# Experimental context limits transfer of microprotein sequence scores

Robin Gounder[1], Samuel Schmidt[2,3]

[1]University of Cambridge

[2]Vaionex Corporation

[3]Cornell University

Correspondence: robin@gounder.com

Samuel Schmidt: samuel@vaionex.com

10 November 2026

**Abstract**

Sequence scores offer a way to prioritize microproteins encoded by noncanonical open reading frames and propose edits that alter their accumulation. We reanalysed published measurements to test whether these scores remain useful across reporter contexts, independent peptide datasets and paired sequence edits. Among 1,072 development constructs, an annotation-based composition score correlated negatively with reporter accumulation (Spearman $\rho = -0.292$), whereas models trained on accumulation reached 0.747 using composition and 0.780 with a richer terminal-aware representation. On the same 1,044 peptides across reporters, the within-GFP advantage of the richer representation was uncertain ($\Delta\rho = 0.021$, conditional 95% interval $-0.005$ to 0.046). When applied to microprotein–mCherry, the GFP-trained richer model reached $\rho = 0.560$, compared with 0.663 for composition alone and 0.742 for a model trained on the target reporter. The paired composition-minus-richer contrast was 0.103 (conditional 95% interval 0.078 to 0.132). Tests on independent canonical peptides and 1,743 paired sequence changes showed that model comparisons also depended on input coverage, training exposure and whether the task was ranking peptides, selecting a subset or predicting edit effects. Strong published baselines reduced several apparent advantages. These retrospective benchmarks quantify predictive portability within the tested populations; they do not isolate causal effects of reporter components. Performance in one reporter does not justify sequence editing or use in another assay without testing. Native turnover and peptide-dependent function require separate experiments.

**Keywords:** microproteins; noncanonical open reading frames; protein stability; sequence predictors; reporter assays; cross-assay transfer; reproducibility

# 1 Introduction

Thousands of translated open reading frames fall outside conventional protein annotations, creating a large set of candidate microproteins [1,2]. We refer to these as noncanonical ORFs (nORFs); canonical proteins are those encoded by conventionally annotated coding genes. Sequence scores can help prioritize these candidates, but deciding which peptide to study requires clarity about what a score predicts. Translation, accumulation of a peptide product and peptide-dependent function are different observations. Some noncanonical peptides have established cellular phenotypes or native complex structures [3–5]; an upstream ORF can also affect downstream expression through its RNA or the act of translation [6]. Endogenous-variant and degron-exposure experiments illustrate how particular biological effects can be tested [7, 8].

Existing predictors address different parts of this problem. ShortStop classifies sequences by resemblance to curated microproteins relative to synthetic references [27]. PepScore integrates ORF features and was accompanied by individual expression assays [28]. Recent preprints distinguish support from ribosome profiling and mass spectrometry, or predict detection in MS repositories using translation, conservation and peptide attributes [29, 30]. These are useful but distinct tasks. A score for annotation resemblance or detection is not automatically a predictor of how much peptide accumulates in a reporter assay. Testing such a score also requires the inputs expected by the original model, including native RNA context where applicable.

This distinction also matters when interpreting sequence features. The six-feature XGBoost nORF Score showed no significant association with EggNOG Gene Ontology annotation [9]. Our controlled reanalysis retained compositional differences after exact length matching. Detecting such differences between annotated and noncanonical sequences, however, does not show that the same features predict reporter accumulation or the effect of editing a particular peptide. Those claims require tests against the corresponding measurements.

For predicting accumulation, published degradation models provide more direct comparators. These models score degrons: sequences that can direct a protein toward degradation. QCDPred predicts yeast quality-control degrons from composition, with companion experiments on human reporters [10, 11]; composition-based methods also support internal degron mapping [12]. Guo et al. studied composition, genetic-code and GC biases, disorder and terminal residues using microprotein reporters, substitutions, extensions and paired reporter configurations [13]. The core assays fuse microproteins to fluorescent reporter proteins and measure the resulting signal. Fusions and tags can change which peptide termini are exposed, making construct context a plausible reason for differences between assays. Cytosolic degron maps and terminal stability screens provide the released peptide abundance predictor (PAP) and DEGRONOPEDIA models [14, 18–20]. C-terminal variation and composition-preserving shuffles have also been studied [15, 16], alongside the role of GC constraints in newly evolved sequences [17]. Thus, the importance of composition and sequence context is already established. The practical question is how reliably a score trained in one of these settings can be used in another.

Calibration introduces a related question: does a better numerical fit reflect improved prediction, a change in measurement scale or a biological relationship? Experiment-specific nonlinear calibration and variable gains from added features are documented [21]. MAVE-NN and MoCHI model genotype–phenotype and measurement relationships, while recognizing limits on separating their nonlinearities [22, 23]. Joint abundance–binding and abundance–activation models [24, 25], and experimentally tested multi-mutant improvements [26], further show why predictive accuracy, physical interpretation and function must be evaluated separately.

We ask whether a score that predicts reporter accumulation in one setting remains useful with another reporter or for choosing sequence edits. We first compare annotation-based scores with models trained on reporter measurements. We then hold peptide identity fixed to test prediction across reporter contexts, before testing independent canonical-peptide and paired-edit datasets. Comparisons with strong published models, identical sequence inputs and checks for overlap with training data help explain apparent differences in performance. This provides a controlled account of when the tested scores remain useful and where they fail. The later analyses were developed retrospectively after earlier results were known; detailed exploratory results and candidate follow-up remain in S1 Text.

Our related guide-selection analysis quantified how earlier CRISPR reagent choices change later depletion estimates [61]. Some ORFs, reference resources and biological precedents are shared with that work. Here the primary measurements are reporter accumulation and paired sequence changes, and the endpoints are predictive portability, input coverage and training exposure; we do not repeat its guide-selection estimands. The contribution is a peptide-matched benchmark of these predictions and their limits, rather than a new reporter experiment or evidence of native peptide function.

## 2 Results

We organize the analyses around the same practical question: when does a score remain useful outside the setting in which it was trained? The core benchmark tests whether training on reporter measurements improves prediction over annotation-based scores. We then ask whether that accuracy survives a change of reporter, a change of study or a change from ranking peptides to predicting edits. Finally, we examine how peptide selection, sequence inputs and training overlap affect the comparisons. All analyses use existing data. The later analyses were developed after earlier results were known and were not preregistered or designed as prospective confirmation. S1 Text contains the detailed results, sensitivity analyses and exploratory candidate checks.

### 2.1 Training on the measured outcome improves reporter prediction

We first asked whether a sequence score trained to recognize annotated proteins also predicts reporter accumulation. To separate the training task from the sequence features, we compared annotation-based scores with models trained on the reporter measurements. Model A uses length and amino-acid composition, T adds terminal features, and O adds residue-order features; all use the same grouped validation folds.

Among the 1,072 development constructs, the composition score trained to distinguish annotated proteins from reference sequences correlated negatively with the GFP–microprotein reporter outcome ($\rho = -0.292$, 95% conditional interval $-0.344$ to $-0.237$). ShortStop's mean SAM probability showed little rank correlation among peptides with available scores ($\rho = -0.018$; $n = 1,057$). Models trained on the measured outcome performed substantially better: composition model A reached $\rho = 0.747$, and terminal-aware T reached 0.780 (Fig 1; Table 1). The composition model also predicted reporter scores in the additional library after the strict exclusions ($\rho = 0.719$). A score that distinguishes annotated proteins from reference sequences therefore need not predict reporter accumulation. This result does not assess its accuracy on the original annotation task.

Adding native-order features to T produced a small increase in rank correlation in the primary analysis, $\Delta\rho = 0.0040$ (95% conditional interval $-0.0046$ to 0.0115). The increase varied across contexts. S1 Text and Figs S1–S3 report which peptides were included, how related peptides were grouped, the catalogue matching and all residue-order sensitivity analyses.

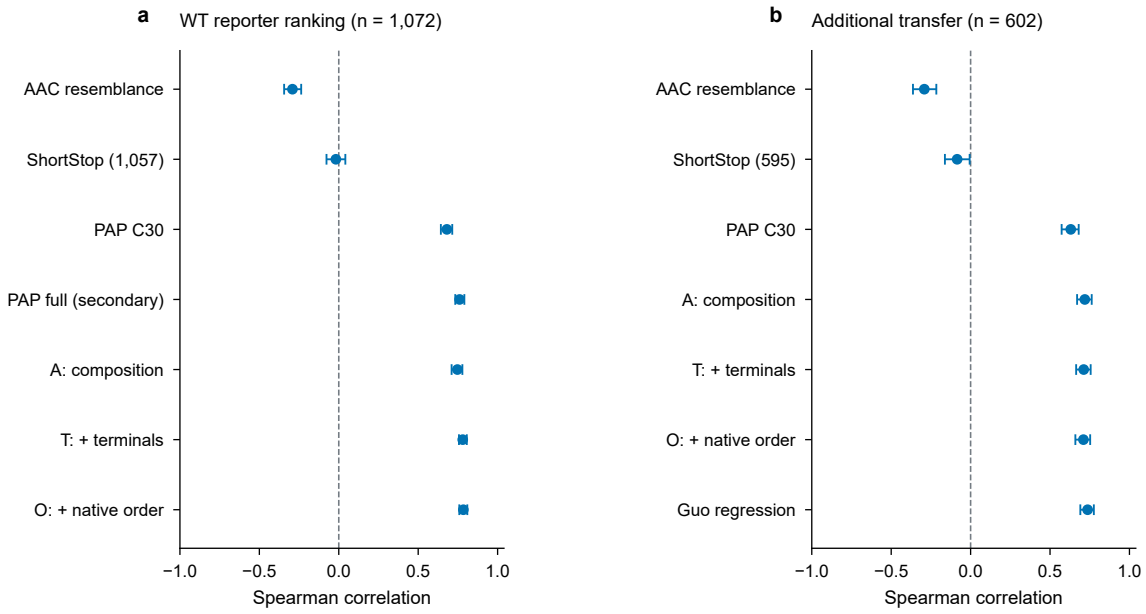


**Fig 1.** Reporter ranking and within-study transfer. Points show Spearman correlation; bars show 95% operational-group bootstrap intervals conditional on fitted predictions. Raw annotation and PAP scores are ranking comparators, not reporter probabilities. A/T/O are group-held-out in panel a and trained on development WTs for transfer in panel b. Coverage exceptions are labelled. The released Guo comparator is training-sequence/group-disjoint on these additional cargos, but not demonstrably prospectively blinded.

**Table 1.** Group-held-out native prediction and strict additional transfer. MAE is shown only for direct endpoint models on reporter scale. Additional sequences belong to the same publication, in a different library.

| | GFP–MP WT, $n = 1,072$ | | Additional strict, $n = 602$ | |
|---|---|---|---|---|
| Model | $\rho$ | MAE | $\rho$ | MAE |
| L: length | 0.207 | 0.2261 | 0.167 | 0.1743 |
| A: composition | 0.747 | 0.1425 | 0.719 | 0.1259 |
| T: plus terminals | 0.780 | 0.1362 | 0.712 | 0.1311 |
| O: plus native order | 0.784 | 0.1354 | 0.710 | 0.1314 |
| T plus shuffled order 1 | 0.770 | 0.1388 | 0.703 | 0.1344 |
| T plus shuffled order 2 | 0.775 | 0.1380 | — | — |
| T plus shuffled order 3 | 0.771 | 0.1378 | — | — |

## 2.2 The same peptides give different results across reporter contexts

Accurate prediction within a reporter assay does not show whether the model will work with a different reporter. We therefore compared the same 1,044 peptides across the Guo reporter contexts, keeping the tested peptide set fixed [13]. Observed rank agreement was $\rho = 0.522$

between the original GFP–MP and MP–mCherry configurations, compared with 0.867 between the two GFP libraries (Fig 2). T trained on the source reporter reached $\rho = 0.560$ on MP–mCherry, versus 0.742 for target-trained T; reverse transfer reached 0.566 on GFP–MP, versus 0.767 for target-trained T (Table 2). The composition-only model performed better than T when applied across the original reporter configurations, reaching 0.663 and 0.648 in the two directions. On this same 1,044-peptide population, the within-GFP T-minus-A contrast was 0.0205 (95% conditional interval −0.0047 to 0.0458), so the within-context increment was uncertain. Across reporters, A-minus-T was 0.1027 (0.0780 to 0.1316) for GFP–MP to MP–mCherry and 0.0819 (0.0581 to 0.1039) in reverse. These paired group-bootstrap contrasts keep fitted predictions fixed and were examined retrospectively. The 1,044-peptide comparison uses its own common-group folds and must be distinguished from the earlier 1,072-construct development estimates.

The richer T representation includes terminal encodings, hydropathy and pretrained PAP components, so its contrast with A does not isolate terminal residues. Its higher within-context point estimates did not imply improved transfer. Reporter identity, orientation, linkers, tags and library conditions all differ between contexts, so their individual effects cannot be separated here. Calibration using target-assay measurements did not close the error gaps. Because that calibration uses target data, it adapts the model rather than independently validating the original predictor.

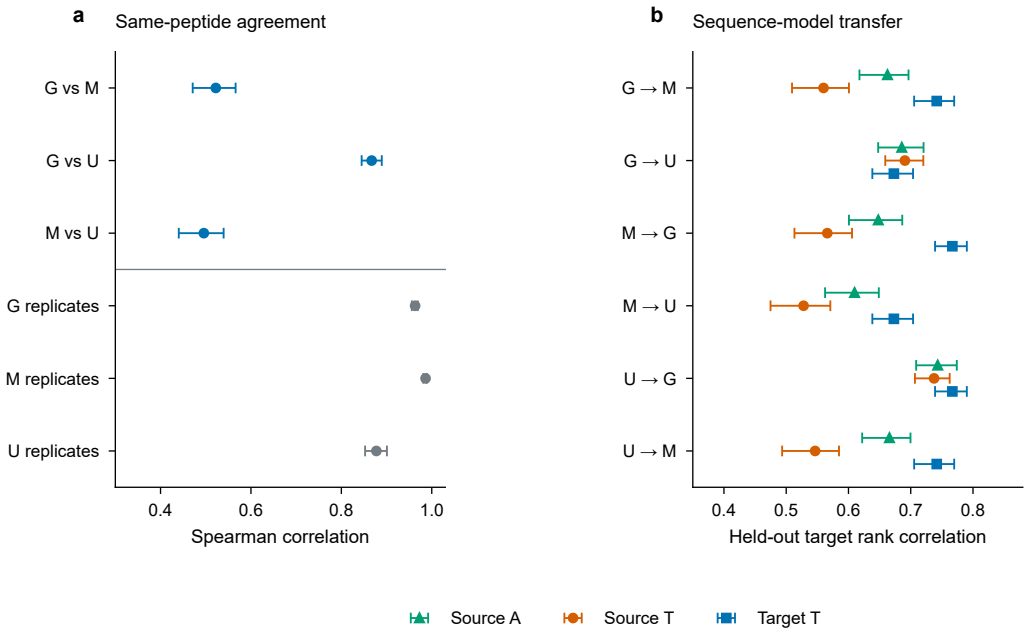


**Fig 2.** Same-peptide agreement and model transfer. a: correlations between context means and within-context biological replicates on the common 1,044-peptide population; missing replicate pairs are omitted as specified in the text. b: all six directed transfers of source-trained A and T versus target-trained T on identical held-out groups. Bars are descriptive 95% operational-group bootstrap intervals conditional on saved predictions. G: original GFP–MP; M: MP–mCherry; U: mutagenesis-library GFP–MP. This is an exploratory same-study comparison.

**Table 2.** Terminal-aware T transfer on the same 1,044 peptides. G: original GFP–MP; M: MP–mCherry; U: mutagenesis-library GFP–MP. Rank transfer uses the source model directly. MAE transfer uses target-scale isotonic calibration learned entirely within outer training groups. The target reference is trained on that target in the same group folds.

| | Target rank correlation | | Target-scale MAE | |
|---|---|---|---|---|
| Transfer | Source-trained | Target-trained | Calibrated transfer | Target-trained |
| G to M | 0.560 | 0.742 | 0.2062 | 0.1572 |
| G to U | 0.691 | 0.673 | 0.1211 | 0.1305 |
| M to G | 0.566 | 0.767 | 0.1820 | 0.1401 |
| M to U | 0.528 | 0.673 | 0.1492 | 0.1305 |
| U to G | 0.737 | 0.767 | 0.1453 | 0.1401 |
| U to M | 0.546 | 0.742 | 0.2065 | 0.1572 |

## 2.3 Performance on independent data depends on the training assay and selection task

The preceding comparison uses reporter contexts from the same study. To test prediction across studies, we next applied the models to independent canonical-protein tiles [14]. These peptide segments test whether models trained on nORF reporters can rank a different sequence population; they do not measure native microprotein function.

Composition models trained in the two nORF reporter configurations reached $\rho = 0.816$ and 0.817 on 14,143 eligible canonical tiles from 383 genes. The corresponding T correlations were 0.812 and 0.656 (Fig 3; Table 3). The difference between A and T was small and uncertain for GFP-trained models, whereas the gap for MP–mCherry-trained models was 0.161 (95% conditional interval 0.149 to 0.174). The A/T difference therefore depended on the training assay; it does not establish a general harmful effect of terminal residues. T also incorporates pretrained PAP features derived from the target study, so its complete representation is not wholly external to that study even though its regression head is trained only on Guo measurements.

The choice of evaluation metric also changed which model performed better. GFP-trained T selected a top decile with higher measured reporter scores despite its slightly lower overall correlation. Overall correlation alone therefore did not identify the model best suited to selecting high-scoring peptides. Published normalized QCDPred reached $\rho = 0.800$, reducing the apparent advantage of the composition model; intervals for the difference in mean scores among the selected top deciles spanned zero.

**Table 3.** nORF models trained only on the source data evaluated on independent canonical tiles. All rows use the same 14,143 held-out tiles from 383 gene groups. Selected means are measured target-assay scores among the highest-scoring 1,415 tiles under each frozen policy; they are not source-scale predictions.

| Source training | Model | External $\rho$ | Selected assay mean |
|---|---|---|---|
| GFP–MP | A | 0.816 | 0.5976 |
| GFP–MP | T | 0.812 | 0.6121 |
| MP–mCherry | A | 0.817 | 0.5970 |
| MP–mCherry | T | 0.656 | 0.5506 |

Ranking unmodified peptides and predicting what happens after an edit are different tasks. A model could do the former well without doing the latter. We therefore tested whether the scores predicted measured changes between paired sequences in the independent Chu study [15].

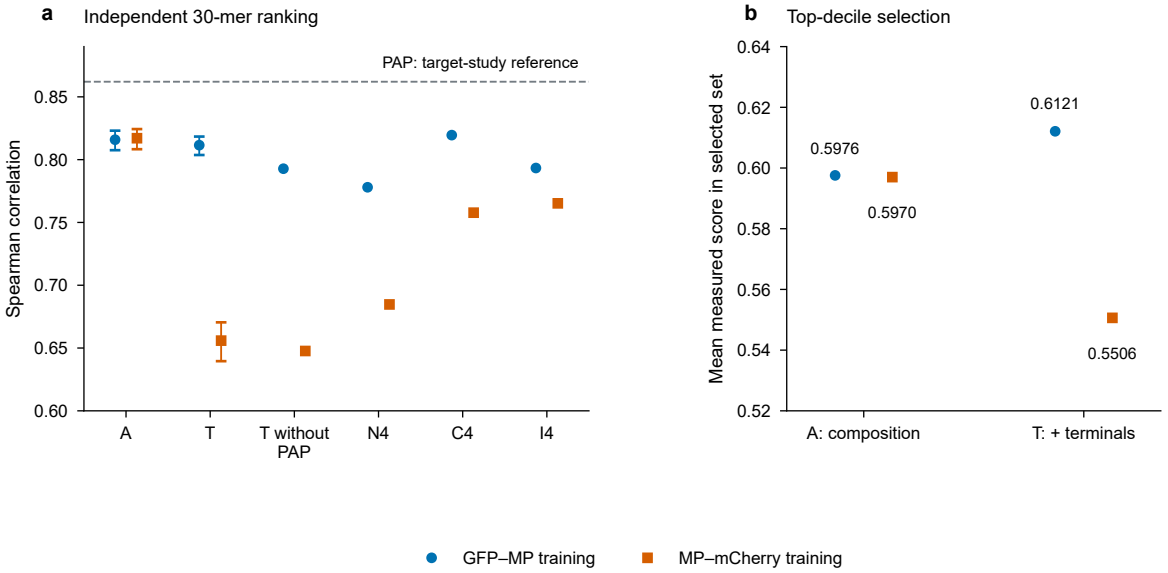


**Fig 3.** Source context and decision metric change the transfer comparison. a: independent-target Spearman correlation for both models trained on the source data. Intervals for A/T are conditional 95% gene-bootstrap intervals; other diagnostics are point estimates. The dashed PAP reference was trained in the target study. N4/C4/I4 each add 80 positional columns to A; T without PAP removes its pretrained components. b: grouped dots show the mean measured target score among the 1,415 peptides selected by each model's highest-scoring 10%. Colors and markers identify the training reporter; labels give the means. Each model selects its own subset. Policy contrasts and uncertainty are reported in the text and S9 Data. The source of the 30-mer assay is independent; the test sequences are canonical tiles rather than native nORFs.

On the 1,743 matched 58-mer pairs retained after common source/PAP exclusions, source-WT calibration improved linear full58 PAP from $\rho = 0.651$ to 0.692. Raw published nonlinear PAP already reached 0.729, and its additional calibration gain was 0.006 (95% conditional interval −0.002 to 0.014; Fig 4; Table 4). Calibration therefore improved this linear model, but the result does not show that calibration generally improves prediction. PAP's author-held-out canonical-tile performance is a target-study reference, not independent external validation of PAP.

**Table 4. Independent paired-effect ranking for frozen baselines.** All entries use the same 1,743 pairs. Calibration is learned solely from Guo GFP–MP WTs. Full58 and C30 inputs are distinct pre-specified representations. Both endpoints of every pair are 58 residues. Values are Spearman correlations with observed RT-minus-WT change; all components/sensitivities are in S10 Data.

| Model and input | Raw | Source-WT calibrated |
|---|---|---|
| QCDPred normalized | 0.602 | 0.622 |
| Human25 normalized | 0.662 | 0.655 |
| Linear PAP, C30 | 0.480 | 0.510 |
| Linear PAP, full58 | 0.651 | 0.692 |
| PAP CNN1, C30 | 0.518 | 0.530 |
| PAP CNN1, full58 | 0.729 | 0.735 |
| PAP CNN7, C30 | 0.524 | 0.533 |
| PAP CNN7, full58 | 0.725 | 0.725 |

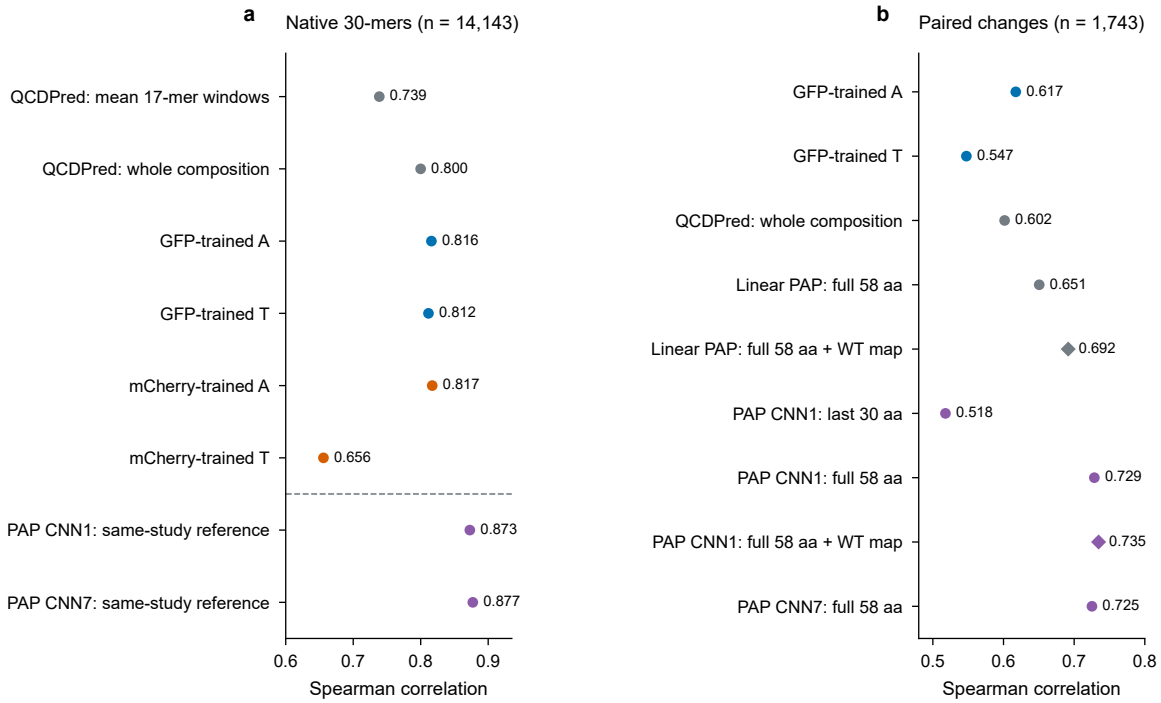


**Fig 4. Published baseline challenge across two fixed cohorts.** (a) Native ranking on 14,143 tiles/383 genes. The dashed separator distinguishes PAP neural models trained in that study from externally trained references. (b) Paired RT-minus-WT ranking on 1,743 independent 58-mer window pairs/1,591 groups after common source/PAP-training exclusions. Diamonds denote Guo-WT-only calibration. Full58 is supported variable-length inference beyond the training 30-mer length. Point estimates are shown; all variants, source contexts and conditional paired contrasts are retained in S10 Data.

## 2.4 Comparisons depend on the tested peptides, model inputs and overlap with training data

Apparent differences between models can reflect the peptides selected for evaluation, the amount of sequence supplied or overlap with training data. We examined these possibilities before interpreting a performance gap as a benefit of particular features or a model architecture.

Changing which peptides were included changed the apparent benefit of terminal features, even though all predictions remained fixed. On all 1,014 valid Leu-to-Ala pairs, T exceeded A by $\Delta\rho = 0.097$. Applying the original terminal-Leu exclusion left 798 pairs and a contrast of $-0.013$ (95% conditional interval $-0.032$ to 0.005). The append-edit advantage persisted. These comparisons were made after the initial analysis, using the source authors' exclusion rule [13]; they do not identify a new causal explanation (Table S1; S1 Text).

We next asked whether the sequence supplied to a model affected its performance. Input coverage also mattered. On the independent pairs, raw CNN1 reached $\rho = 0.518$ on C30 and 0.729 on full58. The full58 representation extrapolates beyond the training length, and this comparison does not identify which residues or mechanisms explain the difference.

A final check asked whether the supposedly independent test sequences had already appeared in the terminal model's development data. At least one sequence in 1,454 of the 1,743 pairs matched that development screen. Neither sequence matched in 289 pairs. In that subset,

DEGRONOPEDIA reached $\rho = 0.537$, compared with 0.521 for PAP on identical C23 inputs and 0.761 for full58 PAP (Fig 5; Table S2). The same-input difference interval spanned zero. These checks identify overlap with the development screen. They cannot determine which sequences entered the individual training split or rule out all earlier training exposure.

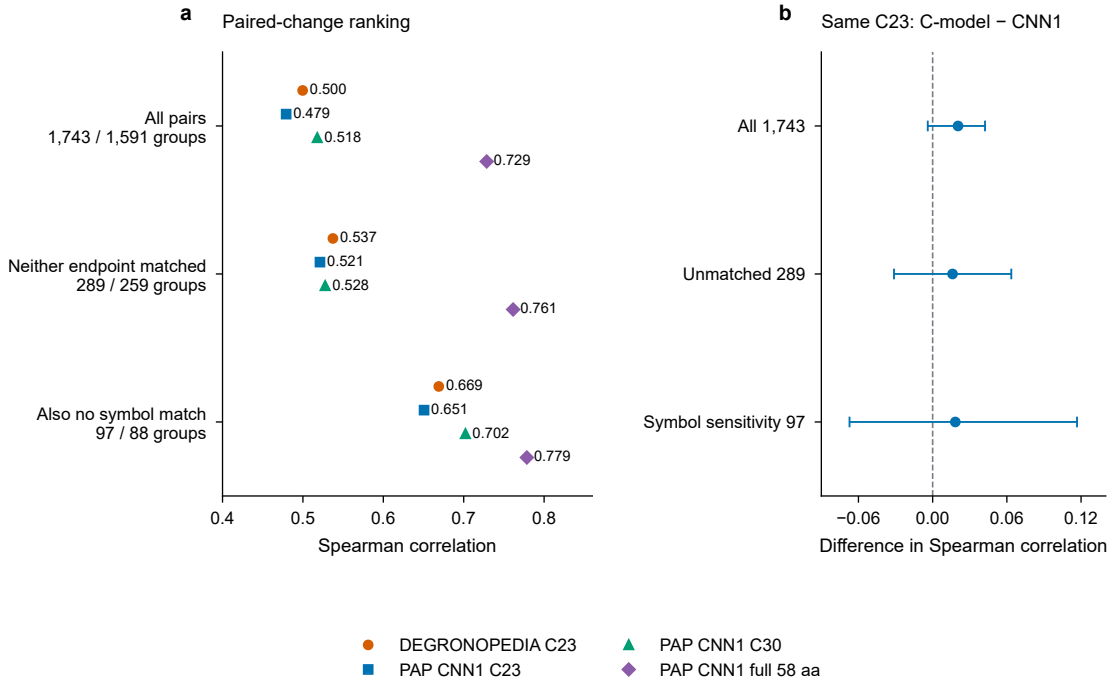


**Fig 5. Overlap with training data and matched sequence inputs change model comparisons.** (a) Paired-effect rank correlations for the C23 terminal model and PAP CNN1 using C23, C30 and full58 on three nested populations. The original 1,743-pair cohort excluded source/PAP training exposure, while 1,454 pairs match an endpoint in the separate C23 development screen. Its actual individual 90/10 split is unavailable. (b) C-model-minus-CNN1 contrasts on identical C23 inputs, with 95% conditional group-bootstrap intervals. The matched-input controls are post hoc; all intervals span zero. The 97-pair available-symbol sensitivity is not alias-complete gene independence. Different cohorts do not identify an exposure effect, and C23/full58 neural inference extrapolates beyond the training length.

## 2.5 Exploratory challenges and candidate follow-up

The main comparisons show that a score's usefulness depends on the assay and task. Two exploratory analyses examine the implications: a separate expression panel tests a further assay change, and sequence matching identifies peptides for experimental follow-up.

The scores failed to predict expression in the separate, selected FLAG panel: primary MP–mCherry-trained T had $\rho = -0.652$ on the strict 20 constructs. Because the panel was selected and used different constructs and an expression-detection assay, this result cannot estimate failure across all microproteins. All model/representation sensitivities, within-study edit analyses and conditional-order controls remain available in S1 Text and Figs S3–S6.

Matching full peptide sequences across studies identified three candidates for follow-up: RP11-620J15.3, JHDM1D-AS1 and the SLCO5A1 uORF peptide. Their existing Leu-to-Ala and append measurements increased reporter output in both source replicates, and the original

study reports interaction-tile annotations for their WT sequences [13, 33]. These matches do not show how either edit affects binding, native protein turnover or peptide-dependent activity. S1 Text and S12–S18 Data retain all 60 candidates, the source-reported positive and negative results, the reassessment of native evidence and the limits imposed by RNA context.

# 3 Discussion

## 3.1 Transfer depends on the experimental question

A microprotein sequence score can perform well in one assay and less well in another. Its usefulness depends on what it was trained to predict, how the peptide was measured and how the score will be used. Composition can distinguish the reference populations studied here and predict reporter accumulation when trained on those measurements. Resemblance to annotated proteins alone does not reliably predict accumulation across assays. Terminal features, calibration and residue order improve prediction in some settings, but no representation performs best in every comparison.

The peptides chosen for evaluation can change the conclusion even when the predictions remain fixed. The pooled terminal-feature advantage for Leu-to-Ala effects disappears under the original source exclusion, whereas the append contrast remains positive [13]. The result from all pairs therefore does not apply equally to the subset defined by the original exclusion rule, and it does not explain the molecular cause of the difference. The intended use of the score also matters: GFP-trained T selects a higher-scoring top decile on canonical tiles despite its slightly lower overall correlation. Ranking unmodified peptides, selecting a subset and predicting an intervention are separate tests.

Guo et al. already measured both reporter orientations, terminal-degron effects and sequence edits, and supplied predictive models [13]. Our contribution is the grouped, peptide-matched prediction comparison and its extension to edit effects, matched inputs and development exposure. Larsen et al. compared 8,645 common nuclear and cytosolic tiles, finding strong overall agreement alongside terminal-specific exceptions, and applied cytosol-trained PAP to the nuclear screen [41]. Context changes therefore need not impair prediction: portability must be measured for the particular construct, population and decision task. The matched-peptide analysis extends these existing reporter and same-sequence context comparisons. Using identical peptides rules out differences in peptide composition between the tested sets, but reporter identity, orientation and library conditions still change together; MP–mCherry also has a variable-length GS linker and an HA tag. The comparison shows how prediction changes across contexts but cannot identify which construct component causes the change. The selected FLAG-expression panel provides another example of failed prediction across assays. Its selection, construct design and expression-detection readout prevent us from generalizing the failure rate to all microproteins [28].

## 3.2 Input coverage and provenance shape predictor comparisons

Comparison with strong published models reduces the apparent benefit of calibration. Calibrating linear PAP on source WT measurements improves its predictions of sequence changes, but uncalibrated nonlinear PAP performs better on the same independent pairs. For the full58 neural model, the interval for the additional calibration gain spans zero. Using the full58 input

rather than C30 is associated with a much larger improvement. Full58 inference extrapolates beyond the training length, and this contrast does not identify a biological cause. Claims about improved calibration therefore need comparisons that account for both model architecture and the sequence supplied to each model.

Experiment-specific nonlinear readouts and separation of measurement from latent biological maps are established methods [21–23]. Interpreting a fitted score as a physical quantity requires independent measurements: Salinas and Ranganathan used affinity-characterized mutants and WT-affinity corrections, while their homologue comparison retained a hypothesized internal-equilibrium mechanism [42]. Joint folding/binding and abundance/activity models have their own physical assumptions and validation targets [24, 25]. Higher-order prediction and domain-extension or ligand maps also use distinct proteins, substitutions and validation settings [43–45]. Our calibration uses source WT reporter measurements, without an independent molecular measurement. It therefore does not convert reporter scores into a physical stability scale that applies across assays.

A test is independent only to the extent that its sequences and related training records have been excluded. The terminal-model audit identifies a smaller set of pairs without known development-screen overlap. Comparing identical C23 inputs also removes most of the apparent performance gap between that model and CNN1. Intervals spanning zero demonstrate neither equivalence nor a universal winner. Higher correlations after exclusions do not establish that exposure caused inflation. Population and decision-metric effects have precedents [26, 46]; our contribution is to compare the models on the same defined peptide sets and provide the records needed to inspect their source and training overlap. Likewise, the small property-feature increment cannot be generalized to all order-aware models. CNN7 retains an order increment on its source study's held-out tiles, while independent paired-effect intervals span zero. Model shuffles test conditional predictive information, not the biological effects of physically shuffling a peptide.

### 3.3 Reporter prediction and native function require different evidence

Sequence descriptors alone do not identify a degradation mechanism. Composition-conditioned motif prioritization, degron scrambles and stabilized full-length products already have experimental precedents [11, 12, 15]. Terminal screens distinguish encoded sequence from exposed ends and pathway dependence from total stabilization [19, 20]. Recognition, ubiquitination and proteasomal engagement are distinct stages [47], potentially affected by accessibility and partner occupancy [8, 48]. Steady-state accumulation also depends on production: under a simple first-order balance, $P_{\mathrm{ss}} = s/k$, where $s$ is production and $k$ includes degradation and dilution. Determining why a particular peptide accumulates therefore requires measurements of both production and turnover.

Annotation, detection and evolutionary origin address further distinct targets [17, 27, 29, 30, 49–51]. A peptide-only reporter test cannot refute a model requiring unavailable native RNA context. HLA presentation, intact-product accumulation and peptide-dependent function also require different evidence [35, 36]; absence from a detection dataset or screen does not establish lack of function.

The three matched candidates suggest experiments; their effects still need to be tested. Synthetic-peptide interactions need not reproduce native stability, concentration or compartmentalization [33], and the edited products were not tested for binding or activity. Characterized MIEF1-MP70 and MRPL33_65 controls have native structural evidence [4], but neither has a

matching primary GFP measurement and edit pair in our linked records. Selected rescue and assembly studies establish functional noncanonical products [3, 5, 52–54], while abundance and functional readouts can diverge [55, 56]. JHDM1D-AS1/DHX15 and RP11-620J15.3 RNA studies do not isolate the nominated peptides [57–59]; the reported DHX15 turnover assay measures DHX15 rather than JHDM1D-AS1 peptide clearance. Tests of RNA architecture, translation and released-peptide activity require controls tailored to the candidate [6, 28, 60].

### 3.4 Practical implications

To predict reporter accumulation, test the score against measurements from the relevant assay and construct. Scores for annotation resemblance or peptide detection answer different questions and cannot be substituted without testing.

When comparing models across studies, specify the sequence given to each model and check for overlap with its training data. Compare strong published models on the same eligible peptides. Report how many peptides receive scores and whether the score is being evaluated for overall ranking or for selecting a subset.

Before using a score to choose sequence edits, test whether it predicts measured changes between WT and edited peptides in the relevant population. Higher reporter output does not by itself show slower native turnover or preserved activity. Those claims require assays that can measure both the original and edited products, together with production, turnover and functional tests.

### 3.5 Limitations and next experiments

Several extensions were retrospective and developed after earlier results were known. We use aggregate assay releases, operational dependence groups and intervals conditional on fitted models, rather than raw-data reconstruction, exhaustive remote-homology control or full retraining. Historical training snapshots are incomplete, and post-hoc intervals are not multiplicity-adjusted discovery tests. Canonical tiles independently test our models trained on the source data, but their PAP-development provenance prevents treating the accompanying PAP scores as independent validation of PAP. The three-row paper/release discrepancy is recorded. The common paired cohort audits source/PAP genes and windows, without certifying independence for every predictor. The terminal-model audit lacks individual split membership, does not resolve every alias or remote homologue, and does not audit N-terminal exposure. Comparisons must retain these coverage and provenance limits. Our literature review focused on the claims made in this paper and cannot establish that no earlier study reached a similar conclusion.

Native validation should first identify the intact or processed product with an edit-compatible measurement and distinguish production from clearance. WT, the individual edits and the unmeasured combined edit/append panel should be tested separately. Binding or activity should be compared across overlapping measured protein amounts, with a predefined equivalence criterion for preservation. RNA, translation and recoded-peptide rescue controls should separate transcript, translation-process and peptide-sequence explanations. A masking hypothesis additionally requires a validated partner perturbation and a clearance measurement. These experiments could show whether the edits change native turnover, binding or activity. The present work identifies candidates for those tests and shows where predictions remain accurate or fail across the assays examined.

# 4 Conclusions

Before using a microprotein sequence score to select peptides or design sequence edits, test whether it predicts the relevant experimental outcome. Composition-based models predict reporter measurements across several settings, but the benefit of terminal features, calibration and residue order depends on the training data, sequence input and assay. Apparent improvements must be compared with strong published models using matched inputs and checks for overlap with training data. Matching peptide sequences across studies also identifies candidates for experiments that distinguish changes in production, turnover, binding and activity. These results show when scores remain useful across the tested assays and when they fail. The effects of the proposed edits on native peptide function remain untested.

# 5 Materials and methods

## 5.1 Analysis hierarchy

The core benchmark evaluates reporter prediction while keeping related peptides in the same validation group. The later analyses test identical peptides across reporter contexts, independent canonical tiles and paired sequence changes. These analyses address the main question of this paper, but they were developed retrospectively after earlier results were known. Population exclusions, input and training-overlap checks, order and calibration sensitivity analyses, the selected FLAG panel and candidate matching are secondary or exploratory analyses. This hierarchy describes their role in the paper; it does not imply preregistration or prospective confirmation. S1 Text reports all outcomes, including failed prediction across contexts.

## 5.2 Controlled catalogue context and clustering reproduction

To check whether catalogue-level sequence differences persist after controlling for length and GC content, the controlled catalogue analysis reconstructed 28,359 TransCODE Phase 2 candidates against hg38, audited current GENCODE v50 coding overlap and retained 27,670 eligible nORFs. Canonical representatives were the longest valid translated isoform per gene (19,563 genes). The protr descriptor definitions were versioned [31]. Default lag/lambda 30 restricted the all-family comparison to proteins of at least 31 residues; no missing long-lag features were imputed. Exact full-protein length matching after dependence filtering yielded 744 comprehensive pairs, 245 Primary pairs and 472 pairs with an additional absolute GC caliper of 0.02. Matched pairs and linked sequence/gene groups stayed in the same evaluation fold.

Clustering used the 1,488 comprehensive matched sequences, all 1,440 descriptors, equal-feature or equal-family weighting, PCA retaining 50 components, $k = 2, 3, 5, 8$, multiple seeds and 80% subsampling. Native sequences were compared with composition-preserving shuffles. A separate computational rerun checked saved PCA memberships and two t-SNE maps. This reproduces the controlled computation; the original cluster memberships could not be recovered. UMAP was not rerun. The original matrix and complete historical settings remain unavailable.

## 5.3 Experimental sources and target definitions

We used the Guo data because they contain reporter measurements, paired reporter contexts and measured sequence edits. This lets us test different uses of a score while retaining exact sequence

and parent links. The Guo supplementary tables supply normalized fusion-reporter scores [13]. The primary development set comprises 1,072 valid wild-type (WT) GFP–microprotein (GFP–MP) cargos; 1,071 have both reported biological replicates. The opposite MP–mCherry orientation supplies 1,046 measured WTs. A separate GFP–MP mutagenesis library contains 1,072 WT baselines, 730 HLA-associated and 71 digest-MS-associated natural sequences, and engineered variants. The additional 801 sequences are absent as exact peptides from the released 1,072-row Guo regression-training table. They remain within the same study, and their availability does not establish prospective blinding of the published models.

We used released means, available biological replicate cells, oligos and source-code metadata. The publisher's article-plus-supplement file subsequently provided full STAR Methods access [13]. The source describes within-bin counts-per-million normalization, five bin weights (0.1, 0.3, 0.5, 0.7, 0.9), their counts-weighted average and replicate-specific min–max scaling. We did not rerun that raw read processing or infer censoring from aggregate scores. Raw read-depth and censoring information were not obtained for this reanalysis. These limitations constrain error and detection interpretation. Throughout, "reporter score" denotes the supplied assay measurement, not an endogenous half-life, thermodynamic folding stability or function.

Exact sequence editing verified 1,014 global Leu-to-Ala pairs and 1,007 extensions by the constant C-terminal sequence DNYNEPKANQ. Observed change was mutant minus WT score in the **same mutagenesis library**. Original and mutagenesis-library WT measurements differ, so the former were not substituted as mutant baselines. Missing outcomes or predictors were not replaced by zero. These all-sequence-valid populations are broader than the source's specified Leu-to-Ala analysis: its STAR Methods exclude WTs containing Leu at either of the last two residues, and Fig. 4F states this restriction for the WT/substitution/extension comparison [13]. After recovering the full methods, we applied that WT-sequence rule to both edit families as a post-hoc sensitivity, retaining all earlier fits and predictions. This retained 798 Leu-to-Ala pairs in 611 operational groups and 802 append pairs in 623 groups. All, retained and omitted strata, mean and both replicate effects, all 20 original predictors and compatible-scale errors are supplied in S12 Data. T-minus-A, O-minus-T and calibrated-minus-raw full-PAP rank contrasts use 1,000 group resamples conditional on the original fits, with no multiplicity adjustment. An independent workbook reader checked all 2,021 membership flags; an independent rank implementation checked 360 point metrics and 18 point contrasts. Bootstrap distributions were not independently replayed. This changes the evaluation population, not the fitted models, and does not reproduce every original assay-processing decision.

The separate Yang assay contains 29 uORFs and eight other ORFs, with native lengths 17–129 residues [28]. We translated all 37 supplied synthesized constructs, verified their reported lengths and terminal DYKDDDDK FLAG tags, and joined measurements by ORF name. Untreated outcomes are the mean of five replicates for the 29 uORFs and four for the other eight. The target is GFP-normalized FLAG-positive cell fraction, not the Guo numerical scale. The source describes the 29 uORFs as randomly selected across PepScores; the eight other ORFs were chosen with high scores. The combined 37-construct set mixes these selection schemes and ORF classes; it is not a population-wide validation sample.

### 5.4 Dependence audit and score provenance

We grouped records linked by identical peptide sequences, a shared parental construct, available gene identities or overlapping catalogue entries. Engineered sequences shared by different par-

ents connected those parents. We exhaustively aligned eligible pairs among native assay peptides, Yang natives, annotation-reference sequences and available ShortStop positive-reference counterparts. Local Smith–Waterman alignment used BLOSUM62, gap-open 10 and gap-extension 1. Primary edges required at least 60% identity over alignment columns and at least 80% coverage of **both** full sequences; a 40% identity graph was a declared sensitivity. Length and composition bounds only removed pairs that could not meet these criteria. Connected components defined split groups. These groups keep known linked records together during validation. They do not rule out distant homology or earlier exposure during pretraining.

The annotation-trained AAC classifier retained the original fixed Extra Trees configuration: 200 trees, maximum depth 10, minimum leaf 10, square-root feature sampling, balanced classes and seed 100. Whole reference pairs and linked reference split groups were removed to a fixed point if connected to any Guo or Yang parent. Seeds 101 and 102 were sensitivity analyses. Higher scores indicate canonical resemblance, opposite to the nORF Score convention [9]; this classifier was not trained on reporter measurements.

ShortStop v1.2.0 (28 October 2026) precomputed scores were joined only after independent GTF translation against hg38 verified exact peptide identity. The primary continuous score was mean SAM probability across released models. Published standard and stringent consensus calls were retained as distinct threshold analyses. Actual supervised-positive identifiers were distinguished from ToBePredicted records, which the inspected training code excludes from supervision. The 1,571 available positive-reference sequences are from the current reviewed release, which does not provide a complete archive of historical training sequences. Thus absence of exact matches does not establish complete training or pretraining separation.

The frozen PAP human30 model was reimplemented from safely parsed published numeric coefficients [14]. Primary transfer uses the final $\min(30, L)$ residues. Composition terms scale residue counts to 30 positions; terminal, pair and triplet terms follow the released positional definition. The full-peptide version uses normalized composition and is evaluated as a secondary analysis; PAP was not refitted. Deposited Guo classification/regression predictions were evaluated only on the additional natural sequences and their strict subset, not against other methods' held-out predictions on Guo's own training rows. Supplied PepScore values were used only for the exact Yang constructs; unavailable transcript, domain and evolutionary inputs were left missing.

### 5.5 Endpoint models, order controls and nested validation

To evaluate predictions on peptides whose related records were absent from training, models used five outer and three inner group folds. Groups were randomly tie-broken with seed 20261031, sorted by descending size and greedily assigned to the smallest fold without using outcomes. Ridge penalties $\{0.1, 1, 10, 100, 1000\}$ were selected by inner mean absolute error (MAE). Standardization and intercept fitting used each training split only. All representations shared the same outer folds. Predictions and fit receipts were saved for every held-out row.

The length model L contains log length and its square. A adds all 20 AAC frequencies. T adds first-five and last-five residue one-hot encodings, whole/N10/C10 hydropathy and PAP composition/terminal components. O adds 40 short-range property-product averages: eight default Moreau–Broto AAindex properties at lags 1–5, standardized across amino acids, then centered and scaled against 199 within-peptide shuffles. Shuffles preserve exact composition

and the first and last five residues. Zero null variance produces a zero residual. This provides features for short cargos without imputing lag-30 values.

Three independent shuffled-order representations add the same 40 standardized features to T using additional shuffled realizations. They test whether residue order adds predictive information after composition and terminal residues are held fixed; the shuffled peptides have no measured outcomes and are not labelled experimental negatives. Raw order features, two extra fold seeds, 40% identity grouping and fixed Extra Trees regression for A/T/O were declared sensitivities. The tree regressors used 200 trees, depth 10, leaf minimum 10 and feature fraction 0.7. No outer-fold winner was substituted for the primary Ridge analysis.

Primary contrasts are O-minus-T rank correlation and T-minus-O MAE, with 1,000 operational-group bootstrap resamples conditional on saved predictions. These intervals keep the fitted models fixed and therefore omit uncertainty from retraining. Additional intervals and subgroup results are descriptive; no secondary significance claims or multiplicity-adjusted discovery list is made. No equivalence test was performed. We report the median absolute difference between biological replicates and half that value to show the scale of replicate variation. Neither is a formal noise floor or a threshold for biological relevance.

### 5.6 Calibration, perturbation prediction and independent transfer

Frozen scores were mapped to reporter scale using isotonic regression fitted to training WTs within each group fold. Direction was determined by training correlation and recorded, including reversed AAC-resemblance maps. Raw score MAE was not compared across incompatible scales. For mutations, each parent's WT and variant were predicted with the **same parent-held-out** model or isotonic map trained on mutagenesis-library WTs. We report raw rank differences as well as calibrated prediction changes. Zero change and an out-of-group training-family mean are error baselines. Replicate-specific outcome changes were analyzed separately.

Secondary mutation-trained Ridge models use mutant-minus-WT changes in A/T/O features and grouped nested evaluation within each family. Opposite-family transfer trains on the other edit family after excluding all held-out parent groups. These tests distinguish using an unmodified-peptide score to predict an edit, training directly on that type of edit, and applying a model trained on a different edit type.

Before scoring Yang outcomes, main predictions and model choices were saved in a validation-release receipt. Primary direct transfer trained on MP–mCherry WTs, whose cargo has a free N terminus and masked C terminus, as an approximate context choice for a C-terminal FLAG construct. Actual FLAG-fusion sequence features were primary; native-only features and GFP–MP training were declared sensitivities. Every training group connected to any Yang construct was removed from these direct-model fits. We did not fit, invert, calibrate or select models on the Yang outcomes. The input representation remains an approximation because neither Guo reporter construct is identical to a FLAG construct. Analyses report all 37, the 27 without exact Guo overlap, and the 20 passing strict separation from all retrieved Guo natural cohorts. Rank correlation and AUROC at the authors' 5% expression threshold are reported; absolute MAE is not reported across assays with different measurement scales.

### 5.7 Exploratory paired reporter-context analysis

To compare contexts on the same peptide population, we joined WT records by both source identifier and exact native amino acid sequence across original GFP–MP (G), original MP–mCherry (M), and mutagenesis-library GFP–MP (U). The complete-mean intersection contains 1,044 peptides in 783 primary dependence groups. Of 1,072 G, 1,046 measured M and 1,072 U records, respectively 28, two and 28 fall outside this intersection. Biological-replicate agreement was computed on available replicate pairs within this same population. The extension was specified after inspecting aggregate paired correlations; it is exploratory and does not supply independent-study validation.

All six directed transfers used the previously defined A and T representations and nested Ridge settings, with identical five-fold assignments across the three contexts. This common-cohort split differs from the earlier 1,072-row development split. Every model excluded the test peptide's entire dependence group across contexts. The target-trained reference used target measurements from the training groups; the source-trained model used source measurements from those groups and was evaluated against the held-out target ranking. For target-scale error, isotonic calibration used source predictions generated by a further group cross-fit within each outer training partition, paired with training-target measurements. Its monotonic direction was learned there. Neither source fitting nor calibration used outer-test target outcomes.

A separate measured-source bridge fitted an isotonic map from paired training measurements and required the test peptide's measured source score; it is not a sequence-only method. Training-target means supplied an MAE baseline. We repeated the sequence analysis with the prior 40% identity graph and split seed 20261032. Descriptive intervals use 1,000 group-bootstrap resamples conditional on saved predictions, including paired contrasts on identical target rows. These analyses do not identify discoveries after adjustment for multiple comparisons.

### 5.8 Independent canonical-peptide transfer extension

To test whether predictions remained useful outside the source study, we specified a comparison with independent canonical peptides after examining the paired-context results but before evaluating the new predictions against the target measurements. We used the author-held-out cytosolic 30-mer library from Voutsinos et al. [14], pinning the repository at commit `9b3eafa6bd9730c558cd018063fa0180edf9db1a`. The released split and original processing log both contain 15,818 test members, whereas the paper states 15,815. We retained all valid released members before exclusions and report this discrepancy. Exact author-training peptide duplicates were absent. Excluding genes shared with the retrieved Guo/annotation inputs and peptide matches at $\geq 60\%$ identity and $\geq 80\%$ coverage of both sequences left 14,143 tiles from 383 ENSG gene groups. The 40% identity sensitivity retained 13,730 tiles. These exclusion rules do not rule out distant homology.

A and T heads were fitted separately to the original 1,072 GFP–MP and 1,046 MP–mCherry WTs, using the benchmark's source-group tuning and training-only scaling. All target predictions and input/model hashes were frozen before joining released abundance scores. Target outcomes did not select the source, sign, penalty or representation. A/T source features were independently reconstructed and matched the saved benchmark exactly. We report Spearman correlation and paired A-minus-T contrasts, with 1,000 gene-cluster bootstrap resamples conditional on fitted predictions. A second source-only tuning seed is a sensitivity analysis; target-scale error was not compared across these different assays.

Secondary diagnostics remove PAP from T or add 80 one-hot columns to A for positions 2–5 (N4), the last four residues (C4), or the four central residues (I4, zero-based slice $L//2-2 : L//2+2$). Equal nominal dimension does not equalize effective complexity or isolate terminal exposure. Source grouped nested-CV results accompany these diagnostics. PAP internal/full scores are target-study-trained held-out references, not equally external competitors; their coefficients also contribute to T's pretrained features.

After initial transfer inspection, a secondary policy comparison reused the earlier benchmark's 10% selection fraction without target optimization. Each frozen score selected its highest-ranked $\lceil 0.1n \rceil$ tiles; ties followed lexical peptide order. We report mean measured assay score among selections and paired policy contrasts, recalculating selection in each gene-bootstrap draw. This is retrospective assay-rank prioritization, not prospective candidate or functional validation (S9 Data).

### 5.9 Frozen published baselines and independent paired changes

After the canonical-tile analysis, we added two secondary challenges with protocols and prediction hashes recorded before their new model–outcome associations. Earlier target results were already known. The original QCDPred code and labeled training tiles were pinned to the authors' repository commit `85979b1b4123b6b5391b617d16551969eda9f56e` [10]. We retained both the mean stability over complete overlapping 17-mers and a whole-cargo composition representation. The latter rescales the sum of residue coefficients by $17/L$ before the published logistic transform; it is a declared alternative representation, not the ordinary 17-mer-window output. No labeled training 17-mer exactly matched the tested windows. This exact-match check does not rule out all homologous sequences. The human25 composition model was retained with first-25 and whole-cargo normalization; its native-tile results are target-study-trained references (S10 Data).

Chu Supplementary Data 1 contains 3,249 matched pairs [15]. Both measured cargos have 58 residues. The RT-derived cargo comprises a retained WT suffix followed by the readthrough sequence; 339 rows retain no WT residues. These are terminal-window replacements, not simple appends to the measured WT58. We required unique gene mapping, standard nonidentical 58/58 sequences, consistent reported sequence/length fields, and at least 60 reads at both endpoints in both replicates. Six Excel-date gene-name artifacts were resolved by transcript/sequence mapping; two inconsistent WT/shared-suffix rows were excluded rather than repaired from outcomes. RT-minus-WT replicate means were checked against all released mean/delta columns. Source gene and sequence exclusions used the preceding 60% identity/80% bilateral-coverage rule, with 40% sensitivity. The common primary cohort additionally excluded PAP-training-linked genes and exact PAP30/human25 training windows from either endpoint. This retained 1,743 pairs in 1,591 operational dependence groups joining gene identity and identical endpoint sequences. The stricter cohort contains 1,738 pairs; the source-head-only sensitivity contains 2,326. Current-transcript terminal matching is an additional sensitivity, not outcome-based membership.

We froze A/T predictions fitted in each original Guo context, linear PAP composition/terminal decompositions, QCDPred and human25. For every frozen baseline, an increasing isotonic map used only the original 1,072 Guo GFP–MP WT outcomes; calibrated changes subtract mapped endpoint scores. No Chu outcome fitted or selected a source, direction, model or calibration. We

retained all 22 original raw/calibrated score variants, clipping and zero mapped effects. Absolute errors on incompatible raw score scales are not compared (S10 Data).

### 5.10 Published neural PAP challenge and conditional order controls

We then challenged these results with published default CNN1 and released kernel-seven CNN7 PAP weights from the pinned Voutsinos release [14]. Other models' outcomes were already inspected, making this a secondary stronger-baseline challenge. The internal channel applies a kernel-one or kernel-seven convolution, global averaging and nonlinear dense layers; a separate nonlinear channel encodes the final five residues. The two channel scores are added. Original model classes and numeric HDF5 weights were used, adapting the Keras backend import and legacy layer loading only. Independent NumPy forward passes verified the preserved graph and weights. CNN1's internal channel is a nonlinear function of composition; CNN7 can retain local order. Differences between separately fitted networks do not isolate order because both fitted channels may differ.

Both final-30-residue and full-cargo inputs were specified and retained. Full58 uses supported variable-length inference, extrapolates beyond the 30-mer training length and is not the ordinary default multi-window CLI aggregate. Every component received its own Guo-GFP-WT-only increasing isotonic map. No target model fitting or score inversion was performed. The authors trained PAP on abundance scores transformed from the sort-seq distribution to the flow-cytometry fluorescence distribution [14]. Here, raw PAP denotes the frozen published output before our additional Guo-WT remapping, rather than an absence of original readout processing. PAP results on its own study's held-out native tiles are labeled target-study-trained references, not equally external competitors.

For each input, two deterministic sets of 19 shuffles preserve composition and the first/final five residues (seeds 20261105 and 20261106). CNN1 native and shuffle-mean scores are analytically identical, checked on fixed samples; CNN7 shuffles were explicitly evaluated. Native-minus-shuffle rank contrasts test information in the prediction conditional on these controls, not experimental outcomes of shuffled peptides. All 40 neural raw/calibrated variants were frozen before new associations. Independent verification checked 62,962 model-input scores, 15,504 sampled shuffled scores, 20 threshold maps and all 340 native/paired metric rows. Float32 calibration outputs were checked against float64 interpolation at the appropriate numerical precision. Intervals use 1,000 paired operational- group bootstrap resamples conditional on fitted predictions, without retraining uncertainty or multiplicity-adjusted discovery. Additional full58-versus-C30 and neural-versus-calibrated-linear intervals are explicitly post hoc, after inspecting point estimates (S10 Data).

### 5.11 Terminal-model exposure audit and matched-input challenge

We next evaluated the released DEGRONOPEDIA C-terminal model [18]. This secondary retrospective challenge followed inspection of other predictors; its protocol preceded the new terminal-model associations (S11 Data). Source and all three released model weights were pinned to commit `2703570145c3c53cf276b705e1a6cc59083dfd12`. We used the original descriptor and native CatBoost inference functions with all 2,997 named model features. The source's unused feature-selection regular expression was not applied. Current dependencies reproduced all 12 rounded author examples; the authors did not specify dependency versions, so we cannot verify

exact agreement with their historical runtime. Only the C-terminal model was applied to the new target.

We recovered the 22,564-row Koren C23 screen through its primary redistribution in Hickey et al. Supplementary Table 5A, whose notes attribute columns A–L to Koren Table S7A [19,32]. All rows contain standard 23-mers and numeric PSI. Exact final-23-residue matches were checked separately for both endpoints of every existing 1,743-pair common-cohort member. A match establishes development-screen exposure, not membership of the unavailable individual 90/10 training split. We retained all pairs, either-endpoint matches, neither-endpoint matches, and a further sensitivity excluding available exact gene-symbol matches. Unresolved aliases and 24 original Excel-date gene-name artifacts prevent certified gene independence. Existing pair eligibility and gene/identical-endpoint groups were unchanged.

We froze scores for 3,302 unique C23 inputs before evaluating terminal-model associations. No retraining, target calibration, direction inversion or target-driven model choice was performed. Higher predicted PSI retained its published stability direction. WT, RT and RT-minus-WT rank correlations, both replicate changes and all frozen comparator scores were retained on every exposure stratum. An independent workbook reader verified every join; 32 original-descriptor preparations reproduced frozen scores exactly.

After inspecting the first terminal-model results, we added explicitly post-hoc PAP CNN1/CNN7 controls on the same C23 inputs. The original C30 and full58 scores remained unchanged. C23 and full58 use supported variable-length inference outside the training 30-mer length; they are explicit input representations rather than the default multi-window CLI output. No new calibration or sign choice was made. Independent NumPy forward passes checked all 6,604 new model-input vectors. All 48 initial and 16 added metric rows were independently recalculated. Contrasts use 1,000 paired operational-group resamples conditional on frozen predictions; they exclude retraining uncertainty and multiplicity-adjusted discovery. Cohort exclusions change the evaluated population and cannot identify a causal effect of development-screen exposure.

### 5.12 Exploratory exact-sequence link to interaction candidates

We linked the 60 distinct full peptide sequences in Sandmann et al. Supplementary Table S2 to all retrieved natural Guo assay records and existing edit pairs [13, 33]. Only whitespace and one terminal stop were normalized. Candidate names and Ribo-seq identifiers were retained as annotations, but records were joined only when their full peptide sequences matched. All 60 candidates, including unmatched records, were retained. Original source Supplementary Table S3 supplied synthetic binding tiles, domain/SLiM annotations, tagged-expression and phosphorylation records. Exact tile occurrences were mapped within each matched WT, and every available edit was annotated for residue changes inside those tiles. A source-associated motif or domain does not establish a causal residue or kinase–substrate relationship. An independent workbook reader checked all 60 candidates, 74 natural-record links, 32 measured edit pairs and 10 tile/edit annotations. This post-hoc descriptive link involved no new fitting, enrichment test, functional-negative labels or edited-binding measurement (S12 Data).

### 5.13 Candidate-specific native evidence and later MS reassessment

We compared all 60 selected candidates with the original human-heart lncORF and uORF tables by exact full-peptide sequence [34]. Reported in-vitro translation, shotgun peptides, SRM targets and positive technical runs were retained as separate source fields; missing coverage and targeted nondetection were not assigned functional labels. An independent reader checked 2,674 source cells and 89 positive SRM run records across the matched panel. For JHDM1D-AS1, the public RefSeq transcript sequence was translated independently to check compatibility with the selected peptide; this did not verify historical plasmid sequence or full isoform identity (S13 Data).

We then matched candidates exactly to full sequences in Deutsch et al. Supplementary Table 9 before using their ORF identifiers to retrieve Tables 3, 6, 7 and 12 [35]. Source-reported final tiers, peptide sharing and missing entries were retained without new tier assignments. For the JHDM1D-AS1 shotgun claim, original ORF, peptide and Universal Spectrum Identifier (USI) linked the community benchmark's full PSM table to its curator-rating table [36]. Its Fig 1c label correction was checked [37]; no study-level plot was used to assign candidate confidence. An independent pandas replay checked all 60 candidates and 4,143 cells from the later tables. These checks use published tables; they do not reinterpret spectra, reprocess raw data or independently repeat the experiments (S14 Data). For a further native-context comparison, all 60 candidates were matched by exact full sequence to brain-atlas Tables S1, S2, S3 and S5 [38]. An independent worksheet-XML reader checked all retained source rows and 407 matched cells, ignoring erroneous declared sheet dimensions without editing the workbook. No score association or new function label was assigned (S15 Data). Additional literature annotations preserve source-specific growth-screen flags and exact product identities from the Shepelev preprint and Schlesinger et al. [6, 39]. All 60 frozen sequences were also compared with the encoded polymer sequences in PDB 5OOL [4]; an independent reader checked the sequence joins and modeled residues. Partner-perturbation enrichment, native structure, peptide detection and cellular activity remain separate evidence fields. S17 Data records these annotations, selected original-source checks and the explicit limits of the scientific reading. Literature validation examples were separated by construct, measurement and intervention rather than pooled as functional labels. S18 Data records primary-source scope for evolutionary age, engineered terminal exposure and endogenous variant readouts. These annotations introduce no new candidate phenotypes or predictor outcomes. The public supporting-data map locates the derived tables, scripts and recorded verification scope for each S1–S18 Data label; original source files are obtained from their providers.

## Generative AI disclosure

OpenAI Codex agents contributed substantially to source discovery, literature synthesis, and mathematical exposition.

## Data and Code Availability

The study data, code, manuscript and appendix are publicly available at `https://github.com/Kohze/microprotein-score-transfer`. The frozen data and analysis release is v0.1.1 (commit `1a22e80f60b46fefa8f370100b2e322457f3f518`). The appendix, S1 Text (PDF), contains supplementary methods, detailed results, figures and tables. The repository README links di-

rectly to the manuscript, appendix, data map and release downloads. The release includes study scripts, frozen derived outcomes and predictions, split and related-sequence records, artwork and source provenance. `SUPPORTING_DATA_MAP.csv` resolves all S1–S18 Data labels to actual public paths. The offline verifier recomputes point metrics and checks file hashes; it does not refit models, recompute bootstrap intervals or reconstruct raw source inputs. Original papers, workbooks, third-party predictor implementations, weights and runtime libraries are obtained from the pinned upstream sources. Original study code is released under the MIT licence; this licence does not cover third-party data, models or publications.

## Funding

Robin Gounder was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training Partnership studentship at the University of Cambridge (studentship reference 2116112).

## Author contributions

Robin Gounder: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing—original draft.

Samuel Schmidt: writing—review and editing.

## Competing interests

The authors declare no competing interests. Samuel Schmidt's affiliation with Vaionex Corporation is disclosed in the author information.

## Research scope

No new participant, animal or laboratory experiments were performed.

## Supporting information

The appendix is available as S1 Text (PDF) in the public repository. S1 Text contains the extended benchmark specifications, complete detailed results, exploratory diagnostics and candidate-source audits, six supporting figures, two supporting tables and the S1–S18 Data guide. All adverse outcomes and sensitivity analyses remain available there.

## References

[1] Mudge JM et al. Standardized annotation of translated open reading frames. Nature Biotechnology. 2022. doi:10.1038/s41587-022-01369-0.

[2] Chothani S et al. An expanded reference catalog of translated open reading frames for biomedical research. Nucleic Acids Research. 2026;54:gkag234. doi:10.1093/nar/gkag234.

[3] Samandi S et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. eLife. 2017;6:e27860. doi:10.7554/eLife.27860.

[4] Brown A, Rathore S, Kimanius D, Aibara S, Bai XC, Rorbach J, Amunts A, Ramakrishnan V. Structures of the human mitochondrial ribosome in native states of assembly. Nature Structural and Molecular Biology. 2017;24:866–869. doi:10.1038/nsmb.3464.

[5] Rathore A et al. MIEF1 Microprotein Regulates Mitochondrial Translation. Biochemistry. 2018;57:5564–5575. doi:10.1021/acs.biochem.8b00726.

[6] Shepelev NM et al. Translation of numerous upstream open reading frames rather than their products is essential for proliferation of human cells of distinct origin. bioRxiv [Preprint]. 2025 July 7;version 1. doi:10.1101/2025.07.07.663426.

[7] Ghosh A et al. Suppressive cancer nonstop extension mutations increase C-terminal hydrophobicity and disrupt evolutionarily conserved amino acid patterns. Nature Communications. 2024;15:9209. doi:10.1038/s41467-024-52779-4.

[8] Hasenjäger S, Bologna A, Essen LO, Spadaccini R, Taxis C. C-terminal sequence stability profiling in Saccharomyces cerevisiae reveals protective protein quality control pathways. Journal of Biological Chemistry. 2023;299:105166. doi:10.1016/j.jbc.2023.105166.

[9] Gounder R. The Elusive Dynamic of the Genome: Novel Sequence Patterns, Splicing and Regulation. PhD thesis, University of Cambridge. 2025. Chapter 2 and Appendix A. doi:10.17863/CAM.129295.

[10] Johansson KE, Mashahreh B, Hartmann-Petersen R, Ravid T, Lindorff-Larsen K. Prediction of quality-control degradation signals in yeast proteins. Journal of Molecular Biology. 2023;435:167915. doi:10.1016/j.jmb.2022.167915.

[11] Mashahreh B et al. Conserved degronome features governing quality control associated proteolysis. Nature Communications. 2022;13:7588. doi:10.1038/s41467-022-35298-y.

[12] Zhang Z et al. Elucidation of E3 ubiquitin ligase specificity through proteome-wide internal degron mapping. Molecular Cell. 2023;83:3377–3392.e6. doi:10.1016/j.molcel.2023.08.022.

[13] Guo Y et al. Intrinsic bias of the genetic code shapes the folding and stability landscapes of microproteins. Molecular Cell. 2026;86:2173–2190.e11. doi:10.1016/j.molcel.2026.04.021. Data: doi:10.17632/jjxy3f462r.1.

[14] Voutsinos V et al. A near-complete map of human cytosolic degrons and their relevance for disease. Science Advances. 2026;12:eadz3483. doi:10.1126/sciadv.adz3483.

[15] Chu CY et al. Protein C-terminal variations impact proteostasis. Nature Communications. 2026;17:2288. doi:10.1038/s41467-026-68979-z.

[16] Casola C, Owoyemi A, Vakirlis N. Degradation determinants are abundant in human non-canonical proteins and minor annotated isoforms. Genome Biology and Evolution. 2026. doi:10.1093/gbe/evag009.

[17] Roginski P, Papadopoulos C, Herman S, Baumann A, Grislain A, Lopes A. Impact of GC content on de novo gene birth. Nature Communications. 2026;17:1268. doi:10.1038/s41467-025-68022-7.

[18] Szulc NA et al. DEGRONOPEDIA: a web server for proteome-wide inspection of degrons. Nucleic Acids Research. 2024. doi:10.1093/nar/gkae238.

[19] Koren I et al. The eukaryotic proteome is shaped by E3 ubiquitin ligases targeting C-terminal degrons. Cell. 2018;173:1622–1635.e14. doi:10.1016/j.cell.2018.04.028.

[20] Timms RT et al. A glycine-specific N-degron pathway mediates the quality control of protein N-myristoylation. Science. 2019;365:eaaw4912. doi:10.1126/science.aaw4912.

[21] Schulze TK et al. Supervised learning of protein variant effects across large-scale mutagenesis datasets. Protein Science. 2026. doi:10.1002/pro.70526.

[22] Tareen A, Kooshkbaghi M, Posfai A, Ireland WT, McCandlish DM, Kinney JB. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. Genome Biology. 2022;23:98. doi:10.1186/s13059-022-02661-7.

[23] Faure AJ, Lehner B. MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. Genome Biology. 2024;25:303. doi:10.1186/s13059-024-03444-y.

[24] Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604:175–183. doi:10.1038/s41586-022-04586-4.

[25] LeBlanc CJ et al. Interpretable biophysical neural networks of transcriptional activation domains separate roles of protein abundance and coactivator binding. Cell Systems. 2026;17:101701. doi:10.1016/j.cels.2026.101701.

[26] Johansson KE, Lindorff-Larsen K, Winther JR. Global analysis of multi-mutants to improve protein function. Journal of Molecular Biology. 2023;435:168034. doi:10.1016/j.jmb.2023.168034.

[27] Miller B et al. ShortStop: a machine learning framework for microprotein discovery. BMC Methods. 2025;2:16. doi:10.1186/s44330-025-00037-4. Precomputed release v1.2.0, 28 October 2026: `https://github.com/brendan-miller-salk/ShortStop/releases/tag/v1.2.0`.

[28] Yang H, Li Q, Stroup EK, Wang S, Ji Z. Widespread stable noncanonical peptides identified by integrated analyses of ribosome profiling and ORF features. Nature Communications. 2024;15:1932. doi:10.1038/s41467-024-46240-9.

[29] He J, Sui J, Chen Q, Hu H, Tan CSH. sORF-Trans2MS: a two-module deep learning framework for sORF translation and MS-supported microprotein prediction. bioRxiv. 2026. Version 1; preprint. doi:10.64898/2026.09.29.755517.

[30] Fierro-Monti I. Translation uniformity, evolutionary conservation, and peptide properties associate with microprotein mass spectrometry detection. Research Square. 2026. Version 1; preprint. doi:10.21203/rs.3.rs-9645389/v1.

[31] Xiao N, Cao DS, Zhu MF, Xu QS. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 2015;31:1857–1859. doi:10.1093/bioinformatics/btv042.

[32] Hickey CM, Breckel C, Zhang M, Theune WC, Hochstrasser M. Protein quality control degron-containing substrates are differentially targeted in the cytoplasm and nucleus by ubiquitin ligases. Genetics. 2021;217:iyaa031. doi:10.1093/genetics/iyaa031. Supplementary Table 5: doi:10.25386/genetics.13273286.

[33] Sandmann CL et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Molecular Cell. 2023;83:994–1011.e18. doi:10.1016/j.molcel.2023.01.023.

[34] van Heesch S et al. The translational landscape of the human heart. Cell. 2019;178:242–260.e29. doi:10.1016/j.cell.2019.05.010.

[35] Deutsch EW et al. Expanding the human proteome with microproteins and peptideins. Nature. 2026;654:813–825. doi:10.1038/s41586-026-10459-x.

[36] Wacholder A et al. Community benchmarking and evaluation of human unannotated microprotein detection by mass spectrometry based proteomics. Nature Communications. 2026;17:1241. doi:10.1038/s41467-025-68002-x.

[37] Wacholder A et al. Author Correction: Community benchmarking and evaluation of human unannotated microprotein detection by mass spectrometry based proteomics. Nature Communications. 2026;17:4882. doi:10.1038/s41467-026-73431-3.

[38] Miller B et al. A microprotein atlas of the human frontal cortex in Alzheimer's disease. Nature Aging. 2026;6:1979–1997. doi:10.1038/s43587-026-01207-x. Source repository: `https://github.com/brendan-miller-salk/brain-microprotein-atlas`.

[39] Schlesinger D et al. A large-scale sORF screen identifies putative microproteins involved in cancer cell fitness. iScience. 2025;28:111884. doi:10.1016/j.isci.2025.111884.

[40] Kondo A et al. Long noncoding RNA JHDM1D-AS1 promotes tumor growth by regulating angiogenesis in response to nutrient starvation. Molecular and Cellular Biology. 2017;37:e00125-17. doi:10.1128/MCB.00125-17.

[41] Larsen FB et al. Comprehensive degron mapping in human transcription factors. Science Advances. 2026;12(45):eaee1766. doi:10.1126/sciadv.aee1766.

[42] Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. eLife. 2018;7:e34300. doi:10.7554/eLife.34300.

[43] Faure AJ, Martí-Aranda A, Hidalgo-Carcedo C, Beltran A, Schmiedel JM, Lehner B. The genetic architecture of protein stability. Nature. 2024;634:995–1003. doi:10.1038/s41586-024-07966-0.

[44] Martí-Aranda A, Lehner B. Seven complete comparative maps of allosteric mutations in a protein family. Nature Communications. 2026;17:5487. doi:10.1038/s41467-026-71005-x.

[45] Hidalgo-Carcedo C, Faure AJ, Martí-Aranda A, Zarin T, Lehner B. Allosteric and energetic remodeling of a PDZ domain by protein domain extensions. Nature Communications. 2026;17:2934. doi:10.1038/s41467-026-69673-w.

[46] Fawzy M, Marsh JA. Understanding the heterogeneous performance of variant effect predictors across human protein-coding genes. Scientific Reports. 2024;14:26114. doi:10.1038/s41598-024-76202-6.

[47] Guharoy M et al. Tripartite degrons confer diversity and specificity on regulated protein degradation in the ubiquitin-proteasome system. Nature Communications. 2016;7:10239. doi:10.1038/ncomms10239.

[48] Guharoy M, Lazar T, Macossay-Castillo M, Tompa P. Degron masking outlines degronons, co-degrading functional modules in the proteome. Communications Biology. 2022;5:445. doi:10.1038/s42003-022-03391-z.

[49] Middendorf L, Ravi Iyengar B, Eicholt LA. Sequence, Structure, and Functional Space of Drosophila De Novo Proteins. Genome Biology and Evolution. 2024;16(8):evae176. doi:10.1093/gbe/evae176.

[50] Ruiz-Orera J et al. Evolution of translational control and the emergence of genes and open reading frames in human and non-human primate hearts. Nature Cardiovascular Research. 2024;3:1217. doi:10.1038/s44161-024-00544-7.

[51] Vakirlis N, Vance Z, Duggan KM, McLysaght A. De novo birth of functional microproteins in the human lineage. Cell Reports. 2022;41:111808. doi:10.1016/j.celrep.2022.111808.

[52] Chen J et al. Pervasive functional translation of non-canonical human open reading frames. Science. 2020;367:1140–1146. doi:10.1126/science.aay0262.

[53] Prensner JR et al. Non-canonical open reading frames encode functional proteins essential for cancer cell survival. Nature Biotechnology. 2021;39:697–704. doi:10.1038/s41587-020-00806-2.

[54] Hofman DA et al. Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. Molecular Cell. 2024;84:261–276.e18. doi:10.1016/j.molcel.2023.12.003.

[55] Salghetti SE, Muratani M, Wijnen H, Futcher B, Tansey WP. Functional overlap of sequences that activate transcription and signal ubiquitin-mediated proteolysis. Proceedings of the National Academy of Sciences USA. 2000;97:3118–3123. doi:10.1073/pnas.97.7.3118.

[56] Matreyek KA et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature Genetics. 2018;50:874–882. doi:10.1038/s41588-018-0122-z.

[57] Yao G et al. Long non-coding RNA JHDM1D-AS1 interacts with DHX15 protein to enhance non-small-cell lung cancer growth and metastasis. Molecular Therapy Nucleic Acids. 2019;18:831–840. doi:10.1016/j.omtn.2019.09.028.

[58] Malmström E et al. The long non-coding antisense RNA JHDM1D-AS1 regulates inflammatory responses in human monocytes. Frontiers in Cellular and Infection Microbiology. 2022;12:934313. doi:10.3389/fcimb.2022.934313.

[59] Liu C et al. LncRNA RP11-620J15.3 promotes HCC cell proliferation and metastasis by targeting miR-326/GPI to enhance glycolysis. Biology Direct. 2023;18:15. doi:10.1186/s13062-023-00370-0.

[60] Devaux A et al. Identification of microprotein-coding intronic polyadenylation isoforms and function in genotoxic anticancer drug response. Genome Biology. 2025;26:366. doi:10.1186/s13059-025-03829-7.

[61] Gounder R, Schmidt S. Guide selection transfers depletion effects across noncanonical ORF screens. Public research manuscript and data/code resource. 2026. `https://github.com/Kohze/norf-guide-selection`. Accessed 10 November 2026.