Latent Encoding of Epistasis in Protein Language Models

Kun Hyung Roh [1] **Vincent Yip** [2] **Sunghoon Rho** [3]

Abstract

Combinations of mutations can create protein functions that are not predictable from single substitutions, known as epistasis, yet most protein language models are evaluated on raw variant scores that mix additive effects and residue-level coupling. We introduce a framework that separates measured protein landscapes into additive single-mutant effects, a monotone global response, and a residual term that captures specific epistasis. A lightweight pairwise model trained on frozen diffusion language model representations learns this residual coupling signal from measured multi-mutant data. We evaluate the approach on 89,706 measured stability double mutants across 149 protein domains and 85,135 measured multi-mutants across 11 functional assays spanning binding and fluorescence. Zero-shot protein language model scores capture limited stability coupling and little functional coupling, whereas the trained representation model recovers functional epistasis across all 11 assays, with median held-out Spearman improving from 0.008 to 0.292. The resulting model selects gain-of-function double mutants from additive-matched candidate pools, where single-mutant selection has no ranking signal. These results define specific epistasis as a measurable target for protein generative models and provide a practical route toward designing combinatorial variants with experimentally useful function.

[1]Department of Bioengineering and Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA [2]Department of Computer Science, Stanford University, Stanford, CA 94305, USA [3]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. Correspondence to: Kun Hyung Roh <rohk@stanford.edu>, Vincent Yip <vyip23@stanford.edu>, Sunghoon Rho <srho@stanford.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Diffusion language models can sample plausible protein sequences, evaluate many variants, and support edit based optimization (Wang et al., 2024). DPLM-Evo extends this toward explicit evolutionary editing (Wang et al., 2026). However, scientific objective in a laboratory is more demanding, as in order to improve a protein's measured property such as binding, fluorescence, catalytic activity, non-intuitive multiple mutations must be made. A generative prior is useful only to the extent that sequence compatibility aligns with the assay being optimized.

Deep mutational scanning (DMS) gives a direct local measurement of this landscape (Notin et al., 2023). Megascale gives a complementary large stability measurement resource (Tsuboyama et al., 2023). Single mutant maps are increasingly available, but dense multi mutant maps remain expensive because the number of combinations grows exponentially with the number of edited positions. Then, we must ask if given single mutant measurements and a protein representation, a model could predict which combinations depart from the additive expectation in a direction that improves the measured function.

The raw multi mutant score gives an incomplete target for this question because a model can perform well on such a score while only recovering the additive sum of single mutant measurements. Recent analyses and valid criticisms show that additive baselines are extremely difficult to beat in stability and multi mutant prediction benchmarks (Visani et al., 2026). Moreover, the residual after additivity still mixes two distinct phenomena. One is a global nonlinear response of the assay scale, and the other is a specific interaction between residues. A meaningful epistasis benchmark must separate these two terms.

In our framework, for every protein, the measured multi mutant function is decomposed into three components: 1) additive sum of measured single mutant effects, 2) the best monotone response curve from the additive score to the measured readout, 3) the residual after these two layers, the "residual specific epistasis". Indeed, if a zero shot model already encodes functional coupling, the discrete second difference of its sequence score should correlate with the specific residual.

A trained coupling head on DPLM hidden states recovers part of this missing signal on measured function assays. On GB1 binding, the trained head improves held out Spearman from 0.02 for zero shot DPLM to 0.37. When used for design, it recovers high binding double mutants from an additively matched pool where additive selection has no ranking power.

The contributions are as follows:

- We define a three layer decomposition for multi mutant protein data that separates additive single mutant effects, a monotone global response, and specific residue coupling.
- We prove that the residual left by the monotone projection is inaccessible to additive models and to monotone transformations of the additive score.
- We show that this decomposition changes the interpretation of apparent epistasis in both stability and function data.
- We train a representation based coupling model that predicts functional specific epistasis on eleven measured assays and for two different language model backbones, and we show it beats both the zero shot score and a supervised identity model that uses no language model at all.

2. Related Work

Protein language models learn constraints from large sequence databases and support mutation effect prediction, representation learning, and generation (Lin et al., 2023). DPLM moves this family toward diffusion based sequence modeling (Wang et al., 2024). DPLM Evo moves the family further toward explicit evolutionary editing (Wang et al., 2026). These methods are powerful priors over protein like sequences. Our work asks when this prior predicts measured function specific residue coupling.

DMS benchmarks such as ProteinGym and Megascale make it possible to test this question on real measurements (Notin et al., 2023). Megascale measures folding stability at large scale (Tsuboyama et al., 2023). ProteinGym aggregates assays for fluorescence, binding, activity, abundance, and related readouts. These datasets differ from natural sequence corpora because they define experimental objectives around a particular protein and assay.

Epistasis has a long history in genotype phenotype maps (Poelwijk et al., 2007). Higher order interaction structure has also been studied in nonlinear genotype phenotype maps (Sailer et al., 2017). A key distinction is global epistasis versus specific epistasis, however, as global epistasis appears when a nonlinear readout maps an additive latent trait to a measured phenotype, the regime modeled by MAVE-NN (Tareen et al., 2022). Specific epistasis remains after this global mapping and reflects interaction between mutations.

The GB1 landscape is a central test case for functional epistasis (Olson et al., 2014). The Wu GB1 landscape gives a dense multi site view of adaptation (Wu et al., 2016). GFP provides a fluorescence landscape (Sarkisyan et al., 2016). GRB2 and PABP provide complementary binding assays (Faure et al., 2022). PABP is also represented by a deep mutational scan of an RNA recognition motif (Melamed et al., 2013). We use these measured landscapes to test whether representation based models can move beyond the additive baseline.

3. Problem Setup and the Three Layer Decomposition

Let x_0 be a wildtype protein sequence. A mutation set S contains substitutions of the form (p, a), where position p is changed to amino acid a. The corresponding variant is x_S. For protein g, the assay returns a scalar measurement $y_g(S)$. Higher values are better after the standard orientation used by the assay.

For each measured single mutation, define the single effect

$$b_g(p, a) = y_g(\{(p, a)\}) - y_g(\emptyset). \tag{1}$$

The additive prediction for a multi mutant S is

$$A_g(S) = \sum_{(p,a)\in S} b_g(p, a). \tag{2}$$

The raw residual $y_g(S) - A_g(S)$ conrains true residue coupling, on top of global nonlinearities in the assay scale. A brightness assay can saturate, or a binding assay can have a floor, and so on.

We therefore define a monotone global response link. Given measured multi mutant pairs $\{A_i, y_i\}_{i=1}^{n_g}$ for protein g, let

$$\widehat{h}_g = \arg\min_{h\in\mathcal{G}_g} \sum_{i=1}^{n_g} \left(y_i - h(A_i)\right)^2, \tag{3}$$

where $\mathcal{G}_g$ is the set of monotone functions on the observed additive scores. The calibrated global prediction is

$$G_i = \widehat{h}_g(A_i). \tag{4}$$

The specific epistasis target is

$$\varepsilon_i = y_i - G_i. \tag{5}$$

This target is the part of measured function that remains after the additive score and the best monotone response to that score have been removed.

4. Method

4.1. Three Layer Decomposition

The decomposition has three layers: The additive layer asks how far the sum of single mutant measurements can go. The global layer asks how much of the remaining error is explained by a monotone calibration curve. The specific layer is the residual that the two earlier layers cannot explain.

The variance explained by the additive layer is

$$R^2_{\text{add}} = 1 - \frac{\sum_i (y_i - A_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{6}$$

The variance explained by the global layer is

$$R^2_{\text{glob}} = 1 - \frac{\sum_i (y_i - G_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{7}$$

The specific share is

$$Q_{\text{spec}} = \frac{\widehat{\text{Var}}(\varepsilon)}{\widehat{\text{Var}}(y)}. \tag{8}$$

Large Q_{spec} means that a learned model has real headroom beyond additive and global response modeling.

4.2. Zero Shot DPLM Coupling

Let $\ell_\phi(x)$ be the DPLM pseudo log likelihood of a sequence. For a double mutant with single mutants x_i and x_j and double mutant x_{ij}, define the zero shot DPLM coupling score as the discrete second difference

$$Z_\phi(i,j) = \ell_\phi(x_{ij}) - \ell_\phi(x_i) - \ell_\phi(x_j) + \ell_\phi(x_0). \tag{9}$$

If the model score is additive across the two mutations, this quantity is zero, and nonzero values measure whether DPLM assigns extra compatibility or incompatibility to the pair. We then compare Z_ϕ with ε within each protein as a zero shot test.

4.3. Model 1

Model 1 predicts specific epistasis from DPLM representations. For a double mutant at positions i and j, we extract frozen DPLM hidden states h_i and h_j from the wildtype sequence. We concatenate projected states, their elementwise product, amino acid identity embeddings for the wildtype and mutant residues, and the zero shot coupling score, giving $z_{ij} = [P_i h_i, P_j h_j, P_i h_i \odot P_j h_j, e_{aa}, Z_\phi(i,j)]$. The predictor is a small multilayer head $\widehat{\varepsilon}_{ij} = f_\theta(z_{ij})$, trained with the Huber objective

$$\mathcal{L}(\theta) = \sum_{(i,j)\in\mathcal{D}_{\text{train}}} \rho_\delta(f_\theta(z_{ij}) - \varepsilon_{ij}). \tag{10}$$

We evaluate two splits, where the held out double split tests whether the model can complete a partially measured combinatorial landscape at known positions, and the held out position split tests transfer to positions not used for training.

4.4. Model 2

Model 2 uses the learned residual in training.. For a candidate pair, the predicted measured function is $\widehat{y}_{ij} = \widehat{h}_g(A_{ij}) + \widehat{\varepsilon}_{ij}$. To show that the selection is not additive in disguise, we evaluate within narrow additive score bins. Inside such a bin, additive selection assigns nearly equal value to every variant. Any ranking skill inside the bin must come from the predicted specific residual.

4.5. Algorithmic Summary

Algorithm 1 gives the complete procedure used in the experiments, where the supervised target is constructed only after the additive single mutant map and the monotone assay response have been fitted. The representation model is therefore trained on the residue coupling left by those two lower dimensional explanations.

Algorithm 1 Specific Epistasis Learning And Additive Matched Design

Require: Wildtype x_0, measured singles, measured multi mutants, frozen protein language model ϕ
Ensure: Held out residual predictions and additive matched design rankings

```
Estimate single mutant effects b_g(p, a) from measured singles
Compute additive scores A_g(S) for each measured multi mutant using Equation (2)
Fit the monotone response ĥ_g by isotonic least squares using Equation (3)
Set the supervised residual target ε_i = y_i − ĥ_g(A_i)
for each train or test pair (i, j) do
    Compute Z_φ(i, j) from the discrete second difference of the PLM score
    Build pair features z_ij from PLM states, amino acid identities, and Z_φ(i, j)
end for
Train f_θ on training pairs by minimizing Σ ρ_δ(f_θ(z_ij) − ε_ij)
Evaluate held out pairs by Spearman correlation with measured ε_ij
Rank additive matched candidates by ĥ_g(A_ij) + f_θ(z_ij)
```

Table 1. Three layer decomposition of measured multi mutant landscapes. Rows report additive fit, monotone global response fit, and residual specific epistasis across stability and function datasets.

Data	Rows	Additive R^2	Global R^2	Specific share
Megascale	89,706	−0.91	0.72	0.28
GFP	11,954	−32.6	0.89	0.11
GB1 (Olson)	40,000	0.70	0.94	0.06
GB1 (Wu)	2,091	−3.43	0.55	0.45
GRB2	39,912	0.43	0.85	0.15
PABP	36,370	−5.50	0.82	0.19

5. Experiments

5.1. Datasets

The stability experiments use Megascale double mutants with measured $\Delta\Delta G$ values and measured constituent single mutants (Tsuboyama et al., 2023). The function experiments use five ProteinGym assays (Notin et al., 2023). They include GFP fluorescence (Sarkisyan et al., 2016), GB1 binding from Olson (Olson et al., 2014), GB1 binding from Wu (Wu et al., 2016), GRB2 binding (Faure et al., 2022), and PABP binding (Melamed et al., 2013).

The three layer decomposition in Table 1 is shown on a named subset for readability. The breadth test extends the held out double evaluation to eleven ProteinGym function assays. For the language model comparison we extract features from two backbones, DPLM-650M (Wang et al., 2024) and ESM-2-650M (Lin et al., 2023), using the same hidden states and the same pseudo log likelihood scoring for both, so the only difference between the two is the pretrained model.

5.2. Global Epistasis Explains Most Non Additivity

Across measured stability and function landscapes, much of the apparent failure of additivity reflects nonlinear assay response rather than residue-specific coupling. Briefly, in Table 1 after the monotone global link, variance explained rises sharply across both stability and function assays. The remaining residual is smaller but still substantial. In GB1 Wu, specific epistasis accounts for 45% of double mutant variance.

The same decomposition exposes a physically interpretable design signal in GB1. Across 149 stability domains, raw additive prediction frequently falls below the mean baseline, whereas monotone calibration recovers a median explained variance of 0.72 (Fig. 2A). In GB1 binding, variants above the additive diagonal represent gain-of-function combinations whose measured activity exceeds the sum of their constituent single mutations (Fig. 2C). The strongest recovered pair, G41 and V54, forms a contact in the folded GB1 structure at approximately 5.5 angstroms (Fig. 2B), linking the residual statistical signal to a local structural interaction.

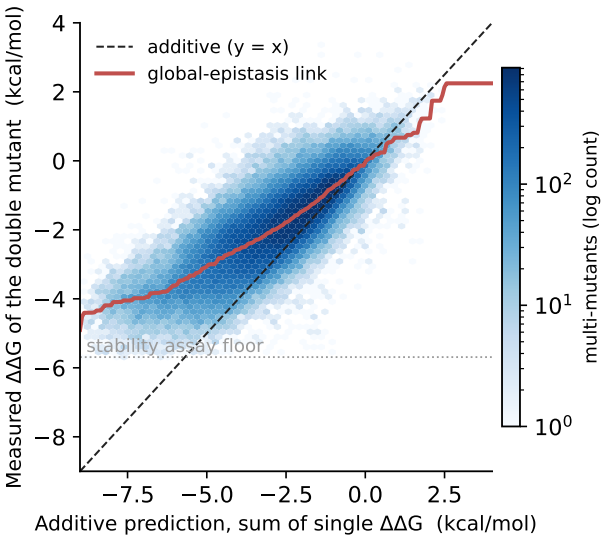

Figure 1. Monotone response calibration for measured stability. Additive double mutant predictions are mapped to measured stability by a fitted monotone global response link, leaving residual specific epistasis.

5.3. Functional Epistasis Is Invisible to Zero Shot Models

We next asked whether DPLM already encodes the specific residual. On Megascale stability, the zero shot second difference has a weak relationship with measured specific epistasis. Across 89,706 doubles from 149 proteins, the median per protein Spearman correlation is approximately 0.25. This suggests that natural sequence statistics contain some stability coupled residue compatibility.

The same test fails on measured function. In GB1 binding, zero shot DPLM reaches Spearman 0.02 on specific epistasis. In GFP fluorescence, it reaches 0.00. Importantly, this can be interpreted that a protein language model can know whether a pair looks natural without knowing whether the pair improves a laboratory assay.

5.4. A Trained Coupling Head Predicts Functional Epistasis

Table 2. Model 1 performance on held out functional epistasis prediction. Spearman correlations are reported for zero shot DPLM and trained coupling heads on the same measured variants.

METHOD	GB1	GFP
ZERO SHOT DPLM	0.02	0.00
MODEL 1 HELD OUT DOUBLES	**0.37**	**0.14**
MODEL 1 HELD OUT POSITIONS	0.13	0.09

Table 2 shows that a trained coupling head extracts functional information that the zero shot score misses. The held

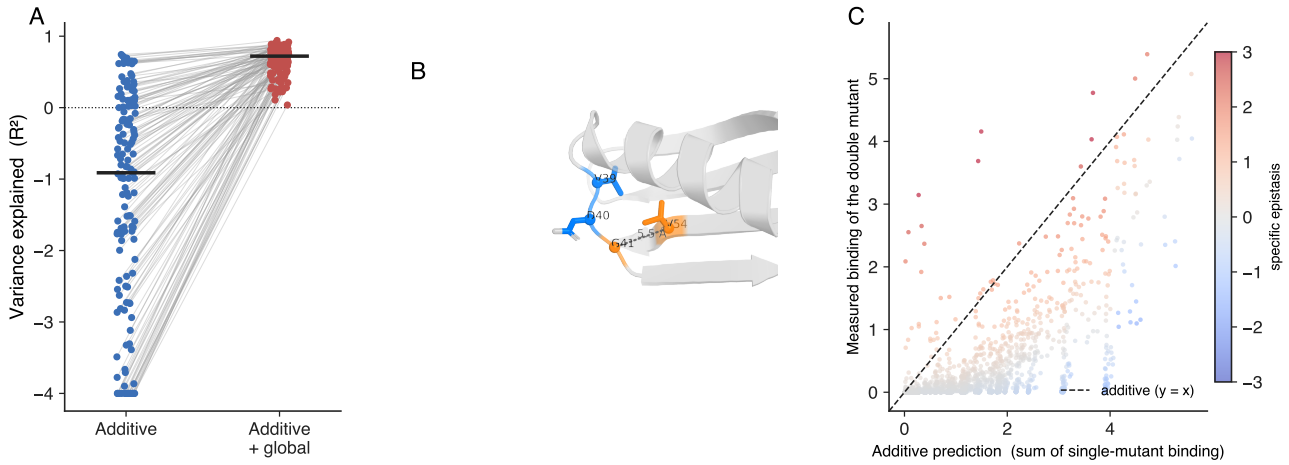

Figure 2. (A) Per protein variance explained on Megascale before and after the monotone global link, with median bars. Raw additivity can be negative because the assay scale is not the additive latent scale, and the link recovers most of it. (B) GB1 in PDB 1PGA. The gain of function pair G41 and V54 sits in a three dimensional contact at about 5.5 angstroms. (C) Additive prediction against measured GB1 binding, colored by specific epistasis. Points above the diagonal are gain of function, individually weak yet strong together.

out double split is the design setting where a lab has partial combinatorial measurements at known positions and wants to complete the landscape. The held out position split is harder and remains modest, but it is still above the zero shot baseline.

5.5. Baselines And Ablations

We then compared the coupling head against a supervised model that uses no language model at all, against a second language model backbone, and against ablations of its own inputs. Table 3 collects the result on the GB1 and GFP held out double task under one identical five fold protocol. These numbers use a fixed shorter training budget so that every method is compared on equal terms, which places the coupling head slightly below its longer trained value in Table 2.

First, surprisingly, the two language model backbones agreed. DPLM and ESM-2 give almost the same held out correlation on both assays, so the recovered functional epistasis is a property of large protein language model representations rather than a quirk of one model. Second, the representation carries information beyond position and amino acid identity. A one hot pairwise model that sees only positions and identities reaches Spearman 0.29 on GB1, and the coupling head exceeds it, so the hidden states encode coupling that the identities alone do not. Third, the ablations locate the signal. Removing the language model states collapses the correlation to near the level of identity alone, while removing the explicit coupling product or the zero shot feature does not hurt and sometimes helps. The useful information lives in the per position representations, and a simpler head than the one we wrote would suffice.

We also compare against MAVE-NN (Tareen et al., 2022), the standard latent additive model with a pairwise interaction term and a global epistasis nonlinearity, run with the same decomposition on the same held out doubles. On GB1 it reaches Spearman 0.23, below the coupling head at 0.34, which we hypothesize this differential is due to the fact that MAVE-NN represents each residue pair interaction as a separate parameter, so it can only score a combination it has already measured. On a held out double split those combinations are unmeasured by construction and the interaction terms it needs are unconstrained, which limits how far it can extrapolate. The coupling head predicts the same held out doubles because the language model representation encodes the relationship between the two positions. MAVE-NN also grows with the square of the protein length, so its pairwise table is small for the GB1 domain but very large for a protein the size of GFP, which is the dash in Table 3.

Across six protein language model backbones, the coupling head gives similar held-out performance on GB1 and GFP, with only modest variation across architecture and scale (Fig. 3C,D). On GB1, the representation-based models consistently exceed the one-hot baseline, indicating information beyond mutation identity alone. On GFP, the same comparison is weaker, and the language model representations remain close to the identity baseline. These results suggest that functional epistasis is partly accessible from general protein representations, but the strength of that signal depends on the assay and protein landscape rather than on parameter count alone.

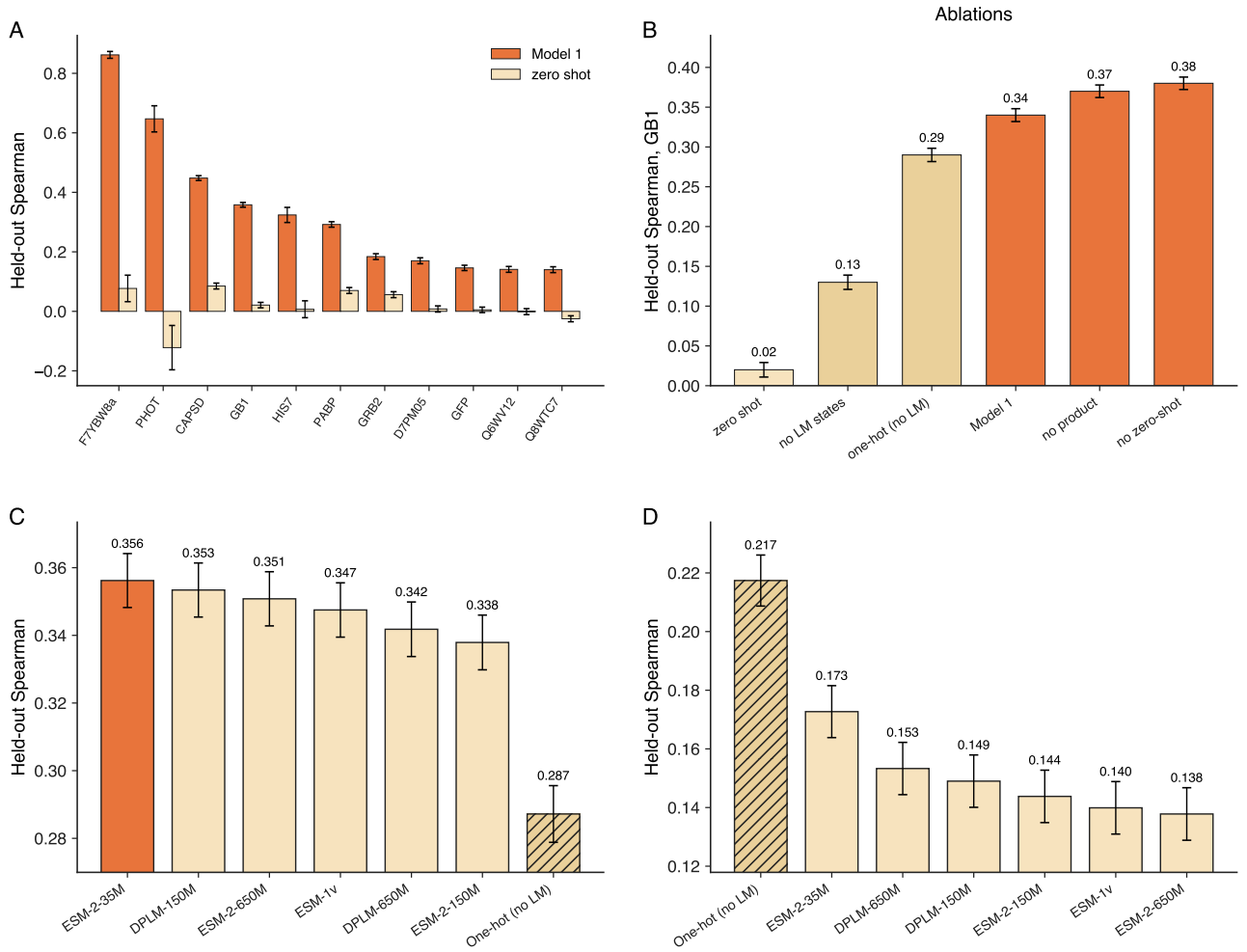

Figure 3. (A) The coupling head beats the zero shot score on all eleven measured function assays. (B) Ablations on GB1 locate the signal in the language model states. Removing those states drops the held out correlation to near the no language model level, while the full representation reaches 0.34, above the zero shot score at 0.02 and the supervised one hot model at 0.29. The explicit coupling product and the zero shot feature are not needed. (C, D) The coupling head ranked across six language model backbones on GB1 and GFP, with the no language model one hot baseline hatched. The ordering is stable and the held out correlation is flat across a 20-fold range of model size. Error bars are analytic standard errors of the Spearman correlation.

5.6. Functional Epistasis Is Specific to Each Protein

Then, we repeated the held out double test on 11 measured function assays that span binding, fluorescence, and other readouts. (Fig 3A) The trained coupling head beats the zero shot score on every one of the eleven assays. The median held out Spearman is 0.29 for the coupling head and 0.008 for the zero shot score. The magnitude of improvement varies across assays and sample sizes, but the sign is consistent across all 11 functional landscapes. Zero-shot scoring provides little sensitivity to functional epistasis, whereas supervised readout of the same protein language model representations recovers a measurable coupling signal.

Then, if the coupling head works on 11 proteins, would a head trained on some of them predict epistasis in a protein it has never seen? We tested this for function, with a leave one assay out protocol that trains on ten function assays and predicts the eleventh. The median cross protein Spearman is 0.03, which is no better than chance. Functional specific epistasis is therefore real and learnable inside a protein, yet it does not transfer between proteins. However, it is of further interest to pursue whether ”families” of proteins share a form of predictable functionally coupled mutations, such as specific types of fluorescent proteins or dehydrogenases, etc.

We then instead separately fine tuned the language model across roughly 120 stability landscapes and evaluated on 25 held out proteins. There the median cross protein Spearman is 0.34, well above the zero shot reference of 0.25. Stability coupled epistasis is shared across proteins because it

Table 3. Spearman correlations are reported under the same 5-fold protocol and on the same measured variants. Rows compare zero shot scores, trained coupling heads, non language model baselines, and feature ablations.

Method	GB1	GFP
No language model		
MAVE-NN (pairwise, global epistasis)	0.23	–
One hot pairwise (supervised)	0.29	0.22
DPLM-650M		
Zero shot second difference	0.02	0.01
Model 1 (coupling head)	0.34	0.15
ESM-2-650M		
Zero shot second difference	0.04	0.01
Model 1 (coupling head)	0.35	0.14
Model 1 ablations (DPLM backbone)		
no coupling product	0.37	0.22
no zero shot feature	0.38	0.15
no language model states	0.13	0.08

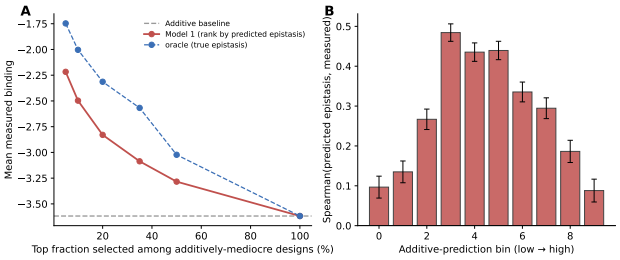

Figure 4. Model 2 design ranking in an additive matched GB1 candidate pool. Predicted specific epistasis is used to rank measured double mutants after controlling for additive score.

reflects the same physics of folding everywhere. Functional epistasis is idiosyncratic because each assay defines its own objective.

5.7. Epistasis Prediction Enables Gain of Function Selection

The GB1 binding landscape provides a closed design setting with dense measurements for double mutant combinations. We evaluate candidates within additive matched pools, where variants have comparable calibrated additive predictions and cannot be separated by single mutant effects alone. Ranking by predicted specific epistasis enriches for higher measured binding, with the top 10 percent of selected variants reaching a mean score of -2.50 compared with -3.62 for the full pool and approximately -2.00 for the oracle top 10 percent. Within additive bins, predicted epistasis correlates with measured function with a median Spearman correlation of 0.28. The selected gain of function variants lie above the additive diagonal in the GB1 landscape (Fig. 2C), and the strongest recovered pair, G41 and V54, forms a local contact in the folded structure (Fig. 2B).

6. Discussion

These results support an evaluation standard for functional protein language models. Evidence for learned functional epistasis should be assessed after additive single mutant effects and global assay response have been removed. Under this decomposition, zero shot DPLM contains a weak stability signal and limited functional signal, whereas a trained representation model recovers part of the residue coupling measured in real landscapes. The resulting predictor gives a practical route toward combinatorial design in settings where additive selection has been exhausted.

This framing also clarifies what the next generation models should do, as a better protein diffusion model should use target specific experimental evidence and should be evaluated on measured multi mutant function, beating additive selection at matched additive score. There are clear limitations, however. Here, we reported the strongest design result using GB1 because it has unusually complete measurements. Cross protein transfer behaves very differently for the two kinds of epistasis. A language model fine tuned across many stability landscapes predicts held out proteins at Spearman 0.34, while the same procedure on function assays reaches only 0.03. Functional epistasis must therefore be learned within each target protein, which is the regime that our held out double results address. The present manuscript reports completed measurements and leaves the wet laboratory test of the selected designs as the necessary next step.

7. Conclusion

We introduce a decomposition for measured function epistasis that removes additive effects and global assay response before evaluating model knowledge of residue coupling. Under this decomposition, natural sequence plausibility gives an incomplete basis for functional design. A trained DPLM representation model predicts part of the missing residue coupling and selects gain of function combinations that additive methods leave unresolved. The broader contribution is a concrete evaluation benchmark for generative protein models that aim to design useful combinatorial variants rather than only plausible sequences.

Impact Statement

This work is intended for protein engineering research. The proposed models nominate variants for laboratory testing. They are not meant to establish clinical safety or therapeutic efficacy without experimental validation.

References

Barlow, R. E. et al. *Statistical Inference under Order Restrictions*. Wiley, 1972.

Faure, A. J. et al. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904): 175–183, 2022.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379 (6637):1123–1130, 2023.

Melamed, D. et al. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(a)-binding protein. *RNA*, 19(11):1537–1551, 2013.

Notin, P. et al. ProteinGym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023.

Olson, C. A. et al. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014.

Poelwijk, F. J. et al. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, 2007.

Sailer, Z. R. et al. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017.

Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Tareen, A. et al. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biology*, 23(1):98, 2022.

Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

Visani, G. et al. Additive baselines furnish no evidence for epistasis learning by MULTI-evolve. *bioRxiv*, 2026.

Wang, X. et al. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Wang, X. et al. Towards a generative protein evolution machine with DPLM-Evo. *arXiv preprint arXiv:2605.00182*, 2026.

Wu, N. C. et al. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016.

A. Additional Figures And Interpretation

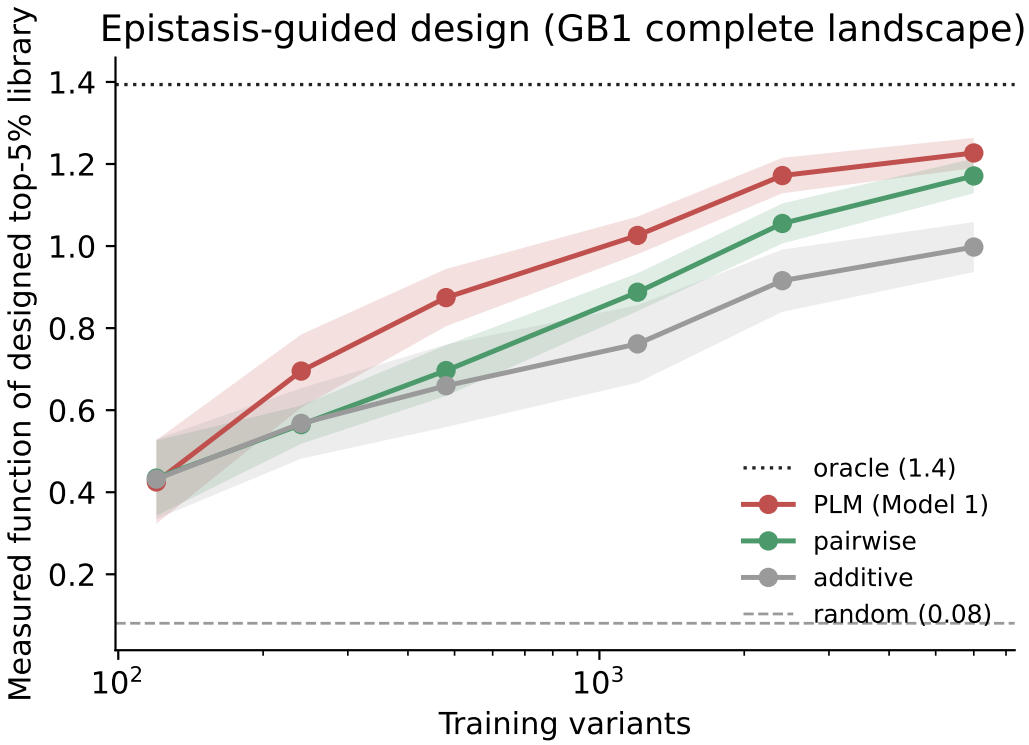

Figure 5. Epistasis guided design in the complete GB1 landscape. The DPLM representation model selects top five percent libraries with higher measured function than additive and pairwise baselines at most training sizes.

B. Algorithmic Connection To Generation

The coupling model acts on the residual $\varepsilon(S) = y(S) - \widehat{h}_g(A_g(S))$.

The combined design score can be written as

$$R_\lambda(S) = \widehat{h}_g(A_g(S)) + \lambda f_\theta(z_S), \tag{11}$$

where $f_\theta(z_S)$ predicts specific epistasis and $\lambda \geq 0$ controls how strongly the design procedure values residue coupling. Setting $\lambda = 0$ gives an additive calibrated generator. Positive λ gives an epistasis aware generator.

Proposition B.1 (Compatibility with additive calibrated generation)**.** *Let B be an additive matched candidate set with $|A_i - A_j| \leq \delta$ for all $i, j \in B$. Let $C_i = c(A_i)$ be any calibrated additive score with c being L_c Lipschitz on the interval containing the bin. For any selector that depends only on C_i, the selected mean of $c(A_i)$ differs from the bin mean by at most $2L_c\delta$. If the selector instead depends on $R_\lambda(S_i) = \widehat{h}_g(A_i) + \lambda\widehat{\varepsilon}_i$, then its excess measured function over the bin mean equals the residual covariance term in Theorem C.8 plus an additive calibration error bounded by $2L\delta$.*

Proof. The first claim follows because all additive scores in B lie in an interval of width δ. Lipschitz continuity gives $|c(A_i) - c(A_j)| \leq L_c\delta$ for every pair in the bin. The difference between the selected average of $c(A_i)$ and the full bin average is therefore bounded by the diameter of the calibrated additive scores, which is at most $L_c\delta$. The looser bound $2L_c\delta$ also covers the case where the selected average and the full average are compared through their deviations from an arbitrary point in the interval.

For the second claim, write $y_i = \widehat{h}_g(A_i) + \varepsilon_i$ for every candidate in the bin. The fitted response $\widehat{h}_g$ is L Lipschitz on the bin by assumption, so the contribution of $\widehat{h}_g(A_i)$ to any selected mean differs from the bin mean by at most $2L\delta$. The remaining contribution is the selected mean of ε_i minus the full bin mean of ε_i. Expanding the empirical covariance between the

selection indicator and ε_i gives exactly the covariance term in Theorem C.8. Thus additive calibrated generation can control the coarse design region, while epistasis prediction supplies the ranking signal within that region. □

C. Mathematical Justification For The Decomposition

The decomposition used in the main text is an empirical projection problem. The additive score $a_i = A_g(S_i)$ is a one dimensional summary of each variant. The global layer is the best monotone transformation of that summary. The specific epistasis target is the part of the measured function that remains after this projection.

C.1. Empirical Model Class

Fix one protein and one assay. Let $y \in \mathbb{R}^n$ be the vector of measured multi mutant values. Let $a \in \mathbb{R}^n$ be the vector of additive predictions computed from measured single mutants. Define the isotonic response class

$$\mathcal{G}(a) = \{u \in \mathbb{R}^n \mid a_i \le a_j \Rightarrow u_i \le u_j\}. \tag{12}$$

This set contains exactly the empirical predictions obtainable by applying an arbitrary monotone scalar function to the additive score. Ties in a are handled by the same order constraint in both directions, which forces the fitted values to be equal on tied additive scores.

Lemma C.1 (Geometry of the response class)**.** *The set $\mathcal{G}(a)$ is a closed convex cone in $\mathbb{R}^n$. It contains every constant vector. For every $y \in \mathbb{R}^n$, the Euclidean projection*

$$\widehat{u} = \operatorname*{arg\,min}_{u \in \mathcal{G}(a)} \|y - u\|_2^2 \tag{13}$$

exists and is unique.

Proof. For each ordered pair (i, j) with $a_i \le a_j$, the constraint $u_i \le u_j$ is a closed half space in $\mathbb{R}^n$. Hence $\mathcal{G}(a)$ is the intersection of finitely many closed half spaces and is closed and convex. If $u \in \mathcal{G}(a)$ and $\lambda \ge 0$, then $\lambda u \in \mathcal{G}(a)$, so $\mathcal{G}(a)$ is a cone. Constant vectors satisfy every order constraint with equality. In finite dimension, every nonempty closed convex set admits a Euclidean projection of y. Strict convexity of $u \mapsto \|y - u\|_2^2$ gives uniqueness. □

Definition C.2 (Specific epistasis residual)**.** The fitted global response is the projection $\widehat{u} = P_{\mathcal{G}(a)} y$. The empirical specific epistasis vector is

$$\varepsilon = y - \widehat{u}. \tag{14}$$

C.2. Projection Theorem

Theorem C.3 (Exact removal of monotone additive structure)**.** *Let $\widehat{u} = P_{\mathcal{G}(a)} y$ and $\varepsilon = y - \widehat{u}$. Then*

$$\langle \varepsilon, v - \widehat{u} \rangle \le 0 \quad \text{for every} \quad v \in \mathcal{G}(a). \tag{15}$$

Moreover,

$$\langle \varepsilon, \widehat{u} \rangle = 0 \quad \text{and} \quad \langle \varepsilon, \mathbf{1} \rangle = 0. \tag{16}$$

Consequently, no prediction vector in $\mathcal{G}(a)$ has smaller squared error than $\widehat{u}$.

Proof. The first order optimality condition for Euclidean projection onto a closed convex set gives

$$\langle y - \widehat{u}, v - \widehat{u} \rangle \le 0 \quad \text{for all} \quad v \in \mathcal{G}(a). \tag{17}$$

For completeness, take any $v \in \mathcal{G}(a)$ and $t \in [0, 1]$. Convexity gives $\widehat{u} + t(v - \widehat{u}) \in \mathcal{G}(a)$. Since $\widehat{u}$ minimizes squared error over $\mathcal{G}(a)$, the function

$$\psi(t) = \|y - \widehat{u} - t(v - \widehat{u})\|_2^2 \tag{18}$$

has a minimum at $t = 0$ from the right. Therefore $\psi'(0) \ge 0$, which gives

$$-2 \langle y - \widehat{u}, v - \widehat{u} \rangle \ge 0. \tag{19}$$

This is the displayed inequality because $\varepsilon = y - \widehat{u}$.

Since $\mathcal{G}(a)$ is a cone, both 0 and $2\widehat{u}$ belong to $\mathcal{G}(a)$. Substituting these two choices gives $\langle \varepsilon, \widehat{u} \rangle \geq 0$ and $\langle \varepsilon, \widehat{u} \rangle \leq 0$, hence $\langle \varepsilon, \widehat{u} \rangle = 0$. Since every constant vector belongs to $\mathcal{G}(a)$, both $\widehat{u} + t\mathbf{1}$ and $\widehat{u} - t\mathbf{1}$ lie in $\mathcal{G}(a)$ for all real t. Applying the same directional condition in both directions gives $\langle \varepsilon, \mathbf{1} \rangle = 0$. The final claim follows from the definition of $\widehat{u}$ as the minimizer over $\mathcal{G}(a)$. □

Theorem C.3 gives the geometric justification for the decomposition. The residual ε is the component left after the best possible monotone calibration curve has already been used.

Proposition C.4 (Block structure of the fitted response)**.** *Order the observations so that $a_1 \leq a_2 \leq \cdots \leq a_n$. The projection $\widehat{u} = P_{\mathcal{G}(a)} y$ is piecewise constant on contiguous blocks of this ordering. There exists a partition*

$$\{1, \ldots, n\} = B_1 \cup B_2 \cup \cdots \cup B_K \tag{20}$$

into consecutive index sets such that

$$\widehat{u}_i = |B_k|^{-1} \sum_{j \in B_k} y_j \quad \text{for} \quad i \in B_k, \tag{21}$$

and the block means are nondecreasing in k. Therefore the global layer is not an additional learned epistasis model. It is the least squares monotone recalibration of the additive score.

Proof. After sorting by a, the feasible set is the usual isotonic cone. The pool adjacent violators construction solves the projection onto this cone (Barlow et al., 1972). Starting from singleton blocks, whenever two adjacent block means violate monotonicity, the two blocks are replaced by their pooled mean. This pooling step preserves feasibility with all outside blocks and decreases the objective by the least squares identity

$$\sum_{i \in B} (y_i - \bar{y}_B)^2 + \sum_{i \in C} (y_i - \bar{y}_C)^2 \geq \sum_{i \in B \cup C} (y_i - \bar{y}_{B \cup C})^2 \tag{22}$$

whenever $\bar{y}_B > \bar{y}_C$ for adjacent blocks B and C. The procedure terminates after finitely many poolings and yields contiguous blocks with nondecreasing means. The resulting vector is feasible, and the projection is unique by Theorem C.1. Therefore it is the fitted response $\widehat{u}$. On each block the fitted value is the empirical mean, since any other constant value on that block would increase squared error while preserving the block structure. □

Corollary C.5 (Pythagorean variance identity)**.** *Let $\bar{y} = n^{-1} \sum_i y_i$ and let $\mathbf{1}$ be the all ones vector. Then*

$$\|y - \bar{y}\mathbf{1}\|_2^2 = \|\widehat{u} - \bar{y}\mathbf{1}\|_2^2 + \|\varepsilon\|_2^2 . \tag{23}$$

Therefore

$$R^2_{\text{glob}} = 1 - \frac{\|\varepsilon\|_2^2}{\|y - \bar{y}\mathbf{1}\|_2^2}. \tag{24}$$

Proof. Write $y - \bar{y}\mathbf{1} = (\widehat{u} - \bar{y}\mathbf{1}) + \varepsilon$. Expanding the squared norm gives the cross term $2 \langle \widehat{u} - \bar{y}\mathbf{1}, \varepsilon \rangle$. This term is zero because Theorem C.3 gives both $\langle \widehat{u}, \varepsilon \rangle = 0$ and $\langle \mathbf{1}, \varepsilon \rangle = 0$. □

C.3. Identifiability Against Additive Explanations

Proposition C.6 (Optimality against every monotone additive model)**.** *Let $m(a)$ be any monotone scalar function evaluated on the observed additive scores, and let $v_i = m(a_i)$. Then*

$$\sum_i (y_i - \widehat{u}_i)^2 \leq \sum_i (y_i - v_i)^2. \tag{25}$$

If a learned model improves on the global prediction on held out rows, then the improvement cannot be represented by any monotone function of the additive score alone.

Proof. The vector v belongs to $\mathcal{G}(a)$ by monotonicity of m. Since $\widehat{u}$ is the Euclidean projection of y onto $\mathcal{G}(a)$, it has squared error no larger than any other vector in $\mathcal{G}(a)$. This proves the inequality. On held out rows, the same statement applies to the fitted monotone response evaluated at test additive scores. Therefore any held out gain at matched additive score must come from information outside the monotone additive model class. □

C.4. Discrete Mixed Difference Of A Protein Language Model

Let F be any real valued score on subsets of a finite mutation set. For two mutations i and j, define

$$\Delta_{ij}F = F(\{i,j\}) - F(\{i\}) - F(\{j\}) + F(\emptyset). \tag{26}$$

Proposition C.7 (Additive annihilation)**.** *If $F(S) = c + \sum_{m\in S} r_m$ for all subsets $S \subseteq \{i,j\}$, then $\Delta_{ij}F = 0$. Conversely, for a function on two binary mutation indicators, $\Delta_{ij}F = 0$ is equivalent to the existence of constants c, r_i, r_j such that $F(S) = c + \sum_{m\in S} r_m$ on all four subsets.*

Proof. If $F(S) = c + \sum_{m\in S} r_m$, direct substitution gives $\Delta_{ij}F = 0$. Conversely, set $c = F(\emptyset)$, $r_i = F(\{i\}) - F(\emptyset)$, and $r_j = F(\{j\}) - F(\emptyset)$. These choices match F on $\emptyset$, $\{i\}$, and $\{j\}$. The condition $\Delta_{ij}F = 0$ implies $F(\{i,j\}) = F(\{i\}) + F(\{j\}) - F(\emptyset) = c + r_i + r_j$. Hence the additive representation matches all four vertices of the two mutation cube. □

We use this operator with $F = \ell_\phi$, the DPLM pseudo log likelihood. The resulting $Z_\phi(i,j)$ removes every additive contribution to the DPLM sequence score and isolates the model implied pair compatibility.

C.5. Matched Additive Selection

The design claim requires a stronger control than overall correlation. It must show that the learned residual ranks variants inside a set where additivity is essentially constant.

Theorem C.8 (Selection gain inside an additive bin)**.** *Let B be a set of candidate variants such that $|A_i - A_j| \le \delta$ for all $i, j \in B$. Suppose the fitted global link is L Lipschitz on the interval containing $\{A_i\}_{i\in B}$. Let I_i be the indicator that candidate i is selected by a rule that depends only on $\widehat{\varepsilon}_i$. If $\alpha = |B|^{-1}\sum_{i\in B} I_i > 0$, then*

$$\frac{\sum_{i\in B} I_i y_i}{\sum_{i\in B} I_i} - \frac{1}{|B|}\sum_{i\in B} y_i = \frac{\widehat{\mathrm{Cov}}_B(I,\varepsilon)}{\alpha} + r_B, \tag{27}$$

where $|r_B| \le 2L\delta$.

Proof. Write $y_i = \widehat{h}(A_i) + \varepsilon_i$ for $i \in B$. Since $\widehat{h}$ is L Lipschitz and all additive scores in B lie in an interval of width δ, any two values of $\widehat{h}(A_i)$ inside the bin differ by at most $L\delta$. Hence the selected mean of $\widehat{h}(A_i)$ and the full bin mean differ by at most $2L\delta$. For the residual term,

$$\frac{\sum_{i\in B} I_i \varepsilon_i}{\sum_{i\in B} I_i} - \frac{1}{|B|}\sum_{i\in B} \varepsilon_i = \frac{\widehat{\mathrm{Cov}}_B(I,\varepsilon)}{\alpha}, \tag{28}$$

which follows by expanding the empirical covariance and using $\alpha = |B|^{-1}\sum_{i\in B} I_i$. Combining the two parts gives the claim. □

Theorem C.8 formalizes the matched additive control used in Figure 4. If the additive bin is narrow, then the selected library improves only when the selection indicator has positive covariance with the specific residual.

C.6. Noise And Residual Scale

Let the observed measurement be $y_i = y_i^\star + \eta_i$, where η_i are independent mean zero errors with variance at most σ_η^2. Let $\varepsilon^\star$ be the specific residual that would be obtained from noiseless measurements with the same projection rule, and let $\widehat{Q}_{\mathrm{spec}}$ be the empirical variance ratio reported in the paper.

Proposition C.9 (Noise floor interpretation)**.** *If the fitted global response is held fixed and only measurement noise is varied, then*

$$\mathbb{E}\left[\widehat{\mathrm{Var}}(\varepsilon)\right] = \widehat{\mathrm{Var}}(\varepsilon^\star) + \sigma_\eta^2 - \frac{1}{n}\sigma_\eta^2. \tag{29}$$

Thus a residual standard deviation substantially above the known measurement noise scale is evidence for nonzero specific epistasis.

Proof. With the projection held fixed, the residual is $\varepsilon_i = \varepsilon_i^\star + \eta_i$. The empirical sample variance expands into the variance of $\varepsilon^\star$, the empirical variance of η, and a cross term. The cross term has expectation zero by independence and mean zero noise. The expected empirical variance of n independent noise terms is $(1 - n^{-1})\sigma_\eta^2$. □

C.7. Higher Order Decomposition

For a complete landscape over k mutable sites, let $X = (X_1, \ldots, X_k)$ be uniformly distributed over the measured allele grid, and let $f(X)$ be the measured function. The Hoeffding decomposition (Hoeffding, 1948) writes

$$f(X) = f_\emptyset + \sum_i f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) + \sum_{i<j<l} f_{ijl}(X_i, X_j, X_l) + \cdots . \tag{30}$$

The components are uniquely determined by the constraints

$$\mathbb{E}[f_U(X_U) \mid X_V] = 0 \quad \text{whenever} \quad V \subsetneq U. \tag{31}$$

Proposition C.10 (Orthogonal interaction orders)**.** *Under the uniform measure on a complete grid, the terms of different index sets in the Hoeffding decomposition are pairwise orthogonal in L^2. Hence the total variance decomposes as*

$$\mathrm{Var}(f(X)) = \sum_{\emptyset \neq U \subseteq \{1,\ldots,k\}} \mathrm{Var}(f_U(X_U)). \tag{32}$$

Proof. Let $U \neq W$. Without loss of generality, choose an index in U that is not in W, or symmetrically exchange the two sets. Conditioning on all variables in W and on all variables in U except that index, the centering constraint for f_U gives conditional mean zero. Therefore $\mathbb{E}[f_U f_W] = 0$. Summing the orthogonal components gives the variance identity. □

The additive layer estimates the first order terms. A pairwise model estimates the second order terms. The third order residual is the component left after the first and second order terms have been fitted.

C.8. Leakage And Evaluation

Every quantity in the decomposition is computed within a protein and on measured variants. The single mutant effects used in A_i come only from constituent single mutants. The global response link is used to define the residual target, not to evaluate generative novelty. Model 1 evaluation uses out of fold predictions for held out double mutants or group held out positions. The design test then ranks measured held out candidates by predicted residual inside additive matched bins. These choices ensure that the reported gain is not an artifact of target leakage or additive memorization.