MECHANISTIC INTERPRETABILITY FINDINGS THAT GENERALIZE: CHAPERONE-DEPENDENT FOLDING IN ESMFOLD2

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability findings in scientific models are often supported by post hoc plausibility rather than held-out tests. We introduce a workflow that turns interpretability outputs into statistically testable predictions: feature terms discovered in examples sharing a biological property should have stronger effects in unseen members than in controls. We apply this workflow to an SAE over ESMFold2 representations and to GroEL/GroES clients: proteins that require this chaperone system for folding *in vivo*. Perturbations in 25 discovery clients identify 67 stable SAE terms. Their ablations have larger effects in eight held-out clients than in 19 matched controls for 56 out of 67 terms ($p = 2 \times 10^{-12}$), and their selected-to-random effect ratio is $\sim 24\%$ higher in clients than in controls. When multiple selected terms are ablated together, client-control separation increases from approximately 3.8- to 7.3-fold across a set of randomly subsampled terms. After these tests, interpretations of the selected terms connect to known client properties and motivate new hypotheses about client-cavity charge patterning, compensatory core packing, and delayed interface closure. Our contribution is a cohort-conditioned discovery-and-evaluation workflow that turns latent feature interactions into held-out tests of biological selectivity. The framework accommodates multiple discovery methods and uses the selected interventions to nominate biological hypotheses.

1 INTRODUCTION

The generalization of foundation models suggests they model mechanisms hidden behind their training data (McGrath et al., 2022; Rives et al., 2021; Rao et al., 2020; Jumper et al., 2021; Candido et al., 2026). A protein structure predictor supervised only on native structures may also encode how proteins reach those structures in a cell, and interpretability methods are increasingly used to read such implicit knowledge out of learned representations (Murdoch et al., 2019; Roscher et al., 2020). But these findings are usually supported by a few examples and a plausible story, which gives no evidence that they hold for examples that were not examined (Adebayo et al., 2020; 2022).

An interpretability finding can instead be stated as a falsifiable prediction: take a group of examples that share a mechanism, discover a behavior in some of them, and require that behavior to be stronger in the remaining members of the group than in matched examples that lack the mechanism. A finding that survives this test has been shown to generalize; one that does not may be spurious to the examples that produced it. If successful, this provides a way to use the implicit knowledge in foundation models to generate scientific hypotheses while separating reproducible structure from post hoc explanation.

Chaperone-assisted protein folding provides a natural test bed for this framework. The GroEL/GroES chaperonin folds certain proteins that cannot fold alone (clients), enclosing non-native intermediates in a chamber that shields them from aggregation (Houry et al., 1999; Li et al., 2009). Its *in vivo* clients have been identified experimentally (Fujiwara et al., 2010) and share this folding requirement despite unrelated sequences and native folds. ESMFold2 (Candido et al., 2026) predicts those native folds while receiving no supervision about chaperone dependence, so any encoding of that shared biology must have emerged implicitly.

ESMFold2 predicts structure from representations produced by a frozen ESMC-6B encoder, and a sparse autoencoder (SAE) released with the model decomposes these representations into interpretable

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.

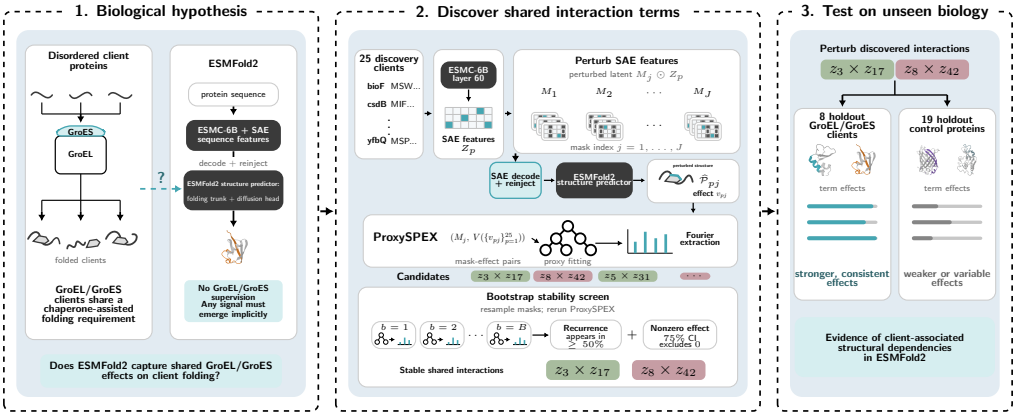


Figure 1: **Discovering SAE feature interactions in ESMFold2 associated with GroEL/GroES-assisted folding.** (1) We test whether ESMFold2 captures structural dependencies shared by GroEL/-GroES clients despite receiving no chaperone supervision. (2) We perturb SAE features in 25 discovery clients, propagate the edits through ESMFold2, and use ProxySPEX with bootstrap screening to identify stable interactions. (3) We hold the panel fixed and test it in eight holdout clients and 19 controls.

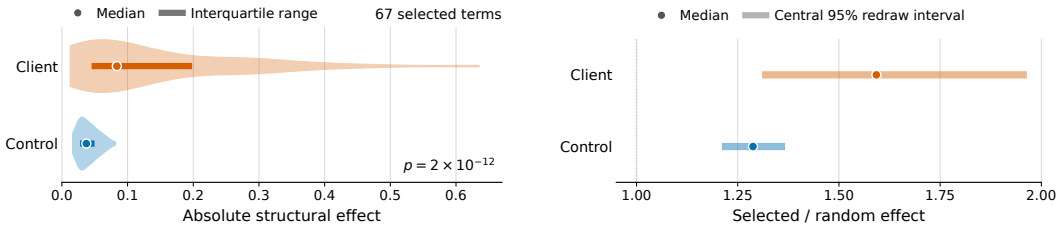
features associated with recognizable sequence and structural patterns (Candido et al., 2026). These features provide natural directions for intervention, but their structural effects can depend strongly on context: a Gly-rich region, for example, may matter differently depending on the surrounding secondary structure or chemical environment. Such context dependence is particularly relevant for GroEL/GroES client folding, which is associated with combinations of sequence and structural properties rather than a single defining signature. We therefore study marginal, pairwise, and higher-order effects, asking which individual features and feature combinations have structural consequences that recur across clients.

Figure 1 summarizes how we search for SAE feature combinations with reproducible structural effects across GroEL/GroES clients and test them on unseen proteins. We perturb combinations of SAE features in a discovery cohort and search for marginal, pairwise, and three-way effects that recur across proteins. Repeating this discovery analysis across resampled perturbations filters the candidates to a stable panel, which is fixed before evaluation on unseen clients and matched controls.

The evaluation separates into three questions. First, **generalization and client specificity**: do the selected terms have larger structural effects in unseen GroEL/GroES clients than in matched controls? Second, **selection specificity**: are the effects specific to the terms selected by the discovery procedure, rather than a generic consequence of perturbing SAE features? Third, **composition**: when several selected multi-feature (interaction) terms are ablated together, do their effects continue to accumulate preferentially in clients? Only after these tests do we use feature descriptions and prior literature to interpret the selected terms and formulate experiments.

Our workflow identifies 67 stable SAE terms: 28 marginals, 38 pairs, and one triple. Their effects are larger in held-out clients for 56 of 67 terms ($p = 2.24 \times 10^{-12}$), and the selected-to-random effect ratio is approximately 24% higher in clients than in controls. Jointly ablating random subsets of the selected terms increases the client-to-control effect ratio from 3.8 for subsets of one term to 7.3 for subsets of five.Together, these results show that the selected terms generalize beyond discovery, capture more than generic perturbation sensitivity, and become more client-selective when combined. After validation, interpretations of the selected terms connect to previously reported hydrophobic-exposure and small-residue properties and motivate hypotheses about compensatory core packing, client-cavity electrostatics, and late interface closure. Together, the held-out interventions indicate that ESMFold2 representations contain structural dependencies associated with GroEL/GroES client status despite receiving no chaperone supervision.

Finally, we introduce GroEL/GroES-dependent protein folding as a biological benchmark for interaction-discovery methods and compare ProxySPEX with marginal LASSO, sparse polynomial regression, RuleFit, and FIGS.



(a) (b)

Figure 2: **Selected SAE terms generalize to unseen GroEL/GroES clients and beat random terms of the same order.** We test the 67 terms discovered in 25 clients (28 single features, 38 pairs, one triple) by jointly ablating each term's SAE features in 8 held-out clients and 19 matched controls. Structural effect is the normalized loss of agreement between ESMFold2's prediction and the experimental structure caused by an ablation. **(a)** Each point is one term's mean absolute effect within a cohort. Effects are larger in held-out clients for 56 of 67 terms (one-sided paired Wilcoxon signed-rank, $p = 2 \times 10^{-12}$). **(b)** Ratio of the selected panel's mean effect to that of each of 1,000 random panels with the same mix of interaction orders, summarized by the median and central 95% interval across redraws. Selected terms' advantage over random terms is about 25% larger in held-out clients than in controls, so the effect is specific to both the selected terms and the client cohort.

2 Background

Protein structure prediction. Protein language models (pLMs) learn sequence representations that encode substantial structural and functional information (Rives et al., 2021; Rao et al., 2020). ESMFold predicts three-dimensional structure directly from such representations (Lin et al., 2023), without the multiple-sequence alignments and templates used by AlphaFold2 (Jumper et al., 2021). ESMFold2 retains this representation-to-structure paradigm while using a frozen pLM encoder, ESMC-6B, followed by a folding trunk and diffusion head (Candido et al., 2026). Because structure is computed from the representation alone, an intervention on that representation is an intervention on everything the folding module sees.

Mechanistic interpretability and interaction discovery. Recent work has begun to mechanistically interpret protein models by decomposing representations into biologically meaningful features (Adams et al., 2025; Simon and Zou, 2025; Gujral et al., 2025), tracing computations and sparse circuits across layers and through generative protein models (Lu et al., 2026; Tsui et al., 2026a), and intervening on learned concepts to alter predictions (Parsan et al., 2025; Shen et al., 2025). Sparse autoencoders (SAEs) (Cunningham et al., 2023), in particular, expose latent features associated with sequence motifs, structural elements, and functional annotations; recently LLMs have been used to scale their description and interpretation (Simon and Zou, 2025; Candido et al., 2026). However, most analyses study SAE features individually even though model behavior may depend on combinations of features. Interaction-attribution methods such as k-Shapley indices, SPEX, and ProxySPEX provide tools for recovering such joint effects (Fumagalli et al., 2024; Kang et al., 2025; Butler et al., 2025). We not only use these tools to identify interacting SAE features, but also to test whether the discovered interactions capture biological structure that generalizes across proteins.

GroEL/GroES-dependent folding. GroEL/GroES assists proteins prone to misfolding or aggregation by enclosing non-native intermediates in a protected chamber where folding can continue (Houry et al., 1999; Li et al., 2009). Proteome-wide surveys mapped its clients (Kerner et al., 2005), and depletion experiments identified 57 Class IV *E. Coli* proteins that require this assistance in vivo (Fujiwara et al., 2010). Client proteins share several properties: Ala/Gly enrichment, exposed hydrophobic and amphipathic-helical regions, vulnerable α/β architectures, and acidic composition. These were established independently of the work reported here (Fujiwara et al., 2010; Houry et al., 1999; Li et al., 2009; Georgescauld et al., 2014; To et al., 2022). We use these properties in Section 4.4 to ask whether the selected terms recover known biology.

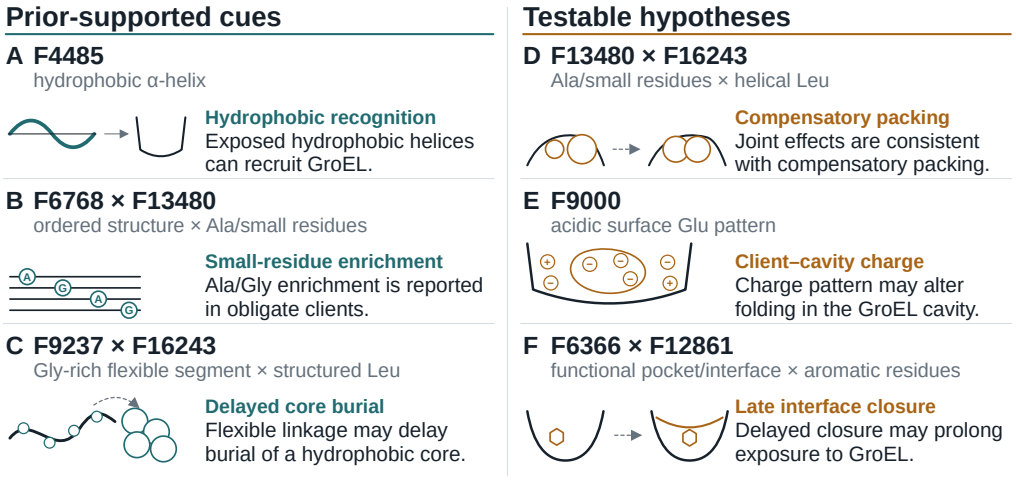


Figure 3: **Selected SAE terms connect literature-linked client properties to candidate folding mechanisms.** Panels A-C relate feature descriptions to previously reported properties of GroEL/-GroES clients, including exposed hydrophobic helices and Ala/Gly-biased sequence composition. Panels D-F propose testable novel mechanisms involving compensatory core packing, client-cavity charge patterning, and delayed interface closure. The × symbol corresponds to a selected feature pair, not a physical contact between the features.

3 Methodology

We want to discover SAE feature interactions that reflect what a cohort of proteins has in common, rather than behaviors specific to individual proteins, and to test whether they generalize before interpreting them. The workflow therefore separates cohort definition, intervention, discovery, and evaluation, producing a panel whose terms and features are fixed before any holdout protein is evaluated (Figure 1).

Concretely, we use the following workflow:

1. **Cohorts.** Define a discovery cohort that shares the behavior of interest, together with held-out members and a matched control cohort.
2. **Interventions.** Choose an interpretable feature basis, define the ablation of a feature set, and specify a scalar effect measure.
3. **Stable discovery.** Search over feature sets for effects that recur across the discovery cohort, retaining terms that are stable under resampling.
4. **Evaluation.** Freeze the selected panel and test whether its effects transfer to held-out members, differ from matched controls, exceed comparable random terms, and continue to separate the cohorts when several terms are ablated together.

Stages 1–4 are instantiated below for GroEL/GroES-assisted folding. Full details for every stage are in Appendices A and B.

3.1 Cohorts and interventions

Discovery and holdout clients. We draw the discovery cohort from experimentally reported GroEL/GroES clients and reserve sequence-dissimilar clients for holdout evaluation. Of the 57 Class IV (chaperonin-dependent) *E. coli* clients reported by Fujiwara et al. (2010), 33 have audited PDB targets (Berman et al., 2000). MMseqs2 clustering (Steinegger and Söding, 2017) identifies 23 of these as sequence singletons. We select eight singletons as holdouts using a covariate-matching rule that does not use intervention responses, proxy predictions, or interaction coefficients (Appendix A). Because these holdouts are sequence singletons, transfer to them cannot be explained by close sequence homology to the 25 remaining discovery clients.

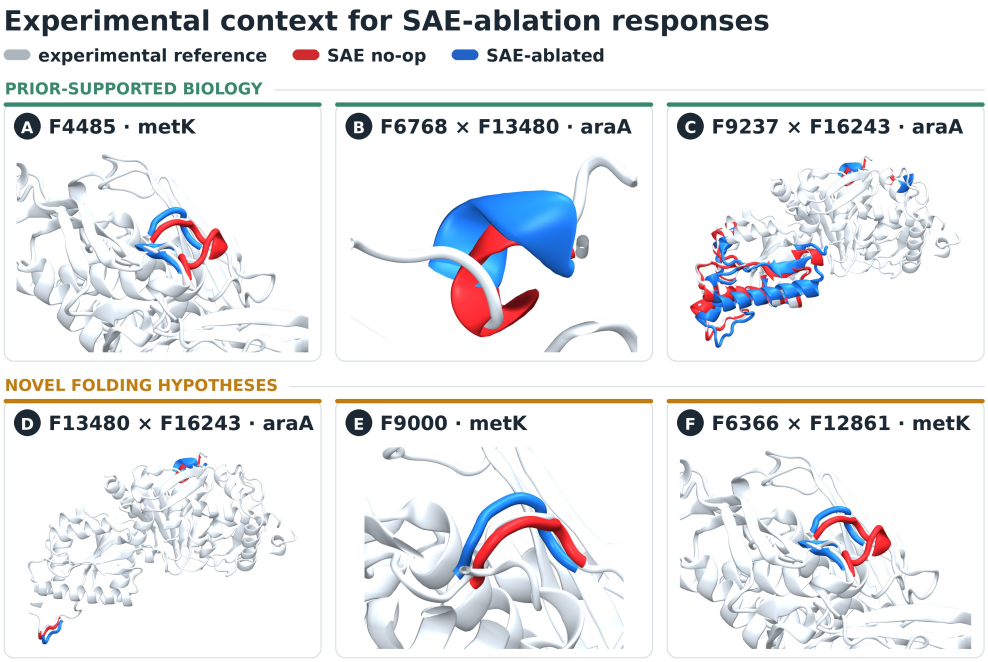


Figure 4: **Structural examples of selected SAE interventions.** The overlays show where selected SAE ablations alter predictions for two held-out client proteins: AraA and MetK, providing structural context for the hypotheses in Figure 3. White marks where the two predictions agree, while red and blue highlight the no-op and ablated predictions where they diverge, respectively.

Controls. Controls are cytosolic, PDB-backed *E. Coli* K-12 proteins with reviewed UniProtKB/Swiss-Prot entries (UniProt Consortium, 2025). We match them to the discovery cohort on sequence length, SAE-feature profiles, sequence diversity, and PDB context to reduce differences unrelated to GroEL/GroES client status. The resulting 19 proteins lack literature evidence of GroEL/GroES dependence.

Latent interventions. We intervene in the SAE basis so that ablations act on labeled latent features rather than arbitrary directions in representation space. Using a pretrained top-K SAE on layer-60 representations of ESMC-6B, ESMFold2's encoder (Candido et al., 2026), we define the feature basis $\mathcal{F}$ from discovery clients alone. $\mathcal{F}$ contains 78 SAE features whose released descriptions are non-generic and that are active on at least 5% of residues in at least 80% of discovery clients.

To ablate a feature set $S \subseteq \mathcal{F}$, we encode the layer-60 representation from amino acid sequence, zero the coordinates in S at every residue, decode the edited SAE representation, and recompute layers 61–80 before passing the result through the folding trunk and diffusion model. The no-op baseline follows the same encode–decode–reinsert path without zeroing any feature. Appendix A.4 quantifies the resulting distortion from SAE encoding and decoding.

Noise-averaged structural effect. For protein p, ablation set $S \subseteq \mathcal{F}$, and conditional diffusion draw r, let $t_{p,r}(S)$ denote the TM-score (Zhang and Skolnick, 2004) between the predicted structure and the experimental PDB target. We use $R = 3$ draws per condition for diffusion sampling and first average their TM-scores $\bar{t}_p(S) = \frac{1}{R}\sum_{r=1}^{R} t_{p,r}(S)$. We then define the signed, headroom-normalized structural effect as

$$v_{p,S} = \frac{\bar{t}_p(\varnothing) - \bar{t}_p(S)}{1 - \bar{t}_p(\varnothing)}, \tag{1}$$

where $\varnothing$ denotes the matched SAE no-op. Positive values indicate reduced agreement with the experimental target. The three draws share one upstream ESMC/SAE/folding-trunk representation and differ only through conditional diffusion sampling.

3.2 STABLE DISCOVERY

Mask evaluation and interaction extraction. For an ablation set S, we summarize its effect across the discovery cohort by

$$V(S) = Q_{0.25}\left(\{v_{p,S}\}_{p=1}^{25}\right),$$

where $Q_{0.25}$ denotes the 25th percentile across proteins. This lower quantile favors feature sets whose effects extend across the cohort rather than sets whose aggregate effect is driven by a few proteins.

Because the 2^{78} possible ablation sets cannot be enumerated, we estimate this set function from sampled perturbations. We uniformly sample 2,000 binary masks $M_j \in \{0,1\}^{78}$, each defining an ablation set S_j, and evaluate every mask on all 25 discovery clients, yielding 50,000 structural effects. ProxySPEX (Butler et al., 2025) estimates this lower-quartile target by fitting a LightGBM quantile model to all 50,000 mask-protein responses, then extracts marginal, pairwise, and three-way k-SII terms from the fitted proxy (Gorji et al., 2025). We choose ProxySPEX due to efficiency in scaling to higher dimensional bases for expensive black box (foundation models) queries. See Section 5 for a study of alternative methods for interaction discovery.

Bootstrap stability screening. We retain only terms that are reproducibly recovered under resampling of the discovery perturbations. Following the PCS stability principle (Yu and Kumbier, 2020; Gan et al., 2025), we repeat proxy fitting and extraction over 1,000 bootstrap resamples of the evaluated masks. Each bootstrap replicate resamples complete masks, keeping all 25 protein responses for a mask together. We retain terms recovered in at least 50% of refits and whose coefficient, conditional on recovery, has a central 75% bootstrap interval excluding zero. This yields the selected panel $\mathcal{S}^*$, consisting of 67 stable terms: 28 marginals, 38 pairs, and 1 triple. We hold the selected panel fixed before any holdout evaluation.

3.3 EVALUATION

Generalization and client specificity. For each frozen term $s^* \in \mathcal{S}^*$, we jointly ablate its constituent SAE features in eight held-out clients and 19 matched controls. This evaluates the direct structural effect of removing the selected feature set via SAE ablation, not re-estimating the k-SII interaction coefficient on the holdout proteins.

For each s^*, we compute the cohort-mean structural effect $|v_{ps^*}|$ in each group and compare the 67 pairs using a one-sided exact Wilcoxon signed-rank test for larger effects in held-out clients. The selected term is the paired unit. Larger effects in held-out clients than in controls would be consistent with transfer beyond the discovery cohort and preferential effects in clients. However, because term effects reuse proteins and terms sometimes reuse SAE features, ***the paired differences may be dependent, which the Wilcoxon calculation does not account for***. We therefore report the p-value as ***nominal and exploratory***, rather than as calibrated evidence of generalization across a protein population.

Selection specificity. After constructing $\mathcal{S}^*$, we fix a comparison bank containing the 50 remaining marginals, 150 random pairs, and 50 random triples from the same feature basis, excluding the selected terms in $\mathcal{S}^*$. We then randomly draw 1,000 panels of 28 marginals, 38 pairs, and one triple without replacement. The same panels are used for clients and controls. Within each cohort, we divide the selected panel's mean absolute effect by the random panel's mean absolute effect. The central 95% range describes variation across random-panel composition, not uncertainty over proteins.

Joint ablations. We sample 20 outcome-blind five-term panels from the 39 selected multi-feature terms in $\mathcal{S}^*$ and evaluate nested prefixes $k \in \{1,2,3,5\}$. For each prefix, we ablate the union of its SAE features, counting shared features once. We compare the measured joint effect with an overlap-corrected reference constructed from the signed effects of the same terms measured separately using inclusion-exclusion (e.g. for two terms A and B, this is $v_p(A) + v_p(B) - v_p(A \cap B)$). Signed effects are combined before taking absolute values. Observation and reference agree at $k = 1$ by construction.

AI-assisted biological interpretation. We provide the selected terms, released SAE descriptions, and the GroEL/GroES literature cited here to GPT-5.6 Sol in Codex Ultra mode (OpenAI, 2026), instructing it to relate each term to a reported client-associated property, propose an experimentally

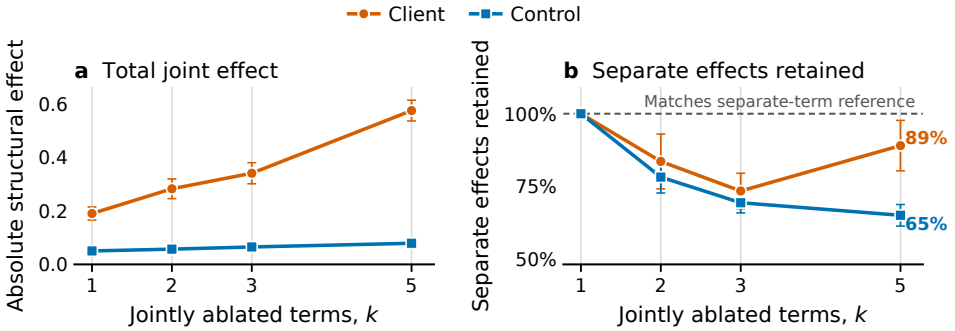


Figure 5: **Joint ablations widen client-control separation.** For 20 randomly sampled nested panels of five, we jointly ablate the features in the first k selected terms. **(a)** From $k = 1$ to $k = 5$, mean absolute structural effects increase 3.0-fold in eight held-out clients but only 1.6-fold in 19 controls. **(b)** We ask how much of the effect measured for these terms separately remains when they are ablated together, after correcting for shared features. A value of 100% means the joint effect equals this separate-term reference, while lower values mean the joint ablation produces less change than the separate measurements suggest. Clients retain more of the reference effect than controls at every $k > 1$, reaching 89% versus 65% at $k = 5$. Error bars summarize variation across the same 20 panels.

testable hypothesis, or abstain. We *first* statistically evaluate in Sections 4.1 to 4.3, then *interpret* already-selected terms in Section 4.4 after validation.

4 Experimental Results

We evaluate whether the selected and fixed 67-term panel $\mathcal{S}^*$ generalizes preferentially to unseen clients, exceeds order-matched random terms, remains client-selective when several terms are ablated together, and yields biologically meaningful hypotheses about GroEL/GroES-assisted folding.

4.1 Selected terms show larger effects in held-out clients than in controls

Across $\mathcal{S}^*$, cohort-mean absolute effects average 0.14 in held-out clients and 0.04 in controls. Figure 2a shows larger cohort-mean and cohort-median absolute effects in clients than controls, with a broader upper tail (one-sided exact Wilcoxon signed-rank: $W = 2125, p = 2 \times 10^{-12}$). Effects are larger in clients for 56/67 terms ($\sim 84\%$ of selected terms).

These results show a cohort-level preference for held-out clients, with substantial variation across individual proteins. The largest raw TM-score changes occur in araA, metK, and yneB (Appendix B.3)

4.2 Selected terms exceed order-matched random terms in held-out clients

The selected-to-random effect ratio was higher in held-out clients than in controls (Figure 2b), with medians of 1.59 and 1.29, respectively. Across paired redraws of random panels, this ratio was approximately $\sim 24\%$ higher in clients, supporting selection specificity: the selected terms in $\mathcal{S}^*$ outperform order-matched random terms by a larger margin in held-out clients than in controls. Protein-specific selection results are in Appendix B.3.

4.3 Joint ablations increasingly separate held-out clients from controls

We next ask whether client specificity persists when several selected terms are ablated together. From the 39 selected multi-feature terms, we sample 20 five-term panels (without using holdout outcomes) and ablate the union of features in the first $k \in \{1, 2, 3, 5\}$ terms.

Joint effects accumulate more strongly in clients. From $k = 1$ to $k = 5$, mean absolute structural effects increased from 0.190 to 0.574 in clients, but only from 0.050 to 0.079 in controls (Figure 5a). The client-control contrast consequently widened from 3.8-fold to 7.3-fold. Thus, ablating additional

Method	# terms	Avg. term size	Nomimal p-value	Ratio contrast (%)
ProxySPEX	67	1.60	2.24×10^{-12}	23.9 (12.5)
LASSO (marginal)	55	1.00	1.38×10^{-9}	NA
Sparse polynomial	38	1.97	1.27×10^{-6}	-0.5 (14.5)
RuleFit	59	1.44	2.54×10^{-12}	20.7 (11.8)
FIGS	13	1.08	4.03×10^{-3}	49.2 (41.1)

Table 1: **Interaction-discovery methods on the GroEL/GroES benchmark.** All methods use the same 50,000 discovery effects and direct ablation in 8 held-out clients and 19 controls. p-values are one-sided paired Wilcoxon tests (client $>$ control). Ratio contrast is $100(R_{\text{client}}/R_{\text{control}} - 1)$, with redraw SD in parentheses. LASSO is NA because too few unselected marginals remain to compare. Differences in panel size and order composition limit direct method ranking

selected terms substantially strengthens client-control separation. Protein-specific joint effect results are in Appendix B.4.

Joint ablations preserve more of the separately measured effect in clients. Figure 5b compares the measured joint effect with a reference constructed from the same terms ablated separately, after correcting for features shared among terms. A value of 100% means that the joint effect equals this separate-ablation reference, and lower values mean that some separately measured effect is lost when the terms are ablated together. At every $k > 1$, this percentage is higher in clients than in controls, reaching 89% versus 65% at $k = 5$. Together, the two panels show that adding selected terms increases client-control separation because structural effects continue to grow in clients but saturate earlier in controls. Protein-specific joint effect results are in Appendix B.4.

4.4 Selected-term interpretation motivates testable hypotheses about client folding

Having shown that the selected panel generalizes, we next ask what biological hypotheses it suggests. Using the released SAE descriptions and focused GroEL/GroES literature with GPT-5.6 Sol in Codex Ultra mode (OpenAI, 2026), we identified literature-linked components, proposed testable hypotheses, or abstained. Figures 3 and 4 present six examples. Appendix B.5 reports all 67 terms, including ambiguous interpretations and abstentions. These are interpretations for an already validated panel of terms, and they do not provide additional statistical or mechanistic validation.

Several terms align with prior biology. F4485 is annotated as a hydrophobic α-helix, consistent with GroEL recognition of exposed hydrophobic or amphipathic helices (Hlodan et al., 1995; Wang et al., 1999; Li et al., 2009); F6768 $\times$ F13480 combines annotations for ordered structure and Ala/small residues, echoing the reported Ala/Gly enrichment of clients (Fujiwara et al., 2010); and F9237 $\times$ F16243 combines Gly-rich flexibility with structured Leu and motivates a hypothesis in which a flexible connector delays burial of a hydrophobic element.

Other terms motivate concrete experiments. F13480 $\times$ F16243 suggests testing whether compensatory small-residue/Leu edits restore core packing; F9000 suggests testing whether the spatial pattern of surface Glu residues affects folding through GroEL cavity electrostatics (Tang et al., 2008; Motojima et al., 2012); F6366 $\times$ F12861 suggests testing whether ligand or partner binding shortens the exposure of a late-closing interface. Figure 4 shows where these SAE ablations alter representative AraA and MetK predictions, providing structural context and intuition for potential experiments.

5 GroEL/GroES as a Benchmark for Interaction Discovery

Our scientific setting can also serve as a biological benchmark for interaction-discovery methods. Each method receives the same 50,000 structural effects (2,000 SAE masks $\times$ 25 discovery clients) and must return a sparse panel of marginal, pairwise, and three-way terms. All methods then pass through the same bootstrap stability screen, and the resulting panel is fixed for evaluation. It is evaluated by direct intervention in the ESMFold2 for the eight held-out clients and 19 controls. Here,

we focus on their effects in held-out clients versus in controls (*generalization*) and their comparison to random, unselected terms of similar interaction order (*selection specificity*).

We compare ProxySPEX with marginal LASSO (Tibshirani, 1996) (no interactions), a sparse degree-three polynomial model with a max order of three on the Walsh basis (O'Donnell, 2014), RuleFit (Friedman and Popescu, 2008), and FIGS (Tan et al., 2025). For the sake of computation, we restricted the comparison to methods that do not require query access, so the interaction discovery comes from the 50,000 mask and structural effect pairing. Experiment details are in Appendix D.

Table 1 describes some diagnostics and results: all five selected panels show larger cohort-mean effects in clients than controls. We therefore also compare their enrichment over order-matched random terms. The selected-to-random effect ratios for ProxySPEX and RuleFit are 23.9% and 20.7% higher in clients than in controls, respectively, while sparse polynomial shows no difference (-0.5%). FIGS has the largest estimated ratio contrast but also the greatest variation across random-panel redraws. LASSO cannot be scored for specificity: it selects 55 of the 78 SAE features, leaving too few unselected features for a size-matched random panel. Note that the selected panels differ in size and order composition, which limits how directly the methods can be compared.

6 Discussion

We develop a workflow that selects SAE feature interactions shared across a cohort of GroEL/GroES-assisted proteins and tests them on held-out clients and matched controls before interpreting them. The selected terms transfer to holdout clients and exceed order-matched random terms in clients more than in controls, and they become more client-selective when ablated together. Together this suggests ESMFold2 encodes structural dependencies of GroEL/GroES client folding despite receiving no chaperone supervision. Interpreting the terms recovers established client properties and yields new hypotheses about compensatory core packing, client-cavity electrostatics, and late interface closure.

Our results complement recent work that discovers (Adams et al., 2025; Simon and Zou, 2025; Gujral et al., 2025), intervenes on (Parsan et al., 2025; Shen et al., 2025), and traces (Lu et al., 2026; Tsui et al., 2026b) biological features in protein models, by testing model-derived hypotheses on unseen members of a biological class. This proposed workflow applies wherever a model has learned a natural system we want to understand. Foundation models absorb far more of that system than they are trained to predict, and our workflow converts that implicit knowledge into out-of-sample testable hypotheses, needing only an interpretable basis to perturb and an output whose change can be scored. Sparse autoencoders now exist for single-cell, chemical, and climate models, among others, where the same workflow could drive scientific discovery via mechanistic interpretability.

Future work. Our interventions establish structural effects within ESMFold2, but do not establish mechanisms of cellular folding. Future work could use the selected ablations to guide amino-acid substitutions that may alter GroEL/GroES dependence. Whereas sequence optimization often starts with a reward that directly scores a desired property, selected ablations could provide a more indirect route to sequence design: identifying candidate mutations that may elicit a biological property through a foundation model's learned associations. Testing folding yield with and without GroE depletion would assess whether these mutations have the hypothesized effect.

Joint ablations produce larger structural effects and stronger client-control separation than individual terms, suggesting that combinations may provide targets for understanding GroEL/GroES-assisted folding behavior. Some panels in the joint effect experiments also produce effects larger than their separate-term reference (Appendix B.4). Future work could test whether shared features, residue-level activation patterns, or proxy predictions help identify these larger combinations before evaluating them directly.

Finally, tracing the selected features through the remaining encoder layers and folding trunk could identify the downstream computations responsible for their structural effects.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025. doi: 10.1101/2025.02.06.636901. URL https://www.biorxiv.org/content/early/2025/02/08/2025.02.06.636901.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020. URL https://arxiv.org/abs/1810.03292.

Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation, 2022. URL https://arxiv.org/abs/2212.04629.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235.

J Buchner, M Schmidt, M Fuchs, R Jaenicke, R Rudolph, F X Schmid, and T Kiefhaber. Groe facilitates refolding of citrate synthase by suppressing aggregation. *Biochemistry*, 30(6):1586–1591, 1991. ISSN 0006-2960. doi: 10.1021/bi00220a020.

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. Proxyspex: Inference-efficient interpretability via sparse feature interactions in llms, 2025. URL https://arxiv.org/abs/2505.17495.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://www.biorxiv.org/content/early/2026/06/04/2026.06.03.729735.

Kausik Chakraborty, Manal Chatila, Jyoti Sinha, Qiaoyun Shi, Bernhard C Poschner, Martin Sikor, Guoxin Jiang, Don C Lamb, F Ulrich Hartl, and Manajit Hayer-Hartl. Chaperonin-catalyzed rescue of kinetically trapped states in protein folding. *Cell*, 142(1):112–122, 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.05.027.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), Sept 2008. ISSN 1932-6157. doi: 10.1214/07-aoas148. URL http://dx.doi.org/10.1214/07-AOAS148.

Kei Fujiwara, Yasushi Ishihama, Kenji Nakahigashi, Tomoyoshi Soga, and Hideki Taguchi. A systematic survey of in vivo obligate chaperonin-dependent substrates. *The EMBO journal*, 29(9): 1552–1564, 2010. ISSN 1460-2075. doi: 10.1038/emboj.2010.52.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. KernelSHAP-IQ: Weighted least square optimization for shapley interactions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 14308–14342. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/fumagalli24a.html.

Luqin Gan, Tarek M. Zikry, and Genevera I. Allen. Are machine learning interpretations reliable? a stability study on global interpretations, 2025. URL https://arxiv.org/abs/2505.15728.

Florian Georgescauld, Kristina Popova, Amit J Gupta, Andreas Bracher, John R Engen, Manajit Hayer-Hartl, and F Ulrich Hartl. Groel/ES chaperonin modulates the mechanism and accelerates the rate of TIM-barrel domain folding. *Cell*, 157(4):922–934, 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.03.038.

Ali Gorji, Andisheh Amrollahi, and Andreas Krause. Shap values via sparse fourier representation, 2025. URL https://arxiv.org/abs/2410.06300.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025. doi: 10.1073/pnas.2506316122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2506316122.

R Hlodan, P Tempst, and F U Hartl. Binding of defined regions of a polypeptide to groel and its implications for chaperonin-mediated protein folding. *Nature structural biology*, 2(7):587–595, 1995. ISSN 1072-8368. doi: 10.1038/nsb0795-587.

Walid A. Houry, Dmitrij Frishman, Christoph Eckerskorn, Friedrich Lottspeich, and F. Ulrich Hartl. Identification of in vivo substrates of the chaperonin groel. *Nature*, 402(6758):147–154, 1999. doi: 10.1038/45977. URL https://doi.org/10.1038/45977.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Kannan Ramchandran, and Bin Yu. Spex: Scaling feature interaction explanations for llms, 2025. URL https://arxiv.org/abs/2502.13870.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Michael J Kerner, Dean J Naylor, Yasushi Ishihama, Tobias Maier, Hung-Chun Chang, Anna P Stines, Costa Georgopoulos, Dmitrij Frishman, Manajit Hayer-Hartl, Matthias Mann, and F Ulrich Hartl. Proteome-wide analysis of chaperonin-dependent protein folding in escherichia coli. *Cell*, 122(2):209–220, 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.05.028.

Yali Li, Xinfeng Gao, and Lingling Chen. Groel recognizes an amphipathic helix and binds to the hydrophobic side*. *Journal of Biological Chemistry*, 284(7):4324–4331, 2009. ISSN 0021-9258. doi: https://doi.org/10.1074/jbc.M804818200. URL https://www.sciencedirect.com/science/article/pii/S0021925820710343.

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. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Kevin Lu, Jannik Brinkmann, Stefan Huber, Aaron Mueller, Yonatan Belinkov, David Bau, and Chris Wendler. Two stages of folding: Convergent mechanisms in ai protein folding trunks, 2026. URL https://arxiv.org/abs/2602.06020.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47), November 2022. ISSN 1091-6490. doi: 10.1073/pnas.2206625119. URL http://dx.doi.org/10.1073/pnas.2206625119.

Fumihiro Motojima, Yuko Motojima-Miyazaki, and Masasuke Yoshida. Revisiting the contribution of negative charges on the chaperonin cage wall to the acceleration of protein folding. *Proceedings of the National Academy of Sciences*, 109(39):15740–15745, 2012. doi: 10.1073/pnas.1204547109. URL https://www.pnas.org/doi/abs/10.1073/pnas.1204547109.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900654116.

OpenAI. GPT-5.6 System Card. Technical report, OpenAI, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf.

Ryan O'Donnell. *Boolean Functions and the Fourier Expansion*, page 1–25. Cambridge University Press, 2014.

Nithin Parsan, David J. Yang, and John J. Yang. Towards interpretable protein structure prediction with sparse autoencoders, 2025. URL https://arxiv.org/abs/2503.08764.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. doi: 10.1101/2020.12.15.422761. URL https://www.biorxiv.org/content/early/2020/12/15/2020.12.15.422761.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. ISSN 2169-3536. doi: 10.1109/access.2020.2976199. URL http://dx.doi.org/10.1109/ACCESS.2020.2976199.

Matthew Shen, Aliyah Hsu, Abhineet Agarwal, and Bin Yu. Adaptive test-time intervention for concept bottleneck models, 2025. URL https://arxiv.org/abs/2503.06730.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, 2025. doi: 10.1038/s41592-025-02836-7. URL https://doi.org/10.1038/s41592-025-02836-7.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Yan Shuo Tan, Chandan Singh, Keyan Nasseri, Abhineet Agarwal, James Duncan, Omer Ronen, Matthew Epland, Aaron Kornblith, and Bin Yu. Fast interpretable greedy-tree sums. *Proceedings of the National Academy of Sciences*, 122(7):e2310151122, 2025. doi: 10.1073/pnas.2310151122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2310151122.

Yun-Chi Tang, Hung-Chun Chang, Kausik Chakraborty, F Ulrich Hartl, and Manajit Hayer-Hartl. Essential role of the chaperonin folding compartment in vivo. *The EMBO journal*, 27(10):1458–1468, 2008. ISSN 1460-2075. doi: 10.1038/emboj.2008.77.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 01 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Philip To, Yingzi Xia, Sea On Lee, Taylor Devlin, Karen G. Fleming, and Stephen D. Fried. A proteome-wide map of chaperone-assisted protein refolding in a cytosol-like milieu. *Proceedings of the National Academy of Sciences*, 119(48):e2210536119, 2022. doi: 10.1073/pnas.2210536119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2210536119.

Darin Tsui, William Deinzer, Daniel Saeedi, and Amirali Aghazadeh. Circuit tracing in autoregressive protein language models, 2026a. URL https://arxiv.org/abs/2606.16044.

Darin Tsui, Kunal Talreja, Daniel Saeedi, and Amirali Aghazadeh. Protein circuit tracing via cross-layer transcoders, 2026b. URL https://arxiv.org/abs/2602.12026.

UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617, 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1010.

Z Wang, H p Feng, S J Landry, J Maxwell, and L M Gierasch. Basis of substrate binding by the chaperonin groel. *Biochemistry*, 38(39):12537–12546, 1999. ISSN 0006-2960. doi: 10.1021/bi991070p.

Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020. doi: 10.1073/pnas.1901326117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1901326117.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, 2004. ISSN 1097-0134. doi: 10.1002/prot.20264.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, 33(7):2302–2309, 2005. ISSN 1362-4962. doi: 10.1093/nar/gki524.

AI USE STATEMENT

In this work, we used generative AI tools to propose and refine biological hypotheses (explicitly for interpretation generation; Sections 3.3 and 4.4 and appendix B.5), to design and provide feedback on research methodology and experiments, to implement methods, to interpret results, and to support qualitative and thematic analysis of the selected-term interpretations. We did not use generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide ingredients for or write proofs, or to clean or reformat the protein cohort data; translation is not applicable to this work. Additionally, we used generative AI tools for code development, figure preparation, literature search and synthesis, and manuscript drafting and editing.

The authors reviewed all AI-assisted work, specifically AI-assisted code and verified all reported numbers and figures. Biological interpretations were checked against the cited literature, and unsupported mappings were retained as abstentions. All AI-assisted text was reviewed and edited by the authors, who take responsibility for the final contents of the paper.

A EXPERIMENTAL SETUP AND INTERACTION DISCOVERY

A.1 PROTEIN COHORTS

Discovery and holdout clients. We began with the 57 *E. coli* Class IV proteins reported by (Fujiwara et al., 2010); 33 had audited PDB targets (Berman et al., 2000). MMseqs2 components were defined at 20% sequence identity and 80% bidirectional coverage (Steinegger and Söding, 2017). All five two-protein MMseqs2 components were retained in the discovery cohort, leaving 23 singleton candidates. We repeatedly drew groups of eight from the 23 singleton candidates, without replacement within each draw. We retained the first group that contained exactly one electron-microscopy reference and two heteromeric PDB contexts, included at most one protein from each of four frozen fold/function strata, and balanced sequence length and non-generic labeled-SAE feature counts at 20% and 25% residue coverage between holdout and discovery ($|\mathrm{SMD}| \leq 0.15$). For each of these three covariates, the holdout range was also required to span the full cohort's Q_1–Q_3 interval. The first qualifying group occurred at trial 68,072. No intervention response, proxy prediction, or interaction coefficient entered selection.

The 25 discovery genes were *bioF*, *csdB*, *dapA*, *dusB*, *dusC*, *eutB*, *frdA*, *ftsE*, *gdhA*, *hemB*, *kdsA*, *metF*, *nanA*, *nfo*, *nuoC*, *pepQ*, *pyrC*, *pyrD*, *sdhA*, *serC*, *tatD*, *tldD*, *tldE*, *ycfH*, and *yfbQ*. The eight holdout genes were *araA*, *argE*, *asd*, *deoA*, *fabF*, *ltaE*, *metK*, and *yneB*. These eight proteins were excluded from feature filtering, proxy fitting, term extraction, and term selection in the reported pipeline.

Controls. Controls were drawn from manually curated UniProtKB/Swiss-Prot entries for cytosolic, PDB-backed *E. coli* K-12 proteins. We excluded known Class IV proteins, mapped GroEL interactors, close full-length Class IV homologs, and a small set of proteostasis proteins. Eligible proteins were restricted to the discovery-client length and PDB-quality ranges, deduplicated by sequence component, and matched to discovery clients on sequence length and low-dimensional summaries of SAE-feature coverage and activation. Neither folding outcomes nor intervention responses entered the selection procedure. One initially selected protein, *yedJ*, was removed because its PDB-chain alignment coverage was below the prespecified 0.95 threshold.

The final 19 controls were *rnd*, *glpK*, *glnA*, *rbsR*, *mtnN*, *bioD1*, *purD*, *nfsA*, *cysI*, *murE*, *gabT*, *cysJ*, *pdxK*, *ycjR*, *guaD*, *nemA*, *allC*, *prpB*, and *ybdL*. They are label-screened controls lacking identified GroEL/GroES dependence, not experimentally certified non-clients.

A.2 ESMFOLD2 AND SAE INTERVENTIONS

The frozen model artifacts and inference settings are listed in Table 2. Exact software environments and cohort manifests are recorded with the accompanying artifacts.

Table 2: Frozen model and inference configuration.

Component	Configuration
ESMFold2	`biohub/ESMFold2` Revision: `8fc3ff471022fdce52c77030685eb775de0c00a3`
ESMC-6B	`biohub/ESMC-6B` Revision: `45b0fa5d7fb06faefbd5e3b89bdcef35d564e79a`
Sparse autoencoder	`biohub/ESMC-6B-sae-layer60-k64-codebook16384` Revision: `99752fe6e4d25fbb26f887db2d9225e7a577da73`
Intervention	Output of transformer block 59 (residual labeled layer 60); model dimension 2,560; top-K $K = 64$; codebook size 16,384
Structure generation	Monomer input; seed 0; 20 folding loops; 100 diffusion steps; 3 diffusion samples; language-model dropout 0

For each residue, the SAE maps the layer-60 representation x_r to a sparse code $z_r \in \mathbb{R}^{16384}$. For an ablation set $S \subseteq \mathcal{F}$, we set the coordinates in S to zero at every sequence position, decode the edited code, and return it as the output of transformer block 59. ESMC blocks 60–79 are then recomputed before the edited representations pass through the ESMFold2 folding trunk and diffusion module. The amino-acid sequence is unchanged.

The no-op baseline follows the same SAE encode–decode–reinsert path without zeroing a feature. It is therefore a lossy SAE round trip rather than vanilla ESMFold2, and all effects are measured relative to this no-op. Structural agreement TM-score (Zhang and Skolnick, 2005) was computed with TM-align through `tmtools` ; TM-score was normalized by the audited experimental target-chain length. As stated in Equation (1), no denominator floor or clipping was applied to the headroom-normalized response.

A.3 Feature basis, ProxySPEX, and stability screening

The feature basis $\mathcal{F}$ was defined using discovery clients only. An SAE feature was eligible when its released description was not generic and it was active on at least 5% of residues in at least 20 of the 25 discovery clients. This produced 78 candidate SAE features. Features outside this basis remained available to the SAE and were not ablated.

We generated two batches of 1,000 masks, using seeds 0 and 1. The first batch included the empty and all-feature masks; all other masks independently included each of the 78 features with probability 0.5. The same 2,000 masks were evaluated on every discovery client (25 clients), yielding 50,000 ESMFold2 interventions.

ProxySPEX (Butler et al., 2025) fit a LightGBM (Ke et al., 2017) quantile proxy model from the 78 mask indicators to the signed structural response, with quantile level $\alpha = 0.25$. Protein identity and protein covariates were not used for proxy fitting. 5-fold mask-grouped cross-validation selected a learning rate of 0.1 from [0.01, 0.1], maximum depth 3 from [3, 5], and 500 boosting iterations from [500, 1000]. ProxySPEX then extracted marginal, pairwise, and three-way k-SII terms from the fitted proxy model, via Gorji et al. (2025). The proxy model was used to nominate candidate terms; every effect reported for the holdout cohorts was obtained from a new, direct ESMFold2 intervention.

For each of 1,000 bootstrap replicates, we resampled complete mask rows with replacement. The empty and full-ablation masks were included exactly once per replicate. We then refit LightGBM and re-extracted the interaction terms (up to, but including, order 3). A term s was retained when

$$\widehat{\Pr}(s \text{ appears}) \geq 0.50 \quad \text{and} \quad 0 \notin [\widehat{q}_{0.125}(\beta_s), \widehat{q}_{0.875}(\beta_s)], \tag{2}$$

where the coefficient quantiles condition on the term appearing in a replicate. This screen produced the frozen panel $\mathcal{S}^*$ of 67 terms: 28 marginals, 38 pairs, and 1 triple.

A.4 SAE no-op baseline fidelity

Because the SAE no-op is a lossy encode–decode–reinsert path, we evaluate its fidelity in a separate matched-draw diagnostic. Figures 6 and 7 compare vanilla ESMFold2 with the SAE no-op using the same fixed folding seed and one matched conditional diffusion draw per protein. This diagnostic is separate from the reported intervention effects, for which three conditional diffusion draws are averaged within each condition before the effect is computed. This is mostly for visual context. Across the 25 discovery clients, mean TM-score to the experimental chain was 0.978 for vanilla ESMFold2 and 0.973 for the SAE no-op (mean paired change, -0.0046); no-op TM-scores ranged from 0.927 to 0.995. Mean no-op pLDDT was 94.9, with a mean change of -2.15 points. Thus, the SAE "round trip" modestly reduced confidence while preserving high global structural agreement. All downstream effects are contrasts against the matched SAE no-op, not against vanilla ESMFold2.

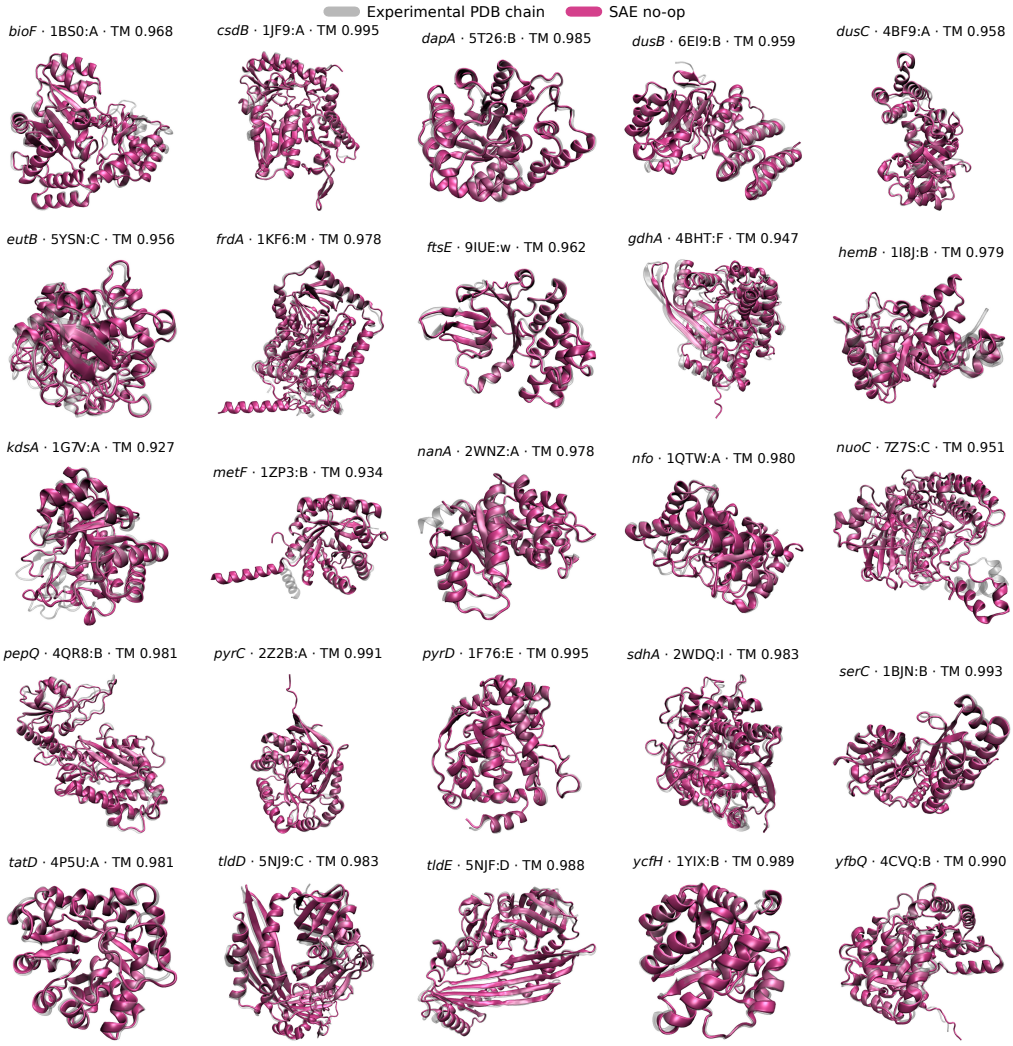


Figure 6: **Matched-draw SAE no-op predictions retain high global structural agreement with experimental targets across the 25 discovery proteins.** Each panel overlays one SAE no-op prediction (magenta) with its TM-aligned experimental PDB chain (gray). Labels report gene, PDB ID and chain, and TM-score normalized by experimental-chain length. This single-draw, fixed-seed diagnostic is separate from the three-draw estimator used for all reported intervention effects.

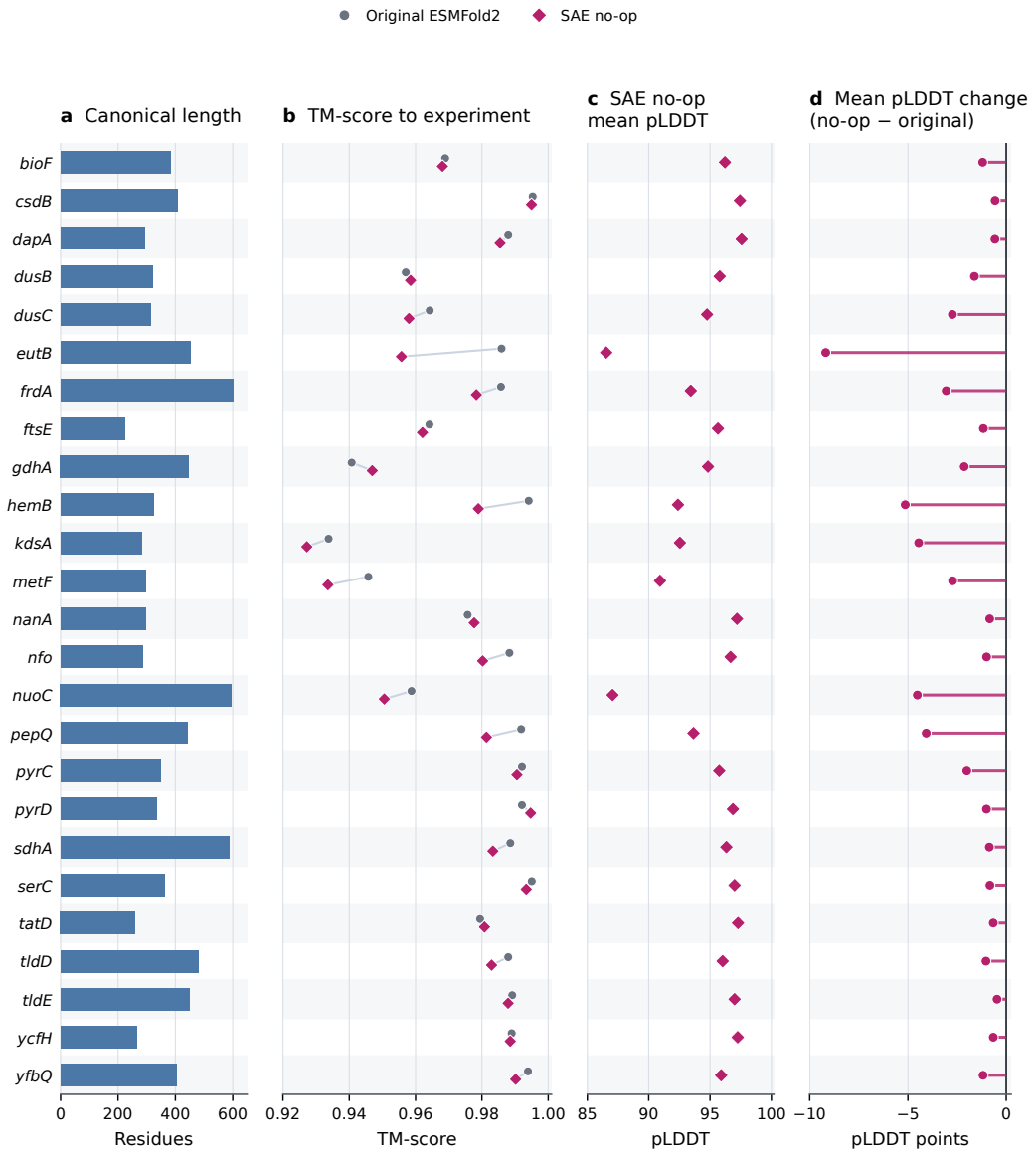


Figure 7: **Quantitative matched-draw fidelity of the SAE no-op path.** (a) Canonical sequence length. (b) TM-score of vanilla ESMFold2 and SAE no-op predictions against the experimental chain; lines join the matched predictions. (c) Mean no-op pLDDT. (d) No-op minus vanilla mean pLDDT. Each path uses one matched fixed-seed draw in this diagnostic; downstream intervention effects use three conditional draws per condition.

B HOLDOUT EVALUATION AND BIOLOGICAL INTERPRETATION

B.1 DIRECT HELD-OUT EVALUATION

For retained term s^*, we performed a fresh joint ablation in each of the eight held-out clients and 19 controls. Let

$$e_{\mathcal{C}s^*} = \frac{1}{|\mathcal{C}|} \sum_{p \in \mathcal{C}} |v_{ps^*}| \tag{3}$$

be the mean absolute effect for cohort $\mathcal{C}$. For a pair or triple, v_{ps^*} is the direct whole-set ablation effect as seen in Equation (1). Specifically, imagine keeping all relevant SAE features active except for the features that compose s^* (zeroing them out). It is neither the fitted k-SII coefficient nor a pure factorial interaction contrast.

Figure 2a compares the resulting 67 paired term-level effects. The one-sided exact Wilcoxon signed-rank test treats the selected term as the paired unit and tests whether $e_{\text{client},s^*} > e_{\text{control},s^*}$. It yielded $p = 2 \times 10^{-12}$; 56 of 67 terms favored clients. Because the terms reuse the same proteins and sometimes the same SAE features, these observations are not independent samples from a protein population. The test supports a difference between the selected-term effects within the fixed holdout cohorts; with eight holdout clients, separate population-level uncertainty over proteins remains limited.

B.2 ORDER-MATCHED RANDOM PANELS

The random-panel analysis asks whether the selected panel exceeds other evaluated terms of the same orders. Eligible pools excluded selected terms and required complete direct effects in both cohorts, leaving 50 marginals, 150 pairs, and 50 triples to sample random panels from. Each of 1000 redraws sampled without replacement 28 marginals, 38 pairs, and one triple. For panel $\mathcal{P}$, we computed

$$E_{\mathcal{C}}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{s \in \mathcal{P}} e_{\mathcal{C}s}, \qquad R_{\mathcal{C}b} = \frac{E_{\mathcal{C}}(\mathcal{S}^*)}{E_{\mathcal{C}}(\mathcal{P}_b)}. \tag{4}$$

For clarity, $b = 1, \ldots, 1000$, and the CI's in Figure 2b are from the draws of random order-matched panels of interactions. In held-out clients, the median selected-to-random ratio was 1.593, with a central 95% redraw range of $[1.310, 1.964]$. In controls, the corresponding ratio was 1.288, with range $[1.211, 1.367]$. These ranges describe variation across order-matched panels, not confidence intervals over proteins.

B.3 PROTEIN-SPECIFIC EFFECT AND RATIO RESULTS

The effect and ratio plots in the main text aggregate over held-out client proteins. Figure 8 displays the held-out client proteins individually, repeating the direct-effect and selected-to-random analyses for each client, using the same 19-protein control reference. Figure 9 displays the non headroom normalized effect and ratio plots; however, discovery still occurred with the normalized value function.

As seen in Figure 8, the largest effects occur in araA, asd, metK, and yneB, and most significant ratios occur in araA, deoA, fabF, and yneB. As seen in Figure 9, the largest effects occur in araA, metK, and yneB, and most significant ratios occur in araA, deoA, fabF, and yneB.

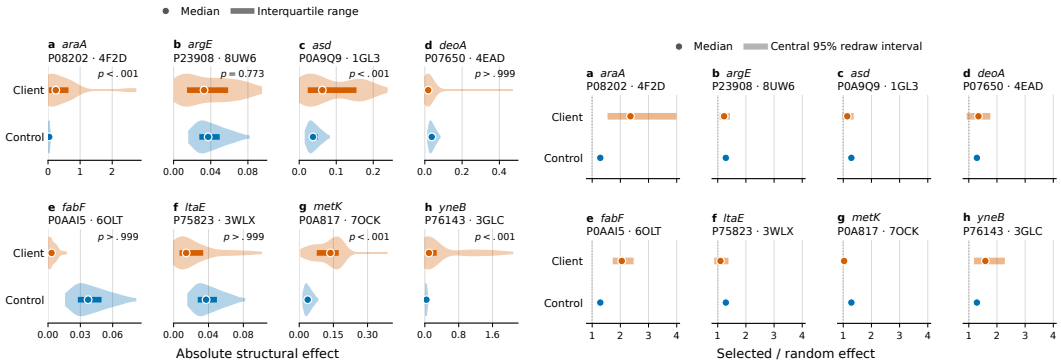


Figure 8: **Protein-specific normalized effects of the 67 selected terms and selected-to-random effect ratios.** On the left, aach panel compares one held-out client protein (orange) with the per-term mean across 19 controls (blue) for the same 67 terms. Violins show term-level distributions, dots mark medians, and thick bars mark interquartile ranges. Horizontal scales vary across panels. Displayed p-values are nominal one-sided paired Wilcoxon tests across matched terms and are not adjusted across proteins. On the right, for each client, the fixed mean absolute effect of the 67 selected terms is divided by the mean effect of each of 1,000 random panels with the same 28/38/1 order composition. Dots mark medians and bars mark central 95% redraw intervals. The blue row repeats the pooled 19-control comparison, and the dotted line marks a ratio of one. These intervals quantify random-panel composition, not protein-population uncertainty.

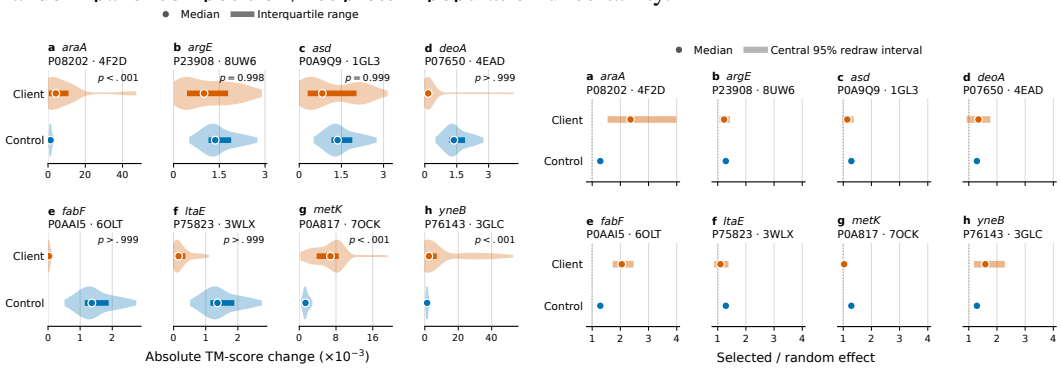


Figure 9: **Protein-specific non-normalized effects of the 67 selected terms and selected-to-random effect ratios.** On the left, aach panel compares one held-out client protein (orange) with the per-term mean across 19 controls (blue) for the same 67 terms. Violins show term-level distributions, dots mark medians, and thick bars mark interquartile ranges. Horizontal scales vary across panels. Displayed p-values are nominal one-sided paired Wilcoxon tests across matched terms and are not adjusted across proteins. On the right, for each client, the fixed mean absolute effect of the 67 selected terms is divided by the mean effect of each of 1,000 random panels with the same 28/38/1 order composition. Dots mark medians and bars mark central 95% redraw intervals. The blue row repeats the pooled 19-control comparison, and the dotted line marks a ratio of one. These intervals quantify random-panel composition, not protein-population uncertainty.

B.4 Protein-specific joint results

The effect and ratio plots in the main text aggregate over held-out client proteins. Figure 10 displays the held-out client proteins individually, repeating the joint effect analyses for each client, using the same 19-protein control reference. Figure 11 displays the non headroom normalized effect and ratio plots; however, discovery still occurred with the normalized value function.

As seen in Figure 10, the largest joint effects occur in araA, asd, ltaE, metK, and yneB, and most significant effects retained occur in araA, argE, deoA, fabF, and ltaE. As seen in Figure 11, the largest joint effects occur in araA, metK, and yneB, and most significant effects retained occur in araA, argE, asd, deoA, fabF, and ltaE. Across both figures, effects retained even surpass 100%, suggesting a synergy between terms.

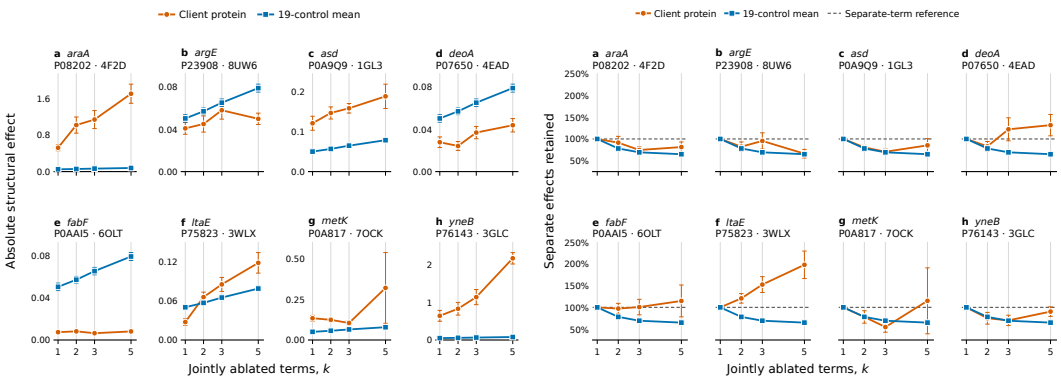


Figure 10: **Protein-specific accumulation of normalized joint-ablation effects.** On the left, each panel shows one held-out client (orange) and the protein-balanced mean across 19 controls (blue). Points show the mean absolute structural effect across 20 prespecified nested panels; bars show standard errors across panels. The intervention ablates the union of features in the first $k \in \{1, 2, 3, 5\}$ terms, so k counts selected multi-feature terms rather than unique SAE features. Vertical scales vary across proteins and begin at zero. On the right, the percentage is the directly measured joint-ablation magnitude divided by an overlap-corrected reference constructed from the same terms ablated separately. Thus, 100% means that the joint effect equals that constituent reference; values below 100% mean that some separately measured effect does not carry over, and values above 100% mean that the joint effect exceeds the reference but do not by themselves establish biological synergy. Bars are paired-panel bootstrap standard errors from 50,000 resamples of the 20 fixed panels. Protein-level estimates are shown to characterize heterogeneity, not to claim a common response in every client.

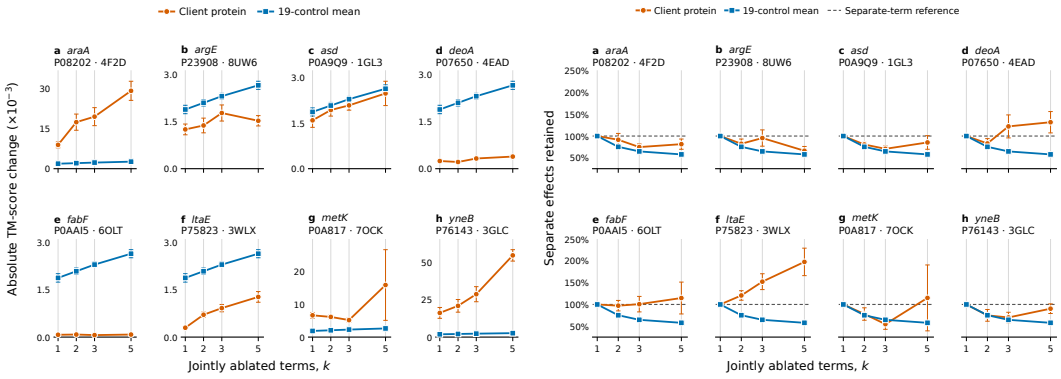


Figure 11: **Protein-specific accumulation of non-normalized joint-ablation effects.** On the left, each panel shows one held-out client (orange) and the protein-balanced mean across 19 controls (blue). Points show the mean absolute structural effect across 20 prespecified nested panels; bars show standard errors across panels. The intervention ablates the union of features in the first $k \in \{1, 2, 3, 5\}$ terms, so k counts selected multi-feature terms rather than unique SAE features. Vertical scales vary across proteins and begin at zero. On the right, the percentage is the directly measured joint-ablation magnitude divided by an overlap-corrected reference constructed from the same terms ablated separately. Thus, 100% means that the joint effect equals that constituent reference; values below 100% mean that some separately measured effect does not carry over, and values above 100% mean that the joint effect exceeds the reference but do not by themselves establish biological synergy. Bars are paired-panel bootstrap standard errors from 50,000 resamples of the 20 fixed panels. Protein-level estimates are shown to characterize heterogeneity, not to claim a common response in every client.

B.5 Interpretation protocol and scope

Biological interpretation was performed only after the term panel and held-out evaluations were frozen. GPT-5.6 Sol in Codex Ultra mode (OpenAI, 2026) was given all 67 selected terms, the Biohub API descriptions and saved full metadata for their component features, and a focused set of GroEL/GroES references. It was instructed to identify a literature-linked component, formulate a testable hypothesis, or abstain. Figures 3 and 4 show a representative subset, while Table 3 reports

every term. These interpretations translate validated structural effects into experimental hypotheses. They do not independently validate the SAE descriptions or establish cellular folding mechanisms.

Specifically, we use the following prompt:

Prompt given to the model for interpretation

```
For each of the 67 selected terms, use the Biohub SAE labels and
descriptions to identify connections to established GroEL/ES client
properties reported in foundational, widely cited, and directly
relevant biological literature. Consider composition, structural
architecture, hydrophobic exposure, aggregation propensity, and
folding kinetics.
A connection to a known property does not require prior validation
of the exact SAE feature or interaction. For each term, report
separately:
1. Known-property connection: Identify the published property, cite
a primary reference, and explain whether the connection concerns
one component or the combined term.
2. New hypothesis: State what additional claim about localization,
coupling, or mechanism remains untested, and propose a concrete
experimental prediction.
3. Limits or abstention: Flag ambiguous annotations and unsupported
assumptions; abstain where no defensible connection or hypothesis
is available.
A term may have both a known-property connection and a new
hypothesis. Distinguish broad literature correspondence from
evidence for the specific mechanism, and do not abstain merely
because the exact SAE term has not previously been studied.
Generate a table for with SAE term(s), interpretation, and
relationship to prior work. Also, select a subset of 3 well
established/reinforced interpretations, and 3 interesting new novel
hypothesis.
```

SAE descriptions are model-derived and can be polysemantic. Moreover, a fitted k-SII coefficient is an allocated contribution in the discovery game: a negative three-way coefficient does not by itself establish a physical conditional reversal or determine the sign of a direct joint intervention. Accordingly, the interpretations in Figures 3 and 4 and Table 3 should be treated as hypotheses for mutational or folding experiments, not as inferred contacts or established GroEL/GroES mechanisms.

Finally, the interventions are causal with respect to ESMFold2's computation, not a cellular folding trajectory. The reported evaluation uses one model and SAE, eight held-out clients, label-screened rather than certified non-client controls, and three conditional diffusion samples per condition, averaged at the TM-score level before differencing or normalization. These choices constrain the scope of the mechanistic and population-level claims.

Table 3: **Literature-grounded agentic AI generated interpretations of all 67 selected SAE terms.** The table includes 28 marginals, 38 pairs, and one triple. Each row reports an interpretation based on the released SAE descriptions and the closest primary-literature precedent, or records an explicit abstention. A cited correspondence may concern only one component and does not validate the combined term. Yellow-shaded rows mark three hypotheses prioritized for experimental follow-up. Each builds on prior biological precedent, but the proposed localization or coupling remains untested. Shading does not imply literature-wide novelty.

SAE term	Interpretation	Closest literature or disposition
Small-residue, boundary, and packing hypotheses		

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F13958	Conditional on client-localized cap activation: a polar cap stabilizes an incompletely packed intermediate; improving the cap-to-core packing transition reduces the time during which adjacent hydrophobic surfaces require GroEL protection.	Kinetic-trap and segmental-folding precedents: (Kerner et al., 2005; Chakraborty et al., 2010; Georgescauld et al., 2014). Polar-cap localization and cap-to-core coupling remain untested.
F15534	If the bacterial activations map to genuine domain-start helices, inefficient helix nucleation at the boundary creates a slow folding intermediate that GroEL rescues.	Domain-scale kinetic-trap precedent: (Kerner et al., 2005; Chakraborty et al., 2010; Georgescauld et al., 2014). A domain-start-helix bottleneck is new and conditional on residue mapping.
F2494	A flexible boundary extension interferes with productive core closure, extending the lifetime of a GroEL-rescuable folding intermediate.	Slow-folding and segmental-structure precedents: (Kerner et al., 2005; Georgescauld et al., 2014). Flexible boundary extensions are not a reported client signature.
F3013	A subset of activated Gly/Pro loops slows closure of the adjacent core, generating kinetic traps whose occupancy determines GroEL rescue.	The Gly component weakly corresponds to Class-IV Ala/Gly enrichment (Fujiwara et al., 2010); a local Gly/Pro-loop trap is not established. Kinetic precedent: (Georgescauld et al., 2014).
F6656	Boundary-localized small-residue runs increase conformational search time before the neighboring domain consolidates, increasing GroEL rescue relative to composition-matched runs located away from that boundary.	Only the small-residue component overlaps the weak Class-IV Ala/Gly association (Fujiwara et al., 2010); boundary localization and conformational-search mechanism are new.
F13480 × F16243	If small residues and helical Leu co-map to one core, mismatched single edits should increase kinetic exposure, aggregation and GroE rescue, whereas a compensatory double edit that restores packing should reduce dependence at matched native stability.	F13480 has a component-level correspondence to the weak Class-IV Ala/Gly association (Fujiwara et al., 2010). Small-residue/Leu core packing and compensatory coupling are the prioritized novel hypothesis.
F13958 × F14234	If F13958 marks a cap directly adjoining a soluble F14234 core, inefficient cap formation may prolong core exposure and increase GroE-dependent folding. The proposed interaction is between boundary closure and the underlying structured segment.	Segmental folding supplies a kinetic precedent (Georgescauld et al., 2014); a polar cap adjoining a soluble ordered core and controlling its exposure is untested.
F1552 × F13480	If the alanine signal lies within a charged amphipathic helix, residue patterning may tune presentation of a GroEL-binding face. An alanine-supported helix with clustered hydrophobic residues is predicted to bind GroEL differently from a composition-matched helix with dispersed hydrophobic residues.	Separate precedents are Ala/Gly enrichment (Fujiwara et al., 2010) and amphipathic hydrophobic-face recognition (Wang et al., 1999; Li et al., 2009); their patterning and coupling are new.
F6620 × F6768	A weakly capped ordered element repeatedly frays into a nonnative intermediate; cap stabilization should reduce GroE dependence preferentially when the adjacent element has suboptimal packing.	Kinetic-trap and segmental-folding precedents: (Chakraborty et al., 2010; Georgescauld et al., 2014). Edge-to-element adjacency, fraying, and nonadditivity are untested.

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F6620 × F9237	Gly-mediated flexibility and charged helix-edge chemistry jointly delay helix nucleation, increasing the interval during which GroE can capture a folding intermediate.	F9237 corresponds at component level to Gly enrichment (Fujiwara et al., 2010); the charged-edge/Gly nucleation mechanism has only general kinetic precedent (Georgescauld et al., 2014).
F6768 × F13480	Small-residue enrichment affects GroE dependence specifically when it creates an underpacked ordered core, rather than through total Ala/Gly abundance alone.	Only F13480 directly corresponds to Class-IV Ala/Gly enrichment (Fujiwara et al., 2010); an underpacked ordered core and local coupling are new.
F8337 × F13480	Small-residue composition at a domain-edge cap controls productive nucleation; strengthening a weak cap should lower GroE dependence without changing global domain topology.	F13480 has a component-level Ala/Gly correspondence (Fujiwara et al., 2010); cap localization and cap-controlled nucleation have only general kinetic precedent (Georgescauld et al., 2014).
F9237 × F13958	Excess flexibility at a polar cap prolongs an incompletely closed folding intermediate; a cap-restoring edit should reduce GroE dependence in a flexibility-dependent manner.	F9237 has a component-level Gly correspondence (Fujiwara et al., 2010); flexible-cap closure has only general segmental-folding precedent (Georgescauld et al., 2014).
F9237 × F16243	A flexible connector delays closure over a Leu-rich core; reducing nonessential flexibility should lower GroE dependence most when core packing is marginal.	Separate precedents are Gly enrichment (Fujiwara et al., 2010) and hydrophobic exposure (Wang et al., 1999); connector-to-Leu-core closure is untested.
Hydrophobic exposure, α/β structure, interfaces, and aggregation		
F16243	At activated buried leucines in verified soluble client helices, hydrophobic packing geometry may control the lifetime of nonnative exposed patches and hence the magnitude of GroEL-assisted folding.	Nonnative hydrophobic-surface recognition is established (Houry et al., 1999; Wang et al., 1999); Leu identity, clustering, and exposure are not. Broad hydrophobic residues were not enriched in Class IV (Fujiwara et al., 2010).
F2888	Some activated functional pockets are slow to consolidate in the apo protein, and early ligand/cofactor binding bypasses the same kinetic bottleneck that GroEL normally rescues.	Most validated Class-IV clients are metabolic enzymes (Fujiwara et al., 2010); no study establishes functional-pocket localization as the causal determinant.
F4485	Conditional broad correspondence: if the activation marks an exposed soluble nonnative helix, its hydrophobic face is a plausible GroEL recognition element. The transmembrane or signal-peptide alternative prevents an SAE-specific assignment.	Conditional recognition precedent only: exposed nonnative hydrophobic or amphipathic helices can bind GroEL (Hlodan et al., 1995; Wang et al., 1999; Li et al., 2009). The membrane/signal-peptide alternative is excluded.
F4975	An activated interaction scaffold undergoes premature intermolecular association during folding; GroEL sequestration permits formation of an assembly-competent monomer.	GroEL anti-aggregation is established in model substrates and obligate clients are aggregation-prone (Buchner et al., 1991; Fujiwara et al., 2010); coiled-coil or repeat scaffolds are not a known client signature.

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F6366	In a subset of clients the activated interface becomes native only after GroEL-assisted folding, and premature partner binding instead traps an assembly-incompetent intermediate.	Metabolic enzymes are common among Class-IV clients (Fujiwara et al., 2010); late interface maturation or premature partner binding is an untested localization-specific mechanism.
F9207	An activated beta edge exposes an aggregation-prone surface before sheet closure; GroEL protects this intermediate, and improved edge shielding reduces the need for rescue.	The closest precedent is nonnative hydrophobic exposure in α/β substrates (Houry et al., 1999; Wang et al., 1999); a Pro-containing β-edge or shielding mechanism is not established.
F12861 × F14234	If F12861 aromatics map to a soluble F14234 core, delayed burial of an aromatic patch during core formation may increase GroEL capture and GroE-dependent folding. Hydrophobic surface recognition is established; this aromatic/core interaction is new.	Hydrophobic-surface recognition is established (Houry et al., 1999; Wang et al., 1999); aromatic identity, soluble-core localization, delayed burial, and the pair interaction are untested. Exclude membrane hits.
F12861 × F16243	If activated aromatic and leucine residues form the same surface in a folding intermediate, their joint hydrophobic presentation may determine GroEL capture more strongly than either residue class alone. The spatially organized patch, rather than total aromatic/leucine abundance, is the proposed determinant.	Clustered hydrophobic exposure is a recognition precedent (Wang et al., 1999); neither aromatic/Leu abundance nor their spatially organized pair is established as a client signature (Fujiwara et al., 2010).
F6366 × F12861	If an exposed aromatic functional interface is a late-closing capture surface, conservative aromatic edits should alter GroEL occupancy, while ligand or partner occupancy that shields the interface should advance protection and reduce GroE rescue.	Separate precedents are metabolic-enzyme enrichment (Fujiwara et al., 2010) and hydrophobic recognition (Wang et al., 1999). Aromatic interface exposure and late pocket closure are prioritized but high-risk hypotheses.
F9237 × F12861	A Gly-containing loop gates burial of an aromatic patch; reducing excess loop flexibility should shorten hydrophobic exposure and reduce GroE dependence.	Separate precedents are Gly enrichment (Fujiwara et al., 2010) and exposed hydrophobic recognition (Wang et al., 1999); loop-gated aromatic burial and spatial coupling are new.
F953 × F6768	A cationic amphipathic segment exposed before an ordered element packs is a transient GroEL-capture surface; improved packing should reduce capture and offset the effect of patch charge on folding partitioning.	Amphipathic hydrophobic-face recognition is established (Wang et al., 1999; Li et al., 2009); positive charge and ordered-element timing are not known dependence determinants.
Electrostatic and acidic conditional hypotheses		
F11911	Conditional on mapped activation in a client surface patch: concentrating positive charge prolongs nonproductive contacts with the negatively charged GroEL cavity, slowing productive folding or release relative to a dispersed patch.	Positive-charge enrichment was not reported for validated Class-IV clients (Fujiwara et al., 2010). Cavity-charge studies motivate, but do not establish, the proposed client-side mechanism (Tang et al., 2008; Motojima et al., 2012).

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F6620	Conditional on actual activation in a soluble, charged client helix edge, local charge patterning changes nonnative intrachain contacts and a kinetic folding trap that GroEL rescues.	Kinetic-trap and confinement precedents: (Chakraborty et al., 2010; Georgescauld et al., 2014). The soluble charged helix-edge interpretation must be mapped before testing local charge patterning.
F9000	The spatial arrangement of mapped client Glu sites may couple to GroEL cavity charge independently of whole-protein net charge; composition-preserving Glu relocation should interact with a characterized cavity-charge mutation in folding or release.	Class-IV clients have a global acidic pI distribution, but no local Glu-tract signature (Fujiwara et al., 2010). Cavity-charge work (Tang et al., 2008; Motojima et al., 2012) motivates this prioritized client/cavity coupling hypothesis.
F583 × F10856	If the Asp-rich tract is the domain-boundary linker, local negative charge may suppress premature association of adjacent folding domains. Neutralizing that charge is predicted to increase GroE-dependent folding most when linker geometry favors premature domain contacts.	The closest precedents are the global acidic distribution of Class-IV clients (Fujiwara et al., 2010) and segmental folding (Georgescauld et al., 2014); local Asp-linker coupling is new.
F583 × F12861	If activated Asp residues flank an exposed aromatic patch, the acidic sequence may act as a local solubility buffer that reduces unproductive association during folding. Removing local negative charge is predicted to increase GroE dependence more strongly when the aromatic patch is retained.	Exposed hydrophobic recognition is established (Wang et al., 1999), whereas local acidic buffering is not. Spatial adjacency, aromatic identity, and charge-by-patch coupling remain untested.
F583 × F16243	If the Asp-rich segment flanks a Leu-containing core, its negative charge suppresses intermolecular association while that core folds; removing charge will increase GroE dependence especially when core packing is weakened.	Hydrophobic exposure is a recognition precedent (Wang et al., 1999); an adjacent Asp-rich solubility buffer and its interaction with Leu-core packing are untested.
F583 × F9237	Within a shared flexible loop, negative charge and backbone flexibility jointly set the competition between soluble folding and aggregation; charge removal should be most harmful in the more flexible loop background.	F9237 has a component-level correspondence to Gly enrichment (Fujiwara et al., 2010); shared-loop localization and charge-by-flexibility coupling are new.
F6366 × F9000	If activated Glu residues surround a functional interface, their local charge may prevent off-pathway association before ligand/partner-assisted closure. Neutralization should raise GroE dependence mainly without interface stabilization.	Metabolic-enzyme enrichment (Fujiwara et al., 2010) and cavity-charge effects (Tang et al., 2008; Motojima et al., 2012) are separate precedents; acidic-interface buffering is untested.
F9000 × F14234	If activated Glu residues are exposed near an incompletely packed soluble core, their charge may limit off-pathway association. Neutralization should increase GroE dependence especially in a mildly destabilized core.	Global client acidity and cavity-charge studies provide only broad context (Fujiwara et al., 2010; Tang et al., 2008); local Glu exposure beside a soluble underpacked core is untested.
F9000 × F9237 × F16243	If mapped Glu, Gly and helical Leu sites participate in one folding unit, Glu-mediated solubilization may buffer slow closure of a Gly-linked Leu-containing core. Neutralizing selected surface Glu residues should increase GroE dependence most in the flexible, weakly packed background; reducing flexibility or restoring packing should suppress that effect.	Gly enrichment, hydrophobic recognition, and cavity charge are separate precedents (Fujiwara et al., 2010; Wang et al., 1999; Tang et al., 2008). Their three-way localization and epistasis are untested.
Broad, label-sensitive, or low-information terms retained as controls		

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F10583	Abstain because the two Biohub annotations disagree: an IAPP-family motif and a class-independent structural-boundary detector suggest different biological explanations.	Abstain: the IAPP/signal-peptide and generic-boundary annotations conflict, and neither is a validated bacterial GroEL-client property.
F10749	Mature-chain or domain-boundary activation is insufficient to distinguish a GroEL client from an ordinary folded bacterial protein.	Generic mature-domain or boundary activation is not the multidomain α/β property reported by (Houry et al., 1999); retained as a broad control.
F14234	The mixed ordered-core/transmembrane label cannot identify an interpretable client-associated property without determining which component activates in the soluble clients.	Abstain: generic orderedness is not a client determinant, and the transmembrane reading is outside the soluble-client interpretation of (Houry et al., 1999).
F15613	Abstain because a sulfatase-specific catalytic motif and a generic helix-boundary signal are incompatible explanatory starting points.	Abstain: sulfatase-loop and generic helix-cap annotations conflict. Metabolic-enzyme enrichment (Fujiwara et al., 2010) does not validate this motif.
F1949	Abstain because the PMT-family description conflicts with a much broader lipobox/basic-motif interpretation in the saved metadata.	Abstain: the plant/protist PMT motif conflicts with lipobox/basic-localization annotations and is not transferable to this bacterial cohort without mapping.
F4066	The annotation is too broad to specify a folding mechanism or a client-associated property.	Abstain: the annotation is too broad. Metabolic-enzyme enrichment (Fujiwara et al., 2010) does not make a generic functional region a client determinant.
F4863	No biological client property can be assigned to the supplied near-null annotation.	Abstain: near-null, diffuse activation supplies no biological client property; retain as a representation/-calibration control.
F5526	Abstain: soluble charged terminal motifs and hydrophobic multipass membrane segments imply different localization and folding contexts, and the annotation sources disagree.	Abstain: Biohub sources conflict between a charged C-terminal helix/loop and a multipass membrane bundle; neither can be chosen from the feature ID.
F5707	Abstain because the motif-specific API description is not supported by the broader helical/assembly annotation in the saved metadata.	Abstain: an INRE-like motif and a broad charged assembly segment imply different family, localization, and mechanism assignments.
F6205	The feature provides no localized substrate recognition or folding-bottleneck mechanism.	Abstain: weak domain-level activation does not establish the α/β or TIM-barrel architectures enriched in substrate surveys (Houry et al., 1999; Kerner et al., 2005).
F6768	Ordered secondary structure alone cannot discriminate GroEL-dependent from spontaneously folding proteins.	Abstain: generic ordered helices, strands, or packing residues do not identify the specific folds enriched in (Houry et al., 1999; Kerner et al., 2005; Fujiwara et al., 2010).
F7771	Abstain: the annotations disagree about activation structure, while neither supplies a specific client-associated property.	Abstain: the API describes weak nonspecific activation while the metadata describes foldedness; neither yields a specific client property.

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F8412	Abstain because generic valine identity and an extended aliphatic membrane-helix detector imply different mechanisms; neither demonstrates a client recognition surface.	Abstain: isolated Val and extended aliphatic membrane-helix readings conflict. Broad hydrophobic residues were not compositionally enriched in Class IV (Fujiwara et al., 2010).
F9857	Abstain: inconsistent structural emphasis and broad targeting/junction examples do not establish one client-associated mechanism.	Abstain: targeting/disorder and helix-boundary readings disagree, and cross-taxon targeting examples are not transferable without residue-level validation.
F2888 × F5707	No defensible joint GroE-client property follows from a conserved pocket and this short motif without the affected proteins and mapped sites. The motif does not by itself identify a cobalamin enzyme, a fold, or a chaperonin-dependent active-site assembly pathway. The independent annotation-source conflict further prevents assigning a motif-specific mechanism.	Abstain: metabolic-enzyme enrichment supports only broad protein class (Fujiwara et al., 2010); the F5707 annotation conflict prevents a motif-specific joint interpretation.
F4863 × F16243	F4863 has no specific biological interpretation in the Biohub description, so its pairing cannot support a specific joint client property or mechanism.	Abstain: F4863 is near-null. F16243 alone has only conditional hydrophobic-exposure precedent (Wang et al., 1999); the pair has no supported mechanism.
F4863 × F5526	The two Biohub sources assign incompatible meanings to F5526. A terminal-helix folding mechanism and a multipass-membrane signal imply different biology; neither can be selected from the feature ID or coefficient. Independently, F4863 is a nearly null nonspecific background signal.	Abstain: a near-null partner and the unresolved F5526 annotation conflict make the pair a representation/label-sensitivity control.
F4863 × F6620	F4863 has no specific biological interpretation in the Biohub description, so its pairing cannot support a specific joint client property or mechanism.	Abstain: F4863 supplies no interpretable biological variable, so a helix-edge/background coupling cannot be assigned.
F4863 × F9237	F4863 has no specific biological interpretation in the Biohub description, so its pairing cannot support a specific joint client property or mechanism.	Abstain: F9237 alone corresponds to the Class-IV Gly association (Fujiwara et al., 2010); the near-null F4863 partner supplies no joint mechanism.
F5526 × F13480	The two Biohub sources assign incompatible meanings to F5526. A terminal-helix folding mechanism and a multipass-membrane signal imply different biology; neither can be selected from the feature ID or coefficient.	Abstain on the pair: F13480 alone corresponds to Ala/Gly enrichment (Fujiwara et al., 2010), whereas F5526 has irreconcilable soluble-terminal versus membrane-bundle annotations.
F5526 × F16243	The two Biohub sources assign incompatible meanings to F5526. A terminal-helix folding mechanism and a multipass-membrane signal imply different biology; neither can be selected from the feature ID or coefficient.	Abstain: the F5526 conflict prevents choosing a soluble-helix or membrane interpretation; Leu/hydrophobic recognition precedent (Wang et al., 1999) does not resolve it.
F5526 × F6620	The two Biohub sources assign incompatible meanings to F5526. A terminal-helix folding mechanism and a multipass-membrane signal imply different biology; neither can be selected from the feature ID or coefficient.	Abstain: F5526 annotation conflict and potentially overlapping charged/helix encodings prevent separation of two biological variables.
F5526 × F9237	The two Biohub sources assign incompatible meanings to F5526. A terminal-helix folding mechanism and a multipass-membrane signal imply different biology; neither can be selected from the feature ID or coefficient.	Abstain on the pair: F9237 alone corresponds to Gly enrichment (Fujiwara et al., 2010); F5526 remains annotation-conflicted.
F5707 × F9237	A short family-associated sequence block and generic glycine signal do not specify a shared folding intermediate or GroE-recognition determinant. A glycine-flanking-motif mechanism would be invented without localization. The independent annotation-source conflict further prevents assigning a motif-specific mechanism.	Abstain on the pair: F9237 alone corresponds to Gly enrichment (Fujiwara et al., 2010); F5707 motif, family, and localization assignments conflict.

Continued on next page

Table 3 continued

SAE term	Interpretation	Closest literature or disposition
F5769 × F9237	An aromatic-hotspot/flexible-hinge hypothesis would depend on selecting one of two incompatible annotations. The snapshot instead raises membrane topology or targeting explanations that cannot be evaluated from the pair alone.	Abstain: F9237 has a Gly correspondence (Fujiwara et al., 2010), but F5769 conflicts between isolated helical aromatics and extended membrane-associated runs.
F583 × F13480	The pair supplies two broad composition axes without specifying their arrangement, a folded domain, or an intermediate. Known Class IV Ala/Gly enrichment supports only part of F13480 and does not establish a joint Asp/small-residue client property.	Abstain on a joint mechanism: only the F13480 component corresponds to Ala/Gly enrichment (Fujiwara et al., 2010); arrangement and acidic context are unspecified.
F583 × F5526	F5526 is described as a family-restricted charged C-terminal motif by the API but as hydrophobic multipass membrane helices by the full metadata. The two imply different mechanisms; neither is validated in the client.	Abstain: the F5526 annotation conflict plus unverified adjacency to an Asp-rich segment precludes a specific terminal-helix or membrane mechanism.
F6205 × F9000	The broad low-amplitude domain sensor supplies too little localization to turn an acidic-composition annotation into a specific joint client mechanism.	Abstain: weak domain-level activation gives no localization for F9000. Global client acidity (Fujiwara et al., 2010) does not establish a local acidic-domain mechanism.
F6656 × F9000	Both features can describe the same low-complexity acidic sequence. The annotations do not separate two biological variables or identify a distinct joint client property.	Abstain: both features can encode the same low-complexity acidic sequence, so the term may reflect representational redundancy rather than two coupled variables.
F8412 × F12861	The API describes F8412 as a Val identity detector across contexts, but the saved full metadata instead requires extended aliphatic helices and is weak on ordinary soluble cores. That conflict prevents a grounded Val/aromatic joint-mechanism assignment.	Abstain: residue-detector and membrane-helix readings conflict. Hydrophobic recognition (Wang et al., 1999) does not validate Val/aromatic clustering, and broad hydrophobes were not Class-IV enriched (Fujiwara et al., 2010).
F9000 × F16088	The features may encode overlapping charge composition. No specific relationship between a linker and a second structural element is provided.	Abstain: the features may overlap in charge composition, F16088 differs between disordered-linker and helical-cap readings, and no second structural variable is identified.
F9237 × F10734	F10734 is a nonspecific mid-chain signal in the API but an extended hydrophobic terminal-segment sensor in the full metadata. A Gly-coupled mechanism cannot be grounded while this conflict is unresolved.	Abstain on the pair: F9237 alone corresponds to Gly enrichment (Fujiwara et al., 2010); F10734 conflicts between nonspecific interior and hydrophobic terminal-region readings.

C COMPUTATIONAL RESOURCES

Structure-generation and intervention jobs used one GPU per worker. GPU inference used NVIDIA H100 GPUs, while vanilla and SAE no-op baseline calculations used NVIDIA A100-SXM4 GPUs. Each GPU worker was allocated eight CPU cores and 80 GB of host memory. The maximum recorded PyTorch CUDA allocation was 18.7 GiB.

The 50,000 discovery interventions required approximately 420 H100 GPU-hours. Direct evaluation of the 67 selected terms required approximately 13 H100 GPU-hours, and construction of the 250-term primary random bank required approximately 65 H100 GPU-hours. The completed multi-method benchmark archive used approximately 26 H100 GPU-hours for 99 additional masks, and the joint ablation experiment used approximately 20 H100 GPU-hours for 65 fresh masks. The completed GPU archives therefore account for approximately 545 H100 GPU-hours in total, plus fewer than

Method	Candidate representation	CV-selected setting	Stable terms (1/2/3)
ProxySPEX	LightGBM; k-SII to order 3	lr .1, depth 3, 500 trees	28/38/1
Marginal LASSO	78 standardized mask indicators	$\alpha/\alpha_{\max} = .025$	55/0/0
Sparse polynomial	Walsh products to degree 3	$\alpha/\alpha_{\max} = .206$	4/31/3
RuleFit	linear terms and boosted-tree rules	100-rule budget; depth 4	35/22/2
FIGS	sum of greedy shallow trees	32-split budget; depth 3	12/1/0

Table 4: **Discovery specifications for the GroEL/GroES benchmark.** Counts report stable marginals, pairs, and triples after the common bootstrap screen.

two A100 GPU-hours for baseline calculations. With at most 24 parallel workers, each of the two discovery batches completed in approximately 11–15 hours.

ProxySPEX and all comparator fitting, bootstrap screening, and downstream analyses were CPU-only. The principal ProxySPEX fit used 16 CPU cores and 64 GB RAM; its 1,000 bootstrap refits used four cores and 32 GB RAM per job. Comparator fits used four cores and 16 GB RAM, and their 4,000 bootstrap refits used one core and 8 GB RAM per job. No protein-model inference was performed during model fitting or bootstrap screening.

D GroEL/GroES benchmark experiment details

Common discovery data. All methods used the same 2,000 binary masks over the 78-feature basis and the same noise-averaged signed effects on 25 discovery clients. ProxySPEX fit a LightGBM conditional-quantile model at $\alpha = 0.25$ to the 50,000 mask–protein rows. The four CPU comparators instead minimized squared error to the empirical 25th percentile across the 25 proteins for each mask. Thus, the biological target is shared, but the fitting losses are not identical. The benchmark compares complete discovery pipelines rather than isolating the estimator alone. Five frozen mask-level folds were used only for hyperparameter selection.

Term extraction and stability. LASSO coefficients were treated as marginal terms. For the remaining comparators, each fitted Boolean surrogate was converted exactly to its Fourier spectrum and then to k-SII terms through order three; spectrum predictions were checked against the fitted surrogate on all sampled masks and additional synthetic masks. Native RuleFit rules and FIGS paths were not themselves treated as evidence of interaction. Each method then used the same 1,000 mask-bootstrap draws. Hyperparameters selected by cross-validation were held fixed during resampling. A term was retained if it appeared in at least 50% of refits and its central 75% coefficient interval, conditional on appearance, excluded zero.

Held-out evaluation. Selected panels were frozen before evaluation on the eight held-out clients and 19 controls. Every selected term was evaluated by a fresh direct ablation of its complete feature set using the same three-draw, means-first structural effect estimator as ProxySPEX. Generalization was summarized by a one-sided paired Wilcoxon test over term-level cohort means. Because the panels differ in size and order composition, the results compare complete discovery pipelines, not equally sized term sets.

Order-matched comparisons. For the benchmark, all methods used one common measured reference bank. Singleton comparisons drew from the 78 candidate singletons after excluding that method's selected marginals; pair and triple comparisons drew from 146 pairs and 50 triples after excluding the union of all five selected panels. For each method, 1,000 panels were sampled without replacement to match its selected order counts, and the same draw was used for clients and controls. Let R_C be the selected-to-random effect ratio for cohort C. Table 1 reports the mean of $100\{R_{\text{client}}/R_{\text{control}} - 1\}$ across paired redraws, with the redraw standard deviation in parentheses. These redraw distributions describe reference-panel composition, not uncertainty over proteins. A matched ratio is unavailable for marginal LASSO because its 55 selected singletons leave only 23 unselected singletons.

The common benchmark bank differs slightly from the primary ProxySPEX analysis in Section 4.2: the primary analysis uses its prespecified 50/150/50 bank, whereas the benchmark has 146 eligible

pairs after excluding every method's selected panel. ProxySPEX is therefore recomputed on the common bank for Table 1. This does not alter any conclusion.