A MULTI-AXIS OUT-OF-DISTRIBUTION BENCHMARK FOR PROTEIN UNDERSTANDING TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein representation learning has advanced rapidly, driven by large-scale language models and structure encoders. However, most existing benchmarks evaluate models under independent and identically distributed (IID) settings or rely on a single out-of-distribution (OOD) axis, such as sequence identity. In real-world applications, test proteins frequently diverge from the training distribution along multiple biological dimensions—sequence, structure and function—yet current benchmarks do not systematically measure model robustness under these diverse shifts. To address this gap, we introduce **POOR** (**P**rotein **O**ut-**O**f-distribution **R**obustness), a multi-axis OOD benchmark for protein understanding that spans 10 protein tasks, each annotated with 24–28 OOD scenario flags covering sequence/evolutionary OOD, structural OOD, and functional OOD. We further propose **ProDPR** (**Pro**tein **D**istributional **P**erformance **R**etention), a unified OOD evaluation metric that enables comparison of model generalization difficulty across both single-axis OOD scenarios and their high-dimensional combinations. Using this framework, we benchmark protein language models at multiple scales and structure-aware models, and further examine the effects of different fine-tuning regimes. We find that stronger in-distribution performance does not necessarily imply greater OOD robustness: model scale and structural inductive bias confer distinct, shift-dependent advantages, while fused multi-axis shifts expose substantially larger failures than conventional single-axis evaluation.

1 INTRODUCTION

Proteins are central to nearly all biological processes, and accurately predicting their structures, functions, interactions, and biophysical properties remains a fundamental goal in computational biology (Alberts et al., 2014; Brändén & Tooze, 1999). Recent advances in protein representation learning have substantially improved predictive performance across a wide range of downstream tasks. Large protein language models, such as ESM-2 (Lin et al., 2023), ProtTrans (Elnaggar et al., 2022), and ProtGPT2 (Ferruz et al., 2022), learn rich evolutionary representations from large-scale sequence corpora, while structure-based and multimodal models, such as GearNet (Zhang et al., 2023b), SaProt (Su et al., 2024), and ESM-3 (Hayes et al., 2025), further incorporate three-dimensional or multimodal protein information. These advances have led to strong performance on standard benchmarks for protein structure, function, interactions, and molecular properties.

Despite this progress, an important question remains: *how reliably do these models generalize when test proteins differ substantially from those encountered during training?* Real-world protein prediction is inherently non-IID. A model trained primarily on well-characterized proteins may be required to annotate a newly sequenced protein with no close homologs, recognize the function of a structurally novel protein, or make predictions for unusually long, disordered, or functionally rare proteins. Existing evaluation protocols often characterize generalization using either conventional dataset splits or a single predefined notion of novelty, such as sequence identity, family membership, or temporal separation (Xu, 2019; Zhou et al., 2019). While such protocols provide valuable measurements of specific forms of extrapolation, they capture only a partial view of the distribution shifts encountered in protein modeling.

A more faithful evaluation therefore requires a **multi-axis perspective on protein OOD generalization**. We characterize protein distribution shift along three biologically distinct axes. **Se-**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

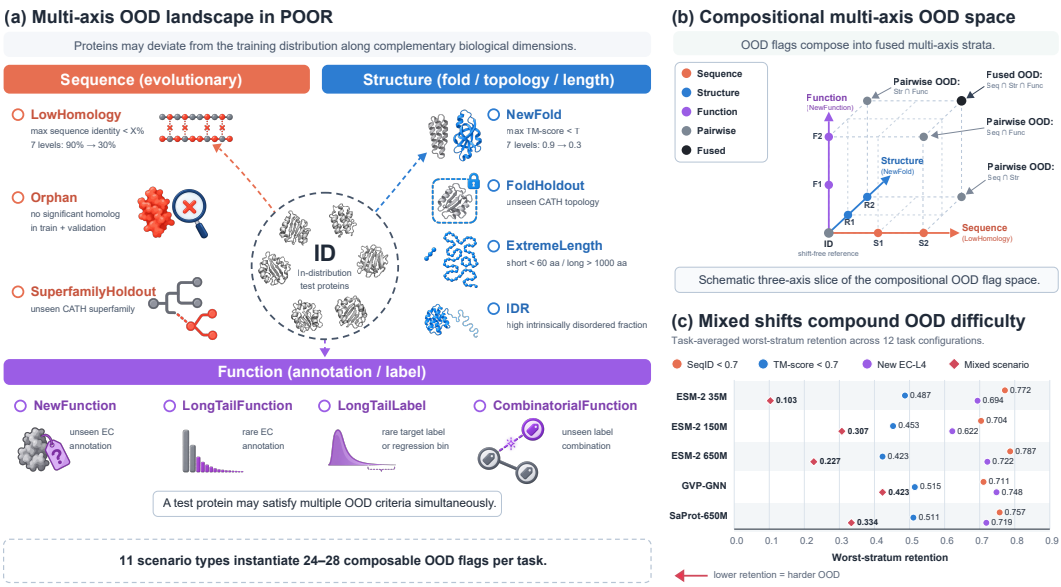


Figure 1: **Overview of the POOR benchmark.** (a) POOR characterizes 11 OOD scenario types across three complementary biological dimensions—sequence, structure, and function—while allowing a test protein to satisfy multiple OOD criteria. (b) Single-axis shifts can be composed into pairwise and multi-axis fused OOD scenarios with graded severity. (c) Task-averaged worst-stratum retention across the twelve task configurations shows that fused distribution shifts compound generalization difficulty far beyond the hardest single axis, for protein language models and graph neural networks alike.

quence/evolutionary OOD captures increasing evolutionary distance from the training distribution, ranging from reduced sequence identity to the absence of detectable homologs and held-out evolutionary families. **Structural OOD** captures novelty in protein geometry and structural organization, including low structural similarity, unseen folds, extreme sequence lengths, and intrinsically disordered proteins. **Functional OOD** captures shifts in annotation and functional space, including unseen enzyme functions, rare labels, and previously unobserved label combinations. Importantly, these axes are neither equivalent nor mutually exclusive: sequence similarity does not uniquely determine structural or functional similarity, and a single protein may simultaneously exhibit novelty along multiple biological dimensions. Consequently, evaluating each shift in isolation may substantially underestimate the difficulty of realistic OOD generalization. Figure 1 illustrates this multi-axis view and how POOR evaluates both individual and intersecting distribution shifts.

In this work, we present **POOR**, a comprehensive benchmark for evaluating OOD generalization across **10 protein understanding tasks**, spanning protein structure, function, interactions, binding, and physicochemical properties. For each task, POOR annotates test examples with **24–28 biologically grounded OOD scenario flags** spanning sequence/evolutionary, structural, and functional shifts. Crucially, these scenarios are represented as non-exclusive flags rather than mutually exclusive test partitions, so the same protein can exhibit multiple forms of novelty simultaneously. This design enables POOR to evaluate not only isolated OOD factors but also their naturally intersecting, high-dimensional combinations.

Beyond benchmark construction, we introduce **ProDPR**, a unified metric for quantifying OOD robustness across heterogeneous protein tasks and distribution shifts. ProDPR measures the predictive utility retained under a given OOD shift relative to a shift-free (in-distribution) reference, enabling consistent comparison across both individual OOD axes and their high-dimensional combinations. Building on POOR and ProDPR, we systematically evaluate protein language models across different scales, structure-aware models, graph neural networks, and different fine-tuning regimes to characterize how model capacity, input modality, and architectural inductive biases affect OOD generalization. Our experiments show that failure modes are axis-dependent: structural novelty is the dominant failure axis for sequence-based function annotation (F_{max} drops of 70–81% relative to

InD at TM<0.7). Fused multi-axis shifts are consistently the hardest condition, with worst-stratum retention as low as 0.103. Moreover, predictive performance does not determine robustness—on GO-CC, SaProt-650M achieves higher F_{max} than GVP-GNN (0.607 vs. 0.457) yet retains less of its shift-free utility (0.505 vs. 0.680). Full fine-tuning nearly memorizes the training set without improving OOD retention, whereas a linear probe on frozen backbones benefits more from scale. Together, these results reveal robustness trade-offs that are largely obscured by conventional IID or single-axis evaluation, establishing POOR and ProDPR as a unified framework for benchmarking, diagnosing, and ultimately improving the OOD generalization of protein models.

2 Related work

Protein benchmarks and datasets

Protein benchmarks have played a central role in representation learning. TAPE (Rao et al., 2019) introduced a transfer-learning suite spanning multiple sequence-level tasks, while PEER (Xu et al., 2022) expanded evaluation to a broader collection of protein understanding problems, largely under random IID splits. FLIP (Dallago et al., 2021) and ProteinGym (Notin et al., 2023) focus on mutation effects and fitness landscapes, whereas ATOM3D (Townshend et al., 2020), ProteinShake (Kucera et al., 2023), and ProteinWorkshop (Jamasb et al., 2024) extend benchmarking to structure-based tasks. Although some of these suites provide biologically motivated or similarity-aware splits for selected tasks, their non-IID protocols remain task-specific and are not designed for systematic comparison across OOD conditions.

Individual tasks are further supported by specialized resources. Structural classification and remote homology rely on SCOP/SCOPe (Murzin et al., 1995; Fox et al., 2014) and CATH (Sillitoe et al., 2021); function annotation is organized around Gene Ontology and CAFA (Ashburner et al., 2000; Radivojac et al., 2013; Zhou et al., 2019); protein–ligand prediction commonly uses PDBbind (Wang et al., 2004); protein–protein interaction and interface modeling use resources such as SKEMPI and PINDER (Moal & Fernández-Recio, 2012; Kovtun et al., 2024); and enzyme-property prediction builds on BRENDA and derived datasets such as TurNuP and CatPred-DB (Chang et al., 2021; Kroll et al., 2023; Boorla & Maranas, 2025). These resources define the tasks and label spaces underlying POOR, but their evaluation protocols are typically based on random splits or a single task-specific non-IID split. POOR instead re-organizes established tasks under explicit, graded, and composable OOD scenarios.

OOD evaluation in protein understanding

OOD evaluation has increasingly appeared in protein understanding, but existing protocols are usually developed for individual tasks. Homology-aware splits are used in function prediction (Gligorijević et al., 2021; Yu et al., 2023), enzyme modeling (Kroll et al., 2023; Boorla & Maranas, 2025), and interaction prediction (Bernett et al., 2024); remote-homology benchmarks use SCOP/CATH hierarchy holdouts (Hou et al., 2018; Rao et al., 2019); mutation-effect benchmarks extrapolate across mutational regimes (Dallago et al., 2021; Hsu et al., 2022a); and interaction models adopt cold-target or both-cold settings (Pahikkala et al., 2015; Huang et al., 2021). These studies consistently show that IID evaluation can overestimate generalization, but their OOD definitions are typically binary, threshold-based, and task-dependent.

Existing protocols can be broadly grouped into three biological OOD dimensions. **Sequence and evolutionary OOD** is most commonly defined through sequence-identity thresholds, homology-aware partitioning, or family/superfamily holdouts (Gligorijević et al., 2021; Nallapareddy et al., 2023; Teufel et al., 2023; Hou et al., 2018). Protein language models also exhibit substantial degradation at low sequence identity (Kabir et al., 2023), while SPECTRA (Ektefaie et al., 2024) evaluates generalization continuously over train–test sequence overlap. **Structural OOD** evaluates generalization to novel folds or geometries, commonly through CATH topology holdouts or structural-similarity filtering in inverse folding and related tasks (Ingraham et al., 2019; Dauparas et al., 2022; Hsu et al., 2022b; Zhang et al., 2023a). Fold-level structural holdouts have likewise been used to evaluate function prediction (Gligorijević et al., 2021). **Functional OOD** captures novelty in annotations, labels, or interaction partners. CAFA evaluates proteins with limited prior knowledge (Radivojac et al., 2013), zero-shot function prediction considers GO terms unseen during training (Kul-

manov & Hoehndorf, 2022; Kulmanov et al., 2024), and interaction models use cold-partner settings to test unseen proteins or compounds (Pahikkala et al., 2015; Huang et al., 2021). Other forms of distribution shift include temporal evaluation, as used in CAFA and time-based docking or binding benchmarks (Zhou et al., 2019; Stärk et al., 2022; Buttenschoen et al., 2024). Related efforts such as DrugOOD (Ji et al., 2023) and PLINDER (Durairaj et al., 2024) introduce stronger OOD or leakage-control protocols for molecular affinity and protein–ligand docking, but remain centered on specific task families.

Overall, existing protein OOD evaluations are largely single-task and single-axis, with heterogeneous definitions that are difficult to compare or compose. Real test proteins, however, can simultaneously deviate from the training distribution in sequence, structure, and function.

3 THE OUT-OF-DISTRIBUTION BENCHMARK AND METRIC

POOR comprises ten protein understanding tasks (twelve task configurations) spanning protein structure, function, interactions, binding, and physicochemical properties; task definitions, data sources, and split construction protocols are detailed in Appendix B. This section defines POOR's OOD scenario annotations, quantifies their pairwise overlap, and introduces the unified retention metric used throughout our evaluation.

3.1 OOD SCENARIO DESIGN

POOR evaluates robustness along three biologically distinct families of distributional shift: sequence/evolutionary, structural, and functional novelty. Rather than assigning each test protein to a single partition, POOR annotates it independently on every scenario, so that the same protein can exhibit several forms of novelty at once and their combinations can be evaluated jointly. Within each family, shift severity is graded rather than binary: similarity-based scenarios compare each test protein against the union of the training and validation sets at up to seven thresholds, and are complemented by discrete holdouts of entire evolutionary or structural groups. Table 1 defines every scenario, and Appendix C reports its per-task coverage.

Table 1: **OOD scenario definitions.** All scenarios are boolean flags annotated on the test set. Each task carries 24–28 scenario flags in total.

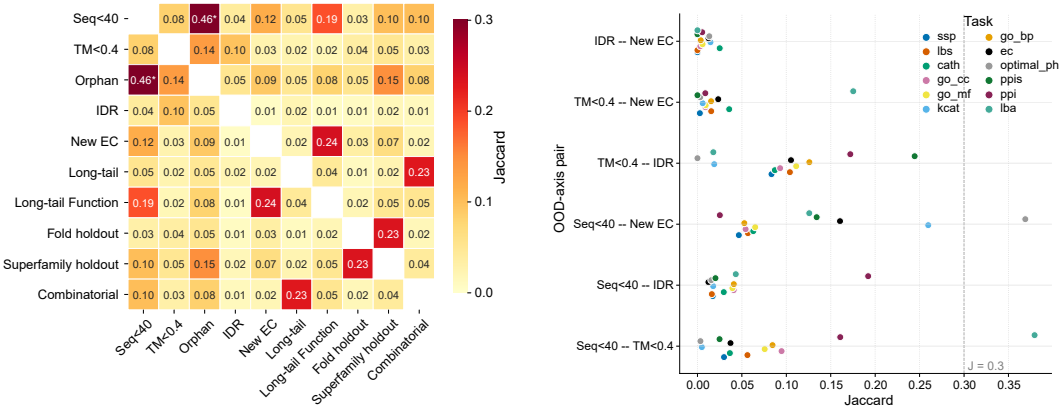
Family	Scenario	Criterion (test sample vs. train+val)	Tool	Thresholds / tasks
Sequence/evol.	LowHomology	max sequence identity below $X\%$	MMseqs2 (Steinegger & Söding, 2017) (easy-search)	X=90→30, step 10; all tasks
	Orphan	no significant hit (e-value $> 10^{-3}$ or none)	MMseqs2 (easy-search)	all tasks
	SuperfamilyHoldout	CATH superfamily (C.A.T.H) absent from training	CATH v4.4 + TED (Sillitoe et al., 2021)	all tasks
Structural	NewFold	max TM-score below T (symmetric mean for LBA)	Foldseek (van Kempen et al., 2024) (easy-search)	T=0.9→0.3, step 0.1; all tasks
	FoldHoldout	CATH topology (C.A.T) absent from training	CATH v4.4 + TED domain annotation	all except Fold
	ExtremeShort/Long	sequence length < 60 / > 1000 aa	length statistics	all tasks
	IDR	residues in long disordered regions $> 30\%$ ($> 10\%$ for pH, LBA)	metapredict v3 (Emenecker et al., 2021); flDPnn (Hu et al., 2021) (PPIS)	all tasks
Functional	NewFunction	*all* EC numbers of the protein (level L4/L3) unseen in training (all-not-in); for Fold, EC7 (L1) carriers are reserved to test before splitting	SIFTS/UniProt EC annotation	L4: all except Fold; L3: all except Fold, EC; L1 prior: Fold (336 domains)
	LongTailFunction	*all* EC numbers (L3/L4) rare in training (frequency ≤ 5 / ≤ 10, including 0)	EC frequency statistics	all tasks (L3 except EC)
	Combinatorial Function	label pair never co-occurs in train+val	label-pair statistics	EC, GO-MF/CC/BP
	LongTail (label)	the task's *own predicted label* rare in training; for regression tasks, the label value falls in a bin with ≤ 50 / ≤ 100 training samples	label-frequency / value-histogram statistics	Fold, EC, GO, LBS, PPIS, SSP; bins: k_{cat}, pH, LBA

The three families divide novelty in ways that mirror how proteins actually diverge from well-characterized data. Sequence/evolutionary scenarios form a ladder from near-duplicates to remote

homology, terminating at *Orphan* proteins with no detectable training homolog, together with holdouts of entire CATH superfamilies. Structural scenarios pair a TM-score ladder—0.5 is the canonical boundary between similar and distinct folds—with holdouts of entire CATH topologies, and further isolate proteins with extreme lengths or high disorder content, which are common in nature but underrepresented in standard datasets. Functional scenarios test whether models learn mechanistic principles rather than label statistics, probing unseen EC numbers, long-tailed labels, and label combinations never observed during training. Two reference annotations complete the design: the *Default* flag marks test samples produced by the normal split, and the *InD* flag marks the shift-free subset carrying no OOD annotation on any axis, which serves as the reference population for the retention metric. Scenarios that do not apply to a task are omitted, leaving 24–28 scenarios per task, extreme-length proteins are set aside or flagged in place so that length shifts are evaluated in isolation, and pair- and complex-level tasks (PPI, LBA) aggregate per-chain signals into sample-level flags.

3.2 Multi-axis overlap analysis

A prerequisite for multi-axis OOD evaluation is that individual axes capture complementary rather than redundant distributional shifts. We quantified pairwise overlap between OOD scenarios by computing Jaccard similarity $J(A, B) = |A \cap B|/|A \cup B|$ over the test set of each task. Figure 2 summarizes the results.



(a) (b)

Figure 2: **OOD-axis overlap analysis.** (a) Cross-task average Jaccard similarity between ten representative OOD axes; * marks values exceeding the 0.3 color-scale maximum. (b) Per-task Jaccard similarity for six selected axis pairs; the dotted grey line marks $J = 0.3$.

Across tasks, the average overlap between axes is low (Figure 2a). The highest value occurs between Seq<40 and Orphan (0.46), an inherent subset relation—Orphan proteins are by construction a limiting subset of the low-identity set; the next-highest values are NewEC–Long-tail EC (0.24), Fold holdout–Superfamily holdout (0.23), and Long-tail label–Combinatorial (0.23), while pairs involving TM<0.4 or IDR stay at or below 0.14 and 0.10, respectively. The per-task view tells the same story (Figure 2b): the largest overlaps are Seq<40–TM<0.4 on LBA (0.38) and Seq<40–NewEC on pH (0.37), and IDR–NewEC is near zero on every task (max 0.03). POOR's OOD axes therefore mark largely non-overlapping test populations and are suitable for multi-axis evaluation.

3.3 A unified OOD evaluation metric

Comparing OOD robustness across POOR's tasks is nontrivial for two reasons. First, the tasks rely on heterogeneous metrics—Fmax for multi-label function annotation, accuracy for classification, Spearman correlation for regression—whose raw values live on different scales and are differentially sensitive to label priors, so a fixed performance drop does not carry the same meaning across tasks. Second, the OOD scenarios defined above are not mutually exclusive: a test protein may be simultaneously remote in sequence, novel in fold, and extreme in length, and robustness under such fused high-dimensional shifts is precisely what a multi-axis benchmark must quantify. A single raw

metric on a single OOD subset cannot serve either purpose, because it conflates the effect of the shift itself with differences in subset composition and label distribution.

We therefore introduce **ProDPR** (Protein Distributional Performance Retention), a unified robustness measure built on three principles. *Null-adjusted utility*: every task metric is centered by the performance of a task-specific non-informative predictor, such as a label-prior baseline, so that scores from heterogeneous metrics become commensurable. *Shift isolation by reference*: rather than reporting raw performance on an OOD subset, ProDPR compares each OOD stratum against a reference of shift-free test proteins—by default the InD pool, which is shift-free on *every* descriptor axis; when only a subset of descriptor axes is evaluated, an alternative Auto pool additionally matches the reference to the stratum's composition on the remaining OOD axes, so that the comparison isolates the target shift rather than composition differences. The resulting per-stratum *retention* quantifies the fraction of effective predictive utility a model preserves when the shift is applied in isolation. *Explicit aggregation over strata*: test proteins are assigned to biologically interpretable strata spanning both single-axis scenarios and their combinations; strata without sufficient support are discarded, and per-stratum retentions are combined with macro weights by default, so that rare but hard scenarios are not swamped by the in-distribution majority. The same construction applies unchanged whether a stratum is defined by one axis or by an intersection of several, which is what makes fused multi-axis evaluation a special case rather than an add-on.

ProDPR yields a single robustness score per task, which we macro-average across tasks for a benchmark-level summary. Because averages can hide catastrophic but localized failures, we additionally report the *worst-stratum retention*, the minimum retention over all strata. Together these two quantities answer the two questions POOR is designed to ask: how much utility current models retain under realistic distribution shift, and where they break first. Formal definitions of the descriptor space, the matching procedure, and the estimators are deferred to Appendix A.

4 Experiments and Analysis

We benchmark five representative models spanning two architectural families: four protein language models—ESM-2 at three scales (35M, 150M, 650M) (Lin et al., 2023) and the structure-aware SaProt-650M (Su et al., 2024)—and the purely structure-based GVP-GNN (Jing et al., 2021). The four PLMs are evaluated with frozen backbone weights: residue representations are extracted once and cached, projected to 480 dimensions, and fed into a randomly initialized six-layer Transformer adapter with a task-specific output head, which is the only component trained on each task's training split (Appendix D), whereas GVP-GNN is trained end-to-end on each task. Unless stated otherwise, every model is scored with each task's designated metric (Table 3) on the Default test subset and on each OOD stratum, and robustness is summarized by ProDPR together with the worst-stratum retention (Section 3.3). We first profile a single representative model across all tasks and scenarios (Section 4.1), then compare single-axis against fused multi-axis shifts (Section 4.2), benchmark all five models under fused OOD conditions (Section 4.3), and finally isolate the effects of model scale and fine-tuning regime within a single model family (Section 4.4).

4.1 OOD generalization landscape

We first profile a single representative model, ESM2-650M, across all ten tasks (twelve task configurations) and twelve OOD scenarios per task—the ten representative axes of Figure 2 plus the two extreme-length axes. Table 2 reports the raw task metric on each OOD scenario (mean $\pm$ std over five runs). Reading one row of Table 2 ranks the scenarios by generalization difficulty for that task; comparing these rankings across tasks reveals which shift family dominates each task's failure modes.

Three patterns emerge. First, structural novelty is the dominant failure axis for sequence-based function annotation: at TM$<$0.7, F_{max} drops by 70–81% relative to InD—far below the corresponding low-sequence-identity scenarios. GO-CC is the exception among the annotation tasks, degrading mildly under structural novelty and instead bottoming out on long-tail labels; for fold classification, the discrete superfamily holdout is much harder than any similarity-graded stratum. Second, extreme lengths and holdout scenarios are the hardest conditions for the enzyme-property and site-prediction tasks—for example, extreme-short chains form the hardest scenario for k_{cat} (Spearman 0.143), and

Table 2: In-distribution (InD) and out-of-distribution (OOD) performance of ESM2-650M across 12 protein prediction tasks. Results are reported as mean ± standard deviation over five runs. Arrows indicate the preferred metric direction. For each task, the hardest available OOD setting is shown in **bold**; the InD reference column is excluded from this selection. Unavailable task–scenario combinations are denoted by an em dash.

Task	Metric	InD	Seq. 0.7	TM-score 0.7	IDR	Orphan	Extreme Short	Extreme Long	Fold Holdout	Superfamily Holdout	New EC	Long Tail†	Long Tail EC†	Combinatorial Function
EC	F_{max} ↑	0.908 ±.015	0.724 ±.013	0.272 ±.016	0.389 ±.048	0.349 ±.018	**0.187** ±.026	0.643 ±.130	0.204 ±.048	0.467 ±.035	0.520 ±.010	0.193 ±.028	0.632 ±.010	0.406 ±.024
Fold	Accuracy↑	0.943 ±.003	0.818 ±.007	0.428 ±.019	0.613 ±.034	0.587 ±.012	0.587 ±.022	—	—	**0.297** ±.017	0.702 ±.036	0.337 ±.019	0.713 ±0.033	—
GO-BP	F_{max} ↑	0.904 ±.003	0.544 ±.004	0.168 ±.006	0.643 ±.010	**0.153** ±.008	0.480 ±.007	0.335 ±.026	0.165 ±.015	0.223 ±.006	0.386 ±.019	0.325 ±.003	0.514 ±0.008	0.314 ±.004
GO-CC	F_{max} ↑	0.739 ±.011	0.566 ±.012	0.454 ±.016	0.568 ±.031	0.459 ±.015	0.547 ±.014	0.507 ±.026	0.460 ±.020	0.449 ±.025	0.556 ±.014	**0.361** ±.010	0.572 ±.007	0.405 ±.007
GO-MF	F_{max} ↑	0.840 ±.009	0.606 ±.007	**0.213** ±.017	0.638 ±.023	0.251 ±.017	0.344 ±.007	0.468 ±.028	0.347 ±.047	0.324 ±.022	0.618 ±.014	0.362 ±.024	0.652 ±0.013	0.436 ±.013
k_{cat}	Spearman ρ ↑	0.278 ±.194	0.351 ±.021	0.229 ±.173	0.457 ±.242	0.299 ±.105	**0.143** ±0.077	0.277 ±.058	0.286 ±.176	0.215 ±.086	0.333 ±.030	0.539 ±.068	0.294 ±.015	—
LBA	Spearman ρ ↑	0.668 ±.006	0.222 ±.025	0.310 ±.031	0.591 ±.024	0.117 ±.058	0.326 ±.300	0.536 ±.024	0.221 ±.055	0.173 ±.078	**−0.012** ±.074	0.527 ±.035	0.509 ±.036	—
LBS	AUROC↑	0.944 ±.001	0.927 ±.001	0.826 ±.004	0.820 ±.006	0.844 ±.005	**0.644** ±.010	0.900 ±.004	0.885 ±.012	0.923 ±.006	0.942 ±.001	0.810 ±.009	0.949 ±0.001	—
pH	Spearman ρ ↑	0.666 ±.022	0.391 ±.015	0.251 ±.038	0.422 ±.073	0.183 ±.033	**0.069** ±.094	0.565 ±.027	0.311 ±.044	0.280 ±.022	0.272 ±.028	0.523 ±.053	0.249 ±.049	—
PPI	AUROC↑	0.717 ±.013	0.771 ±.006	**0.701** ±.013	0.705 ±.006	0.716 ±.013	0.742 ±.013	0.714 ±.018	0.852 ±.049	0.762 ±.024	0.868 ±.022	—	0.846 ±.015	—
PPIS	AUROC↑	0.793 ±.006	0.769 ±.003	0.684 ±.007	0.737 ±.004	0.700 ±.006	0.631 ±.010	**0.575** ±.018	0.721 ±.009	0.710 ±.003	0.780 ±.003	0.581 ±.013	0.751 ±.005	—
SSP	Residue acc.↑	0.861 ±.002	0.770 ±.001	**0.680** ±.002	0.840 ±.002	0.715 ±.002	—	0.683 ±.001	0.765 ±.006	0.736 ±.002	0.777 ±.001	0.717 ±.002	0.779 ±0.002	—

EC: Enzyme Commission; GO-BP/CC/MF: Gene Ontology biological process/cellular component/molecular function; LBA: ligand-binding affinity; LBS: ligand-binding site; PPI: protein–protein interaction; PPIS: protein–protein interaction site; SSP: secondary-structure prediction; IDR: intrinsically disordered region. NewEC uses L4, except Fold (L1). † Long Tail is the long-tail scenario defined on each task's own labels, whereas Long Tail EC refers to the long-tail scenario defined on EC functional annotations; for each task, we select the available threshold (and EC level, where applicable) that yields the lowest seed-averaged ProDPR relative to the full InD reference, and report the corresponding original metric as mean ± sample standard deviation. Candidate test subsets must contain more than 100 samples, except for non-EC Long Tail in LBA, pH, PPI, and PPIS. "—" denotes unavailable or ineligible results.

extreme-length proteins likewise form the hardest scenario for pH, LBS, and PPIS. Third, pair-level interaction prediction is comparatively robust on every axis: no scenario degrades PPI by more than 0.05 AUROC relative to InD, and several scenarios exceed InD, most notably under functional novelty. We note that scenario-level AUROC on PPI is additionally shaped by the composition of the negative class, which inflates OOD-scenario AUROC relative to InD (Appendix E).

4.2 Single-axis versus fused multi-axis OOD shifts

The strata examined above isolate one shift at a time, but proteins in the wild routinely depart from the training distribution along several axes at once. To quantify this compounding effect, we evaluate each model on three complementary single-axis strata—low sequence identity (SeqID< 0.7), low structural similarity (TM-score< 0.7), and unseen level-4 EC annotations (NewEC-L4)—one axis per biological dimension of the POOR taxonomy, together with their intersection, i.e., the fused stratum of test proteins flagged on all three axes simultaneously. Every stratum is scored against a shift-free reference, and Figure 1c reports the resulting worst-stratum retention, task-averaged over the twelve task configurations, so that the comparison reflects where each model breaks first rather than its average case. If a fused shift were no harder than its hardest component, fused retention would track the minimum of the three single-axis retentions; a systematic drop below that minimum indicates super-additive difficulty, meaning that the axes compound rather than merely co-occur.

Fused shifts are consistently the hardest condition: for every model, the fused stratum yields the lowest worst-stratum retention of the four conditions—0.103 (ESM-2 35M), 0.307 (ESM-2 150M), 0.227 (ESM-2 650M), 0.334 (SaProt-650M), and 0.423 (GVP-GNN)—falling below each model's hardest single axis by 0.09–0.38 (on average 0.20), and in the worst case leaving ESM-2 35M with only about one-tenth of its shift-free utility. The degradation under the joint shift is also sharpest for protein language models: the compounding gap is smallest for the purely structure-based GVP-GNN (0.423 fused vs. 0.515 on its hardest single axis), whose fused worst case is dominated by the structural axis alone, whereas the PLMs degrade far more sharply. Jointly shifted subsets therefore exhibit lower retention than the corresponding marginal OOD subsets, and evaluating one axis at a time systematically overestimates robustness under realistic, jointly shifted proteins.

4.3 Cross-model comparison under multi-axis OOD shifts

We compare five models using their task-specific test metrics and ProDPR under joint multi-axis stratification (Figure 3). The two views answer complementary questions: how accurately a model predicts, and how much utility it retains relative to its shift-free reference after accounting for the null baseline.

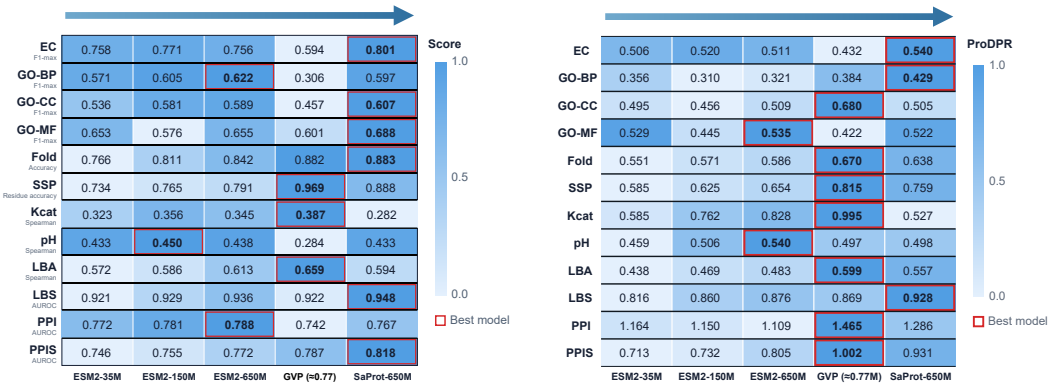


Figure 3: **Predictive performance and utility retention across models.** (a) Task-specific metrics on the full test set. (b) ProDPR (InD reference pool) aggregated over joint sequence, structure, and NewEC-L4 strata; Fold uses NewEC-L1. Cell values are unnormalized; colors are normalized independently within each row for within-task comparison. Red outlines indicate row maxima. ProDPR above one indicates average utility exceeding the shift-free reference after null adjustment.

Predictive performance does not determine retention. ProDPR reveals ranking reversals that raw metrics alone cannot capture: on GO-CC, SaProt-650M achieves higher F_{max} than GVP-GNN (0.607 vs. 0.457) yet lower ProDPR (0.505 vs. 0.680), and the same reversal appears on PPIS (AUROC 0.818 vs. 0.787; ProDPR 0.931 vs. 1.002). High scores can likewise coexist with substantial degradation—on Fold, ESM2-650M attains 0.842 accuracy yet retains only 0.586 of its shift-free utility—and the non-monotonic retention across ESM-2 scales on EC (0.506, 0.520, 0.511) further cautions against treating either predictive performance or model scale as a proxy for robustness. ProDPR therefore adds a reference-relative view of distribution-shift sensitivity rather than another measure of absolute accuracy.

Model rankings also differ between the two views and across tasks. On absolute performance, no single model dominates (Figure 3a): SaProt-650M attains the best raw score on six of the twelve task configurations and GVP-GNN on three, with the rest split across ESM-2 scales; scaling is not universally helpful, as GO-MF dips non-monotonically at 150M across ESM-2 scales ($0.653 \rightarrow 0.576 \rightarrow 0.655$) and GVP-GNN collapses on pH (0.284 Spearman), a physicochemical task poorly served by a purely geometric encoder. Retention re-ranks the field (Figure 3b): macro-averaged ProDPR improves monotonically within the ESM-2 family ($0.600 \rightarrow 0.617 \rightarrow 0.646$), yet the highest overall belongs to GVP-GNN (0.736, best per-task retention on seven of twelve tasks), followed by SaProt-650M (0.677); on PPI all five models exceed reference parity (1.109–1.465), a super-retention regime we trace to the negative-pair composition of the PPI test set rather than to genuine robustness. Which paradigm wins depends on where the task's supervision signal lives: sequence-driven function-annotation tasks (EC, GO) split retention across SaProt-650M, ESM-2-650M, and GVP-GNN, whereas every structure- and interaction-centric task (Fold, SSP, k_{cat}, LBA, LBS, PPI, PPIS) is won in retention by a structure-aware encoder—GVP-GNN on six of them, SaProt-650M on LBS. Task difficulty is likewise axis-dependent, with GO-BP, GO-MF, pH, and LBA the hardest (mean retention 0.360–0.509) and PPI, LBS, and PPIS the most robust (0.837–1.235).

4.4 Model scale and fine-tuning regime shape overfitting

The cross-model comparison above changes model architecture and input modality simultaneously. We next isolate two factors that can be controlled within one pretrained model family: parameter count and downstream training regime. On the POOR-EC task, we evaluate ESM-2 at five scales (8M, 35M, 150M, 650M, and 3B parameters) using either full fine-tuning or a linear probe. Both regimes use a batch size of 32; full fine-tuning updates the entire network with a learning rate of 10^{-5}, whereas the linear probe freezes the pretrained backbone and trains only the prediction head with a learning rate of 10^{-3}. We report F-max on the training split and the Default test subset

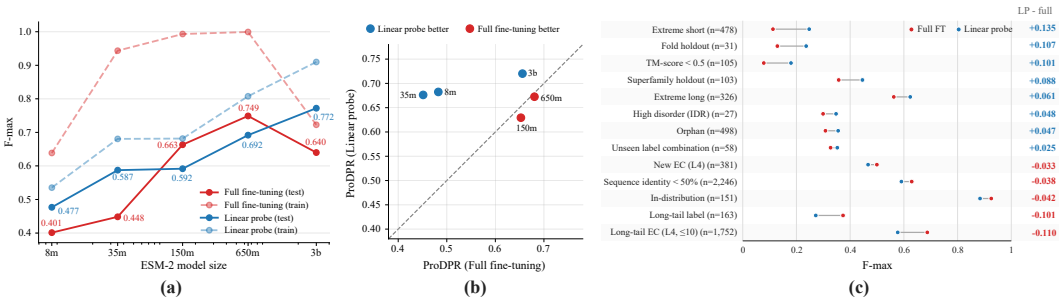


Figure 4: **Effects of model scale and training regime on POOR-EC generalization.** (a) Train and Default-test F-max for full fine-tuning (Full FT) and linear probing across five ESM-2 scales. Solid lines denote test performance and translucent dashed lines denote training performance; test values are annotated. (b) ProDPR (Auto reference pool) for the same model–training combinations. Higher values indicate that more effective predictive utility is retained across OOD strata. (c) F-max on representative subsets for ESM-2-650M. Rows are ordered by the difference between linear probing and full fine-tuning (LP − full); sample counts are shown in parentheses. Blue differences favor the linear probe and red differences favor full fine-tuning.

($n = 4{,}243$), together with ProDPR over the POOR-EC OOD strata. This controlled comparison tests whether increased capacity translates into transferable task knowledge or instead amplifies adaptation to the training distribution.

Figure 4 summarizes the results, which decompose into four observations. First, stronger fitting capacity does not imply stronger test performance: full fine-tuning nearly memorizes the training set at 35M, 150M, and 650M (train F-max 0.944, 0.993, 0.999) yet tests far lower (0.448, 0.663, 0.749), and at the two smallest scales it is even outperformed on the test set by the linear probe (8M: 0.401 vs. 0.477; 35M: 0.448 vs. 0.587); the ProDPR gap in Figure 4b mirrors this, with the linear probe retaining substantially more utility at 8M and 35M (0.68 vs. 0.48 and 0.68 vs. 0.45). Second, the overfitting of full fine-tuning is itself scale-dependent: the train–test gap shrinks from 0.495 at 35M to 0.330 at 150M and 0.250 at 650M, suggesting that intermediate-scale models preserve more pretrained knowledge under full adaptation—plausibly because the smallest models are more prone to catastrophic forgetting—yet they still trail the linear probe on a majority of OOD subsets (eight of thirteen in Figure 4c, with the linear probe's largest gains on extreme-short proteins, +0.135 F-max, topology holdouts, +0.107, and TM-score$<$ 0.5, +0.101), so that aggregate ProDPR is nearly tied at these scales (0.65 vs. 0.63 at 150M; 0.68 vs. 0.67 at 650M). Third, the performance drop of full fine-tuning at 3B (test F-max 0.640) is better explained by *underfitting* than by overfitting: under the same fixed training recipe its training F-max reaches only 0.723, indicating that the shared optimization configuration, rather than excess capacity, is the bottleneck at this scale. Finally, the linear probe improves monotonically with model scale on the test set ($0.477 \rightarrow 0.587 \rightarrow 0.592 \rightarrow 0.692 \rightarrow 0.772$) while its ProDPR remains comparatively stable (0.63–0.72, peaking at 3B), indicating that probing a frozen backbone rarely overfits and more faithfully reflects the pretrained model's inherent capability. Where the two regimes differ, their failure modes differ as well: the linear probe is stronger under structural novelty and extreme inputs, whereas full fine-tuning better exploits task-specific supervision for rare or familiar functional labels (long-tail EC +0.110, long-tail task labels +0.101); model scale and training regime should therefore be reported together in OOD evaluation.

5 Conclusion

We presented **POOR**, a multi-axis out-of-distribution benchmark for protein understanding, together with **ProDPR**, a reference-based metric that quantifies how much predictive utility models retain under distribution shift. Systematic evaluation across representative protein models shows that in-distribution performance is a poor proxy for OOD robustness, that robustness depends jointly on the shift axis and the model's inductive bias, and that fused multi-axis scenarios, in which several OOD factors co-occur, are consistently the hardest condition. We expect POOR and ProDPR to serve as an extensible testbed for shift-aware method development, moving reported progress in protein modeling closer to open-world reliability.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, assist with translation, clean and reformat datasets, and support qualitative and thematic data analysis. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or interpret results. Generating synthetic data sets and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, draft parts of the research paper, and edit the research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, New York, NY, 6th edition, 2014. ISBN 9780815344322.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Judith Bernett, David B. Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2):bbae076, 2024. doi: 10.1093/bib/bbae076.

Veda Sheersh Boorla and Costas D. Maranas. CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature Communications*, 16(1):2072, 2025. doi: 10.1038/s41467-025-57215-9.

Carl-Ivar Brändén and John Tooze. *Introduction to Protein Structure*. Garland Science, New York, NY, 2nd edition, 1999. doi: 10.1201/9781136969898.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/D3SC04185A.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research*, 49(D1):D498–D508, 2021. doi: 10.1093/nar/gkaa1025.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

José M. Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research*, 47(D1):D482–D489, 2019. doi: 10.1093/nar/gky1114.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Timothy F. Huddy, Sam Pellock, Doug Tischer, Ferdinand Chan, Brian Koepnick, Hahnbeom Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, Emanuele Rossi, Guoqing Zhou, Srimukh Veccham, Clemens Isert, Yuxing Peng, Prabindh Sundareson, Mehmet Akdel, Gabriele Corso, Hannes Stärk, Gerardo Tauriello, Zachary Carpenter, Michael Bronstein, Emine Kucukbenli, Torsten Schwede, and Luca Naef. PLINDER: The protein-ligand interactions dataset and evaluation resource, 2024.

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G. Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6(12):1512–1524, 2024. doi: 10.1038/s42256-024-00931-6.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Ryan J. Emenecker, Daniel Griffith, and Alex S. Holehouse. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical Journal*, 120(20):4312–4319, 2021. doi: 10.1016/j.bpj.2021.08.039.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, 2022. doi: 10.1038/s41467-022-32007-7.

Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, 2014. doi: 10.1093/nar/gkt1240.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, 2021. doi: 10.1038/s41467-021-23303-9.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Jie Hou, Badri Adhikari, and Jianlin Cheng. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 2018. doi: 10.1093/bioinformatics/btx780.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, 40(7):1114–1122, 2022a. doi: 10.1038/s41587-021-01146-5.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970, 2022b.

Gang Hu, Akila Katuwawala, Kui Wang, Zhonghua Wu, Sina Ghadermarzi, Jianzhao Gao, and Lukasz Kurgan. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications*, 12(1):4438, 2021. doi: 10.1038/s41467-021-24773-7.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H. Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 15820–15831, 2019.

Arian R. Jamasb, Alex Morehead, Chaitanya K. Joshi, Zuobai Zhang, Kieran Didi, Simon V. Mathis, Charles Harris, Jian Tang, Jianlin Cheng, Pietro Liò, and Tom L. Blundell. Evaluating representation learning on the protein structure universe. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sTYuRVrdK3.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, Houtim Lai, Wei Liu, Junzhou Huang, Shuigeng Zhou, Ping Luo, Peilin Zhao, and Yatao Bian. DrugOOD: Out-of-distribution dataset curator and benchmark for AI-Aided drug discovery – a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023. doi: 10.1609/aaai.v37i7.25970.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

Anowarul Kabir, Asher Moldwin, and Amarda Shehu. A comparative analysis of transformer-based protein language models for remote homology prediction. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB '23)*, pp. 1–9, 2023. doi: 10.1145/3584371.3612942.

Wolfgang Kabsch and Chris Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: 10.1002/bip.360221211.

Daniel Kovtun, Mehmet Akdel, Alexander Goncearenco, Guoqing Zhou, Graham Holt, David Baugher, Dejun Lin, Yusuf Adeshina, Thomas Castiglione, Xiaoyun Wang, Céline Marquet, Matt McPartlon, Tomas Geffner, Gabriele Corso, Hannes Stärk, Zachary Carpenter, Emine Kucukbenli, Michael Bronstein, and Luca Naef. PINDER: The protein interaction dataset and evaluation resource, 2024. bioRxiv preprint.

Alexander Kroll, Yvan Rousset, Xiao-Pan Hu, Nina A. Liebrand, and Martin J. Lercher. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature Communications*, 14(1):4139, 2023. doi: 10.1038/s41467-023-39840-4.

Tim Kucera, Carlos Oliver, Dexiong Chen, and Karsten Borgwardt. ProteinShake: Building datasets and benchmarks for deep learning on protein structures. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 58277–58289. Curran Associates, Inc., 2023.

Maxat Kulmanov and Robert Hoehndorf. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics*, 38(Supplement 1): i238–i245, 2022. doi: 10.1093/bioinformatics/btac256.

Maxat Kulmanov, Francisco J. Guzmán-Vega, Paula Duek Roggli, Lydie Lane, Stefan T. Arold, and Robert Hoehndorf. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence*, 6(2):220–228, 2024. doi: 10.1038/s42256-024-00795-w.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Iain H. Moal and Juan Fernández-Recio. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, 28(20):2600–2607, 2012. doi: 10.1093/bioinformatics/bts489.

Alex Morehead, Chen Chen, Ada Sedova, and Jianlin Cheng. DIPS-Plus: The enhanced database of interacting protein structures for interface prediction. *Scientific Data*, 10(1):509, 2023. doi: 10.1038/s41597-023-02409-3.

Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995. doi: 10.1016/S0022-2836(05)80134-2.

Vamsi Nallapareddy, Nicola Bordin, Ian Sillitoe, Michael Heinzinger, Maria Littmann, Vaishali P. Waman, Neeladri Sen, Burkhard Rost, and Christine A. Orengo. CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models. *Bioinformatics*, 39(1):btad029, 2023. doi: 10.1093/bioinformatics/btad029.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan J. Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023 Datasets and Benchmarks Track)*, pp. 64331–64379, 2023.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug-target interaction predictions. *Briefings in Bioinformatics*, 16(2):325–337, 2015. doi: 10.1093/bib/bbu010.

Predrag Radivojac, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M. Yunes, Ameet S. Talwalkar, Susanna Repo, Michael L. Souza, Damiano Piovesan, Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W. A. Buchan, Kevin Bryson, David T. Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K. Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M. Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E. Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaßner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas A. Hopf, Stefanie Kaufmann, Michael Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N. Wass, Michael J. E. Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis A. I. Kourmpetis, Aalt D. J. van Dijk, Cajo J. F. ter Braak, Yuanpeng Zhou, Qingtian Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C. Babbitt, Steven E. Brenner, Christine Orengo, Burkhard Rost, Sean D. Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013. doi: 10.1038/nmeth.2340.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John F. Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 9686–9698, 2019.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P. Waman, Paul Ashford, Harry M. Scholes, Camilla S. M. Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, Mahnaz Abbasian, Sean Le Cornu, Su Datt Lam, Karel Berka, Ivana Hutařová Varekova, Radka Svobodova, Jon Lees, and

Christine A. Orengo. CATH: increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1):D266–D273, 2021. doi: 10.1093/nar/gkaa1079.

Rajarshi SinhaRoy, Christian Clauß, Ivan Ivanikov, and Georg Künze. Dataset EnzyBase12k: A curated dataset for predicting enzyme pH optima using pHoptNN, 2026. Zenodo dataset.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. EquiBind: Geometric deep learning for drug binding structure prediction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20503–20521, 2022.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, volume 2024, pp. 6987–7009, 2024.

Felix Teufel, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. GraphPart: homology partitioning for biological sequence analysis. *NAR Genomics and Bioinformatics*, 5(4):lqad088, 2023. doi: 10.1093/nargab/lqad088.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on molecules in three dimensions, 2020.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. doi: 10.1038/s41587-023-01773-0.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022. doi: 10.1093/nar/gkab1061.

Martin Šošić and Mile Šikić. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017. doi: 10.1093/bioinformatics/btw753.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004. doi: 10.1021/jm030580l.

Ulrike Wittig, Renate Kania, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Heidrun Sauer-Danzwith, Saqib Mir, Olga Krebs, Meik Bittkowski, Elina Wetsch, Isabel Rojas, and Wolfgang Müller. SABIO-RK–database for biochemical reaction kinetics. *Nucleic Acids Research*, 40(D1):D790–D796, 2012. doi: 10.1093/nar/gkr1046.

Jinbo Xu. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(34):16856–16865, 2019. doi: 10.1073/pnas.1821309116.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022 Datasets and Benchmarks Track)*, pp. 35156–35173, 2022.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023. doi: 10.1126/science.adf2465.

Xing Zhang, Hongmei Yin, Fei Ling, Jian Zhan, and Yaoqi Zhou. SPIN-CGNN: Improved fixed backbone protein design with contact map-based graph construction and contact graph neural network. *PLoS Computational Biology*, 19(12):e1011330, 2023a. doi: 10.1371/journal.pcbi.1011330.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=to3qCB3tOh9.

Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, Larry Davis, Tunca Dogan, Volkan Atalay, Ahmet S. Rifaioglu, Alperen Dalkıran, Rengul Cetin Atalay, Chengxin Zhang, Rebecca L. Hurto, Peter L. Freddolino, Yang Zhang, Prajwal Bhat, Fran Supek, José M. Fernández, Branislava Gemovic, Vladimir R. Perovic, Radoslav S. Davidović, Neven Sumonja, Nevena Veljkovic, Ehsaneddin Asgari, Mohammad R. K. Mofrad, Giuseppe Profiti, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, Florian Boecker, Heiko Schoof, Indika Kahanda, Natalie Thurlby, Alice C. McHardy, Alexandre Renaux, Rabie Saidi, Julian Gough, Alex A. Freitas, Magdalena Antczak, Fabio Fabris, Mark N. Wass, Jie Hou, Jianlin Cheng, Zheng Wang, Alfonso E. Romero, Alberto Paccanaro, Haixuan Yang, Tatyana Goldberg, Chenguang Zhao, Liisa Holm, Petri Törönen, Alan J. Medlar, Elaine Zosa, Itamar Borukhov, Ilya Novikov, Angela Wilkins, Olivier Lichtarge, Po-Han Chi, Wei-Cheng Tseng, Michal Linial, Peter W. Rose, Christophe Dessimoz, Vedrana Vidulin, Saso Dzeroski, Ian Sillitoe, Sayoni Das, Jonathan Gill Lees, David T. Jones, Cen Wan, Domenico Cozzetto, Rui Fa, Mateo Torres, Alex Warwick Vesztrocy, Jose Manuel Rodriguez, Michael L. Tress, Marco Frasca, Marco Notaro, Giuliano Grossi, Alessandro Petrini, Matteo Re, Giorgio Valentini, Marco Mesiti, Daniel B. Roche, Jonas Reeb, David W. Ritchie, Sabeur Aridhi, Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Da Chen Emily Koo, Richard Bonneau, Vladimir Gligorijević, Meet Barot, Hai Fang, Stefano Toppo, Enrico Lavezzo, Marco Falda, Michele Berselli, Silvio C. E. Tosatto, Marco Carraro, Damiano Piovesan, Hafeez Ur Rehman, Qizhong Mao, Shanshan Zhang, Slobodan Vucetic, Gage S. Black, Dane Jo, Erica Suh, Jonathan B. Dayton, Dallas J. Larsen, Ashton R. Omdahl, Liam J. McGuffin, Danielle A. Brackenridge, Patricia C. Babbitt, Jeffrey M. Yunes, Paolo Fontana, Feng Zhang, Shanfeng Zhu, Ronghui You, Zihan Zhang, Suyang Dai, Shuwei Yao, Weidong Tian, Renzhi Cao, Caleb Chandler, Miguel Amezola, Devon Johnson, Jia-Ming Chang, Wen-Hung Liao, Yi-Wei Liu, Stefano Pascarelli, Yotam Frank, Robert Hoehndorf, Maxat Kulmanov, Imane Boudellioua, Gianfranco Politano, Stefano Di Carlo, Alfredo Benso, Kai Hakala, Filip Ginter, Farrokh Mehryary, Suwisa Kaewphan, Jari Björne, Hans Moen, Martti E. E. Tolvanen, Tapio Salakoski, Daisuke Kihara, Aashish Jain, Tomislav Šmuc, Adrian Altenhoff, Asa Ben-Hur, Burkhard Rost, Steven E. Brenner, Christine A. Orengo, Constance J. Jeffery, Giovanni Bosco, Deborah A. Hogan, Maria J. Martin, Claire O'Donovan, Sean D. Mooney, Casey S. Greene, Predrag Radivojac, and Iddo Friedberg. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):244, 2019. doi: 10.1186/s13059-019-1835-8.

A PROTEIN OOD EVALUATION WITH PRODPR

Protein out-of-distribution (OOD) generalization is inherently multi-factorial. A test protein may differ from the training distribution in sequence space, structure space, annotation space, or in a combination of these factors. We therefore do not model protein OOD as a one-dimensional similarity curve. Instead, we define protein OOD evaluation over a high-dimensional descriptor space and

derive ***ProDPR***, short for Protein Distributional Performance Retention, as a finite-sample estimator of an OOD-space retention functional.

A.1 Problem setup and the ProDPR estimator

Let P_t^{te} denote the test distribution for task t over protein-label pairs (X, Y). A model f is evaluated by a task-specific utility functional $\mathcal{M}_t(f; P)$, where larger values indicate better performance. In practice, $\mathcal{M}_t$ is instantiated by the corresponding empirical metric, such as Fmax for multi-label GO prediction, accuracy for classification, Spearman correlation for regression, or TM-score/LDDT for structure prediction. To place heterogeneous tasks and metrics on a common scale, we define the null-adjusted effective utility

$$\Delta_t(f; P) = \mathcal{M}_t(f; P) - \mathcal{M}_t(f_{\mathrm{null}}; P), \tag{1}$$

where f_{null} is a task-specific non-informative predictor, such as a label-prior baseline.

Each protein x is mapped to a task-specific OOD descriptor

$$\mathbf{z}_t(x) = [\mathbf{c}_t(x), \mathbf{a}_t(x)] \in \mathcal{Z}_t, \tag{2}$$

where $\mathbf{c}_t(x)$ contains continuous or discretized novelty coordinates, and $\mathbf{a}_t(x)$ contains binary biological scenario annotations. For example, $\mathbf{c}_t(x)$ may include sequence novelty, structural novelty, length novelty, and label-cardinality novelty, whereas $\mathbf{a}_t(x)$ may encode IDR, orphan, long-tail, new-function, or combinatorial-function annotations. Sequence and structural novelty are computed from maximum similarity to the training set:

$$\nu_{\mathrm{seq}}(x) = 1 - \max_{z \in \mathcal{D}_t^{\mathrm{tr}}} s_{\mathrm{seq}}(x, z), \qquad \nu_{\mathrm{str}}(x) = 1 - \max_{z \in \mathcal{D}_t^{\mathrm{tr}}} s_{\mathrm{str}}(x, z), \tag{3}$$

where s_{seq} can be computed from coverage-corrected sequence identity, and s_{str} can be computed from a symmetric structural similarity score such as the symmetric mean $\frac{1}{2}\big(\mathrm{TM}(x, z) + \mathrm{TM}(z, x)\big)$.

At the population level, an ideal OOD evaluation functional would measure the expected retained utility over the descriptor space $\mathcal{Z}_t$. Let P_t^z denote the conditional test distribution of (X, Y) given $\mathbf{z}_t(X) = z$, and let Q_t^z denote a reference distribution that removes the OOD shift represented by z while preserving relevant nuisance factors. Given an evaluation measure μ_t over $\mathcal{Z}_t$, the population form of ProDPR is

$$\mathcal{R}_t(f; \mu_t) = \int_{\mathcal{Z}_t} \frac{\Delta_t(f; P_t^z)}{\Delta_t(f; Q_t^z) + \epsilon} \, d\mu_t(z), \tag{4}$$

where $\epsilon > 0$ is a small constant for numerical stability. This quantity measures the expected fraction of effective predictive utility retained across the protein OOD descriptor space.

Stratification. Since exact conditioning on a high-dimensional continuous–discrete descriptor is statistically infeasible, we estimate Eq. 4 by partitioning $\mathcal{Z}_t$ into a finite collection of biologically interpretable OOD strata:

$$\mathcal{P}_t = \{A_{t,1}, A_{t,2}, \ldots, A_{t,L_t}\}, \qquad A_{t,\ell} \subseteq \mathcal{Z}_t. \tag{5}$$

Concretely, given a selected set of descriptor axes, every test protein is assigned to exactly one cell of the joint cross of those axes—its sequence-novelty bin, its structure-novelty bin, and its joint binary pattern over the selected scenario axes—so each stratum may correspond to a joint sequence–structure novelty region, a biological scenario pattern, or an interaction between novelty and scenario annotations. The cell in which every selected axis sits at its shift-free state (the highest-similarity bins and no active scenario flag) defines the reference population rather than an OOD region and is excluded from the strata. The induced test subset of a stratum is

$$\mathcal{E}_{t,\ell} = \left\{x \in \mathcal{D}_t^{\mathrm{te}} : \mathbf{z}_t(x) \in A_{t,\ell}\right\}. \tag{6}$$

Only strata with sufficient support are retained:

$$\Omega_t = \left\{\ell : |\mathcal{E}_{t,\ell}| \geq n_{\min}\right\}. \tag{7}$$

Cells with fewer than $n_{\min}$ proteins are discarded; an axis whose proteins all fall into discarded cells may be reintroduced as a single-axis stratum when it retains at least $n_{\min}$ otherwise-uncovered proteins. Each protein contributes to at most one stratum, so the retained strata are mutually disjoint, although they need not cover the whole test set.

Reference pools. For each valid stratum $\ell \in \Omega_t$, we construct a reference subset $\mathcal{B}_{t,\ell}$ whose role is to provide the performance level expected after removing the evaluated OOD shift. All pools are built from test proteins with completely computed OOD annotations. We define three reference pools:

- the *InD pool*: test proteins at the shift-free state of *every* descriptor axis (all novelty axes in their highest-similarity bins and all scenario flags inactive);
- the *Default pool*: all test proteins with fully computed OOD annotations, used without further filtering;
- the *Auto pool*: proteins that are shift-free on the *selected* axes, preferring the highest-similarity bins, constructed by a tiered relaxation within the Default pool (Section A.2).

Under the InD and Default pools the reference subset is the whole pool, $\mathcal{B}_{t,\ell} = \mathcal{R}_t$, identical and deterministic across strata. Under the Auto pool, $\mathcal{B}_{t,\ell}$ is a background-matched control sample drawn per stratum (Section A.2). Unless stated otherwise, **all experimental results in this paper use the InD pool in full**.

The retained utility of model f on OOD stratum ℓ is then

$$r_t(f, \ell) = \frac{M_t(f, \mathcal{E}_{t,\ell}) - M_t(f_{\text{null}}, \mathcal{E}_{t,\ell})}{M_t(f, \mathcal{B}_{t,\ell}) - M_t(f_{\text{null}}, \mathcal{B}_{t,\ell}) + \epsilon}. \tag{8}$$

Finally, the empirical ProDPR score is the finite-stratum approximation of Eq. 4:

$$\text{ProDPR}_t(f) = \sum_{\ell \in \Omega_t} \mu_t(\ell) r_t(f, \ell), \qquad \sum_{\ell \in \Omega_t} \mu_t(\ell) = 1, \qquad \mu_t(\ell) \geq 0. \tag{9}$$

By default, we use macro weights $\mu_t(\ell) = 1/|\Omega_t|$, assigning equal importance to all valid OOD strata. Alternatively, μ_t can be specified according to the intended biological use case or the desired OOD composition of the target protein dataset. Thus, ProDPR is not a post-hoc aggregation of separate OOD scores; it is a finite-sample estimator of a single high-dimensional OOD retention functional.

We also report the worst-stratum retention within each task:

$$\text{WorstStratum}_t(f) = \min_{\ell \in \Omega_t} r_t(f, \ell), \tag{10}$$

which captures failure modes that may be hidden by average performance.

A.2 Background-matched control sampling

This section describes the reference construction used under the Auto pool, the only pool mode that involves sampling. Its purpose is to isolate the evaluated axes: the reference should sit at the shift-free state of the *selected* axes while reproducing the stratum's composition on all *remaining* descriptor dimensions, so that the retention ratio in Eq. 8 attributes utility differences to the evaluated shift alone.

Auto pool construction. Candidates are Default-annotated test proteins with no active flag on any selected scenario axis, preferring proteins in the highest-similarity novelty bins; when this strictest definition leaves too few proteins for a stable denominator, the similarity requirements are progressively relaxed until at least $n_{\min}$ proteins are available. The Auto pool therefore contains only proteins at the shift-free state of the selected axes, while their values on unselected axes are left unrestricted—those dimensions are balanced by the matching step below.

Background profile. When evaluation uses only a subset of the available descriptor axes, we summarize the remaining dimensions by the background profile

$$H_t(x) = \left(\mathbf{a}_t^{\text{bg}}(x), b_{\text{seq}}^{\text{bg}}(x), b_{\text{str}}^{\text{bg}}(x) \right), \tag{11}$$

where $\mathbf{a}_t^{\text{bg}}(x)$ is the joint binary pattern over the scenario axes *not* selected for evaluation, and $b_{\text{seq}}^{\text{bg}}(x)$, $b_{\text{str}}^{\text{bg}}(x)$ are the novelty bins of the sequence/structure axes when these are disabled as evaluation axes. When every available axis is selected, no background dimensions remain and the reference is the whole pool.

The reference distribution is obtained by balancing the background-profile distribution of the evaluated stratum. In population form, this corresponds to

$$Q_{t,\ell} = \sum_h P_{t,\ell}(H_t = h)\, P_{\mathcal{R}_t}\left(\cdot \mid H_t = h\right), \tag{12}$$

where $P_{t,\ell}$ is the conditional distribution induced by $\mathcal{E}_{t,\ell}$, and $P_{\mathcal{R}_t}$ is the distribution induced by the Auto pool. Empirically, this balanced reference is implemented by stratified sampling:

$$\mathcal{B}_{t,\ell} = \bigcup_h \mathrm{Sample}\left(\mathcal{R}_{t,h}, |\mathcal{E}_{t,\ell,h}|\right), \tag{13}$$

where

$$\mathcal{E}_{t,\ell,h} = \{x \in \mathcal{E}_{t,\ell} : H_t(x) = h\}, \qquad \mathcal{R}_{t,h} = \{x \in \mathcal{R}_t : H_t(x) = h\}. \tag{14}$$

For each background profile h present in the stratum, an equal number of controls is drawn from pool proteins sharing that exact profile, so the only systematic difference between a stratum and its reference lies on the evaluated axes. Sampling is without replacement; profile groups whose candidate set is no larger than the request are taken in full, and groups with no matching candidates in the pool fall back to sampling from the whole pool.

Monte Carlo averaging. A single matched draw is noisy when profile cells are small. We therefore draw $K = 50$ independent matched references per stratum, average the reference metrics of both the model and the null predictor over the K draws, and compute the retention in Eq. 8 once from the averaged metrics. This approximates the expectation under the balanced reference distribution in Eq. 12 and removes single-draw Monte Carlo noise.

B Task and split construction

B.1 Task definitions and data sources

POOR comprises ten protein understanding tasks spanning three prediction paradigms—protein-level classification, protein-level regression, and per-residue prediction. The protein function annotation task is distributed over three independent datasets covering the molecular function (GO-MF), cellular component (GO-CC), and biological process (GO-BP) aspects of the Gene Ontology (Ashburner et al., 2000), yielding twelve task configurations in total. Every sample pairs an amino-acid sequence with a 3D structure file, so that sequence-based and structure-based models are evaluated on identical splits. Table 3 defines the prediction unit, input, output, and data source of each task; Table 5 reports split sizes and sequence-length profiles, and Table 8 details the provenance of the associated structures (experimental versus AlphaFold-predicted (Varadi et al., 2022)).

Several task definitions warrant emphasis. EC and GO are both chain-level multi-label annotation problems derived from SIFTS residue-level PDB–UniProt mappings, with one representative chain retained per UniProt entry; EC labels are truncated at hierarchical level 3 to avoid parent–child leakage along the EC hierarchy. For LBS, a residue is labeled as binding if it lies within 4.5 Å of any ligand heavy atom, and each chain–ligand pair constitutes one sample. For PPI, each chain is represented by its unbound (monomer) structure, preferring experimental apo structures, then AlphaFold models, with holo monomers as a last resort. For PPIS, the interface labels of the same chain observed across multiple complexes are unioned into a single chain-centric sample. For SSP, only high-quality X-ray chains (resolution ≤ 2.5 Å, R-free ≤ 0.25) are retained to ensure reliable hydrogen-bond-based DSSP assignments. For pH, training and validation structures are AlphaFold models while test structures are predominantly experimental, making this split simultaneously a provenance-shift evaluation (Table 8).

B.2 Split construction protocol

All tasks are built with a common three-stage protocol: (i) curation and quality control, (ii) redundancy reduction, and (iii) task-specific splitting with explicit leakage guards. Table 4 summarizes the per-task split rules, anti-leakage guards, and quality-control checks; final split sizes are reported in Table 5.

Table 3: **Task definitions and data sources.** Level is the prediction granularity. All samples pair an amino-acid sequence with a structure file (experimental PDB structures unless noted as AlphaFold/AF). Label counts refer to the training label space; GO class counts are given for the MF/CC/BP datasets respectively. Split sizes and value ranges are reported in Table 5.

Task	Level	Input	Output (labels)	Data source (structures)
Fold	domain	sequence + structure	CATH class–architecture–topology (1,163 classes)	CATH v4.4 S95 (Sillitoe et al., 2021) (PDB (Berman et al., 2000), exp.)
EC	chain	sequence + structure	EC number (level 3), multi-label (260 classes)	SIFTS (Dana et al., 2019)–UniProt (The UniProt Consortium, 2023) (PDB, exp.)
GO-MF/CC/BP	chain	sequence + structure	GO terms, multi-label (447/264/884 classes)	SIFTS–UniProt (PDB, exp.)
PPI	pair	sequence + structure of both chains	binary interaction	PINDER 2024-02 (Kovtun et al., 2024) (apo exp. / AlphaFold monomers)
PPIS	residue	sequence + structure	binary interface label	DIPS-Plus (Morehead et al., 2023) (PDB, exp.)
SSP	residue	sequence + structure	DSSP 9-state (Kabsch & Sander, 1983)	RCSB PDB, X-ray ≤ 2.5 Å (exp.)
k_{cat}	protein	seq. + struct. + reactant SMILES/ECFP4	$\log_{10} k_{\mathrm{cat}}$	CatPred-DB (Boorla & Maranas, 2025) (BRENDA (Chang et al., 2021)/SABIO-RK (Wittig et al., 2012); AlphaFold v6)
pH	chain	sequence + structure	optimal pH	EnzyBase12k (SinhaRoy et al., 2026) (AF train/val, PDB test)
LBA	complex	(multi-chain) struct. + ligand SMILES/ECFP4	$\mathrm{pK} = -\log_{10} K$ (K_d/K_i)	PDBbind v2020.R1 (Wang et al., 2004) (exp.)
LBS	residue	chain struct. + ligand SMILES/ECFP4	binary binding-site label	RCSB PDB mirror (exp.)

Table 4: **Split construction protocols.** "Temporal" splits order samples by PDB deposition/release date so that the test set strictly postdates training. Guards are anti-leakage rules applied on top of the base split.

Task	Redundancy control	Base split (ratio)	Additional guards
Fold	CATH S95 (pre-filtered)	topology-stratified random (70:10:20)	EC7 (L1) domains reserved to test pre-split
EC, GO	95% seq-id clustering	temporal (72:8:20 / 63:7:30)	GO test-label evidence filter; label-frequency filter
k_{cat}	UniProt-disjoint	random (70:10:20)	reactant Tanimoto guard (≥ 0.6)
pH	95% seq-id analysis	provenance-based (PDB → test; AF → train/val, 9:1)	—
LBA	ligand Tanimoto clustering (≥ 0.6)	within-cluster temporal (70:10:20)	dissimilar-ligand fallback to train
LBS	ligand-wise 95% seq-id	temporal (70:10:20)	val/test ligand Tanimoto filter (> 0.6)
PPI	pair-level 95% seq-id	temporal, PDB-disjoint (70:10:20)	per-split 1:1 negative resampling
PPIS	95% seq-id, 80% coverage	random (72:8:20)	—
SSP	95% seq-id clustering	temporal (72:8:20)	—

Quality control. Three checks are applied benchmark-wide. *Structure–sequence consistency:* the sequence extracted from each structure file is aligned to the annotated sequence by global alignment with edlib (Šošić & Šikić, 2017); mismatched entries are corrected by re-extraction (LBA and LBS reach 100% agreement), filtered at an identity threshold (0.90 for EC/GO; 0.95 for SSP), or, for pH, retained with quantified agreement (median identity 0.96) as a documented limitation. *Residue-level label alignment:* for the per-residue tasks (LBS, PPIS, SSP), labels are re-mapped onto the structure-derived sequence by global alignment, unresolved or unannotated residues are masked with -1, and $\mathrm{len}(\text{labels}) = \mathrm{len}(\text{sequence})$ is enforced for every sample. *Sequence-legality screening:* modified residues are mapped to their standard parents (e.g., MSE → M, SEC → C), and non-standard symbols (X/B/Z/J/U) are removed before disorder prediction and similarity search.

Redundancy reduction. Training-pool redundancy is reduced with MMseqs2 clustering at 95% sequence identity (80% coverage) for the chain-level tasks (EC/GO, PPIS, PPI, SSP), and in a ligand-wise manner for LBS so that the same chain bound to different ligands is preserved. Fold relies on the CATH S95 list, which is already non-redundant at 95% identity, while k_{cat} and LBA instead enforce protein-level disjointness and ligand-similarity guards, respectively. Representative selection prioritizes boundary-precise structure extraction, reviewed UniProt status, resolution, and annotation richness.

Splitting with leakage guards. Split rules are task-specific and chosen to reflect realistic deployment. EC/GO, LBS, PPI, and SSP use temporal splits by PDB deposition or release date. Fold uses stratified random splitting at the topology level, with all EC7-carrying domains reserved exclusively for testing before the split and superfamily holdouts annotated post-split. PPIS uses random splitting after 95%-identity deduplication, and k_{cat} splits at the UniProt level so that no protein crosses splits. LBA clusters complexes by complete-linkage ligand Tanimoto similarity (≥ 0.6) and splits temporally within clusters. pH splits by structure provenance, assigning all experimental structures to test and AlphaFold models to train/validation, which makes the pH test set simultaneously a provenance-shift evaluation. Two further chemical-similarity guards apply to ligand-aware tasks: for k_{cat} and LBA, validation/test samples whose reactants or ligands are too dissimilar from training (maximum Tanimoto < 0.6) are moved back into training rather than left unevaluable, whereas LBS filters such validation/test ligands out. For PPI, negative pairs are resampled separately within each split from that split's own protein pool at a 1:1 ratio, excluding all known BIO/XTAL pairs, which prevents cross-split leakage through negative examples.

Extreme-length handling and sanity checks. Chains longer than 2,000 residues are removed outright. Proteins of 60–1,000 residues form the normal pool; shorter or longer proteins are relocated to the test set as the ExtremeShort/ExtremeLong scenarios, except in Fold and LBA, where they remain in place and are flagged, and in SSP and PPI, where only extreme-long chains/pairs are relocated while extreme-short ones are split normally and flagged in place. All splits are verified to be pairwise disjoint at the sample-identifier level (and at the PDB level for structure-derived tasks).

Pair- and complex-level OOD aggregation. Two tasks annotate samples that comprise more than one chain, and their scenario flags are aggregated as follows. On PPI (pair-level), the sequence-homology, structural-similarity, Orphan, IDR, and length axes use *any-chain* semantics: a pair is flagged whenever either chain triggers the criterion—for example, a pair is low-homology if the maximal training identity of *either* chain falls below $X\%$, orphan if either chain has no significant training hit, and extreme-long if either chain exceeds 1,000 residues. Functional axes instead *pool* the two chains: EC numbers (NewFunction, LongTailFunction) and CATH/TED labels (Fold/SuperfamilyHoldout) are unioned across both chains, and the all-not-in and holdout criteria are applied to the union. On LBA (complex-level), every axis is computed on the whole complex: the multi-chain sequences are concatenated for homology search, structural novelty is measured on the intact complex with Foldseek-Multimer using the symmetric mean of the two directional TM-scores (Eq. 3), length and disorder are computed over the full multi-chain protein, and functional annotations are pooled over all chains.

C Dataset statistics and quality details

C.1 Dataset scale overview

Table 5 summarizes the scale and basic properties of each task in POOR. The benchmark contains 429,645 protein samples in total, split into 303,516 training, 37,642 validation, and 88,487 test instances. Tasks span three prediction paradigms: classification (fold, EC, GO, PPI), regression (enzyme kinetics, optimal pH, binding affinity), and per-residue prediction (binding site, protein-protein interface, secondary structure). Sequence lengths vary substantially across tasks, from median 133 aa in fold classification to 379 aa in enzyme kinetics. Four tasks (Fold, PPI, LBA, SSP) retain extreme-length proteins (< 60 or >1000 aa) in their training/validation splits—for Fold and SSP these are exclusively short (<60 aa) chains—which ensures that models encounter length distributional shifts at test time rather than extrapolating from uniformly short training data; for the remain-

ing tasks, extreme-length proteins appear only in the test set, while train/validation stay within the normal 60–1,000 aa range.

Table 5: **Dataset scale overview.** POOR comprises 429,645 samples across 10 tasks. CLS=classification, REG=regression, PER=per-residue. Median sequence lengths (min–max in aa) are shown. "Extreme" indicates whether train/val contain proteins $<$60 or $>$1000 aa.

Task	Type	Train	Val	Test	Total	Labels/Range	Median Len.
Fold (CATH)	CLS	43,473	6,132	12,911	62,516	1,163 classes	133 (9–960)
EC	CLS	15,389	1,687	5,047	22,123	260 classes	298 (3–1,997)
GO-MF	CLS	18,179	2,084	5,609	25,872	447 classes	236 (3–1,997)
GO-CC	CLS	13,878	1,730	5,613	21,221	264 classes	222 (3–1,997)
GO-BP	CLS	15,232	1,742	5,142	22,116	884 classes	228 (3–1,997)
PPI	CLS	58,968	8,394	16,264	83,626	pos/neg	280 (3–1,997)
k_{cat}	REG	15,479	2,150	5,002	22,631	$[-6.00, 6.00]$	379 (16–1,928)
pH	REG	7,794	866	2,955	11,615	$[1.50, 12.50]$	361 (23–1,476)
LBA	REG	8,366	930	2,367	11,663	$[0.40, 15.22]$	231 (1–1,504)
LBS	PER	71,093	7,938	17,225	96,256	binary	291 (4–1,991)
PPIS	PER	6,043	672	1,980	8,695	binary	220 (7–1,484)
SSP	PER	29,622	3,317	8,372	41,311	9 classes	222 (31–1,836)
Total		**303,516**	**37,642**	**88,487**	**429,645**		

C.2 OOD SCENARIO COVERAGE MATRIX

Table 6 reports the proportion of test-set samples falling into each OOD scenario, per task. Coverage is defined as the fraction of the full test set (including the Default/ID subset) that satisfies the OOD criterion. A dash ($-$) indicates that the scenario is not applicable to the task. The matrix reveals a highly heterogeneous coverage landscape. LBA is the least sequence-diverged task (24.3% of its test set below 90% identity, 10.2% below 30%), whereas PPIS, SSP, and PPI are the most remote at the strict rungs (30.4%, 24.1%, and 22.5% below 30%); Orphan proteins are most frequent in PPIS (23.1%). Structural novelty is densest in PPI (50.2% below TM 0.9, 23.2% below TM 0.7) and LBA (18.9% below TM 0.7), and sparsest in k_{cat} (2.1% below TM 0.7). Functional novelty concentrates where enzyme annotations are central: NewEC L4 affects 21.9% of the pH and 17.1% of the k_{cat} test sets, long-tail EC L4 labels ($\leq$10 training occurrences) cover 74.7% of the pH test set, and unseen label combinations reach 15.7% in GO-BP. CATH+TED-derived fold and superfamily holdouts are annotated on all tasks and are most populated in PPIS (3.3%/11.9%) and pH (3.1%/10.5%). The purely in-distribution residue varies widely: the InD subset (no flag on any axis) spans 1.7% (pH) to 46.1% (LBA) of the test set.

C.3 LABEL LONG-TAIL DISTRIBUTION

Table 7 quantifies the degree of class imbalance in POOR's classification and per-residue tasks. Label counts are computed on the training label space and match the class counts in Table 3. The multi-label GO tasks are markedly skewed: the 20% most frequent terms cover 77.9%/82.6%/68.8% of GO-MF/CC/BP training label assignments, and their Zipf exponents α (estimated from $\log f \propto -\alpha \log r$) are shallow ($-1.28$/$-1.31$/$-1.00$), indicating heavy-tailed distributions. EC spans 260 level-3 classes with 7.7% singleton labels and a steeper Zipf slope (-1.60); after re-splitting at the class–architecture–topology granularity, fold classification spans 1,163 classes with 14.0% singletons, the heaviest singleton tail among the classification tasks. The per-residue task PPIS has 22.8% positive (interface) residues in training, an imbalance of roughly 1:3.4 that reflects the biological rarity of interface residues and necessitates threshold-robust evaluation metrics (e.g., AUPR, MCC) rather than raw accuracy. Regression tasks (bottom panel) have approximately balanced, unimodal label distributions with comparable moments between train and test, indicating that label shift is not a confounding factor for these tasks.

C.4 PROTEIN STRUCTURE SOURCE PROPORTION

Table 8 breaks down the origin of structural annotations in each split. Overall, 74.9% of POOR samples are associated with experimentally determined structures (X-ray, NMR, cryo-EM) and 25.1%

Table 6: **OOD scenario coverage matrix.** Each cell shows the percentage of the full test set belonging to the corresponding OOD scenario. A dash (−) means the scenario is not applicable to the task. Sequence-identity, TM-score, and Orphan flags are computed within the Default subset; samples set aside before splitting (Default = False, e.g., relocated extreme-length proteins) are uncomputed (NaN) on these axes and counted as not belonging. NewEC and LongTailEC follow all-not-in semantics against the training reference with frequency thresholds including 0. BP=GO-BP, CC=GO-CC, MF=GO-MF.

OOD Scenario	k_{cat}	pH	Fold	EC	BP	CC	MF	LBA	LBS	PPI	PPIS	SSP
Test set size	5,002	2,955	12,911	5,047	5,142	5,613	5,609	2,367	17,225	16,264	1,980	8,372
Default (ID) subset	88.2%	91.9%	100.0%	84.1%	80.1%	75.8%	82.2%	100.0%	79.7%	91.9%	85.1%	99.0%
InD subset (no OOD flag)	7.3%	1.7%	14.0%	3.0%	3.3%	3.4%	3.5%	46.1%	30.2%	12.0%	6.8%	13.4%
Extreme short (<60 aa)	1.2%	0.9%	5.9%	9.5%	14.7%	19.4%	13.4%	0.6%	11.5%	4.4%	13.2%	3.5%
Extreme long (>1000 aa)	10.6%	7.1%	–	6.5%	5.1%	4.8%	4.4%	12.8%	8.9%	8.1%	1.7%	1.0%
SeqID <90%	78.7%	86.0%	79.4%	79.0%	74.6%	70.3%	76.7%	24.3%	45.9%	65.5%	73.6%	82.3%
SeqID <80%	68.3%	81.1%	70.6%	71.9%	66.8%	62.6%	68.7%	21.0%	41.3%	60.6%	68.6%	73.3%
SeqID <70%	61.2%	75.4%	62.8%	65.7%	59.2%	56.1%	61.2%	19.2%	37.3%	55.5%	65.6%	66.3%
SeqID <60%	52.9%	68.5%	54.6%	57.1%	50.5%	49.3%	52.4%	17.3%	32.2%	49.8%	60.7%	60.1%
SeqID <50%	42.9%	56.6%	43.3%	44.5%	41.1%	41.8%	42.3%	15.5%	26.5%	42.9%	53.8%	51.7%
SeqID <40%	24.4%	40.5%	30.0%	28.7%	31.1%	33.3%	30.9%	12.4%	18.5%	33.8%	44.4%	39.2%
SeqID <30%	7.7%	21.8%	18.2%	14.4%	20.6%	24.0%	20.0%	10.2%	10.8%	22.5%	30.4%	24.1%
Orphan (no hit)	3.5%	14.3%	15.3%	9.9%	16.2%	18.9%	16.1%	7.6%	7.6%	18.6%	23.1%	18.7%
TM <0.9	19.1%	31.3%	31.2%	27.0%	33.3%	34.9%	32.8%	32.4%	17.5%	50.2%	38.9%	34.2%
TM <0.8	6.8%	13.9%	13.6%	11.4%	17.5%	20.6%	16.2%	28.0%	7.6%	32.9%	21.5%	16.2%
TM <0.7	2.1%	5.8%	6.5%	5.9%	9.9%	12.8%	9.4%	18.9%	4.0%	23.2%	11.3%	8.3%
TM <0.6	0.3%	1.4%	3.8%	3.0%	6.2%	7.8%	5.6%	12.9%	2.2%	17.5%	5.0%	4.1%
TM <0.5	0.1%	0.3%	2.7%	2.1%	4.2%	5.3%	3.6%	8.5%	1.4%	13.4%	2.4%	2.2%
TM <0.4	0.1%	0.1%	2.1%	1.1%	2.9%	3.6%	2.5%	5.7%	1.1%	9.3%	1.4%	1.5%
TM <0.3	0.0%	0.1%	1.8%	0.6%	1.9%	2.6%	1.6%	3.3%	0.8%	5.8%	1.0%	1.2%
IDR-containing	1.0%	1.3%	3.1%	0.5%	3.1%	3.1%	3.0%	3.9%	0.8%	23.0%	1.4%	1.9%
Fold holdout	0.9%	3.1%	–	0.6%	1.0%	1.0%	0.6%	1.6%	0.1%	0.3%	3.3%	0.2%
Superfamily holdout	3.2%	10.5%	5.9%	2.0%	2.9%	3.3%	3.3%	3.1%	0.8%	1.6%	11.9%	1.1%
NewEC L3	0.1%	0.7%	–	–	0.2%	0.2%	0.0%	0.3%	0.1%	0.1%	0.5%	0.0%
NewEC L4	17.1%	21.9%	–	7.5%	2.3%	2.4%	3.3%	1.9%	1.6%	1.2%	7.3%	2.9%
NewEC L1	–	–	2.6%	–	–	–	–	–	–	–	–	–
Long-tail EC L3 ≤5	0.6%	4.0%	2.9%	–	1.1%	0.8%	0.6%	1.0%	0.2%	0.1%	3.1%	0.4%
Long-tail EC L3 ≤10	1.5%	7.5%	3.2%	–	1.6%	1.7%	1.9%	1.6%	0.4%	0.4%	6.9%	0.9%
Long-tail EC L4 ≤5	30.5%	59.7%	11.3%	22.9%	7.2%	7.7%	8.6%	8.0%	4.7%	2.7%	26.0%	8.8%
Long-tail EC L4 ≤10	43.2%	74.7%	17.0%	34.7%	9.9%	9.8%	13.1%	14.5%	6.6%	4.3%	33.3%	13.4%
Long-tail (label-freq)	–	–	8.4%	3.2%	13.3%	4.2%	5.3%	–	1.9%	–	2.4%	5.2%
Long-tail bin ≤50	0.1%	0.8%	–	–	–	–	–	0.8%	–	–	–	–
Long-tail bin ≤100	2.7%	2.0%	–	–	–	–	–	2.7%	–	–	–	–
Combinatorial	–	–	–	1.1%	15.7%	8.4%	8.0%	–	–	–	–	–

with AlphaFold-predicted models. This balance is heavily task-dependent: the k_{cat} and PPI tasks are almost entirely prediction-based (100% and 96.5% predicted in test, respectively), because their source databases (UniProt-based enzyme kinetics, PINDER interaction models) adopted AlphaFold as the reference structure. Conversely, fold classification (CATH), all GO/EC function tasks, LBA, LBS, PPIS, and SSP use exclusively or predominantly experimental structures from RCSB PDB and curated subsets (PDBbind, DIPS-PLUS). The pH task is a notable hybrid: its training and validation splits are 100% AlphaFold-derived, but its test set is 93.0% experimental, drawn from the manually curated Enzyme Optimum pH database. This structural provenance heterogeneity is important for interpreting benchmark results: performance drops on k_{cat}/PPI may partly reflect model sensitivity to predicted-structure artifacts, whereas drops on CATH/LBA more directly indicate failures of generalization against biological novelty.

D Training configuration and task-specific prediction heads

Pretrained protein language models. We evaluated three ESM2 variants (35M, 150M, and 650M parameters) and SaProt-650M using the same adapter-tuning protocol. The three ESM2 encoders contain 12, 30, and 33 Transformer layers, respectively, with residue embedding dimensions of 480, 640, and 1280. SaProt-650M contains 33 Transformer layers with a hidden dimension of 1280. ESM2 received the amino-acid sequence as input, whereas SaProt used structure-aware tokens

Table 7: **Label long-tail distribution.** Label counts are computed on the training label space and match the class counts in Table 3. Head 20% coverage = proportion of samples covered by the 20% most frequent labels. Tail 50% coverage = proportion covered by the 50% least frequent labels. Singletons = labels appearing exactly once in training. Zipf α estimated from rank–frequency plot.

Task	#Labels	Head 20%	Tail 50%	Singletons (%)	Zipf α
EC	260	75.4%	4.9%	7.7%	-1.596
GO-MF	447	77.9%	7.0%	0.0%	-1.277
GO-CC	264	82.6%	5.7%	0.0%	-1.308
GO-BP	884	68.8%	11.9%	0.0%	-0.995
SSP	9	53.6%	7.7%	0.0%	-1.763
Fold	1,163	86.6%	3.8%	14.0%	-1.445

Binary task class imbalance

Task	Train Pos%	Test Pos%	Type
PPI	50.0%	50.0%	Pair-level
PPIS	22.8%	22.9%	Per-residue

Regression label distribution

Task	Train Mean±Std	Test Mean±Std	Range
LBA	6.26 ± 1.97	6.11 ± 1.80	$[0.40, 15.22]$
k_{cat}	0.88 ± 1.67	0.93 ± 1.62	$[-6.00, 6.00]$
pH	7.13 ± 1.28	7.39 ± 1.08	$[1.50, 12.50]$

Table 8: **Experimental versus predicted structure proportion.** For each task and split, we report the number of samples and the percentage of experimentally determined structures (Exp) versus AlphaFold-predicted structures (Pred). PPI is evaluated at pair-level: a pair is predicted if either chain is predicted. BP=GO-BP, CC=GO-CC, MF=GO-MF.

	Train		Val		Test		
Task	N	Exp% / Pred%	N	Exp% / Pred%	N	Exp% / Pred%	Source
k_{cat}	15,479	0.0 / 100.0	2,150	0.0 / 100.0	5,002	0.0 / 100.0	AlphaFold v6 (UniProt)
pH	7,794	0.0 / 100.0	866	0.0 / 100.0	2,955	93.0 / 7.0	AF v6 + RCSB PDB
Fold	43,473	100.0 / 0.0	6,132	100.0 / 0.0	12,911	100.0 / 0.0	RCSB PDB (CATH S95)
EC	15,389	100.0 / 0.0	1,687	100.0 / 0.0	5,047	100.0 / 0.0	RCSB PDB (SIFTS)
GO-BP	15,232	100.0 / 0.0	1,742	100.0 / 0.0	5,142	100.0 / 0.0	RCSB PDB (SIFTS)
GO-CC	13,878	100.0 / 0.0	1,730	100.0 / 0.0	5,613	100.0 / 0.0	RCSB PDB (SIFTS)
GO-MF	18,179	100.0 / 0.0	2,084	100.0 / 0.0	5,609	100.0 / 0.0	RCSB PDB (SIFTS)
LBA	8,366	100.0 / 0.0	930	100.0 / 0.0	2,367	100.0 / 0.0	PDBbind v2020.R1
LBS	71,093	100.0 / 0.0	7,938	100.0 / 0.0	17,225	100.0 / 0.0	RCSB PDB local mirror
PPI	58,968	10.7 / 89.3	8,394	4.8 / 95.2	16,264	3.5 / 96.5	PINDER (pair-level)
PPIS	6,043	100.0 / 0.0	672	100.0 / 0.0	1,980	100.0 / 0.0	DIPS-PLUS (RCSB PDB)
SSP	29,622	100.0 / 0.0	3,317	100.0 / 0.0	8,372	100.0 / 0.0	RCSB PDB (X-ray)

Aggregate

Category	Samples	Proportion
Experimental	321,792	74.9%
Predicted	107,853	25.1%
Total	429,645	100.0%

obtained by pairing each amino acid with its Foldseek 3Di state. During downstream training, all pretrained backbone parameters were frozen. Residue representations were extracted under gradient tracking disabled, cached, and used as fixed input features; therefore, these runs did not perform full-model fine-tuning. Although the command-line configuration retains a PEFT-type field, it is inactive in adapter mode; consequently, neither LoRA nor AdaLoRA was applied in the reported experiments.

For every pretrained backbone, the residue embeddings were first projected to 480 dimensions and then passed through a randomly initialized six-layer Transformer adapter. Each adapter layer used 20 attention heads, a 1920-dimensional feed-forward sublayer, GELU activation, and dropout 0.1. A LayerNorm operation was applied to the adapter output. Sequence-level tasks used masked mean pooling over valid residue positions, followed by a linear prediction head. Residue-level tasks applied the linear head independently to each valid residue. For ligand-conditioned tasks, a 2048-bit radius-2 Morgan fingerprint (ECFP4) was linearly projected to 480 dimensions and fused as an ad-

Table 9: Task-specific heads and optimization objectives shared by the adapter and GVP pipelines. Spearman correlation is used to evaluate regression tasks and to compute ProDPR, but it is not used as the training loss.

Tasks	Prediction level	Output head	Training objective	Evaluation metric
EC, GO-BP/CC/MF	protein, multi-label	linear, C logits	weighted BCE (adapter); BCE (GVP)	$F_{\max}$
k_{cat}, affinity, optimal pH	protein, scalar	linear, one output	MSE	Spearman ρ
Ligand-binding site, PPIS	residue, binary	linear, one logit	binary cross-entropy	residue AUROC
PPI	protein pair, binary	linear, one logit	binary cross-entropy	AUROC
Fold	protein, multi-class	linear, C logits	cross-entropy	accuracy
SSP	residue, multi-class	linear, C logits	cross-entropy	accuracy

ditional token. For protein–protein interaction prediction, the residue representations of the two proteins were concatenated with a separator token, processed jointly by the adapter, and mean-pooled before classification.

The adapter and task head were optimized with AdamW using a learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$, and zero weight decay. We used a step-wise cosine-annealing schedule over the maximum 50 epochs, bfloat16 training, and gradient-norm clipping at 1.0. Validation was performed after every epoch, and early stopping used a patience of five epochs. For EC and GO prediction, checkpoints were selected by maximum validation $F_{\max}$; for PPI prediction, they were selected by maximum validation AUROC; all regression and residue-level tasks used minimum validation loss. The reported ESM2 and SaProt results aggregate five independently trained runs with random seeds $\{2024, 2025, 2026, 2027, 2028\}$.

GVP baseline trained from scratch. The GVP model was randomly initialized for every task and trained end-to-end; no pretrained sequence or structure encoder was used. Each protein structure was represented as a residue graph constructed from backbone $\{\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{O}\}$ coordinates with the 30 nearest neighbors per residue. Node scalar features contained a 128-dimensional learned amino-acid embedding and six backbone-dihedral features. Node vector features contained the virtual side-chain direction and the forward and backward backbone orientations. Edge scalar features consisted of 16 radial-basis encodings of inter-residue distance and 16 positional features, while the normalized inter-residue direction formed a single edge vector channel. These inputs were projected to node dimensions $(128, 16)$ and edge dimensions $(128, 1)$ for scalar and vector channels, respectively, and processed by three GVP message-passing layers with dropout 0.1.

For ordinary protein-level tasks, mean pooling over residue scalar states was followed by LayerNorm, dropout 0.1, and a linear output layer. Residue-level tasks applied the same head at every residue. For PPI, the two mean-pooled protein representations h_A and h_B were combined using the symmetric feature

$$[h_A + h_B,\ |h_A - h_B|,\ h_A \odot h_B]\,, \tag{15}$$

which was passed to the binary-classification head. For ligand-conditioned tasks, the 2048-bit ECFP4 fingerprint was transformed by a Linear–GELU–LayerNorm block to 128 dimensions. The resulting ligand vector was concatenated either with the mean-pooled protein representation for affinity and k_{cat} prediction or with every residue representation for ligand-binding-site prediction. Depending on the task head, the complete GVP model contained approximately 0.775–1.038 million trainable parameters.

GVP was optimized in float32 with AdamW at a learning rate of 10^{-4} and a weight decay of 10^{-5}; no learning-rate scheduler was used. Gradients were clipped to a norm of 1.0. The maximum protein length was 768 residues and the maximum number of epochs was 50. The EC and GO-BP tasks used a batch size of 8, whereas all remaining tasks used a batch size of 2. Training used the training split only, checkpoint selection minimized validation loss, and early stopping used a patience of five epochs with a minimum improvement of 10^{-4}. The reported GVP results use seed 2024.

For multi-label EC and GO tasks in the adapter pipeline, the binary cross-entropy loss used a training-set-derived positive-class weight $w_c = N_c^- / \max(N_c^+, 1)$, clipped to the interval $[1, 100]$. GVP used unweighted binary cross-entropy for these tasks. Except for the stated GVP GO-MF configuration, model selection used the validation split before evaluating the held-out test set. Changing the regression component of ProDPR from MAE to Spearman correlation affected only post-training evaluation; it did not alter the checkpoints, training objectives, or optimization settings described above.

E NEGATIVE-PAIR COMPOSITION EXPLAINS SUPER-PARITY RETENTION ON PPI

On PPI, all five models exceed reference parity (ProDPR 1.109–1.465; Section 4.3), and several OOD strata attain AUROC values above the InD reference. We trace this counter-intuitive pattern to the composition of the negative class rather than to genuine robustness.

Two negative types of unequal difficulty. PPI negatives are of two kinds: *random* pairs ($4{,}185$; proteins paired at random) and *xtal* pairs ($3{,}947$; crystal-packing contacts that physically touch in the crystal lattice but are not biological interactions). Xtal negatives share the geometric and physicochemical signature of true interfaces and are far harder to reject: for a representative model (ESM-2-35M with a linear probe trained additionally for this analysis, distinct from the adapter pipeline of Appendix D), per-stratum AUROC against random negatives alone reaches 0.61–0.72, versus only 0.52–0.58 against xtal negatives alone.

Hard negatives concentrate in the InD subset. Because xtal pairs are annotated from crystal structures of well-studied proteins, they concentrate on in-distribution proteins: 66.0% of InD negatives are xtal, versus 21.5–50.3% on the OOD strata (Table 10). Stratum-level AUROC therefore conflates shift sensitivity with the negative-type mix, and OOD strata with random-heavy negatives appear artificially strong.

Table 10: **PPI AUROC decomposed by negative-pair type.** Results for a representative model (ESM-2-35M linear probe, best validation checkpoint) on the PPI test set. "xtal (%)" is the share of crystal-packing negatives among all negatives of the stratum. The xtal-only column isolates the hard-negative regime; the New EC L4 and Long-tail xtal cells rest on 23 and 62 negatives, respectively, and are noisy.

Stratum	n_{neg}	xtal (%)	AUROC (all)	AUROC (xtal)	AUROC (random)
InD	756	66.0	0.603	0.580	0.645
IDR-containing	2,031	50.3	0.600	0.537	0.663
Extreme short	386	40.2	0.638	0.518	0.718
Extreme long	732	44.3	0.574	0.528	0.611
SeqID <30%	2,022	45.1	0.593	0.549	0.629
NewEC L4	107	21.5	0.581	0.528	0.596
Long-tail EC L4 $\leq$5	236	26.3	0.613	0.615	0.613

Controlling for negative type reverses the picture. With evaluation restricted to xtal negatives, InD outperforms every OOD stratum except the noisy long-tail cell (AUROC 0.580 vs. 0.518–0.549; the long-tail stratum reaches 0.615 on only 62 xtal negatives)—a genuine shift-induced degradation of 5–11% that the mixed-negative AUROC hides; consistently, the bio–xtal score gap shrinks from 0.028 on InD to 0.007–0.013 on OOD strata. Random negatives show the opposite trend: they are *easier* on OOD strata (e.g., AUROC 0.718 on extreme-short proteins vs. 0.645 on InD), because novel proteins rarely participate in recorded interactions, so novelty itself acts as a negative cue—the bio–random score gap widens from 0.054 on InD to 0.152 on extreme-short proteins.

Implication. For pair-level tasks, the negative-pair type is a confound that is not part of the matching profile $H_t(x)$ (Eq. 11). We report headline PPI numbers under the standard mixed-negative protocol for comparability across tasks, but retention on PPI should be read together with Table 10; stratifying AUROC by negative type—or adding the negative type to the matching profile—yields the more conservative, degradation-consistent view.