [H_a^{\mathrm{mask}(i)}]_i. \tag{1}$$

Retaining local and protein-wide context. Using h_i^0 and H_a^{WT} from Eq. 1, we retain the masked-position state, the wild-type state at position i, a protein-wide summary, and the normalized sequence position. After projection into a shared interaction space, these form

$$\begin{aligned} Z_{a,i} = \big[& P_{\mathrm{mask}}(h_i^0) + e_{\mathrm{mask}};\ P_{\mathrm{WT}}([H_a^{\mathrm{WT}}]_i) + e_{\mathrm{WT}}; \\ & P_{\mathrm{global}}(\mathrm{MeanPool}(H_a^{\mathrm{WT}})) + e_{\mathrm{global}};\ P_{\mathrm{pos}}(i/L_a) + e_{\mathrm{pos}} \big]. \end{aligned} \tag{2}$$

The token set $Z_{a,i}$ contains masked, wild-type, protein-wide, and positional context. Here L_a is sequence length, P denotes learned projections, and e denotes token-type embeddings. Neither h_i^0 nor $Z_{a,i}$ depends on the assay description.

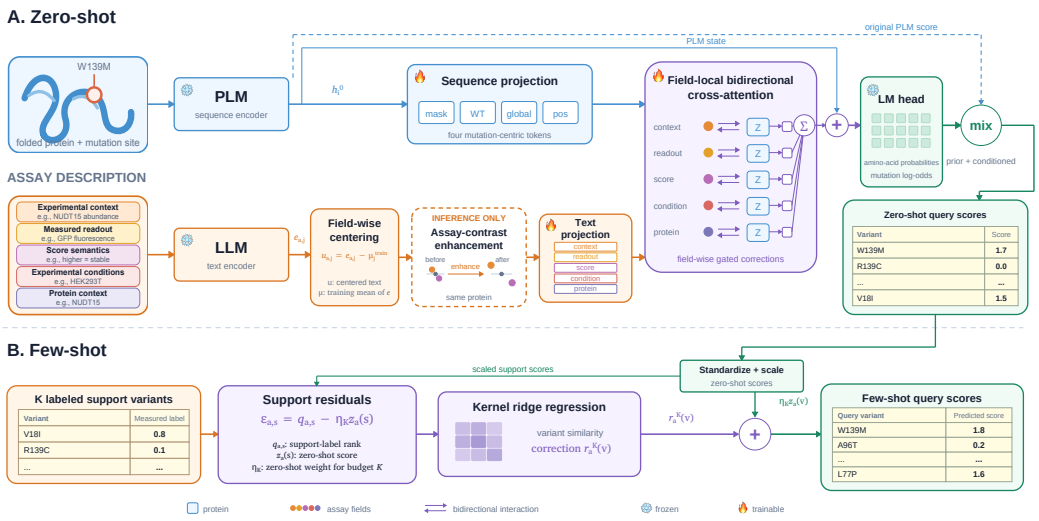

Figure 1: **Overview of LAVEP.** The sequence and mutation branch produces an assay-independent protein representation, while the language branch encodes five fields of the assay description. Field-local bidirectional interaction yields the zero-shot variant-effect score. In the few-shot setting, labeled support variants refine this score through a kernel model fitted to support-set residuals.

3.3 Natural-Language Assay Representation

Describing the assay with five semantic fields. An assay description specifies what was measured and how the reported values should be interpreted. We retain five complementary fields: experimental context, measured readout, score semantics, experimental conditions, and protein information:

$$d_a = \left(d_a^{\text{context}}, d_a^{\text{readout}}, d_a^{\text{score}}, d_a^{\text{condition}}, d_a^{\text{protein}}\right). \tag{3}$$

Appendix A defines the scope of each field; an example appears in Section 5.1.

Encoding and normalizing the semantic fields. The five fields in Eq. 3 are independently encoded by the frozen Qwen3-Embedding-8B text encoder (Zhang et al., 2025):

$$e_{a,j} = E_{\text{text}}(d_a^j) \in \mathbb{R}^{4096}, \qquad j \in \{\text{context}, \text{readout}, \text{score}, \text{condition}, \text{protein}\}. \tag{4}$$

The embeddings $e_{a,j}$ in Eq. 4 share some content across assays within each field, which contributes little to distinguishing one assay from another. We therefore subtract the mean of each field across the training assays:

$$\mu_j^{\text{train}} = \frac{1}{|\mathcal{A}_{\text{train}}|} \sum_{b \in \mathcal{A}_{\text{train}}} e_{b,j}, \qquad u_{a,j} = e_{a,j} - \mu_j^{\text{train}}. \tag{5}$$

The training means μ_j^{train} are reused for unseen assays. Each centered embedding is projected into the protein–language interaction space to form a text token:

$$t_{a,j} = P_{\text{text}}(u_{a,j}) + e_j^{\text{text}}, \qquad j \in \{\text{context}, \text{readout}, \text{score}, \text{condition}, \text{protein}\}. \tag{6}$$

Here, P_{text} is learned and e_j^{text} identifies the field type.

3.4 Language-Guided Steering of PLM Mutation Log-Odds

Bidirectional protein–language interaction. The same assay field can imply different effects for different mutations and protein contexts. Conversely, which sequence features matter depends on what the assay measures. We therefore let the text token $t_{a,j}$ from Eq. 6 interact with the protein tokens $Z_{a,i}$ from Eq. 2 through bidirectional cross-attention:

$$\begin{aligned} \bar{t}_{a,i,j} &= \text{LN}\left(t_{a,j} + \text{MHA}_{T \leftarrow Z}\left(Q = t_{a,j}, K = Z_{a,i}, V = Z_{a,i}\right)\right), \\ \overline{Z}_{a,i}^{(j)} &= \text{LN}\left(Z_{a,i} + \text{MHA}_{Z \leftarrow T}\left(Q = Z_{a,i}, K = \bar{t}_{a,i,j}, V = \bar{t}_{a,i,j}\right)\right). \end{aligned} \tag{7}$$

The resulting text and protein states remain field-specific.

Constructing field-specific corrections. Different semantic fields need not be equally informative for every mutation. For each field j, we combine the updated text and protein states in Eq. 7 to form a feature $c_{a,i,j}$. A field-specific branch F_j then proposes a hidden-state correction, while a mutation-dependent scalar gate $g_{a,i,j}$ modulates its contribution:

$$c_{a,i,j} = \left[\mathrm{Concat}\left(\overline{Z}^{(j)}_{a,i,1:3}\right); \bar{t}_{a,i,j}\right], \qquad g_{a,i,j} = \sigma(G_j(c_{a,i,j})), \qquad r_{a,i,j} = g_{a,i,j} F_j(c_{a,i,j}). \tag{8}$$

Here, F_j determines the correction proposed by semantic field j, whereas G_j determines how strongly that field should influence the current mutation. Maintaining separate field-specific branches preserves the provenance of each semantic signal without collapsing all assay information into a single representation.

We add the five corrections from Eq. 8 to the original PLM state h_i^0 (Eq. 1):

$$\widetilde{h}_{a,i} = h_i^0 + \eta \sum_{j=1}^{5} r_{a,i,j}, \qquad \ell_a(i, w \to m) = \log p_a(m \mid i) - \log p_a(w \mid i). \tag{9}$$

The resulting state $\widetilde{h}_{a,i}$ is passed through the frozen ESM-2 language-model head, $p_a(\cdot \mid i) = \mathrm{Softmax}\Big(H_{\mathrm{LM}}(\widetilde{h}_{a,i})\Big)$, thereby translating field-specific assay semantics into an assay-conditioned mutation log-odds while retaining the pretrained PLM as the underlying sequence prior. Here, η is a fixed residual scale.

3.5 Training Across Assays

For a variant $v_{a,n}$ with $K_{a,n}$ substitutions, we sum the mutation log-odds from Eq. 9 to obtain its score:

$$s_{a,n} = \sum_{k=1}^{K_{a,n}} \ell_a(i_k, w_k \to m_k). \tag{10}$$

The model is **trained jointly across assays**, while all variant comparisons remain within assay. Let $\mathcal{P}_a$ denote pairs of measured variants from assay a, and let $p_i^0 = \mathrm{Softmax}(H_{\mathrm{LM}}(h_i^0))$ be the original PLM distribution. Using the variant scores in Eq. 10, we optimize

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{a,(r,s)\in\mathcal{P}_a}\left[\mathrm{softplus}\left(-\frac{\mathrm{sign}(y_{a,r}-y_{a,s})(s_{a,r}-s_{a,s})}{\tau}\right)\right] \\ + \lambda_{\mathrm{KL}}\mathbb{E}_{a,i}\left[D_{\mathrm{KL}}\left(p_i^0 \,\|\, p_a(\cdot \mid i)\right)\right],\end{aligned} \tag{11}$$

where τ controls the ranking temperature and λ_{KL} controls regularization toward the original ESM-2 distribution p_i^0. The ranking term preserves the observed ordering within each assay, while shared parameters across assays allow the model to learn transferable associations between assay semantics and mutation effects. The KL term anchors these assay-conditioned corrections to the pretrained protein prior, preventing language conditioning from unnecessarily overriding it.

3.6 Inference on Unseen Assays

Assay contrast. Assays of the same protein share much of their background, which can obscure differences in what they measure. Before inference, we therefore contrast the centered embeddings from Eq. 5 across assays of protein p. Let $\mathcal{A}_{p,j}$ contain the available assays with text in field j. For the context, readout, score-semantics, and condition fields, we amplify each assay's deviation from the field mean:

$$\mu_{p,j} = \frac{1}{|\mathcal{A}_{p,j}|}\sum_{b\in\mathcal{A}_{p,j}} u_{b,j}, \qquad \widehat{u}_{a,j} = \mathcal{N}_{u_{a,j}}\left(\mu_{p,j} + \gamma(u_{a,j} - \mu_{p,j})\right). \tag{12}$$

Here $\gamma > 1$ sets the contrast strength, and $\mathcal{N}_u(x) = \|u\|_2 x/\|x\|_2$ restores the original norm. We also attenuate overlap between the context and readout/score deviations as in Eq. 18 (Appendix C.3), yielding the final embeddings $\widetilde{u}_{a,j}$. The protein-information field is unchanged. For proteins with only one available assay, we set $\widetilde{u}_{a,j} = u_{a,j}$. Otherwise, $\widetilde{u}_{a,j}$ replaces $u_{a,j}$ in Eq. 6.

Zero-shot prediction. Using these text embeddings, let $s_a^{\mathrm{cond}}(v)$ be the score from Eq. 10, and let $s^{\mathrm{PLM}}(v)$ be the same mutation-score sum using the original distribution p_i^0. We combine them so that assay language can change variant rankings while retaining the protein prior:

$$s_a^{\mathrm{ZS}}(v) = (1-\alpha)s^{\mathrm{PLM}}(v) + \alpha s_a^{\mathrm{cond}}(v). \tag{13}$$

Few-shot residual adaptation. Given K labeled support variants, we fit a KRR model to correct the zero-shot score in Eq. 13. We standardize these scores using support-set statistics to obtain $z_a(v)$. For a support variant s, $q_{a,s}$ is the rank of its measured label $y_{a,s}$ within assay a's support set, standardized to zero mean and unit variance. For proteins with multiple assays, the regressor can use support variants from their other assays, with reduced weight for cross-assay similarities. It learns a residual to the zero-shot score:

$$\epsilon_{a,s} = q_{a,s} - \eta_K z_a(s), \qquad s_a^{\mathrm{FS}}(v) = \eta_K z_a(v) + r_a^K(v). \tag{14}$$

Here, $r_a^K(v)$ is the kernel prediction fitted to the support targets $\epsilon_{a,s}$. The anchor η_K keeps predictions closer to the zero-shot score when support is scarce; cross-assay sharing increases as K grows. The steering model remains fixed. The kernel construction and scheduling details are given in Appendix C.3.

4 Theoretical Analysis

Limit of a shared ranking. We analyze why assay conditioning is necessary. Consider two assays a and b performed on the same wild-type protein, with measured variant sets $\mathcal{V}_a$ and $\mathcal{V}_b$. Let $\mathcal{O}_{ab} = \mathcal{V}_a \cap \mathcal{V}_b$ contain at least two shared variants. On this overlap, let r_a^{ab} and r_b^{ab} be the centered and ℓ_2-normalized experimental rank vectors, and let $\rho_{ab} = (r_a^{ab})^\top r_b^{ab}$ be their Spearman correlation. Because an assay-agnostic predictor receives no assay information, it must induce one shared predicted ranking $\widehat{r}^{ab}$ on $\mathcal{O}_{ab}$.

Theorem 1 (Assay-Agnostic Ranking Ceiling)**.** *For any assay-agnostic predictor with centered and normalized predicted rank vector $\widehat{r}^{ab}$,*

$$\boxed{\frac{\rho(\widehat{r}^{ab}, r_a^{ab}) + \rho(\widehat{r}^{ab}, r_b^{ab})}{2} \le \sqrt{\frac{1+\rho_{ab}}{2}}.} \tag{15}$$

Why the ceiling appears. The bound is *structural*, not specific to a particular PLM. If the shared variants have similar rankings in both assays ($\rho_{ab} \approx 1$), one common ranking may perform well. As the two fitness landscapes diverge, the best attainable average correlation necessarily decreases. For example, $\rho_{ab} = 0$ yields a ceiling of $1/\sqrt{2} \approx 0.707$, while perfectly reversed rankings yield zero. The result applies only to the shared variants and does not require the complete variant sets to match. It therefore motivates using $f(x, v, d_a)$ instead of $f(x, v)$ whenever mutation effects depend on the assay.

Resolving the ranking conflict. For each assay $c \in \{a, b\}$, define its zero-shot correction $C_c(v) = s_c^{\mathrm{cond}}(v) - s^{\mathrm{PLM}}(v)$. Equation 13 then gives $s_c^{\mathrm{ZS}}(v) = s^{\mathrm{PLM}}(v) + \alpha C_c(v)$.

Corollary 1 (Assay-Conditioned Ranking Resolution)**.** *Suppose both assays have strict rankings on $\mathcal{O}_{ab}$. For each $c \in \{a, b\}$, let $\mathcal{P}_c$ contain every pair (r, s) in the overlap for which assay c prefers r to s. If*

$$s^{\mathrm{PLM}}(r) - s^{\mathrm{PLM}}(s) + \alpha\big(C_c(r) - C_c(s)\big) > 0 \quad \textit{for all } (r, s) \in \mathcal{P}_c, \tag{16}$$

then the zero-shot scores recover both assay rankings exactly. Their average Spearman correlation is one and, if $\rho_{ab} < 1$, strictly exceeds the assay-agnostic ceiling in Theorem 1.

The condition asks the assay-specific correction to reverse the PLM's incorrect pairwise preferences without reversing the correct ones. Proofs are given in Appendix B.

5 Experiments

5.1 Experimental Setup

Benchmark and Data Splits. Deep mutational scanning (DMS) assays measure the functional effects of large libraries of protein variants in parallel. We evaluate our method on the ProteinGym

DMS substitutions benchmark (Notin et al., 2023a), which standardizes 217 assays covering 186 proteins and 2,465,767 variants across diverse phenotypes. We partition the assays into training, validation, and test sets, targeting a 70/15/15 split while keeping homologous proteins in the same partition. The homology-aware grouping and assignment procedure is detailed in Appendix C.1.

Example assay description. The five fields in Eq. 3 are illustrated below for a ProteinGym Spike RBD–ACE2 binding assay.

> **Example: Spike RBD–ACE2 binding assay (Starr et al., 2020)**
> **Experimental context.** Yeast-display assay of SARS-CoV-2 Spike RBD substitutions and their effects on ACE2 binding.
> **Measured readout.** ACE2-binding signal or interaction enrichment.
> **Score semantics.** Higher scores indicate stronger or better retained ACE2 binding; lower scores indicate weaker binding.
> **Experimental conditions.** ACE2 is the interaction partner; dose, temperature, and pH are not specified.
> **Protein information.** SARS-CoV-2 Spike RBD, with mutations at residues 331–531; Spike binds a host receptor to initiate viral entry.

Baselines. We compare against eight representative baselines, grouped by their use of ProteinGym labels and target-assay information. **Label-free assay-agnostic methods** include ESM2-650M (Lin et al., 2023), VESM-650M (Dinh et al., 2026), VESPA (Marquet et al., 2022), MSA Transformer (Rao et al., 2021), and PoET-200M (Truong Jr. & Bepler, 2023); they use no ProteinGym assay labels or assay-specific information. **Per-assay supervised methods** include ProteinNPT (Notin et al., 2023b), which trains a separate predictor for each target assay. **Cross-assay adaptive methods** include METALIC (Beck et al., 2025) and PRIMO (Teufel et al., 2025), which learn across labeled non-test assays and adapt predictions to the target assay.

Evaluation Metrics. We evaluate performance using Spearman's rank correlation (ρ), AUC, MCC, and Top-10% Recall, with higher values indicating better performance. Metrics are computed using evaluation scripts adapted from ProteinGym (Notin et al., 2023a); definitions and aggregation procedures are provided in Appendix C.2.

5.2 Zero-Shot and Few-Shot Results

We assess whether assay descriptions improve variant effect prediction on unseen assays and whether limited target-assay labels provide further gains. In zero-shot evaluation (Table 1), LAVEP leads on all four reported metrics, showing that assay language can improve the protein sequence prior without target-assay labels.

Table 1 also shows that LAVEP has the highest Spearman correlation at every support budget. Its gain over zero-shot scoring becomes clear at $K = 8$ (0.516 versus 0.485). In contrast, METALIC and PRIMO-TTT fall below their zero-shot scores at $K = 4$ and 8; METALIC remains below at $K = 16$. This pattern is consistent with the difficulty of adapting a predictor from very few labels. Because assay descriptions state what is measured in natural language, the shared steering module in Eqs. 8–9 can score variants in unseen assays before any target labels are available. The module is trained across assays, while **ESM-2 and the 8B text encoder remain frozen** throughout training and inference. At test time, support labels fit only the residual in Eq. 14, a lighter-weight adaptation than fine-tuning the full neural predictor.

5.3 Assay Text and Cross-Assay Rankings

Assay-text intervention. To test whether the improvement comes from assay-relevant text, we hold the protein sequence, mutation, and trained model fixed while changing only the test-time assay information. We compare the matched description with a meaning-preserving rerender that retains the same evidence but changes its wording and order. A no-text control disables the text-conditioned update entirely. Two mismatch controls replace the assay-specific fields while retaining the target protein context: one draws descriptions from random assays, whereas the other favors donors from the same protein or a related selection type. The latter is a more conservative test of

Table 1: **Variant effect prediction.** Zero-shot Spearman correlation (ρ), AUC, MCC, and Top-10% Recall (R@10%), followed by few-shot Spearman correlation at six support budgets. PRIMO uses test-time training for $K > 0$; dashes indicate unavailable modes. The remaining few-shot metrics are reported in Appendix D.

Method	**Zero-shot** ($K = 0$)				**Few-shot** ρ (K)					
	ρ	AUC	MCC	R@10%	4	8	16	32	64	128
Label-free assay-agnostic methods										
ESM2-650M	0.384	0.721	0.312	0.212	–	–	–	–	–	–
VESM-650M	0.455	0.759	0.362	0.210	–	–	–	–	–	–
VESPA	0.425	0.746	0.350	0.200	–	–	–	–	–	–
MSA Transformer	0.395	0.729	0.325	0.216	–	–	–	–	–	–
PoET-200M	0.445	0.755	0.360	0.222	–	–	–	–	–	–
Per-assay supervised methods										
ProteinNPT	–	–	–	–	0.204	0.311	0.382	0.463	0.525	0.600
Cross-assay adaptive methods										
METALIC	0.340	0.699	0.279	0.182	0.226	0.269	0.323	0.407	0.481	0.551
PRIMO	0.425	0.738	0.338	0.179	0.364	0.393	0.438	0.484	0.535	0.557
LAVEP	**0.485**	**0.774**	**0.379**	**0.245**	**0.487**	**0.516**	**0.534**	**0.552**	**0.583**	**0.610**

whether the model distinguishes related assays. Figure 2A shows that paraphrasing barely changes performance, while removing conditioning or assigning unrelated text causes larger losses. Related-assay mismatches have a smaller effect. These results suggest that the model uses assay-relevant information rather than relying on a particular phrasing; the no-text control also indicates that the conditioning pathway contributes to prediction.

Multi-assay variant effect prediction. Table 1 shows that LAVEP improves general variant effect prediction on unseen assays, and the text intervention indicates that assay-relevant descriptions contribute to this gain. We next ask a more specific question: can the description guide the model to rank the *same variants* differently when the measured property changes? Only two proteins in the held-out test set have multiple eligible assays: KCNE1 and RL40A. To broaden the range of phenotype contrasts, we also include two additional benchmark pairs: NUDT15 and LDLR.

For variants shared by assays a and b, we change the assay description while keeping the variants fixed, and ask whether the resulting change in predicted rankings agrees with the experimentally observed change. To quantify this agreement, let $\widetilde{s}$ and $\widetilde{y}$ denote within-assay percentile ranks of predicted and measured effects, respectively. We define

$$\rho_{\Delta}^{(a,b)} = \mathrm{Spearman}(\widetilde{s}_{\cdot,a} - \widetilde{s}_{\cdot,b}, \widetilde{y}_{\cdot,a} - \widetilde{y}_{\cdot,b}) \,, \tag{17}$$

Higher ρ_{Δ} means that the model better captures how the effect of a shared variant changes between assays. Pairwise results are averaged within each protein, then across the four proteins.

Figure 2B shows a clear zero-shot advantage for LAVEP. Across the evaluated label budgets, it also achieves the strongest overall performance. Together with Table 1 and the text-intervention experiment, these results suggest that language guidance improves general prediction while also steering the ranking of identical variants toward the property measured by each assay.

Assay text versus target labels. We next ask how many target-assay labels are needed to close the gap provided by the description. Across 33 unseen assays, we compare text-conditioned zero-shot predictions with the frozen PLM followed by the same few-shot kernel correction, using identical support/query splits. Figure 2C shows a zero-shot text advantage of 0.079 Spearman at four labels and 0.042 at 16 labels. The difference is small at 32 labels; with 64 or more, the no-text few-shot model performs better. Text conditioning remains useful after adaptation: adding it to the few-shot model improves performance at every tested budget.

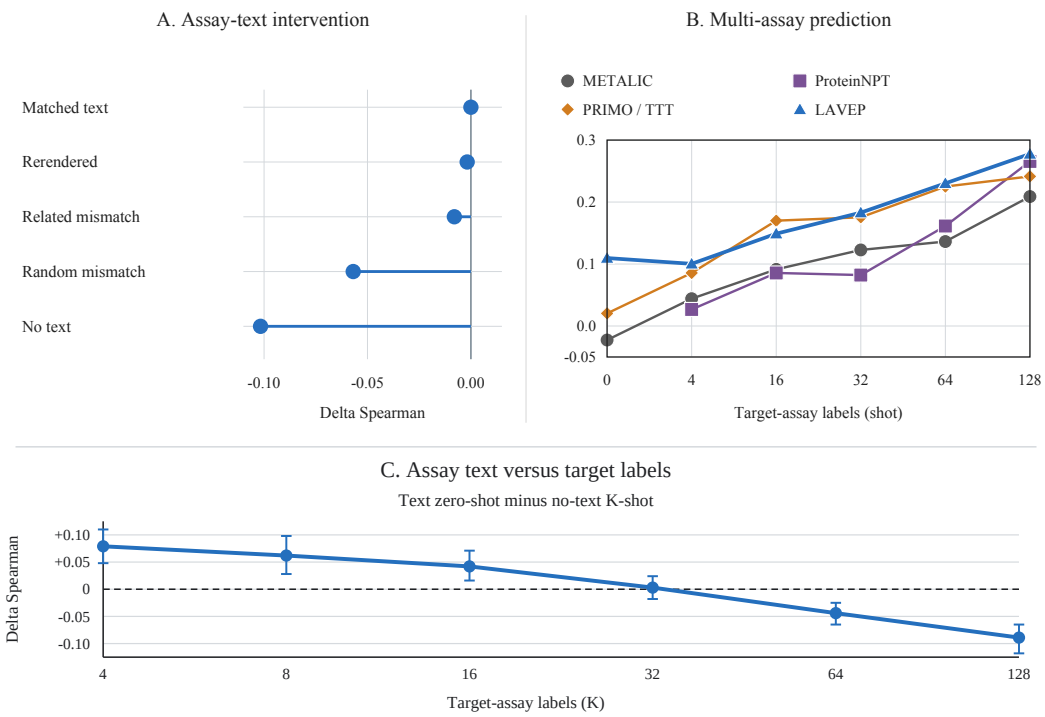

Figure 2: **Assay text and cross-assay ranking.** (A) Change in function-balanced Spearman correlation relative to matched assay text; the no-text condition disables text conditioning, while the two mismatch conditions retain target protein context. (B) Unweighted mean ρ_Δ across KCNE1, RL40A, NUDT15, and LDLR by label budget; ProteinNPT has no zero-shot mode. (C) Difference in function-balanced Spearman correlation between text-conditioned zero-shot LAVEP and the no-text frozen PLM with K-shot kernel correction across 33 unseen assays. Points use matched query variants and three support splits; bars show 95% protein-bootstrap intervals. Positive values favor assay text. Exact intervention values and per-protein ρ_Δ results are provided in Appendix D and Appendix D.3, respectively.

6 Conclusion and Limitations

Conclusion. Protein variant effects depend on what an assay measures, yet a sequence-only score cannot distinguish different experimental objectives for the same protein. We presented LAVEP, which uses natural-language assay descriptions as an open-vocabulary task specification for a frozen protein language model. By interacting semantic fields with mutation-specific protein context, it steers a shared sequence prior toward assay-conditioned rankings without requiring target-assay labels; a kernel residual supports adaptation when labels become available. Our analysis formalizes the ranking ceiling faced by any shared prediction when assays disagree and shows how assay-conditioned corrections can overcome it. On unseen ProteinGym assays, LAVEP improves zero-shot prediction across complementary measures and remains effective with limited support labels. Text interventions and comparisons of shared variants across assays further indicate that language guides rankings toward the measured property. These findings show the potential of language-guided protein language models to adapt general sequence priors to new experimental objectives.

Limitations. Assay descriptions provide an interpretable task specification but may not capture all biological and technical factors that shape variant effects, limiting language-based adaptation when key determinants are missing. Moreover, ProteinGym is a retrospective benchmark, and the generalization of language-guided prediction to prospective protein engineering campaigns remains unclear. Finally, this study is computational, and prospective wet-lab validation is needed to assess its practical utility.

AI USE STATEMENT

Generative Al tools were used to assist with language editing, literature searches, research discussions, and code development. All AI-assisted text, code, and analyses were critically reviewed and verified against the underlying sources, data, and implementations. All methodological and interpretive decisions were made by the authors, who take full responsibility for the final content of the work.

ETHICS STATEMENT

This is a computational study using publicly available deep mutational scanning data from ProteinGym and MaveDB. We did not collect data from human participants or conduct new animal or laboratory experiments. Variant-effect predictions may be unreliable outside the evaluated assays and could be misused to prioritize harmful mutations. We present them as research results, not as clinical evidence or experimental validation. Applications involving pathogens, clinical interpretation, or protein engineering should undergo appropriate review and independent experimental validation.

REPRODUCIBILITY STATEMENT

Code and data are available at `https://anonymous.4open.science/r/LAVEP-7497`. The main paper describes the model and training procedure. The appendix provides the assay splits and evaluation protocol (Appendix C.2), implementation settings (Appendix C.3), and proofs (Appendix B).

REFERENCES

Jacob Beck, Shikha Surana, Manus McAuliffe, Oliver Bent, Thomas D. Barrett, Juan Jose Garau-Luis, and Paul Duckworth. METALIC: Meta-learning in-context with protein language models. In *International Conference on Learning Representations*, 2025.

Melissa A. Chiasson, Nathan J. Rollins, Jason J. Stephany, Katherine A. Sitko, Kenneth A. Matreyek, Marta Verby, Song Sun, Frederick P. Roth, Daniel DeSloover, Debora S. Marks, Allan E. Rettie, and Douglas M. Fowler. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife*, 9:e58026, 2020.

Tuan Dinh, Seon-Kyeong Jang, Noah Zaitlen, and Vasilis Ntranos. Compressing the collective knowledge of ESM into a single protein language model. *Nature Methods*, 23:772–784, 2026.

Xiao Fei, Michail Chatzianastasis, Sarah Almeida Carneiro, Hadi Abdine, Lawrence P. Petalidis, and Michalis Vazirgiannis. Prot2Text-V2: Protein function prediction with multimodal contrastive alignment. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Douglas M. Fowler, Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence–function relationships. *Nature Methods*, 7(9):741–746, 2010.

Ruihan Guo, Chaoran Cheng, Zhanghan Ni, Neil He, Bangji Yang, and Ge Liu. MutAtlas: A PDB-wide energy-guided atlas of protein mutation effects. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306. PMLR, 2026.

Xiao-Yu Guo, Yi-Fan Li, Yuan Liu, Xiaoyong Pan, and Hong-Bin Shen. Ab-initio amino acid sequence design from protein text description with ProtDAT. *Nature Communications*, 16:10544, 2025. doi: 10.1038/s41467-025-65562-w.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, 40(7):1114–1122, 2022.

Ala Jararweh, Oladimeji Macaulay, David Arredondo, Yue Hu, Luis E. Tafoya, Kushal Virupakshappa, and Avinash Sahu. Protein2Text: Resampling mechanism to translate protein sequences into human-interpretable text. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 918–937. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-industry.68.

Élise Laine, Yasaman Karami, and Alessandra Carbone. GEMME: A simple and fast global epistatic model predicting mutational effects. *Molecular Biology and Evolution*, 36(11):2604–2619, 2019.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Nuowei Liu, Jiahao Kuang, Yanting Liu, Tao Ji, Changzhi Sun, Man Lan, and Yuanbin Wu. Protein design with dynamic protein vocabulary. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *Nature Machine Intelligence*, 7:580–591, 2025b. doi: 10.1038/s42256-025-01011-z.

Yizhen Luo, Zikun Nie, Massimo Hong, Suyuan Zhao, Hao Zhou, and Zaiqing Nie. MutaPLM: Protein language modeling for mutation explanation and engineering. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024.

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human Genetics*, 141(10):1629–1647, 2022.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303. Curran Associates, Inc., 2021.

Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S. Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):e112, 2014.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N. Gomez, Debora S. Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16990–17017. PMLR, 2022.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023a.

Pascal Notin, Ruben Weitzman, Debora S. Marks, and Yarin Gal. ProteinNPT: Improving protein property prediction and design with non-parametric transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33529–33563. Curran Associates, Inc., 2023b.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8844–8856. PMLR, 2021.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, Hongyuan Lu, and Fajie Yuan. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02836-0.

Felix Teufel, Aaron W. Kollasch, Yining Huang, Ole Winther, Kevin K. Yang, Pascal Notin, and Debora S. Marks. Few-shot protein fitness prediction via in-context learning and test-time training. In *NeurIPS 2025 Workshop on AI for Science: The Reach and Limits of AI for Scientific Discovery*, 2025.

Timothy F. Truong Jr. and Tristan Bepler. PoET: A generative model of protein families as sequences-of-sequences. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77379–77415. Curran Associates, Inc., 2023.

Ruben Weitzman, Peter Mørch Groth, Lood van Niekerk, Aoi Otani, Yarin Gal, Debora Susan Marks, and Pascal Notin. Protriever: End-to-end differentiable protein homology search for fitness prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 66397–66418. PMLR, 2025.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. ProtST: Multi-modality learning of protein sequences and biomedical texts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38749–38767. PMLR, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, Jike Wang, Qiang Zhang, and Huajun Chen. Advancing biomolecular understanding and design following human instructions. *Nature Machine Intelligence*, 7(7):1154–1167, 2025. doi: 10.1038/s42256-025-01064-0.

A Assay Description Fields

Each field in Eq. 3 answers a different question about the assay. **Experimental context** gives an overall account of the experiment: the target, biological system, assay objective, and measurement approach. It can summarize information that is expanded in the other fields. **Measured readout** identifies the phenotype or signal being observed and how it is measured, such as binding, expression, activity, or growth. **Score semantics** explains the reported phenotype score, including its direction: what a larger or smaller value means for the measured phenotype. It describes the score convention without revealing any variant-level measurements. **Experimental conditions** specifies the setting in which the readout is measured, including an interaction partner or treatment and, when reported, its concentration, temperature, or pH. Missing quantitative details are left unspecified. **Protein information** identifies the assayed protein, organism, mutated region, and relevant functional context. In particular, the readout states *what is measured*, while score semantics states *how the reported value is interpreted*.

The assay descriptions draw on ProteinGym assay metadata, with UniProt protein annotations and source-study titles used when available. They do not include variant-level scores or labels.

B Proofs

B.1 Proof of the assay-agnostic ranking ceiling

Theorem 1 (Assay-Agnostic Ranking Ceiling). *For any assay-agnostic predictor with centered and normalized predicted rank vector $\widehat{r}^{ab}$,*

$$\frac{\rho(\widehat{r}^{ab}, r_a^{ab}) + \rho(\widehat{r}^{ab}, r_b^{ab})}{2} \le \sqrt{\frac{1+\rho_{ab}}{2}}.$$

Proof. Spearman correlation is the Pearson correlation between rank vectors. Since r_a^{ab}, r_b^{ab}, and $\widehat{r}^{ab}$ are centered and normalized, we may therefore write $\rho(\widehat{r}^{ab}, r_c^{ab}) = (\widehat{r}^{ab})^\top r_c^{ab}$ for $c \in \{a, b\}$. It follows that

$$\begin{aligned}\frac{\rho(\widehat{r}^{ab}, r_a^{ab}) + \rho(\widehat{r}^{ab}, r_b^{ab})}{2} &= \frac{(\widehat{r}^{ab})^\top (r_a^{ab} + r_b^{ab})}{2} \\ &\le \frac{\|\widehat{r}^{ab}\|_2 \|r_a^{ab} + r_b^{ab}\|_2}{2} = \frac{\|r_a^{ab} + r_b^{ab}\|_2}{2},\end{aligned}$$

where the inequality is Cauchy–Schwarz and the last equality uses $\|\widehat{r}^{ab}\|_2 = 1$. The remaining norm is determined by the agreement between the two assay rankings:

$$\begin{aligned}\|r_a^{ab} + r_b^{ab}\|_2^2 &= \|r_a^{ab}\|_2^2 + \|r_b^{ab}\|_2^2 + 2(r_a^{ab})^\top r_b^{ab} \\ &= 2 + 2\rho_{ab}.\end{aligned}$$

Substitution gives $\|r_a^{ab} + r_b^{ab}\|_2/2 = \sqrt{(1+\rho_{ab})/2}$, which proves the stated ceiling. □

B.2 Proof of assay-conditioned ranking resolution

Corollary 1 (Assay-Conditioned Ranking Resolution). *Suppose assays a and b have strict rankings on $\mathcal{O}_{ab}$. Let $\mathcal{P}_c$ contain every pair (r, s) preferred by assay $c \in \{a, b\}$, and let $C_c(v) = s_c^{\mathrm{cond}}(v) - s^{\mathrm{PLM}}(v)$. If every $(r, s) \in \mathcal{P}_c$ satisfies*

$$s^{\mathrm{PLM}}(r) - s^{\mathrm{PLM}}(s) + \alpha\big(C_c(r) - C_c(s)\big) > 0,$$

then the zero-shot scores recover both rankings exactly. Their average Spearman correlation is one and, when $\rho_{ab} < 1$, exceeds the assay-agnostic ceiling in Theorem 1.

Proof. By Eq. 13, each preferred pair satisfies

$$s_c^{\mathrm{ZS}}(r) - s_c^{\mathrm{ZS}}(s) = s^{\mathrm{PLM}}(r) - s^{\mathrm{PLM}}(s) + \alpha\big(C_c(r) - C_c(s)\big) > 0.$$

Because $\mathcal{P}_c$ contains every preferred pair, the predicted ranking matches assay c's ranking. Both Spearman correlations are therefore one. When $\rho_{ab} < 1$, Theorem 1 bounds every assay-agnostic average below one, proving the strict improvement. □

C Experimental Details

C.1 Datasets and Data Processing

Our main benchmark is the ProteinGym DMS substitutions collection, comprising 217 assays, 186 proteins, and 2,465,767 variant measurements. It contains both single- and multiple-substitution variants. The split is made at the assay level, with homologous proteins kept in the same partition. Proteins are connected when they share a UniProt identifier or meet the MMseqs2 sequence-similarity criterion of at least 30% identity and 80% coverage (E-value at most 10^{-5}). Connected homology groups are assigned as units toward a 70/15/15 training/validation/test ratio. The result contains 152 training, 32 validation, and 33 test assays; no homology group crosses partitions. Model selection uses validation assays. For few-shot evaluation, support budgets are $K \in \{4, 16, 32, 64, 128\}$. Support sets are nested within each assay, and support variants are removed from that assay's query set before evaluation.

The controlled comparison of different properties on the same protein uses KCNE1 and RL40A assays from the held-out ProteinGym test set. It also uses two additional pairs obtained from MaveDB: NUDT15 protein abundance versus thiopurine activity, and LDLR uptake versus cell-surface abundance. For these added pairs, we retain missense substitutions with finite scores, match variants across the two assays of each protein, and compare only the shared variants.

C.2 Evaluation Metrics and Aggregation

All four metrics are first computed separately within each test assay, using its measured scores y_i and predictions s_i, with higher values oriented toward a stronger measured effect. Let $b_i \in \{0, 1\}$ denote ProteinGym's binary experimental label.

Spearman correlation. Let R_i^y and R_i^s denote the ranks of y_i and s_i, respectively, with average ranks assigned to ties; $\overline{R}^y$ and $\overline{R}^s$ are their respective means. Spearman correlation is the Pearson correlation of these rank vectors:

$$\rho = \frac{\sum_i (R_i^y - \overline{R}^y)(R_i^s - \overline{R}^s)}{\sqrt{\sum_i (R_i^y - \overline{R}^y)^2 \sum_i (R_i^s - \overline{R}^s)^2}}.$$

AUC. Let $P = \{i : b_i = 1\}$ and $N = \{j : b_j = 0\}$. The area under the ROC curve is the fraction of positive–negative pairs correctly ordered by the predictions, giving half credit to tied predictions. Here $\mathbf{1}[\cdot]$ is one when its condition holds and zero otherwise:

$$\mathrm{AUC} = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} \left(\mathbf{1}[s_i > s_j] + \tfrac{1}{2} \mathbf{1}[s_i = s_j] \right).$$

MCC. Predictions are binarized at their within-assay median, $\widehat{b}_i = \mathbf{1}[s_i \geq \mathrm{median}(s)]$. For true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the Matthews correlation coefficient is

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

Top-10% Recall. Let q_y and q_s be the within-assay 90th percentiles of measured and predicted scores. Including values tied at either threshold, the recall is

$$\mathrm{Recall}_{10\%} = \frac{\sum_i \mathbf{1}[y_i \geq q_y]\, \mathbf{1}[s_i \geq q_s]}{\sum_i \mathbf{1}[y_i \geq q_y]}.$$

Metrics that are undefined for an assay, such as AUC with only one binary class, are excluded from that metric's aggregate.

Following the ProteinGym evaluation script (Notin et al., 2023a), we aggregate each metric by first averaging assay-level values for the same protein within a phenotype category, then averaging across proteins in that category, and finally taking an unweighted mean across the available categories: activity, binding, expression, organismal fitness, and stability. In few-shot evaluation, metrics are computed on query variants after removing the labeled support variants.

C.3 Implementation Details

Table 2 summarizes the configuration used for the reported results. The field weights follow the order in Eq. 3. The checkpoint and hyperparameters were selected on validation assays.

Table 2: **Key implementation settings for LAVEP.**

Setting	Value
Encoders and assay-conditioning module	
Protein / text encoders	Frozen ESM-2 650M / Qwen3-Embedding-8B
Trainable parameters	28.6 million
Hidden / bottleneck dimensions	512 / 128
Attention heads / text dropout	8 / 0.1
Training	
Training seed / duration	2026 / 10 epochs, 500 steps per epoch
Batch size / learning rate	256 variant pairs / 2×10^{-4}
Weight decay / ranking temperature	10^{-4} / 1.0
KL regularization weight	0.05
Selected checkpoint	Epoch 7
Zero-shot inference	
Interpolation weight α	0.8
Field-centering scale γ / interaction ξ	4.0 / -0.3375
Field weights, multi-assay proteins	$(1, 2, 2, 1.5, 1)$
Field weights, single-assay proteins	$(0.75, 1.75, 1.75, 1.25, 0.75)$
Few-shot adaptation	
KRR temperature / ridge coefficient	0.2 / 0.1
Residual mixing / cross-assay weight	1.0 / 0.5
Zero-shot anchor	0.35
Schedule reference K_0 / power	10 / 6

Assay-contrast adjustment. The first step is Eq. 12. To limit repeated emphasis on the same assay differences, let $\overline{u}_{p,j}$ be the mean of $\widehat{u}_{b,j}$ over $b \in \mathcal{A}_{p,j}$ and set $\delta_{a,j} = \widehat{u}_{a,j} - \overline{u}_{p,j}$. We then adjust the context, readout, and score-semantics fields:

$$\begin{aligned} \widetilde{u}_{a,\text{context}} &= \mathcal{N}_{\widehat{u}_{a,\text{context}}}\left(\widehat{u}_{a,\text{context}} + \frac{\xi}{2}(\delta_{a,\text{readout}} + \delta_{a,\text{score}})\right), \\ \widetilde{u}_{a,j} &= \mathcal{N}_{\widehat{u}_{a,j}}\left(\widehat{u}_{a,j} + \xi\delta_{a,\text{context}}\right), \qquad j \in \{\text{readout}, \text{score}\}. \end{aligned} \tag{18}$$

The negative coefficient ξ reduces overlap between these fields while $\mathcal{N}$ preserves their norms. We set $\widetilde{u}_{a,\text{condition}} = \widehat{u}_{a,\text{condition}}$ and $\widetilde{u}_{a,\text{protein}} = u_{a,\text{protein}}$; for a protein with only one available assay, all fields retain their original $u_{a,j}$.

Few-shot kernel residual. For each assay, a single seeded permutation defines nested support sets of $K \in \{4, 16, 32, 64, 128\}$ labeled variants. We standardize the assay's zero-shot scores using the mean and standard deviation of its support scores, and standardize the ranks of its support labels within that assay. The resulting $z_a(v)$ and $q_{a,s}$ give the residual target in Eq. 14.

The kernel feature $\phi_a(v)$ is built from the model's mutation-specific representation and an effect representation formed by multiplying it elementwise by the ESM-2 decoder-weight difference between the mutant and wild-type amino acids. For a variant with several substitutions, we average the corresponding position-wise representations. We normalize each of the two resulting vectors, concatenate them, and normalize the concatenation. Thus $\|\phi_a(v)\|_2 = 1$. For a query variant v from assay a and support variant s from assay b, the kernel is

$$k_K\big((a, v), (b, s)\big) = \omega_{ab}(K) \exp\left(\frac{\phi_a(v)^\top \phi_b(s) - 1}{\tau}\right), \qquad \omega_{ab}(K) = \begin{cases} 1, & a = b, \\ w_K, & a \neq b, \end{cases} \tag{19}$$

Here τ controls how rapidly kernel similarity decreases as the normalized features diverge, while $\omega_{ab}(K)$ reduces the contribution of support variants from another assay. We pool support variants

from all available assays of the same protein. Writing S for this pooled set and ϵ_S for its assay-wise residual targets, KRR predicts

$$r_a^K(v) = k_K\big((a,v),S\big)^\top(\mathbf{K}_{SS} + \lambda I)^{-1}\epsilon_S. \tag{20}$$

The schedule is

$$g_K = \frac{(K/K_0)^p}{1+(K/K_0)^p}, \qquad \eta_K = 1-(1-\eta_\infty)g_K, \qquad w_K = w_\infty g_K, \tag{21}$$

K_0 sets the support budget at which g_K reaches one half, and p controls the sharpness of this transition. The parameters η_∞ and w_∞ are the limiting zero-shot anchor and cross-assay weight. Consequently, the prediction starts close to the standardized zero-shot score at small K and gradually gives more weight to the fitted residual and support variants from other assays of the same protein. The neural model is not updated during this step.

D Additional Experimental Results

D.1 Module Ablation

We assess protein-guided text grounding, text-guided sequence conditioning, and residual gating using controls that remove each component, replace cross-attention with late fusion, or omit the numeric position coordinate. Each ablated architecture is retrained independently under the same protocol as the full model.

Table 3 shows lower Spearman correlation for every ablation, supporting the contribution of each component. Because the reduced architectures are retrained, their remaining pathways can adapt and may partially compensate for a missing component, which could explain the modest decreases. Together with the assay-text intervention, this suggests that assay-relevant text is a major source of the performance gain, while the architectural components help the model use that information effectively.

Table 3: **Module ablation.** Single-seed function-balanced Spearman results for the full LAVEP model and five independently retrained architectural controls.

Variant	Text grounding $T \leftarrow Z$	Sequence conditioning $Z \leftarrow T$	Residual gate	Spearman ρ
Full model	✓	✓	✓	**0.4855**
w/o text grounding	×	✓	✓	<u>0.4820</u>
w/o sequence conditioning	✓	×	✓	0.4755
Late fusion	×	×	✓	0.4782
Ungated residual	✓	✓	×	0.4751
w/o numeric coordinate	✓	✓	✓	0.4721

D.2 Complete Few-Shot Results

Table 4 reports all four metrics for the four few-shot methods compared in Table 1.

D.3 Complete Multi-Assay Results

Table 5 reports the complete multi-assay ρ_Δ results for all compared methods and available label budgets, including the per-protein values underlying Figure 2B.

Table 4: **Complete few-shot results on the held-out ProteinGym assays.** Each row reports four metrics from the same evaluation episode, measured on query variants after excluding the support set.

Method	K	Spearman ρ	AUC	MCC	Top-10% Recall
ProteinNPT	4	0.204	0.615	0.168	0.167
	8	0.311	0.679	0.250	0.184
	16	0.382	0.721	0.298	0.221
	32	0.463	0.755	0.375	0.246
	64	0.525	0.790	0.417	0.264
	128	0.600	0.824	0.479	0.354
METALIC	4	0.226	0.625	0.181	0.133
	8	0.269	0.646	0.199	0.166
	16	0.323	0.680	0.248	0.187
	32	0.407	0.720	0.319	0.202
	64	0.481	0.760	0.366	0.253
	128	0.551	0.793	0.425	0.332
PRIMO-TTT	4	0.364	0.711	0.302	0.190
	8	0.393	0.722	0.310	0.220
	16	0.438	0.740	0.340	0.225
	32	0.484	0.765	0.380	0.230
	64	0.535	0.785	0.405	0.273
	128	0.557	0.804	0.435	0.290
LAVEP	4	0.487	0.774	0.383	0.236
	8	0.516	0.787	0.404	0.255
	16	0.534	0.794	0.414	0.265
	32	0.552	0.799	0.423	0.271
	64	0.583	0.816	0.450	0.262
	128	0.610	0.829	0.475	0.321

Table 5: **Cross-assay rank-change agreement for each protein.** Each protein contributes equally to the macro mean.

Method	K	KCNE1	RL40A	NUDT15	LDLR	Mean ρ_Δ
ProteinNPT	4	0.067	−0.020	0.044	0.016	0.027
	16	0.103	0.048	0.103	0.088	0.086
	32	0.196	0.061	0.086	−0.014	0.082
	64	0.386	0.140	0.155	−0.036	0.161
	128	0.476	0.271	0.319	−0.003	0.266
METALIC	0	−0.020	−0.040	−0.066	0.036	−0.023
	4	0.056	0.066	0.056	−0.002	0.044
	16	0.149	0.146	0.096	−0.025	0.091
	32	0.180	0.171	0.129	0.011	0.123
	64	0.301	0.174	0.066	0.005	0.136
	128	0.476	0.227	0.111	0.023	0.209
PRIMO / PRIMO-TTT	0	0.070	0.092	−0.075	−0.006	0.020
	4	0.100	0.047	0.123	0.072	0.086
	16	0.309	0.155	0.204	0.012	0.170
	32	0.284	0.191	0.176	0.050	0.175
	64	0.370	0.217	0.245	0.068	0.225
	128	0.487	0.184	0.200	0.094	0.241
LAVEP	0	0.109	0.043	0.276	0.010	0.110
	4	0.075	0.061	0.269	−0.004	0.100
	16	0.185	0.141	0.228	0.042	0.149
	32	0.259	0.173	0.247	0.052	0.183
	64	0.320	0.217	0.315	0.069	0.230
	128	0.403	0.311	0.291	0.106	0.278