Beyond Nativeness: Viral Proteins in Protein Language Models

Arthur Bigot [1 2] **Harmon Bhasin** [3] **Core Francisco Park** [4 5] **Eugene Shakhnovich** [2] **Dianzhuo Wang** [2 3]

Abstract

Protein language models are trained on highly imbalanced datasets, raising the question of how they represent underrepresented biological sequences. Using viral proteins as a case study across ESM model families, we identify a dominant *nativeness axis* in embedding space, aligned with masked-reconstruction perplexity, that orders sequences from well-modeled cellular proteins through viral proteins to shuffled and random sequences. Scaling contracts this axis unevenly across viral families. Despite this, protein language model embeddings retain viral-specific signal: viral proteins remain linearly separable beyond zero-shot perplexity and shallow sequence features. Together, these results suggest that pLM representations are structured by a general notion of nativeness while preserving information specific to distinct biological groups.

1. Introduction

Protein language models (pLMs) are powerful tools for learning general-purpose protein representations. Trained on large sequence databases, they are used for structure prediction (Lin et al., 2023a), inverse folding (Hsu et al., 2022), and function prediction (Dallago et al., 2021; Feldman et al., 2025; Meier et al., 2021a; Notin et al., 2023). Their strong performance has motivated a growing interest in understanding what these models learn, including recent work using mechanistic interpretability (Adams et al., 2025; Simon & Zou, 2025; Silberg et al., 2025). However, much less is known about how pLMs represent biological groups that are functional but underrepresented and evolutionarily distinct from cellular proteins.

Viral proteins provide a useful case study for how pLMs represent underrepresented biological groups. Compared with proteins from cellular organisms, they are less abundant in standard pretraining data and are shaped by different evolutionary constraints, including host dependence, high mutation rates, compact genomes, and multifunctionality (Simmonds et al., 2019; Gurev et al., 2026; Tokuriki et al., 2009). Benchmark results reflect this gap: pLMs struggle on viral mutation-effect prediction, despite strong performance on standard, mostly non-viral benchmarks (Gurev et al., 2026).

In this work, we analyze viral protein representations across model scales to answer two questions: *First, is viral separation mainly explained by lower nativeness, meaning that viral proteins are less well modeled by the distribution learned during pretraining? Second, beyond nativeness, do embeddings retain viral-specific information?* We answer the first question by analyzing the embedding geometry across model families through principal component analysis (PCA) and masked-reconstruction perplexity, which we use as a model-relative measure of nativeness. We answer the second question by comparing linear probes trained on mean-pooled embeddings with shallow sequence-based classification baselines and zero-shot classifier based on masked-reconstruction perplexity. Our main contributions are summarized as follows:

- We identify a dominant nativeness axis in pLM representation space that closely tracks reconstruction difficulty and explains much of the viral to cellular protein shift.
- We show that scaling reduces viral displacement along this axis unevenly across human viral families, with some families moving toward the native region and others remaining persistently displaced.
- We show that embeddings retain viral signal beyond nativeness: linear probes outperform perplexity-only classifiers and shallow baselines across model families and scales.

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland [2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA [3]TwentyTwo, USA [4]Harvard University, Cambridge, MA, USA [5]Prior Computers, USA. Correspondence to: Arthur Bigot <bigot.arthur@gmail.com>, Eugene Shakhnovich <shakhnovich@chemistry.harvard.edu>, Dianzhuo Wang <john@twentytwo.bio>.

Proceedings of the ICML 2026 Workshop on Generative and Agentic AI for Biology, Seoul, Korea. 2026. Copyright 2026 by the author(s).

2. Related Work

pLM scaling laws Language model loss and downstream performance improve predictably as model size, data, and compute are scaled (Kaplan et al., 2020; Hoffmann et al., 2022). In protein language models, scaling has similarly been associated with stronger biological representations and the emergence of capabilities useful for structure prediction, fitness prediction, function prediction, and protein generation (Rives et al., 2021a; Lin et al., 2023b; Nijkamp et al., 2023; Hayes et al., 2025; Wang et al., 2024; Huot et al., 2025). However, it remains unclear how scaling changes the representation of underrepresented and evolutionarily distinct groups such as viral proteins.

Perplexity as a proxy for protein nativeness pLMs have been shown to encode biological constraints (Simmonds et al., 2019; Gurev et al., 2026; Tokuriki et al., 2009) beyond simple amino acid statistics, including structural contacts, family structure, and useful residue-level representations learned from sequence-only objectives (Rao et al., 2021; Rives et al., 2021a). As a result, pseudo-log-likelihood and pseudo-perplexity scores from masked pLMs have been widely used as zero-shot proxies for mutational effects, functional tolerance, viral escape, and generated-protein filtering (Salazar et al., 2020; Meier et al., 2021b; Hie et al., 2021; Verkuil et al., 2022; Madani et al., 2023; Notin et al., 2023). These works motivate our use of masked-reconstruction perplexity as a model-relative notion of nativeness. We show that perplexity corresponds to a dominant geometric axis in embedding space that organizes entire biological groups.

Biological signal in pLM representations Recent work using tools from mechanistic interpretability, such as sparse autoencoders (SAEs), have shown that pLM representations contain latent features associated with binding sites, structural motifs, functional domains, thermostability, localization, and family-specific constraints (Adams et al., 2025; Simon & Zou, 2025; Silberg et al., 2025). However, these studies do not examine viral proteins, leaving open whether pLM representations encode viral-specific signal.

Viral protein representations Viral proteins are among the most difficult sequences for pLMs to model (Vieira et al., 2026), yet viral pLM embeddings are still widely used. pLM embeddings have been used to annotate viral proteins beyond remote homology (Flamholz et al., 2024) and to improve virus-host association and host-range prediction, often outperforming sequence-based baselines (Liu et al., 2024; Panigrahi et al., 2025; Wang et al., 2024; 2026). Ofer & Linial (2025) even showed that viral and cellular proteins are linearly separable from mean-pooled ESM2 embeddings alone. Although these works establish that viral proteins are both atypical and informative in pLM representation space, they do not explain what drives their separation from cellular proteins.

3. Data and Methods

3.1. Experimental setting

Protein language models We evaluate three ESM-family pLMs spanning over three orders of magnitude in parameter count (Rives et al., 2021b). **ESM2** (Lin et al., 2023a) is a masked language model trained on UniRef_{50} (Suzek et al., 2015) (8M, 35M, 150M, 650M, 3B, and 15B). **ESMC** (EvolutionaryScale Team, 2024) is the current state-of-the-art representation model in the ESM family, also trained with masked language modeling (300M, 600M, and 6B). **ESM3** (Hayes et al., 2025) is a multimodal generative model jointly trained over sequence, structure, and function tracks (1.4B OPEN, 1.4B SMALL, 7B MEDIUM, and 98B LARGE). ESM3-OPEN is the only model trained without viral sequences; all others use virus-inclusive training data.

For each model and each input sequence, we take the final-layer embedding of every amino acid residue and mean-pool these residue embeddings, excluding BOS and EOS tokens. This gives one sequence embedding per protein.

Masked-reconstruction perplexity We use masked-reconstruction perplexity (PPL) as a sequence-level score. For each sequence $\mathbf{x}$, we mask a fraction $p{=}0.15$ of residue positions (excluding BOS/EOS), then compute the log-likelihood of masked tokens. The per-sequence perplexity is

$$\mathrm{PPL}(\mathbf{x}) \;=\; \exp\left(\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} -\log p_\theta\big(x_i \mid \mathbf{x}_{\setminus\mathcal{M}}\big)\right), \quad (1)$$

where $\mathcal{M}$ is the set of masked positions and p_θ denotes the conditional probability distribution over amino acids at each position. Results are averaged over three independent mask samples per sequence. For ESM3, PPL is computed with the model provided sequence input only. This directly matches the masked-reconstruction objective used to train ESM2 and ESMC, and is aligned with the sequence-track component of ESM3's masked denoising objective.

3.2. Sequence datasets

Pretraining dataset coverage (Table 1) To measure the data imbalance between cellular and viral proteins in ESM pretraining data, we count UniRef_{50} cluster representatives by biological group, the same clustering level used to pretrain ESM2, ESMC, and ESM3 (Suzek et al., 2015). Query details are provided in Appendix Section A.

Multi-group biological dataset (Figures 1 and 2). To evaluate these models, we gather ten biological groups span-

ning the tree of life: six cellular groups (bacteria, archaea, plants, fungi, insects, and human non-viral proteins) and four viral groups (bacterial (bacteriophage), human, plant, and invertebrate viruses). Cellular groups are drawn manually from Swiss-Prot (UniProt Consortium, 2025). Non-human viral groups are assembled using host and lineage annotations: plant and invertebrate viruses are selected with UniProt virus-host taxonomy filters, while bacteriophages are identified using phage-relevant viral lineages and organism-name matching. The human-virus group (Section 3.2) is a curated reference set defined for the human classification dataset. Full query strings, shared length/composition/deduplication filters, and per-group post-filter sizes are in Appendix Section B.

Biologically meaningless controls (Figure 1) We generate three biologically meaningless sequence controls: *position-shuffled cellular sequences*, *position-shuffled viral sequences*, and *i.i.d. uniform random sequences*. *Position-shuffled cellular sequences* and *position-shuffled viral sequences* are random permutations of the residue positions of the corresponding biological pool, preserving length and per-sequence amino acid composition while destroying positional structure. *i.i.d. uniform random sequences* are drawn position-wise from the uniform distribution over the 20 standard amino acids and length-matched to the viral pool.

Human viral/cellular classification dataset (Figure 3) The human viral group is a curated set of 5,203 human-infecting viral proteins spanning 32 annotated viral families, assembled from Swiss-Prot (UniProt Consortium, 2025) and `NP_` entries from NCBI Virus (Brister et al., 2015; Goldfarb et al., 2025). The cellular group is a length-decile-stratified random multi-kingdom sample of 5,197 sequences drawn from the cellular fraction of Swiss-Prot, matched in size to the human viral group. To prevent evaluation leakage from sequence homology, we pool all 10,400 sequences and cluster them with MMseqs2 (Steinegger & Söding, 2017) at the standard 30% identity and 80% bidirectional coverage. Whole clusters are assigned to train, validation, and test at a 60/20/20 split. All linear probe and zero-shot scores (defined in Section 3.3) reported in Figure 3 are computed on the held-out test split. Full query strings, shared length/composition filters, and family composition are in Appendix Section C.

3.3. Analysis methods

Embedding geometry For each model, we pool the sequence embeddings of all biological groups (cellular and viral) and biologically meaningless controls into a single matrix and apply principal component analysis (PCA) jointly. We examine PC_1, the direction that explains the largest variance in the embedding space, to assess whether viral separation concentrates along a single dominant axis.

Per-family scaling analysis (Figure 2) To measure how scale changes viral nativeness, we compute, for each human viral family, the fraction of sequences with $\mathrm{PPL}<5$. We use this fixed threshold across all ESMC model sizes because cellular groups already fall below it at smaller model scales. We restrict the analysis to families with at least 50 sequences and plot the eight families with the highest native-like fraction at ESMC-6B.

Embedding linear probe For each pLM, we standardize the train-split mean-pooled embeddings and fit an ℓ_2-regularized logistic regression head to predict the viral/cellular label. Linear probes are trained on the train split and AUC-ROC is reported on the held-out test split.

Zero-shot perplexity classifier We use the negated per-sequence perplexity, $s(\mathbf{x})=-\mathrm{PPL}(\mathbf{x})$, as a viral-group score and report its AUC-ROC on the same test split. To make AUC values directly comparable to the embedding linear probe regardless of which group has higher mean perplexity, we report $\max(\mathrm{AUC},\ 1-\mathrm{AUC})$.

Baselines We use three logistic-regression baselines on the same split: (i) sequence length (1 feature), (ii) amino acid composition (20 features), and (iii) dipeptide composition over adjacent residue pairs (400 features). The maximum AUC across these defines the *baseline ceiling* shown in Figure 3.

Table 1. **Pretraining data is strongly dominated by cellular proteins.** UniRef_{50} cluster counts per biological group. ESM2, ESMC, and ESM3 are all pretrained at this clustering level.

Biological group	UniRef_{50} clusters
Cellular groups	
Bacteria	30.7 M
Archaea	864.9 k
Plants	6.0 M
Fungi	6.3 M
Insects	2.4 M
Viral groups	
Bacteriophage	266.1 k
Plant viruses	3.4 k
Invertebrate viruses	6.6 k
Human viruses (18 fam.)	114.3 k
All cellular	46.3 M
All viral	390.3 k
Ratio cellular / viral	**119×**

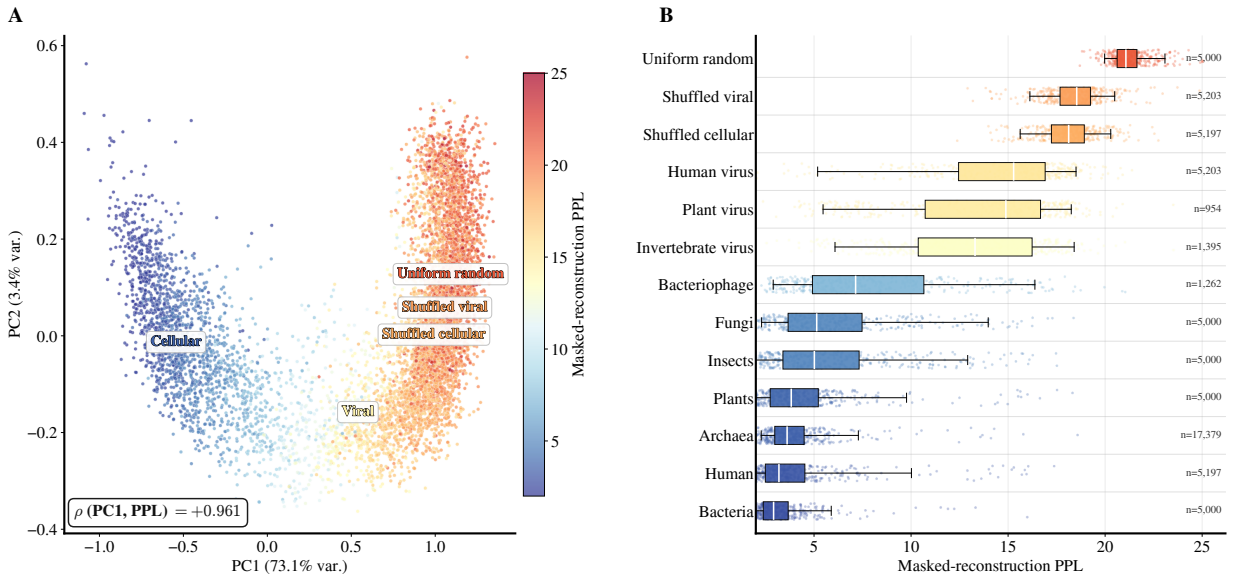

Figure 1. **A dominant nativeness axis organizes pLM representation space across the tree of life. (A)** PCA of ESMC-600M mean-pooled embeddings over ten biological groups and three biologically meaningless controls (Section 3.2). Points correspond to individual sequences from the ten biological groups, while the overlaid labels show higher-level category centroids: cellular, viral, shuffled, and random. Each label is placed at the median coordinate of its category in (PC_1, PC_2) space and colored by that category's mean PPL. Individual points are colored by masked-reconstruction perplexity. Sequences order from well-modeled cellular proteins (low PPL, left), through the viral region, to out-of-distribution shuffled and random sequences (high PPL, right). PC_1 explains 73.1% of variance and is strongly aligned with PPL (ρ=+0.961). **(B)** The same ordering is visible in the group-wise perplexity distributions: cellular groups have the lowest PPL, viral groups are shifted upward, and shuffled/random controls occupy the highest-PPL range. Per-group ESMC-600M masked-reconstruction perplexity (15% mask) is shown as overlaid box and strip plots, ordered by median, with n reported for each group.

4. Results

4.1. A dominant nativeness axis organizes pLM representation space

Viral proteins occupy a displaced yet structured region of pLM representation space. In ESMC-600M, PC_1 explains 73.1% of the variance of the pooled embedding distribution (Figure 1A) and closely tracks masked-reconstruction perplexity (Spearman ρ=+0.961). This axis orders sequences from well-reconstructed cellular proteins, through viral proteins, to poorly reconstructed biologically meaningless controls, including shuffled and random sequences. The alignment also holds within individual groups: PCA refits on each group alone recover the same PPL-aligned PC_1 (Appendix Table 3). PC_1 thus captures a continuum from in-distribution to out-of-distribution sequences, with viral proteins occupying an intermediate region.

We refer to PC_1 as the *nativeness axis*. A sequence is native to a pLM to the extent that it matches the statistical patterns learned during pretraining.

The same sequence ordering is visible when sequences are grouped by biological group (Figure 1B). Cellular groups cluster at low masked perplexity, whereas viral groups are systematically shifted upward. Bacteriophage proteins are closest to cellular proteins, whereas invertebrate, plant, and human viruses are progressively more displaced. Biologically meaningless controls occupy the high-perplexity end of the same ordering. Viral proteins are therefore not extreme outliers in pLM space; they are less native than cellular proteins under the learned prior, yet more structured than biologically meaningless sequences.

The nativeness axis generalizes across ESM families (Appendix Figure 6): PC_1 explains 54.3% (ESM2-650M) and 67.3% (ESM3-OPEN) of the variance and remains strongly correlated with PPL (ρ=+0.926, +0.935). This pattern is therefore not specific to one model architecture or training setup. It also extends beyond the masked-LM objective: an autoregressive (ProGen2) and a discrete-diffusion (EvoDiff) pLM show the same $PC_1 \approx$PPL alignment ($\rho = +0.90$ and +0.95) and five-tier ordering (Appendix L).

This axis likely reflects both data imbalance and evolutionary constraints. ESM pretraining is dominated by cellular proteins ($\sim$ 119:1; Table 1), biasing the model toward them, while viral proteins are shaped by distinct evolutionary pressures. Viral sequences are thus less native both because they are underrepresented in training and because they follow different constraints.

A sequence-novelty control shows that the cellular-viral gap

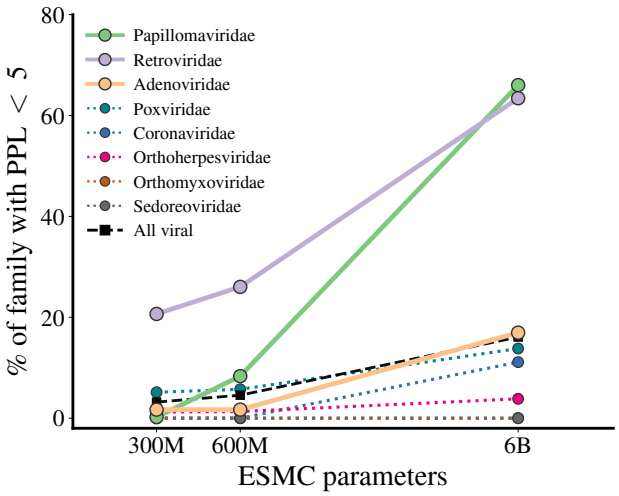

Figure 2. **Scaling contracts the nativeness axis heterogeneously across human viral families.** The fraction of sequences within each viral family whose masked-reconstruction perplexity falls below the native-like threshold PPL<5, as a function of ESMC parameter count. The top three families by ESMC-6B nativization rate are drawn with a solid line whereas the rest are represented by dotted lines. The mean across all human viral sequences is represented by the dashed grey line.

is not explained by pretraining data exposure alone. We evaluate n=1,723 cellular Swiss-Prot proteins released after the ESMC-600M checkpoint (Appendix Section G), and therefore absent from pretraining. Their median PPL is 5.3, which is much closer to the pre-release cellular reference (3.2) than to the viral reference (15.3). Thus, nativeness likely reflects compatibility with the cellular-dominated protein prior: newly released cellular proteins remain native-like, while viral proteins remain displaced.

Together, these results show that the pLM representation space contains a dominant reconstruction-aligned nativeness axis, PC_1, that accounts for much of the cellular-to-viral shift, while placing viral proteins between well-modeled cellular proteins and biologically meaningless controls.

4.2. Model scale contracts nativeness heterogeneously across human viral families

Model scale improves viral nativeness, but not uniformly across viral biological groups. In Section 4.1, we showed that viral proteins as a whole occupy a shifted region along the nativeness axis. We now refine this analysis within a single biological group, focusing on human viruses, to examine how scaling acts across finer-grained viral families.

On average, scaling makes human viral proteins only slightly more native-like, but the effect differs strongly across families (Figure 2). Aggregating across all human viral sequences, the fraction of native-like viral proteins increases from $\sim 5\%$ at 300M to $\sim 17\%$ at 6B. However, Papillomaviridae and Retroviridae gain roughly 60% with scale, whereas Orthomyxoviridae, Orthoherpesviridae, and Sedoreoviridae remain mostly outside the native-like region even at 6B. This pattern suggests that scaling reduces reconstruction difficulty primarily for families already closer to the learned protein prior, while leaving more displaced families poorly reconstructed. This is consistent with the broader scaling picture, where larger models achieve lower reconstruction loss and learn richer representations (Kaplan et al., 2020; Hoffmann et al., 2022; Rives et al., 2021a; Lin et al., 2023b), but here the contraction of the viral shift is selective rather than uniform.

What may distinguish the families that become native-like is whether their proteins have cellular homologs, and Retroviridae are a clear example. Many retroviral proteins have cellular counterparts: reverse transcriptases appear in both retroviruses and eukaryotic retrotransposons (Menéndez-Arias et al., 2017; Lescot et al., 2016), and LTR retrotransposons share key structural features with retroviruses (Eickbush & Jamburuthugoda, 2008). This overlap means retroviral protein domains are already present in the cellular training distribution, making Retroviridae more compatible with a cellular-trained pLM.

4.3. Viral identity is linearly encoded beyond reconstruction difficulty

Next we ask whether pLM embeddings capture biologically meaningful viral signal, or whether viral separability is simply driven by reconstruction difficulty. For each model, we evaluate three types of classifiers on the same held-out human viral/cellular test split (Section 3.2): a linear probe trained on mean-pooled embeddings, a zero-shot classifier using masked-reconstruction perplexity (PPL) alone, and the best of three baselines based on length, amino acid composition, or dipeptide composition (Section 3.3).

Figure 3 shows that viral identity is linearly accessible from pLM embeddings as linear probes can recover viral identity with near-ceiling accuracy across the ESM family. The linear probe AUC remains above the shallow-feature baselines at every scale and reaches the $\mathrm{AUC} \in [0.97, 1.00]$ ceiling band for larger models, even under a homology-controlled train/test split. Two additional robustness checks support that this separability reflects viral-specific signal rather than a dataset artifact. First, when the negative class is restricted to human host proteins, AUC remains at least 0.95 for all models with more than 35M parameters. Second, leave-one-family-out cross-validation over the 13 viral families with at least 50 sequences yields a mean held-out AUC above 0.96 (Appendix K). This conclusion is not restricted to the ESM family: applying the same probe analysis to ProGen2 and EvoDiff yields AUCs of 0.984 and 0.986, respectively (Appendix L).

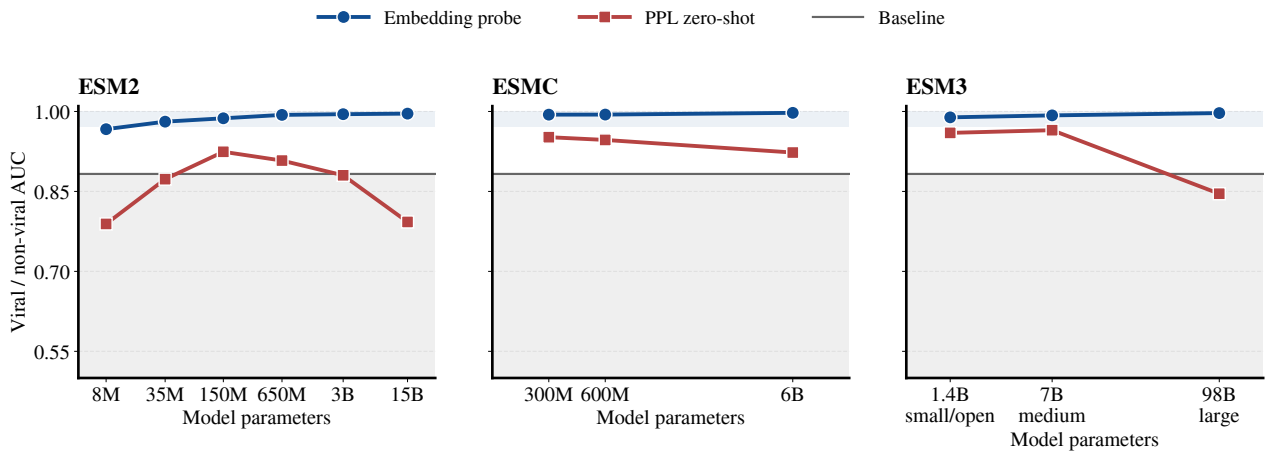

Figure 3. **Embedding linear probes capture more than perplexity: probe AUC scales to ceiling while PPL-based zero-shot classification does not.** For each ESM pLM family, we report human viral vs. cellular AUC-ROC on the human viral/cellular classification dataset (Section 3.2) using two readouts of the same model: a linear probe (logistic regression) on mean-pooled embeddings (blue circles) and a PPL-based zero-shot classifier (red squares, score$=-$PPL). The blue band at the top marks the near-ceiling region AUC$\in[0.97, 1.00]$. The grey *shallow-feature region* at the bottom spans AUC values attainable from raw sequence statistics alone. Its upper edge is the *best baseline* (Section 3.3).

This separability is not explained by reconstruction difficulty alone: the linear embedding probe and PPL-only classifier diverge with scale. PPL-based discrimination is weaker and non-monotonic: it improves at intermediate scales, but drops again for ESM2-15B and ESM3-LARGE. This drop is consistent with scale making some viral proteins easier to reconstruct, moving them into the low-PPL, more native-like region and reducing the viral/cellular separation available from PPL alone. In contrast, the embedding linear probe remains near ceiling at these same scales. Thus, larger models can make some viral proteins look more native under the reconstruction objective while still preserving a linearly accessible viral signal in the embeddings.

The gap is sharpest in the low-false-positive setting relevant for sequence screening, where the embedding probe far exceeds the PPL-only classifiers across all model families and scales (Appendix Figure 7). At 1% false positive rate (FPR), the embedding linear probe reaches 88.3% true positive rate (TPR) for ESM2-15B, 96.7% for ESMC-6B, and 90.6% for ESM3-large, compared with only 26.9%, 39.2%, and 36.1% for the PPL-only classifier. At 0.1% FPR, the linear probe improves from 6.2% TPR for ESM2-8M to 55.4% for ESM2-15B, and from 47.9% for ESMC-300M to 83.4% for ESMC-6B. Therefore, scaling improves the practical sequence screening value of the embedding representation, even when PPL alone becomes less reliable as a separator.

Finally, the linear probe also exceeds the best baseline (dipeptide composition over adjacent residue pairs), indicating that it is not merely exploiting simple sequence-level statistics. Together, these results support our third contribution: pLM embeddings carry a residual viral signal beyond masked-reconstruction perplexity and shallow sequence features.

5. Discussion

Viral proteins remain displaced relative to a cellular-dominated prior ESM3-OPEN was trained without viral sequences, yet it still exhibits a dominant nativeness axis aligned with masked-reconstruction perplexity, similar to the other ESM models, and viral proteins remain linearly separable from cellular proteins in embedding space. This suggests that the displaced position of viral proteins in representation space does not require direct viral exposure during pretraining. More broadly, it supports that viral proteins are positioned relative to a cellular-dominated protein prior: whether viral data are absent in pretraining, as in ESM3-OPEN, or present but strongly underrepresented, as in other ESM models, viral proteins remain in a shifted region rather than being fully absorbed into the native protein manifold.

This framing gives nativeness a practical interpretation. Low nativeness may indicate that pLMs capture the constraints governing a protein family less well. Consistent with this, recent viral mutation-effect benchmarks (Gurev et al., 2026) suggest that sequence-only pLMs are less reliable on viral proteins than on standard, mostly non-viral benchmarks. Nativeness may therefore be useful as a *diagnostic tool*: low-nativeness families may require more cautious zero-shot predictions.

Implications for biosecurity Our results suggest that pLM embeddings may carry information that is relevant for viral screening and could complement traditional homology-

based methods. Homology screening primarily detects sequence similarity to known proteins, whereas embedding-based approaches may capture broader functional and biophysical signals beyond direct similarity (Abel et al., 2026; Wittmann et al., 2025). While we do not evaluate a deployed screening system here, these findings raise the possibility that pLM-based methods may eventually help detect engineered or highly diverged variants that are less easily captured by standard homology-based filters (Wittmann et al., 2025; Wang et al., 2026).

Limitations Our main analysis focuses on the ESM family of protein language models. In Appendix L we provide preliminary evidence that the nativeness axis ($\rho(\mathrm{PC}_1, \mathrm{PPL}) = +0.90$ and $+0.95$) and the residual linear viral signal (probe AUC 0.984 and 0.986) also emerge in a decoder-only autoregressive model (ProGen2) and a discrete-diffusion model (EvoDiff); the heterogeneous per-family scaling reproduces as well, though with an objective-dependent family ranking. A systematic survey across architectures and scales remains future work.

Future work A natural next step is to fine-tune these models on viral sequences and study the consequences for the nativeness axis. Early results show that fine-tuning lowers the perplexity on viral sequences and moves them towards the cellular/native region without degrading linear probe classification performance.

A second direction is to understand the mathematical and statistical origin of the nativeness axis. Our results suggest that masked pLMs may develop a dominant direction that tracks model fit, but the mechanism that produces this geometry is unclear. Formalizing it would help clarify whether the PC_1–PPL alignment is an incidental empirical feature of ESM models, a consequence of high-dimensional mixture geometry, or a more general property of masked sequence models trained on imbalanced biological data.

A third direction is cross-domain comparison, such as multilingual language models. Underrepresented languages and dialects may provide an analog of viral proteins in pLMs: both are valid structured sequences, but they are under-sampled and may be governed by partly different statistical regularities from the dominant training distribution. Studying whether multilingual LMs exhibit a comparable low-density or high-loss axis could test whether the nativeness-axis phenomenon is specific to proteins or a broader property of large masked sequence models.

Impact Statement

This work has potential positive and negative societal implications. On the positive side, understanding how pLMs represent viral proteins may improve evaluation and screening methods for biological foundation models. On the negative side, methods that better distinguish viral-like from cellular-like sequences could be misused in dual-use settings. We therefore view this work as contributing to safer biological model evaluation rather than capability deployment.

References

Abel, Gary R., J., Alexanian, T., Bartling, C., Beal, J., Curtis, S., Flyangolts, K., Foner, L., Forry, S. P., Godbold, G. D., Horvitz, E., Hu, B., Hudson, C. M., Jagla, C., Lababidi, R., Lin-Gibson, S., Magalis, B., Pannu, J., Rivera, S., Ross, D. J., Wittmann, B. J., and Diggans, J. Beyond sequence similarity: Toward function-based screening of nucleic acid synthesis. *Frontiers in Bioengineering and Biotechnology*, 14:1832724, 2026. doi: 10.3389/fbioe.2026.1832724. URL https://doi.org/10.3389/fbioe.2026.1832724.

Adams, E., Bai, L., Lee, M., Yu, Y., and Alquraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 460–476. PMLR, 2025. URL https://proceedings.mlr.press/v267/adams25a.html.

Alamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., and Yang, K. K. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023. doi: 10.1101/2023.09.11.556673.

Baltimore, D. Expression of animal virus genomes. *Bacteriological Reviews*, 35(3):235–241, 1971.

Brister, J. R., Ako-adjei, D., Bao, Y., and Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Research*, 43(D1): D571–D577, 2015. doi: 10.1093/nar/gku1207.

Dallago, C., Mou, J., Johnston, K. E., Wittmann, B. J., Bhattacharya, N., Goldman, S., Madani, A., and Yang, K. K. FLIP: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, 2021. 2021.11.09.467890.

Eickbush, T. H. and Jamburuthugoda, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Research*, 134(1–2):221–234, 2008. doi: 10.1016/j.virusres.2007.12.010.

Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 55–65, 2019.

EvolutionaryScale Team. ESM cambrian: Revealing the mysteries of proteins with unsupervised learning. https://www.evolutionaryscale.ai/blog/esm-cambrian, 2024.

Feldman, J., Maechler, A., Wang, D., and Shakhnovich, E. Biophysically grounded deep learning improves protein–protein $\delta\delta$ g prediction. *bioRxiv*, pp. 2025–12, 2025.

Flamholz, Z. N., Biller, S. J., and Kelly, L. Large language models improve annotation of prokaryotic viral proteins. *Nature Microbiology*, 9(2):537–549, 2024. doi: 10.1038/s41564-023-01584-8.

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D. H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., Murphy, M. R., Schafer, S. S., Souvorov, A., Spurrier, B., Strope, P. K., Sun, H., Vatsan, A. R., Wallin, C., Webb, D., Brister, J. R., Hatcher, E., Kimchi, A., Klimke, W., Marchler-Bauer, A., Pruitt, K. D., Thibaud-Nissen, F., and Murphy, T. D. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, 53 (D1):D243–D257, 2025. doi: 10.1093/nar/gkae1038.

Gurev, S., Youssef, N., Jain, N., Mehrotra, A., Leung, S. R. M., Jackson, A., and Marks, D. Evaluating variant effect prediction across viruses. *bioRxiv*, 2026. doi: 10.1101/2025.08.04.668549. URL https://www.biorxiv.org/content/early/2026/01/12/2025.08.04.668549.1.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Hie, B., Zhong, E. D., Berger, B., and Bryson, B. Learning the language of viral evolution and escape. *Science*, 371 (6526):284–288, 2021. doi: 10.1126/science.abd7331.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hoogeboom, E., Gritsenko, A. A., Bastings, J., Poole, B., van den Berg, R., and Salimans, T. Autoregressive diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pp. 8946–8970. PMLR, 2022.

Huot, M., Wang, D., Liu, J., and Shakhnovich, E. I. Predicting high-fitness viral protein variants with bayesian active learning and biophysics. *Proceedings of the National Academy of Sciences*, 122(24):e2503742122, 2025.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lescot, M., Hingamp, P., Kojima, K. K., Villar, E., Romac, S., Veluchamy, A., Boccara, M., Jaillon, O., Iudicone, D., Bowler, C., Wincker, P., Claverie, J.-M., and Ogata, H. Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *The ISME Journal*, 10:1134–1146, 2016. doi: 10.1038/ismej.2015.192.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure. *Science*, 379(6637):1123–1130, 2023a.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023b. doi: 10.1126/science.ade2574.

Liu, D., Young, F., Lamb, K. D., Robertson, D. L., and Yuan, K. Prediction of virus-host associations using protein language models and multiple instance learning. *PLOS Computational Biology*, 20(11):e1012597, 2024. doi: 10.1371/journal.pcbi.1012597.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L. J., Xiong, C.,

Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29287–29303, 2021a.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303, 2021b.

Menéndez-Arias, L., Sebastián-Martín, A., and Álvarez, M. Viral reverse transcriptases. *Virus Research*, 234: 153–176, 2017. doi: 10.1016/j.virusres.2016.12.019.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

Notin, P., Kollasch, A. W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, R., and Marks, D. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ofer, D. and Linial, M. Protein language models expose viral immune mimicry. *Viruses*, 17(9):1199, 2025. doi: 10.3390/v17091199.

Panigrahi, S., Ansaldi, M., and Ginet, N. Phage evolutionary relationships emerge from protein language model-based proteome representation. *NAR Genomics and Bioinformatics*, 7(4):lqaf134, 2025. doi: 10.1093/nargab/lqaf134.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fylclEqgvgd.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118, 2021a. doi: 10.1073/pnas.2016239118.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118 (15):e2016239118, 2021b.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, 2020. doi: 10.18653/v1/2020.acl-main.240.

Silberg, J., Simon, E., and Zou, J. Towards functional annotation with latent protein language model features. *bioRxiv*, 2025. doi: 10.1101/2025.10.02.680154. 2025.10.02.680154.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. Prisoners of war—host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, 17(5):321–328, 2019. doi: 10.1038/s41579-018-0120-2.

Simon, E. and Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Steinegger, M. and Söding, J. 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.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., and Tawfik, D. S. Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences*, 34 (1):53–59, 2009. doi: 10.1016/j.tibs.2008.10.009.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

Verkuil, R., Kabeli, O., Du, Y., Wicky, B. I. M., Milles, L. F., Dauparas, J., Baker, D., Ovchinnikov, S., Sercu, T., and Rives, A. Language models generalize beyond natural proteins. *bioRxiv*, 2022. 2022.12.21.521521.

Vieira, L. C., Lin, S., and Wilke, C. O. Intrinsic dataset features drive mutational effect prediction by protein language models. *bioRxiv*, 2026. doi: 10.64898/2026.03.08.710389. URL https://www.biorxiv.org/content/early/2026/03/10/2026.03.08.710389.

Wang, D., Huot, M., Mohanty, V., and Shakhnovich, E. I. Biophysical principles predict fitness of sars-cov-2 variants. *Proceedings of the National Academy of Sciences*, 121(23):e2314518121, 2024. doi: 10.1073/pnas.2314518121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2314518121.

Wang, D., Huot, M., Zhang, Z., Jiang, K., Shakhnovich, E. I., and Esvelt, K. M. Without safeguards, ai-biology integration risks accelerating future pandemics. *Frontiers in Microbiology*, 16:1734561, 2026.

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768):82–87, 2025.

A. UniProt queries for pretraining coverage

Table 1 reports UniRef_{50} (Suzek et al., 2015) cluster counts at 50% sequence identity, the granularity at which ESM2, ESMC, and ESM3 are pretrained. All queries were issued on 2026-04-22 against the UniProt REST endpoint and are logged for reproducibility.

Cellular per-taxon counts. Each cellular group is queried by NCBI taxonomy ID (e.g. `taxonomy_id:2` for Bacteria, `taxonomy_id:2157` for Archaea). Per-taxon rows are *not additive*: a single UniRef_{50} cluster can contain sequences from multiple taxa and hence be counted in more than one row.

Viral per-taxon counts. The UniRef REST endpoint does not accept host filters (`virus_host_id`), so host-scoped viral groups such as "plant-infecting viruses" cannot be counted directly at UniRef_{50} resolution. We therefore report per-taxon viral entries only for groups queryable by lineage: Caudoviricetes (queried as `lineage:"Caudoviricetes"`) and an 18-family sum of human-infecting virus families enumerated explicitly.

B. Multi-group biological dataset: assembly details

Cellular groups. For each cellular group we issue a single UniProt query of the form

```
reviewed:true AND taxonomy_id:<ID> AND (existence:1 OR existence:2 OR
existence:3) AND NOT keyword:KW-0181,
```

restricting to manually reviewed Swiss-Prot entries with experimental, transcript, or homology evidence (PE1–PE3). The final clause excludes proteins flagged `KW-0181` (COMPLETE PROTEOME): UniProt attaches this keyword to every entry belonging to an organism whose proteome has been designated a reference proteome, and these few heavily sampled species (human, *E. coli*, *S. cerevisiae*, *A. thaliana*, ...) each contribute tens of thousands of reviewed Swiss-Prot entries, whereas most cellular taxa contribute only dozens. Without the filter a per-taxon query would therefore be dominated by the proteomes of a few model organisms rather than sample broadly across the tree of cellular life, biasing the pooled cellular amino acid distribution accordingly.

Non-human-host viruses. These groups are queried by host taxon: `virus_host_id:33090` for plant-infecting and `virus_host_id:6656 AND NOT virus_host_id:9606` for invertebrate-infecting viruses.

Bacteriophages. Assembled in a two-step pass: we download all reviewed viral sequences (`taxonomy_id:10239`) and retain those whose lineage annotation lies within a curated list of phage-relevant clades: Caudoviricetes, Duplodnaviria, Microviridae, Inoviridae, Tectiviridae, Leviviridae, Cystoviridae, Corticoviridae, Plasmaviridae, Caudovirales and organism name containing the substring "phage".

Local filtering. Downloaded sequences are filtered to length $\in [50, 1022]$ residues (the common pLM context window, minus BOS/EOS), $< 5\%$ non-standard amino acids, and exact-sequence-deduplicated by SHA-256 hash. If more than $5{,}000$ sequences survive these filters we draw a uniform random subsample of size $5{,}000$ with random seed 42; otherwise we keep all surviving sequences.

Post-filter pool sizes. Plant viruses 954, invertebrate viruses $1{,}395$, bacteriophage $1{,}262$; bacteria, plants, fungi, insects each capped at the $5{,}000$ subsample; archaea is retained in full at $17{,}379$ sequences; human viruses use the curated $5{,}203$-sequence pool of Section 3.2.

C. Human classification dataset: assembly and composition

Viral group. The viral positives are the union of two sources. First, $6{,}044$ Swiss-Prot (UniProt Consortium, 2025) entries returned by

```
reviewed:true AND virus_host_id:9606 AND (existence:1 OR existence:2 OR
existence:3),
```

restricting to manually reviewed entries with experimental, transcript, or homology evidence (PE1–PE3). Second, 620 RefSeq NP_ (Goldfarb et al., 2025) entries manually exported from NCBI Virus (Brister et al., 2015) under a human-host filter, capturing curated viral proteins not yet in Swiss-Prot.

Cellular group. Cellular negatives are drawn from the non-viral fraction of Swiss-Prot (reviewed entries from any taxon other than viruses). Sequences are bucketed into length deciles of the viral pool and a uniform random subsample is drawn from each decile so that the cellular length distribution matches the viral one.

The human-virus pool is composed of 32 viral families. The pool also covers all seven Baltimore classes, a standard virology taxonomy (Baltimore, 1971) that groups viruses by the form of their genome and the route used to transcribe it to mRNA: double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), positive-sense single-stranded RNA (+ssRNA), negative-sense single-stranded RNA (−ssRNA), and the two reverse-transcribing classes ssRNA-RT (e.g. retroviruses) and dsDNA-RT (e.g. hepadnaviruses). Therefore our dataset spans the full range of viral replication strategies rather than a single DNA or RNA lineage.

Table 2. **Family composition of the human-virus pool.** Per-family sequence counts and percentages after exact deduplication and length/composition filtering. The top five families account for 70.1% of the pool; 230 sequences (4.4%) carry no family-level annotation.

Family	n	%
Orthomyxoviridae	1,026	19.7
Orthoherpesviridae	990	19.0
Poxviridae	681	13.1
Retroviridae	484	9.3
Papillomaviridae	468	9.0
Sedoreoviridae	234	4.5
Hepadnaviridae	194	3.7
Adenoviridae	177	3.4
Paramyxoviridae	163	3.1
Rhabdoviridae	106	2.0
Filoviridae	80	1.5
Coronaviridae	54	1.0
Pneumoviridae	52	1.0
Kolmioviridae	33	0.6
Arenaviridae	30	0.6
Polyomaviridae	28	0.5
Flaviviridae	24	0.5
Hantaviridae	20	0.4
Hepeviridae	19	0.4
Anelloviridae	19	0.4
Parvoviridae	17	0.3
Phenuiviridae	15	0.3
Astroviridae	13	0.2
Picornaviridae	11	0.2
Caliciviridae	11	0.2
Peribunyaviridae	8	0.2
Nairoviridae	5	0.1
Togaviridae	4	0.1
Spinareoviridae	2	< 0.1
Picobirnaviridae	2	< 0.1
Circoviridae	2	< 0.1
Tobaniviridae	1	< 0.1
(unannotated)	230	4.4
Total	**5,203**	**100.0**

D. Embedding extraction

For locally executed checkpoints we read the final-layer hidden states directly from the `transformers/esm` forward pass. For Forge-API-only ESM3 configurations (SMALL, MEDIUM, LARGE) we encode each sequence via `tokens = client.encode(ESMProtein(sequence=s))` and call `client.logits(tokens,`

Table 3. Within-group PCA refit. For each group, PCA is fit on that group's ESMC-600M embeddings alone, and Spearman ρ between each of the first three principal components and masked-reconstruction PPL is reported. var_1 is the fraction of variance captured by PC_1 of the refit. The final column identifies the PC carrying the largest $|\rho|$ with PPL.

Group	n	var_1 (%)	ρ_{PC_1}	ρ_{PC_2}	ρ_{PC_3}	$\arg\max_k \lvert\rho_{\mathrm{PC}_k}\rvert$
Biological groups						
Archaea	17,379	28.6	+0.849	+0.021	+0.052	PC_1
Bacteria	5,000	30.1	+0.874	+0.001	−0.137	PC_1
Fungi	5,000	51.5	+0.963	−0.041	−0.157	PC_1
Insects	5,000	44.5	+0.955	−0.053	−0.210	PC_1
Plants	5,000	38.7	+0.934	+0.061	−0.095	PC_1
Human non-viral	5,197	41.5	+0.905	+0.061	−0.105	PC_1
Bacteriophage	1,262	54.3	+0.945	+0.138	−0.037	PC_1
Plant virus	954	59.5	+0.893	+0.047	−0.061	PC_1
Invertebrate virus	1,395	47.9	+0.905	+0.054	+0.103	PC_1
Human virus	5,203	50.0	+0.830	−0.112	−0.032	PC_1
Pooled groups						
Cellular-only (6 groups)	42,576	36.8	+0.910	−0.144	+0.055	PC_1
Viral-only (4 groups)	8,814	56.6	+0.894	−0.048	+0.024	PC_1
Biologically meaningless controls						
Shuffled cellular	5,197	19.6	−0.111	−0.398	−0.097	PC_2
Shuffled viral	5,203	22.9	−0.070	−0.269	−0.146	PC_2
Random uniform	4,999	34.4	−0.031	−0.018	+0.010	PC_1

`LogitsConfig(return_embeddings=True))`, which returns the same final-layer activations as the open weights would. Mean-pooling over the residue axis (excluding BOS/EOS) is applied identically to both sources.

Embedding dimensionality d varies across the model registry: 320 (ESM2-8M), 480 (ESM2-35M), 640 (ESM2-150M), 1,280 (ESM2-650M), 2,560 (ESM2-3B), 5,120 (ESM2-15B); 960 (ESMC-300M), 1,152 (ESMC-600M), 2,560 (ESMC-6B); and 1,536, 2,560, 6,144 for Forge ESM3 SMALL, MEDIUM, LARGE. The pooling recipe is identical across all models.

E. Perplexity vocabulary

ESM2, ESMC, and the ESM3 sequence track all use a 33-token sequence vocabulary (Lin et al., 2023a; EvolutionaryScale Team, 2024; Hayes et al., 2025) including the 20 standard amino acids, the 5 IUPAC ambiguity codes `X, B, U, Z, O`, and 8 additional special and formatting tokens including `<cls>`, `<pad>`, `<eos>`, and `<mask>`.

The softmax denominator in Section 3.1 is taken over this full vocabulary. Consequently the absolute upper bound on PPL is $|\mathcal{V}|$=33 (uniform over all tokens). Therefore we can observe a PPL greater than 20 reflecting a small leakage of probability mass onto the 13 non-amino acid tokens. That's the case for the random uniform PPL.

F. Within-group PCA: PC1 is the nativeness axis

Figure 1A fits PCA on the pooled embeddings of all 13 sequence pools: ten biological groups plus three synthetic controls, corresponding to the higher-level categories cellular, viral, shuffled, and random. Because these groups have different centroids and perplexity ranges, PCA1 may be biased. TThis bias, called anisotropy bias, was documented for transformer representations in Ethayarajh (2019). The reported $\rho(\mathrm{PC}_1, \mathrm{PPL})$=+0.961 could therefore be inflated by the cellular–viral–random group separation alone, rather than reflecting a continuous per-sequence nativeness coordinate.

We test this directly by *refitting PCA inside each group* and recomputing $\rho(\mathrm{PC}_k, \mathrm{PPL})$ on that refit for $k \in \{1, 2, 3\}$. All embeddings are ESMC-600M mean-pooled representations, as in Figure 1A. Table 3 reports the within-group refit for every natural population and for the three biologically meaningless controls, together with two pooled anchors (cellular-only and viral-only) that exclude the other group entirely. We also report, for each group, the PC on which $|\rho|$ is largest over the top-3 components.

Two observations follow.

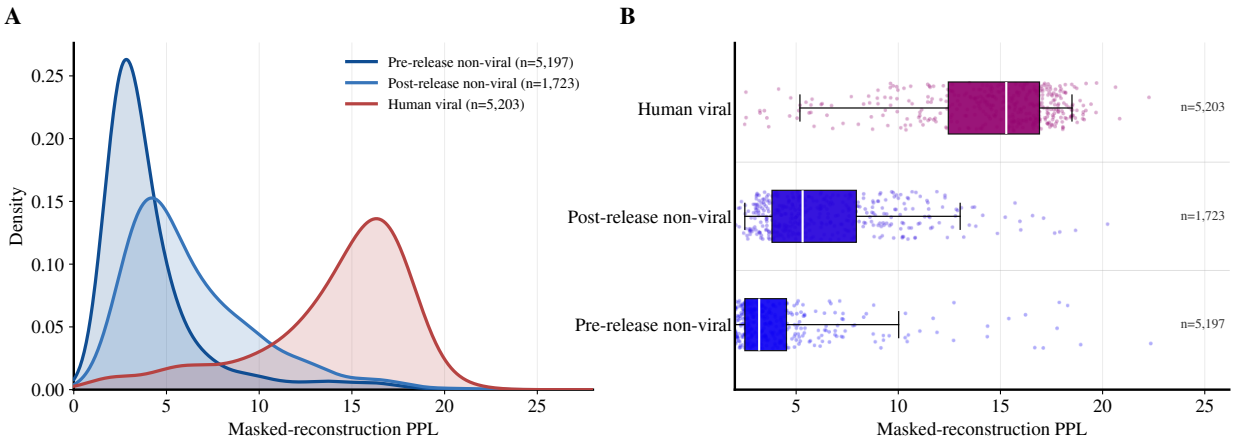

Figure 4. **Post-release cellular proteins remain native-like under ESMC-600M.** Masked-reconstruction perplexity for three sequence groups: cellular proteins already present in sequence databases before the checkpoint release ($n = 5{,}197$), post-release cellular Swiss-Prot proteins created on or after 2025-01-01 ($n = 1{,}723$), and the curated human-virus pool from Section 3.2 ($n = 5{,}203$).

(i) The alignment survives every single group refit. Across all ten biological groups, PC_1 of the group refit is the PPL-aligned axis, with $\rho(\mathrm{PC}_1, \mathrm{PPL}) \in [+0.830, +0.963]$. PC_2 and PC_3 carry near-zero PPL signal in every case ($|\rho| \leq 0.21$).

(ii) The alignment is not a generic property of PCA on embeddings. Fitting PCA on i.i.d. uniform random sequences yields a PC_1 that is still a real top direction of variance (34.4%) but carries no PPL signal on any of its first three components ($|\rho| \leq 0.03$).

G. Post-release cellular sequence-novelty control

The goal is to test whether the cellular–viral perplexity gap is primarily explained by training exposure. If exposure were the dominant factor, then cellular proteins released after the model checkpoint, and therefore absent from that model's pretraining data, should move toward the viral distribution. We find that they do not: post-release cellular proteins remain much closer to the cellular reference distribution than to viral proteins.

Dataset. We use ESMC-600M, released as `esmc-600m-2024-12`. Because the checkpoint predates entries created in 2025, we query UniProtKB Swiss-Prot (UniProt Consortium, 2025) for reviewed, cellular entries (`NOT taxonomy_id:10239`) with `date_created` on or after 2025-01-01. These sequences were not available when the released checkpoint was produced. We apply the same filters used for the main sequence pools: length in $[50, 1022]$, fewer than 5% non-standard amino acids, and exact-sequence deduplication by SHA-256 hash (Section B). This yields $1{,}723$ post-release cellular sequences.

Scoring. We compute masked-reconstruction perplexity with the same protocol used throughout the paper (Section 3.1): 15% masking, three independent mask samples per sequence, and the ESMC-600M sequence model. As reference distributions, we reuse the ESMC-600M perplexities from the human viral/cellular classification dataset (Section 3.2), pooling the train, validation, and test splits. This gives $5{,}197$ cellular proteins that were already present in sequence databases before the checkpoint release, and $5{,}203$ curated human-virus proteins.

Result. The post-release cellular proteins remain native-like (Figure 4). Median PPL is 3.2 for the pre-release cellular reference set, 5.3 for the post-release cellular set, and 15.3 for the viral set. Thus the post-release distribution is shifted upward relative to the pre-release cellular reference distribution, but the shift is small compared with the viral displacement. cellular proteins absent from the checkpoint's pretraining data remain far more native to the model than viral proteins, indicating that nativeness reflects compatibility with the learned cellular-dominated protein prior rather than raw exposure alone.

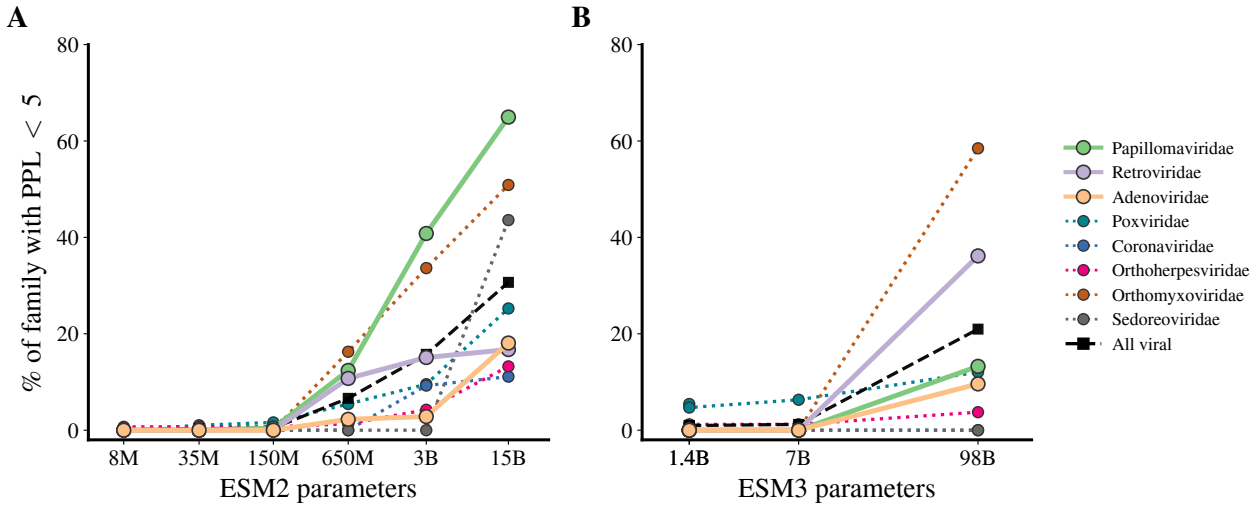

Figure 5. **Per-family nativization across ESM2 and ESM3.** Fraction of each viral family with masked-reconstruction PPL< 5 as a function of parameter count. The threshold is the same fixed PPL< 5 cut used in Figure 2

H. ESM2 and ESM3 per-family nativization

We verify that the same heterogeneous scaling is present in the other two ESM families rather than an ESMC-specific effect. Figure 5 We repeat the same analysis for the other two ESM architectures: ESM2 (8M $\rightarrow$ 15B, six scales) and ESM3 (SMALL $\rightarrow$ OPEN $\rightarrow$ MEDIUM $\rightarrow$ LARGE, four scales). We find that the same families that become increasingly native-like under ESMC scaling (Papillomaviridae and Retroviridae) also rise under ESM2 and ESM3 scaling.

I. ESM2 and ESM3 nativeness axis

We verify that the same *nativeness axis* is present in the other two ESM families rather than an ESMC-specific effect, we repeat the PCA + PPL analysis on one checkpoint from each of ESM2 and ESM3 (Figure 6).

We find that the qualitative claims of main Figure 1A survive in both other architectures (Figure 6). **(i) Low-dimensionality.** PC_1 alone captures 54.3% of the mean-pooled embedding variance on ESM2-650M and 67.3% on ESM3-OPEN. **(ii) PC_1-PPL alignment.** The Spearman rank correlation between PC_1 and masked-reconstruction perplexity is $\rho{=}{+}0.926$ on ESM2-650M and $\rho{=}{+}0.935$ on ESM3-OPEN. Both values are similar to the ESMC $\rho{=}{+}0.961$. **(iii) Group ordering on PC_1.** The cellular groups sits at one end of PC_1 at low PPL; the three biologically meaningless controls (shuffled cellular, shuffled viral, i.i.d. uniform random) sit at the high-PPL end; and the human-virus cloud lies between the cellular groups and the biologically meaningless tail, exactly as in main Figure 1B.

J. Low-FPR classification performance

We report TPR at fixed low FPR thresholds to evaluate performance in the setting relevant for screening applications. Across all model families and scales, embedding linear probes substantially outperform PPL-based classifiers. The gap is largest at very low FPR (0.1% and 1%), where PPL-based discrimination remains limited even for large models, while embedding linear probes approach near-ceiling performance.

These results confirm that the viral signal captured by embeddings is not only stronger in aggregate metrics such as AUC, but also more effective in practical low-false-positive settings.

K. Probe robustness: harder negatives and held-out families

We run two additional robustness checks for the linear-probe results in Section 4.3. These controls test whether the near-ceiling viral/cellular separability reflects viral-specific signal rather than a coarse dataset contrast or memorization of family-specific motifs.

Harder negatives. In the main classification dataset, cellular negatives are drawn from a length-matched multi-kingdom Swiss-Prot sample. A probe could therefore in principle exploit broad differences between human-infecting viruses and distant cellular taxa such as bacteria, plants, or fungi. To test this possibility, we re-evaluate each trained probe after restricting the held-out negative class to *Homo sapiens* proteins, matching the host infected by the viral positives.

Probe performance is essentially unchanged (Figure 8). Human-negative AUC-ROC remains at least 0.95 for every pLM with at least 35M parameters, and only the smallest model, ESM2-8M, drops to 0.886. Thus, the probe separates viral proteins from same-host cellular proteins, not only from distant cellular kingdoms. Because the held-out human-negative pool is small (n=38; the multi-kingdom test split is only $\sim$3.7% human), this estimate is noisier than the full-pool AUC.

Held-out families. We next test whether the probe relies on family-specific motifs. We run leave-one-family-out cross-validation over the 13 viral families with at least 50 sequences. For each split, one viral family is removed entirely from training and used only for evaluation.

Held-out family performance remains close to the full-probe performance (Figure 9). The median held-out AUC-ROC is 0.984, and only three evaluations fall below 0.92. All three correspond to Retroviridae, consistent with the endogenous-retrovirus overlap with host genomes discussed in Section 4.2 and with Retroviridae being closest to the cellular manifold. These results indicate that the probe captures a family-general viral signal rather than memorizing family-specific sequence motifs.

L. Cross-architecture generalization: autoregressive and diffusion pLMs

The main analysis focuses on the ESM family, whose models are trained with masked or span-denoising reconstruction objectives. We next ask whether the same three results hold outside this family. We evaluate two non-ESM, sequence-only generative pLMs: **ProGen2-base** (764M parameters), a decoder-only autoregressive Transformer (Nijkamp et al., 2023), and **EvoDiff OA-DM** (640M parameters), a discrete order-agnostic diffusion model (Alamdari et al., 2023). Both models are evaluated on the same 13-group sequence pool ($n = 66{,}790$) and the same held-out human viral/cellular classification split used in the main experiments.

Perplexity has a different meaning for each objective: masked-token reconstruction for ESM, causal next-token likelihood for ProGen2, and the order-agnostic ELBO for EvoDiff (Hoogeboom et al., 2022). These values are therefore not comparable in absolute scale across architectures. We only compare within-model quantities: the alignment between PC_1 and perplexity, the ordering of sequence groups, and the relative performance of embedding probes, perplexity-only classifiers, and shallow sequence baselines.

Models and embeddings. The two added models differ from ESM in both architecture and training objective. ProGen2-base is a decoder-only Transformer trained by left-to-right next-residue prediction, whereas ESM models use bidirectional masked reconstruction. EvoDiff OA-DM is a dilated-convolutional ByteNet model with an explicit diffusion-timestep input, trained to reconstruct residues in an order-agnostic generation process rather than under a fixed mask rate.

For each model, we extract one embedding per protein by mean-pooling residue-level hidden states over amino-acid positions only, excluding special and padding tokens. For ProGen2, we mean-pool the final-layer hidden states of the clean sequence after removing BOS/EOS tokens, giving a 1536-dimensional embedding. For EvoDiff, we mean-pool the pre-decoder ByteNet representation of the clean sequence at the fully generated timestep $t = L$, captured before the final projection layer, giving a 1280-dimensional embedding. PCA, linear probes, and all embedding analyses use these vectors without further modification. Perplexity is computed using each model's native objective: exact causal per-residue likelihood for ProGen2 and a 24-sample OA-ARDM ELBO estimate for EvoDiff.

L.1. The nativeness axis appears to extend beyond masked language modeling

Figure 10 repeats the main nativeness-axis analysis for ESMC-600M, ProGen2-base, and EvoDiff OA-DM. The same structure appears outside the ESM family. PC_1 remains strongly aligned with reconstruction difficulty, with $\rho(\mathrm{PC}_1, \mathrm{PPL}) = +0.903$ for ProGen2 and $+0.951$ for EvoDiff. PC_1 explains 54.5% of variance for ProGen2 and 27.4% for EvoDiff, compared with 73.1% for ESMC-600M.

The group ordering is also preserved. In both non-ESM models, cellular proteins have the lowest perplexity, followed by human viral proteins, shuffled cellular sequences, shuffled viral sequences, and random sequences. The group-mean

perplexities follow this ordering for ProGen2 ($7.2 < 14.0 < 18.3 < 18.7 < 22.1$) and for EvoDiff ($10.5 < 16.1 < 18.1 < 18.6 < 21.9$). Thus, as in ESM, viral proteins occupy an intermediate region between well-modeled cellular proteins and biologically meaningless controls.

The PC_1–PPL alignment is strongest within the biological groups. On the shared PC_1, the within-group correlations for cellular and viral proteins are $+0.82$ and $+0.86$ for ProGen2, and $+0.91$ and $+0.71$ for EvoDiff. The weaker viral correlation for EvoDiff suggests that the diffusion model distributes viral variation across more directions, but the dominant axis still tracks reconstruction difficulty. Together, these results show that the nativeness axis is not an artifact of the masked-LM objective. It also appears in autoregressive and order-agnostic diffusion pLMs.

L.2. Scale contracts nativeness heterogeneously outside the ESM family, but the family ranking depends on objective

The fixed $\mathrm{PPL} < 5$ native-like threshold used in Figure 2 is not meaningful across objectives. ProGen2 already assigns cellular proteins a median perplexity below 5 at 764M parameters, whereas EvoDiff does not reach this scale because its score is an ELBO-based perplexity bound. We therefore define a model-family-specific native-like threshold τ as the 90th percentile of cellular perplexity at the reference scale, giving $\tau = 14.2$ for ProGen2-base and $\tau = 16.2$ for EvoDiff-640M. This threshold is then held fixed across scales within each architecture. Consequently, the vertical axis of Figure 11 should not be compared in absolute value with Figure 2; only within-architecture trends and family rankings are interpretable.

For ProGen2, evaluated across four scales from 151M to 6.4B parameters, scaling again contracts nativeness selectively rather than uniformly. The fraction of viral sequences below τ increases from 20% at 151M to 52% at 6.4B, while the cellular median perplexity decreases from 6.5 to 2.7. However, the families that become native-like are not the same as in ESMC. In ESMC, Papillomaviridae and Retroviridae move most strongly toward the native region, whereas Orthomyxoviridae, Orthoherpesviridae, and Sedoreoviridae remain displaced. In ProGen2, the strongest nativizers are instead Orthomyxoviridae and Sedoreoviridae. Orthomyxoviridae drops from median PPL 17.8 to 1.8, with the fraction below τ increasing from 13% to 82%, and Sedoreoviridae increases from 0% to 71% below τ.

Other patterns are shared across objectives. Papillomaviridae nativizes in both ESMC and ProGen2, increasing from 4% to 69% below τ in ProGen2. Orthoherpesviridae remains displaced in both objectives, as do Adenoviridae, Coronaviridae, and Poxviridae, whose median perplexities do not decrease with scale ($\Delta\mathrm{PPL}$ between $+0.4$ and $+1.3$). Retroviridae is the most consistent case: it is already largely native-like at the smallest ProGen2 scale, with 76% of sequences below τ, and plateaus thereafter. Thus, scaling nativizes some viral families and not others outside the ESM family, but the detailed family ranking is objective-dependent. The family ordering reported in Section 4.3 should therefore be interpreted as partly ESM-specific.

EvoDiff shows a complementary pattern across its two available scales, 38M and 640M. The cellular median perplexity decreases substantially, from 14.5 to 8.0, whereas viral family medians remain nearly flat, with per-family changes between -0.1 and -0.7. As a result, the viral–cellular gap widens rather than closes with scale. Because only two EvoDiff scales are available, this should be interpreted as a direction-of-effect result rather than a full scaling law. The largest viral shift is again observed for Papillomaviridae ($\Delta = -0.7$), consistent with its behavior under ESM.

These cross-architecture results also support the homology-based interpretation of Retroviridae from Section 4.3. Retroviridae is the most native-like viral family at the smallest scale of each architecture: ESMC-300M (21% below the $\mathrm{PPL} < 5$ threshold), ProGen2-151M (76% below τ), and EvoDiff-38M (rank 1). It also remains among the most native-like families at the largest scale of each architecture, ranking first for ProGen2-6.4B and EvoDiff-640M, and second for ESMC-6B. Unlike families such as Papillomaviridae or Orthomyxoviridae, Retroviridae does not require scale to become native-like; it starts near the native region. This is consistent with the sequence-level mechanism proposed in Section 4.3: retroviral protein domains overlap the cellular training distribution through eukaryotic retrotransposons. The Retroviridae result is therefore not only reproduced outside ESM, but strengthened by the fact that its nativeness is stable across training objectives.

L.3. Viral identity remains linearly encoded beyond reconstruction difficulty outside the ESM family

We next repeat the probe-versus-perplexity comparison of Section 4.3. outside the ESM family (Figure 12). For each scale of ProGen2 and EvoDiff, we train a logistic-regression probe on mean-pooled embeddings and compare it with a perplexity-only zero-shot classifier, $s = -\mathrm{PPL}$, and the same shallow-feature baselines. All scores are computed on the same held-out human viral/cellular test split.

The embedding probe remains near ceiling across both non-ESM architectures and improves with scale. ProGen2 increases

from AUC-ROC 0.975 at 151M parameters to 0.993 at 6.4B, while EvoDiff increases from 0.965 at 38M to 0.986 at 640M. These values fall within the range of the 13 ESM embedding probes (0.967–0.997) and remain well above the best shallow-feature baseline (AUC 0.883, dipeptide composition). In contrast, the perplexity-only classifier is substantially weaker. For ProGen2, its AUC decreases from 0.82 at 151M to 0.72 at 6.4B, despite the embedding probe improving over the same scale range. For EvoDiff, the PPL classifier improves from 0.78 to 0.88, but remains at or below the shallow-feature baseline. Thus, as in the ESM family, scaling can make the embedding representation more linearly informative while leaving perplexity-based discrimination weak or even worse at larger scales.

The signal generalizes to held-out viral families. Leave-one-family-out evaluation over the 13 viral families with at least 50 sequences gives mean held-out AUC 0.976 for ProGen2 and 0.970 for EvoDiff, close to the full-probe performance. This mirrors the family-level robustness analysis in Appendix K: the probe is not simply memorizing the dominant viral families, but recovers a viral signal that transfers across family boundaries. Together, these results show that residual viral separability beyond reconstruction difficulty is not specific to the masked-LM ESM family.

Summary. Across one autoregressive and one diffusion pLMs, the nativeness axis and residual viral embedding signal both reproduce outside the ESM family. Scaling also remains heterogeneous across viral families, but the family ranking changes across objectives. These results suggest that the main findings are not artifacts of masked language modeling.

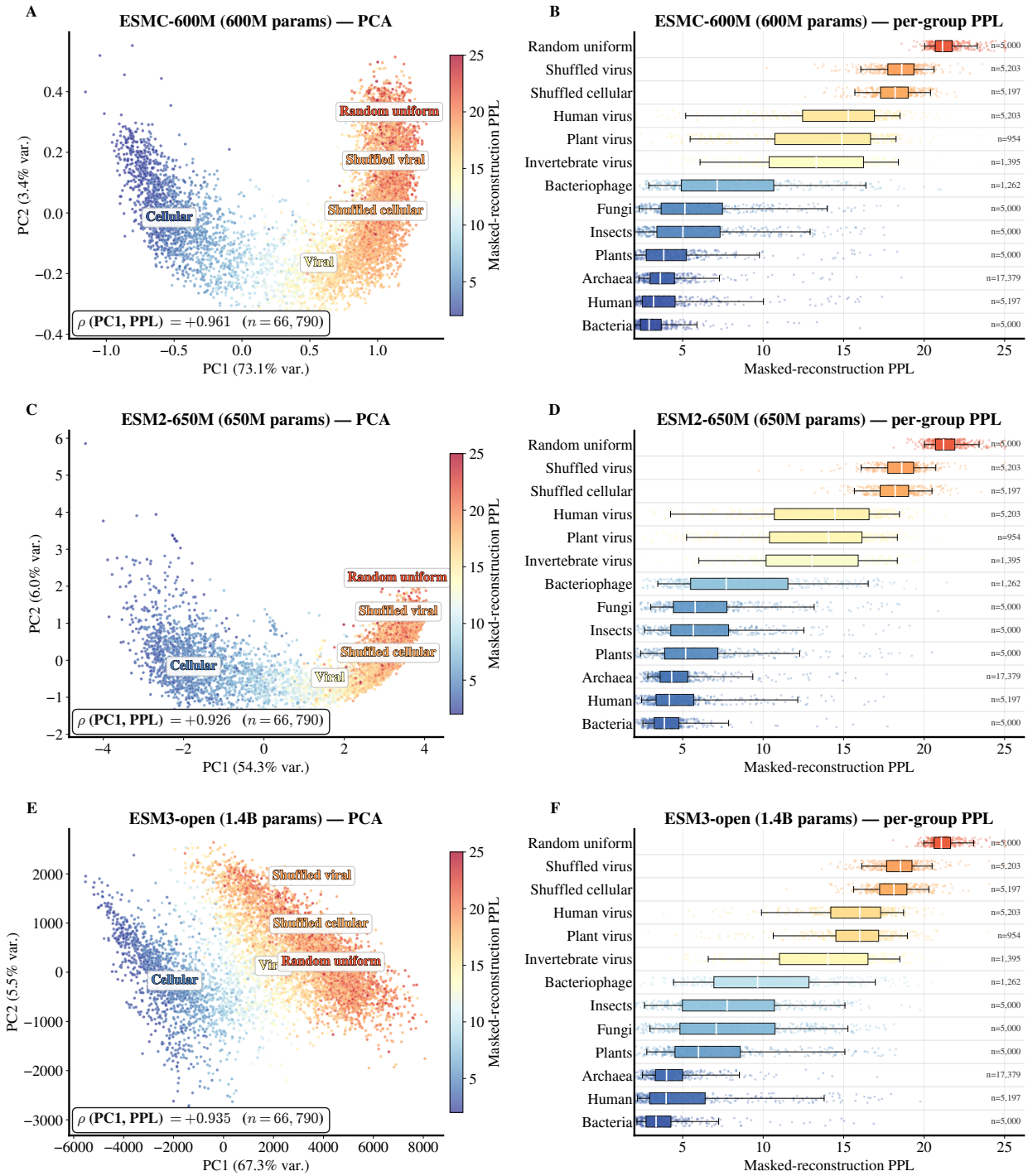

Figure 6. **The nativeness axis reproduces on ESM2-650M and ESM3-OPEN.** PCA of mean-pooled embeddings, points coloured by masked-reconstruction PPL with the same colourmap as main Figure 1A.

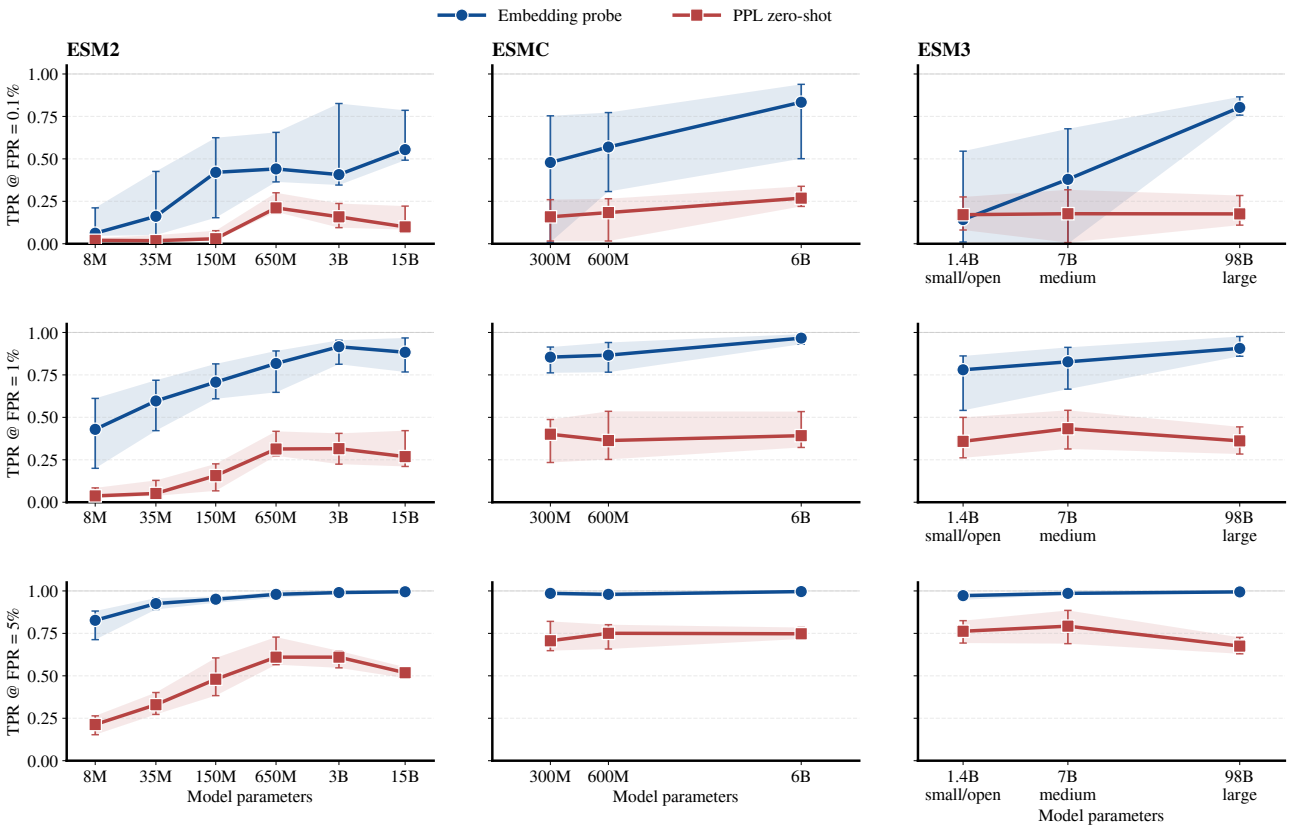

Figure 7. True positive rate (TPR) at low false positive rates (FPR) for embedding linear probes (blue) and PPL-based classifiers (red) across ESM2, ESMC, and ESM3 model families. Rows correspond to operating points at 0.1%, 1%, and 5% FPR. Shaded regions and error bars indicate 95% percentile bootstrap confidence intervals over 2,000 resamples of the held-out human test set (n=2,080). Embedding linear probes consistently achieve substantially higher TPR than PPL-based classifiers across all scales, with the gap most pronounced in the low-FPR setting relevant for screening.

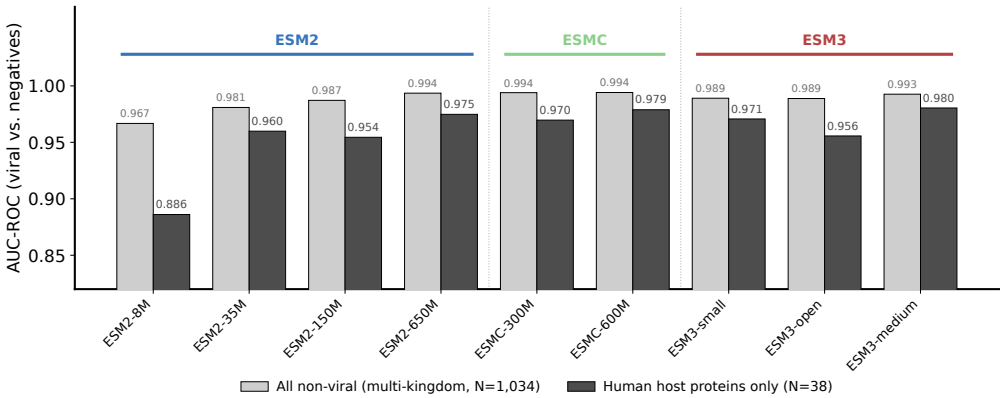

Figure 8. **The probe distinguishes viral proteins from the human host itself, not only from distant cellular kingdoms.** Per-model probe AUC-ROC on the held-out human-virus test split, evaluated against the full multi-kingdom Swiss-Prot negative pool (light, n=1,034) and against *Homo sapiens* negatives only (dark, n=38). Human-only AUC remains at least 0.95 for every model with at least 35M parameters; only ESM2-8M drops, to 0.886.

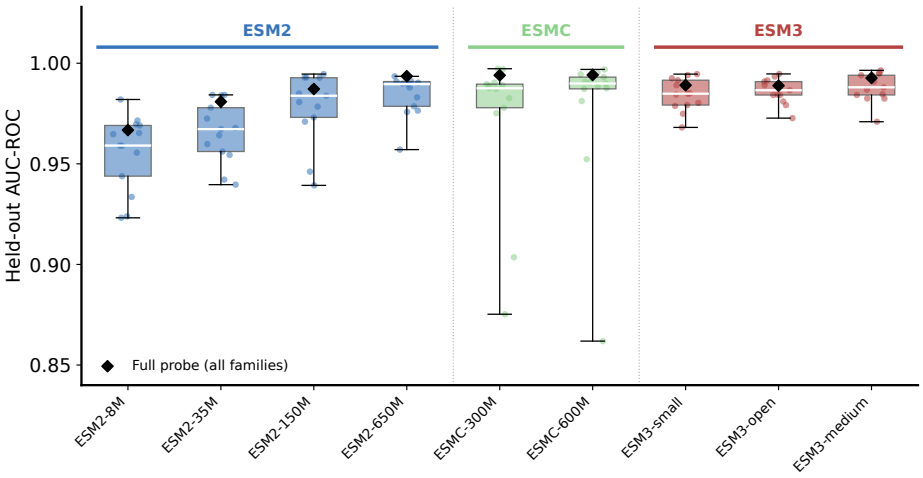

Figure 9. **Held-out viral families remain separable, indicating that the probe does not simply memorize family-specific motifs.** Leave-one-family-out cross-validation across the 13 viral families with at least 50 sequences. For each pLM, the box and jittered points show held-out AUC-ROC across families; the black diamond marks the model's full within-distribution probe AUC. Held-out AUC closely tracks the full probe. The lowest ESMC evaluations correspond to Retroviridae, down to 0.862.

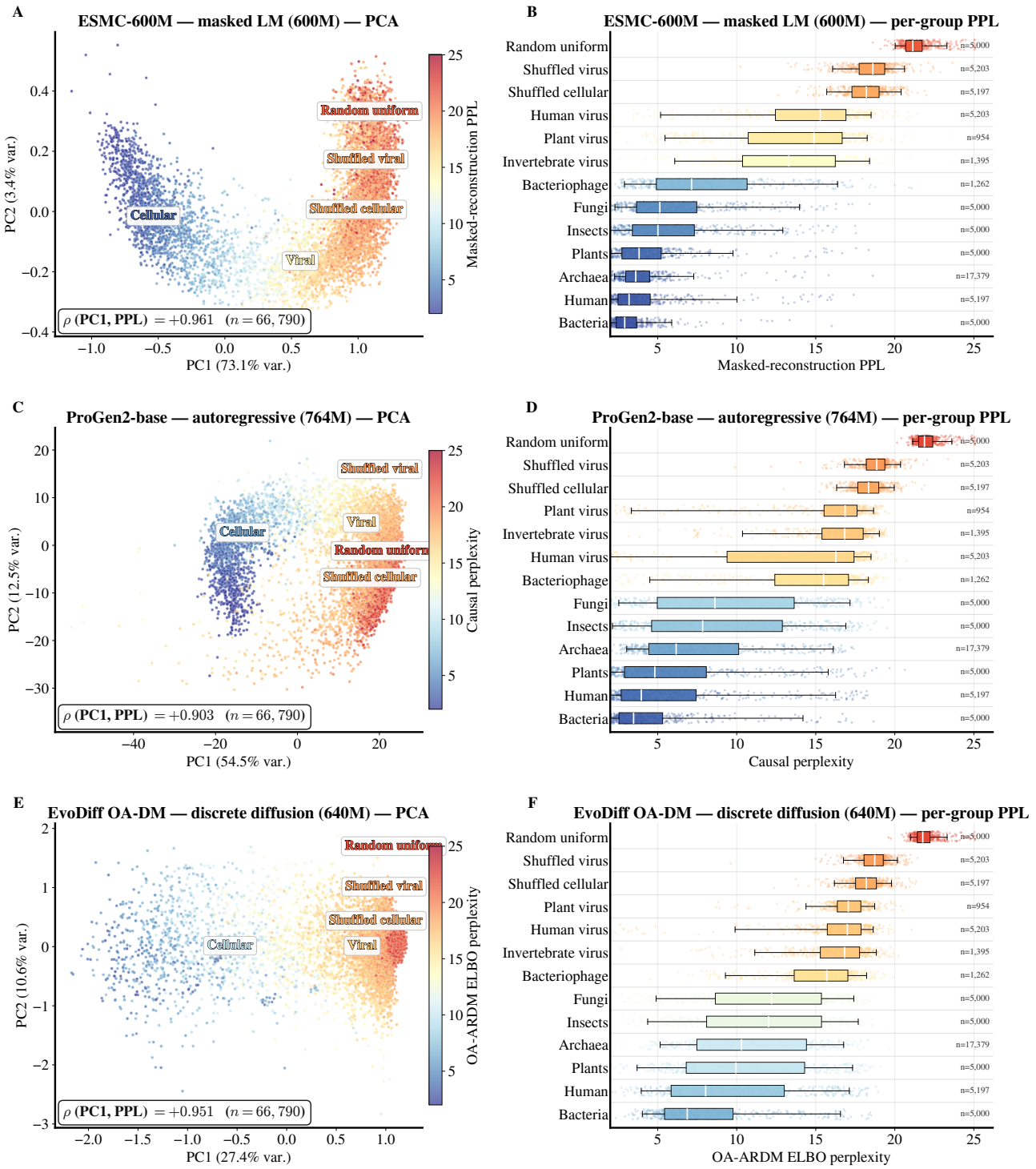

Figure 10. **The nativeness axis appears beyond the masked-LM objective.** Figure 1 analysis repeated for three pLMs with different training objectives, evaluated on the same 13-group sequence pool ($n = 66{,}790$): ESMC-600M (masked LM; reference), ProGen2-base (autoregressive), and EvoDiff OA-DM (order-agnostic discrete diffusion). *Left:* PCA of mean-pooled sequence embeddings, colored by each model's native per-residue perplexity score: masked-reconstruction perplexity for ESMC, causal perplexity for ProGen2, and the OA-ARDM ELBO-based perplexity estimate for EvoDiff. Insets report $\rho(\mathrm{PC}_1, \mathrm{PPL})$. *Right:* group-wise perplexity distributions, shown as box and strip plots and ordered by median. PC_1 is strongly aligned with perplexity in all three models ($\rho = +0.961$, $+0.903$, and $+0.951$), and the same ordering from cellular proteins to viral proteins to shuffled and random controls is preserved. Because perplexity is objective-specific, absolute values should not be compared across rows; the shared color scale is used only for visual consistency.

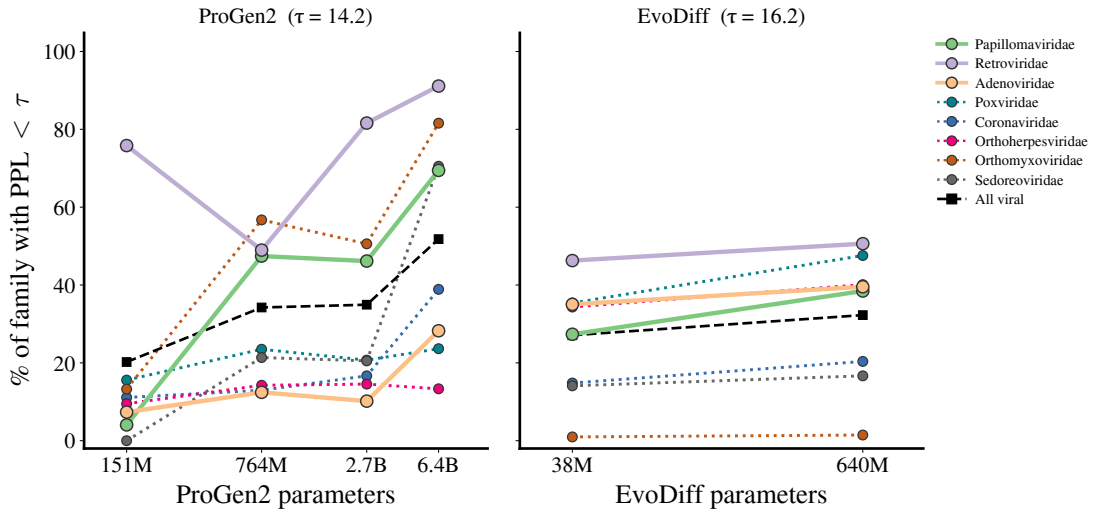

Figure 11. **Scaling remains heterogeneous outside the ESM family, but the family ranking depends on the objective.** Fraction of each human viral family below a model-family-specific native-like threshold τ, plotted as a function of parameter count for ProGen2 (autoregressive; four scales) and EvoDiff (diffusion; two scales). For each architecture, τ is defined as the 90th percentile of cellular perplexity at the reference scale and is then held fixed across scales. Solid lines mark the three families that nativize most strongly under ESMC-6B in Figure 2; colors are matched across figures. ProGen2 shows heterogeneous nativization, but its strongest nativizers are Orthomyxoviridae and Sedoreoviridae, two families that remain displaced under ESMC. EvoDiff shows a weaker viral-family response over its two available scales. Since τ is architecture-specific, the vertical axis is not directly comparable to the fixed $\mathrm{PPL} < 5$ axis in Figure 2.

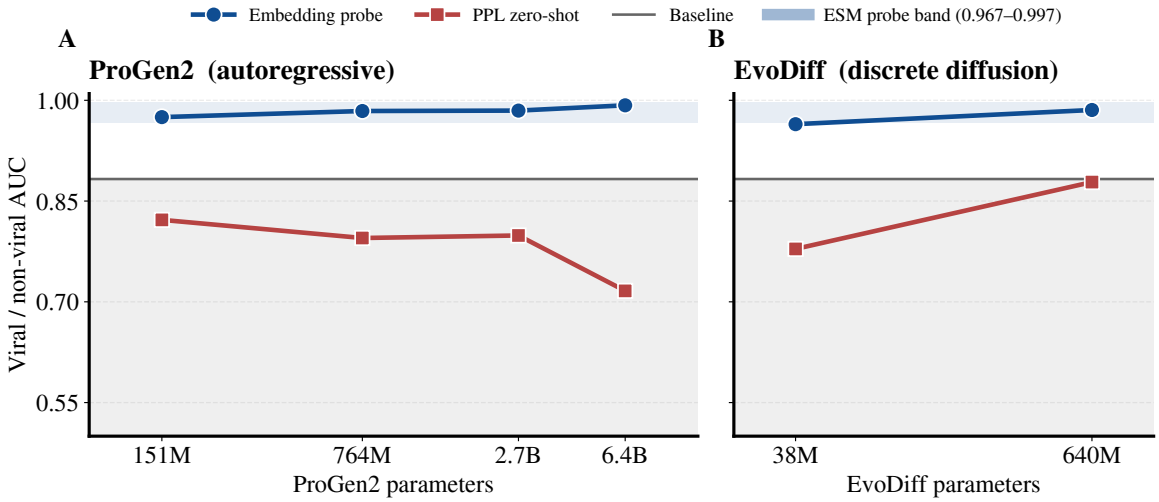

Figure 12. **Embedding probes outperform perplexity-only classifiers across scale outside the ESM family.** Human viral/cellular AUC-ROC on the held-out test split, as a function of model scale, for **(A)** ProGen2 (autoregressive; 151M–6.4B parameters) and **(B)** EvoDiff OA-DM (discrete diffusion; 38M–640M parameters). Blue points show logistic-regression probes trained on mean-pooled embeddings from each scale. Red points show the corresponding perplexity-only zero-shot classifier ($s = -\mathrm{PPL}$, reported as $\max(\mathrm{AUC}, 1 - \mathrm{AUC})$ as in Section 4.3). The gray band shows the shallow-feature baseline range, whose upper edge is the best length, amino-acid-composition, or dipeptide baseline ($\mathrm{AUC} = 0.883$). The blue band shows the range of ESM embedding-probe AUCs from Figure 3 (0.967–0.997). For both non-ESM architectures, the embedding probe is near ceiling and lies within the ESM range at every scale, whereas the perplexity-only classifier is substantially weaker. In ProGen2, this gap widens with scale: the embedding probe improves, while the perplexity-only classifier drops to AUC 0.72 at 6.4B. EvoDiff perplexity denotes the OA-ARDM ELBO-based estimate.