NOT A FITNESS BOTTLENECK: WHY PROTEIN LANGUAGE MODELS FALL SHORT ON DISORDERED-PROTEIN ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (PLMs) are increasingly used as frozen encoders to predict the conformational properties of intrinsically disordered proteins (IDPs), yet how much ensemble physics evolutionary pretraining captures remains unclear. Here we measure *evolutionary sufficiency* (ES), the fraction of a directly supervised model's out-of-distribution accuracy that a frozen PLM representation recovers, for five polymer-physics properties computed from simulated ensembles of nearly 10,000 disordered sequences across six kingdoms, with viral sequences held out. We first show that ES is graded rather than binary, as frozen PLMs adapted to disordered regions recover nearly all of the supervised accuracy for chain dimensions but progressively less for asphericity, the Flory exponent, and its prefactor. We then test and reject a fitness-bottleneck explanation for this shortfall, because the embeddings predict these properties several times better than the PLM's own likelihood and encode net charge almost perfectly, whereas mean pooling hides the charge patterning on which the hardest properties depend. Instead, the gap has a readout component, which a position-aware probe partly closes and which grows with model size, and a pretraining-data component, as encoders pretrained on simulated ensembles reach the supervised ceiling while frozen. Once the backbone is fine-tuned, however, evolutionary initialisation matters little. ES thus offers a probe-aware diagnostic for when frozen PLM features suffice for biophysical prediction.

1 INTRODUCTION

Intrinsically disordered proteins and regions (IDRs) are found in all kingdoms of life and make up a substantial fraction of eukaryotic proteomes (Uversky, 2002; Tompa, 2002; Dyson & Wright, 2005). Rather than folding into a single structure, an IDR populates an ensemble of rapidly interconverting conformations, such that its behaviour is described by ensemble-average properties rather than by atomic coordinates (Holehouse & Kragelund, 2024). Several such properties are routinely measured and simulated, of which the radius of gyration (R_g) and end-to-end distance (R_{ee}) report on overall size and the asphericity (Δ) reports on shape, ranging from 0 for a sphere to 1 for a rod. The apparent Flory exponent (ν) and prefactor (A_0), obtained by fitting intramolecular distances to a polymer scaling law, in turn report on how compaction depends on sequence separation. These quantities are coarse, but they link sequence chemistry, notably the amount and patterning of charge (Das & Pappu, 2013), to function and to the ability of IDRs to form condensates.

Two routes now exist for predicting these properties from sequence. Coarse-grained force fields such as CALVADOS and Mpipi reproduce experimental ensemble dimensions and have been applied to entire proteomes (Tesei et al., 2021; Joseph et al., 2021), and networks trained on such simulations predict them in milliseconds (Lotthammer et al., 2024). In parallel, protein language models (PLMs) pretrained on evolutionary sequence databases are increasingly used as general-purpose frozen encoders for the same task (Lin et al., 2023; Elnaggar et al., 2022). PLMs are attractive because they require no simulation, yet their pretraining never observes a conformation, and it therefore remains unclear which ensemble properties their representations capture and whether a shortfall reflects the representation itself, the probe placed on top of it, or the pretraining signal.

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.

One hypothesis offers a principled answer, based on the view that masked language modelling on extant sequences approximates learning the evolutionary fitness landscape (Jiao et al., 2026). If a PLM representation carries physical information only insofar as fitness does, properties that vary independently of selection should be unrecoverable from it at any model scale, a hypothesis we refer to as the *fitness bottleneck*. This hypothesis is appealing because it would predict in advance which properties a PLM will fail on, and it also makes concrete, directly testable predictions.

Here, we test it by measuring *evolutionary sufficiency* (ES), which we define as the out-of-distribution accuracy of a frozen PLM representation expressed as a fraction of a direct-supervision baseline trained on the same data. We use five polymer-physics targets from a cross-taxon dataset of simulated IDR ensembles (Velasquez & Rahman, 2026), and hold out all viral sequences as a test set. To separate the representation from everything around it, we compare frozen evolutionary, simulator-pretrained, and randomly initialized backbones under shallow and deep probes, alongside end-to-end fine-tuning and architecture-matched supervised baselines. Every comparison uses one cluster-aware split, untruncated sequences, three seeds and paired bootstrap intervals.

We first show that ES is graded rather than binary, as a frozen IDP-ESM2-8M, an ESM2 model further pretrained on disordered regions (Quinn et al., 2025), recovers 95% of the supervised ceiling for chain dimensions but 89% for the Flory exponent and 77% for the prefactor, and neither a 19-fold larger IDP-ESM2 nor a 375-fold larger PLM from a different family changes this ordering under the standard probe. We then show that the fitness bottleneck does not account for this ordering, because frozen embeddings predict the Flory exponent three to four times better than PLM log-likelihood does and encode net charge almost perfectly, whereas no representation, pretrained or random, linearly exposes charge patterning after mean pooling. Finally, we show that the gap has a readout component, which a position-aware probe partly closes and which grows with model size, and a pretraining-data component, such that two models pretrained on simulations from the same coarse-grained model family reach the supervised ceiling while frozen, most strongly for sequences resembling their pretraining set. In contrast, once the backbone is fine-tuned on about 7,400 labelled ensembles, evolutionary initialisation adds at most 0.02 R^2 over random initialisation.

In summary, we make four contributions.

1. A probe-aware definition of evolutionary sufficiency and a matched protocol for measuring it, [1]with random-feature, architecture-matched, and fine-tuning controls (§3–4).
2. A test and rejection of the fitness-bottleneck hypothesis for IDP ensemble physics, using log-likelihood baselines, linear decoding of charge features and a descriptor ablation that identifies charge patterning as the information the hardest targets require (§5.2).
3. A decomposition of the remaining gap into readout and pretraining-data components, with simulator pretraining closing the gap most strongly near its pretraining domain (§5.3–5.4).
4. Practical guidance on when frozen evolutionary features suffice and when deeper probes, fine-tuning or simulator-pretrained encoders are needed, and evidence that truncating the baseline's inputs or reading out by mean pooling can change the apparent sufficiency (§5.1, §6).

2 Related work

Predicting IDR ensemble properties. Coarse-grained force fields parameterised against experiments, such as CALVADOS and Mpipi, reproduce the global dimensions of disordered proteins (Tesei et al., 2021; Tesei & Lindorff-Larsen, 2023; Joseph et al., 2021) and have been used to simulate entire proteomes (Tesei et al., 2024). Their outputs have in turn trained fast sequence-based predictors, from recurrent networks for ensemble dimensions (Lotthammer et al., 2024) to generative models of full ensembles (Janson et al., 2025; Novak et al., 2026). Two encoders have been pretrained directly on simulated ensembles, namely SeqDance on dynamic properties from molecular dynamics, including IDRome (Hou et al., 2026b), and GeoGraph on geometric and contact-network descriptors of IDRome ensembles (Quinn et al., 2025). BENDER extends simulated ensembles across kingdoms (Velasquez & Rahman, 2026), which makes it possible to test generalization to

[1]Code, split definitions, per-sequence predictions and the scripts to produce the tables and figures are available at `https://anonymous.4open.science/r/ES-1430/`.

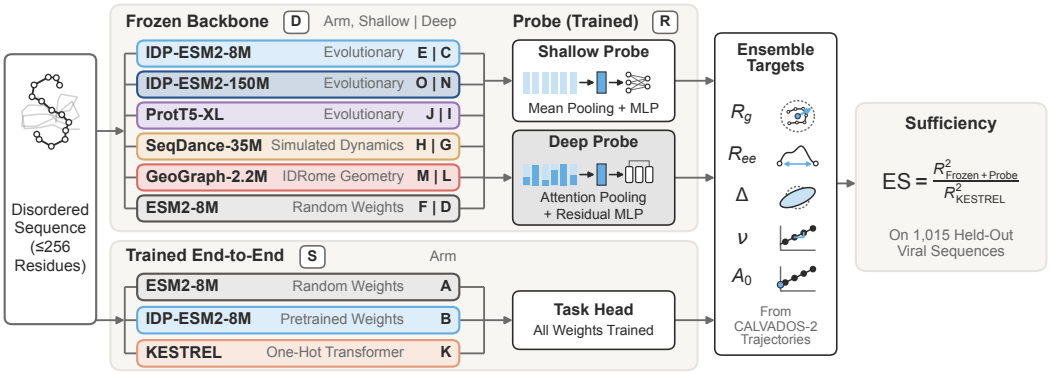

Figure 1: **Configurations compared.** Frozen backbones share two probes (arm labels shallow | deep), three models are trained end-to-end, and all predict ensemble properties computed from CALVADOS-2 trajectories, with KESTREL as the ES denominator. Chips mark the choices that test pretraining data (D), readout (R) and supervision (S), as defined in §3.3.

taxa, such as viruses, that are under-represented in pretraining corpora (Ding & Steinhardt, 2024; Bigot et al., 2026; Sawhney et al., 2025).

What protein language models encode. PLMs trained by masked language modelling capture residue contacts and structural information without supervision (Rives et al., 2021; Rao et al., 2021), and their likelihoods predict mutational effects zero-shot (). Recent work has examined when this works, showing that likelihood-based fitness prediction succeeds or fails depending on how strongly the model prefers the wild-type sequence (Gordon et al., 2025), that scaling does not reliably improve fitness prediction (Hou et al., 2026a), and that masked language modelling can be read as inverse reinforcement learning on the fitness landscape (Jiao et al., 2026). Closest to our analysis, Li et al. (2024) found that almost all downstream protein tasks benefit from PLM pretraining, but that for most of them performance does not scale with pretraining and instead rests on low-level features learned early in pretraining. We ask the complementary question for physical targets that pretraining never observes, and test whether a fitness-based account explains which of them frozen representations recover.

Probing methodology. Linear probes are the standard tool for asking what a representation encodes (Alain & Bengio, 2016), but their conclusions depend on the probe's expressivity and on suitable controls (Hewitt & Liang, 2019; Belinkov, 2021). We follow this literature by reporting accuracy under both a shallow and a deep probe, by comparing against random-weight representations read out by the same probes, and by measuring accuracy relative to a supervised model on the same data.

3 Evolutionary Sufficiency and Competing Explanations

3.1 An Operational Definition

Asking whether a PLM "knows" a physical property is only meaningful relative to what could be learned about that property from the same data without pretraining. We therefore define sufficiency against a supervised reference and make its dependence on the probe explicit.

Let Y be a physical target, $\mathcal{D}_{\text{tr}}$ a training set and $\mathcal{D}_{\text{ood}}$ an out-of-distribution test set. For a frozen representation T and a probe class $\mathcal{P}$ (a predictor trained on T with the backbone held fixed), let $R^2_{T,\mathcal{P}}(Y)$ be the accuracy on $\mathcal{D}_{\text{ood}}$ after fitting the probe on $\mathcal{D}_{\text{tr}}$, and let $R^2_{\text{sup}}(Y)$ be the accuracy of a model trained end-to-end from raw sequence on the same data. We define

$$\text{ES}_{\mathcal{P}}(Y) \;=\; \frac{R^2_{T,\mathcal{P}}(Y)}{R^2_{\text{sup}}(Y)}. \tag{1}$$

Three properties of this definition matter for interpretation. First, ES is not bounded by one, because a representation pretrained on signals related to Y can exceed a model trained from scratch on $\mathcal{D}_{\text{tr}}$ alone. Second, ES depends on the probe, and we therefore report $\text{ES}_{\text{shallow}}$ (mean pooling followed by a two-layer MLP, the common protocol) and ES_{deep} (learned attention pooling followed

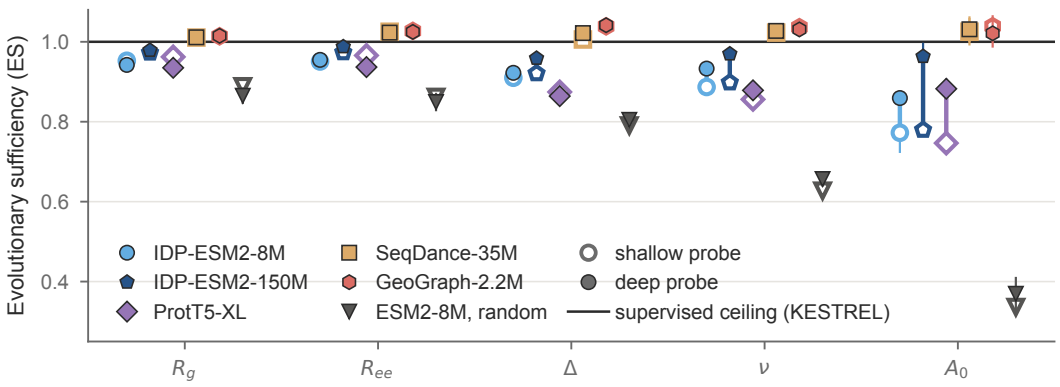

Figure 2: **Evolutionary sufficiency is graded, depends on the readout and is closed by simulator pretraining.** ES (viral R^2 relative to KESTREL on the same split, mean $\pm$ s.e.m. over three seeds) for each frozen backbone read out by the shallow probe (open markers) and the deep probe (filled markers), with vertical lines joining the two probes. Targets are ordered by ES of the shallow IDP-ESM2-8M probe, and the horizontal line marks the supervised ceiling (ES $= 1$).

by a four-layer residual MLP), calling their difference the *readout gap*. Third, ES depends on the denominator, for which we use a one-hot transformer, and we check that the PLM architecture trained from random initialisation reaches the same ceiling (§5.1), ruling out a weak baseline.

ES measures how much of what supervision can learn a frozen representation already provides, but it does not say why a shortfall exists, and we therefore use two further quantities. The *pretraining gain* $\Delta_{\text{pre}} = R^2_{T,\mathcal{P}} - R^2_{T_0,\mathcal{P}}$, where T_0 is the same architecture with random weights, measures what pretraining adds to the representation under a fixed readout, whereas the *fine-tuning gain* Δ_{ft}, the difference in R^2 between fine-tuning from the pretrained and from the random initialisation, measures what the pretrained weights are worth once the whole network is trained.

3.2 The fitness-bottleneck hypothesis

If frozen PLM representations fall short on some targets, the most economical explanation would be one that follows from how PLMs are trained. Masked language modelling on extant sequences can be viewed as maximum-entropy inverse reinforcement learning on the evolutionary fitness landscape (Jiao et al., 2026; Ziebart et al., 2008), so that the log-likelihood approximates fitness up to a selective-stringency constant, $\log p_{\text{PLM}}(s) \approx \beta F(s) - \ln Z(\beta)$. Suppose, in addition, that the representation carries no information about Y beyond what fitness provides, such that

$$\textbf{(H}_{\textbf{FB}}\textbf{)} \qquad I(T;\, Y \mid F) \approx 0. \tag{2}$$

The chain rule of mutual information, $I(T;Y) = I(T;Y \mid F) + I(F;Y) - I(F;Y \mid T)$, then gives $I(T;Y) \leq I(F;Y)$, meaning that the representation can know about Y at most what fitness does and that properties that vary independently of selection would be unrecoverable at any scale. We refer to this as the fitness bottleneck, from which three testable predictions follow.

P1 *Log-likelihood suffices.* A regressor on $\log p_{\text{PLM}}(s)$ and sequence length should do at least as well as a regressor on T for every target (Corollary 1, Appendix K).

P2 *Selection-invisible variables are absent.* The hypothesis was motivated by the contrast between the fraction of charged residues (FCR), which is constrained in disordered regions (Zarin et al., 2019; 2021), and net charge per residue (NCPR), which sets electrostatic chain expansion (Das & Pappu, 2013). Applied to NCPR as the target, Corollary 1 requires NCPR to be no more predictable from T than from $\log p$ and length.

P3 *Readout invariance.* Because the missing information is absent from T, neither a more expressive probe nor a larger model should close a target-specific gap.

We operationalise F by $\log p_{\text{PLM}}$, as the IRL argument implies. A violation of P1 could in principle reflect a poor fitness approximation rather than a failure of H_{FB}, but in that case $\log p_{\text{PLM}}$ is not a usable fitness proxy either, and the hypothesis can no longer predict which properties a PLM will fail on. Moreover, H_{FB} constrains frozen representations only and says nothing about what a PLM can learn when its weights are updated on labelled data.

3.3 COMPETING EXPLANATIONS

A shortfall in ES need not reflect missing information, and if H_{FB} fails, three non-exclusive explanations remain, each with a distinct signature that the controls in §4 can detect.

(R) Readout. The information may be present in T but out of reach of a shallow probe on a mean-pooled embedding, because properties that depend on residue order, such as charge patterning (κ, SCD), are not linear functions of an average over positions. This explanation predicts $\mathrm{ES}_{\mathrm{deep}} > \mathrm{ES}_{\mathrm{shallow}}$, with the readout gap concentrated on targets sensitive to patterning, and random-initialisation controls under the same probe bound what the architecture alone provides.

(D) Pretraining data. A representation reflects what its pretraining signal exposes, and two published encoders were pretrained on IDRome (Tesei et al., 2024), which contains simulations of different (human) sequences with the same coarse-grained model family as our labels. SeqDance learned dynamic properties from these and other molecular-dynamics trajectories (Hou et al., 2026b), whereas GeoGraph learned to predict ensemble geometry, including the five properties we study, and contact-graph descriptors (Quinn et al., 2025). This explanation predicts that both reach $\mathrm{ES} \approx 1$ while frozen, with their advantage largest on sequences resembling their pretraining set.

(S) Supervision. With enough labelled ensembles, the initialisation may matter little once the backbone is trained, in which case $\Delta_{\mathrm{ft}} \approx 0$ even where Δ_{pre} is large.

Table A.1 in Appendix A summarises each prediction, the experiment that tests it and the outcome.

4 EXPERIMENTAL SETUP

4.1 DATA AND SPLITS

We use BENDER (Velasquez & Rahman, 2026), a dataset of IDR conformational ensembles from bacteria, plants, fungi, mammals, protists and viruses, simulated with the CALVADOS-2 coarse-grained model (Tesei et al., 2021). BENDER's ensembles were validated against a cross-taxon set of experimental SAXS radii of gyration that includes viral IDPs, so the viral labels are experimentally anchored. For each ensemble we predict R_g, R_{ee}, ν, Δ and A_0.

Two choices matter for everything that follows, the first being sequence length. A sequence truncated to fit a model's input window is paired with a label computed for the full chain, which penalises the length-dominated targets on which PLMs look strongest, and we therefore restrict every model to sequences of at most 256 residues (9,931 of 10,272) rather than truncating longer ones (Appendix B). The second is the split, in which all 1,015 viral sequences form the out-of-distribution (OOD) test set, such that no model sees a viral sequence or kingdom label during training. BENDER retains one representative per sequence cluster, and we split the non-viral representatives 80/10/10 at random, so no cluster appears in more than one split. Each seed (42, 67, 93) fixes the split and the initialisation, giving 7,441 training, 737 validation and 738 test sequences.[2]

Two of the encoders below, SeqDance and GeoGraph, were pretrained on IDRome (Tesei et al., 2024), 28,058 human IDRs simulated with CALVADOS. Only 39 BENDER sequences (0.4%) also appear in IDRome, and because IDRome is human-only, none of them is viral.

4.2 REPRESENTATIONS, READOUTS AND BASELINES

To separate the pretraining signal, the readout, and the training regime, we compare the configurations in Figure 1. The evolutionary PLMs are IDP-ESM2-8M and IDP-ESM2-150M, which are ESM2-8M and ESM2-150M (Lin et al., 2023) further trained by masked language modelling on IDP-Euka-90, about 30 million eukaryotic IDR sequences (Quinn et al., 2025), and ProtT5-XL (Elnaggar et al., 2022), a much larger model from a different family, whereas SeqDance-35M (Hou et al., 2026b) and GeoGraph-2.2M (Quinn et al., 2025) represent simulator pretraining. As a control without pretraining, we use the ESM2-8M architecture with random weights, redrawn per seed.

[2] 1,546 non-viral sequences without a cluster label are kept together as one block, which falls in the training set for all three seeds and leaves the viral test set unaffected (Appendix B).

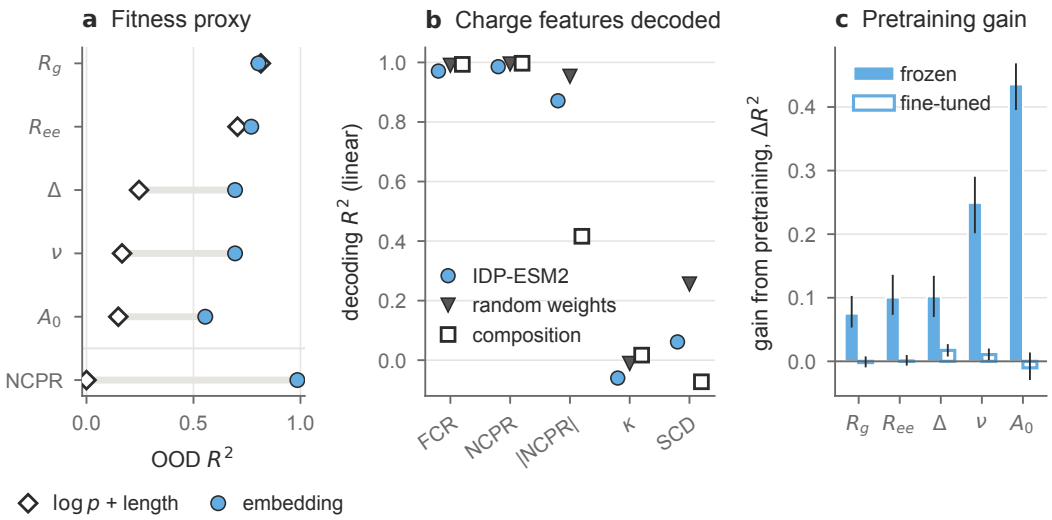

Figure 3: **The shortfall is not a fitness bottleneck.** **a**, Viral R^2 of gradient-boosted trees on $\log p_{\mathrm{PLM}}$ and length (open diamonds) and on the mean-pooled IDP-ESM2-8M embedding (filled circles) for the five targets, and for NCPR (embedding read out linearly, values below zero drawn at zero). Under the fitness bottleneck, each diamond would lie at or beyond its circle (Corollary 1). **b**, Linear decoding of charge descriptors from mean-pooled IDP-ESM2-8M, random-weight and composition features. **c**, Gain in viral R^2 from evolutionary pretraining with the backbone frozen (deep probe, C − D) and fine-tuned (B − A), with paired bootstrap 95% intervals.

The probing protocols are detailed in Appendix M; fine-tuned models update the whole backbone with the shallow head, and KESTREL is trained with its original procedure. All hyperparameters, including learning rates, were selected on validation loss for seed 42 and then fixed, and the viral set was never used for selection (Appendix C).

4.3 Log-likelihood and decoding analyses

To test the fitness bottleneck, we estimate $\log p_{\mathrm{PLM}}(s)$ with a deterministic masked pseudo-log-likelihood that scores every residue once (Appendix D), and regress each target on length, $\log p$, both, the mean-pooled embedding or hand-crafted sequence descriptors (composition, charge and hydropathy patterning), using ridge regression and gradient-boosted trees (GBT). To ask which charge features a representation encodes, we fit ridge probes from mean-pooled embeddings to FCR, NCPR, $|\mathrm{NCPR}|$, κ (Das & Pappu, 2013), SCD and SHD. FCR and NCPR depend only on composition, whereas κ and SCD also depend on how charges are arranged, and we compare IDP-ESM2-8M, random-weight embeddings, and amino-acid composition as representations.

4.4 Evaluation

We report R^2 on the viral test set as the mean ± s.e.m. over three seeds. As all models share the same 1,015 viral sequences, we compare them with paired bootstrap 95% intervals (2,000 resamples, per-seed differences averaged), calling a difference significant when its interval excludes zero.

5 Results

5.1 Frozen evolutionary representations recover a graded fraction of the supervised ceiling

ES is only as reliable as its denominator, so we first asked whether the supervised reference is a fair ceiling. KESTREL and the PLM architecture trained from scratch (A) reach the same accuracy on every target (Table 1), with no difference exceeding 0.012 R^2 and every paired interval including zero. The reference is, however, sensitive to preprocessing, since in an earlier pipeline that truncated

Table 1: OOD R^2 on 1,015 held-out viral sequences (mean ± s.e.m., three seeds), with the corresponding ES shown in Figure 2.

	R_g	R_{ee}	ν	Δ	A_0
Trained end-to-end					
KESTREL (K)	$.972 \pm .001$	$.951 \pm .002$	$.900 \pm .006$	$.866 \pm .005$	$.886 \pm .017$
ESM2-8M, scratch (A)	$.974 \pm .009$	$.963 \pm .009$	$.907 \pm .004$	$.863 \pm .001$	$.890 \pm .017$
IDP-ESM2-8M, fine-tuned (B)	$.972 \pm .006$	$.963 \pm .005$	$.918 \pm .004$	$.880 \pm .003$	$.880 \pm .008$
Frozen, evolutionary					
IDP-ESM2-8M, shallow (E)	$.926 \pm .002$	$.904 \pm .004$	$.798 \pm .013$	$.788 \pm .009$	$.683 \pm .033$
IDP-ESM2-8M, deep (C)	$.916 \pm .006$	$.908 \pm .006$	$.840 \pm .013$	$.799 \pm .017$	$.762 \pm .021$
IDP-ESM2-150M, shallow (O)	$.946 \pm .006$	$.926 \pm .006$	$.809 \pm .001$	$.798 \pm .004$	$.690 \pm .004$
IDP-ESM2-150M, deep (N)	$.952 \pm .005$	$.939 \pm .007$	$.873 \pm .004$	$.830 \pm .004$	$.853 \pm .018$
ProtT5-XL, shallow (J)	$.936 \pm .002$	$.919 \pm .003$	$.771 \pm .004$	$.757 \pm .007$	$.661 \pm .015$
ProtT5-XL, deep (I)	$.909 \pm .012$	$.891 \pm .012$	$.791 \pm .007$	$.749 \pm .010$	$.781 \pm .014$
Frozen, random weights					
ESM2-8M, shallow (F)	$.864 \pm .012$	$.820 \pm .013$	$.565 \pm .002$	$.685 \pm .003$	$.298 \pm .002$
ESM2-8M, deep (D)	$.841 \pm .021$	$.808 \pm .024$	$.591 \pm .014$	$.697 \pm .007$	$.326 \pm .032$
Frozen, simulator-pretrained					
SeqDance-35M, shallow (H)	$.983 \pm .001$	$.975 \pm .002$	$.921 \pm .001$	$.871 \pm .002$	$.907 \pm .017$
SeqDance-35M, deep (G)	$.983 \pm .001$	$.973 \pm .003$	$.925 \pm .004$	$.885 \pm .001$	$.913 \pm .017$
GeoGraph-2.2M, shallow (M)	$.986 \pm .001$	$.977 \pm .001$	$.933 \pm .003$	$.900 \pm .003$	$.920 \pm .015$
GeoGraph-2.2M, deep (L)	$.987 \pm .001$	$.975 \pm .001$	$.928 \pm .003$	$.902 \pm .003$	$.904 \pm .021$
No learned representation					
Sequence descriptors, GBT	$.985 \pm .000$	$.970 \pm .001$	$.860 \pm .003$	$.823 \pm .002$	$.850 \pm .002$

long sequences (and differed in split), KESTREL scored 0.026 R^2 lower on R_g, 0.023 on R_{ee} and 0.080 on A_0, enough to make the length-dominated targets look recoverable (ES ≈ 1, Appendix B).

Against this reference, frozen IDP-ESM2-8M read out by the standard shallow probe recovers a graded fraction of the ceiling (Figure 2), namely 0.95 for R_g and R_{ee}, 0.91 for Δ, 0.89 for ν and 0.77 for A_0, such that no target reaches one. The same ordering holds for IDP-ESM2-150M, which is 19 times larger (0.97, 0.97, 0.92, 0.90 and 0.78), and for ProtT5-XL, which is 375 times larger and from a different model family (0.96, 0.97, 0.87, 0.86 and 0.75), so under this probe scale does not close the gap, and the gradient recurs within each kingdom (Appendix F), whereas contact-network descriptors are recovered almost completely (ES 0.97 to 0.98 for four of five with IDP-ESM2-150M, Appendix H).

A graded shortfall does not mean that evolutionary pretraining is uninformative. Relative to the same architecture with random weights, pretraining adds 0.233 R^2 on ν [0.180, 0.280] and 0.385 on A_0 [0.318, 0.431] under the shallow probe, and 0.249 and 0.436 under the deep probe (Figure 3c), compared with 0.06 to 0.08 on R_g. Taken together, evolutionary pretraining injects substantial information about the targets it recovers least completely, but not enough to match direct supervision.

5.2 The fitness bottleneck does not account for the shortfall

If the representation knew about each target only what fitness does (H_{FB}), the PLM's own log-likelihood, together with length, should predict each target at least as well as the full embedding (P1, Corollary 1). This holds for the length-dominated targets (Figure 3a, Table D.1), for which $\log p$ and length reach 0.78 to 0.82 R^2 on R_g against 0.80 to 0.84 for the embedding, but it fails for the others. On ν, $\log p$ and length reach 0.17 while the embedding reaches 0.56 with ridge regression and 0.69 with boosted trees, a three- to four-fold difference, and Δ and A_0 show the same pattern. The embedding therefore carries information about these targets that the fitness proxy does not.

The second prediction (P2) fails as well, because net charge per residue, the variable whose supposed invisibility to selection motivated the hypothesis, is linearly decodable from frozen IDP-ESM2-8M embeddings with $R^2 = 0.99$ on viral sequences (Figure 3b, Table E.1), whereas $\log p$ and length predict it no better than its mean ($R^2 \leq 0.00$). The embeddings also recover $|\mathrm{NCPR}|$ ($R^2 = 0.87$), which composition itself cannot express linearly ($R^2 = 0.42$). In contrast, no representation exposes charge patterning linearly, as κ is not recovered by any representation ($R^2 \leq 0.02$) and SCD only

weakly ($R^2 \leq 0.26$, with Spearman $\rho = 0.20$ for IDP-ESM2-8M, 0.50 for random weights and 0.43 for composition). Because the same limitation applies to random-weight embeddings and to composition, it reflects mean pooling rather than evolutionary pretraining. Patterning is, however, what the hardest targets require, since boosted trees on hand-crafted descriptors reach 0.86 on ν and 0.85 on A_0, above every frozen evolutionary configuration except the deep-probed IDP-ESM2-150M (Table 1), and without κ, SCD and SHD only 0.59 and 0.28 (Table D.1).

Linear decoding from a mean-pooled embedding is only a lower bound on what the per-residue representation contains, and the role of $\log p$ as a fitness proxy is discussed in §3.2 and Appendix K. Taken together, a fitness bottleneck does not explain the shortfall.

5.3 A POSITION-AWARE READOUT RECOVERS PART OF THE SHORTFALL

If the information is present but hard to reach, a more expressive probe should narrow the gap on the targets that depend most on sequence arrangement (explanation R). Replacing the shallow probe with the deep probe on the same frozen IDP-ESM2-8M embeddings raises R^2 by 0.042 on ν [0.015, 0.068] and 0.079 on A_0 [0.038, 0.127], while leaving R_g (-0.010 [-0.022, 0.000]), R_{ee} and Δ unchanged. ES under the deep probe is 0.94 to 0.96 for R_g, R_{ee}, ν and Δ and 0.86 for A_0 (Figure 2), so for IDP-ESM2-8M the readout accounts for about 40% of the shortfall on ν and on A_0, and the ordering between ν and the length-dominated targets largely collapses.

Part of this gain comes from probe capacity alone, since the same change of probe also helps random-weight embeddings, albeit less, by 0.026 on ν [0.008, 0.044] and 0.028 on A_0 [0.005, 0.050]. The gain is thus about 1.6 times larger on ν, and 2.8 times larger on A_0, for pretrained than for random features, indicating that the pretrained representation holds order-dependent information that the shallow probe cannot use, and per sequence the shallow probe errs most on highly charged sequences (Appendix G). For ProtT5-XL the readout effect is smaller and more target-specific ($+0.120$ on A_0 but $+0.020$ [-0.005, 0.046] on ν), and the deep probe lowers R_g and R_{ee} by about 0.03, which we attribute to overfitting its 1,024-dimensional features. In contrast, the readout gap grows with size within the IDP-ESM2 family, since for IDP-ESM2-150M the deep probe adds 0.064 on ν and 0.162 on A_0, such that the larger model exceeds IDP-ESM2-8M by 0.033 and 0.091 under the deep probe but by at most 0.010 on ν, Δ and A_0 under the shallow probe, and reaches ES of 0.96 to 0.99 while remaining below KESTREL on R_g, ν and Δ (Appendix J). Prediction P3 therefore fails with respect to the probe and, once the readout is position-aware, with respect to scale within the IDP-ESM2 family, whereas ProtT5-XL never exceeds IDP-ESM2-8M on ν or Δ.

5.4 SIMULATOR-MATCHED PRETRAINING CLOSES THE GAP, AND FINE-TUNING MAKES INITIALISATION NEARLY IRRELEVANT

If the remaining gap reflects what the pretraining signal exposes (explanation D), encoders pretrained on simulated ensembles should close it. Both SeqDance and GeoGraph, read out with either probe, match or slightly exceed the supervised ceiling on every target (ES 1.01 to 1.04, Figure 2). Under the deep probe, they exceed frozen IDP-ESM2-8M by 0.085 and 0.088 R^2 on ν ([0.065, 0.108] and [0.069, 0.110]) and by 0.151 and 0.142 on A_0, are indistinguishable from each other on ν and A_0, and barely depend on the probe (at most 0.016 R^2). GeoGraph does so with 2.2 million parameters, 1,400 times fewer than ProtT5-XL, and also exceeds IDP-ESM2-150M under the deep probe (by 0.055 on ν), although it was trained to predict these properties on IDRome.

Because both encoders were pretrained on human IDRs, we asked whether their advantage depends on similarity to that pretraining set. Relative to the architecture-matched scratch model (A), both gain most in mammals (SeqDance by $+0.128$ on ν and $+0.238$ on A_0, GeoGraph by $+0.155$ and $+0.327$) and next in plants. Elsewhere, SeqDance's advantage is close to zero, while GeoGraph keeps a smaller one ($+0.03$ to $+0.08$), and both gain little on viruses ($\approx +0.02$ on ν, Figure F.1 and Appendix F). The 39 sequences shared with IDRome are too few to explain this ordering, which instead points to transfer that decays with distance from the pretraining domain. Because per-kingdom test sets are small (44 to 270 sequences per seed), we treat only this ordering as robust.

Finally, we asked whether the evolutionary initialisation still matters once the backbone is trained (explanation S). Fine-tuning IDP-ESM2-8M end-to-end (B) recovers the full ceiling, raising ν by 0.078 over the frozen deep probe, and differs from the same architecture trained from random

weights (A) by at most 0.017 R^2, namely $+0.011$ on ν [0.002, 0.020] and $+0.017$ on Δ [0.008, 0.027], with no significant difference on R_g, R_{ee} or A_0. The fine-tuning gain ($\Delta_{\mathrm{ft}} \leq 0.02$, Figure 3c) is thus an order of magnitude smaller than the pretraining gain for frozen features ($\Delta_{\mathrm{pre}} \approx 0.23$ on ν). Taken together, with about 7,400 labelled ensembles, the pretraining signal matters mainly for frozen backbones, and simulator pretraining brings them to the ceiling.

6 Discussion

Why ν and A_0 are hardest. Both are fitted from how intramolecular distances scale with sequence separation and depend on how charged and hydrophobic residues are arranged along the chain, which mean pooling largely discards. Consistent with this, κ, SCD and SHD add 0.27 and 0.58 R^2 on ν and A_0 to composition-based descriptors, no mean-pooled representation linearly exposes κ or SCD, and a deeper probe helps pretrained more than random features. Pooling may therefore hide patterning that pretraining encodes, or evolutionary statistics may reflect patterning only indirectly, and the residual deep-probe gap (ES of 0.86 to 0.97 on ν and A_0) suggests both contribute.

What the negative result on the fitness bottleneck means. The information-theoretic argument is valid, but its premise fails, because frozen embeddings predict ν, Δ and A_0 several times better than the model's own likelihood and encode net charge almost perfectly. Unlike fitness (Meier et al., 2021; Gordon et al., 2025), which properties a PLM recovers cannot be read off from selection.

Pretraining data matters more than scale. A 19-fold larger IDP-ESM2 narrows the gap on ν, Δ and A_0 only with the deep probe, and a 375-fold larger PLM from another family does not narrow it, whereas two encoders 85 to 1,400 times smaller, pretrained on simulated ensembles, close it even with a shallow probe, most strongly for mammalian sequences. Once the whole network is trained on about 7,400 labelled ensembles, initialization matters little, consistent with reports that PLM pretraining gains often rest on low-level features (Li et al., 2024), so the pretraining signal matters most for frozen feature extractors.

Practical guidance. (1) Report ES with its baseline and readout, and check baselines for artifacts such as truncation. (2) Frozen evolutionary features suffice for chain dimensions, whereas ν, Δ and A_0 call for a position-aware probe, fine-tuning or an encoder pretrained on simulated ensembles. (3) fine-tune the whole model, since the choice of initialization then changes R^2 by at most 0.02.

Limitations. First, we studied global properties of isolated chains, not local structure or interactions with folded domains or other chains. Second, we fine-tuned only IDP-ESM2-8M with one label budget, and initialization may matter more for larger models or fewer labels.

7 Conclusion

Under a matched protocol, frozen evolutionary PLM representations recover nearly all of what supervision learns about chain dimensions but progressively less about Δ, ν and A_0. A fitness bottleneck does not account for this shortfall, since the embeddings carry far more information about these properties than the model's own likelihood, whereas a position-aware probe narrows it, and pretraining on simulated ensembles closes it. Taken together, these results show that frozen evolutionary PLMs capture a large but graded share of disordered-ensemble physics, and reporting this sufficiency together with its baseline and readout, while matching the pretraining signal to the target physics, should make them more reliable tools for studying sequence–ensemble relationships.

USE OF LARGE LANGUAGE MODELS

We used Claude (Anthropic), through the Claude Code agent, to give feedback on the experimental design, refine hypotheses, formulate the formal statements of Appendix K, implement and run analyses (including data filtering and splits). We did not use it to generate or alter data, and translation and qualitative analysis do not apply. All numbers, tables, and figures were regenerated from the released result files, and the authors reviewed all AI-assisted code, analyses, and text and take full responsibility for the content of this paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. URL https://arxiv.org/abs/1610.01644. Version Number: 4.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: Mutual Information Neural Estimation. 2018. doi: 10.48550/ARXIV.1801.04062. URL https://arxiv.org/abs/1801.04062. Version Number: 5.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances, 2021. URL https://arxiv.org/abs/2102.12452. Version Number: 4.

Arthur Bigot, Harmon Bhasin, Core Francisco Park, Eugene Shakhnovich, and Dianzhuo Wang. Viral Proteins Reveal Geometry of Protein Language Models, 2026. URL https://arxiv.org/abs/2606.12609. Version Number: 1.

Rahul K. Das and Rohit V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, August 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1304749110. URL https://pnas.org/doi/full/10.1073/pnas.1304749110.

Frances Ding and Jacob Steinhardt. Protein language models are biased by unequal sequence sampling across the tree of life. *BioRxiv*, pp. 2024.03. 07.584001, 2024.

H. Jane Dyson and Peter E. Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197–208, March 2005. ISSN 1471-0072, 1471-0080. doi: 10.1038/nrm1589. URL https://www.nature.com/articles/nrm1589.

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, October 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3095381. URL https://ieeexplore.ieee.org/document/9477085/.

Cade Gordon, Amy Lu, and Pieter Abbeel. Protein language model fitness is a matter of preference. In *International Conference on Learning Representations*, volume 2025, pp. 39759–39775, 2025.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks, 2019. URL https://arxiv.org/abs/1909.03368. Version Number: 1.

Alex S. Holehouse and Birthe B. Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, 25(3):187–211, March 2024. ISSN 1471-0072, 1471-0080. doi: 10.1038/s41580-023-00673-0. URL https://www.nature.com/articles/s41580-023-00673-0.

Alex S. Holehouse, Rahul K. Das, James N. Ahad, Mary O.G. Richardson, and Rohit V. Pappu. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal*, 112(1):16–21, January 2017. ISSN 00063495. doi: 10.1016/j.bpj.2016.11.3200. URL https://linkinghub.elsevier.com/retrieve/pii/S0006349516342692.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding language model scaling for protein fitness prediction. *Nature Computational Science*, 6(7):778–788, July 2026a. ISSN 2662-8457. doi: 10.1038/s43588-026-01010-z. URL https://www.nature.com/articles/s43588-026-01010-z.

Chao Hou, Haiqing Zhao, and Yufeng Shen. Protein language models trained on biophysical dynamics inform mutation effects. *Proceedings of the National Academy of Sciences*, 123(4): e2530466123, January 2026b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2530466123. URL https://pnas.org/doi/10.1073/pnas.2530466123.

Giacomo Janson, Alexander Jussupow, and Michael Feig. Deep generative modeling of temperature-dependent structural ensembles of proteins, March 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.03.09.642148.

Xiaoran Jiao, Shengdong Lin, Jigang Fan, Zhanming Liang, Weian Mao, Hao Chen, and Chunhua Shen. Evolutionary profiles for protein fitness prediction. *Bioinformatics*, 42 (8):btag525, August 2026. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btag525. URL https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btag525/8738351.

Jerelle A. Joseph, Aleks Reinhardt, Anne Aguirre, Pin Yu Chew, Kieran O. Russell, Jorge R. Espinosa, Adiran Garaizar, and Rosana Collepardo-Guevara. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nature Computational Science*, 1(11):732–743, November 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00155-3. URL https://www.nature.com/articles/s43588-021-00155-3.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pp. 2024.02.05.578959, 2024.

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, March 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Jeffrey M. Lotthammer, Garrett M. Ginell, Daniel Griffith, Ryan J. Emenecker, and Alex S. Holehouse. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nature Methods*, 21(3):465–476, March 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-023-02159-5. URL https://www.nature.com/articles/s41592-023-02159-5.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function, July 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.07.09.450648.

Borna Novak, Jeffrey M. Lotthammer, Ryan J. Emenecker, and Alex S. Holehouse. Accurate predictions of disordered protein ensembles with STARLING. *Nature*, 652(8108):240–250, April 2026. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-026-10141-2. URL https://www.nature.com/articles/s41586-026-10141-2.

Eoin Quinn, Marco Carobene, Jean Quentin, Sebastien Boyer, Miguel Arbesú, and Oliver Bent. GeoGraph: Geometric and Graph-based Ensemble Descriptors for Intrinsically Disordered Proteins, 2025. URL https://arxiv.org/abs/2510.00774. Version Number: 2.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021. ISBN 2640-3498.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2016239118. URL https://pnas.org/doi/full/10.1073/pnas.2016239118.

Rajan Sawhney, Barbra D. Ferrell, Thibaut Dejean, Zachary Schreiber, William Harrigan, Shawn W. Polson, K. Eric Wommack, and Mahdi Belcaid. Fine-tuning protein language models unlocks the potential of underrepresented viral proteomes. *PeerJ*, 13:e19919, September 2025. ISSN 2167-8359. doi: 10.7717/peerj.19919. URL https://peerj.com/articles/19919.

Lucas Sawle and Kingshuk Ghosh. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *The Journal of Chemical Physics*, 143(8):085101, August 2015. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.4929391. URL https://pubs.aip.org/jcp/article/143/8/085101/73623/A-theoretical-method-to-compute-sequence-dependent.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, 2017. URL https://arxiv.org/abs/1703.01365. Version Number: 2.

Giulio Tesei and Kresten Lindorff-Larsen. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Research Europe*, 2:94, January 2023. ISSN 2732-5121. doi: 10.12688/openreseurope.14967.2. URL https://open-research-europe.ec.europa.eu/articles/2-94/v2.

Giulio Tesei, Thea K. Schulze, Ramon Crehuet, and Kresten Lindorff-Larsen. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proceedings of the National Academy of Sciences*, 118(44):e2111696118, November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2111696118. URL https://pnas.org/doi/full/10.1073/pnas.2111696118.

Giulio Tesei, Anna Ida Trolle, Nicolas Jonsson, Johannes Betz, Frederik E. Knudsen, Francesco Pesce, Kristoffer E. Johansson, and Kresten Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626(8000):897–904, February 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-07004-5. URL https://www.nature.com/articles/s41586-023-07004-5.

Peter Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, October 2002. ISSN 09680004. doi: 10.1016/S0968-0004(02)02169-2. URL https://linkinghub.elsevier.com/retrieve/pii/S0968000402021692.

Vladimir N. Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, 11(4):739–756, April 2002. ISSN 0961-8368, 1469-896X. doi: 10.1110/ps.4210102. URL https://onlinelibrary.wiley.com/doi/10.1110/ps.4210102.

Juan Velasquez and Taseef Rahman. BENDER: A Cross-taxon IDP Simulation Database Reveals Conserved Sequence-Ensemble Laws Across the Tree of Life, August 2026. URL http://biorxiv.org/lookup/doi/10.64898/2026.08.18.745604.

Taraneh Zarin, Bob Strome, Alex N Nguyen Ba, Simon Alberti, Julie D Forman-Kay, and Alan M Moses. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife*, 8:e46883, July 2019. ISSN 2050-084X. doi: 10.7554/eLife.46883. URL https://elifesciences.org/articles/46883.

Taraneh Zarin, Bob Strome, Gang Peng, Iva Pritišanac, Julie D Forman-Kay, and Alan M Moses. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *eLife*, 10:e60220, February 2021. ISSN 2050-084X. doi: 10.7554/eLife.60220. URL https://elifesciences.org/articles/60220.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A SUMMARY OF HYPOTHESES AND TESTS

Table A.1: Hypotheses, predictions and tests, with outcomes summarised here and reported in §5.

Hypothesis	Prediction	Test	Outcome
H_{FB}, P1	$\log p$ + length $\approx T$ for all Y	log-likelihood regressors	rejected (§5.2)
H_{FB}, P2	NCPR not decodable from T	linear decoding of charge features	rejected (§5.2)
H_{FB}, P3	gap invariant to probe and scale	shallow vs. deep probe, 8M vs. 150M vs. 3B	rejected (§5.3)
(R) Readout	$\mathrm{ES}_{\mathrm{deep}} > \mathrm{ES}_{\mathrm{shallow}}$ on ν, A_0	probe comparison, per-sequence errors, descriptor ablation	supported (§5.3, App. G)
(D) Pretraining data	simulator-pretrained ES ≈ 1, domain-dependent	SeqDance, GeoGraph, per-kingdom	supported (§5.4)
(S) Supervision	$\Delta_{\mathrm{ft}} \approx 0$	fine-tuned vs. scratch	supported (§5.4)

B DATA, FILTERING AND SPLITS

Sequences. We use the BENDER sequences with all five geometric targets (10,272 sequences), of which 9,931 remain after restricting to at most 256 residues (Table B.1), the excluded sequences being 257 to 1,181 residues long.

Table B.1: Sequences per kingdom before and after the length filter, and the test-set size per seed (seeds 42 / 67 / 93). Viral sequences are all used as the OOD test set.

Kingdom	All	$\leq$ 256 aa	Test set per seed
Bacteria	2,850	2,812	262 / 270 / 249
Plants	2,480	2,396	210 / 209 / 239
Fungi	1,507	1,406	92 / 90 / 86
Mammals	1,361	1,274	130 / 121 / 109
Protists	1,049	1,028	44 / 48 / 55
Viruses	1,025	1,015	OOD (1,015)

Splits. BENDER retains one representative per sequence cluster, so every cluster label occurs once, and we shuffle cluster labels with the seed and assign 80%, 10% and 10% of them to training, validation and test. For 1,546 non-viral sequences, all from a subset of BENDER whose scaling fits were recalculated, the cluster label is missing, and the split treats them as a single block. This block was assigned to training for all three seeds, which is why the split sizes (7,441 / 737 / 738) are identical across seeds even though the test sets differ (70 of 738 test sequences are shared between seeds 42 and 67). Because the block is over-represented in protists (554 of 1,028) and fungi (528 of 1,406), their in-distribution test sets are small, whereas the viral test set is unaffected and all models use the same training sets.

Overlap with IDRome. SeqDance and GeoGraph were pretrained on IDRome, which shares 39 sequences (0.4%) with BENDER. Since IDRome contains only human IDRs, none of these is viral, and removing them does not change any conclusion that relies on the viral test set.

Truncation and the supervised baseline. KESTREL's original pipeline truncates sequences to 256 residues while keeping the labels of the full chains, and it evaluates 1,025 viral sequences instead of 1,015 and uses a different split. Table B.2 compares its viral R^2 with the matched run, which differs by 0.023 to 0.080 on R_g, R_{ee} and A_0 but by only 0.002 on ν and Δ, and we did not isolate which of the three differences is responsible.

C MODEL AND TRAINING DETAILS

Fine-tuning. Arms A and B attach the shallow head to the ESM2-8M backbone (random or IDP-ESM2-8M weights) and train all parameters with AdamW (weight decay 10^{-2}), a head learning rate

Table B.2: KESTREL on the viral test set under its original pipeline (Velasquez & Rahman, 2026) (truncated sequences) and under the matched protocol (sequences ≤ 256 residues, three seeds).

	R_g	R_{ee}	ν	Δ	A_0
Original pipeline ($n = 1{,}025$)	.946	.928	.898	.864	.806
Matched protocol ($n = 1{,}015$)	.972	.951	.900	.866	.886

of 10^{-3}, 5% linear warm-up followed by cosine decay, batch size 32, up to 40 epochs, and patience 6. The backbone learning rate was selected as in Table C.1.

KESTREL. We retrain KESTREL (Velasquez & Rahman, 2026) on our splits with its original architecture and training procedure, namely a four-layer pre-norm transformer (width 256, eight heads) over one-hot sequences with sinusoidal positions, a 16-dimensional kingdom embedding (zero for viral sequences) and separate heads for geometric and contact-network targets, trained jointly with batch size 256 for up to 100 epochs with patience 15, and only its learning rate was selected (Table C.1).

Selection and early stopping. Every model is early-stopped on a three-epoch moving average of validation loss (mean-squared error on training-set-standardised targets), and learning rates were selected by the same criterion on seed 42 and then fixed.

Table C.1: Learning-rate selection on the seed-42 validation set (best smoothed validation loss, lower is better), with selected values in bold.

Model	Learning rate	Validation loss
A (scratch, backbone LR)	$\mathbf{1 \times 10^{-4}}$ / 3×10^{-4}	**0.145** / 0.171
B (IDP-ESM2-8M, backbone LR)	2×10^{-5} / $\mathbf{5 \times 10^{-5}}$	0.123 / **0.118**
K (KESTREL)	2 / **5** / 10 / 20×10^{-4}	0.231 / **0.216** / 0.228 / 0.357

D LOG-LIKELIHOOD AND DESCRIPTOR BASELINES

We estimate $\log p_{\mathrm{PLM}}(s)$ with IDP-ESM2-8M in seven forward passes, where pass k masks every residue whose position is congruent to k modulo 7, so that each residue is masked once at a rate close to the pretraining rate, and $\log p$ is the sum of the log-probabilities of the true residues at masked positions. Unlike random masking, the estimate has no sampling noise, and it agrees with perplexities from five random 15% masks (Spearman $\rho = 0.85$ on the viral set). The regressors are RidgeCV over $\alpha \in [10^{-3}, 10^{4}]$ on standardised features and histogram gradient-boosted trees (learning rate 0.05, up to 1,000 iterations, early stopping on the validation split), and Table D.1 gives the results, with an s.e.m. across seeds of at most 0.02.

The same regressors are fitted to hand-crafted sequence descriptors of the kind computed by local-CIDER (Holehouse et al., 2017), in three nested sets, namely the 20 amino-acid fractions and length, these together with FCR, NCPR, $|\mathrm{NCPR}|$ and mean hydropathy, which are functions of composition given explicitly, and finally all of these together with the patterning descriptors κ, SCD and SHD, all computed from sequence and taken from the dataset. The difference between the last two sets isolates what patterning adds, which with boosted trees is 0.27 R^2 on ν, 0.58 on A_0 and 0.11 on Δ, compared with 0.03 to 0.06 on R_g and R_{ee}. With all descriptors, boosted trees reach ES of 1.01 and 1.02 on R_g and R_{ee} and 0.95 to 0.96 on ν, Δ and A_0, whereas ridge regression on the same features stays far lower on ν and A_0, indicating that these targets depend nonlinearly on the descriptors.

E CHARGE-DESCRIPTOR DECODING

FCR is the fraction of charged residues and NCPR the net charge per residue, whereas κ (Das & Pappu, 2013) measures the segregation of oppositely charged residues relative to a well-mixed sequence of the same composition, and SCD (Sawle & Ghosh, 2015) is a sequence-separation-weighted sum over charge pairs, with SHD the analogous hydropathy decoration. All values are taken from the dataset, and the probes are RidgeCV models on standardised mean-pooled representations, trained on the training split and evaluated on the viral set. Because SCD is heavy-tailed (up to about 1,050), we also report Spearman correlations (Table E.2).

Table D.1: Predicting each target from the PLM's pseudo-log-likelihood, from the mean-pooled embedding and from hand-crafted sequence descriptors, reported as OOD R^2 (mean of three seeds) for ridge / gradient-boosted trees.

Features	R_g	R_{ee}	ν	Δ	A_0
length	.735 / .769	.607 / .635	.090 / .023	.104 / .088	.205 / .069
$\log p$	.404 / .415	.303 / .303	.169 / .146	.203 / .231	.202 / .135
$\log p$ + length	.783 / .815	.673 / .706	.168 / .166	.208 / .246	.218 / .149
embedding (mean-pooled)	.840 / .804	.809 / .770	.560 / .694	.601 / .694	.424 / .555
composition + length	.713 / .934	.588 / .887	.103 / .540	.162 / .689	.197 / .359
+ FCR, NCPR, $\lvert\mathrm{NCPR}\rvert$, hydropathy	.829 / .954	.766 / .912	.292 / .590	.530 / .717	.046 / .275
+ κ, SCD, SHD	.861 / .985	.799 / .970	.432 / .860	.609 / .823	.263 / .850

Table E.1: Linear decoding of charge descriptors from mean-pooled representations (OOD R^2, mean of three seeds).

Representation	FCR	NCPR	$\lvert\mathrm{NCPR}\rvert$	κ	SCD	SHD
IDP-ESM2-8M	.971	.986	.871	$-.060$	.061	.079
ESM2-8M, random	.990	.995	.953	$-.011$	.256	.087
Composition (20-d)	.993	.997	.416	.017	$-.073$	.141

Integrated gradients. Decoding shows what a representation encodes but not what a trained readout uses. To ask the latter, we computed integrated gradients (Sundararajan et al., 2017) of the ν output on the 1,015 viral sequences from a zero baseline, with respect to the one-hot input of KESTREL and to the token embeddings of IDP-ESM2-8M, with gradients passing through the frozen backbone and the shallow probe, and took the norm of each residue's integrated-gradient vector as its attribution. Both models come from single seed-42 runs that precede the matched protocol, namely KESTREL trained with its original pipeline (truncated sequences and its own split) and the shallow IDP-ESM2-8M probe of our initial pipeline, so this analysis is qualitative. Charged residues (R, K, D and E) receive on average 6.84 times the attribution of uncharged residues in KESTREL (median 6.44) but only 1.17 times in IDP-ESM2-8M (median 1.14). Hence, although net charge is almost perfectly decodable from the frozen embedding, the shallow readout's predictions of ν are barely more sensitive to charged residues than to any other, which is consistent with the readout component of the gap.

F Per-kingdom analysis

Figure F.1 and Table F.1 compare the two frozen simulator-pretrained encoders (deep probe) with the scratch model (A) within each kingdom, using the in-distribution test sets for non-viral kingdoms, whose sizes vary by seed (Table B.1).

ES within kingdoms. Figure F.2 asks whether the gradient of §5.1 is specific to the viral test set. Because the in-distribution test sets are small, we pool the test predictions of the three seeds within each kingdom (135 to 1,015 distinct sequences) and compute ES as the ratio of the pooled R^2 to KESTREL's on the same sequences, which on viral sequences reproduces the values in Figure 2. R_g and R_{ee} reach ES ≥ 0.91 in every kingdom, for IDP-ESM2-8M and ProtT5-XL under both probes. Under the shallow probe, ES for ν and A_0 is below that for R_g in every kingdom, with one exception (ν in fungi for IDP-ESM2-8M, where the interval is wide), so the gradient is not a property of viral sequences alone, although its depth varies. For IDP-ESM2-8M the shortfall on ν and A_0 is largest in mammals (ES 0.65 and 0.38) and protists (0.73 and 0.76), and the deep probe closes it in bacteria and fungi (0.94 to 0.99) but not in mammals or protists. Notably, mammals are also the kingdom in which the simulator-pretrained encoders gain most (Table F.1). Although KESTREL receives the kingdom label for non-viral sequences and the probes do not, the scratch model (A), which also lacks the label, matches KESTREL in distribution on every target except A_0 (Table I.1), so the label is unlikely to explain these differences.

Table E.2: Spearman correlation between decoded and true descriptors on the viral set (mean of three seeds), complementing Table E.1.

Representation	FCR	NCPR	|NCPR|	κ	SCD	SHD
IDP-ESM2-8M	.987	.991	.863	.126	.199	.040
ESM2-8M, random	.995	.997	.924	.258	.495	.091
Composition (20-d)	.997	.999	.644	.247	.427	.198

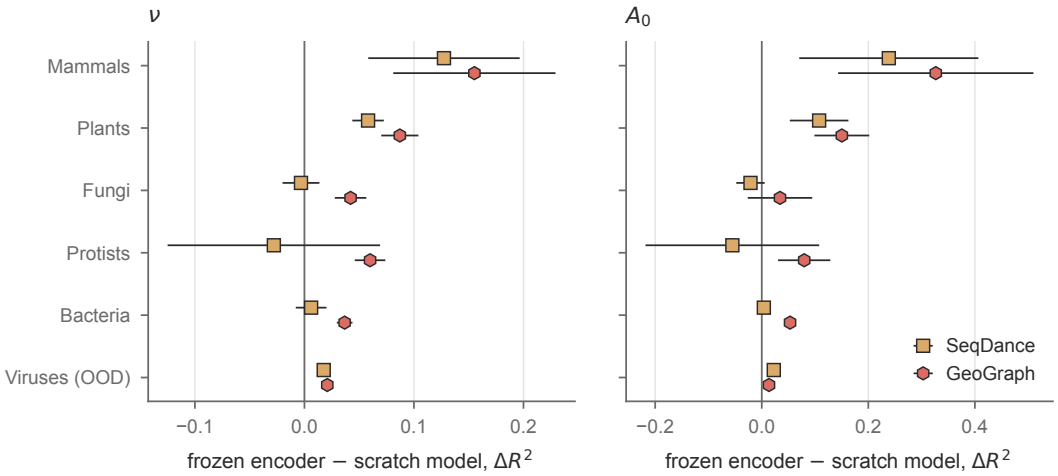

Figure F.1: Advantage of frozen SeqDance and GeoGraph (deep probe) over the scratch model (A) by kingdom, given as ΔR^2 for ν and A_0 (mean $\pm$ s.e.m. over three seeds, computed on the test sets for non-viral kingdoms and on the OOD set for viruses).

G WHERE THE FROZEN PROBES ERR

To locate the readout gap of §5.3, we compared per-sequence errors on the viral set. For each configuration we averaged the predictions of the three seeds, and defined the excess error of a frozen probe as its absolute error minus KESTREL's on the same sequence, in units of the target's standard deviation on the viral set (Figure G.1). The shallow probe's excess error grows with the charge of a sequence, whatever its sign, as it correlates with FCR (Spearman $\rho = 0.14$ for ν and 0.22 for A_0) and with $|\mathrm{NCPR}|$ (0.13 and 0.20) and is largest for sequences with NCPR above 0.15 or below -0.05. The tenth of sequences with the largest excess error on ν is correspondingly enriched in both acidic and basic residues. Since FCR and NCPR are linearly decodable from the same embeddings ($R^2 \geq 0.97$, Appendix E), the shallow probe has access to the charge of these sequences but does not convert it into ν and A_0 as accurately as direct supervision does.

The deep probe removes this dependence for net-positive sequences, reducing the excess error for NCPR above 0.15 to 0.014 s.d. on ν and 0.015 on A_0, against 0.124 and 0.186 for the shallow probe, while its correlation with FCR falls to 0.00 and 0.05. In contrast, net-negative sequences are not helped (0.154 and 0.220 s.d., against 0.134 and 0.230), such that they make up 18% of the viral set but account for 58% of the deep probe's total excess error on ν and 66% on A_0, and the sequences the deep probe predicts worst are enriched in acidic and depleted in basic residues. Because the sign of the net charge is decodable from the same embeddings (NCPR, $R^2 = 0.99$), this residual does not reflect missing charge information, and it may instead reflect features that neither probe extracts, such as charge patterning, which no representation exposes linearly (Appendix E), although we did not resolve its cause.

Compression of the extremes. Figure G.2 plots predicted against true values of ν and A_0 on the viral set. Both frozen IDP-ESM2-8M probes compress the range of these targets, with a slope of predicted on true of 0.81 for ν and 0.62 to 0.64 for A_0, against 0.92 and 0.85 for KESTREL and 0.90 and 0.84 for GeoGraph. The compression is concentrated in the most compact sequences, since the 41 viral sequences with $\nu < 0.4$ are predicted to be expanded by 0.07 (shallow probe) and 0.06 (deep probe), against 0.03 for KESTREL and 0.02 for GeoGraph, and the 24 sequences with $A_0 > 1$ are underestimated by about 0.31, against 0.09 and 0.12. The deep probe therefore raises

Table F.1: Advantage of frozen SeqDance and GeoGraph (deep probe) over the scratch model (A) by kingdom, given as ΔR^2 averaged over three seeds (test sets for non-viral kingdoms and the OOD set for viruses, with n the mean test size). R_g and R_{ee} differ by at most 0.02 (SeqDance) and 0.04 (GeoGraph) in every kingdom.

		SeqDance			GeoGraph		
Kingdom	n	ν	Δ	A_0	ν	Δ	A_0
Mammals	120	+.128	+.083	+.238	+.155	+.089	+.327
Plants	219	+.058	+.077	+.108	+.087	+.111	+.150
Fungi	89	−.004	+.002	−.021	+.042	+.047	+.034
Protists	49	−.027	−.011	−.055	+.060	+.073	+.080
Bacteria	260	+.006	+.007	+.004	+.037	+.042	+.053
Viruses	1,015	+.018	+.022	+.023	+.021	+.039	+.014

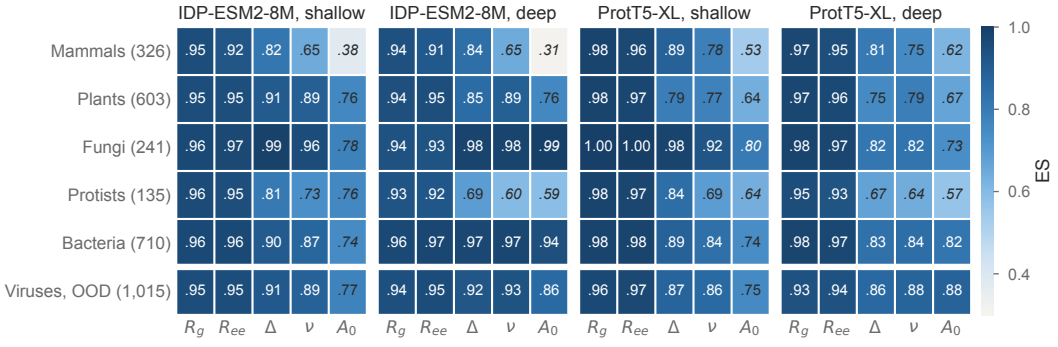

Figure F.2: ES within each kingdom for the frozen evolutionary PLMs. ES is the R^2 of each configuration divided by KESTREL's, computed on the test predictions of the three seeds pooled within a kingdom (distinct sequences in parentheses, with viral sequences forming the OOD set). Italic values have a paired-bootstrap 95% interval wider than 0.4.

R^2 by reducing scatter rather than by correcting the compression, whereas an encoder pretrained on simulated ensembles largely avoids it.

H CONTACT-NETWORK DESCRIPTORS

BENDER also provides five descriptors of the residue contact networks computed from the same trajectories, namely global efficiency, fragmentation index, average clustering, transitivity and degree assortativity. We trained both probes on these descriptors for IDP-ESM2-150M and SeqDance with the data, splits, and seeds of the main analysis, and used the contact-network heads that KESTREL trains jointly with its geometric heads as the reference (Table H.1). GeoGraph is omitted because it was pretrained to predict contact-graph descriptors. SeqDance was pretrained to predict residue-pair interaction frequencies from simulations, which are the raw material of contact networks, so its results should be read with that in mind, whereas IDP-ESM2-150M was pretrained on sequences alone.

Frozen IDP-ESM2-150M recovers these descriptors almost completely, with ES of 0.97 to 0.98 under the shallow probe for all but the fragmentation index, which varies little across sequences (s.d. 0.008) and is poorly predicted by every model ($R^2 \leq 0.42$). Global efficiency and assortativity largely track chain size (Spearman $|\rho| \geq 0.88$ with R_g), whereas clustering and transitivity are more closely related to chain shape (ρ of -0.52 to -0.58 with ν and Δ), yet the latter two are also recovered with ES of 0.97 to 0.98, above the 0.90 and 0.92 that the same model and probe reach for ν and Δ. The deep probe brings clustering and transitivity to the ceiling (ES 0.99 and 1.00, with differences from KESTREL of -0.009 [-0.020, 0.002] and 0.000 [-0.012, 0.013]) but slightly lowers global efficiency and assortativity. SeqDance matches the ceiling to within 0.01 on every descriptor, exceeds it on clustering and transitivity, and exceeds IDP-ESM2-150M under both probes (by 0.012 to 0.083 R^2, all intervals excluding zero). Taken together, contact-network topology is recovered more completely by a frozen evolutionary PLM than the scaling descriptors ν and A_0 are.

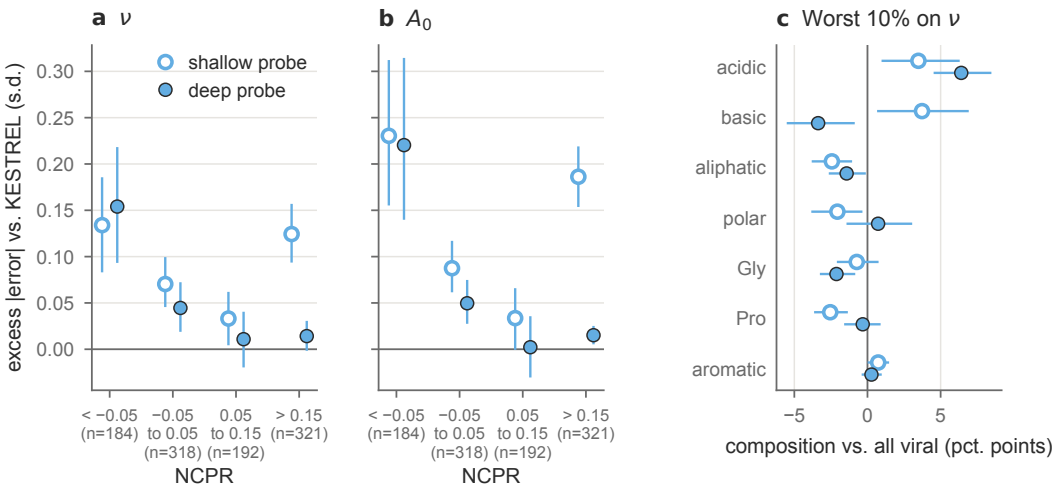

Figure G.1: Where the frozen IDP-ESM2-8M probes err on the viral set. Excess error is a probe's absolute error minus KESTREL's on the same sequence, in units of the target's standard deviation, with predictions averaged over three seeds. **a**, **b**, Mean excess error by net charge per residue for ν and A_0 (bootstrap 95% intervals). **c**, Residue-class composition of the 10% of sequences with the largest excess error on ν, minus the viral-set mean (acidic DE, basic KR, aliphatic AILMV, polar STNQ, aromatic FWY).

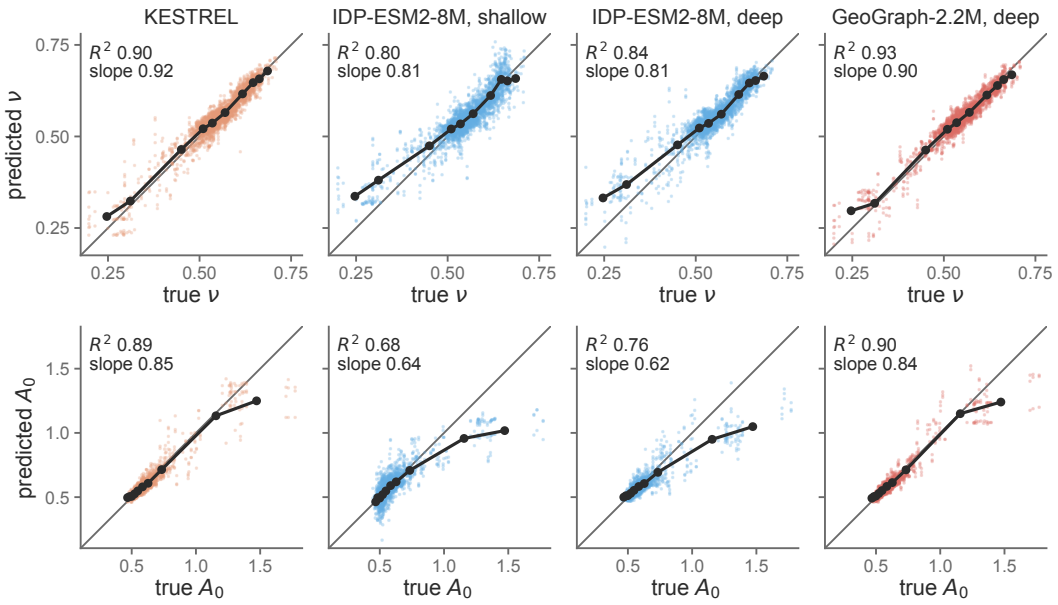

Figure G.2: Predicted against true ν and A_0 on the viral set for KESTREL, the frozen IDP-ESM2 probes and frozen GeoGraph (deep probe), with the predictions of all three seeds pooled. The dark line joins the mean prediction in bins of the true value, which are narrower in the tails (quantiles 0, 1, 3, 10, 25, 50, 75, 90, 97, 99 and 100%), and each panel gives R^2 and the slope of predicted on true, averaged over seeds.

I In-distribution results

Table I.1 reports R^2 on the non-viral test sets, where the ordering of configurations matches the viral results. R^2 for ν, Δ and A_0 is lower in distribution than on viral sequences for every model, which suggests that the viral set spans a narrower, easier range of these targets rather than that any model generalises better out of distribution.

Table H.1: Contact-network descriptors on the viral test set, R^2 (mean $\pm$ s.e.m., three seeds). KESTREL trains its contact-network heads jointly with the geometric ones, whereas the probes are trained on the network descriptors alone.

	Global efficiency	Fragmentation index	Average clustering	Transitivity	Degree assortativity
KESTREL (K)	$.982 \pm .000$	$.408 \pm .020$	$.862 \pm .003$	$.844 \pm .002$	$.838 \pm .002$
IDP-ESM2-150M, shallow (O)	$.952 \pm .002$	$.357 \pm .013$	$.832 \pm .005$	$.829 \pm .004$	$.820 \pm .001$
IDP-ESM2-150M, deep (N)	$.935 \pm .008$	$.325 \pm .021$	$.853 \pm .002$	$.844 \pm .001$	$.795 \pm .003$
SeqDance-35M, shallow (H)	$.967 \pm .006$	$.421 \pm .017$	$.879 \pm .003$	$.859 \pm .003$	$.833 \pm .004$
SeqDance-35M, deep (G)	$.974 \pm .004$	$.408 \pm .014$	$.875 \pm .002$	$.861 \pm .002$	$.838 \pm .003$

Table I.1: In-distribution test R^2 (non-viral kingdoms, mean $\pm$ s.e.m. over three seeds).

	R_g	R_{ee}	ν	Δ	A_0
KESTREL (K)	$.984 \pm .001$	$.965 \pm .003$	$.816 \pm .006$	$.754 \pm .007$	$.771 \pm .015$
ESM2-8M, scratch (A)	$.983 \pm .002$	$.964 \pm .003$	$.819 \pm .019$	$.756 \pm .007$	$.715 \pm .068$
IDP-ESM2-8M, fine-tuned (B)	$.983 \pm .001$	$.970 \pm .004$	$.859 \pm .007$	$.784 \pm .004$	$.802 \pm .035$
IDP-ESM2-8M, shallow (E)	$.944 \pm .008$	$.917 \pm .016$	$.688 \pm .020$	$.676 \pm .009$	$.536 \pm .056$
IDP-ESM2-8M, deep (C)	$.934 \pm .006$	$.911 \pm .013$	$.706 \pm .021$	$.673 \pm .010$	$.558 \pm .067$
IDP-ESM2-150M, shallow (O)	$.946 \pm .003$	$.919 \pm .005$	$.692 \pm .018$	$.678 \pm .007$	$.558 \pm .055$
IDP-ESM2-150M, deep (N)	$.959 \pm .003$	$.940 \pm .007$	$.754 \pm .023$	$.697 \pm .008$	$.664 \pm .065$
ESM2-8M random, shallow (F)	$.912 \pm .016$	$.882 \pm .017$	$.453 \pm .045$	$.533 \pm .035$	$.287 \pm .050$
ESM2-8M random, deep (D)	$.897 \pm .011$	$.859 \pm .015$	$.433 \pm .022$	$.539 \pm .018$	$.268 \pm .041$
SeqDance-35M, shallow (H)	$.986 \pm .002$	$.973 \pm .001$	$.865 \pm .007$	$.787 \pm .011$	$.833 \pm .011$
SeqDance-35M, deep (G)	$.985 \pm .001$	$.972 \pm .001$	$.863 \pm .012$	$.796 \pm .003$	$.842 \pm .024$
ProtT5-XL, shallow (J)	$.969 \pm .001$	$.943 \pm .007$	$.659 \pm .048$	$.655 \pm .024$	$.514 \pm .083$
ProtT5-XL, deep (I)	$.958 \pm .004$	$.928 \pm .003$	$.654 \pm .017$	$.605 \pm .018$	$.541 \pm .055$
GeoGraph-2.2M, shallow (M)	$.988 \pm .001$	$.977 \pm .002$	$.877 \pm .010$	$.808 \pm .009$	$.855 \pm .026$
GeoGraph-2.2M, deep (L)	$.987 \pm .002$	$.976 \pm .004$	$.893 \pm .003$	$.828 \pm .002$	$.897 \pm .016$

J ALL PAIRED COMPARISONS

Paired differences in viral R^2 between configurations are shown in Table J.1.

K FORMAL STATEMENTS

We state the fitness-bottleneck hypothesis precisely and derive the testable predictions used in §5.2 and §5.3.

Setup. Let s be a random sequence and Y a real-valued target with $0 < \mathrm{Var}(Y) < \infty$. The frozen representation $T = T(s)$, the fitness $F = F(s)$ and the length $L = |s|$ are measurable functions of s. For any random variable X, define the optimal coefficient of determination

$$R^{2*}(Y \mid X) \;=\; 1 - \frac{\mathbb{E}\big[(Y - \mathbb{E}[Y \mid X])^2\big]}{\mathrm{Var}(Y)}, \tag{3}$$

the largest R^2 that any measurable predictor of Y from X can achieve. Two standard facts follow from the optimality of conditional expectation under squared loss, namely (i) $R^{2*}(Y \mid g(X)) \le R^{2*}(Y \mid X)$ for measurable g, with equality when g is a bijection with measurable inverse, and (ii) $R^{2*}(Y \mid X) \le R^{2*}(Y \mid X, Z)$.

Conditional mutual information vanishes exactly when the conditional distributions factorise, so the exact form of H_{FB}, $I(T; Y \mid F) = 0$, is equivalent to conditional independence $T \perp Y \mid F$.

Proposition 1 (fitness bottleneck). If $T \perp Y \mid F$, then $R^{2*}(Y \mid T) \le R^{2*}(Y \mid F)$.

Proof. Conditional independence gives $\mathbb{E}[Y \mid T, F] = \mathbb{E}[Y \mid F]$ almost surely, so that by (ii) $\mathbb{E}[(Y - \mathbb{E}[Y \mid T])^2] \ge \mathbb{E}[(Y - \mathbb{E}[Y \mid T, F])^2] = \mathbb{E}[(Y - \mathbb{E}[Y \mid F])^2]$, and dividing by $\mathrm{Var}(Y)$ gives the claim. □

Table J.1: Paired differences in viral R^2 between configurations (row $b - a$), with paired-bootstrap 95% intervals underneath (2,000 resamples of the 1,015 sequences, per seed differences averaged over three seeds). Bold differences have intervals that exclude zero.

	Comparison	R_g	R_{ee}	ν	Δ	A_0
Supervised ceiling						
A − K	ESM2-8M scratch vs. KESTREL	+.002 [−.011, .015]	+.012 [−.002, .028]	+.007 [−.003, .017]	−.003 [−.014, .007]	+.004 [−.018, .026]
Frozen evolutionary PLMs vs. the ceiling						
E − K	IDP-ESM2-8M shallow vs. KESTREL	**−.046** [−.059, −.032]	**−.047** [−.060, −.034]	**−.102** [−.131, −.074]	**−.078** [−.099, −.057]	**−.203** [−.251, −.171]
N − K	IDP-ESM2-150M deep vs. KESTREL	**−.021** [−.031, −.007]	−.012 [−.025, .006]	**−.027** [−.051, −.005]	**−.036** [−.055, −.017]	−.034 [−.093, .013]
C − A	IDP-ESM2-8M deep vs. scratch	**−.058** [−.070, −.045]	**−.055** [−.067, −.042]	**−.067** [−.091, −.044]	**−.064** [−.084, −.045]	**−.129** [−.157, −.100]
Pretraining gain, frozen (Δ_{pre})						
E − F	IDP-ESM2-8M vs. random, shallow	**+.062** [.045, .082]	**+.084** [.055, .118]	**+.233** [.180, .280]	**+.103** [.068, .135]	**+.385** [.318, .431]
C − D	IDP-ESM2-8M vs. random, deep	**+.075** [.053, .103]	**+.100** [.073, .136]	**+.249** [.202, .290]	**+.102** [.070, .135]	**+.436** [.395, .469]
Readout (deep vs. shallow probe)						
C − E	IDP-ESM2-8M	−.010 [−.022, .000]	+.004 [−.012, .021]	**+.042** [.015, .068]	+.011 [−.009, .031]	**+.079** [.038, .127]
N − O	IDP-ESM2-150M	+.006 [−.002, .015]	**+.013** [.003, .026]	**+.064** [.046, .082]	**+.033** [.015, .050]	**+.162** [.111, .209]
D − F	ESM2-8M, random weights	**−.023** [−.050, −.002]	−.012 [−.041, .012]	**+.026** [.008, .044]	+.012 [.000, .025]	**+.028** [.005, .050]
I − J	ProtT5-XL	**−.027** [−.042, −.012]	**−.028** [−.045, −.012]	+.020 [−.005, .046]	−.009 [−.035, .017]	**+.120** [.086, .154]
G − H	SeqDance-35M	+.001 [−.002, .003]	−.002 [−.006, .001]	+.004 [−.004, .013]	**+.014** [.005, .024]	+.006 [−.007, .021]
L − M	GeoGraph-2.2M	+.001 [−.002, .004]	−.002 [−.005, .000]	−.005 [−.012, .002]	+.002 [−.003, .008]	**−.016** [−.029, −.002]
Model scale and family						
O − E	IDP-ESM2 150M vs. 8M, shallow	**+.020** [.007, .031]	**+.022** [.003, .040]	+.010 [−.012, .032]	+.010 [−.007, .026]	+.007 [−.025, .044]
N − C	IDP-ESM2 150M vs. 8M, deep	**+.036** [.026, .046]	**+.031** [.020, .042]	**+.033** [.013, .051]	**+.031** [.013, .051]	**+.091** [.035, .131]
C − I	IDP-ESM2-8M vs. ProtT5-XL, deep	+.007 [−.006, .020]	**+.017** [.003, .032]	**+.049** [.021, .081]	**+.051** [.025, .078]	−.020 [−.054, .027]
N − I	IDP-ESM2-150M vs. ProtT5-XL, deep	**+.043** [.030, .056]	**+.048** [.033, .065]	**+.082** [.056, .111]	**+.082** [.057, .109]	**+.071** [.016, .113]
Simulator vs. evolutionary pretraining						
H − E	SeqDance vs. IDP-ESM2-8M, shallow	**+.056** [.046, .067]	**+.071** [.054, .087]	**+.123** [.099, .149]	**+.083** [.064, .103]	**+.225** [.183, .274]
G − C	SeqDance vs. IDP-ESM2-8M, deep	**+.067** [.058, .078]	**+.065** [.054, .077]	**+.085** [.065, .108]	**+.086** [.069, .106]	**+.151** [.127, .173]
L − C	GeoGraph vs. IDP-ESM2-8M, deep	**+.071** [.061, .081]	**+.067** [.055, .079]	**+.088** [.069, .110]	**+.103** [.084, .124]	**+.142** [.115, .172]
L − N	GeoGraph vs. IDP-ESM2-150M, deep	**+.035** [.029, .042]	**+.035** [.027, .045]	**+.055** [.038, .074]	**+.072** [.055, .090]	**+.051** [.006, .116]
G − I	SeqDance vs. ProtT5-XL, deep	**+.074** [.061, .088]	**+.082** [.068, .098]	**+.134** [.112, .161]	**+.137** [.114, .163]	**+.132** [.106, .168]
L − G	GeoGraph vs. SeqDance, deep	+.004 [−.001, .007]	+.002 [−.003, .007]	+.003 [−.005, .012]	**+.017** [.007, .028]	−.009 [−.025, .011]
Fine-tuning						
B − A	evolutionary vs. random init. (Δ_{ft})	−.002 [−.010, .008]	+.000 [−.007, .010]	**+.011** [.002, .020]	**+.017** [.008, .027]	−.010 [−.029, .014]
B − C	fine-tuned vs. frozen deep	**+.056** [.046, .067]	**+.055** [.044, .067]	**+.078** [.056, .102]	**+.081** [.062, .102]	**+.119** [.085, .157]

Corollary 1 (testable form). If, in addition, $\log p_{\mathrm{PLM}}(s) = \beta F(s) - \ln Z(\beta)$ with $\beta \neq 0$, then

$$R^{2*}(Y \mid T) \;\leq\; R^{2*}(Y \mid \log p_{\mathrm{PLM}}) \;\leq\; R^{2*}(Y \mid \log p_{\mathrm{PLM}}, L). \tag{4}$$

Proof. $\log p_{\mathrm{PLM}}$ is an invertible affine function of F, so $R^{2*}(Y \mid F) = R^{2*}(Y \mid \log p_{\mathrm{PLM}})$ by (i), and adding L cannot decrease R^{2*} by (ii), which together with Proposition 1 gives the claim. □

Corollary 1 holds for any Y, including sequence descriptors, and gives prediction P1 when Y is a physical target and prediction P2 when $Y = \mathrm{NCPR}$. When the IRL relation holds only approximately, F is no longer a function of $\log p_{\mathrm{PLM}}$ and the corollary does not follow, in which case, as noted in §3.2, the hypothesis no longer identifies in advance which properties a PLM will miss.

From population statements to data. A fitted predictor h evaluated on some distribution satisfies $R^2(h) \leq R^{2*}(Y \mid X)$ on that distribution, so the held-out R^2 of a regressor on T is a lower bound on $R^{2*}(Y \mid T)$, and Corollary 1 would require $R^{2*}(Y \mid \log p, L)$ to be at least as large. On the in-distribution test sets, where training and test distributions match, boosted trees on $(\log p, L)$ reach $R^2 = 0.08$ for ν while boosted trees on the mean-pooled embedding reach 0.57, and for NCPR, regressors on $(\log p, L)$ reach $R^2 \leq 0.00$ while a linear probe on the embedding reaches 0.98 (and 0.99 on viral sequences). Reconciling these results with Corollary 1 would require the best predictor from two scalar features to exceed flexible regressors trained on 7,441 sequences by about 0.5 R^2 for ν and by about 1.0 for NCPR, which we regard as implausible, and we therefore conclude that $T \perp Y \mid F$, with fitness measured by the PLM's likelihood, does not hold for these targets.

Proposition 2 (readout gap). Let $\mathcal{P}_{\mathrm{s}} \subseteq \overline{\mathcal{P}}_{\mathrm{d}}$ be probe classes acting on the per-residue embeddings, where the bar denotes closure under pointwise limits. For any target and any evaluation distribution, $\sup_{p \in \mathcal{P}_{\mathrm{s}}} R^2(p) \leq \sup_{p \in \mathcal{P}_{\mathrm{d}}} R^2(p)$.

Proof. The claim follows immediately from the inclusion, since R^2 is continuous under pointwise convergence of bounded predictors (dominated convergence). □

The deep probe contains the shallow one in this sense, because a zero attention query gives uniform weights, i.e. mean pooling, setting the residual blocks to zero makes them identities, and the final LayerNorm acts approximately linearly on inputs padded with large constant coordinates, so that the output head can represent the shallow two-layer MLP to arbitrary accuracy, and the population readout gap $\mathrm{ES}_{\mathrm{deep}} - \mathrm{ES}_{\mathrm{shallow}}$ is therefore non-negative. Negative estimates, such as -0.010 for R_g with IDP-ESM2-8M and about -0.03 for R_g and R_{ee} with ProtT5-XL (Table J.1), mean that the fitted deep probe fell short of the best predictor in its class, for example by overfitting high-dimensional features, rather than a loss of accessible information.

L WHY WE TEST THE FITNESS BOTTLENECK BY REGRESSION

A direct test would estimate $I(T; Y \mid \log p)$, and Table L.1 reports two such estimates for IDP-ESM2-8M on the 1,015 viral sequences, with T the mean-pooled embedding and $\log p$ the strided pseudo-log-likelihood of Appendix D. Both use the chain rule $I(T; Y \mid \log p) = I([T, \log p]; Y) - I(\log p; Y)$. KSG (Kraskov et al., 2004) is applied with $k = 5$ nearest neighbours to a 50-component PCA of the standardised embedding, whereas MINE (Belghazi et al., 2018) trains a statistics network with hidden layers of 256 and 128 units using Adam (learning rate 3×10^{-4}, batch 256, 2,000 steps) and averages the Donsker–Varadhan bound over the last 20% of steps, with five seeds. The two estimators agree within 0.14 nats on the one-dimensional term $I(\log p; Y)$ but differ by one to two nats on the high-dimensional term. KSG places $I([T, \log p]; Y)$ below $I(\log p; Y)$ for every target, which is impossible, since adding variables cannot reduce mutual information, so its conditional estimate of zero reflects the downward bias of nearest-neighbour estimates in 51 dimensions rather than an absence of information, and the same holds for $k = 3$ and 10. MINE instead estimates 0.8 to 1.3 nats, but a difference of two variational bounds carries no one-sided guarantee. Neither estimator can therefore decide the question, and we rely instead on the held-out prediction test of §5.2, which Corollary 1 justifies (Appendix K) and whose uncertainty is quantified across seeds. Its conclusion, that the embedding carries information about the targets beyond $\log p$, agrees in direction with MINE. Our preliminary analysis, which used the embedding of the original ESM2-8M and conditioned on a random-mask log-perplexity, showed the same pattern, with KSG at 0 to 0.03 nats and MINE at 0.93 to 1.70 nats.

Table L.1: Estimates of $I(T; Y \mid \log p)$ in nats for IDP-ESM2-8M on the viral set, from the chain rule. KSG ($k = 5$, 50-component PCA) varies by less than 10^{-5} between runs, and MINE values are means over five seeds, with the s.d. for the conditional term. The KSG conditional estimate is clipped at zero, and its unclipped value is given below it.

Term	Estimator	R_g	R_{ee}	ν	Δ	A_0
$I(\log p; Y)$	KSG	.893	.629	.239	.289	.250
	MINE	.754	.534	.178	.203	.247
$I([T, \log p]; Y)$	KSG	.053	.045	.093	.081	.091
	MINE	2.098	1.848	1.385	1.339	1.082
$I(T; Y \mid \log p)$	KSG	0	0	0	0	0
	KSG, unclipped	$-.840$	$-.584$	$-.146$	$-.208$	$-.160$
	MINE	$1.344 \pm .022$	$1.314 \pm .076$	$1.207 \pm .008$	$1.137 \pm .010$	$.835 \pm .041$

M SHALLOW AND DEEP PROBES

Both probes read the same frozen per-residue embeddings $\mathbf{e}_1, \ldots, \mathbf{e}_L \in \mathbb{R}^D$ from the final layer of the backbone, with special tokens excluded, cached once in 16-bit precision and never updated, and both predict the five targets jointly after standardising each on the training split. The width D ranges from 256 for GeoGraph to 1,024 for ProtT5-XL (Table M.1). The shallow probe is used in arms E, F, H, J, M and O and the deep probe in arms C, D, G, I, L and N, and the difference in ES between them is the readout gap of §3.

Shallow probe. The shallow probe averages the embeddings over residues and applies a two-layer MLP,

$$\bar{\mathbf{e}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{e}_i, \qquad \hat{\mathbf{y}} = W_2\,\mathrm{Dropout}\big(\mathrm{SiLU}(W_1\bar{\mathbf{e}} + \mathbf{b}_1)\big) + \mathbf{b}_2, \tag{5}$$

with $W_1 \in \mathbb{R}^{128\times D}$, $W_2 \in \mathbb{R}^{5\times 128}$ and dropout 0.1, which gives $128D + 773$ parameters. It has the form of the feature head of GeoGraph (Quinn et al., 2025) with a fixed hidden width of 128, and like the mean-pooled probes commonly used to evaluate PLM embeddings, it sees only the average of the contextual residue embeddings, so information about how residues are arranged along the chain reaches it only insofar as it survives averaging. The fine-tuned models (A and B) attach the same head to the mean-pooled output of the backbone.

Deep probe. The deep probe replaces the average with attention pooling by a learned query $\mathbf{w} \in \mathbb{R}^D$ and the two-layer MLP with a residual network of width 512,

$$\alpha_i = \frac{\exp(\mathbf{w}^\top \mathbf{e}_i)}{\sum_{j=1}^{L}\exp(\mathbf{w}^\top \mathbf{e}_j)}, \qquad \mathbf{h}_0 = W_{\mathrm{in}}\sum_{i=1}^{L}\alpha_i\,\mathbf{e}_i + \mathbf{b}_{\mathrm{in}}, \tag{6}$$

$$\mathbf{h}_{k+1} = \mathbf{h}_k + \mathrm{Linear}\Big(\mathrm{Dropout}\big(\mathrm{SiLU}(\mathrm{Linear}(\mathrm{LN}(\mathbf{h}_k)))\big)\Big), \qquad k = 0, \ldots, 3, \tag{7}$$

$$\hat{\mathbf{y}} = \mathrm{Linear}_{128\to 5}\Big(\mathrm{SiLU}\big(\mathrm{Linear}_{512\to 128}(\mathrm{LN}(\mathbf{h}_4))\big)\Big), \tag{8}$$

where $W_{\mathrm{in}} \in \mathbb{R}^{512\times D}$, each residual block has two 512×512 linear layers with dropout 0.1, LN denotes layer normalisation, and padding positions receive zero attention weight. It has $513D + 2{,}173{,}189$ parameters, about 56 times as many as the shallow probe on IDP-ESM2-8M. The attention weights let the probe emphasise particular residues and the residual layers let it combine features nonlinearly, but the pooled vector is still a weighted average of residue embeddings, so the deep probe tests a more flexible readout of the same representation rather than a different representation.

Training. Both probes are trained with AdamW (weight decay 10^{-4}), batch size 32, mean-squared error on the standardised targets, gradient clipping at norm 1 and a cosine learning-rate decay over the maximum number of epochs (Table M.1). Training stops once a three-epoch moving average of the validation loss has not improved for the patience window, the weights with the best moving average are restored, and predictions are mapped back to physical units with the training-split means and standard deviations. The probe settings are shared by every backbone and were not tuned on the matched protocol, and the probe initialisation, like the split, is set by the seed.

Table M.1: Probe settings and trainable parameters for each backbone width D.

	Shallow probe	Deep probe
Pooling	mean	learned attention
Head	MLP, width 128	residual MLP, width 512
Learning rate	3×10^{-3}	10^{-3}
Maximum epochs, patience	200, 20	300, 30
GeoGraph ($D = 256$)	33,541	2,304,517
IDP-ESM2-8M ($D = 320$)	41,733	2,337,349
ESM2-8M, random weights ($D = 320$)	41,733	2,337,349
SeqDance-35M ($D = 480$)	62,213	2,419,429
IDP-ESM2-150M ($D = 640$)	82,693	2,501,509
ProtT5-XL ($D = 1{,}024$)	131,845	2,698,501

N ProtT5, GeoGraph and IDP-ESM2-150M details

We use the encoder of `Rostlab/prot_t5_xl_uniref50` in 32-bit precision. Sequences are written as space-separated residues with U, Z, O and B mapped to X, which yields one token per residue and a final `</s>`, as we verified for every sequence, and we exclude `</s>` from pooling. Because the released checkpoint is stored in a format that recent library versions load only with newer PyTorch, we converted its encoder weights once to safetensors, with a conversion that fails if any encoder weight is missing, and reloaded and checked the converted model before use. Embeddings are computed in batches of eight, and with a 1,024-dimensional embedding the shallow probe has 131,845 parameters and the deep probe 2,698,501. The shallow arm (J) agrees with our earlier run of the same probe (viral R^2 for ν of 0.771 versus 0.774), and both arms are reported in Table 1.

GeoGraph. We use the published checkpoint `InstaDeepAI/GeoGraph` (2,185,894 parameters in four layers of width 256), loaded with a strict check that every weight is present. GeoGraph was trained on Human-IDRome to predict 14 sequence-level descriptors, including R_g, R_{ee}, Δ, ν and A_0, and 7 residue-level contact-graph descriptors. Its tokenizer adds no special tokens, so each token is one residue, as we verified for every sequence. We use its final-layer embeddings with the same two probes as for the other backbones (33,541 and 2,304,517 parameters for the 256-dimensional embedding), and do not use its own prediction heads.

IDP-ESM2-150M. We use `InstaDeepAI/IDP-ESM2-150M` (30 layers of width 640), which was further pretrained on IDP-Euka-90, the same eukaryotic IDR sequences as IDP-ESM2-8M, loaded with a strict check that all 485 learned tensors are present. Tokenisation and pooling follow IDP-ESM2-8M, and with the 640-dimensional embedding the shallow probe has 82,693 parameters and the deep probe 2,501,509. Both arms (O shallow, N deep) use the same splits and seeds as the other frozen arms and are reported in Table 1.

O Compute and software

Arms A, B, C, G, H, K and the ProtT5 arms, together with the contact-network runs of Appendix H, were run on a GPU cluster, each run on a single NVIDIA GeForce GTX 1080 Ti with 11 GB of memory, using PyTorch 2.5.1 (CUDA 12.1 build) and Hugging Face Transformers 4.57.6. Arms D, E, F, L, M, N, O and the analyses of Appendices D–E were run on Apple silicon (PyTorch 2.14, Transformers 5.17, MPS backend). Some comparisons therefore cross platforms (C − E, C − D), although the pretraining gains measured within one platform (E − F) and across platforms (C − D) agree closely ($+0.233$ and $+0.249$ on ν), and rerunning D, E and F on the cluster would remove this source of variation entirely.