

Constructed negatives determine the population a benchmark can rank over

Justin Li

jli89@tulane.edu

Tulane University <https://orcid.org/0009-0004-1055-213X>

Research Article

Keywords: negative sampling, positivity, benchmark evaluation, partial identification, AUROC, post-translational modification, protein-protein interaction

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11252843/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Constructed negatives determine the population a benchmark can rank over

Justin Li¹

¹ Division of Biomedical Informatics & Genomics, Tulane University, New Orleans, LA, USA

Correspondence: Justin Li (jli89@tulane.edu)

Abstract

Benchmarks often build negatives only from well-annotated entities, leaving parts of the target with zero sampling probability. We write W for the share of a declared target's candidate negatives excluded. Assuming nothing about the excluded region, no strict ordering of two models is identified once W exceeds one half, as in all eight DeepMVP tasks; we therefore report the break-even regional difference that would overturn an observed gain. On reconstructed negatives, a protein-level embedding harmed every task's unsampled region, reversing its gain in seven tasks. The measured regional difference was 2.9–12.8 times the heterogeneity within the sampled region, a reference scale that flagged only two reversals. UPNA-PPI's released negatives omit 73.0% of proteins, and its length-three rule excludes pairs even where every protein keeps partners. Benchmarks should report the target and sampling support at the evaluated unit, and measure part of the excluded region.

Introduction

A benchmark's comparison concerns a population. Its measured ordering extends to a deployment population only when the relationship between the two is known. Park and Marcotte noted that test negatives, unlike training negatives, must represent that population ¹. In biological prediction, negative construction often determines this relationship: curated resources record observed positives, while many remaining candidates have never been assayed.

Negative-set construction affects leakage, degree imbalance and evaluation outcomes in biological prediction ^{2–12}. Related work in machine learning examines test-set reuse, adaptive overfitting, distribution shift and sampled ranking metrics ^{13–17}. Because unannotated candidates may be undiscovered positives, these are positive–unlabelled problems, and annotation that depends on how well an entity has been studied is not selected completely at random ^{18–20}. CLASPP, for example, reports that a conventionally sampled test set favours models with a low false-positive rate ²¹. These findings establish that evaluation design matters, not that they share one statistical problem or one correction.

Here we examine a specific design: drawing negatives only from entities above an annotation-depth threshold. DeepMVP ²² uses well-annotated proteins as negative donors for modification-site prediction. The motivation is to reduce false negatives among residues whose modification status is unknown. But proteins below the threshold then contribute no negative examples. UPNA-PPI ⁵, a protein-interaction benchmark, uses a different eligibility rule that can likewise exclude entire proteins from its negative set.

This is a positivity violation, not covariate shift, and the difference matters because the two call for different remedies. Importance weighting requires positive sampling probability wherever the target distribution has mass ²³. When probabilities are positive but unequal, as under covariate shift, weighting can address the difference, although limited overlap, especially in high dimensions, increases variance and trimming changes the target ^{24,25}. Where a probability is exactly zero, however, the importance weight is undefined, so that region's contribution cannot be recovered without additional information or assumptions. Support must therefore be checked at the unit being evaluated: including every protein does not necessarily cover every protein pair.

Our earlier audit found that a protein-level feature could help under threshold-drawn negatives and harm under a reconstruction using a broader negative pool ²⁶. Here we ask what such a rule allows a benchmark to establish before that reconstruction is available. We separate the missing-region contribution from differences within the sampled region, derive a worst-case AUROC interval, and find the target populations that can pass it.

Without assumptions about the unsampled region, no strict ordering of two models is identified once more than half of the target's candidate negatives are excluded; all eight DeepMVP tasks exceed that threshold. The informative quantity is then the break-even regional difference that would overturn an observed ordering, which we scale by heterogeneity among sampled strata. Where a reconstruction covers part of the excluded region, it can measure that difference for a particular comparison: an augmentation that helps within the sampled region reverses on reconstructed negatives in seven of eight tasks, and in every task the missing region differs from the sampled one more than sampled strata differ among themselves. We also examine support at both protein and pair levels.

Results

Threshold sampling is a positivity violation

We first define the population whose coverage is being measured. For a modification task, the operational target contains candidate residues lacking a recorded annotation, weighted uniformly across the reference proteome. These are candidate negatives, not experimentally verified absences. Partition them by annotation depth k , and let w_k be the target's negative mass in stratum k . Throughout the AUROC analysis, the positive distribution and the two compared scoring functions remain fixed.

AUROC is an average over positive–negative pairs. Consequently, the difference between two models, or between an augmented model and its baseline, satisfies

$$\Delta_{\text{deploy}} = \sum_k w_k \cdot \Delta_k \quad (1)$$

The deployment effect is the average stratum effect under deployment's own weights. This identity concerns evaluation distributions; it does not require retraining either model.

A donor threshold T leaves observed strata $\mathbf{O}$ with depth at least T and unobserved strata $\mathbf{U}$ below it. Their missing weight is

$$W = \sum_{k \in \mathbf{U}} w_k.$$

For this operational target, W requires candidate counts and the release's annotation map, without new labels or model predictions. It ranges from 0.605 to 0.977 across the eight tasks (Fig. 1b). This proteome-wide target is our reference choice, not a scope DeepMVP claims; a benchmark that declared a narrower target would have a smaller W . The derived scopes below restrict the negative population; they do not change the positive reference or establish protein-specific prediction accuracy.

The support failure does not depend on the metric, but (1) relies on AUROC's linearity. For average precision the same decomposition leaves residuals of 23%–90% of the pooled value (Methods), so the bounds below are AUROC bounds only.

The error separates missing support from distribution differences

Missing mass alone does not determine the error. Let $\Delta_{\mathbf{O}}$ and $\bar{\Delta}_{\mathbf{U}}$ denote target-weighted effects conditional on the observed and unobserved regions. Equation (1) gives

$$\Delta_{\text{deploy}} - \Delta_{\text{O}} = W \cdot (\bar{\Delta}_{\text{U}} - \Delta_{\text{O}}) \quad (2)$$

The error relative to the observed region is its difference from the missing region, multiplied by how much target mass is missing. A benchmark's reported effect Δ_{bench} need not equal Δ_{O} , because it may choose different negative weights within O. Therefore

$$\Delta_{\text{deploy}} - \Delta_{\text{bench}} = W \cdot (\bar{\Delta}_{\text{U}} - \Delta_{\text{O}}) + (\Delta_{\text{O}} - \Delta_{\text{bench}}) \quad (3)$$

The first, missing-region term depends on data the benchmark never samples. The second, within-O term concerns a distribution difference inside the sampled region. Reweighting can address the within-O term only if sampling support and the required inclusion probabilities are available at the relevant unit; matching stratum counts alone is insufficient when selection also acts within strata.

In the controlled interaction experiment described below, the within-O term never exceeds 0.006 in high-fidelity cells (Fig. 1c), a favourable case because negatives are sampled uniformly among admissible pairs. For PTM, within-O shifts reach 0.0286 in absolute AUROC, so observed-region effects cannot simply be replaced by the published reading (Methods).

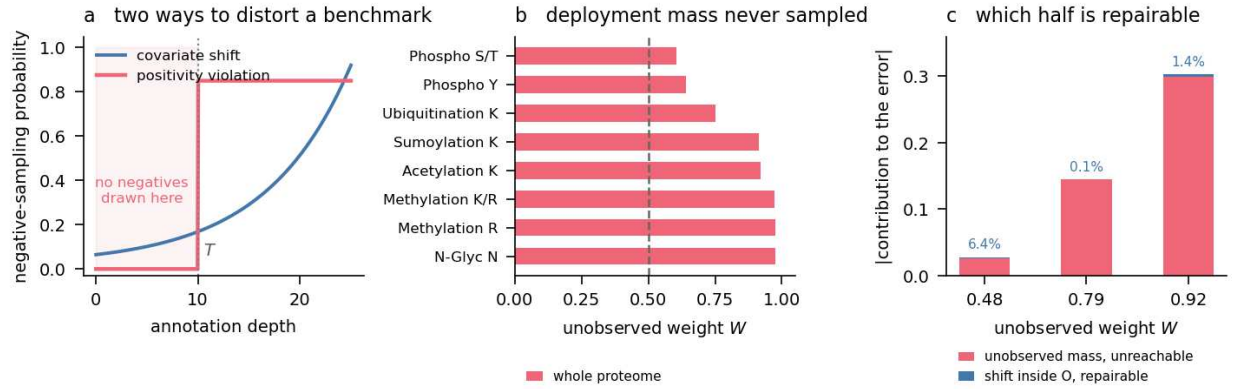


Figure 1. Sampling support and two contributions to error. (a) Schematic negative-sampling probabilities that remain positive or become zero below threshold T . (b) Reference-proteome candidate-negative weight W for eight PTM tasks; the dashed line marks the half-mass screen. Passing that screen would permit, but not certify, a sufficiently large AUROC gap. (c) Absolute magnitudes of the missing-region and within-region terms in the controlled STRING experiment at three prefilter severities. The within-region term contributes 6.4%, 0.1% and 1.4% of the summed absolute magnitudes; these are not fractions of a signed total. Positives and compared scoring functions remain fixed within each decomposition.

Suppose the observed-region effect is positive. A reversal requires the unknown contribution in (2) to outweigh it. Using only $\bar{\Delta}_{\text{U}} \geq -1$ yields

$$\text{a reversal requires } W > \Delta_{\text{O}} / (1 + \Delta_{\text{O}}). \quad (4)$$

Below this threshold, a negative deployment effect is excluded; equality can leave a tie. For a negative observed effect, interchange the model labels and use its magnitude. Applying the published gaps as a scenario in which $\Delta_{\text{bench}} = \Delta_{\text{O}}$ gives no guarantee on any of the eight full-proteome targets: their W values all exceed one half. The scenario misses the required thresholds, 0.010–0.063, by 13×–99×. These multipliers illustrate the condition, rather than estimating the unmeasured within-O shift.

A coverage screen is necessary but does not certify a ranking

After orienting the comparison between two fixed models so that $\Delta_{\text{O}} > 0$, write $D_{\text{O}} = \Delta_{\text{O}}$. The unrestricted missing-region mean then gives the sharp identified interval

$$[(1 - W)D_o - W, (1 - W)D_o + W].$$

The interval excludes a negative ordering only when

$$D_o \geq W / (1 - W). \quad (5)$$

A strict positive ordering requires a strict inequality. This is a worst-case statement over missing effects in $[-1, 1]$; additional restrictions on model scores can narrow the interval (Methods). Since an AUROC difference cannot exceed one,

$$W > 0.5 \Rightarrow \text{no strict ordering is identified by this unrestricted interval.} \quad (6)$$

The converse is weaker: W below one half merely permits a sufficiently large gap to pass. For example, $W = 0.4$ and $D_o = 0.01$ give an interval from -0.394 to 0.406 . A population passing the coverage screen can still leave its actual model comparison unresolved.

The eight full-proteome targets require gaps of 1.53 – 42.5 , whereas the largest gap in the audit's feature-channel table is 0.073 . None of its 42 pairwise comparisons passes the unrestricted criterion (Fig. 2a). This shows neither that those orderings reverse nor that larger samples would help: more data from the sampled population cannot reveal an unobserved region's effect.

The published DeepMVP–MusiteDeep comparison²⁷ provides a familiar scale. Reported serine/threonine AUROCs are 0.95 versus 0.83 , or 0.95 versus 0.89 after MusiteDeep is retrained on DeepMVP's data. The required gap of 1.53 is 12.8 and 25.5 times the observed gaps of 0.12 and 0.06 , so the unrestricted screen offers no full-target guarantee for these tools; the reconstruction below concerns a different model pair and cannot supply one.

Reconstruction measures a particular comparison on a particular population

The worst-case interval assumes nothing about the excluded region. To replace it with a measurement, define $c = |\bar{\Delta}_U - \Delta_O|$: the magnitude of the effect difference between the missing and observed regions for the same model pair and target. Equation (2) then gives $|\Delta_{\text{deploy}} - \Delta_O| = Wc$, so the sign of Δ_O persists whenever c is below the break-even value $c^* = |\Delta_O|/W$. Like Rosenbaum's Γ ²⁸ or the robustness value of Cinelli and Hazlett²⁹, c^* states how large an unmeasured difference must be to overturn a conclusion and leaves its plausibility to evidence. Exceeding c^* permits a reversal but does not establish one; once both signed regional effects are measured, their signed mixture answers the question directly.

We reanalysed the comparison in which the earlier audit found harm²⁶: a site-level sequence model with and without a frozen ESM2 embedding. This embedding is pooled over each protein, so all its candidate sites share the same vector. Both models are trained under the threshold rule and scored on the audit's reconstruction.

We write R for these reconstructed evaluation samples and W_R for their own missing-region share. Negatives are partitioned by the donor rule's merged annotation depth. Three protein-disjoint splits are analysed separately, each with two further initialization-specific scores. Within each split, the same fitted models are evaluated on O and R with positives held fixed, and W_R , $\Delta_{O,R}$ and $\bar{\Delta}_{U,R}$ come from that evaluation, never from the reference proteome's W . Table 1 separates coverage of reference populations (A), a conditional scenario (B) and the effects measured here (C).

The augmentation helps within O and harms in the unsampled region in every task: $\Delta_{O,R}$ is positive in every split, and $\bar{\Delta}_{U,R}$ ranges from -0.047 to -0.269 , with protein-bootstrap intervals excluding zero (Table 1, panel C; intervals in Methods). The overall sign then follows W_R . Seven tasks reverse across R in all three ensemble-scored splits (Fig. 2b), whereas phosphorylation Y, with $\bar{\Delta}_{U,R}$ of -0.093 , stays positive only because its W_R is 0.235 . The across-task mean Δ_R (-0.101) reproduces the harm reported in the earlier audit²⁶.

The six initialization-specific comparisons per task give the same direction pattern, except that one of six ubiquitination comparisons retains its sign. Because $\bar{\Delta}_{U,R} < \Delta_{O,R}$ in every run, a reversal is algebraically the event $c_R > c_R^*$ (Table 1), so the informative quantities are the sizes. Scaled by h , the largest deviation of a within-O depth tier from $\Delta_{O,R}$, break-even values are $0.66\text{--}4.27h$ but measured differences $2.9\text{--}12.8h$ (Fig. 2c); with two to four tiers c_R/h stays above 2.3, and its bootstrap intervals exclude one in every task (Methods). With three tiers, five of the seven reversals have c_R^*/h above one (with four, three do), so reading c^*/h above one as reassurance would have been wrong for most reversals; here it would have had to exceed about 13, the largest c_R/h . That is one experiment, not a general threshold, but it argues for measuring part of the excluded region. Bootstrap intervals for Δ_R also exclude zero in every task (Methods).

These are measurements on R, which covers a subset of proteins and no depth-zero negatives, and c must be recomputed for any other model pair, split or target (Methods).

Reporting the population and the comparison separately

We invert (6) over the explicitly chosen family of negative populations consisting of proteins with annotation depth at least d_{min} . Column (A) of Table 1 reports the largest such population passing the $W \leq 0.5$ screen. It is a reproducible boundary within that family, rather than a uniquely determined deployment target or certification of a particular ranking.

The retained populations range from 66.1% of candidate-bearing proteins for phosphorylation S/T to 1.5%, or 300 proteins, for N-glycosylation (Table 1). A comparison on them must still satisfy (5) with its own observed-region effect.

Column (B) illustrates the stricter gap-dependent condition using the published reading and an explicit assumption of zero within-O shift. It retains only the donor pools, 0.5%–26.7% of candidate-bearing proteins, and is a sensitivity scenario rather than a benchmark-only certificate. For the same published gaps on the full reference targets, the break-even differences Δ_{bench}/W are 0.010–0.083; judging them requires the published models' own within-O heterogeneity.

Table 1. Coverage screens and measured effects answer different questions. In the reference target, W is the excluded candidate-negative share and d_{min} is the smallest annotation depth passing $W \leq 0.5$. (A) gives the retained fraction of candidate-bearing proteins; it does not certify a ranking. (B) applies $W \leq \Delta_{bench}/(1 + \Delta_{bench})$, assuming Δ_{bench} equals the target-weighted observed effect. It is a conditional scenario, not an established guarantee. (C) concerns the ESM2 augmentation versus its baseline on reconstruction R. W_R , $\Delta_{O,R}$, $\bar{\Delta}_{U,R}$, c_R , c_R^* and Δ_R are means of three separately evaluated split-specific ensemble models; within each split $c_R = |\bar{\Delta}_{U,R} - \Delta_{O,R}|$ and the break-even value $c_R^* = |\Delta_{O,R}|/W_R$, below which the sign must persist. The last column counts splits whose signed effect persists when all reconstructed negatives are included. It is descriptive, not a confidence level. The identity is checked per split: products of the displayed means need not reproduce mean Δ_R . Per-initialization values and effects recomputed for every depth restriction are archived. None of the measured columns certifies a different model pair or unmeasured proteins.

Reference candidate population

Task	W	Δ_{bench}	d_{min}	(A) screen %	(B) scenario %
Phosphorylation S/T	0.605	0.0165	1	66.1%	26.2%
Phosphorylation Y	0.642	0.0532	2	57.7%	26.7%
Ubiquitination K	0.750	0.0497	3	41.5%	18.4%
Sumoylation K	0.913	0.0667	5	12.0%	5.9%
Acetylation K	0.921	0.0578	5	10.6%	4.4%
Methylation K/R	0.974	0.0378	7	2.1%	0.9%
Methylation R	0.977	0.0534	7	2.1%	0.9%
N-Glycosylation N	0.977	0.0100	6	1.5%	0.5%

(C) Empirical reconstruction comparison

Task	W_R	$\Delta_{O,R}$	$\bar{\Delta}_{U,R}$	c_R	c_R^*	Δ_R	Sign persists
Phosphorylation S/T	0.446	+0.0312	-0.0978	0.1291	0.0701	-0.0266	0/3
Phosphorylation Y	0.235	+0.0480	-0.0933	0.1413	0.2039	+0.0149	3/3
Ubiquitination K	0.615	+0.0453	-0.0470	0.0922	0.0736	-0.0114	0/3
Sumoylation K	0.768	+0.0355	-0.0967	0.1321	0.0455	-0.0663	0/3
Acetylation K	0.815	+0.0273	-0.2243	0.2516	0.0334	-0.1776	0/3
Methylation K/R	0.941	+0.0729	-0.2694	0.3423	0.0778	-0.2488	0/3
Methylation R	0.921	+0.0622	-0.2495	0.3117	0.0680	-0.2238	0/3
N-Glycosylation N	0.860	+0.0158	-0.0818	0.0976	0.0178	-0.0666	0/3

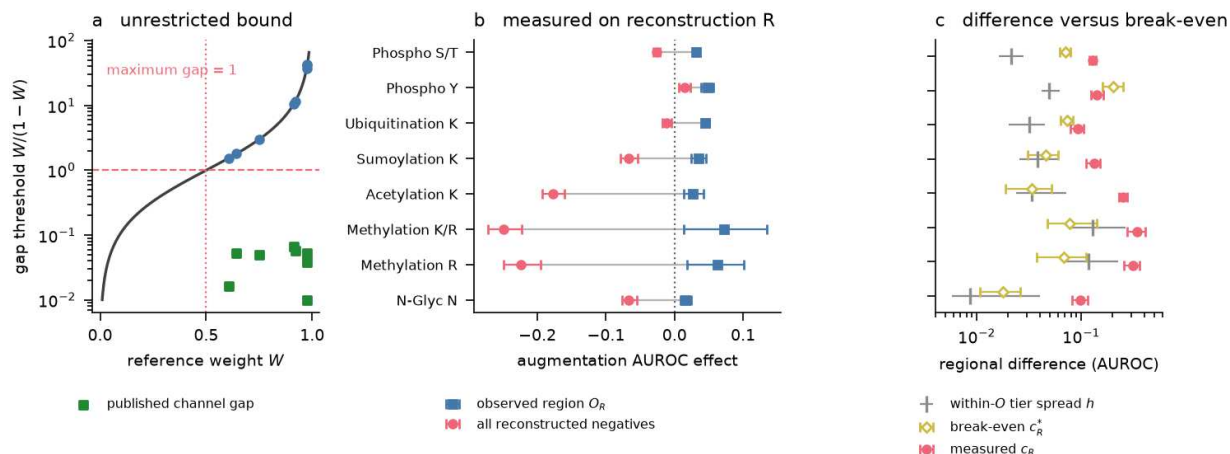


Figure 2. Coverage conditions and measured effects are different quantities. (a) The unrestricted gap threshold $W/(1 - W)$, with circles on the bound and squares showing the selected published channel gaps. Above $W = 0.5$ the threshold exceeds one. (b) Signed ESM2-augmentation effects within O (squares) and across reconstruction R (circles), averaged over three separately scored splits, with 95% protein-bootstrap intervals for the split mean. Seven tasks reverse sign; phosphorylation Y retains a positive effect. (c) Measured regional difference c_R (circles) and break-even value c_R^* (diamonds) beside the largest deviation h of a within- O depth tier from $\Delta_{O,R}$ (grey), all with bootstrap intervals; logarithmic scale. c_R^* and h use only the sampled region; c_R requires the reconstruction. The panels concern this model pair and do not certify another.

Shape restrictions on Δ_k could narrow the interval, but the three tested in synthetic populations with known truth fail when the assumed direction is wrong or too few observed strata remain (Extended Data Fig. 1; Methods).

The prefilter excludes proteins while the length-three rule excludes pairs

UPNA-PPI⁵ scores protein pairs with a configuration model, retains the ten million least probable pairs, and then requires zero length-three paths. The two stages remove support at different levels. The configuration probability increases with degree product, so the prefilter imposes a cutoff τ . In a graph with minimum partner degree one, every protein with degree greater than τ has no eligible partner. Across four human networks, τ ranges from 2 to 42; 52%–70% of proteins are excluded, carrying 93%–98% of positive-edge degree mass (Extended Data Table 1).

Contrastive-L3 alone leaves every protein at least one eligible partner on these networks (Fig. 3a). That establishes marginal protein coverage, not pair coverage. Yet it excludes 11.52%, 4.45%, 93.91% and 95.81% of nonedge pairs on STRING ≥ 700 , STRING ≥ 900 , BioGRID and HIPPIE, respectively. Those pairs have zero inclusion probability under the rule.

Table 2. UPNA-PPI's released negatives by degree in an independent network. Counts concern proteins matched to BioGRID; mean negatives is the number of released negative pairs incident on a protein in that bin.

BioGRID degree	Proteins	In negatives	Excluded	Mean negatives
1	445	407	8.5%	2800.3
2	443	412	7.0%	2137.6
3–4	784	736	6.1%	1459.9
5–9	1,570	1,393	11.3%	780.8
10–24	2,902	1,419	51.1%	217.3
25–49	2,625	69	97.4%	13.6
50–99	3,111	25	99.2%	4.3
100–249	3,395	9	99.7%	2.1
250+	1,651	2	99.9%	0.5

The external check supports the predicted degree dependence but cannot locate the proprietary network's exact cutoff; two proteins above BioGRID degree 250 do appear.

A taxonomy requires a declared target and sampling unit

A shortcut, a distribution shift and a support violation describe different mechanisms. DUD-E's decoys permit discrimination unrelated to binding³¹; pair-input evaluation can also reward structure unrelated to the intended generalisation³. Unequal sampling probabilities can create a target-distribution mismatch, while zero probability prevents weighting-based recovery of a target region. These mechanisms can coexist. Table 3 distinguishes the unit on which support is assessed and the evidence available for each rule.

Table 3. Negative construction and sampling support. Rules are grouped by the support they leave at the evaluated unit: full, not established, or violated. Inclusion means eligibility under the randomised sampling rule, not appearance in one finite draw; recoverability also requires a declared target, positive inclusion probabilities at that unit and appropriate weights. The evidence for each row is given in parentheses.

Rule	Unit	Finding	Support
Random pairing ² (published rule)	Pair	Any noninteracting pair in the declared pool can be drawn	Full within the pool
Random pairing baseline ⁴ (computed)	Pair	Every eligible nonedge has positive probability; the negative/positive degree ratio still varies 200-fold	Full; imbalance is neither a violation nor proof of shift
Degree-distribution-balanced sampling ⁴ (computed counts, published rule)	Pair	Degree matching changes pair probabilities and flattens the degree ratio	Target-specific correction; full support not established
TCR–epitope shuffling ⁶ (published rule)	Pair	Epitope negative counts track positive counts; pair support depends on shuffle constraints	Not established; needs separate checking
Similarity-based curation ²¹ (released code)	Candidate window	Near-motif candidates share label -1 , but downstream sampling can still admit a representative	Severe thinning; an exact zero is not established
Contrastive-L3 alone ⁵ (computed)	Pair	Every protein keeps partners, yet 4.45%–95.81% of nonedge pairs are ineligible	Violation for an all-nonedge target
Configuration-model prefilter ⁵ (computed)	Protein and pair	Degree $> \tau$ removes every partner of a protein	Violation: whole proteins excluded
PTM donor threshold (computed)	Candidate residue	All candidate negatives below T are ineligible; $W = 0.605\text{--}0.977$	Violation for the reference target

Random pairing and degree-balanced sampling illustrate why a target is essential. Their negative/positive degree ratios differ by up to 200-fold, and degree matching flattens the ratio across four orders of magnitude of degree; this changes composition but does not, by itself, show that either design estimates the intended deployment distribution. CLASPP's code admits representatives by cluster label, including the shared rejection label, so we do not classify its near-motif region as structurally unobserved.

Outside biology, uniformly sampled recommendation metrics can reverse method comparisons even in expectation while every candidate remains eligible¹⁷; the proposed corrections reduce bias or mean squared error rather than guaranteeing exact recovery. Metric distortion under full support should not be conflated with a region that the sampling rule excludes.

Controlled features show when support loss changes conclusions

We combined real network topology and the prefilter rule with controlled features so the mechanism could be tested without attributing an effect to a particular embedding (Methods). At high fidelity of the protein-level degree channel and UPNA-PPI's sampling fraction, benchmark benefit accompanies deployment harm in 20 of 20 replicates on three networks and 15 of 20 on STRING ≥ 900 (Extended Data Fig. 2a). On BioGRID and HIPPIE, benchmark gains of about 0.11 AUROC accompany deployment losses of about 0.31. Random-sampling controls agree with deployment within 0.0008 and produce no reversals in 80 replicates.

Severity alone is insufficient: at moderate feature fidelity the same prefilter yields an understatement rather than a reversal. The experiment shows an attainable reversal mechanism, not that UPNA-PPI's own trained predictor reverses. On $\text{STRING} \geq 700$ at the most severe setting, an observed-region effect of +0.244 is overturned by an unseen-region effect of only -0.082 (Methods), so a reversal does not require the extreme value -1 used in the worst-case interval.

Discussion

Constructed negatives define an evaluation population before any model is scored. Our analysis separates three questions: which candidate regions a rule excludes, which comparisons can be certified without observing those regions, and what a reconstruction actually measures.

A benchmark should first state its candidate universe, weighting and positive reference. It can then report W , the annotation-depth boundary d_{min} passing the half-mass screen, and the number of surviving strata, all from candidate and annotation counts. The boundary is conditional on the chosen family of targets and marks where a sufficiently large gap might pass, not where every comparison holds.

A ranking claim additionally needs its observed-region gap and a treatment of within-region sampling. Where W exceeds one half, the unrestricted interval identifies no strict sign, and the informative report is the break-even difference c^* that the unmeasured region would need to overturn the observed one, with the within-O spread h as a reference scale rather than a bound. Effects broadly rise with depth through the threshold rather than jumping at it (Extended Data Fig. 3), so tiers above the threshold see only the flatter upper part of a gradient whose steepest section the rule excludes. Where a ranking matters, part of the excluded region should be measured. Where W is smaller, the actual gap must still clear its threshold. An unidentified ordering should be reported as unresolved, not as a demonstrated reversal.

Reconstruction measurements need equally explicit scope. The positive reference, model pair, training split and negative target determine c , so measuring it for an augmented model and a baseline establishes neither DeepMVP's ordering against MusiteDeep nor an effect across unmeasured proteins. Bootstrap intervals describe resampling of the tested proteins, not training variability, which three splits and two initializations describe only descriptively; numerical residuals check the implementation, not population validity.

The interaction example adds a separate requirement: check support at the evaluated unit. Contrastive-L3 retains every protein yet excludes many protein pairs. The configuration prefilter goes further by removing every partner of high-degree proteins. An audit based only on protein counts would miss the former, so both marginal coverage and item-level eligibility should be reported.

For negative-set design, maintaining non-zero inclusion probability across the intended target preserves the possibility of correction. It does not remove variance or label uncertainty. A filtered sampler with a uniform component, or a documented stratified design, can preserve support; the appropriate mixture and sample size depend on the required precision and false-negative sensitivity. Donor thresholds do buy cleaner negatives: under the assumption most favourable to them, modelled contamination falls from 21% to 8% at $T = 10$, yet a single required site already puts five tasks' W above one half (Extended Data Fig. 2c; Methods). Conversely, a deliberately restricted target can be legitimate if its scope is stated and users do not infer broader performance.

The effect in the missing region remains unidentified without data or assumptions. The tested shape restrictions failed when observed strata provided too little resolution or the assumed direction was wrong; monotone instrumental-variable and treatment-response work^{32–34} suggests other formulations. External measurements in under-annotated proteins could instead extend empirical coverage.

Our evidence has defined limits. The PTM effect analysis uses reconstructed candidate negatives and three split-specific model pairs; it neither verifies their biological absence nor covers the reference proteome. The interaction experiment uses controlled features, and UPNA-PPI's proprietary positive network prevents

exact replication of its pipeline. Positives are also concentrated in well-studied proteins; holding them fixed isolates support loss on the negative side but leaves that selection in place, which is one reason a protein-level feature can act as a shortcut²⁰. The bounds concern AUROC and do not transfer to average precision. GO/CAFA is omitted because no single applicable negative-construction release was identified. These limits constrain generalisation but leave a practical reporting principle: state the target, identify what the sampling rule cannot observe, and distinguish a coverage screen from evidence for a particular comparison.

Methods

Notation

The symbols are defined for the target specified in each analysis; quantities indexed by R refer to reconstructed PTM evaluations.

- Target: the population of candidate negatives an evaluation is meant to cover; in the PTM analyses, unannotated residues of the target type, weighted uniformly across the reference proteome.
- k, w_k : an annotation-depth stratum, and the target's candidate-negative mass in it.
- Δ_k : the AUROC difference between the two compared models with negatives restricted to stratum k and positives fixed; Δ_{deploy} is the difference on the whole target.
- T, O, U : the donor threshold; the observed strata (depth at least T); the unobserved strata (depth below T).
- W : the unobserved mass, $\sum_{k \in U} w_k$.
- $\Delta_O, \bar{\Delta}_U$: target-weighted effects within O and within U.
- Δ_{bench} : the effect a benchmark reports, which may weight negatives within O differently from the target.
- D_O : Δ_O after orienting the comparison so that it is positive.
- c : $|\bar{\Delta}_U - \Delta_O|$, the regional difference for one model pair and one target.
- $d_{\min}$: the smallest annotation depth whose depth-restricted target has $W \leq 0.5$.
- R: the reconstructed evaluation samples; $W_R, \Delta_{O,R}, \bar{\Delta}_{U,R}, \Delta_R$ and c_R are the corresponding quantities computed on R.
- τ : the degree-product cutoff of UPNA-PPI's configuration-model prefilter.

Identification model and attainable bounds

For (4)–(6), let the unknown conditional difference range freely over $[-1, 1]$. The resulting interval is $[(1-W)D_O - W, (1-W)D_O + W]$. Its endpoints are attainable in this unrestricted score-distribution model: missing negatives can score below all positives for one model and above all positives for the other, or vice versa. Mixtures attain intermediate values while leaving observed-region scores unchanged. This construction does not assert that these assignments remain possible under additional known constraints on the two scoring functions. Such constraints would require a narrower identification model. The interval's lower endpoint is nonnegative iff $D_O \geq W/(1-W)$; strict ranking requires $>$, with equality allowing a tie. $W \leq 0.5$ is only an existence screen for a sufficiently large observed gap, not a certificate for the gap at hand.

Synthetic populations

The decomposition and the bounds built on it were first evaluated where the truth is known by construction. Populations carry the three structural conditions of the real case and nothing else: labels defined on items grouped into entities; negatives constructed by a rule that draws them only from entities carrying at least T positives; and an entity-constant feature that partially encodes how many positives its entity has. Each entity receives a latent character z driving both its items' propensity to be positive and, through that, the number

of positives it accumulates, so annotation depth is an observed consequence rather than an input. The entity feature is z recovered at a controllable fidelity r^2 , measured against log depth rather than assumed. Splits are entity-disjoint. Models are a baseline using item-local features and an augmented model adding the entity channel, both multilayer perceptrons with two hidden layers (64, 32 units), L_2 penalty 10^{-3} , early stopping. Twenty seeds per cell at 2,500 entities. The stratum decomposition (1) is exact by construction; the residual is at machine precision in every configuration tested (maximum 2.4×10^{-16} over the threshold grid, 1.1×10^{-16} over the network grid).

Shape restrictions. Can assumptions about the missing region narrow the interval? We tested three in these populations. Assuming that Δ_k is non-decreasing in depth bounds unseen effects by the shallowest observed stratum. Its direction can be wrong: at low T , before the harm mechanism engages, sparse strata perform better rather than worse, invalidating the bound in 19 of 20 replicates (Extended Data Fig. 1a).

Reading direction from the observed gradient raises validity at $T = 2$ from 5% to 90%. At $T \geq 10$ it does not help: mean extrapolation error is 0.036, compared with 0.010 for the endpoint anchor, which uses the shallowest observed stratum's effect for every unseen stratum. Curvature fitting is less stable still. These results evaluate the specified procedures in the controlled experiment; they do not rule out all externally justified shape restrictions.

The number of nonempty observed strata falls from 7 to 5, 3, 2 and 1 at $T = 0, 2, 5, 10$ and 20 (Extended Data Fig. 1c). Two distinct strata are needed to fit a slope and three to fit curvature. Thus the specified fits lose the resolution they require at the higher thresholds. Reporting surviving strata complements W : one measures observed resolution, the other excluded mass. Neither alone establishes an extrapolation assumption.

Post-translational modification tasks

The eight tasks use the earlier audit's reconstructed site tables and per-site predictions, based on deepmvp.ptmax.org releases and a reference proteome of 20,413 sequences³⁵. Candidate-bearing proteins contain at least one residue of the target type. For each protein we count candidate residues minus recorded positive sites, not experimentally established negatives. The annotation depth used for donor eligibility merges serine/threonine and tyrosine phosphorylation, and uses the combined methylation task for the arginine subset. This same depth definition is used for reference counts and prediction stratification.

Reference coverage. `deployment_scope.py` and `scope_reanalysis.py` scan minimum depths from zero through $T = 10$. At each step they recompute W from retained reference candidate counts. Column (A) is the first depth with $W \leq 0.5$; none of the reported boundary cases equals exactly one half. Column (B) instead uses $\Delta_{\text{bench}}/(1+\Delta_{\text{bench}})$, where the published gaps are 0.0100–0.0667. This is explicitly conditional on the target-weighted observed effect equalling the published gap. The reference universe and the nested depth family are design choices; neither is inferred from the observed ranking.

Reconstruction and models. The earlier audit²⁶ rebuilt each task from the released positive annotations and this reference proteome. Candidate negatives are the unannotated residues of the target type in proteins carrying at least one positive of that type. Two constructions differ only in the donor rule: the threshold construction (`replica` in file names) draws negatives only from proteins carrying at least ten sites of the merged modification type, which reproduces the released data; the reconstruction R (`rebuilt`) removes that restriction. Negatives are sampled uniformly at five per positive where the pool allows.

Proteins were clustered at 40% identity with CD-HIT and whole clusters assigned to three protein-disjoint partitions shared by all tasks, so the two constructions' test partitions contain the same proteins and differ only in their negatives.

The baseline follows DeepMVP's published architecture²²: three one-dimensional convolutional layers (512, 512 and 128 channels, kernel 5) and a bidirectional GRU over the residue's sequence window. The

augmented model adds a frozen ESM2 embedding (650M parameters) mean-pooled over the whole protein, projected to 128 dimensions and concatenated with the recurrent output, so it is constant across a protein's candidate sites. Each model was trained twice under the threshold construction (Adam, learning rate 10^{-3} , early stopping on validation AUROC), and its ensemble score is the mean of the two. Trained models, per-site predictions and code are archived with the audit ²⁶.

Empirical reconstruction effects. `scope_reanalysis.py` pairs the baseline with the actual ESM2-augmented condition, trained under replica sampling (the benchmark's threshold rule) and evaluated on rebuilt samples (the audit's reconstruction, R). Predictions are matched one to one by protein, site and label. Split IDs 0, 1 and 2 are analysed separately. The recorded ensemble score is the primary score; `init_0` and `init_1` provide six additional run-level measurements per task, not six independent datasets. File manifests record model condition and SHA-256 hashes. Shuffled controls have a distinct index key and cannot replace augmented predictions. We do not pool predictions from different fitted models to estimate one AUROC.

For a negative with score s , its AUROC contribution is the fraction of fixed positive scores greater than s plus half the tied fraction. Subtracting the baseline contribution from the augmented contribution gives a per-negative effect. Its mean over each split's observed and unseen negatives supplies $\Delta_{O,R}$ and $\bar{\Delta}_{U,R}$. Their counts supply W_R , and the mean over all that split's negatives is Δ_R . Every identity is evaluated within a split before averaging. The maximum residual is 5.6×10^{-17} . Means and ranges across splits are descriptive; intervals come from the protein bootstrap below, and no population sign guarantee is inferred from numerical precision.

Changing empirical scope. At every minimum depth we retain the same positives and recompute the negative subset, W_R , $\bar{\Delta}_{U,R}$, c_R and the signed Δ_R . We record both sign agreement with $\Delta_{O,R}$ and the sufficient symmetric condition $|\Delta_{O,R}| > W_R c_R$. A failed sufficient condition does not override the directly measured sign. The full 792-row table includes all eight tasks, three splits, three score versions and eleven minimum depths. The first depth at which all six initialization-specific comparisons preserve the observed sign is 4, 0, 4, 7, 9, 10, 10 and 9 in Table 1's task order. These are empirical scan outcomes; they are not percentages of the reference proteome or out-of-sample guarantees.

Bootstrap intervals. `bootstrap_effects.py` resamples proteins with replacement within each task-split, carrying every positive and negative site of a drawn protein with its draw count, and recomputes W_R , $\Delta_{O,R}$, $\bar{\Delta}_{U,R}$, c_R , c_R^* and Δ_R from weighted AUROC contributions with ties counted as half. The three splits are resampled independently and averaged replicate by replicate, giving 95% percentile intervals for the split means in Table 1 (2,000 replicates; Extended Data Table 2). Split-specific intervals are wider, and some cross zero. In the methylation and N-glycosylation tasks the observed region comes from only 8–23 proteins per split, which widens their $\Delta_{O,R}$ intervals; one of 2,000 replicates for N-glycosylation split 0 drew none of its eight observed-region proteins and is excluded. The intervals describe variability among the tested proteins, not uncertainty about unmeasured proteins or the reference target. The fitted models are held fixed, so training variability enters only through the three splits and two initializations.

Break-even scale. `calibrate_cstar.py` divides each split's observed negatives into three tiers by protein depth with roughly equal negative counts, keeping each protein within one tier, and computes each tier's effect with positives fixed. The scale h is the largest absolute deviation of a tier effect from $\Delta_{O,R}$; like c_R^* , it needs no unsampled negatives. Split means of c_R^*/h are 0.66–4.27 and of c_R/h 2.9–12.8; split-specific values range from 0.2 to 5.2 and from 1.7 to 21.3. The scale depends on the tiering: with two or four tiers, mean c_R^*/h is 1.0–8.7 or 0.62–3.4, and mean c_R/h stays above 2.3 for every tiering (Extended Data Table 3). The same protein bootstrap, with tiers held fixed and replicates paired with Extended Data Table 2, gives intervals for h and the ratios; the lower end of the c_R/h interval is at least 1.3 in every task for two to four tiers (`results_calibration_bootstrap.tsv`). Because a maximum over noisy tier effects is biased upwards, we also report a negative-weighted root-mean-square deviation of tier effects from $\Delta_{O,R}$; with it, mean c_R is 3.4–58.7 times the scale across tierings (`results_calibration_runs.tsv`). In the methylation and N-glycosylation tasks a tier can hold two to four proteins, so h is imprecise.

Depth profile. Extended Data Fig. 3 recovers the mean effect of the unsampled negatives at each depth exactly, by differencing consecutive totals of the archived minimum-depth scan; R has no depth-zero negatives.

Reconstruction coverage. Reference W and empirical W_R are kept separate. For example, N-glycosylation's reconstructed source table covers 2,977 proteins, while the reference candidate universe contains 20,123. Prediction files cover test subsets of that reconstruction, and provide no negatives at annotation depth zero. The coverage table reports per-split site and protein counts. No regional effect is extrapolated from those files to all reference proteins. Within-stratum AUROCs that change both the positive and negative sets are not used to calculate c_R . A full-target effect would require additional coverage or justified transport assumptions; machine-precision reconciliation within R verifies the arithmetic, not those assumptions.

Metric comparison. `metric_linearity.py` provides a separate arithmetic diagnostic on the pooled baseline scores, not a model-ranking estimate. It holds all positives fixed, includes all nonempty depth strata, and compares the pooled metric with the negative-count-weighted stratum metrics. The AUROC residual is at most 2.2×10^{-16} . Average-precision residuals are 0.173–0.363 (23%–90% of pooled values; Extended Data Fig. 2b). Precision depends on class ratio and rank structure, so the weighted AUROC identity does not extend to average precision. The updated calculation uses merged donor depths for phosphorylation and methylation.

False-negative trade. The expected contamination among a threshold- T benchmark's negatives is the candidate-weighted posterior that an unannotated site is positive, using the earlier audit's sensitivity model. The model is driven by the observed positive rate among candidates at each depth; its anchor is the stratum with the highest such rate rather than the deepest stratum, because the rate is not monotone in depth. At $\alpha = 0$, all depth–rate differences are attributed to missing annotation; this is the scenario most favourable to the donor threshold. Results are also reported at $\alpha = 0.5$. These model-based rates do not identify the true biological negative population used by an assay.

Donor thresholds seek cleaner negatives. Under the sensitivity assumption most favourable to that rationale, assigning the entire depth–rate relationship to missing annotation, expected contamination falls from 21% without a threshold to 8% at $T = 10$ (Extended Data Fig. 2c). These are model-based estimates, not measured false-negative rates.

Requiring one annotated site already raises mean W to 0.504, for a contamination reduction of 3.8 percentage points. Five tasks exceed one half at $T = 1$, where T is the donor threshold applied to the full reference target (unlike Table 1's d_{min} , which restricts the target at DeepMVP's threshold of ten): acetylation 0.569, methylation K/R 0.586, methylation R 0.609, sumoylation 0.605 and N-glycosylation 0.796. Those targets therefore fail the unrestricted ranking screen at the first threshold step. Higher thresholds further reduce support, without resolving a particular ordering.

The benefit also depends on the annotation model. Under a less extreme attribution of depth differences to missing annotation, proteins with no recorded sites are the cleanest candidate negatives, and initially excluding them raises expected contamination from 2.8% to 5.4%. The earlier sensitivity analysis explores this dependence²⁶. A design decision should therefore report both the assumed contamination model and the support it removes. Maintaining support and modelling contamination are alternatives to treating a donor threshold as cost-free protection.

Interaction networks

Four human networks were used: STRING physical subnetwork³⁶ v12.0 at combined score ≥ 700 and ≥ 900 , HIPPIE³⁷ (rolling release, fetched 2026-08-28) at its high-confidence mark of 0.63, and BioGRID release 4.4.246 physical interactions restricted to both interactors being human. Each was reduced to a two-column edge list, symmetrised, and stripped of self-loops and duplicate edges. Node counts after processing range from 8,706 to 20,088 and maximum degrees from 255 to several thousand, a tenfold range of density.

Contrastive-L3. For each protein we count partners inducing zero length-3 paths. The test reduces to a boolean matrix cube: length-3 walks and length-3 paths differ only by degenerate cases, every one of which requires the edge (u, v) , and negative candidates are non-adjacent by construction, so on exactly the pairs that matter walks and paths coincide. Computed blockwise by rows so that the full pair space is never materialised, and verified against brute-force enumeration of simple paths on random graphs.

Configuration-model prefilter. The configuration model's interaction probability is monotone increasing in the degree product, so keeping the N least probable pairs is keeping the N pairs of smallest degree product. The cutoff τ is found by binary search over the degree histogram rather than by enumerating pairs, at the ratio the source rule uses (10^7 of $\approx 1.56 \times 10^8$ pairs, 6.4%). All pairs tied at the selected degree-product cutoff are retained in the reimplement, so its realised retained fraction can exceed 6.4%, especially on sparse networks. Known edges are then excluded. The exact cutoffs are $\tau = 2$ on both STRING networks, 15 on BioGRID and 42 on HIPPIE; these differ from the degree-bin labels used to display the results.

Protein versus pair support. `support_unit_audit.py` reads the saved per-protein eligible-partner counts. Their sum divided by two is the number of eligible unordered pairs. The candidate nonedge denominator is $[n(n-1) - \sum \text{degree}]/2$, so one minus their ratio gives the excluded nonedge share. Every protein can have a nonempty partner set while some pairs remain ineligible. Degree mass instead counts positive-edge endpoints and is reported separately.

Similarity curation. We inspected CLASPP's released data-curation code at commit 83ac3dbe2ae6ac66dc22fbc658112c66fd369316. `get_easy_neg_data_5.py` assigns the shared label -1 to candidates failing its seven-mismatch test against known 21-mers. `sample_negative_data` in `getData/DataProcess.py` samples a representative per label without excluding -1 . This establishes thinning rather than exclusion of every member of that region; inclusion probabilities also depend on upstream ordering and other negative categories. We do not infer a region-wide exact zero from a finite realised sample.

Degree-distribution-balanced sampling. Reimplemented from the published source: node pairs are indexed by degree sum, one unused pair of the same sum is taken per positive edge, and neighbouring sums are searched outward when a bucket is exhausted. Two departures are noted rather than silently corrected — the source enumerates all node pairs without excluding self-pairs, which can emit a protein paired with itself as a negative, and its exhaustive enumeration is replaced here by sampling a degree composition with probability proportional to the product of stratum sizes, which is the same distribution over pairs without materialising n^2 of them.

Released negatives. UPNA-PPI's negative set (`data/Topological_Negatives`, 3,063,605 pairs over 5,037 proteins) was analysed as published and joined to BioGRID physical human degrees by gene symbol, matching 16,926 of the 17,975 proteins in its released universe. No part of that analysis reimplements the rule.

Consequence experiment

Real topology and the real rule are combined with controlled features so that ground truth is known and any effect is attributable to the sampling rule rather than to an embedding's idiosyncrasies. Positives are network edges; benchmark negatives are non-edges admitted by the prefilter; deployment negatives are non-edges drawn uniformly. The protein-level channel encodes standardised log degree at fidelity r^2 , and the pair-local channel is drawn from the label alone, identically for both evaluations, so it carries no information about which rule produced a pair. The split is protein-disjoint: test pairs join two proteins neither of which appears in any training pair. Twenty seeds per cell. Training pairs are capped uniformly at 50,000 across networks so that no cross-network comparison reflects training-set size; a sensitivity analysis at 40,000 moves the estimates by about 0.01 against effects of 0.30, and shifts them towards zero, so the cap is conservative.

For uniform deployment negatives, omitting a known fraction of nonedge pairs fixes the missing pair mass. Our implementation retains all degree-product ties at the cutoff and removes known edges, so a nominal prefilter fraction need not equal the retained fraction of deployment negatives. The measured excluded pair mass, W , is 0.888–0.937, 0.793–0.801 and 0.484–0.502 at nominal prefilter fractions of 0.064, 0.2 and 0.5, respectively. This relation concerns a hard restriction of the pool. It does not imply that evaluation must sample nearly every candidate: a non-zero uniform component mixed with filtered sampling can restore support while keeping a small sample. Known inclusion probabilities then permit weighting, with precision and contamination still requiring assessment.

`reversal_bound.py` evaluates identity (2) on $\text{STRING} \geq 700$ with three seeds per cell. At nominal fraction 0.064 and $r^2 = 0.90$ ($W = 0.917$), $\Delta_O = +0.244$ and $\bar{\Delta}_U = -0.082$, giving a deployment effect of -0.055 .

Within-O shift. The two terms of (3) — the unobserved mass and the shift inside O — are measured separately using `shift_check.py` on the consequence experiment's own $\text{STRING} \geq 700$ network at fidelity $r^2 = 0.90$: (W , unobserved-mass term, within-region shift term) = (0.479, -0.026 , $+0.002$), (0.788, -0.145 , $+0.0001$), (0.918, -0.299 , -0.004) across three prefilter severities. Quoted throughout as $|\text{shift}|/(|\text{mass}| + |\text{shift}|)$, the shift term runs 0.1% to 6.4% there — largest at the mildest severity, where the total is itself smallest — and never exceeds 0.006 in absolute value; taken instead against the signed total the two terms sum to, the same three cells give approximately 7.4%, 0.1% and 1.4%, the conventions diverging only where the terms have opposite signs. At the low fidelity $r^2 = 0.30$, where no reversal occurs, it is 36% to 52% of a sum that is itself under 0.011. This measurement is a favourable case for the approximation $\Delta_{\text{bench}} \approx \Delta_O$: these experiments sample uniformly among admissible pairs, close to the deployment conditional, whereas a more structured within-O rule (a fixed ratio of negatives per entity, say) could carry a much larger second term.

The PTM diagnostic is recomputed by `shift_check_ptm.py` for each actual augmentation–baseline model pair and split. Both evaluation files use the same positive identities and the reconstructed file's positive scores. The negative distribution changes from the replica evaluation to the rebuilt evaluation. The decomposition uses the rebuilt sample's W_R , not the reference W . Across 24 split-task cells, the largest absolute within-O term is 0.0286 and the maximum decomposition residual is 1.2×10^{-16} . Score drift can be compared only when the identical negative sites are present in both files; incomplete matches are recorded as unavailable, not used as an uncertainty allowance. In particular, subtracting a file-drift value from an effect is not treated as a confidence bound.

Reproducibility

`scope_reanalysis.py` produces Table 1 and its per-run, per-depth and coverage files; `prediction_files.py` supplies condition-aware indexing, and `measured_c.py` provides a compact per-split check. `deployment_scope.py` scans reference coverage only. `ranking_identification.py` and `ptm_guarantee.py` retain the published-gap worst-case scenarios. `metric_linearity.py` tests the metric identity. `shift_check_ptm.py` and `shift_check.py` measure the two error terms on PTM and controlled PPI evaluations, respectively. `cl3_exclusion.py` counts eligible partners; `support_unit_audit.py` derives pair coverage from those counts; `ddb_sampling.py` measures degree composition. `identification_bound.py` and `ppi_consequence.py` generate the controlled experiments, and `reversal_bound.py` evaluates identity (2) within them. `bootstrap_effects.py` gives the protein bootstrap intervals for Table 1(C), and `calibrate_cstar.py` the within-O scale in Fig. 2c. `threshold_tradeoff.py` calls `fn_sensitivity.py`, reused from the first audit and archived separately. `make_figures.py` generates the main and Extended Data figures from the archived results. Model-condition selection, scope changes and support-unit counterexamples have executable regression checks.

Use of large language models

Claude (Anthropic) was used to polish the manuscript text for language, clarity and readability. The author reviewed and edited the AI-assisted text and takes full responsibility for the content.

Data availability

PTM reconstructions and predictions derive from our earlier audit, archived at Zenodo (data 10.5281/zenodo.21670043, release v1.2 10.5281/zenodo.22102323). Interaction networks are public: STRING physical v12.0, HIPPIE (rolling release; v2.4 as of the 2026-08-28 fetch date, DOI unversioned by the resource), and BioGRID release 4.4.246. UPNA-PPI's negatives are public at github.com/alxndgb/UPNA-PPI; its positives are proprietary (QIAGEN BKB) and were not used. Per-split and per-initialization PTM results, all depth scans, coverage counts and input-file hashes are archived alongside the analysis code; the interaction-network sweep is archived as per-cell aggregates over twenty seeds (including the per-replicate reversal counts reported in Results), with per-seed results regenerable from the fixed seeds using `ppi_consequence.py --sweep`.

Code availability

All analysis code specific to this paper, including every script named throughout Methods, is available under the MIT license at github.com/JustinLi2002/constructed-negatives and archived at Zenodo (doi:10.5281/zenodo.23146595; release v1.4.0, commit ccd0c89, the version analysed here). Components reused verbatim from the first paper's audit (`fn_sensitivity.py`, via `threshold_tradeoff.py`) are archived separately at Zenodo (10.5281/zenodo.21993819).

Acknowledgements

The author thanks Hong-Wen Deng for his support and encouragement.

Author contributions

J.L. conceived and designed the study, performed all analyses and wrote the manuscript.

Funding

J.L. declares no relevant funding.

Competing interests

The author declares no competing interests.

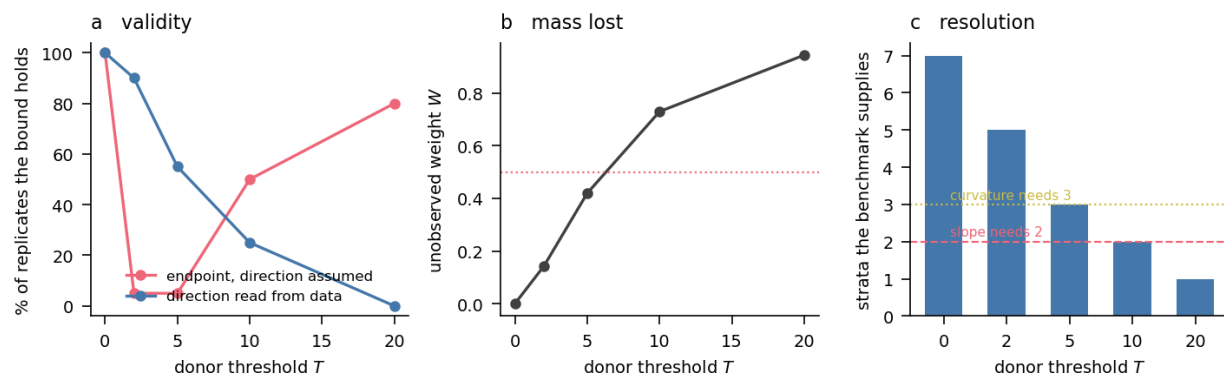
References

1. Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein–protein interactions. *Bioinformatics* 2011;27:3024–8.
2. Bernett J, Blumenthal DB, Grimm DG, *et al.* Guiding questions to avoid data leakage in biological machine learning applications. *Nat Methods* 2024;21:1444–53.
3. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* 2012;9:1134–6.
4. Li P, Shao B, Zhao G, Liu Z-P. Negative sampling strategies impact the prediction of scale-free biomolecular network interactions with machine learning. *BMC Biol* 2025;23:123.
5. Chatterjee A, Ravandi B, Haddadi P, *et al.* Topology-driven negative sampling enhances generalizability in protein–protein interaction prediction. *Bioinformatics* 2025;41:btaf148.

6. Dens C, Laukens K, Bittremieux W, Meysman P. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat Mach Intell* 2023;5:1060–2.
7. Graber D, Stockinger P, Meyer F, *et al.* Resolving data bias improves generalization in binding affinity prediction. *Nat Mach Intell* 2025;7:1713–25.
8. Tourne N, De Waele G, Vermeirssen V, Waegeman W. How negative sampling shapes the performance of transcription factor binding site prediction models. *Bioinformatics* 2026;42:btac048.
9. Ben-Hur A, Noble WS. Choosing negative examples for the prediction of protein–protein interactions. *BMC Bioinformatics* 2006;7(Suppl 1):S2.
10. Neumann D, Roy S, Minhas FUA, Ben-Hur A. On the choice of negative examples for prediction of host–pathogen protein interactions. *Front Bioinform* 2022;2:1083292.
11. Lannelongue L, Inouye M. Pitfalls of machine learning models for protein–protein interaction networks. *Bioinformatics* 2024;40:btac012.
12. Cabrera-Andrade A, Robalino-Díaz J, Dueñas-Espín I, Zambrano AK, Villegas-Ch W. BioGraphBench-PPI enables audit-first evaluation of negative-sampling contracts and structural baselines in protein interaction networks. *Front Bioinform* 2026;6:1966469.
13. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 2023;4:100804.
14. Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet classifiers generalize to ImageNet? *PMLR* 2019;97:5389–400.
15. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. The reusable holdout: preserving validity in adaptive data analysis. *Science* 2015;349:636–8.
16. Koh PW, Sagawa S, Marklund H, *et al.* WILDS: a benchmark of in-the-wild distribution shifts. *PMLR* 2021;139:5637–64.
17. Krichene W, Rendle S. On sampled metrics for item recommendation. *KDD* 2020:1748–57.
18. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. *KDD* 2008:213–20.
19. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. *Mach Learn* 2020;109:719–60.
20. Bekker J, Robberechts P, Davis J. Beyond the selected completely at random assumption for learning from positive and unlabeled data. *Lect Notes Comput Sci* 2020;11907:71–85.
21. Gravel N, Zhou Z, Fang R, *et al.* CLASPP: a unified model for predicting post-translational modifications. *PLoS Comput Biol* 2026;22:e1014616.
22. Wen B, Wang C, Li K, *et al.* DeepMVP: deep learning models trained on high-quality data accurately predict PTM sites and variant-induced alterations. *Nat Methods* 2025;22:1857–67.
23. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* 2012;21:31–54.
24. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009;96:187–99.
25. D'Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. *J Econom* 2021;221:644–54.
26. Li J, Yu J. Data supporting: Negative sampling determines whether protein-level features help or harm PTM site prediction. *Zenodo* 2026; doi:10.5281/zenodo.21670043 (analysis code: doi:10.5281/zenodo.21993819).
27. Wang D, Liu D, Yuchi J, *et al.* MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res* 2020;48:W140–6.

28. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* 1987;74:13–26.
29. Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. *J R Stat Soc B* 2020;82:39–67.
30. Oughtred R, Rust J, Chang C, *et al.* The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 2021;30:187–200.
31. Chen L, Cruz A, Ramsey S, *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* 2019;14:e0220113.
32. Manski CF, Pepper JV. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* 2000;68:997–1010.
33. Manski CF. Monotone treatment response. *Econometrica* 1997;65:1311–34.
34. Manski CF. *Partial Identification of Probability Distributions*. Springer, 2003.
35. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–31.
36. Szklarczyk D, Kirsch R, Koutrouli M, *et al.* The STRING database in 2023. *Nucleic Acids Res* 2023;51:D638–46.
37. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningful biological interactions. *Nucleic Acids Res* 2017;45:D408–14.

Extended Data



Extended Data Fig. 1. Specified shape restrictions in controlled synthetic populations. (a) Fraction of twenty replicates in which each upper bound holds, comparing an assumed monotonic direction with direction estimated from the observed gradient. The latter improves validity at $T = 2$, but its extrapolation is less accurate at high T . (b) Unobserved weight across the threshold grid. (c) Number of nonempty observed strata, with reference lines at the two required for a slope and three for curvature. The results assess these fitted procedures and do not rule out externally justified restrictions.

Extended Data Table 1. The prefilter at the ratio UPNA-PPI uses, on four networks. τ is the degree-product cutoff, not the first fully excluded degree-bin label. Boundary ties are included. Degree mass refers to positive edges and is not the missing negative-pair weight.

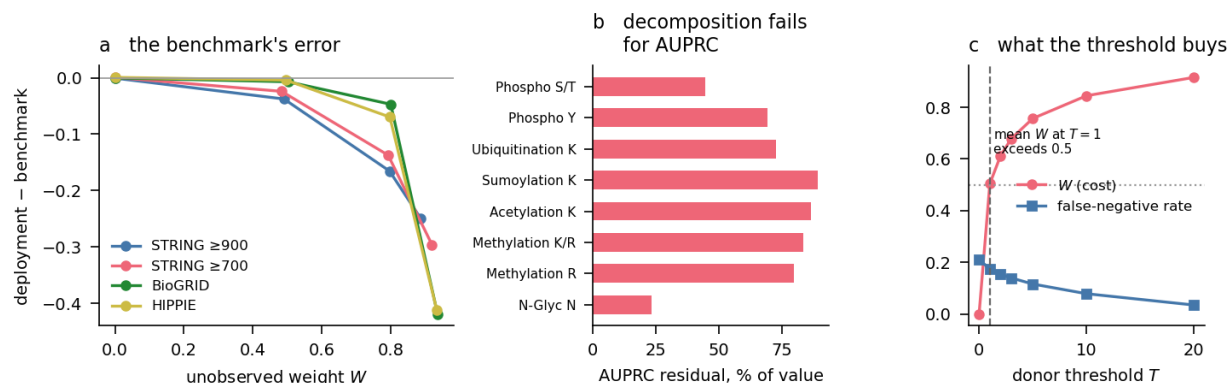
Network	Proteins	Excluded	Share	Degree mass	τ
STRING physical ≥ 900	8,706	5,518	63.4%	0.950	2
STRING physical ≥ 700	10,746	7,521	70.0%	0.974	2
BioGRID	20,088	13,125	65.3%	0.977	15
HIPPIE	18,740	9,737	52.0%	0.927	42

Extended Data Table 2. Protein bootstrap intervals for the reconstruction effects in Table 1(C). Entries are 95% percentile intervals for the mean of the three split-specific ensemble comparisons, from 2,000 replicates that resample proteins with all their sites within each split. They describe variability among the tested proteins, not uncertainty about unmeasured proteins.

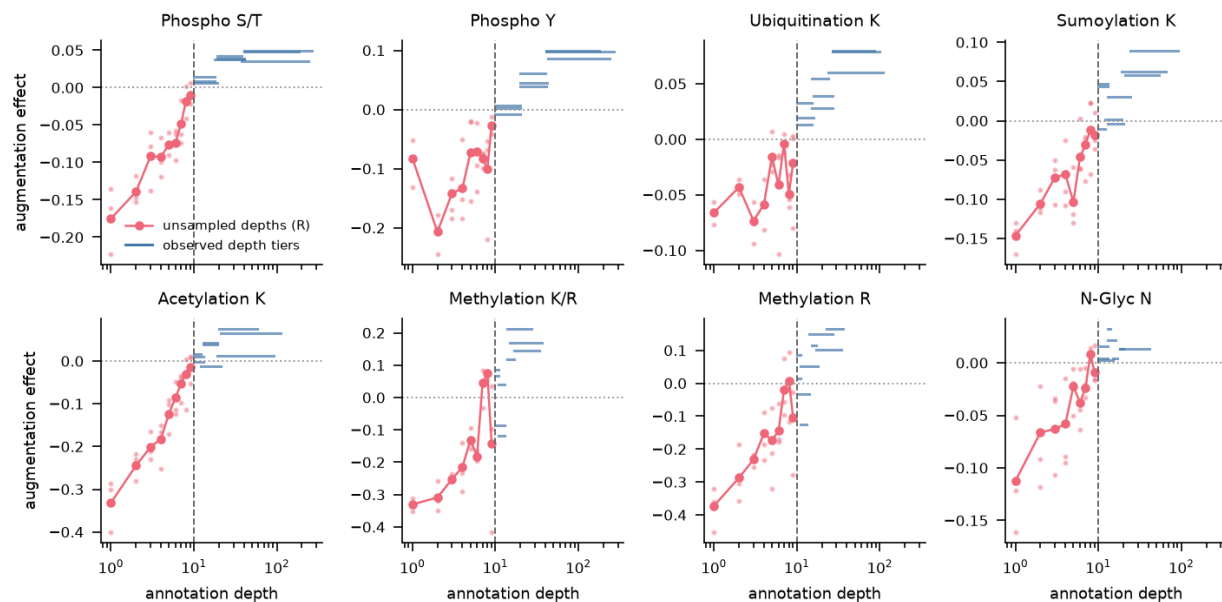
Task	$\Delta_{O,R}$	$\bar{\Delta}_{U,R}$	c_R	c_R^*	Δ_R
Phosphorylation S/T	0.028 to 0.035	-0.108 to -0.088	0.119 to 0.139	0.062 to 0.079	-0.031 to -0.022
Phosphorylation Y	0.039 to 0.057	-0.112 to -0.076	0.124 to 0.162	0.160 to 0.252	0.006 to 0.024
Ubiquitination K	0.039 to 0.051	-0.059 to -0.035	0.079 to 0.106	0.063 to 0.084	-0.018 to -0.005
Sumoylation K	0.024 to 0.046	-0.114 to -0.079	0.112 to 0.152	0.031 to 0.060	-0.079 to -0.054
Acetylation K	0.014 to 0.042	-0.246 to -0.202	0.228 to 0.276	0.019 to 0.052	-0.192 to -0.161
Methylation K/R	0.014 to 0.135	-0.298 to -0.240	0.274 to 0.408	0.047 to 0.141	-0.273 to -0.223
Methylation R	0.018 to 0.101	-0.280 to -0.215	0.256 to 0.359	0.038 to 0.111	-0.249 to -0.195
N-Glycosylation N	0.009 to 0.024	-0.098 to -0.066	0.081 to 0.114	0.011 to 0.026	-0.077 to -0.055

Extended Data Table 3. Within-O scale for the break-even difference. h is the largest deviation of a depth-tier effect from $\Delta_{O,R}$ with three tiers, as a mean over splits with its protein-bootstrap interval. Ratios are means of split-specific values with two, three and four tiers; the last column is the bootstrap interval of c_R/h with three tiers, from the same replicates as Extended Data Table 2.

Task	h (interval)	c_R^*/h , 2 / 3 / 4 tiers	c_R/h , 2 / 3 / 4 tiers	c_R/h interval
Phosphorylation S/T	0.021 (0.016 to 0.028)	6.30 / 3.42 / 2.84	11.21 / 6.12 / 4.99	4.70 to 8.90
Phosphorylation Y	0.049 (0.042 to 0.062)	6.24 / 4.27 / 3.41	4.42 / 3.00 / 2.38	2.28 to 3.93
Ubiquitination K	0.032 (0.020 to 0.045)	4.29 / 2.33 / 2.04	5.34 / 2.91 / 2.56	2.10 to 6.14
Sumoylation K	0.038 (0.025 to 0.060)	4.62 / 1.26 / 0.96	13.32 / 3.56 / 2.66	2.26 to 7.99
Acetylation K	0.034 (0.024 to 0.072)	2.85 / 1.06 / 0.63	23.51 / 8.31 / 4.89	3.95 to 19.63
Methylation K/R	0.128 (0.077 to 0.264)	1.25 / 0.82 / 0.63	4.78 / 3.29 / 2.39	1.45 to 8.18
Methylation R	0.117 (0.066 to 0.225)	1.03 / 0.66 / 0.62	5.78 / 2.92 / 2.68	1.46 to 10.99
N-Glycosylation N	0.009 (0.006 to 0.040)	8.68 / 1.99 / 1.31	53.35 / 12.75 / 8.70	3.89 to 70.41



Extended Data Fig. 2. Consequence experiment, metric diagnostic and contamination sensitivity. (a) Error in the controlled feature comparison against missing pair weight on four networks, with a no-prefilter control. The features are constructed; these effects are not measurements of UPNA-PPI's own predictor. (b) Average-precision decomposition residual relative to its pooled value for each PTM task, using the same fixed positive set; the corresponding AUROC residual is at machine precision. (c) Mean model-based false-negative contamination against mean W as the donor threshold changes. Under the displayed sensitivity assumption, $T = 1$ already puts mean W above one half. These contamination rates depend on the annotation model and are not measured biological false-negative rates.



Extended Data Fig. 3. Augmentation effect by annotation depth on reconstructed negatives. Circles give the mean effect of unsampled negatives at each depth below the donor threshold (dashed line, $T = 10$), pooled over three splits, with light points for individual splits; they are recovered exactly from the archived depth scan. Horizontal bars give the effects of the three observed depth tiers in each split, from which h is computed. R contains no depth-zero negatives. Effects broadly rise with depth through the threshold, so tiers above it see only the flatter upper part of the gradient.