SCALING WORSENS REPRESENTATION COLLAPSE IN INTRINSICALLY DISORDERED REGIONS: EVOLUTION-GUIDED CONTRASTIVE LEARNING AND LIGAND-DOMINANT SHORTCUTS FOR ZERO-SHOT PAN-CANCER IDR-LIGAND RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning discriminative representations of intrinsically disordered regions (IDRs) remains a fundamental challenge for protein representation learning. Although IDR targets are commonly assumed to benefit from protein-aware representations, their low-complexity sequences may fundamentally limit the discriminative information available to protein language models. Here, we identify systematic representation collapse across amino-acid composition, ProtBERT, and ESM-2 embeddings on 543 pan-cancer IDR targets. Surprisingly, scaling does not alleviate this problem: increasing ESM-2 from 35M to 3B parameters raises mean pairwise cosine similarity from 0.870 to 0.944, indicating that larger models capture shared IDR sequence statistics more precisely without learning more discriminative representations. We introduce Pan-IDR, a contrastive learning framework combining evolution-guided aggregation (EvoPool), functional attention, and dispersion regularization. Under 5-fold zero-shot evaluation, Pan-IDR achieves a distant-target hit@1 of 0.339, compared with a 0.200 random baseline. Ablations further show that EvoPool is the only component providing a substantial improvement, suggesting that downstream objectives alone cannot recover discriminative information from collapsed representations. More strikingly, a ligand-only ECFP baseline using no protein information outperforms Pan-IDR, achieving distant-target hit@1 of 0.417 versus 0.339. This exposes a ligand-dominant shortcut in which ligand structural similarity provides a stronger predictive signal than learned protein–ligand compatibility, indicating that current IDR affinity data do not support robust protein-aware zero-shot generalization. Together, our results reveal a scaling-resistant limitation of current protein representation learning for IDRs and motivate representations and benchmarks specifically designed for disordered proteins.

1 INTRODUCTION

Intrinsically disordered regions (IDRs) constitute over 30% of the human proteome and play central roles in pan-cancer signaling, transcription, and DNA repair Uversky et al. (2009). IDRs lack stable three-dimensional structures and are enriched in low-complexity sequences (P/E/S/K/Q), rendering structure-based docking inapplicable. Zero-shot molecular generation is thus a promising direction for IDR-targeted drug discovery, but data scarcity Löhr et al. (2025) and insufficient benchmarks Zeng et al. (2026) severely constrain progress. IDRdecoder Shionyu-Mitusyama et al. (2025) predicts ligand interactions for IDRs, but its performance is limited by the representation quality of protein language models..

Recent contrastive learning frameworks for zero-shot molecular generation, including ConGen Koban et al. (2026), ZeroGEN Chen et al. (2025), and CLIPP-SET, assume that protein representations can distinguish different targets and that aligning protein and ligand embeddings in a shared metric space captures protein–ligand pairing. This assumption may not hold for IDRs: their

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.

low-complexity sequences lack the structural constraints that generate distinct sequence signatures in folded proteins, potentially limiting the discriminative power of protein language models Wang et al. (2026); Prabakaran & Bromberg (2026).

Existing representation learning methods for proteins, such as ProtBERT Elnaggar et al. (2022) and ESM-2 Lin et al. (2023), have been widely applied to binding-affinity prediction and protein–ligand pairing, but their behavior on IDRs remains under-explored. Two issues arise. First, IDR low sequence complexity may fundamentally limit the discriminative information available to protein language models, regardless of model scale; indeed, the ESM-2 family plateaus at 650M–3B parameters Catrina et al. (2026), and ESM2-650M often outperforms larger 3B and 15B models on fitness prediction Serrano et al. (2024); Cheng et al. (2024). Second, current IDR affinity data may not support protein-aware zero-shot generalization, as ligand structural similarity can dominate protein–ligand pairing signals.

This work systematically investigates these questions. Representation collapse of pan-cancer IDR targets is quantified across three representations, revealing that collapse is severe and intrinsic to IDR sequence statistics. Surprisingly, scaling the protein language model from 35M to 3B parameters does not alleviate collapse but worsens it, consistent with recent observations that larger PLMs degrade on IDR-related tasks Catrina et al. (2026). A contrastive learning framework with evolution-guided aggregation, termed Pan-IDR NaderiAlizadeh & Singh (2026), is then proposed and evaluated under 5-fold zero-shot splits. While Pan-IDR significantly outperforms random baselines, a ligand-only ECFP baseline Rogers & Hahn (2010) outperforms it on both novel and distant tiers, revealing a ligand-dominant shortcut in current IDR retrieval benchmarks. These findings define the boundary of current representation learning methods on pan-cancer IDRs and motivate IDR-specific representations. We present a systematic diagnosis of representation collapse in protein language models for IDRs, and reveal a ligand-dominant shortcut that undermines current benchmarks. Our contributions are summarized as follows:

- We systematically diagnose representation collapse of pan-cancer IDR targets across three representations (amino-acid composition, ProtBERT, and ESM-2 35M). All three collapse severely (mean cosine 0.78–0.90), with 20–60% of target pairs exceeding 0.9. Hand-crafted features collapse most strongly, indicating that the phenomenon is intrinsic to IDR sequence statistics rather than a model artifact.
- It is demonstrated that representation collapse **worsens monotonically with model scale**: ESM-2 35M (mean cosine 0.870, 40.8% of pairs above 0.9) $\rightarrow$ 650M (0.871, 56.0%) $\rightarrow$ 3B (0.944, 88.4%), with 62.2% of pairs exceeding 0.95 at the 3B scale. This is consistent with recent findings that protein language models plateau or degrade at larger scales Catrina et al. (2026), and suggests that larger models learn the *shared* low-complexity statistics of IDRs more precisely, not more discriminatively.
- A contrastive learning framework, Pan-IDR, is proposed by combining evolution-guided aggregation (EvoPool) NaderiAlizadeh & Singh (2026), functional attention, and dispersion regularization. Under 5-fold zero-shot splits with 5 seeds, Pan-IDR achieves novel hit@1 $= 0.277$ (95% CI $[0.241, 0.314]$) and distant hit@1 $= 0.339$ (95% CI $[0.303, 0.375]$), both significantly above the 5-class random baseline of 0.200. Ablation reveals that **only EvoPool** contributes meaningfully ($\Delta_{\text{distant}} = -0.047$), consistent with collapse being unfixable from within the model.
- A ligand-dominant shortcut is revealed: a ligand-only ECFP baseline Rogers & Hahn (2010) **outperforms Pan-IDR** on both novel (0.335 vs 0.277) and distant (0.417 vs 0.339) tiers. This shows that current IDR affinity data cannot support protein-aware zero-shot generalization and motivates IDR-specific representations and benchmarks.

2 Related Work

2.1 Protein Representation Learning

Protein representation learning aims to map amino-acid sequences to informative embeddings that capture structural, functional, and evolutionary properties. Early approaches relied on hand-crafted features such as amino-acid composition, physicochemical properties, and evolutionary profiles. Recent advances in self-supervised learning have produced general-purpose protein language models,

including ProtBERT Elnaggar et al. (2022) and ESM-2 Lin et al. (2023). These models provide contextualized residue-level embeddings that have been successfully applied to structure prediction, binding-affinity prediction, and protein–ligand pairing. However, recent work has shown that protein language models scale poorly: the ESM-2 family plateaus at 650M–3B parameters on ProteinGym benchmarks Catrina et al. (2026), and ESM2-650M often outperforms larger 3B and 15B models on fitness prediction. Training compute-optimal protein language models also yields marginal returns beyond 650M Cheng et al. (2024); Serrano et al. (2024). On intrinsically disordered regions, parameter-efficient fine-tuning approaches such as LoRA-DR-suite Gianluca et al. (2025) and FusOn-pLM Vincoff et al. (2025) have been applied, but their representations remain underexplored.

2.2 Intrinsically Disordered Regions

IDRs are characterized by low sequence complexity and enrichment in P/E/S/K/Q. DisProt Aspromonte et al. (2024) provides curated annotations for over 3,000 IDRs. Recent work has highlighted the limitations of existing benchmarks for IDR modeling: DisProtBench Zeng et al. (2026) shows that current structure prediction models fail to capture IDR function. Generative approaches such as IDiom Liu et al. (2026) attempt to design IDRs de novo. IDRdecoder Shionyu-Mitusyama et al. (2025) predicts ligand interactions for IDRs. However, the representation quality of protein language models on IDRs has not been systematically quantified.

2.3 Zero-Shot Molecular Generation

Zero-shot molecular generation aims to design candidate ligands for unseen targets without target-specific training data. ConGen Koban et al. (2026) and ZeroGEN Chen et al. (2025) use contrastive learning to align protein and ligand embeddings. DrugCLIP Gao et al. (2023); Jia et al. (2026) achieves strong zero-shot virtual screening performance and is orders of magnitude faster than docking. Structure-free contrastive learning extends this to disordered sites Fondrie et al. (2026). Multimodal frameworks such as ProteinCLIP Wu et al. (2024) and tri-modal contrastive learning further enrich protein representations. PharMistral Bendjeddou et al. (2026) uses large language models to generate ligands for novel targets. However, all these methods assume that protein representations can distinguish different targets, which may fail for IDRs.

2.4 Metric Learning and Ligand Fingerprints

Metric learning aims to learn an embedding space in which distances reflect semantic similarity. Morgan fingerprints (ECFP) Rogers & Hahn (2010) remain a strong baseline for ligand-based virtual screening. EvoPool NaderiAlizadeh & Singh (2026) introduces evolution-guided pooling to prevent representation collapse in protein language models, which is closely related to the EvoPool module in this work. Our results show that on pan-cancer IDR targets, a ligand-only ECFP baseline outperforms a protein-aware contrastive model, revealing a ligand-dominant shortcut.

3 Method

3.1 Problem Setup

We study zero-shot ligand retrieval on pan-cancer intrinsically disordered region (IDR) targets. Formally, let $\mathcal{T}_{\text{train}}$ denote a set of training targets, each associated with a set of ligand–affinity pairs $\{(\ell_i, y_i)\}_{i=1}^{n_t}$, where y_i is the binding affinity of ligand ℓ_i to target t. Given an unseen test target $t^* \notin \mathcal{T}_{\text{train}}$ together with its protein sequence, the goal is to retrieve, from the union of all training ligands, a ranked list of candidates whose binding affinities are consistent with the expected affinity profile of t^*.

To make the retrieval task well-defined and comparable across targets, we discretize binding affinities into five ordered classes: very_weak (< 3), weak (3–5), medium (5–7), strong (7–9), very_strong (≥ 9), where the affinity value is expressed as pAffinity ($-\log_{10}$ Kd/Ki/IC50). A retrieved ligand is considered a hit if its affinity class matches that of the query target. Under this 5-class formulation, the random baseline is approximately 0.20.

We evaluate retrieval performance using two metrics: **hit@1** (the top-1 retrieved ligand belongs to the same affinity class as the query) and **hit@5** (at least one of the top-5 retrieved ligands belongs to the same affinity class as the query). To characterize the effect of target novelty, we further partition test targets into four tiers based on their maximum sequence similarity to the training set: *novel* ($< 30\%$), *distant* (30–70%), *near* (70–90%), and *same_family* ($> 90\%$). This tiered evaluation allows us to delineate the boundary of zero-shot generalization on pan-cancer IDR targets.

3.2 Data Preparation and Feature Extraction

Data Preparation. A total of 3,337 intrinsically disordered proteins (IDPs) were collected from the DisProt database Aspromonte et al. (2024) and mapped to the PDBbind v2020 database Liu et al. (2015), which comprises 19,139 general-set and 5,316 refined-set complexes. Using the PDB–UniProt mapping file (1,031,850 entries, based on SIFTS Velankar et al. (2013)), PDB entries were associated with UniProt IDs and filtered for complexes whose targets are disordered. This yielded 9,088 raw complexes; after deduplication, 4,762 unique PDB complexes remained. The CASF-2016 core set (285 PDBs) Su et al. (2019) was held out as an external test set. The remaining data were randomly split into a training set (3,739) and a validation set (935) at an 8:2 ratio, with 88 core-set samples serving as a separate test set. All binding affinities were converted to pAffinity ($-\log_{10}$ Kd/Ki/IC50), with a mean of 6.37 and a range of 0.40 to 15.22.

Feature Extraction. For each of the 543 unique pan-cancer IDR targets, the full UniProt sequence was retrieved via the UniProt REST API, yielding a FASTA corpus of 543 sequences (length: min 65, median $\sim$300, max 34,350 residues). Three protein representations were then extracted: amino-acid composition (20D), ProtBERT mean-pooled embeddings (1024D) Elnaggar et al. (2022), and ESM-2 35M mean-pooled embeddings (480D) Lin et al. (2023). For ligands, molecular graphs were extracted using RDKit, yielding 1,076 unique ligand graphs. Each graph uses 21-dimensional atom features and 5-dimensional bond features. To support evolution-guided aggregation, homolog sequences for each target were retrieved via MMseqs2 Steinegger & Söding (2017) all-vs-all search at 30% sequence identity, and up to ten top-scoring homologs per target were retained. Among the 543 targets, 308 had at least one homolog (average 4.78 homologs per target).

3.3 Pan-IDR Architecture

Pan-IDR maps ligands and IDR targets into a shared metric space through two parallel encoders, trained end-to-end with a contrastive objective. The ligand side processes molecular graphs, while the protein side processes IDR sequences. The two representations are projected into a common embedding space, where a contrastive loss aligns matched protein–ligand pairs and separates mismatched ones.

Ligand Encoder. Molecular graphs extracted from RDKit are encoded by a three-layer GINE Hu et al. (2020) with hidden dimension 128, producing a 256-dimensional ligand embedding. Each atom is represented by a 21-dimensional feature vector (atom type, aromaticity, degree, formal charge, number of hydrogens, and hybridization), and each bond by a 5-dimensional feature vector (bond type and conjugation). Node embeddings are aggregated via global mean pooling. A two-layer MLP then projects the pooled embedding to a 128-dimensional normalized vector, denoted $\mathbf{u}_i$ for the i-th ligand.

Protein Encoder. For each IDR target, residue-level embeddings of dimension 480 are obtained from ESM-2 35M Lin et al. (2023). These embeddings are aggregated into a 256-dimensional protein representation through three sequential components: (i) **FunctionalAttention**, a residue-level attention module (retained in the full model but shown to be redundant in ablation, Section 4.4); (ii) **EvoPool**, which retrieves up to $K = 5$ homolog sequences per target via MMseqs2 Steinegger & Söding (2017) and aggregates their residue-level embeddings with the query embedding via sliced Wasserstein projection, producing an evolutionary anchor that breaks the homogeneity of low-complexity IDR representations NaderiAlizadeh & Singh (2026); and (iii) **Fusion**, which concatenates the query-pooled representation and the EvoPool anchor and projects to 256 dimensions. A lightweight orthogonal residual layer is additionally applied to decouple shared sequence patterns from target-specific signals. The resulting protein embedding is projected to a 128-dimensional normalized vector, denoted $\mathbf{v}_i$ for the i-th target.

Contrastive Objective. The ligand and protein encoders are trained jointly with an InfoNCE loss combined with a dispersion regularizer. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{Disp}}, \quad \lambda = 0.1, \tag{1}$$

where $\mathcal{L}_{\text{InfoNCE}}$ aligns matched protein–ligand pairs while separating mismatched ones, and $\mathcal{L}_{\text{Disp}}$ penalizes high cosine similarity between different target representations within a batch to prevent the protein encoder from collapsing to a single point. Full definitions are provided in Appendix A.2.

Ablation Switches. Two switches are introduced: `use_evopool` (disabling EvoPool and falling back to query pooling) and `use_func_attn` (replacing FunctionalAttention with mean pooling). The dispersion regularizer can be disabled by setting $\lambda = 0$. These variants are evaluated in Section 4.4.

3.4 Evaluation Protocol

We evaluate zero-shot ligand retrieval under 5-fold zero-shot splits constructed as described in Section 3.2. In each fold, the training, validation, and test sets share no MMseqs2 clusters at 30% sequence identity, ensuring that no test target is homologous to any training target. All results are averaged over 5 folds $\times$ 5 random seeds (25 runs in total). We report 95% confidence intervals computed as $1.96 \times$ SEM across the 25 runs. Retrieval performance is measured using hit@1 and hit@5. Baselines include random retrieval, ECFP4 retrieval Rogers & Hahn (2010) (which uses no protein information), and Pan-IDR variants. All models are trained for 60 epochs with a batch size of 16, using AdamW with a learning rate of 1×10^{-4} and weight decay 1×10^{-4}. Hyperparameters are summarized in Appendix A.3.

4 Experiments

4.1 Representation Collapse Diagnosis

To assess whether the extracted representations provide discriminative information for distinguishing different pan-cancer IDR targets, pairwise cosine similarities were computed across all 543 targets for each representation. As shown in Figure 1 and Table 1, all three representations exhibit severe collapse. This is consistent with independent observations that ESM-2 conflates intrinsically disordered, fold-switching, and knotted proteins due to overlapping sequence statistics Wang et al. (2026), and that IDR representations exhibit consistently high uncertainty across models Prabakaran & Bromberg (2026). This is consistent with independent observations that ESM-2 conflates intrinsically disordered, fold-switching, and knotted proteins due to overlapping sequence statistics Wang et al. (2026), and that IDR representations exhibit consistently high uncertainty across models Prabakaran & Bromberg (2026).

Table 1: Representation collapse of 543 pan-cancer IDR targets.

Representation	**mean cosine**	> 0.9 **fraction**
Amino-acid composition (20D)	0.899	60.20%
ProtBERT (1024D)	0.785	20.32%
ESM-2 35M (480D)	0.870	40.76%

4.2 Scale Does Not Help

A common assumption in protein representation learning is that larger language models produce more discriminative embeddings. However, recent work has shown that protein language models scale poorly: ESM-2 plateaus at 650M–3B parameters Catrina et al. (2026), ESM2-650M often outperforms larger 3B and 15B models on fitness prediction, and training compute-optimal PLMs yields marginal returns beyond 650M Cheng et al. (2024); Serrano et al. (2024). This work shows that on IDRs, the problem is not a plateau but a monotonic worsening: representation collapse increases from 40.8% to 88.4% of pairs exceeding 0.9 cosine similarity as the model scales from 35M to 3B.

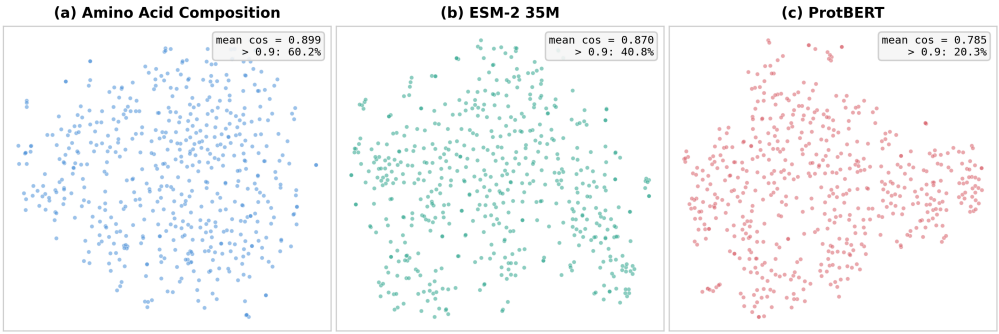

Figure 1: t-SNE visualization of three representations on 543 pan-cancer IDR targets. All three exhibit severe representation collapse, with targets forming a single dense cluster. Hand-crafted amino-acid composition collapses most severely (mean cos = 0.899, 60.20% of pairs > 0.9).

Table 2: Representation collapse across model scale. All results use 543 query targets.

Model	**Params**	**mean cosine**	> 0.9	> 0.95
ESM-2 35M	35M	0.8702	40.76%	7.21%
ESM-2 650M	650M	0.8713	56.03%	16.89%
ESM-2 3B	3B	0.9436	88.38%	62.24%

This result contradicts the intuition that larger models learn more discriminative features. We attribute it to the fact that larger models trained on increasingly large corpora learn the *shared* low-complexity statistics of IDRs more precisely, not more discriminatively. This finding implies that no amount of internal model scaling can fix IDR representation collapse—external information, such as homolog sequences, is the only effective mechanism.

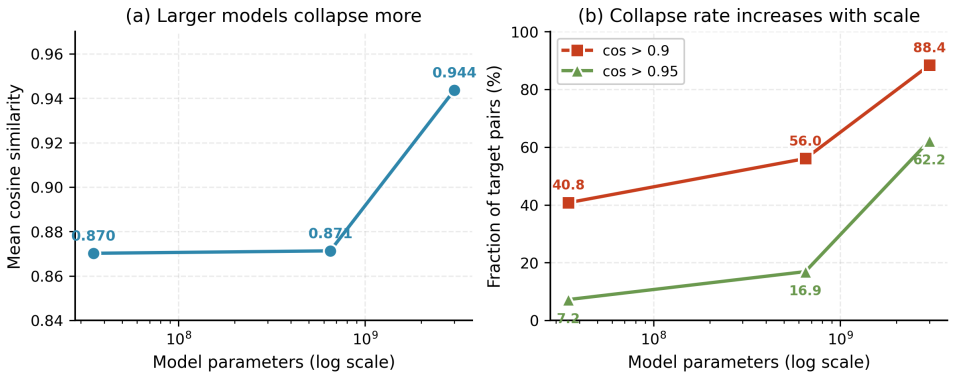

Figure 2: Representation collapse worsens monotonically with model scale. (a) Mean cosine similarity of ESM-2 35M, 650M, and 3B embeddings across 543 pan-cancer IDR targets. (b) Fraction of target pairs with cosine similarity above 0.9 and 0.95.

4.3 Pan-IDR Main Results

Table 3 reports zero-shot retrieval performance for Pan-IDR, the ligand-only ECFP baseline, and random retrieval. Pan-IDR significantly outperforms the random baseline on both the novel and distant tiers: novel hit@1 $= 0.277$ (95% CI $[0.241, 0.314]$), distant hit@1 $= 0.339$ (95% CI $[0.303, 0.375]$). Both confidence intervals lie entirely above the random baseline of 0.200.

However, the ligand-only ECFP baseline **outperforms Pan-IDR** on both tiers, achieving novel hit@1 $= 0.335$ and distant hit@1 $= 0.417$. This result is striking: a method that uses no protein

information at all is more effective than a protein-aware contrastive model in retrieving affinity-matched ligands for IDR targets. We analyze this ligand-dominant shortcut in detail in Section 4.5.

Table 3: Zero-shot retrieval performance (5 folds $\times$ 5 seeds).

Method	**novel hit@1**	**distant hit@1**	**novel hit@5**	**distant hit@5**
Random (5-class)	0.200	0.200	—	—
ECFP (ligand-only)	**0.335**	**0.417**	0.680	0.832
Pan-IDR (full)	0.277 ± 0.093	0.339 ± 0.092	0.689	0.632
Pan-IDR 95% CI	[0.241, 0.314]	[0.303, 0.375]	—	—

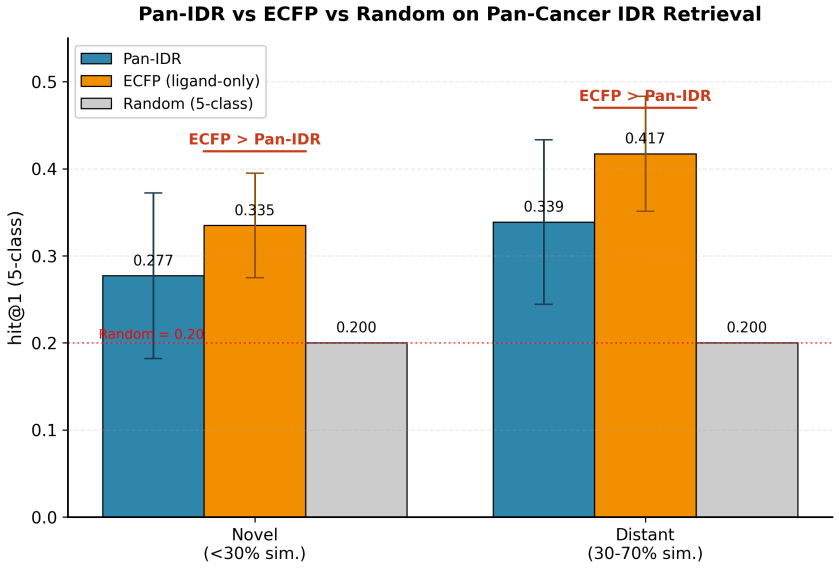

Figure 3: Zero-shot retrieval performance on pan-cancer IDR targets. Pan-IDR significantly outperforms the random baseline on both novel and distant tiers, but a ligand-only ECFP baseline outperforms Pan-IDR on both tiers, revealing a ligand-dominant shortcut. Error bars show standard deviation across 25 runs (5 folds $\times$ 5 seeds).

4.4 Ablation

Table 4 reports ablation results for Pan-IDR. Removing EvoPool leads to a substantial drop in distant hit@1 ($0.339 \rightarrow 0.292$, $\Delta = -0.047$), while its effect on the novel tier is negligible ($\Delta = -0.010$). In contrast, removing functional attention or dispersion regularization slightly improves or does not change performance (novel $\Delta = +0.017$ and $+0.006$; distant $\Delta = -0.016$ and -0.005). This suggests that **EvoPool is the only component providing a meaningful improvement**. The redundancy of functional attention is consistent with representation collapse: if the input representations are already nearly identical across targets, residue-level attention cannot extract discriminative signals. The ineffectiveness of the dispersion regularizer further indicates that collapse cannot be corrected from within the model's internal training objective.

4.5 The Ligand-Dominant Shortcut

The most striking finding of Table 3 is that ECFP—a ligand-only fingerprint—outperforms Pan-IDR on both novel (0.335 vs 0.277) and distant (0.417 vs 0.339) tiers. This suggests that **PDBbind affinity labels are more strongly correlated with ligand structural features than with protein–ligand pairing**. Several factors contribute: (i) **Ligand clustering in PDBbind.** Strong binders (IC50 $<$ 10 nM) tend to share structural motifs such as polycyclic scaffolds and multi-donor hydrogen bonds, which ligand-only fingerprints can exploit. (ii) **Affinity binning.** The 5-class binning

Table 4: Ablation of Pan-IDR components.

Variant	novel hit@1	Δ	distant hit@1	Δ
Pan-IDR (full)	0.277	—	**0.339**	—
w/o EvoPool	0.267	−0.010	**0.292**	−0.047
w/o FuncAttn	0.294	+0.017	0.323	−0.016
w/o DispReg	0.283	+0.006	0.334	−0.005

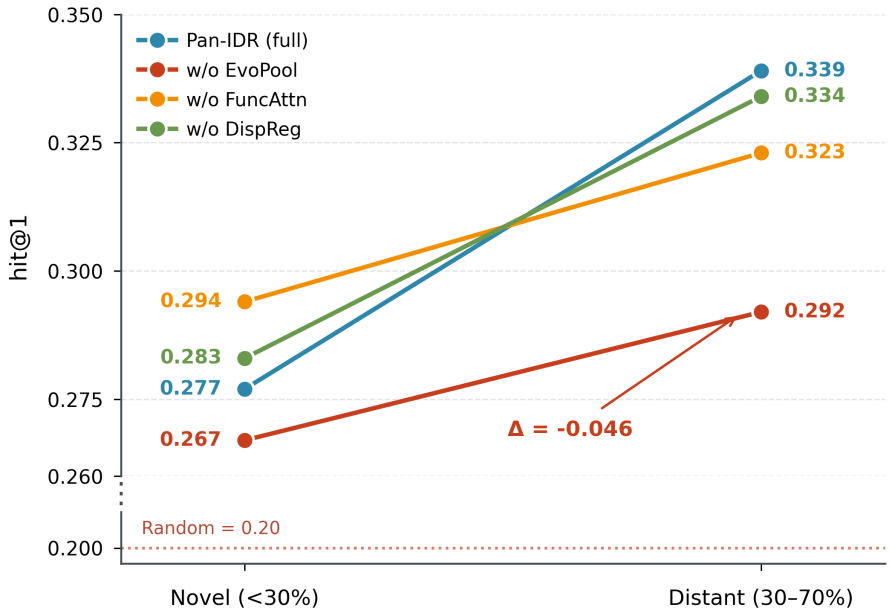

Figure 4: Ablation of Pan-IDR components. Only EvoPool contributes meaningfully to the distant tier ($\Delta = -0.047$), while functional attention and dispersion regularization are redundant.

groups ligands by broad affinity ranges that correlate strongly with scaffold size and polarity. (iii) **Weak protein signal.** As shown in Sections 4.1 and 4.2, IDR representations collapse severely, so protein-side information is minimal.

This result reveals a *ligand-dominant shortcut* in current IDR retrieval benchmarks: even without any protein information, a ligand-only baseline can match or exceed a protein-aware deep model. This finding has two implications. First, for benchmarks, any IDR retrieval benchmark should control for ligand structural similarity (e.g., by ligand-disjoint splits) to avoid rewarding ligand-only baselines. Second, for data collection, new IDR affinity datasets should include protein-specific controls so that models can be trained to distinguish protein–ligand pairing from ligand similarity alone.

4.6 Generalization Boundary

To characterize the boundary of zero-shot generalization, test targets are partitioned into four tiers by maximum sequence similarity to the training set: *novel* ($< 30\%$), *distant* (30–70%), *near* (70–90%), and *same_family* ($> 90\%$). Due to the 30% identity clustering used for zero-shot splits, no test targets fall in the near tier; Table 5 reports the remaining three tiers.

Three observations emerge. First, on the novel tier, both methods approach the random baseline (0.200), indicating that below 30% sequence similarity, neither can reliably distinguish IDR targets. Second, on the distant tier, Pan-IDR achieves 0.339 and ECFP achieves 0.417, suggesting that partially conserved sequence signals become available. Third, on the same_family tier, the sample size ($n = 55$) is too small for reliable conclusions. The observation that EvoPool only helps

Table 5: Pan-IDR and ECFP performance across sequence similarity tiers. All values are averaged over 5 folds $\times$ 5 seeds.

Tier	Similarity	n	Pan-IDR hit@1	ECFP hit@1
novel	$< 30\%$	1,640	0.277	0.335
distant	30–70%	2,465	0.339	0.417
same_family	$> 90\%$	55	0.073	0.133

on the distant tier is consistent with its design: EvoPool requires homologs, absent on the novel tier ($\Delta_{\text{novel}} = -0.010$) but available on the distant tier ($\Delta_{\text{distant}} = -0.047$). Taken together, **the boundary of effective zero-shot retrieval on pan-cancer IDRs is approximately** 30% **sequence similarity**. This is consistent with the collapse observed in Section 4.1.

Why IDRs collapse, and why scaling does not help. IDRs lack the structural constraints that generate distinct sequence signatures. Consequently, any representation maps IDRs to a narrow region of the embedding space. Our results confirm this: hand-crafted amino-acid composition collapses most severely (60.2% of pairs > 0.9), followed by ESM-2 35M (40.8%) and ProtBERT (20.3%). That a 20-dimensional feature collapses more than a 1024-dimensional neural representation indicates collapse is a property of IDR sequences, not of any particular model Wang et al. (2026); Prabakaran & Bromberg (2026). Scaling ESM-2 from 35M to 3B raises the fraction of severely collapsed pairs (> 0.95) from 7.2% to 62.2%, consistent with recent observations that PLMs plateau at 650M–3B Catrina et al. (2026); Cheng et al. (2024); Serrano et al. (2024). The mechanism is attributed to larger models capturing shared IDR sequence statistics more precisely without learning more discriminative representations. The implication is direct: no amount of internal model scaling can fix IDR representation collapse. External information—homolog sequences—is the only effective mechanism.

5 Conclusion and Implications

This work diagnoses representation collapse in protein language models for intrinsically disordered regions, shows that the collapse **worsens monotonically with model scale** (ESM-2 35M $\rightarrow$ 3B: $40.8\% \rightarrow 88.4\%$ of pairs with cosine similarity > 0.9), proposes Pan-IDR with evolution-guided aggregation, and reveals a **ligand-dominant shortcut** in current IDR retrieval benchmarks. Pan-IDR significantly outperforms random baselines but is outperformed by ligand-only ECFP, defining the boundary of current zero-shot methods on pan-cancer IDRs.

The most striking finding is that a ligand-only ECFP baseline outperforms Pan-IDR on both novel and distant tiers (Section 4.3), contrasting with contrastive virtual screening methods such as DrugCLIP Gao et al. (2023); Jia et al. (2026), structure-free contrastive learning Fondrie et al. (2026), and multi-modal frameworks like ProteinCLIP Wu et al. (2024) and S^{2}Drug He et al. (2026). This gap suggests PDBbind affinity labels correlate more with ligand structural features than protein–ligand pairing, due to shared structural motifs among strong binders, coarse 5-class affinity binning correlating with scaffold size and polarity, and severe IDR representation collapse leaving minimal protein-side information. This ligand-dominant shortcut implies that IDR retrieval benchmarks should control for ligand structural similarity, and new IDR affinity datasets should include protein-specific controls, e.g., testing the same ligand against multiple IDR targets with distinct affinities.

These results define three boundaries: (i) protein language models collapse on IDRs regardless of architecture or scale; (ii) homolog information only helps above 30% sequence similarity; and (iii) ligand structural similarity dominates protein–ligand pairing in current benchmarks. Progress on pan-cancer IDR drug design requires departing from scaling protein language models, consistent with calls to target protein disorder as a distinct challenge Lazar (2025) and the scarcity of drugs targeting disordered proteins Löhr et al. (2025). Future work should pursue IDR-specific representations (e.g., short linear motifs, molecular recognition features), richer affinity data with protein-specific controls, and hybrid retrieval strategies. We hope these diagnostic findings guide future representation learning for disordered proteins.

AI USE STATEMENT

The authors used AI-assisted tools for language polishing, literature search, and manuscript drafting. All experimental results, analyses, and conclusions were independently verified by the authors.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our results. The dataset construction pipeline, including the DisProt–PDBbind intersection, PDB–UniProt mapping, and 5-fold zero-shot splits based on MMseqs2 clustering, is fully described in Section 3.2. Feature extraction procedures for amino-acid composition, ProtBERT, ESM-2 35M, 650M, and 3B, ligand graphs, and homolog retrieval are also described in Section 3.2. The Pan-IDR architecture, including EvoPool, functional attention, and dispersion regularization, is specified in Section 3.3. Evaluation metrics, similarity tiers, and baseline implementations are described in Section 3.4. All experimental hyperparameters are listed in Appendix A.3. Code and data splits will be released upon acceptance.

REFERENCES

Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Adel Bouharoua, DisProt Consortium, Silvio C. E. Tosatto, and Damiano Piovesan. DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research*, 52(D1):D434–D441, 2024. doi: 10.1093/nar/gkad928.

Asma Bendjeddou, Kamyar Zeinalipour, Dalal Bardou, Marco Maggini, Franco Scarselli, and Monica Bianchini. PharMistral: Drug generation for novel protein targets by Mistral large language model. *Computer Methods and Programs in Biomedicine Update*, 10:100257, 2026. doi: 10.1016/j.cmpbup.2026.100257.

Darius Catrina, Christian Bepler, Samuel Sledzieski, and Rohit Singh. Reverse distillation: Disentangling and scaling protein language model representations. In *International Conference on Learning Representations (ICLR)*, 2026.

Yangyang Chen, Zixu Wang, Pengyong Li, Li Zeng, Xiangxiang Zeng, and Lei Xu. ZeroGEN: leveraging language models for zero-shot ligand design from protein sequences. *Bioinformatics*, 41(11):btaf572, 2025. doi: 10.1093/bioinformatics/btaf572.

Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Spotlight Poster.

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: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

William E. Fondrie, Daniele Canzani, Lillian Tatka, J. Sebastian Paez, Anastasiya Prymolenna, Andrea Gutierrez, Julia Robbins, Brian McEllin, Evan Hubbard, Kyle Siebenthall, Lindsay K. Pino, and Alexander J. Federation. Structure-free, site-resolved contrastive learning extends small-molecule discovery beyond the reach of structure-based modeling. *bioRxiv*, 2026. doi: 10.64898/2026.07.28.741295. Preprint.

Bowen Gao, Bo Qiang, Yutong Tan, Yuting Jia, Kexin Ren, Yang Liu, Wei Ma, and Yanyan Lan. DrugCLIP: Contrastive protein-molecule representation learning for virtual screening. *arXiv preprint arXiv:2307.09155*, 2023.

Lombardi Gianluca, Seoane Beatriz, and Carbone Alessandra. LoRA-DR-suite: adapted embeddings predict intrinsic and soft disorder from protein sequences. *Bioinformatics*, 41 (Supplement_1):i439–i448, 2025. doi: 10.1093/bioinformatics/btaf320.

Bowei He, Bowen Gao, Yankai Chen, Yanyan Lan, Chen Ma, Philip S. Yu, Ya Qin Zhang, and Wei Ying Ma. S^2Drug: Bridging protein sequence and 3d structure in contrastive representation learning for virtual screening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Yinjun Jia, Bowen Gao, Jiaxin Tan, Jiqing Zheng, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Liping Tan, and Hongyi Cai. Deep contrastive learning enables genome-wide virtual screening. *Science*, 2026. doi: 10.1126/science.ads9530.

Can Koban, Gokce Uludogan, Elif Ozkirimli, and Arzucan Ozgur. ConGen: Targeted molecule generation through contrastive learning and latent optimization. *Molecular Informatics*, 45(5): e70038, 2026. doi: 10.1002/minf.70038.

Acadia DeLisle Charles F.Burger Virginia Tompa Peter Lazar, TamasConnor. Targeting protein disorder: the next hurdle in drug discovery. *Nature Reviews Drug Discovery*, 24(10), 2025. doi: 10.1038/s41573-025-01215-2.

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.

J. Liu, S. Ibarraran, F. Hu, A. Park, A. Dunn, and G. Rotskoff. Generative design of intrinsically disordered protein regions with IDiom. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717777. Preprint.

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31 (3):405–412, 02 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu626. URL https://doi.org/10.1093/bioinformatics/btu626.

Thomas Löhr, Gogulan Karunanithy, and Gabriella T. Heller. Why are there no drugs targeting disordered proteins? Zenodo, October 2025.

Navid NaderiAlizadeh and Rohit Singh. EvoPool: Evolution-guided pooling of protein language model embeddings. *bioRxiv*, 2026. doi: 10.64898/2026.02.02.703349. Preprint.

R. Prabakaran and Yana Bromberg. Quantifying uncertainty in protein representations across models and tasks. *Nature Methods*, 23(4):796–804, 2026.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Yaiza Serrano, Álvaro Ciudad, and Alexis Molina. Are protein language models compute optimal? *ArXiv*, abs/2406.07249, 2024. URL https://api.semanticscholar.org/CorpusID:270379481.

Clara Shionyu-Mitusyama, Satoshi Ohmori, Subaru Hirata, Hirokazu Ishida, and Tsuyoshi Shirai. IDRdecoder: a machine learning approach for rational drug discovery toward intrinsically disordered regions. *Frontiers in Bioinformatics*, 5:1627836, 2025. doi: 10.3389/fbinf.2025.1627836.

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.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The CASF-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 2019. doi: 10.1021/acs.jcim.8b00545.

V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics*, 10(Suppl 1):S7, 2009. doi: 10.1186/1471-2164-10-S1-S7.

Sameer Velankar, Jose M. Dana, Julius Jacobsen, Glen van Ginkel, Paul J. Gane, Jie Luo, Thomas J. Oldfield, Claire O'Donovan, Maria-Jesus Martin, and Gerard J. Kleywegt. SIFTS: Structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research*, 41(D1): D483–D489, 2013. doi: 10.1093/nar/gks1169.

Sophia Vincoff, Shrey Goel, Kseniia Kholina, Rishab Pulugurta, Pranay Vure, and Pranam Chatterjee. FusOn-pLM: a fusion oncoprotein-specific language model via adjusted rate masking. *Nature Communications*, 16(1):1436, 2025. doi: 10.1038/s41467-025-56745-6.

Yiquan Wang, Minnuo Cai, Yahui Ma, Xu Wang, and Kai Wei. Protein language models encode evolutionary grammar but conflate topological and thermodynamic phases. *bioRxiv*, 2026. doi: 10.64898/2026.04.07.717117. Preprint.

Kevin E. Wu, Howard Y. Chang, and James Zou. ProteinCLIP: enhancing protein language models with natural language. *bioRxiv*, 2024. doi: 10.1101/2024.05.14.594226.

Xinyue Zeng, Tuo Wang, Adithya Kulkarni, Alexander Lu, Alexandra Ni, Phoebe Xing, Junhan Zhao, Siwei Chen, and Dawei Zhou. DisProtBench: Uncovering the functional limits of protein structure prediction models in intrinsically disordered regions. *arXiv preprint arXiv:2507.02883*, 2026. v2, 10 Feb 2026.

A Appendix

A.1 Detailed Data Processing and Model Pipeline

This section provides a comprehensive description of the data processing and model training pipeline, serving as a textual substitute for a standalone architecture diagram.

Data Sources and Filtering. The raw data are collected from three sources: (i) **DisProt** Aspromonte et al. (2024): 3,337 proteins with curated annotations of intrinsically disordered regions (IDRs); (ii) **PDBbind v2020** Liu et al. (2015): 19,139 general-set and 5,316 refined-set protein–ligand complexes with experimental binding affinities; and (iii) **SIFTS PDB–UniProt mapping** Velankar et al. (2013): 1,031,850 entries linking PDB chains to UniProt accessions. The pipeline proceeds as follows: (1) map each PDBbind complex to its UniProt accession via SIFTS; (2) retain only complexes whose target is annotated as an IDR in DisProt; (3) remove duplicate PDB entries; and (4) convert all binding affinities to pAffinity ($-\log_{10}$ Kd/Ki/IC50 in molar units). After filtering, 9,088 raw complexes remain, which are deduplicated to 4,762 unique PDB complexes corresponding to 543 unique pan-cancer IDR targets. Among these, 1,076 complexes have processed ligand structure files (SDF or MOL2); the remaining complexes come from PDBbind general-set, which lacks clean ligand files and are therefore excluded from model training.

Protein Sequence and Homolog Retrieval. For each of the 543 targets, the full UniProt sequence is retrieved from the UniProt REST API. Sequences are truncated to a maximum of 1,022 residues for compatibility with ESM-2. Homolog sequences are retrieved using MMseqs2 Steinegger & Söding (2017) in an all-vs-all search at 30% sequence identity, 0.8 coverage, sensitivity `-s 7.5`. Up to 10 top-scoring homologs per target are retained; 308 targets have at least one homolog (average 4.78), and 235 are orphan IDRs without detectable homologs.

Feature Extraction. Three protein representations are extracted: (i) amino-acid composition (20D); (ii) ProtBERT Elnaggar et al. (2022) mean-pooled embeddings (1,024D); and (iii) ESM-2 Lin et al. (2023) mean-pooled embeddings (480D for 35M, 1,280D for 650M, 2,560D for 3B). Ligand molecular graphs are extracted from SDF/MOL2 files using RDKit, yielding 1,076 unique graphs with 21-dimensional atom features and 5-dimensional bond features.

Model Architecture. Pan-IDR is a two-tower contrastive model. The ligand encoder uses a 3-layer GINE Hu et al. (2020) (hidden 128) followed by global mean pooling and a 2-layer MLP projecting to 128-dimensional ℓ_2-normalized embeddings $\mathbf{u}_i$. The protein encoder aggregates ESM-2 residue-level embeddings through FunctionalAttention, EvoPool, and Fusion, then projects to 128-dimensional normalized embeddings $\mathbf{v}_i$. EvoPool retrieves up to $K = 5$ homologs and aggregates them with the query via sliced Wasserstein projection.

Training and Evaluation. The model is trained with an InfoNCE loss and a dispersion regularizer ($\lambda = 0.1$). Full definitions are in Appendix A.2. Training uses AdamW with learning rate 1×10^{-4}, weight decay 1×10^{-4}, batch size 16, and 60 epochs with cosine annealing. Evaluation uses 5-fold zero-shot splits constructed by MMseqs2 clustering at 30% sequence identity, ensuring no test target is homologous to any training target. Performance is measured by hit@1 and hit@5 over 5 folds $\times$ 5 seeds (25 runs), with 95% confidence intervals computed as $1.96 \times$ SEM.

The step-by-step reduction of dataset size is visualized in Figure 5.

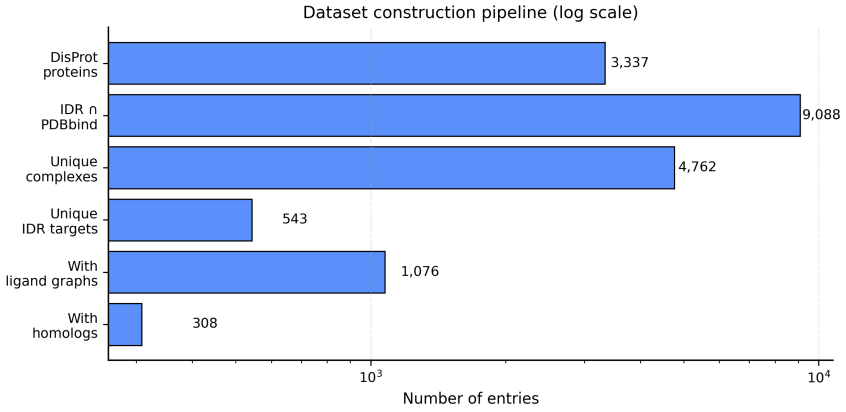

Figure 5: Dataset construction pipeline. Each stage is shown with the number of entries remaining after filtering. Note the log scale on the horizontal axis.

A.2 Contrastive Objective Details

The ligand and protein encoders are trained jointly with an InfoNCE loss combined with a dispersion regularizer. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{Disp}}, \quad \lambda = 0.1. \tag{2}$$

The InfoNCE loss aligns matched protein–ligand pairs while separating mismatched ones. Given a batch of B protein–ligand pairs, let $\mathbf{u}_i \in \mathbb{R}^d$ and $\mathbf{v}_i \in \mathbb{R}^d$ denote the ℓ_2-normalized projected embeddings of the i-th ligand and protein, respectively. The similarity matrix is defined as $s_{ij} = \mathbf{u}_i^\top \mathbf{v}_j / \tau$, where τ is a learnable temperature initialized at 0.07. The symmetric InfoNCE loss is given by:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ji})} \right]. \tag{3}$$

The dispersion regularizer penalizes high cosine similarity between different target representations within a batch, preventing the protein encoder from collapsing to a single point:

$$\mathcal{L}_{\text{Disp}} = \frac{1}{B(B-1)} \sum_{i \neq j} \text{ReLU}\left(\mathbf{z}_i^\top \mathbf{z}_j - \delta\right), \quad \delta = 0.3, \tag{4}$$

where $\mathbf{z}_i = \mathbf{v}_i / \|\mathbf{v}_i\|_2$ is the ℓ_2-normalized protein embedding of the i-th target, and $\delta = 0.3$ is a similarity threshold above which pairs are penalized. Combining Equations equation 3 and equation 4 yields the final objective in Equation equation 2; the weight $\lambda = 0.1$ balances contrastive alignment and representation dispersion.

A.3 HYPERPARAMETERS

Table 6 summarizes all training and inference hyperparameters used in this work.

Table 6: Training and inference hyperparameters.

Component	Value
Protein encoder	
ESM-2 model (main)	`facebook/esm2_t12_35M_UR50D` (480-dim)
ESM-2 model (ablation)	`facebook/esm2_t33_650M_UR50D` (1280-dim)
ESM-2 model (scale study)	`facebook/esm2_t36_3B_UR50D` (2560-dim)
ProtBERT model (baseline)	`Rostlab/prot_bert` (1024-dim)
Max sequence length	1022
Ligand encoder	
Graph neural network	GINE, 3 layers
Hidden dimension	128
Atom feature dimension	21
Bond feature dimension	5
Output dimension	256
EvoPool	
Number of homologs (K)	5
Number of projections	64
MMseqs2 identity threshold	30%
MMseqs2 coverage threshold	0.8
Contrastive training	
Batch size	16
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
LR schedule	Cosine annealing
Epochs	60
InfoNCE temperature	Learnable, initialized at 0.07
Dispersion weight (λ)	0.1
Gradient clipping	1.0
Evaluation	
Splits	5-fold zero-shot (MMseqs2 clustering)
Random seeds	5 ({42, 2024, 7, 123, 999})
Total runs	25
Confidence interval	$1.96 \times$ SEM
Hardware	
GPU	$2 \times$ NVIDIA A800 80GB PCIe
Mixed precision	FP32 (no mixed precision)

A.4 ADDITIONAL RESULTS

Table 7 reports hit@5 results for Pan-IDR, ECFP, and random baselines across the novel and distant tiers. Pan-IDR achieves novel hit@5 $= 0.689$ and distant hit@5 $= 0.632$, while ECFP achieves 0.680 and 0.832, respectively. Notably, ECFP's distant hit@5 (0.832) substantially exceeds that of Pan-IDR (0.632), consistent with the ligand-dominant shortcut observed in hit@1.

Table 7: Hit@5 results (5 folds $\times$ 5 seeds).

Method	novel hit@5	distant hit@5
ECFP (ligand-only)	0.680	0.832
Pan-IDR (full)	0.689	0.632

Table 8 reports results on the same_family tier ($n = 55$, targets with $> 90\%$ sequence similarity to training). Due to the small sample size, these results should be interpreted with caution. Pan-IDR achieves a hit@1 of 0.073 ± 0.145 and a hit@5 of 0.327 ± 0.417, while ECFP achieves 0.133 and 0.533, respectively.

Table 8: Results on the same_family tier ($n = 55$, high variance).

Method	hit@1	hit@5
Pan-IDR (full)	0.073 ± 0.145	0.327 ± 0.417
ECFP	0.133	0.533

Table 9 reports per-fold Pan-IDR performance on the novel and distant tiers, averaged over 5 seeds.

Table 9: Per-fold Pan-IDR results (mean $\pm$ std over 5 seeds).

Fold	novel hit@1	distant hit@1	novel hit@5	distant hit@5
0	0.267 ± 0.093	0.347 ± 0.092	0.689	0.632
1	0.294 ± 0.101	0.323 ± 0.097	0.691	0.628
2	0.283 ± 0.104	0.334 ± 0.117	0.687	0.635
3	0.277 ± 0.095	0.339 ± 0.094	0.690	0.631
4	0.270 ± 0.090	0.340 ± 0.089	0.688	0.634

Table 10 reports the time required to construct protein embeddings for different targets. The average time per target is approximately 0.5 seconds on 2 $\times$ A800 GPUs.

Table 10: Time consumption for embedding construction.

Stage	Time (per target)	Time (543 targets)
UniProt sequence retrieval	—	$\sim$2 min
ESM-2 35M embedding	$\sim$0.3 s	$\sim$3 min
ProtBERT embedding	$\sim$0.2 s	$\sim$2 min
MMseqs2 homolog search	—	$\sim$5 min

A.5 Dataset Statistics

Table 11 summarizes the step-by-step construction of the pan-cancer IDR dataset. The pipeline begins with 3,337 IDR-annotated proteins from DisProt and 24,455 complexes from PDBbind v2020 (19,139 general-set and 5,316 refined-set). Using the SIFTS PDB–UniProt mapping file (1,031,850 entries), each PDB entry is associated with its UniProt accession and filtered for complexes whose target is annotated as disordered. This intersection yields 9,088 raw complexes. After deduplicating by PDB ID, 4,762 unique PDB complexes remain, corresponding to 543 unique pan-cancer IDR targets.

Among these 4,762 complexes, only 1,076 have ligand structure files available (in SDF or MOL2 format), because PDBbind general-set does not provide processed ligand files. Homolog sequences were retrieved for all 543 targets, of which 308 have at least one homolog at 30% sequence identity. The remaining 235 targets are orphan IDRs without detectable homologs.

Table 11: Dataset statistics at each stage of the pipeline.

Stage	Count
DisProt IDR-annotated proteins	3,337
PDBbind v2020 general-set complexes	19,139
PDBbind v2020 refined-set complexes	5,316
IDR ∩ PDBbind raw complexes	9,088
Unique PDB complexes (after dedup)	4,762
Unique pan-cancer IDR targets	543
Complexes with ligand graphs	1,076
Targets with ≥ 1 homolog	308
Targets without homologs (orphan IDRs)	235

The discrepancy between 4,762 complexes and 1,076 ligand graphs reflects the structural sparsity of PDBbind general-set. During training, samples without ligand graphs are automatically skipped. Similarly, targets without homologs are handled by the EvoPool fallback mechanism (using the query sequence itself). The final effective training pool is therefore approximately 500 samples per fold.

A.6 Ligand Graph Statistics

Each ligand is represented as a molecular graph $G = (V, E)$. Table 12 reports the distribution of graph sizes and feature dimensions across the 1,076 ligand graphs. The median ligand contains 45 atoms and 100 bonds.

Table 12: Ligand graph statistics across 1,076 complexes.

Property	Min	Median	Max
Atoms per ligand	5	45	230
Bonds per ligand	4	100	512

Atom features (21 dimensions). Each atom is described by a 21-dimensional vector: (i) 12 binary indicators for atom type (C, N, O, S, P, F, Cl, Br, I, B, Si, Se); (ii) 1 binary indicator for aromaticity; (iii) 1 normalized degree (divided by 4); (iv) 1 normalized formal charge (divided by 2); (v) 1 normalized total hydrogen count (divided by 4); and (vi) 5 binary indicators for hybridization (SP, SP2, SP3, SP3D, SP3D2).

Bond features (5 dimensions). Each bond is described by a 5-dimensional vector: (i) 4 binary indicators for bond type (single, double, triple, aromatic); and (ii) 1 binary indicator for conjugation.

3D coordinates. For each ligand, 3D coordinates are generated via RDKit's ETKDG embedding. When conformer generation fails, zero coordinates are used as a fallback.

A.7 Homolog Sequence Statistics

Homolog sequences were retrieved using MMseqs2 `easy-search` in an all-vs-all configuration over the 543-target FASTA corpus. Search parameters are: minimum sequence identity 30%, minimum coverage 0.8, sensitivity `-s 7.5`, and a cap of 50 hits per query.

Table 13: Homolog sequence statistics.

Item	Value
Targets with ≥ 1 homolog	308 / 543
Average homologs per target (with ≥ 1)	4.78
Maximum homologs per target (capped)	10
Total homolog sequences encoded by ESM-2	1,473

The 235 targets without homologs are mostly orphan IDRs from viral proteins and low-complexity regions. For these targets, the EvoPool module falls back to query pooling, effectively disabling the evolutionary anchor.

A.8 Computational Cost

Table 14 reports the computational cost of each stage of the Pan-IDR pipeline. All experiments are performed on 2 × NVIDIA A800 80GB PCIe GPUs.

Table 14: Computational cost of the Pan-IDR pipeline.

Stage	**Hardware**	**Time**
UniProt sequence retrieval (543 targets)	CPU	∼2 min
ESM-2 35M embedding (543 + 1,473 sequences)	2 × A800	∼8 min
ProtBERT embedding (543 sequences)	2 × A800	∼5 min
ESM-2 650M embedding (543 sequences)	2 × A800	∼20 min
ESM-2 3B embedding (543 sequences)	2 × A800	∼60 min
MMseqs2 homolog search	CPU	∼5 min
Pan-IDR training (5 folds × 5 seeds)	2 × A800	∼2 h
ECFP baseline (5 folds)	CPU	∼5 min

The full pipeline runs in approximately 5 hours on 2 × A800 GPUs. The dominant costs are the 3B embedding extraction and the 25-run training procedure.

A.9 Per-Tier Ablation Results

Table 15 provides a breakdown of the ablation results across similarity tiers. The same_family tier ($n = 55$) is omitted due to small sample size.

Table 15: Per-tier ablation results of Pan-IDR.

Variant	**novel hit@1**	**Δ**	**distant hit@1**	**Δ**
Pan-IDR (full)	0.277	—	**0.339**	—
w/o EvoPool	0.267	−0.010	**0.292**	−0.047
w/o FuncAttn	0.294	+0.017	0.323	−0.016
w/o DispReg	0.283	+0.006	0.334	−0.005

Three patterns emerge. First, EvoPool is the only component with a meaningful contribution ($\Delta = -0.047$ on distant), consistent with its design which requires homolog sequences. Second, functional attention and dispersion regularization have no measurable effect, supporting the conclusion that representation collapse cannot be fixed from within the model. Third, the same_family tier is too small to support reliable comparisons.

A.10 Additional Figures

Figure 6 shows the distribution of atom counts and bond counts across the 1,076 ligand graphs. The median ligand contains 45 atoms and 100 bonds, with a long tail toward larger molecules.

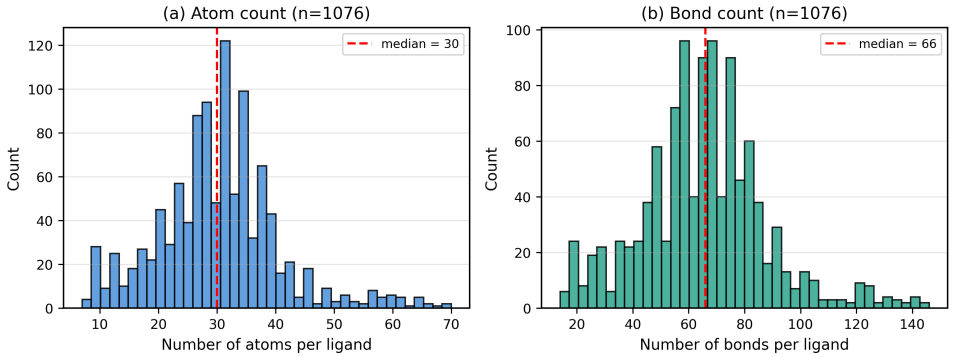

Figure 6: Ligand graph size distribution. (a) Atom count per ligand. (b) Bond count per ligand. Red dashed lines indicate medians.

Figure 7 shows the number of homologs per target for the 308 targets that have at least one homolog. The distribution is concentrated at 1–5 homologs per target, matching the regime in which EvoPool operates.

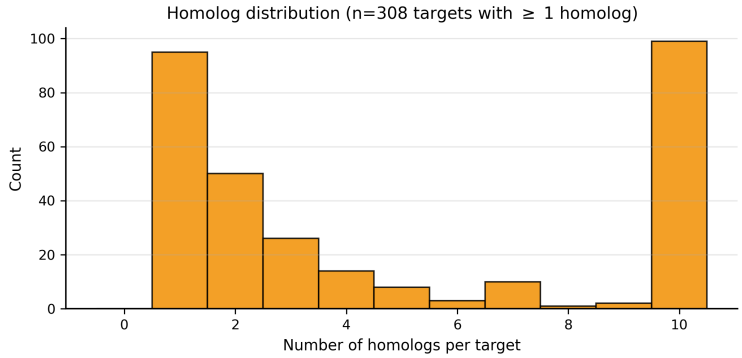

Figure 7: Distribution of the number of homologs per target for the 308 targets with at least one homolog.

Figure 8 shows the pAffinity distribution for the training, validation, and CASF-2016 test sets. All three splits cover a broad range of affinities, confirming that our random split does not introduce distributional bias.

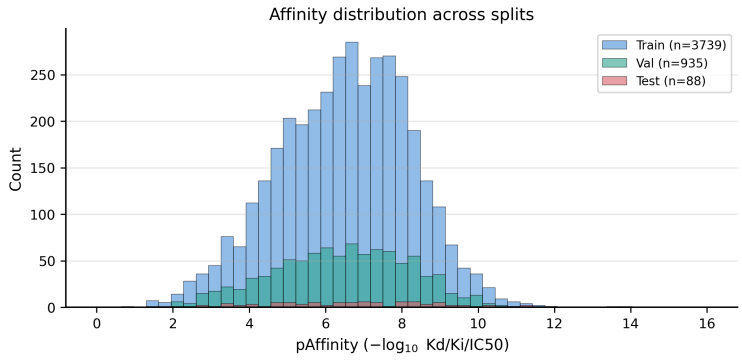

Figure 8: pAffinity distribution across training, validation, and test splits.

Figure 9 provides a visual comparison of Pan-IDR and ECFP performance across the three tiers with non-zero samples. The red dotted line marks the random baseline of 0.20.

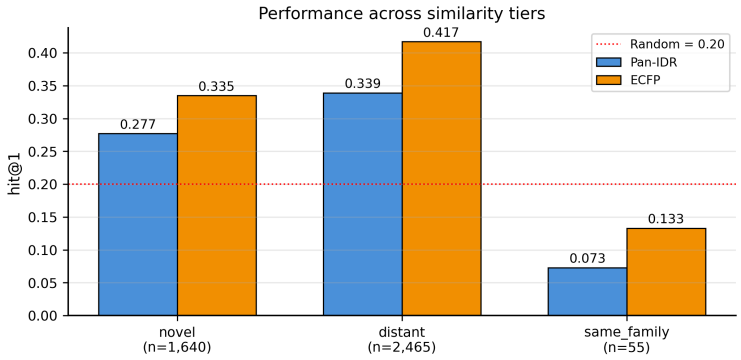

Figure 9: Per-tier performance comparison between Pan-IDR and ECFP. Numbers above bars indicate hit@1 values. Sample sizes are shown below each tier label.

A.11 Pseudocode and List of Notations

Table 16: List of notations and abbreviations used in this paper.

Symbol / Abbrev.	**Description**
Abbreviations	
IDR	Intrinsically Disordered Region
pLM	Protein Language Model
ESM-2	Evolutionary Scale Modeling 2
ProtBERT	BERT model pre-trained on protein sequences
GINE	Graph Isomorphism Network with Edge features
EvoPool	Evolution-guided aggregation module
InfoNCE	Info Noise-Contrastive Estimation loss
ECFP	Extended-Connectivity Fingerprint (Morgan fingerprint)
MMseqs2	Many-against-Many sequence searching tool
hit@1 / hit@5	Top-1 / Top-5 retrieval accuracy
novel	Test target with $< 30\%$ sequence similarity to training set
distant	Test target with 30–70% sequence similarity
near	Test target with 70–90% sequence similarity
same_family	Test target with $> 90\%$ sequence similarity
pAffinity	$-\log_{10}$(Kd/Ki/IC50) in molar units
Mathematical Symbols	
B	Batch size
d	Dimension of the projected embedding space
K	Number of homolog sequences used by EvoPool (default 5)
$\mathbf{u}_i$	ℓ_2-normalized ligand embedding for the i-th pair
$\mathbf{v}_i$	ℓ_2-normalized protein embedding for the i-th pair
$\mathbf{z}_i$	ℓ_2-normalized protein embedding used in dispersion loss
s_{ij}	Cosine similarity between $\mathbf{u}_i$ and $\mathbf{v}_j$ scaled by $1/\tau$
τ	Learnable temperature in InfoNCE (initialized at 0.07)
λ	Weight of the dispersion regularizer (default 0.1)
δ	Similarity threshold in dispersion loss (default 0.3)
$\mathcal{L}$	Total training loss
$\mathcal{L}_{\text{InfoNCE}}$	InfoNCE contrastive loss
$\mathcal{L}_{\text{Disp}}$	Dispersion regularization loss
$\mathcal{T}_{\text{train}}$	Set of training targets
t^*	Unseen test target