Disentangling Protein Family Signals in Protein Language Models: Composition or Motifs?

Sohel Bashar
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
sohelbashar11@gmail.com

M. Saifur Rahman
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
mrahman@cse.buet.ac.bd

Abstract

We test whether family-level separation in protein language model (pLM) embeddings persists after controlling for amino-acid composition. For six Pfam families and four models (ESM-2, ProtBERT, ProtXLNet, ProteinBERT), we compute layer-wise within-family (In) and between-family (Out) cosine similarities for true sequences and composition-preserving shuffles. We report a ratio-based *fidelity* (In/Out) and a difference metric $\Delta = \text{In} - \text{Out}$, and visualize the geometry with t-SNE against a pooled negative bank. Across families and models, shuffled curves closely track true curves in both fidelity and Δ, and frequently match or exceed them. ProtXLNet's fidelity rises with depth but the shuffled curve is typically comparable or higher; ProtBERT's mid-layer spike is mirrored by shuffles; ESM-2 and ProteinBERT are weak overall. t-SNE strips show compact clusters for both true and shuffled sequences with negatives separated. These results indicate that amino-acid composition accounts for much of the apparent family fidelity in current embeddings, highlighting the need for composition-controlled baselines and biologically meaningful evaluation metrics.

1 Introduction

Protein language models (pLMs) trained on large sequence corpora are widely used for annotation and prediction tasks [Elnaggar et al., 2021, Lin et al., 2023, Brandes et al., 2022, Rao et al., 2019, Meier et al., 2021].They have achieved success in secondary-structure prediction, mutational effect modeling, and structure generation. However, despite their empirical performance, there remains limited understanding of what biological signals they actually capture. Proteins are strings over a 20-letter alphabet, and related proteins are grouped into *families* in Pfam based on conserved domains and alignments [Finn et al., 2016, Mistry et al., 2021]. Because many applications touch these families, it is important to clarify what aspects of family structure are reflected in pLM representations.

To separate overall amino-acid composition from residue order, we use a *composition-preserving shuffle* that keeps length and amino-acid counts fixed while randomizing order [Jiang et al., 2008], in the spirit of probing analyses used in NLP embeddings [Alain and Bengio, 2017, Hewitt and Manning, 2019]. We then compare true sequences and their shuffles across layers of several pLMs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences (FM4LS 2025).

Our evaluation covers six diverse Pfam families and four representative models (ESM-2, ProtBERT, ProtXLNet, ProteinBERT). For each family and layer we compute within-family (*In*) and to-negative-bank (*Out*) cosine similarities, summarize separation using a ratio we call *fidelity* (In/Out) and a difference $\Delta = \text{In} - \text{Out}$, and visualize the embeddings with t-SNE[van der Maaten and Hinton, 2008]. This provides a straightforward, composition-controlled setup for examining how family structure appears in current pLM embeddings without training task-specific classifiers.

2 Methods

2.1 Dataset

We use six Pfam families as positives: C2H2 zinc finger (PF00096, n=159); EF-hand (PF00036, n=560); protein kinase domain (PF00069, n=38); ubiquitin (PF00240, n=59); SH3 domain (PF00018, n=55); and class A GPCR (PF00001, n=63). These families were chosen to reflect functional and structural diversity: DNA binding (zinc finger), calcium signaling (EF-hand), catalysis (kinase), post-translational modification (ubiquitin), protein–protein interaction (SH3), and membrane signaling (GPCR). They also vary in sequence length and family size, from compact folds with tens of sequences (ubiquitin, kinase domain) to large repeat-containing or membrane-associated families with hundreds of sequences (EF-hand, GPCR). This diversity helps ensure that observed composition effects are not restricted to a single function, fold type, or dataset size. As a pooled negative bank we include non-target families (e.g., PF00005, PF00072, PF00076, PF00013) to provide a broad background and avoid cherry-picking. For each positive sequence we also create a composition-preserving *shuffle* by permuting residues while keeping length and the amino-acid histogram fixed (per-sequence shuffle), ensuring that residue order is fully randomized while composition is preserved.

2.2 Models and embeddings

We evaluate four representative pLMs that differ in scale, architecture, and training objective. *ESM-2* (33-layer encoder; masked language modeling) [Lin et al., 2023] is among the largest transformer encoders trained directly on protein corpora, and serves as a strong high-capacity baseline. *ProtBERT* (30-layer BERT-style encoder; masked language modeling) [Elnaggar et al., 2021, Devlin et al., 2019] adapts the BERT design from NLP, providing a widely used bidirectional context model for proteins. *ProtXLNet* (36-layer XLNet-style encoder; permutation language modeling with autoregressive factorization) [Elnaggar et al., 2021, Yang et al., 2019] differs in pretraining paradigm, using permutation objectives intended to capture longer-range dependencies. *ProteinBERT* (6-layer compact encoder with global tokens) [Brandes et al., 2022] is far smaller in depth and parameter count but includes global tokens that efficiently summarize the whole sequence.

These models were chosen to span both large-scale transformers trained on billions of residues (ESM-2, ProtBERT, ProtXLNet) and a compact architecture designed for efficiency (ProteinBERT). This coverage allows us to test whether composition-driven effects are consistent across different capacities and objectives rather than being tied to one model family. At each layer ℓ, token embeddings are mean-pooled across residues and then L2-normalized [Bepler and Berger, 2019, Ethayarajh, 2019]. Mean-pooling provides a simple, comparable sequence representation across variable lengths, and normalization ensures that cosine similarities are stable and comparable across models of different scales.

2.3 In/Out similarities and metrics

For family F and layer ℓ,

$$\text{In}(F,\ell) = \frac{1}{|P|} \sum_{(i,j)\in P} \cos\big(x_i^\ell, x_j^\ell\big), \qquad \text{Out}(F,\ell) = \frac{1}{|N|} \sum_{(i,k)\in N} \cos\big(x_i^\ell, x_k^\ell\big),$$

where P are pairs within F and N are pairs between F and the pooled negatives. We summarize separation with *fidelity* $= \text{In}/\text{Out}$ (ratio) and $\Delta = \text{In} - \text{Out}$ (difference). In addition to point estimates, we quantify uncertainty via nonparametric sequence-level bootstrapping: for each family we resample sequences with replacement and recompute metrics to obtain 95% confidence intervals. This provides a statistically grounded view of model behavior that avoids over-interpreting single point estimates.

2.4 Evaluation design

We systematically analyze every combination of model, family, and layer for both true and shuffled sequences. This exhaustive design ensures that the reported patterns are not artifacts of a particular model depth or family size. By computing metrics layer by layer, we capture the dynamics of how representations evolve across the network, which is critical for architectures such as BERT, XLNet, and ESM-2 where depth plays a central role in shaping the learned features. Both fidelity and Δ are evaluated under identical conditions, allowing direct comparisons across models and controls. This uniform setup provides a consistent basis for assessing whether composition-preserving shuffles diverge from true sequences in a statistically meaningful way. The main text reports multi-panel figures for Δ and fidelity and one representative t-SNE strip; the Appendix contains full t-SNE panels (all families $\times$ models) and per-family In vs. Out curves.

3 Results

Δ separation (moved to Appendix): Detailed $\Delta = \text{In} - \text{Out}$ panels for all families and models are provided in Appendix A.2. The qualitative pattern closely parallels both the fidelity ratios and the raw In/Out similarities described below. Across families and depths, shuffled curves often match or exceed their true counterparts.

Fidelity ratios: Figure 1 reports fidelity (In/Out) across families and models. **ESM-2** shows a gradual increase across depth but never departs far from unity, and the shuffled controls shadow the true curve closely, occasionally surpassing it at later layers. **ProtBERT** exhibits a distinct spike around layer $\sim$10; however, the shuffled sequences reproduce the same behavior, indicating that the enrichment arises largely from composition rather than sequence order. **ProtXLNet** produces steadily rising fidelities with depth, but the shuffled curves remain comparable and in many cases higher than the true values—there is no robust true>shuffled advantage. **ProteinBERT** peaks early before declining with depth; in this case, true and shuffled sequences remain close throughout, with only small family-specific gaps. Taken together, these comparisons highlight that the fidelity metric provides no consistent evidence of true sequences outperforming their shuffled counterparts.

Uncertainty quantification: Because pairwise similarities are not independent, we assessed robustness for the fidelity metric using sequence-level bootstrap resampling. Figure 1 shows shaded 95% confidence intervals around the fidelity curves. Across all models and families these intervals were narrow and closely followed the point estimates, confirming that the qualitative trends—shuffled curves shadowing or exceeding true curves—are statistically stable rather than artifacts of sampling. .

Raw In vs Out (PF00096). Figure 2 shows the underlying within-family (In) and to-negatives (Out) cosine similarities for PF00096. This view complements the fidelity ratio by showing the raw components separately. Across layers, the shuffled control follows the true In and Out curves remarkably closely, with only minor deviations, which is, consistent with the fidelity trends. The depth-dependent trends in PF00096 are not unique to this family; additional examples for the other five Pfams are provided in Appendix A.1, where the same pattern of close correspondence between true and shuffled sequences can be observed.

t-SNE visualization: Figure 3 illustrates a geometric view of the embeddings using PF00096 with **ESM-2** as an example. From approximately layer 10 onward, both true (red) and shuffled (pink) sequences form compact clusters in the two-dimensional projection, while the pooled negatives (gray) remain well separated. The modest displacement between the true and shuffled clusters is small compared to the tightness of each cluster itself, underscoring the difficulty of distinguishing them based on order information. This visualization therefore echoes the message of the fidelity and In/Out analyses: clustering behavior attributed to family structure can be largely reproduced by shuffled controls. Full t-SNE panels for all families and models are included in Appendix A.3, which confirm that this observation generalizes across architectures and sequence classes.

Cross-family and architectural trends. When comparing across the six Pfam families—spanning DNA-binding, signaling, catalysis, modification, interaction, and membrane receptor functions—the same qualitative trend emerges. In nearly every setting, shuffled curves track the true curves closely and frequently exceed them. Architectural differences primarily affect the depth

at which separation peaks appear: **ProtBERT** displays a sharp mid-layer spike that is mirrored by the shuffled sequences, **ProtXLNet** exhibits a gradual rise in later layers where the shuffled control often matches or surpasses the true sequences, and **ProteinBERT** peaks early before declining. **ESM-2**, by contrast, shows relatively weak overall separation. Despite these architectural idiosyncrasies, none of the models demonstrate a consistent true>shuffled advantage across families. This repeated pattern suggests that the composition effect is robust to both family choice and model design, reinforcing the central observation that composition-preserving shuffles closely match the true families.

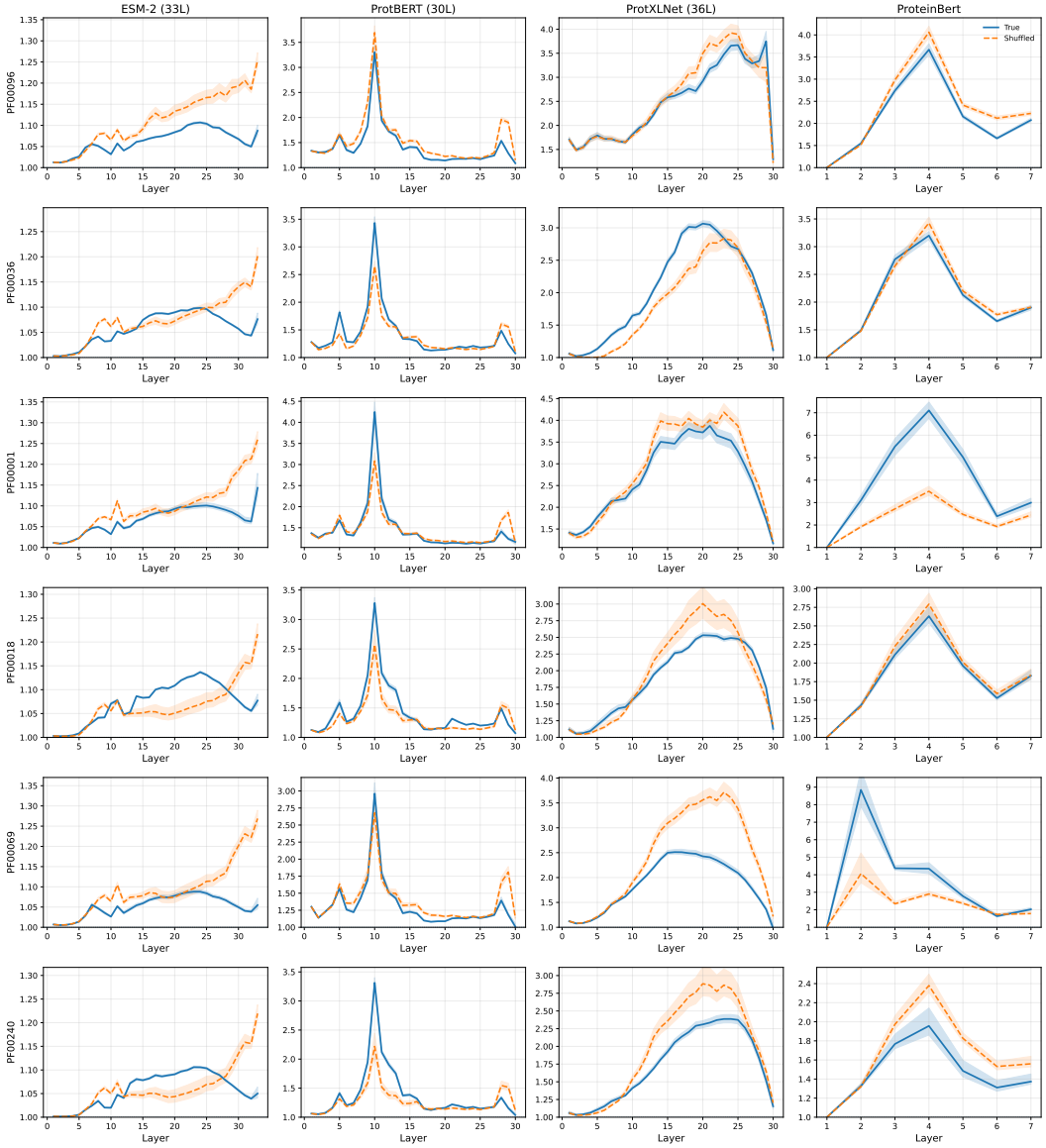


Figure 1: Fidelity (In/Out) across six Pfam families and four models. Solid = true, dashed = shuffled. Shaded regions = 95% bootstrap confidence intervals.

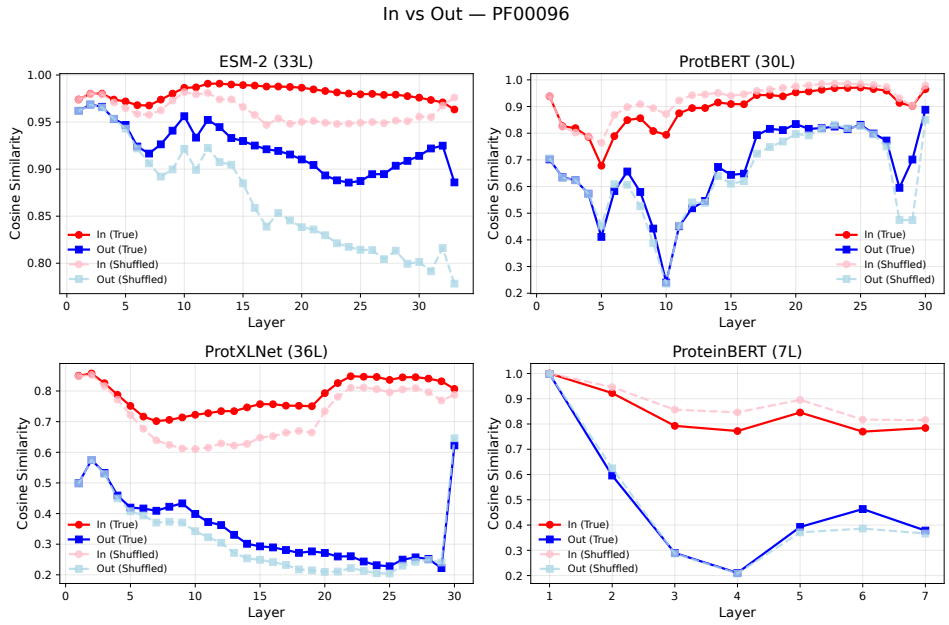


Figure 2: **Raw within- vs between-family cosine similarities (PF00096).** Red circles = In (True), blue squares = Out (True), pink dashed circles = In (Shuffled), light-blue dashed squares = Out (Shuffled). Composition-preserving shuffles closely track the true family across layers.

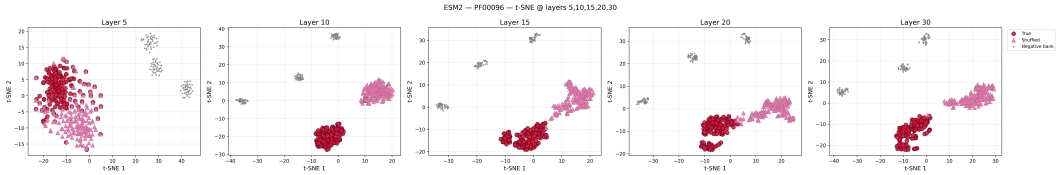


Figure 3: t-SNE for PF00096 with ESM-2 at selected layers (5, 10, 15, 20, 30). True = red circles, Shuffled = pink triangles, Negative bank = gray dots.

4 Conclusion

Protein language models often appear to separate protein families, but our analysis shows that much of this separability arises from amino-acid composition rather than conserved order-dependent motifs. Across six families and four models, composition-preserving shuffles closely track—and often exceed—the fidelity and Δ values of true sequences. Architecture influences the layer depth and scale of these effects (ProtBERT's mid-layer spike, ProtXLNet's gradual rise, ProteinBERT's early peak), but none of the models shows a consistent true $>$ shuffled advantage. These findings caution against interpreting family-level clustering as direct evidence of motif learning. They also raise an open question: is it desirable that pLMs cluster together sequences that are biologically irrelevant but compositionally similar? Addressing this will be key for developing composition-controlled baselines and biologically meaningful benchmarks in future protein language models.

5 Future Work

Our analysis shows that composition accounts for much of the apparent family structure in current protein language models. A natural next step is to design training objectives that make models less sensitive to simple composition statistics and more attentive to order-dependent motifs. One possible direction comes from recent work in nucleotide modeling Refahi et al. [2025],where they proposed a transition-matrix regularization loss (CARMANIA) to discourage models from relying only on base composition. Exploring similar ideas for protein sequences could reveal whether such regularization helps recover deeper functional signals beyond composition.

Code Availability

Code and data of the experiments are available here : `https://github.com/Sohel016/PLM_Fidelity`.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. In *International Conference on Learning Representations (ICLR)*, 2019.

Niv Brandes, Dan Ofer, Yuval Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: A universal deep-learning model of protein sequence and function. *PLoS Computational Biology*, 18(6): e1009839, 2022. doi: 10.1371/journal.pcbi.1009839.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, 2019. arXiv:1810.04805.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, et al. Prottrans: Towards understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. Also available as arXiv:2007.06225.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of ACL*, pages 55–65, 2019. doi: 10.18653/v1/P19-1006.

Robert D Finn, Penelope Coggill, et al. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016. doi: 10.1093/nar/gkv1344.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *NAACL*, 2019.

Minghui Jiang, James Anderson, John Gillespie, and Martin Mayne. ushuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9:192, 2008. doi: 10.1186/1471-2105-9-192.

Zeming Lin, Handong Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Joshua Meier, Roshan Rao, Robert Verkuil, et al. Language models enable zero-shot prediction of the effects of mutations on proteins. *bioRxiv*, page 2021.07.09.450648, 2021. doi: 10.1101/2021.07.09.450648.

Jaina Mistry, Sara Chuguransky, et al. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, et al. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems (NeurIPS) – Datasets and Benchmarks Track*, 2019. arXiv:1906.08230.

Mohammadsaleh Refahi, Mahdi Abavisani, Bahrad A Sokhansanj, James R Brown, and Gail Rosen. Context-aware regularization with markovian integration for attention-based nucleotide analysis. *arXiv preprint arXiv:2507.09378*, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Zhilin Yang, Zihang Dai, et al. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1906.08237.

A Additional Figures

A.1 Raw In vs Out (All Families)

The following panels report raw within-family (**In**) and to-negatives (**Out**) cosine similarities across layers for each positive family and model. Curves for shuffled controls are overlaid (dashed).

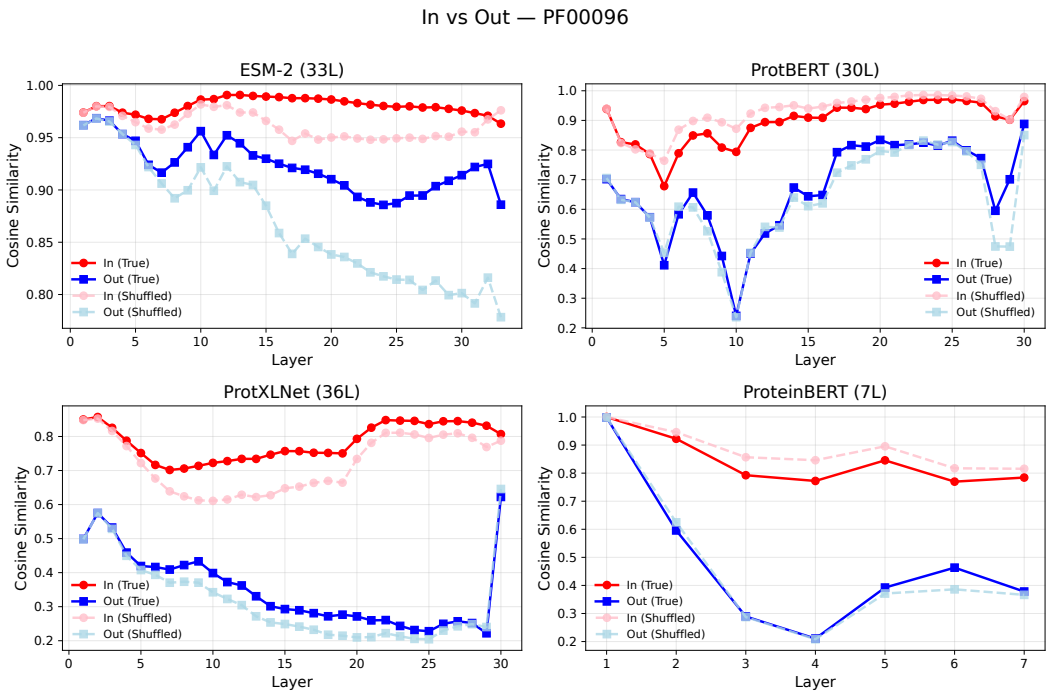


Figure 4: In vs Out for PF00096 (C2H2 zinc finger). Red circles = In (True), blue squares = Out (True), pink dashed circles = In (Shuffled), light-blue dashed squares = Out (Shuffled).

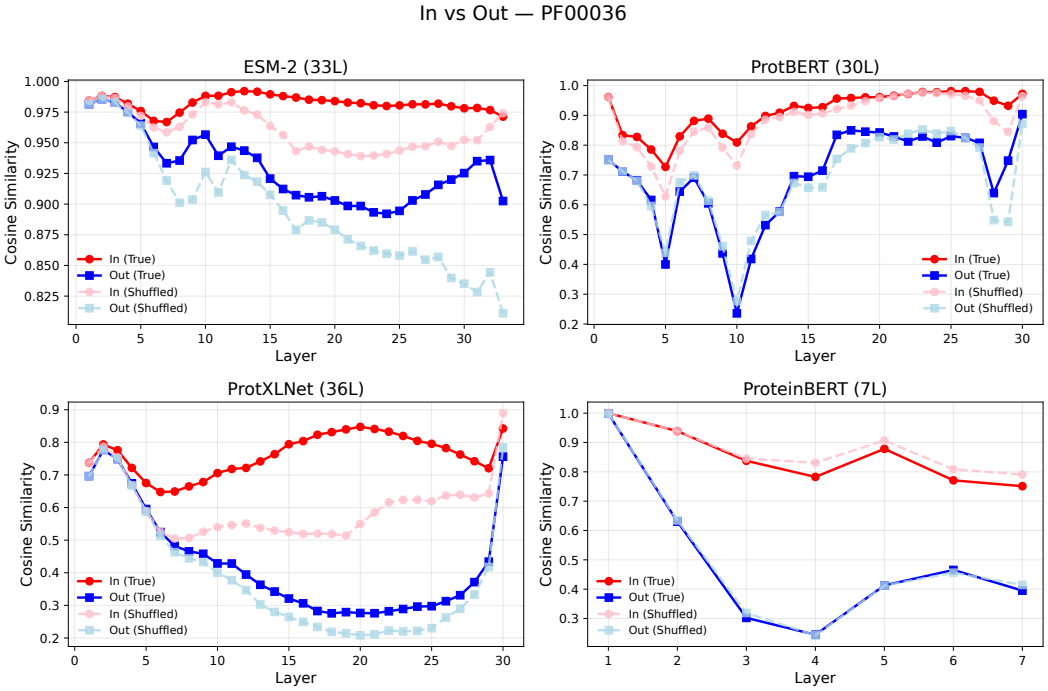


Figure 5: In vs Out for PF00036 (EF-hand). See Fig. 4 for legend mapping.

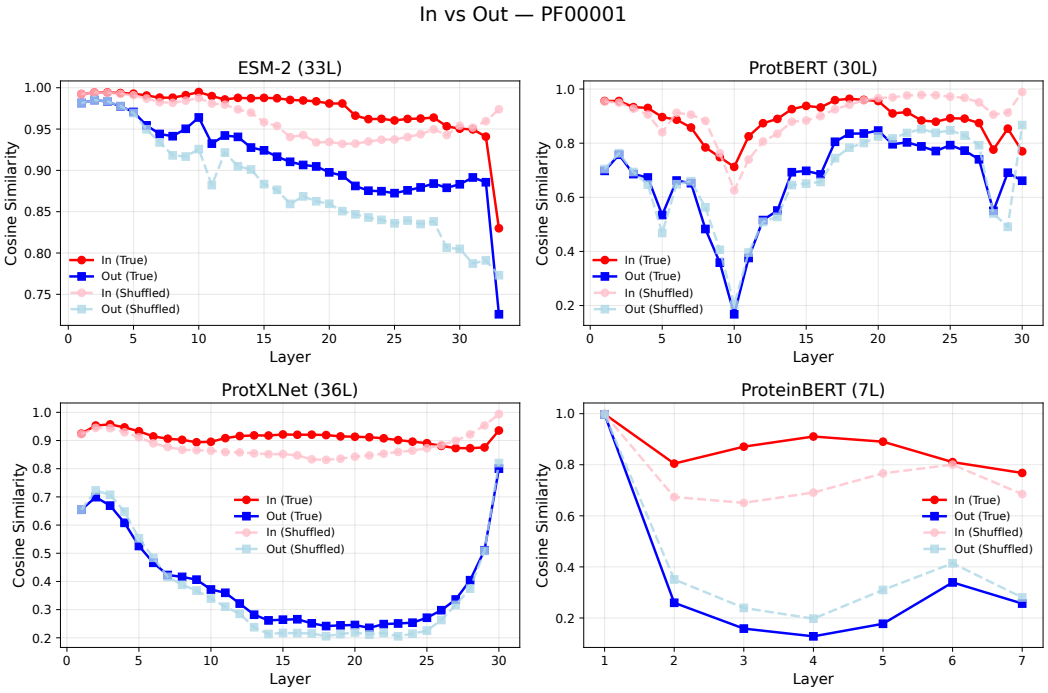


Figure 6: In vs Out for PF00001 (Class A GPCR).

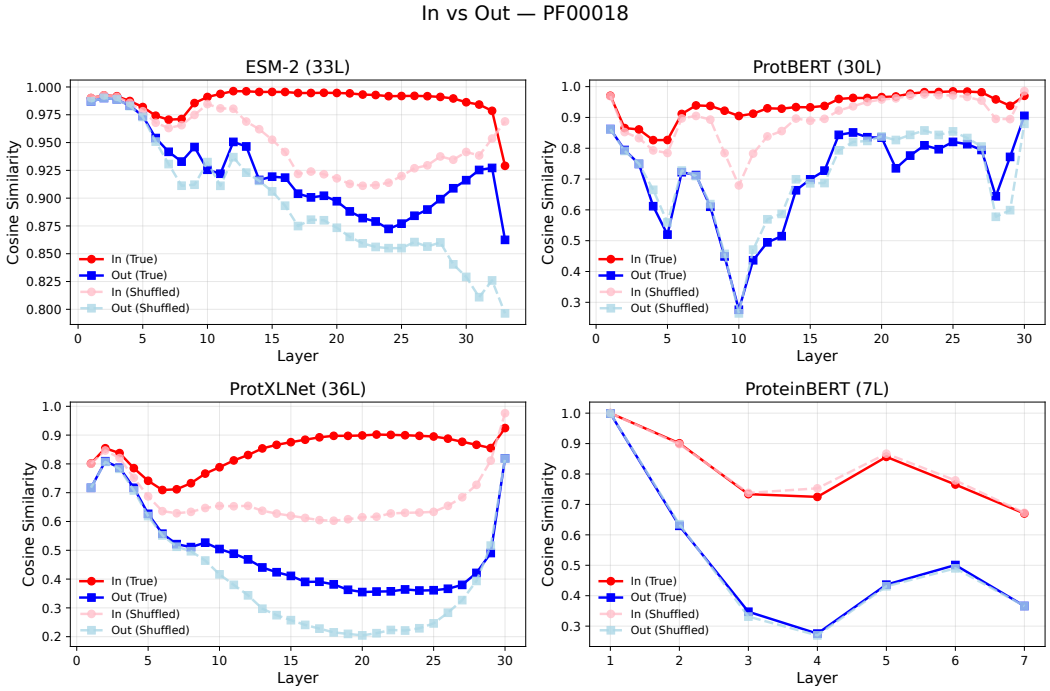


Figure 7: In vs Out for PF00018 (SH3 domain).

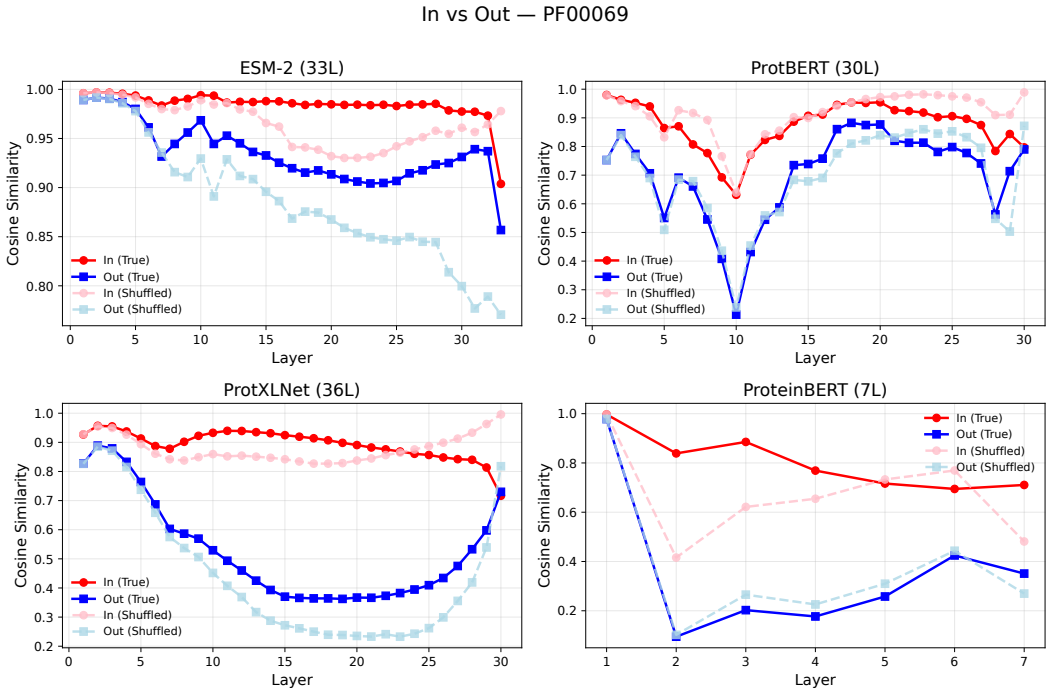


Figure 8: In vs Out for PF00069 (Protein kinase domain).

A.2 Δ panels (All Families)

Figures below report $\Delta = \text{In} - \text{Out}$ across all six families and four models.

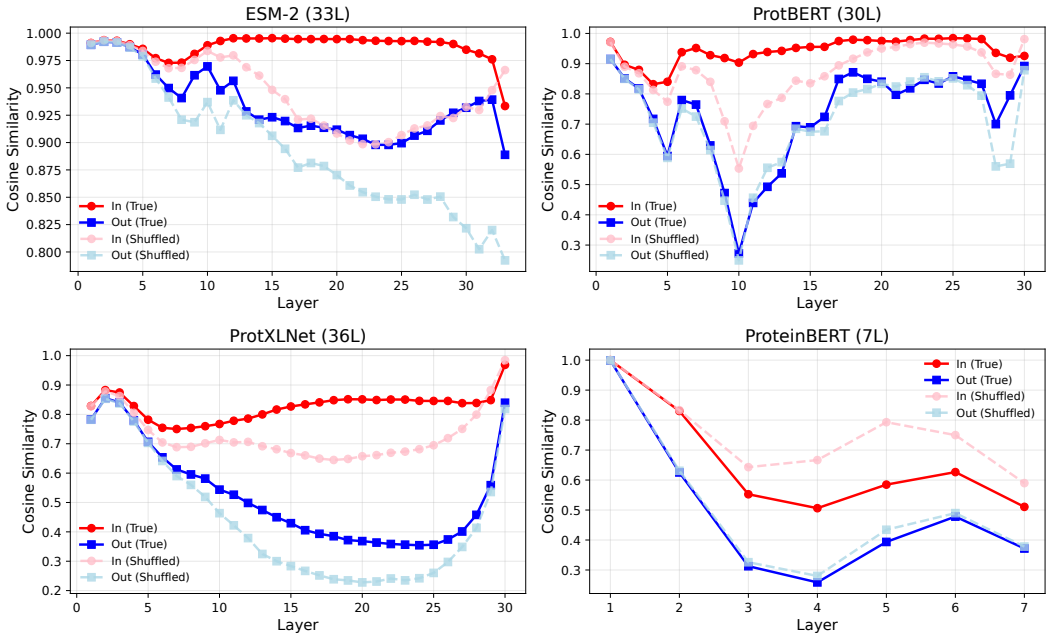
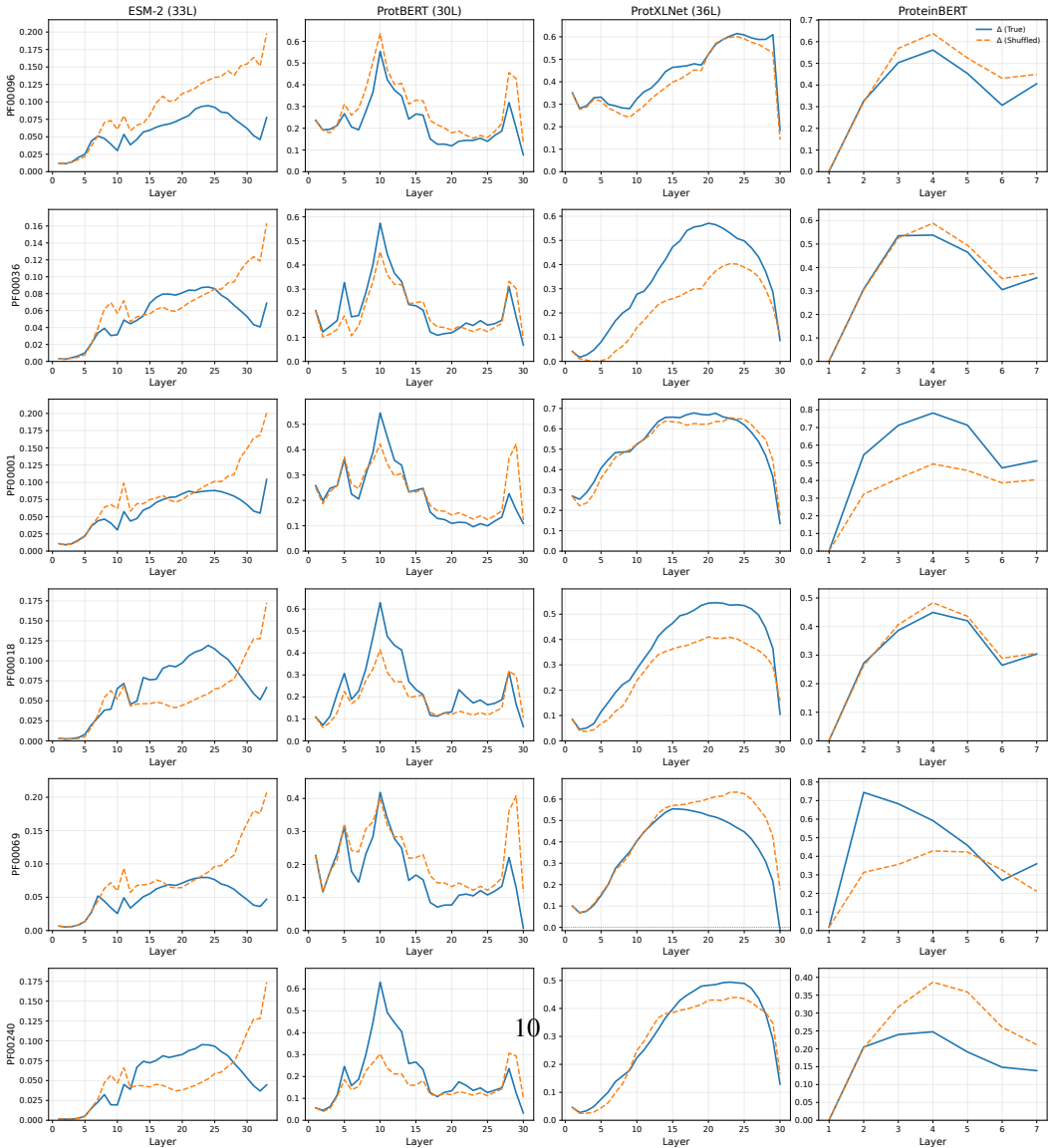


Figure 9: In vs Out for PF00240 (Ubiquitin).



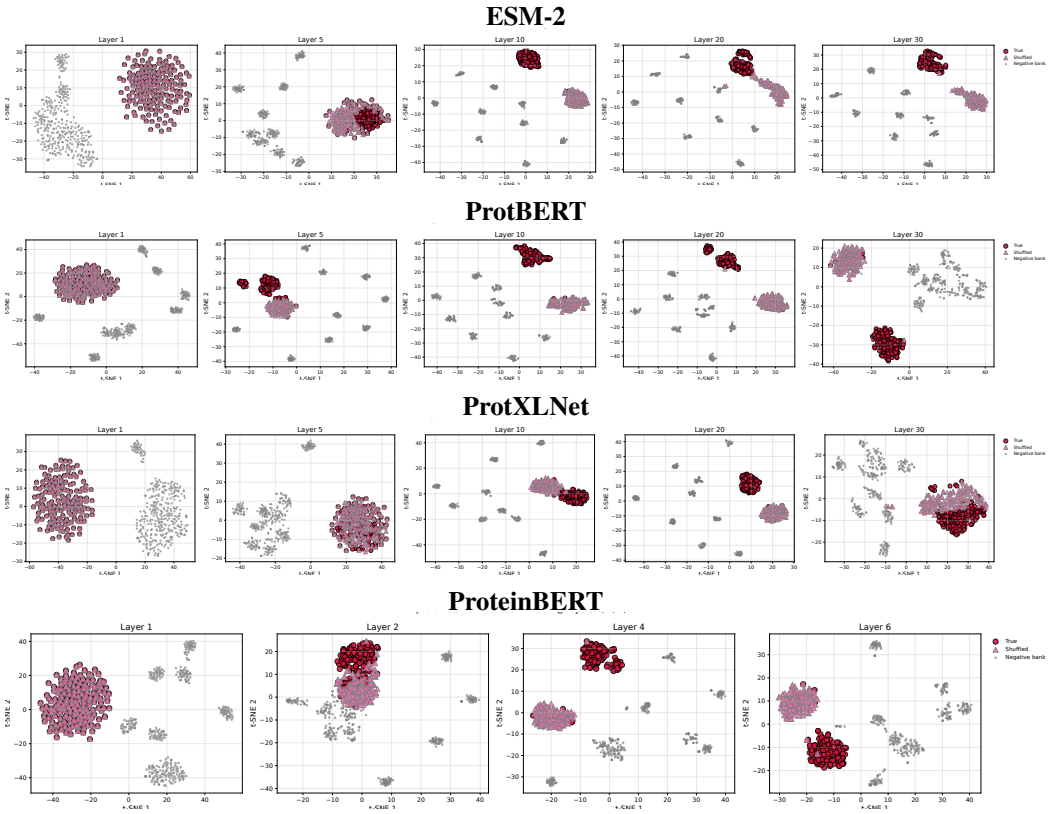


Figure 11: t-SNE for **C2H2 zinc finger** (**PF00096**) across models. Each strip shows selected layers left-to-right. True = red circles, Shuffled = pink triangles, Negatives = gray dots.

A.3 t--SNE panels grouped by family

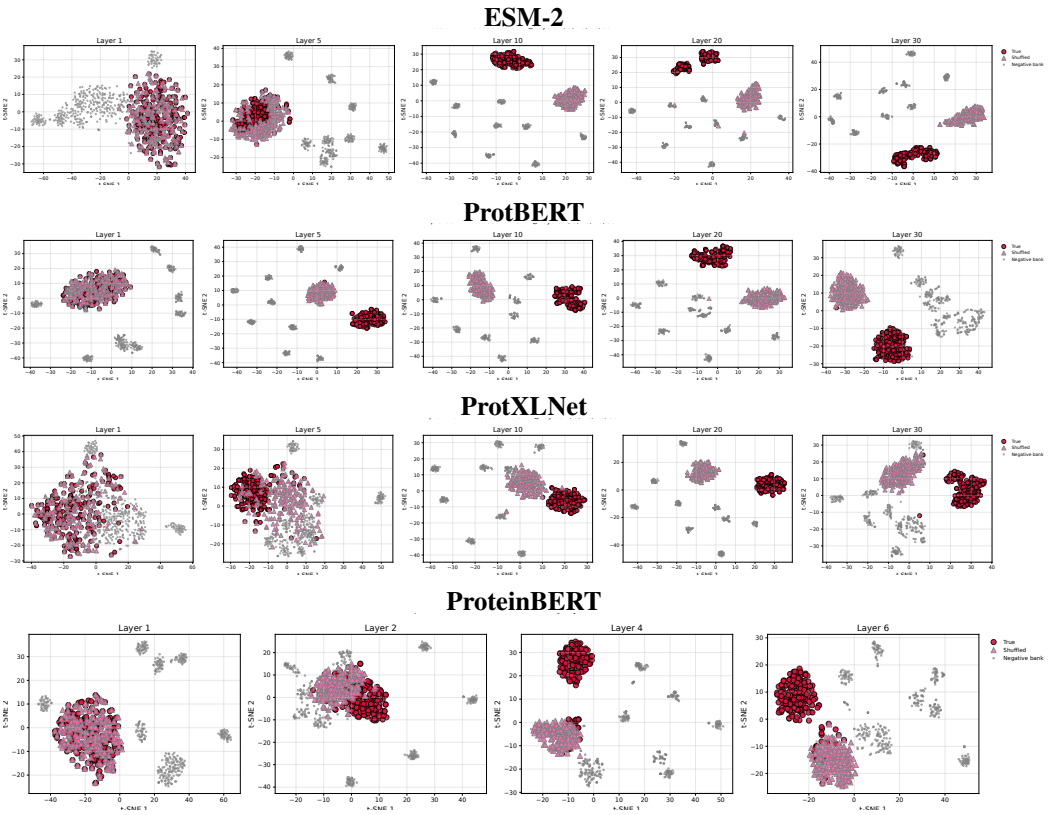


Figure 12: t-SNE for **EF-hand** (**PF00036**) across models. Each strip shows selected layers left-to-right. True = red circles, Shuffled = pink triangles, Negatives = gray dots.

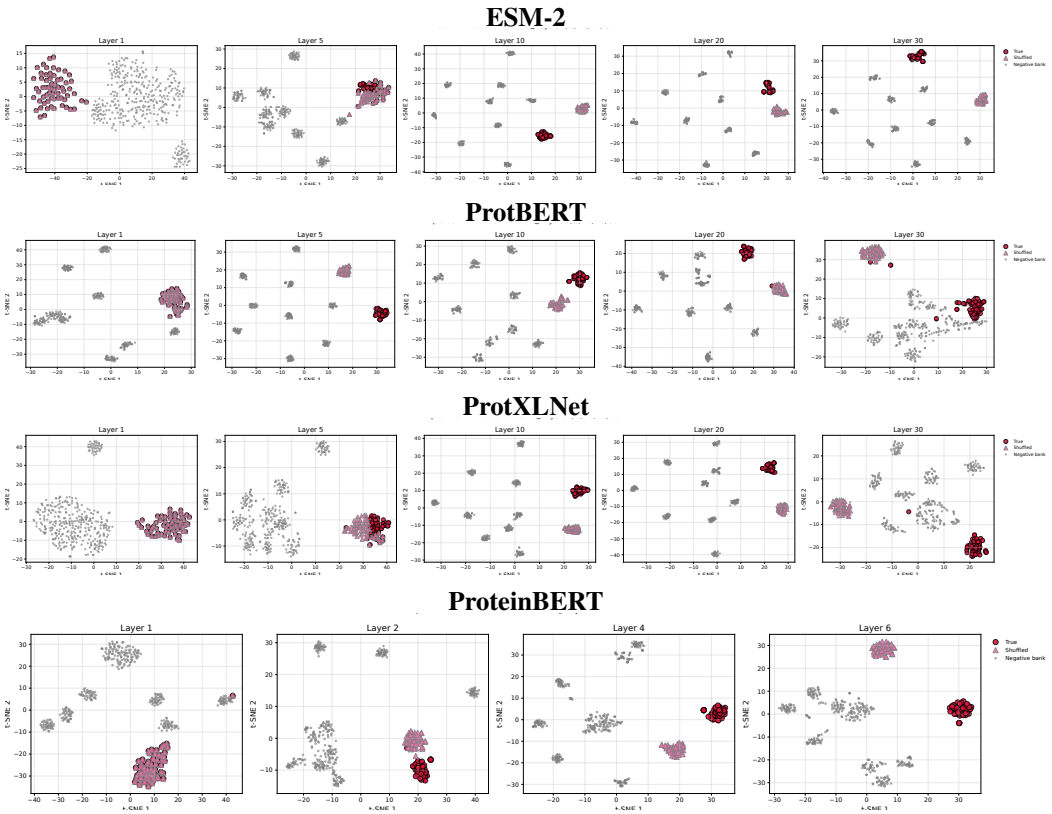


Figure 13: t-SNE for **Class A GPCR** (**PF00001**) across models. Each strip shows selected layers left-to-right. True = red circles, Shuffled = pink triangles, Negatives = gray dots.

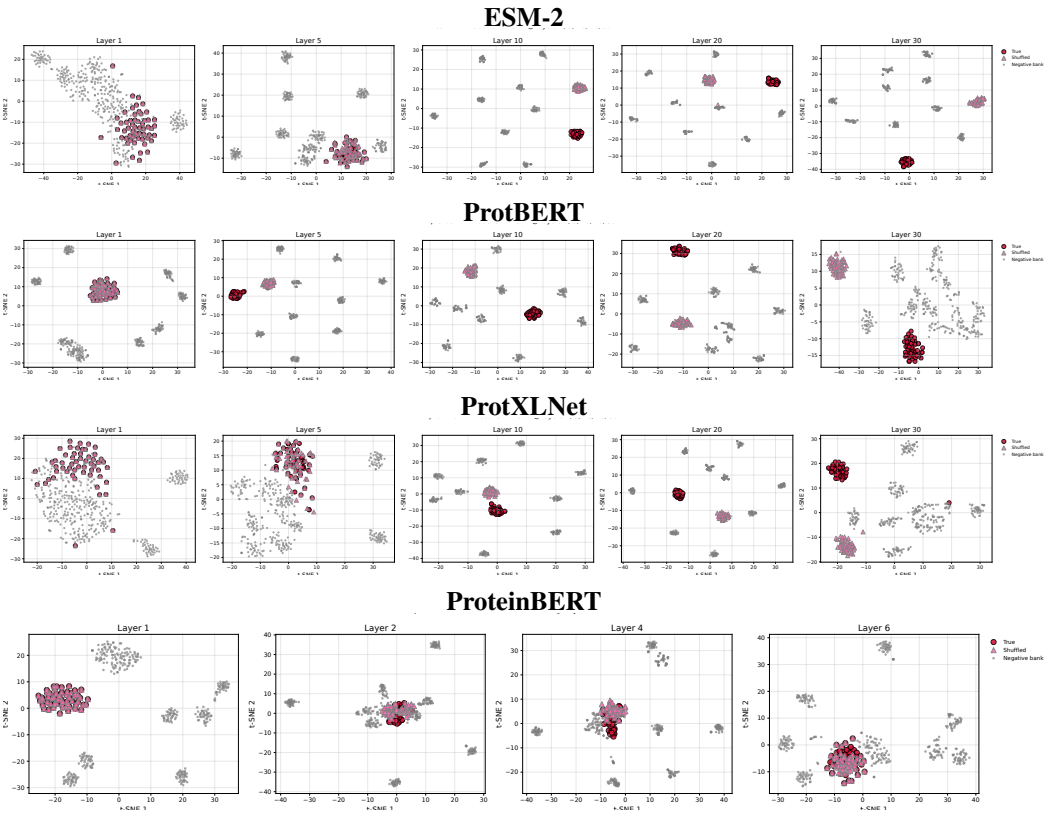


Figure 14: t-SNE for **SH3 domain** (**PF00018**) across models. Each strip shows selected layers left-to-right. True = red circles, Shuffled = pink triangles, Negatives = gray dots.

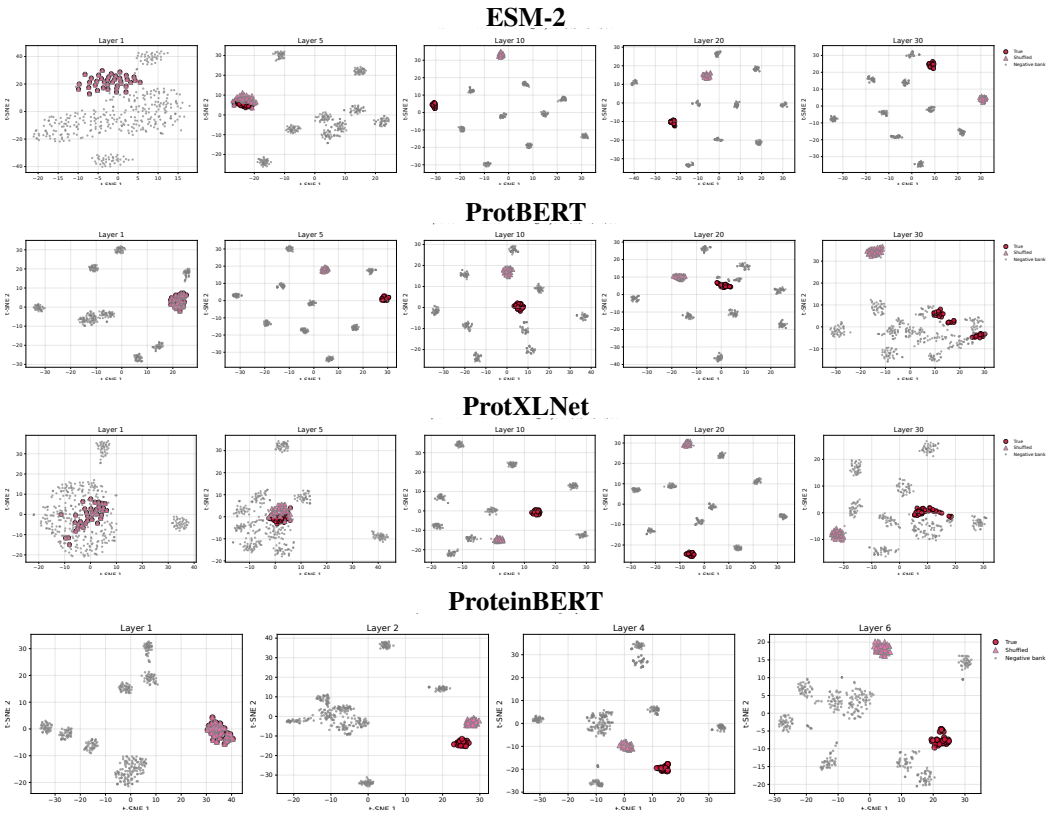


Figure 15: t-SNE for **Protein kinase** (**PF00069**) across models. Each strip shows selected layers left-to-right. True = red circles, Shuffled = pink triangles, Negatives = gray dots.

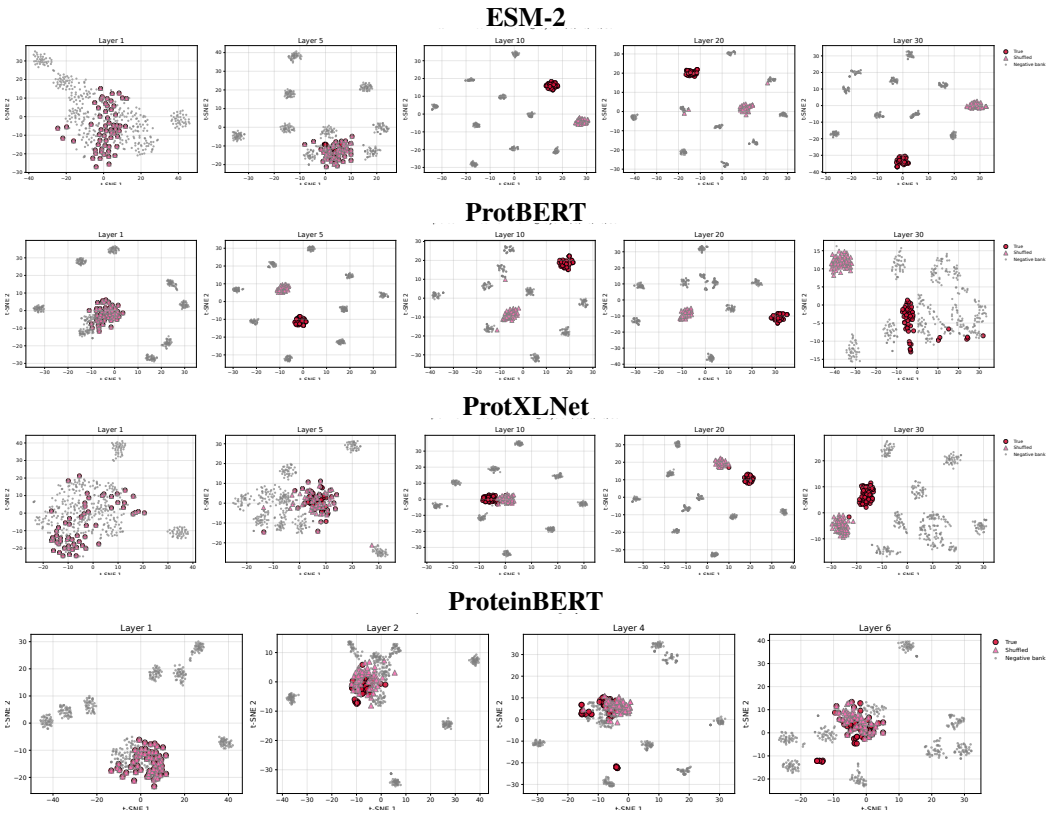


Figure 16: t-SNE for **Ubiquitin** (**PF00240**) across models. Each strip shows selected layers left-to-right. True = red circles, Shuffled = pink triangles, Negatives = gray dots.