Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Center-dependent generalization in public protein-failure data: a benchmark with homology-controlled and laboratory-held-out evaluation

Ganggang Liu

ganggliu@oihtech.com

Soochow University

Research Article

Keywords: protein expression, protein solubility, structural genomics, negative results, benchmark dataset, batch effect, evaluation protocol, protein language model, TargetTrack

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11321374/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background.

Public protein-production records offer a potential resource for predicting experimental failure, but their labels reflect heterogeneous experimental procedures and reporting practices. Sequence redundancy and laboratory-specific conventions may therefore complicate evaluation. We examined whether predictive associations in these records transfer between structural genomics centers after controlling sequence similarity and visible construct-related features.

Results.

We harmonized 2380297 records into six stage-specific label categories, representing 98617 sequence clusters at 30% identity. Affinity-tag usage, tag position and recorded failure rates varied substantially between centers; pooled and within-center associations with sequence length also differed. On identical test subsets, the frozen ESM-2 model achieved a median prevalence-normalized precision-recall area under the curve of 1.85 with same-center training and 1.01 with training from other centers, despite the latter containing 4.0-17.7 times more homology groups. For expression failure, the four held-out center groups showed appreciable enrichment, weak enrichment, performance compatible with the random baseline, and performance below that baseline. The below-baseline result persisted after low-rank adaptation and the tested adversarial intervention. A model using only amino-acid composition remained competitive with the protein language model. Opposite directions relative to baseline were observed between expression and soluble-expression tasks in one center group with the frozen model, but this comparison was inconclusive after low-rank adaptation.

Conclusions.

In these public records, sequence-based prediction of experimental failure was strongly dependent on the evaluation center and task. Homology control and removal of visible construct features did not ensure transfer to held-out centers. The released labels, frozen splits and evaluation code provide a benchmark for assessing this transfer and its uncertainty before predictors are used to prioritize experiments in a new laboratory.

Background

Structural genomics consortia and large-scale production efforts have accumulated records of attempted protein production in which many entries document failure rather than success. Such records are attractive for training sequence-based predictors of experimental outcome, because negative results are otherwise rarely published. They are also, however, administrative records produced by many laboratories under different target-selection criteria, construct conventions, expression protocols and

reporting practices. Whether an association learned from pooled records describes a property of proteins or a property of the contributing laboratories is therefore an evaluation question, not only a modeling question.

Two precautions are standard in this literature. The first is homology-based redundancy reduction, which prevents near-identical sequences from appearing on both sides of a train-test division. The second is removal or masking of construct-derived residues such as affinity tags, which are known to carry label-correlated information. Both precautions address the identity of the *protein* being evaluated. Neither addresses the identity of the *laboratory* that produced the record. A predictor can therefore satisfy both and still be evaluated, in effect, on sequences from laboratories it has already seen.

Prior work

Previous studies have already established that protein-production datasets contain biases that can affect sequence-based prediction. NetSolP identified strong associations between affinity-tag motifs and solubility labels and suggested that these associations could reflect selection or labeling practices in the groups performing the experiments [1]. SoluProt balanced sequence-length distributions and evaluated predictions on independently obtained NESG measurements [2]. PLM_Sol incorporated tag removal into dataset curation and included an in-house experimental evaluation [3]. More recent work has addressed heterogeneous production data and prospective validation through RP3Net [4], while the Aiki-Sol preprint explicitly conditions solubility predictions on centrifugation regimes [5]. These studies provide important precedents for treating experimental context as part of the prediction problem.

Related methods, including PredPPCrys, used experimental-progress records from PepcDB to predict successive stages of protein production and crystallization [6]. Their use of homology reduction addressed dependence between related sequences. Separately, a recent independent evaluation reported that solubility-predictor rankings changed between native and heterologous protein datasets, with simple sequence scores remaining competitive outside some models' training distributions [7]. Thus, neither the existence of dataset artifacts nor a reduction in performance under distribution shift is, by itself, the contribution of the present study.

A second body of work concerns the crystallization-propensity and multi-stage production-propensity predictors, which are the closest methodological precedents for stage-wise prediction from sequence [6, 8–13]. Within the examples we verified in the primary papers, generalization is assessed after homology reduction, and the contributing laboratory is not used as a unit of evaluation. We restrict this statement to the papers we checked directly rather than extending it to the family as a whole; in particular, a critical evaluation published in 2018 [13] cannot establish the evaluation protocol of a method published in 2021 [12], and we do not use it for that purpose. The endpoints of PredPPCrys and related predictors concern cloning, production, purification and crystallization propensity, which are related to, but not identical with, the six stage-specific categories used here.

Leakage in train-test division has also been studied directly in a neighboring domain. Bushuiev et al. demonstrated in protein-interaction data that commonly used splitting strategies based on sequence or metadata similarity introduce major data leakage, with the consequence that evaluation may be overoptimistic [14]; the companion main-conference work provides a non-redundant reference dataset and a near-duplicate detection algorithm, and constructs leakage-free divisions by three-dimensional interface similarity [15]. Our task is sequence-level rather than interface-level, so we use the transitive closure of sequence homology as the unit of division, but we adopt the same diagnostic stance: assume the division leaks, then measure it.

Study contribution

The defensible contribution of this work is the joint analysis of center-associated features and labels, the matched comparison of training sources, and the released evaluation resource. Here we quantify how construct features, recorded outcomes and sequence-label associations vary among the contributing centers, and test whether these associations transfer when training and test centers differ. The central comparison evaluates same-center and cross-center training on identical test sequences, with homologous groups kept separate. We then examine whether tag removal, omission of an explicit length feature, low-rank fine-tuning or the tested adversarial intervention changes the held-out-center results. This design connects descriptive audits of center-associated variation to its consequences for prediction. The six-stage label resource, verified splits and group-level uncertainty estimates make the comparison reproducible. By distinguishing performance on unfamiliar sequences from performance at unfamiliar centers, the benchmark addresses an aspect of generalization that sequence-homology control alone does not establish.

Methods

Data sources and label construction

Records were assembled from five public sources: the Protein Structure Initiative TargetTrack archive [16], the mega-scale folding-stability dataset of Tsuboyama et al. [17], ProteinGym deep mutational scanning substitutions [18], a meta-analysis dataset of de novo designed binders [19], and the Adaptyv Bio EGFR design competition rounds [20]. All raw files were downloaded from the addresses recorded in the repository configuration, verified by MD5 against recorded values, and used read-only; no upstream raw file is redistributed by this work.

Each record was mapped to six stage-specific categories — cloning, expression, soluble expression, purification, stability and binding — under a three-state convention: 1 denotes a recorded success, 0 a recorded or inferred failure, and − 1 an outcome that was not observed. The − 1 state is excluded from both the training loss and all evaluation, so that an unobserved step is never treated as a success. Every count reported in this article states the denominator it was computed over, because records that reached a given step and records that exist in the source differ systematically.

A composite soluble-expression endpoint was defined as expression successful and soluble expression successful, matching the label definition used by the transferred predictor in the baseline comparison. This composite endpoint is reported separately from the stage-specific soluble category throughout, and neither is interpreted as a direct measurement of intrinsic thermodynamic solubility.

Construction and leakage verification of the frozen splits

The unit of division is neither the sequence nor the greedy-clustering cluster, but the **split group**: a connected component of the 30%-identity transitive closure obtained from an all-versus-all search over all unique sequences. Cascaded clustering in MMseqs2 is a greedy set cover, which guarantees only that members satisfy the identity threshold with respect to their own cluster representative, and not that members of different clusters fall below it. Division by cluster therefore leaks, and we measured this rather than assuming it (Table 3).

Table 3
Train-test sequence pairs above 30% identity at each stage of split construction

Procedure	Pairs above 30% identity between train and test
Whole-group assignment by 30% set-cover cluster	10841 (maximum identity 1.00)
Plus merging after pairwise comparison of cluster representatives	2302
Plus all-versus-all transitive closure over all sequences	16 / 1 / 8
Plus iterative repair	0 / 0 / 0

Three values separated by slashes are the counts for the sequence, laboratory and temporal splits respectively. Division by greedy-clustering cluster leaks because the clustering guarantees the identity threshold only against each cluster's own representative.

Taking the transitive closure at 30% identity and 80% coverage, the target space forms a single giant connected component containing 28646 set-cover clusters. We verified that the cause is not bridging by short fragments but multi-domain proteins bridging between families. Groups accounting for more than 1% of the dataset were forced to the training side, since otherwise one component would consume an entire test partition. The cost of this choice is that held-out test sets systematically exclude proteins that sit within large homology networks, and we treat it as a limitation on the applicability of our conclusions rather than as a solved problem.

A final iterative repair step is required because group construction and leakage verification are two independent searches with heuristic k-mer prefiltering. Four splits were then frozen (Table 4): a sequence split testing basic generalization to unfamiliar sequences; a laboratory split in which whole centers are held out; a temporal split; and a cross-target split for binder prediction. In addition, center folds were

constructed for the matched comparison described below, with per-fold homology exclusion applied so that no training group is homologous to any test group in the same fold.

Table 4
The four frozen evaluation splits

Split	What it tests	Test records	Leakage check
Sequence	Generalization to unfamiliar sequences	73219	Pass (0 pairs)
Laboratory	Whether center-associated features have been learned	5843	Pass (0 pairs)
Temporal	Drift over time	606	Pass (0 pairs)
Binding target	Binder prediction for unseen targets	300	Pass (0 pairs)

Analyses of center-associated variation

Four descriptive analyses were specified before the modeling experiments. First, expression failure rates were tabulated per center over records that reached the expression step, restricted to centers with at least 1000 such records. Second, the presence of a His6 motif and its position were tabulated per center over all records of the TargetTrack-derived source, restricted to centers with at least 10000 records; position was expressed as the proportion of tagged records carrying the tag within the terminal region at the N- or C-terminus. Third, the rank area under the receiver operating characteristic curve of sequence length against each stage label was computed pooled across centers and separately within each center, and the pooled value was compared with the within-center median and range. Fourth, an ablation removed construct-derived residues matching 21 pattern classes (His, FLAG, MYC, HA, Strep and T7 tags; TEV, thrombin, Factor Xa and enterokinase sites; GS and GGGGS linkers; and vector cloning-site remnants) and additionally dropped the explicit length feature, after which the same models were retrained and evaluated on the sequence split and on the laboratory split.

Models

Three model levels were evaluated. L0 is a gradient-boosted decision tree (scikit-learn HistGradientBoostingClassifier) using only 20-dimensional amino-acid composition, with construct residues stripped and **no length feature**. L1 uses frozen ESM-2 650M embeddings with a trained classification head. L2 applies low-rank adaptation to ESM-2 650M, updating 5.41M parameters, which is 0.82% of the model, for three epochs. The adversarial variant adds a center-discrimination head trained adversarially against the representation, and we report the accuracy that head attains as a measure of how much center-separable information the representation retains under that specific probe.

No model weights were retained. The training scripts evaluate and discard them, and per-fold predictions are released instead, so that every reported figure can be recomputed without the weights.

Evaluation metric and adjudication rules

The positive class is failure. The primary metric is **PR-AUC/base**, the area under the precision-recall curve divided by the positive-class prevalence, so that scores from folds with different failure prevalence are comparable on a common scale. A value of 1.0 corresponds to the prevalence baseline. Lift at a fixed budget is computed alongside as an interpretability reference, but is not used for adjudication: cluster-stratified bootstrap on a fold with approximately 2500 samples gives an interval 0.63 wide for lift at the top 100 items, against 0.17 for PR-AUC/base, and the two metrics once produced opposite verdicts on the same fold. The timing and rationale of this rule change are recorded in the evaluation changelog released with the code.

Uncertainty is quantified by bootstrap resampling stratified by split group, with 2000 resamples, so that homologous sequences are resampled together rather than independently. Fold-level results are then classified by criteria that separate direction from magnitude (Table 5): the interval determines whether the direction is resolvable, and the point estimate determines whether the magnitude is appreciable. The magnitude threshold used for this fold-level classification is 1.10.

Table 5
Criteria for classifying fold-level results

Classification	Condition
Appreciable enrichment	Interval does not cross 1.0, point estimate at least 1.10, and the adversarial variant's interval likewise does not cross 1.0
Borderline	Meets the above numerically, but the adversarial variant's interval crosses 1.0
Weak but resolvable	Interval does not cross 1.0 and point estimate below 1.10
Compatible with baseline	Interval crosses 1.0
Below baseline	Interval lies entirely below 1.0
Prerequisite: direction robustness	The direction must agree across five independent bootstrap seeds; where it does not, the conservative reading is taken and the cell is recorded as compatible with baseline
The magnitude threshold of 1.10 applies to this fold-level classification only. It is distinct from the value 1.17 reported in Methods as a reference point for subsequent work, which was obtained for one task on one hold-out set and is not a universal threshold.	

A threshold for subsequent work is reported separately and is not the same number. The value 1.17 is the score attained by the composition-only baseline on the composite soluble-expression comparison on

the laboratory hold-out set. It is reported as a reference point for future claims of cross-center signal, paired with the requirement that the bootstrap interval not cross 1.0. Because 1.17 was obtained for one task on one hold-out set, it cannot be transferred as a universal threshold to expression failure or to other folds; the general requirement is comparison against a trivial baseline evaluated **on the same task and the same test items**. Neither 1.10 nor 1.17 defines an economically worthwhile experimental gain, and neither should be read as such.

Direction robustness across resampling seeds. A fold-level direction verdict is accepted only if it agrees across five independent bootstrap seeds; where the five disagree, the conservative reading is taken and the cell is recorded as compatible with the baseline. This rule was added after two cells that both displayed as [1.00, 1.04] at two decimal places received opposite verdicts, the lower bounds being 0.9957 and 1.0010 — a difference of 0.001. Direct testing showed that one of the two had a direction verdict of four positive and one indeterminate across five seeds, and raising the resample count from 2000 to 20000 still returned 1.0007, because the uncertainty lies in the data rather than in the resampling count. Across all 40 fold-level results, exactly one was reclassified by this rule. Where an interval endpoint falls within 0.02 of 1.0, four decimal places are reported. This check assesses the stability of interval-based classification under resampling; it is **not** a replication across independent model-training seeds, and we do not present it as one.

Provenance control

Every artifact carries a fingerprint file recording its own SHA-256 hash, byte count and row count, together with the same three values for all of its upstream inputs. Each downstream script verifies this chain at startup and exits on mismatch rather than continuing. This mechanism was added after a real incident in which, following an upstream rebuild, a downstream script read a stale intermediate artifact and a decontaminated subset silently shrank from 2377 records to 289 while the numbers were produced as usual, with no error raised. All figures quoted in this article are inserted programmatically from frozen output files rather than transcribed by hand.

Use of AI tools

AI tools were used in this work and their use is described here in full. The author conceived the study, directed it throughout, and made the final decision at every step of the study design and of the evaluation protocol, which were developed iteratively in dialogue with these tools. Claude Code (Anthropic) was used to write and run the analysis code, covering data ingestion, label construction, split construction and leakage verification, model training and evaluation, and the provenance checks described above, and to generate the reports from which all results are taken; every figure quoted in this manuscript is inserted programmatically from frozen output files rather than transcribed by hand. Claude (Anthropic) was used for literature searching, for drafting and revising the manuscript text, and for translation between Chinese and English. All cited references were verified against Crossref and against the publishers' full texts; during that verification two DOIs proposed by an AI tool were found to point to

unrelated papers and were removed. The author reviewed all outputs and takes full responsibility for the content of this paper.

Results

Record counts overstate the effective scale of the resource

The harmonized resource contains 2380297 records, corresponding to 354019 unique sequences once deep mutational scanning variants are collapsed to their parent proteins, and to 98617 clusters at 30% identity. Aggregated by production stage, the number of negative records exceeds the number of independent homology clusters they represent by factors ranging from 2.2 to 76 (Table 1).

Table 1
Negative records and homology clusters by production stage

Stage	Negative records	Clusters at 30% identity	Shrinkage
Cloning	18398	5663	3.2
Expression	94165	22771	4.1
Soluble	9995	4640	2.2
Purification	29854	9001	3.3
Stability	153858	3766	41
Binding	163950	2161	76

Shrinkage is the ratio of negative records to clusters, reported to one decimal place below 10 and as an integer at 10 and above. Stage-level aggregation masks differences between sources; see Table 2.

Aggregating by source rather than by stage identifies the cause (Table 2). Full-pipeline production records, in which each entry is an independent target, shrink least, by a factor of 4.8. Deep mutational scanning sources shrink most, by factors of 486 and 6661, because they consist of many single-residue variants of a small set of parent proteins. The two aggregations do not reconcile exactly because the negatives for the stability and binding categories come almost entirely from variant-scanning sources.

Table 2
Negative records and homology clusters by data source

Source	Negative records	Clusters at 30% identity	Shrinkage	Composition
TargetTrack	146678	30438	4.8	Full-pipeline records; each entry an independent target
Tsuboyama et al.	106886	220	486	Single-residue variants of a set of parent domains
ProteinGym	213151	32	6661	Deep mutational scanning of a set of parent proteins
De novo binders	3275	1998	1.6	Designed binders; few targets, designs mutually dissimilar
Adaptyv EGFR	230	40	5.8	Single-target competition designs

Shrinkage spans 1.6 to 6661, that is from 0.21 to 3.82 orders of magnitude. The stage and source aggregations do not reconcile exactly because stability and binding negatives come almost entirely from variant-scanning sources.

What this quantity measures is effective diversity for cross-protein generalization, not data quality. For the purpose variant-scanning datasets are made for — predicting the effect of mutations within one protein — they are excellent resources. For the task of predicting whether an unseen protein will fail, the number of negatives a source contains matters less than the number of independent protein families it covers. Cluster counts quantify diversity under the specified search and grouping procedure and are not an exact count of statistically independent biological samples.

Recorded failure, tag usage and tag position vary between centers

Among the 17 centers that attempted expression and have at least 1000 such records, the recorded expression failure rate ranges from 73.7% down to 0.0%. Three centers have a failure rate of exactly zero, totaling 367452 records; the largest of these, JCSG, accounts for 350702. Tens of thousands of attempts without a single recorded failure is not biologically plausible, and indicates that whether negative labels exist at all is strongly associated with which center performed the work.

Across the TargetTrack-derived source, 21.1% of records contain a His6 motif (n = 939665), but usage across the 12 centers with at least 10000 records ranges from 91.8% to 0.02%. Tag position is equally center-specific (Table 6): the same tag sequence appears almost exclusively at the C-terminus in one center and almost exclusively at the N-terminus in another. Both the presence and the position of a tag are therefore center-associated features that are visible in the sequence itself.

Table 6
His6 usage and tag position by center

Center	Records	His6 usage	N-terminal	C-terminal
NYSGXRC	44727	54.5%	2%	98%
NESG	133249	69.4%	41%	59%
SGPP	20856	55.3%	100%	3%
SSGCID	21715	91.8%	99%	1%
SECSG	14816	16.5%	99%	0%
JCSG	378364	11.5%	100%	0%

Denominator is all records of the TargetTrack-derived source for each center, including those that reached only cloning. Position percentages are over tagged records and are reported independently for each terminus, so they need not sum to 100%.

The association between sequence length and failure also differs between the pooled and within-center views (Table 7). For the purification and stability categories the pooled and within-center values fall on opposite sides of 0.5 (0.476 versus 0.545, and 0.412 versus 0.548). For cloning and expression the apparent association weakens from 0.621 and 0.577 pooled to within-center medians of 0.547 and 0.533, close to uninformative. The pooled association with length is thus largely attributable to differences in convention between centers rather than to a within-center relationship.

Table 7
Rank AUC of sequence length against failure, pooled versus within center

Stage	Pooled across centers	Within-center median	Within-center range
Cloning	0.621	0.547	0.5158–0.6608
Expression	0.577	0.533	0.4926–0.6751
Soluble	0.427	0.441	0.4209–0.5706
Purification	0.476	0.545	0.4703–0.703
Stability	0.412	0.548	0.5043–0.5916

For purification and stability the pooled and within-center values fall on opposite sides of 0.5.

The tag-and-length ablation affects the two divisions asymmetrically. On the sequence split, where training and test data come from the same centers, all six tasks decline, with a mean change of -4.8%. On the laboratory split, where centers are held out, three of four evaluable tasks decline, the largest change is in the opposite direction at + 9.8%, and the mean change is -1.5%. A biological effect of construct residues and length would be expected to reduce performance comparably on both sides. Instead these features are usable when training and test data share centers and are not usable across

centers, which is consistent with an identifying rather than a mechanistic function. The statistical strength of this particular comparison is limited: only four tasks on the cross-center side have sufficient negatives to evaluate, and a fluctuation of 10% in a single task would change the mean. It therefore stands alongside the within-center length analysis as corroboration and does not carry the conclusion on its own.

A transferred predictor and a composition-only baseline both lie close to the prevalence baseline across centers

Against the composite soluble-expression endpoint, the transferred solubility predictor and a composition-only gradient-boosted baseline give similar values, and both fall substantially on the laboratory split relative to the sequence split (Table 8).

Table 8
Transferred solubility predictor versus composition-only baseline, composite soluble-expression endpoint

Method	Sequence split PR-AUC/base	Laboratory split PR-AUC/base
SoluProt, full test set	1.47	1.16
SoluProt, decontaminated subset	1.44	1.15
Amino-acid composition GBDT, tags stripped, no length	1.60	1.17
SoluProt is applied outside its stated domain of applicability and its core identity feature is unavailable for 41% of the sequence-split test set, so these values are a lower bound for it. Contamination at 30% identity against its upstream source is 18.7% for the sequence split and 9.9% for the laboratory split.		

We do not describe this as outperforming the transferred predictor. Its ecoli_usearch_identity feature, the maximum identity to *E. coli* PDB sequences, cannot be computed for 6730 of 16435 sequences, that is 41%, in our sequence-split test set, against 4 of 21, or 19%, in the test examples shipped with the tool; in those cases the tool falls back to the training-set mean. It was trained and designed for naturally occurring proteins expressed heterologously in *E. coli*, whereas our test set contains designed proteins, membrane proteins and many sequences without a significant PDB homolog. We are applying it outside its domain of applicability, so its values here are a lower bound for it. The accurate statement is that on this dataset a gradient-boosted model using only 20-dimensional amino-acid composition, with no length feature and construct residues stripped, matches or exceeds a published tool, while the published tool's core feature degrades severely out of domain.

Two asymmetries in this comparison are unfavorable to the transferred predictor and we state them explicitly. First, it was trained on its own dataset and transferred by us for inference, whereas the composition baseline is trained directly on this dataset; cross-domain use necessarily lowers its score. Second, its PDB-identity feature is unavailable for 41% of our test set.

One candidate asymmetry we examined is not one. The transferred predictor handled length by balancing the length distribution at dataset-construction time, whereas we do not use length as a feature at all. These are different operations — one changes the training distribution, the other the feature set — but their effect in this comparison runs in the same direction, since neither party gains from length.

One point must nevertheless be recorded on the transferred predictor's behalf: our test set is not length-balanced. A decision function fitted on length-balanced training data and then applied to a test set that is not length-balanced may be disadvantaged by that mismatch alone. Quantifying this would require retraining the tool, which is outside the scope of this work, so we record it as a known but unquantified residual asymmetry in our favor. Separately, scores obtained under the length-retaining ablation configuration are higher and must not be placed alongside the transferred predictor's, because what they gain from is precisely the center-associated channel this article identifies.

Because the transferred predictor's training set derives from the same upstream archive as our largest source, we measured contamination at 30% identity: 18.7% for the sequence split and 9.9% for the laboratory split. Every value is therefore reported both for the full test set and for the decontaminated subset. The laboratory-split values are essentially unchanged after decontamination, so contamination did not inflate them, and we record that as a point in favor of that split.

Same-center and cross-center training on identical test items

This comparison holds the test items fixed and changes only the source of the training data. For each center fold, the held-out center's homology groups are divided into three inner parts; the i-th inner part is the test set, the remainder of the held-out center supplies the **same-center** training data, and the other centers supply the **cross-center** training data. Both settings are evaluated on the same inner part, item by item.

The median PR-AUC/base for the frozen protein language model is 1.85 with same-center training and 1.01 with cross-center training, against a measured random control of 1.03; the composition-only model shows the same pattern, with medians of 1.58 and 0.97 (Table 9). Because both settings are evaluated on the same test items, the contrast is not explained by a difference in the evaluated proteins or in their failure prevalence.

Table 9
Median PR-AUC/base by source of training data, identical test items

Source of training data	Composition GBDT	Frozen PLM
Same center as the test items	1.58	1.85
Other centers	0.97	1.01
Measured random control	1.03	1.03

Medians over four center folds. Both settings are evaluated on the same test items, so the contrast contains no test-population difference. These are medians, not a measured causal effect of laboratory identity.

These are medians over folds, not a measured causal effect of laboratory identity. The fixed test set removes a test-population difference, but changing the training centers also changes the training protein population and the associated protocols.

The cross-center training sets are larger, not smaller (Table 10). After redundancy reduction the dataset retains one record per split group, so these are homology-cluster counts rather than record counts inflated by homologous redundancy. The cross-center side has 4.0 to 17.7 times as many training clusters as the same-center side, and nevertheless performs worse on identical test items. A larger cross-center training set rules out "fewer source examples" as a simple explanation; it does not rule out insufficient *relevant* examples, underfitting, or other optimization effects under the configurations tested here. The random control falling near 1.0 indicates that the evaluation is calibrated and that the low cross-center values are not a computational error.

Table 10
Training homology clusters per center fold

Fold	Held-out centers	Same-center clusters	Cross-center clusters	Ratio
1	NESG, RSGI, NATPRO	1650	29284	17.7
2	MCSG, SGPP, SGX, CSMP, MPP, TRANSPORTPDB	5768	23107	4.0
3	NYSGRC, NYCOMPS, SSGCID, SECSG, BSGI, BSGC	4472	25051	5.6
4	CSGID, NYSGXRC, EFI, CESG, TB, SPINE	4932	24360	4.9

One record is retained per split group after redundancy reduction, so these are homology-cluster counts rather than record counts.

One implementation trap in this comparison is worth recording, because we fell into it. The same-center training set was initially defined as the part of the held-out center not belonging to the test group; since the test fold is *all* of that center's groups, the mask was identically empty and the same-center column

could never obtain a value. The error was found before any GPU run by asserting that the mask was non-empty, and the three-part inner division described above is the correction.

The protein language model does not exceed the composition-only baseline

On the same held-out centers and test folds, the composition-only model is compared against the frozen and low-rank-adapted protein language models (Table 11). On the only fold with a stable appreciable signal, fold 1, the composition model attains 1.33 with a 95% interval of [1.25, 1.42], which exceeds the frozen model's 1.20 and overlaps the adapted model's interval of [1.21, 1.37]. Substituting a 650M-parameter protein language model therefore buys nothing beyond 20-dimensional amino-acid composition in this cross-center setting. This is the same observation as the composition baseline matching a published tool above, not an independent second result.

Table 11
Fold-level PR-AUC/base for expression failure

Fold	n	Prevalence	Composition GBDT	Frozen PLM	LoRA-adapted PLM
1	2475	0.306	1.33 [1.25, 1.42]	1.20 [1.13, 1.28]	1.28 [1.21, 1.37]
2	8652	0.032	0.86 [0.81, 0.94]	0.97 [0.90, 1.07]	1.13 [1.0199, 1.3076]
3	6708	0.413	0.77 [0.76, 0.78]	0.79 [0.77, 0.80]	0.84 [0.83, 0.86]
4	7399	0.522	1.02 [0.9957, 1.0399]	1.05 [1.03, 1.07]	1.04 [1.0139, 1.0582]

Held-out centers per fold are given in Table 10. Square brackets are 95% cluster-stratified bootstrap intervals; four decimal places are shown where an endpoint falls within 0.02 of 1.0. Classification by Table 5: fold 1 appreciable enrichment for all three models; fold 2 below baseline, compatible with baseline and borderline respectively; fold 3 below baseline throughout; fold 4 compatible with baseline, then weak but resolvable for both language models.

The composition model falls below the prevalence baseline on fold 2 while the language models do not, which indicates that the language-model representation at least does not make that fold worse. It is not evidence that the language model is better, because the adapted model on fold 2 is itself a borderline case, as described below.

Four distinct regimes across held-out centers

For expression failure with the frozen model, PR-AUC/base ranges from 0.79 to 1.20 across the four held-out center groups (Table 11). Group 1 shows enrichment above the prevalence baseline at 1.20 [1.13, 1.28]; group 2 is compatible with that baseline at 0.97 [0.90, 1.07]; group 3 falls entirely below it at 0.79 [0.77, 0.80]; and group 4 shows a small but statistically resolvable increase at 1.05 [1.03, 1.07]. A

single summary score across centers would obscure these differences. Directions are positive, indeterminate, below baseline and positive; magnitudes range from − 21% to + 20%. Even restricting attention to the two folds with positive direction, the magnitudes differ fourfold, at + 5% against + 20%. This cross-center inconsistency is supported by all four folds and is the principal result of this work.

On the meaning of a below-baseline result. We use "below the prevalence baseline" when the entire interval for PR-AUC/base lies below 1.0. This establishes depletion of failures among high-scoring candidates under this metric. It does not imply that every pairwise ordering is reversed, that inverting all scores would produce a useful predictor, or that any particular fixed-budget decision is harmful. Such claims would require direct tests of those decisions.

The below-baseline result on group 3 is stable across model levels and interventions (Table 14): [0.77, 0.80] for the frozen model, [0.83, 0.86] after low-rank adaptation, and [0.78, 0.81] for the adversarial variant. It appears on one held-out center group of one task, and we do not describe below-baseline transfer as a general property of this kind of data; it is one of four regimes.

Table 14
Interventions applied to the below-baseline fold and what each tests

Intervention	Result on fold 3, expression	What the result tests
None, frozen PLM	0.79 [0.77, 0.80]	Reference
LoRA adaptation, 5.41M parameters, 0.82%, three epochs	0.84 [0.83, 0.86]	One fine-tuning configuration; part of the model is updated
Adversarial center head, accuracy 0.429–0.507	0.80 [0.78, 0.81]	Reduced center discrimination by that classifier, not statistical independence
Neither intervention moves the interval above 1.0. Adversarial-head accuracy near the majority-class rate is not evidence that center information has been removed from the representation.		

Task definition changes the result within one center group

For the composite soluble-expression endpoint, the frozen model is above the prevalence baseline on group 3 at 1.05 [1.03, 1.07], while expression failure on the same group is below it (Table 12). The directions are therefore opposite within one held-out center group (Table 13).

Table 12
Fold-level PR-AUC/base for the composite soluble-expression endpoint

Fold	n	Prevalence	Frozen PLM	LoRA-adapted PLM
1	2539	0.404	1.28 [1.22, 1.35]	1.32 [1.26, 1.40]
2	5388	0.080	0.99 [0.94, 1.07]	1.01 [0.95, 1.08]
3	4693	0.640	1.05 [1.03, 1.07]	1.02 [1.0010, 1.0415]
4	7147	0.611	1.12 [1.10, 1.14]	1.13 [1.11, 1.15]

The adapted-model result on fold 3 is recorded as compatible with baseline despite interval endpoints that do not cross 1.0, because the direction verdict disagrees across the five bootstrap seeds and the conservative reading is taken. Sample sizes differ from Table 11 because eligibility differs between the two endpoints.

Table 13
Direction and magnitude by center fold, expression versus soluble expression

Fold	Frozen: expression	Frozen: soluble expression	Frozen directions	Adapted: expression	Adapted: soluble expression	Adapted directions
1	Positive, 1.20	Positive, 1.28	Agree	Positive, 1.28	Positive, 1.32	Agree
2	Indeterminate, 0.97	Indeterminate, 0.99	Agree	Positive, 1.13	Indeterminate, 1.01	Not comparable
3	Below baseline, 0.79	Positive, 1.05	Opposite	Below baseline, 0.84	Indeterminate, 1.02	Not comparable
4	Positive, 1.05	Positive, 1.12	Agree	Positive, 1.04	Positive, 1.13	Agree

Each row compares the same held-out center group, not an identical set of labeled sequences; eligible sample sizes differ between the two endpoints, 6708 against 4693 for fold 3. The opposite directions on fold 3 are a single observation for the frozen model and are not reproduced as an adjudicable comparison after adaptation.

This comparison concerns the same held-out center group, not an identical set of labeled sequences. The eligible test samples differ between the two tasks: group 3 contributes 6708 eligible expression samples and 4693 eligible soluble-expression samples. An item-matched claim would require a test restricted to sequences observed for both endpoints, which we did not perform.

The three other folds agree in direction between the two tasks and differ only in magnitude. Furthermore, after low-rank adaptation the soluble-expression result on group 3 is classified as inconclusive under the bootstrap-seed criterion: the five seeds disagree, so the conservative reading is taken and the cell is recorded as compatible with the baseline at a point estimate of 1.02, the same sign as the frozen model but not resolvable. The opposite directions are therefore a specific observation for the frozen model in

one center group, and not evidence of a general cross-task reversal. The adapted-model comparison neither contradicts the frozen-model result, since the point estimates share a sign and no fold reverses direction, nor independently corroborates it.

Taken together, these two findings bear on the two most natural remedies. Past good performance at a given center does not license transfer, because direction and magnitude are both inconsistent across centers; and an absence of difficulty on one task does not license an assumption about another, because at least one counterexample exists. No quantity available in advance — center identity, base failure rate, sample size, or task — predicted which regime a given fold would fall into.

Interventions tested

Two interventions were applied to the below-baseline fold, and we describe what each tested rather than treating either as a categorical exclusion (Table 14).

Low-rank adaptation of 5.41M parameters, which is 0.82% of the model, for three epochs leaves the below-baseline result in place at 0.84 [0.83, 0.86]. This tests one fine-tuning configuration; it updates only part of the model and does not exclude better architectures, longer or differently tuned optimization, or protocol-aware models.

The adversarial center-discrimination head reaches an accuracy of 0.429 to 0.507, which is close to majority-class guessing, while the below-baseline result is unchanged at 0.79 against 0.80. This shows reduced center discrimination *by that classifier*, which is not a demonstration of statistical independence between the representation and center membership. Likewise, omitting an explicit length feature does not guarantee that length-related information is absent from a learned representation.

Both interventions leave the result in place, which points to the same open question: the behavior may originate in a label-generating mechanism that varies by center rather than in the sequence representation. The present data cannot adjudicate this, and we do not speculate further.

A borderline case, and the standard applied to this work

Low-rank adaptation on expression fold 2, where the prevalence is 0.032, gives 1.13 [1.0199, 1.3076], an interval that formally does not cross 1.0. With the adversarial head added it returns to [0.96, 1.22], which does cross 1.0, and the lift point estimate of 0.62 points in the opposite direction. The direct basis for classifying this as borderline is that a robust signal does not vanish on the addition of an adversarial head.

Measured against the two-condition standard described in Methods — an interval that does not cross 1.0, together with a point estimate exceeding a trivial baseline evaluated on the same task and test items — the frozen model satisfies both conditions on one of four held-out center groups, and the adapted model on one of the four groups with predictions recorded. Condition (a) alone is insufficient: with sufficiently large n even a 2% increase is resolvable, as fold 4 of this work illustrates at 1.05 [1.03, 1.07],

where the direction is resolvable but the magnitude is below the trivial baseline. This standard is not set only for others.

Discussion

Interpretation of the matched center comparison

On the matched test subsets, the frozen protein language model achieved a median PR-AUC/base of 1.85 when trained on other homology groups from the same centers, compared with 1.01 when trained on data from the remaining centers. The composition-based model showed the same pattern, with medians of 1.58 and 0.97, respectively. The measured random control was 1.03. Because both training settings were evaluated on the same test items, the contrast is not explained by a change in the evaluated proteins or their failure prevalence.

The cross-center training sets contained 4.0-17.7 times more homology groups than the same-center sets. Their poorer performance therefore cannot be attributed simply to having fewer training examples. Increasing the amount of data drawn from other centers did not compensate for the change in experimental and reporting context under the configurations tested here. These results show that predictive associations can be learned within a center group while transferring poorly beyond it.

The comparison does not isolate a single cause of the gap. Center membership also captures differences in target selection, expression protocols, construct design and outcome recording, which are incompletely described in the public records. Together with the tag and length analyses, the matched evaluation supports a substantial contribution from center-associated variation. It does not determine how much of that contribution arises from reporting artifacts and how much reflects genuine differences in experimental outcomes. The practical implication is that performance on new sequences from familiar centers should not be assumed to describe performance in a new laboratory.

Implications of heterogeneous transfer

For expression failure with the frozen model, PR-AUC/base ranged from 0.79 to 1.20 across the four held-out center groups. Group 1 showed enrichment above the prevalence baseline (1.20, 95% CI 1.13–1.28), group 2 was compatible with that baseline (0.97, 0.90–1.07), group 3 fell below it (0.79, 0.77–0.80), and group 4 showed a small, statistically resolvable increase (1.05, 1.03–1.07). These differences would be obscured by a single summary score. In particular, an interval above 1.0 establishes enrichment under this metric, but does not establish that the gain is large enough to improve a given experimental selection strategy.

Task definition also changed the result within one center group. For group 3, the frozen model was below baseline for expression failure but above baseline for the composite soluble-expression endpoint (1.05, 1.03–1.07). This comparison concerns the same held-out centers, not an identical set of labeled sequences: the eligible test samples differed between tasks. Moreover, after low-rank adaptation, the

soluble-expression result was classified as inconclusive under the bootstrap-seed criterion. The opposite directions are therefore a specific observation for the frozen model in one center group, rather than evidence of a general cross-task reversal.

For users, the relevant question is whether the predictor enriches for the intended outcome in the laboratory and task where it will be applied. In the below-baseline group, using high failure scores to exclude candidates could remove proteins less likely to fail under the recorded endpoint. However, the appropriate action at a fixed experimental budget requires a separate evaluation of the proposed selection rule. Neither a favorable result at another center nor success on a related task establishes that benefit. These data support evaluation in the intended setting; they do not establish that transfer is intrinsically impossible or that useful predictors of transfer cannot be developed.

Limitations

The six-stage schema provides a common representation of the records, but does not make the underlying endpoints interchangeable. The center-transfer analyses are dominated by TargetTrack-derived expression labels. Among the production stages used for these analyses, expression failure has the largest number of negative homology groups. Stability and binding measurements contribute to the wider resource, but arise largely from different assays and do not establish center-level transfer across a complete six-stage production pipeline. Likewise, large numbers of mutational measurements on a small set of parent proteins provide limited evidence about performance on unrelated proteins, even though they remain informative for variant-effect prediction.

Solubility-specific conclusions are particularly limited by label provenance. Approximately 99.9% of the negative labels for the separate soluble stage were inferred from recorded experimental progress rather than obtained from explicit solubility measurements. The explicit-negative control subset contained only 10 records, all from one center and none carrying a His6 motif. This subset cannot distinguish a tag-associated biological effect from differences in reporting. An inferred negative may also represent an unrecorded measurement, cessation of work or failure at another step. The soluble stage must therefore be distinguished from the composite soluble-expression endpoint, and neither should be interpreted as a direct measurement of intrinsic thermodynamic solubility.

Homology control also constrains the evaluated protein population. Transitive grouping at 30% identity and 80% coverage produced a large connected component spanning 28646 set-cover clusters. Assigning groups larger than 1% of the dataset to training prevented a single component from dominating a test partition, but systematically excluded proteins in these large homology networks from held-out evaluation. The resulting scores describe the retained test population and may not extend to proteins in large families or densely connected multidomain groups. Cluster counts quantify diversity under the specified search and grouping procedure; they are not an exact count of statistically independent biological samples.

The available groups limit the precision and scope of the transfer analysis. Several centers were combined within each of four held-out groups, so a fold-level result need not describe every constituent laboratory. Cluster-stratified confidence intervals quantify uncertainty within these groups, not uncertainty over all possible laboratories. The temporal test contained only 606 records, and only two species groups met the prespecified size threshold in the species audit. These analyses do not support broad claims about temporal robustness or cross-species transfer. Repeating bootstrap seeds checks the stability of interval-based classification; it does not assess variation across independent model-training runs.

The interventions test selected explanations without exhausting them. Competitive performance of the composition baseline and persistence of the below-baseline result after low-rank adaptation do not exclude better architectures, optimization procedures or protocol-aware models. Similarly, near-majority accuracy of the tested adversarial center classifier shows reduced center discrimination by that classifier, not removal of all center-related information. Omitting an explicit length feature also does not guarantee that length-related information is absent from a learned representation. Public metadata are insufficient to separate target selection, experimental conditions and reporting practices causally. The study therefore establishes limited transfer under the tested conditions, rather than a universal inability to remove laboratory-associated bias.

Finally, all evaluations were retrospective and included no new wet-lab experiments. The transferred SoluProt comparison is affected by differences in applicability, preprocessing and feature availability and should not be read as a ranking of methods in their intended domains. PR-AUC/base measures enrichment relative to prevalence; it neither measures probability calibration nor makes scores fully invariant to class prevalence. Its magnitude does not directly specify the benefit of a fixed-budget selection strategy. The study's magnitude thresholds are operational evaluation criteria, not universal thresholds of experimental utility. Computational job failures and structure-confidence scores from the in-house pipeline cannot substitute for experimentally measured production or binding outcomes. Prospective data collected under documented protocols, with explicit failure recording, are needed to establish the practical benefit of using these predictors in a new laboratory.

Conclusions

Public protein-production records contain learnable associations, but their transfer between laboratories is uneven. In our matched evaluations, training on other centers produced substantially poorer predictions than training on data from the test centers, despite the larger cross-center training sets. Performance on held-out center groups ranged from enrichment above the prevalence baseline to performance below it. Removing recognizable construct features did not ensure transfer, and the below-baseline result in one group persisted after low-rank adaptation and the tested adversarial intervention. Together, these findings show why sequence-homology control must be accompanied by an explicit assessment of transfer between laboratories.

The unified labels, frozen splits and evaluation code provide a reproducible basis for this assessment across the represented failure stages and evaluation settings. For a laboratory deciding which proteins to pursue, the relevant evidence is performance on the intended experimental endpoint under its production protocol. A favorable score on a pooled dataset or at another center does not establish that benefit. Prospective studies with documented protocols and explicit failure records are needed to determine whether model-guided selection improves experimental outcomes, and to distinguish transferable sequence associations from those specific to the conditions represented in the training data.

Abbreviations

PR-AUC
area under the precision-recall curve
PR-AUC/base
PR-AUC divided by the positive-class prevalence
GBDT
gradient-boosted decision tree
PLM
protein language model
LoRA
low-rank adaptation
CI
confidence interval
LLM
large language model

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

GL is the founder and chief executive officer of Open Intelligence Hub, a commercial computational drug-discovery company, which funded this work. Because this is the strongest form the relationship could take, the three consequences are stated explicitly rather than left to be inferred. First, this paper

examines the evaluation protocols of published methods in a field in which the funder operates commercially. Second, data generated by the funder's own pipeline are used in this work (4762 designs, used only as an unlabeled probe set and contributing to no conclusion). Third, a by-product of this work, the calibration of a structure-confidence threshold against designs with real wet-lab outcomes, is directly applicable to the funder's pipeline; it is reported in reports/ds6_oih_inventory.md in the companion repository and is not used in any conclusion of this paper. GL declares no other competing interests.

Figures, tables and additional files

This submission comprises the manuscript file only. No figures and no additional files are submitted, and the text makes no reference to any.

All tables are supplied within the manuscript file using the word-processor table function, not as images or spreadsheets, without color or shading and without thousands separators, with a title of at most 15 words above each table and the notes below it.

The supporting material that would otherwise have been supplied as additional files is instead released in the companion repository and the archived deposit cited under Availability of data and materials, where each item is a file under version control rather than a static copy: the evaluation changelog in configs/stage3_train.yaml; the data inventory and bias audit in reports/gate1_data_inventory.md; the split construction and leakage verification in reports/splits_and_leakage.md; the baseline comparisons in reports/gate2_baselines.md; the full Data and Methods account in reports/paper_data_methods.md; the frozen artifact fingerprints in the .prov.json file accompanying each released artifact; the in-house pipeline inventory in reports/ds6_oih_inventory.md; and the ProteinGym assay list in reports/proteingym_assays_used.md. We will supply any of these as a formatted additional file on request.

Funding

This work was funded by Open Intelligence Hub (Suzhou, China). No public or external grant support was received. The funder is the company of which the author is founder and chief executive officer; the author therefore participated in the conceptualization, design, analysis, decision to publish and preparation of the manuscript in that capacity. No other party had any role in the study.

Author Contribution

G.L. conceived the study, designed the evaluation protocol, directed the analysis, interpreted the results and wrote the manuscript. The author reviewed the manuscript.

Acknowledgements

Not applicable.

Data Availability

This study generated no data of the types for which BMC mandates deposition in a community-endorsed repository: no new protein sequences, DNA or RNA sequences or sequencing data, genetic polymorphisms, linked genotype-phenotype data, macromolecular structures, microarray or gene-expression data, and no crystallographic data were produced. Every protein sequence analyzed is a pre-existing public record obtained from an established archive.The datasets generated and analyzed during the current study are available in the Zenodo repository under DOI 10.5281/zenodo.23189683 (https://doi.org/10.5281/zenodo.23189683), openly accessible under CC BY-SA 4.0. The deposit contains the unified six-stage failure labels, the four frozen evaluation splits, the center-fold assignments and the per-artifact provenance records. The source code and all reports cited in this article are available in the protein-failure-benchmark repository (https://github.com/liugangg/protein-failure-benchmark) and are archived in the same deposit.The five upstream archives from which the records were obtained are cited in the reference list of the manuscript with their own persistent identifiers: PSI TargetTrack (10.5281/zenodo.821654), Tsuboyama et al. (10.5281/zenodo.7844779), ProteinGym (https://github.com/OATML-Markslab/ProteinGym), Overath et al. (10.5281/zenodo.15722219), and the Adaptyv Bio EGFR design competition (https://github.com/adaptyvbio/egfr_competition_1).

References

1. Thumuluri V, Martiny HM, Almagro Armenteros JJ, Salomon J, Nielsen H, Johansen AR. NetSolP: predicting protein solubility in Escherichia coli using language models. Bioinformatics. 2022;38(4):941–6. https://doi.org/10.1093/bioinformatics/btab801.
2. Hon J, Marusiak M, Martinek T, Kunka A, Zendulka J, Bednar D, et al. SoluProt: prediction of soluble protein expression in Escherichia coli. Bioinformatics. 2021;37(1):23–8.

https://doi.org/10.1093/bioinformatics/btaa1102.

3. Zhang X, Hu X, Zhang T, Yang L, Liu C, Xu N, et al. PLM_Sol: predicting protein solubility by benchmarking multiple protein language models with the updated Escherichia coli protein solubility dataset. Brief Bioinform. 2024;25(5):bbae404. https://doi.org/10.1093/bib/bbae404.
4. Tankhilevich E, Martinez Cuesta S, Barrett I, Berg C, Holmberg Schiavone L, Leach AR. RP3Net: a deep learning model for predicting recombinant protein production in Escherichia coli. Bioinformatics. 2026;42(1):btag003. https://doi.org/10.1093/bioinformatics/btag003.
5. Rajagopalan R, Sharma Meda R, Shastry S, Mysore V. Protein solubility depends on centrifugation: Aiki-Sol, a per-regime predictor for E. coli. bioRxiv [Preprint]. 2026. https://doi.org/10.64898/2026.05.14.725067
6. Wang H, Wang M, Tan H, Li Y, Zhang Z, Song J. PredPPCrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. PLoS ONE. 2014;9(8):e105902. https://doi.org/10.1371/journal.pone.0105902.
7. Oeschey J. Published benchmark AUROC does not predict generalisation: an independent evaluation of protein solubility predictors on native and heterologous E. coli proteins. Res Square [Preprint]. 2026. https://doi.org/10.21203/rs.3.rs-10471634/v1.
8. Price WN, Handelman SK, Everett JK, Tong SN, Bracic A, Luff JD, et al. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in Escherichia coli. Microb Inf Exp. 2011;1(1):6. https://doi.org/10.1186/2042-5783-1-6.
9. Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics. 2011;27(13):i24–33. https://doi.org/10.1093/bioinformatics/btr229.
10. Wang H, Feng L, Zhang Z, Webb GI, Lin D, Song J. Crysalis: an integrated server for computational analysis and design of protein crystallization. Sci Rep. 2016;6:21383. https://doi.org/10.1038/srep21383.
11. Meng F, Wang C, Kurgan L. fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. BMC Bioinformatics. 2017;18(1):580. https://doi.org/10.1186/s12859-017-1995-z.
12. Zhu Y, Hu J, Ge F, Li F, Song J, Zhang Y, Yu DJ. Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features. Brief Bioinform. 2021;22(3):bbaa076. https://doi.org/10.1093/bib/bbaa076.
13. Wang H, Feng L, Webb GI, Kurgan L, Song J, Lin D. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. Brief Bioinform. 2018;19(5):838–52. https://doi.org/10.1093/bib/bbx018.
14. Bushuiev A, Bushuiev R, Sedlar J, Pluskal T, Damborsky J, Mazurenko S, Sivic J. Revealing data leakage in protein interaction benchmarks. In: ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design. 2024.

15. Bushuiev A, Bushuiev R, Kouba P, Filkin A, Gabrielova M, Gabriel M et al. Learning to design protein-protein interactions with enhanced generalization. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR). 2024. Dataset references.
16. Berman HM, Gabanyi MJ, Kouranov A, Micallef DI, Westbrook J. Protein Structure Initiative network of investigators. Protein Structure Initiative - TargetTrack 2000–2017 - all data files [Data set]. Zenodo. 2017. https://doi.org/10.5281/zenodo.821654.
17. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ et al. Mega-scale experimental analysis of protein folding stability in biology and protein design [Data set]. Zenodo. 2023. https://doi.org/10.5281/zenodo.7844779.
18. Notin P, Kollasch A, Ritter D, van Niekerk L, Paul S, Spinner H, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. Adv Neural Inf Process Syst. 2023;36:64331–79.
19. Overath MD, Rygaard A, Jenkins TP. Dataset for: predicting experimental success in de novo binder design: a meta-analysis of 3766 experimentally characterised binders [Data set]. Zenodo. 2025. https://doi.org/10.5281/zenodo.15722219.
20. Adaptyv Bio. EGFR protein design competition, rounds 1 and 2. 2024–2025. https://github.com/adaptyvbio/egfr_competition_1; https://github.com/adaptyvbio/egfr_competition_2
21. Liu G. Protein-failure benchmark: unified six-stage labels, leakage-verified frozen splits, and a center-stratified evaluation protocol [Data set]. Zenodo. 2026. https://doi.org/10.5281/zenodo.23189683
22. ProteinGym. additionally requires the original experimental paper for each deep mutational scanning assay to be cited. This work uses 33 assays, which derive from 13 original experimental papers; the assay identifiers and their source papers are listed in reports/proteingym_assays_used.md in the companion repository, and the complete bibliographic entries are distributed by ProteinGym as assays.bib.