Which pLM to choose? A comparative evalutaion

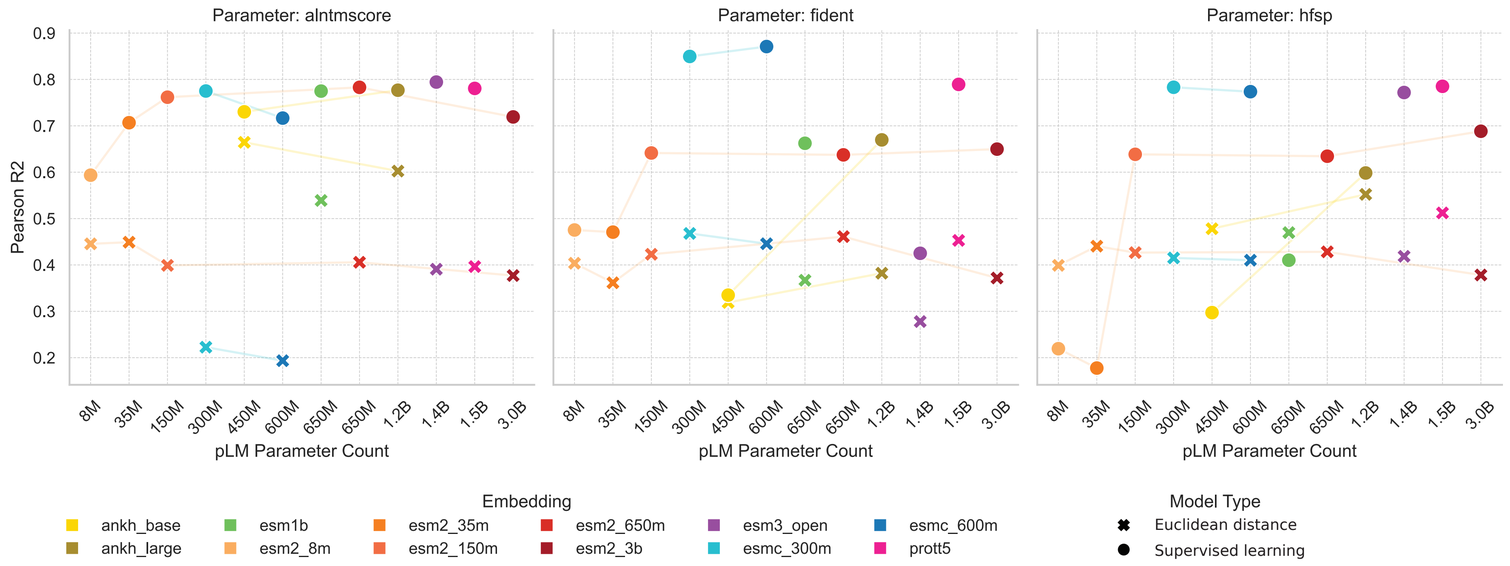
Tobias Senoner[1]*, Ivan Koludarov[1]*, Joshua Günther[1], Amarda Shehu[2], Burkhard Rost[1,3,4]@, Yana Bromberg[4,5,6]@

[1]I12 Chair of Bioinformatics, Technical University of Munich
[2]Department of Computer Science, George Mason University
[3]School of Life Sciences Weihnstephan, Technical University of Munich
[4]Institute for Advanced Study (TUM-IAS), Garching/Munich
[5]Department of Computer Science, Emory University
[6]Department of Biology, Emory University

Background

Protein language models (pLMs) create conflicting maps of sequence space—like early cartographers charting unknown territories. Current models range from 8M to 15B parameters (ProtT5 [1], ESM2 [2], Ankh [3], ESM-C [4]), yet the relationship between size and biological representation quality contradicts NLP scaling laws [5]. We distinguish between raw embedding information (accessible through Euclidean distance) and extractable information (supervised learning).

Model size does not correlate with performance

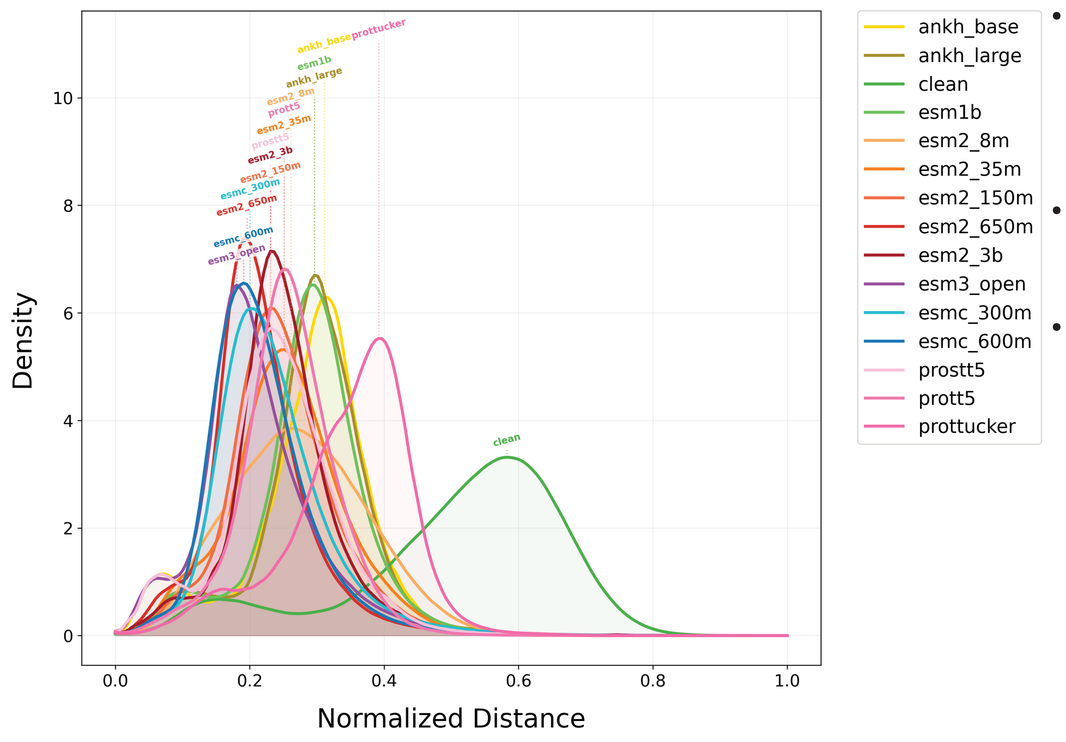


- Smaller models (≤150M) match or outperform billion-parameter variants for raw embeddings
- Supervised learning reverses this trend: larger models tend to unlock richer latent information
- ESM-C [4] models, while small, show remarkable improvements over similar sized older models

Methodology

1. Benchmarked 15 pLMs (8M-3B parameters) on < 2.5M protein pairs from SwissProt-pre2024 (542K proteins) [6]
2. Evaluated three biological similarity metrics: sequence identity (MMseqs2 [7]), structural similarity (Foldseek [8], TM-scores), and functional similarity (HFSP scores [9])
3. Compared Euclidean distances versus supervised learning
4. Analyzed embedding space topology and cross-model correlations

Bimodal distribution reflects learned similarities



- Bimodal distributions expose natural protein organization: functionally similar clusters (left peak) vs divergent sequences (right peak)
- Contrastive learning amplify separation: CLEAN [10] doubles normalized distances
- Near unimodal patterns in ESM-C [4] and ProtT5 [1] suggest fundamentally different organizational principles

Conclusion

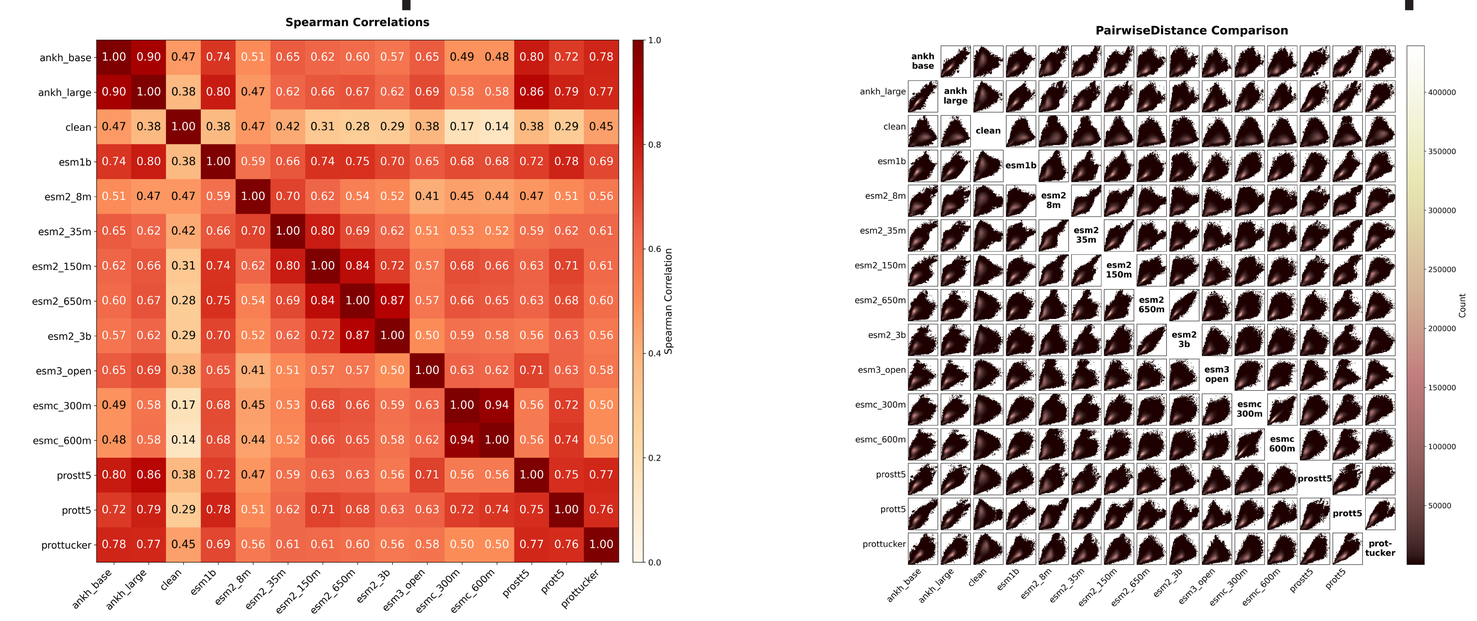
Need immediate insights? Choose foundation models ≤150M parameters (optimal cost-benefit ratio)

Have GPUs and labeled data? Use 1-3B models for supervised learning

Specific task only? Contrastive learn specialist, but accept limited versatility

Future Directions: Progress requires smarter training objectives and better data curation, not merely more parameters.

Pairwise comparison shows different learned space



- Model families mostly maintain signature correlations (ρ>0.75) regardless of size: architecture defines embedding topology
- Surprising: Ankh-large [3] and ProstT5 [1] achieve ρ>0.80 despite different frameworks
- Contrastive training breaks family bonds: CLEAN [10] - ESM1b [11] correlation drops to ρ=0.38

[1] Elnaggar, Ahmed, et al. "Prottrans: Toward understanding the language of life through self-supervised learning." IEEE transactions on pattern analysis and machine intelligence 44.10 (2021): 7112-7127.
[2] Lin, Zeming, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379.6637 (2023): 1123-1130.
[3] Elnaggar, Ahmed, et al. "Ankh: Optimized protein language model unlocks general-purpose modelling." arXiv preprint arXiv:2301.06568 (2023).
[4] ESM Team. (2024). ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. Evolutionary-Scale Website.
[5] Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).
[6] "UniProt: the Universal protein knowledgebase in 2025." Nucleic Acids Research 53, no. D1 (2025): D609-D617.
[7] Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology, 35(11), 1026-1028.
[8] Van Kempen, Michel, et al. "Fast and accurate protein structure search with Foldseek." Nature biotechnology 42.2 (2024): 243-246.
[9] Mahlich, Yannick, et al. "HFSP: high speed homology-driven function annotation of proteins." Bioinformatics 34.13 (2018): i304-i312.
[10] Yu, Tianhao, et al. "Enzyme function prediction using contrastive learning." Science 379.6639 (2023): 1358-1363.
[11] Rives, Alexander, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." Proceedings of the National Academy of Sciences 118.15 (2021): e2016239118.