Trends in Genetics

CellPress

Forum

Rethinking molecular evolution through protein language model embeddings

Rosa Fernández [1,*], Sergi Valverde [1], Aureliano Bombarely [2], Ildefonso Cases [3], Scott A. Handley [4], and Ana M. Rojas [3,*]

Protein language models compress protein sequences into high-dimensional embeddings that capture biochemical, structural, and functional constraints without explicit supervision. We highlight that these embeddings encode rich evolutionary information, enabling new geometry-based views of homology, divergence, and convergence, and calling for a synthesis between classical molecular evolution and systematic evolutionary embedding analysis.

A change of paradigm: The unsupervised evolutionist

A core premise of molecular evolution is that homology, as a hypothesis of shared ancestry, is tested and inferred by comparing sequences through pairwise and multiple alignments. From Dayhoff's mutation matrices to modern likelihood and Bayesian phylogenetics, we have relied on explicit models of substitution and theoretically grounded summary statistics to relate sequence divergence to time, structure, and function. This framework has been remarkably successful, yet it struggles in familiar regimes such as the 'twilight zone' of low sequence identity and the deluge of functionally uncharacterized proteins from genome and metagenome sequencing.

Protein language models (pLMs) have emerged as an alternative way of encoding protein sequences. These models are typically trained on hundreds of millions of natural proteins and summarize each amino acid sequence into a high-dimensional numerical vector (an embedding), usually with hundreds to a few thousand dimensions. These embeddings capture biochemical, structural, functional, and evolutionary constraints but can also reflect higher-order interactions, a limitation for traditional tools. As a result, tasks that were historically framed as homology searches or family classification can now be approached as geometric and topological questions on a landscape of protein representations (Figure 1A). This landscape reflects constraints of structure, function, and evolution, all of which are learned from sequence data. That is, language models reveal systemic properties of protein evolution, not just sequence patterns. Crucially, however, what pLMs learn from natural sequences are the consequences of evolution (i.e., the biochemical, structural, and functional constraints that natural selection and drift have left imprinted on sequence space), rather than the evolutionary process itself; they have no mechanism for tracking the historical sequence of mutations, divergence times, or lineage relationships that produced current proteins.

It is now broadly accepted that pLM embeddings primarily reflect structural and functional similarity rather than evolutionary relatedness *per se*, but the degree to which they faithfully capture these properties, and how this varies across protein families and divergence levels, is model-dependent and remains an active area of investigation. This distinction between modeling evolutionary history and reasoning about its consequences (i.e., function) is precisely why we need a systematic interrogation of embedding spaces, as we advocate here. Proteins that share a fold or function but have diverged beyond the reach of standard alignment methods may lie close together in embedding space, whereas closely related sequences can occupy distinct regions that reflect functional specialization. This shifts the focus of molecular evolution in this context to the study of how lineages trace trajectories through this learned landscape, raising new questions about how to interpret and rigorously validate these spaces against the classical models based on sequence similarity and homology inference that have previously supported the field.

In this forum article, we argue that embeddings invite us to rethink what molecular evolution studies in practice, beyond sequence constraints. Instead of focusing solely on multiple sequence alignments (MSAs), we can ask how evolutionary processes sculpt trajectories through embedding space, how gene duplication and neofunctionalization appear in this representation, and where classical concepts hold or fail under this new lens (see Box 1). For instance, Shaw *et al.* [1] used pLM embeddings to test the ortholog conjecture and found no clear difference in embedding similarity between one-to-one orthologs and paralogs at equivalent sequence divergence, suggesting that embedding space does not straightforwardly distinguish functional conservation from functional divergence.

Embeddings can also partly recapitulate phylogenetic relationships within protein families but show systematic deviations in indel-rich histories. Notably, this is also the regime where alignment-based methods are most error-prone, leaving open the question of which approach better captures the true evolutionary signal [2]. In parallel, pLMs can be sensitive to convergent evolution, pulling independently evolved but functionally similar sequences together in embedding space and revealing a complex sequence basis for convergence beyond what pairwise identity alone can show [3]. Tools such

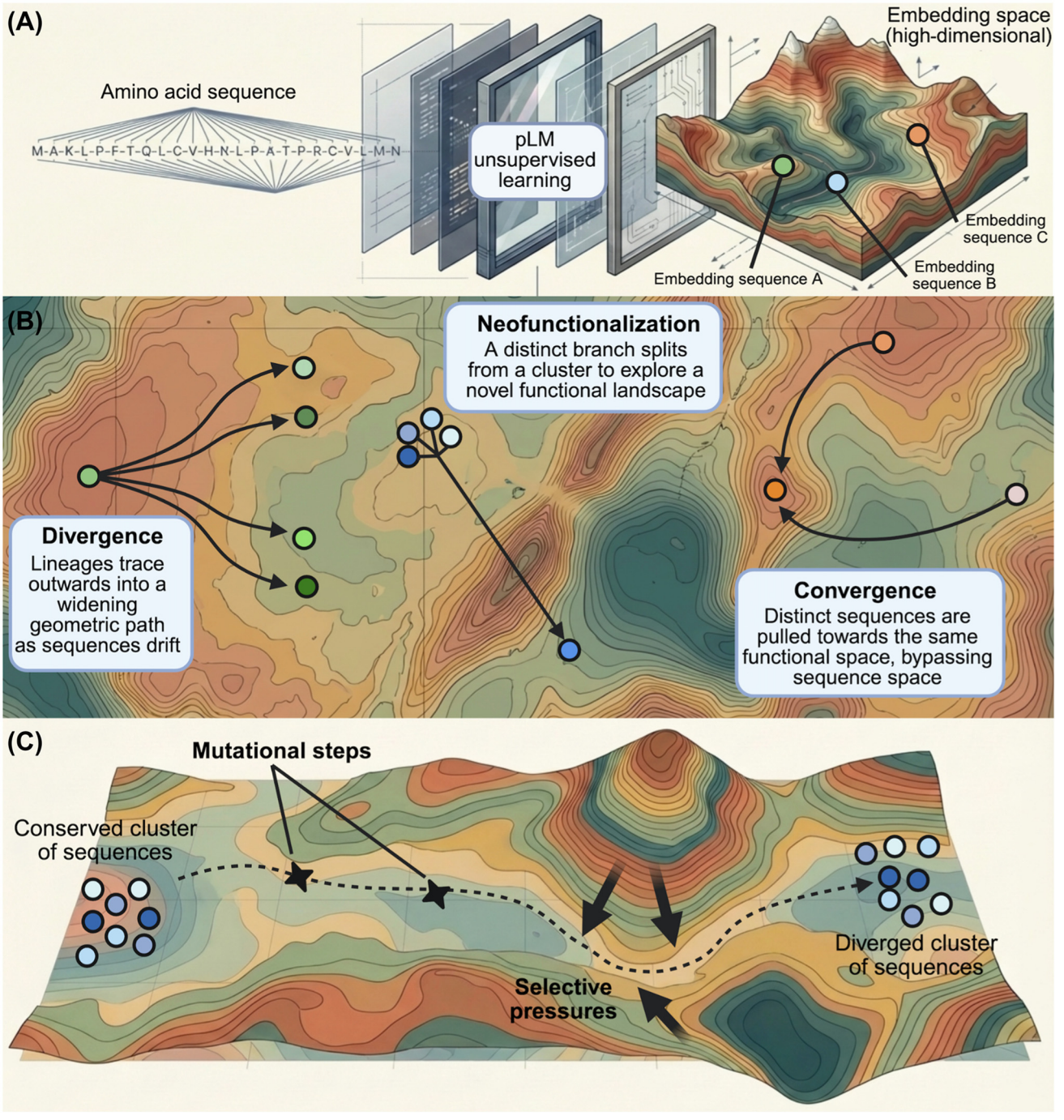

Figure 1. Conceptual view of protein language model (pLM) embedding spaces in molecular evolution. (A) An amino acid sequence is mapped by a pLM into a high-dimensional embedding space, depicted as a rugged landscape in which individual sequences occupy distinct positions. (B) Schematic trajectories in embedding space illustrate how divergence, neofunctionalization, and convergence can appear as lineages drift apart, explore novel regions, or independently approach similar functional neighborhoods despite distant ancestry. (C) A conserved cluster of sequences moves through embedding space via mutational steps under selective pressures, settling into a derived cluster that reflects both historical contingency and current adaptive constraints. This figure was created using BioRender (https://BioRender.com/ss7y7r3).

Box 1. A new research agenda for systematic evolutionary embedding analysis

Protein language models are already useful for annotation, homology search, and structure/function prediction, but their implications for other aspects of molecular evolution, such as genome evolution and phylogenomics, remain largely unexplored. Recognizing that embeddings primarily capture the consequences of evolution rather than its underlying process, we propose several key open questions:

1. Can species-level evolutionary signal be extracted from embeddings?—When we compare embeddings across species, do the resulting distances reflect shared ancestry or mainly structural and functional similarity within protein families, and under what conditions can we meaningfully extract species-level evolutionary signals from them?

2. How do embeddings relate to orthology and paralogy?—Can embedding distances refine or overturn orthology calls in large, domain-shuffled families, and when do they better predict functional conservation than homology or synteny?

3. What does gene repertoire evolution look like in embedding space?—How do processes such as gene duplication, loss, horizontal transfer, and domain rearrangement appear as trajectories or topological features, and can we distinguish these processes geometrically?

4. Can whole-proteome embeddings act as comparative traits?—Do the summaries of a species' embedding distribution correlate with ecology, life history, genome architecture, or developmental complexity, and can they be used for phylogenomic inference?

5. Where do classical models and embeddings disagree, and why?—In which regions of sequence space do explicit evolutionary models and embedding-based inferences give conflicting answers about homology, constraint, or convergence, and what can we learn from those disagreements?

6. Can embeddings capture evolutionary processes such as speciation, diversification, or adaptation?—Do they contain recognizable signatures of lineage splitting, changing diversification dynamics, and adaptive shifts in protein constraints, and how can these be robustly validated against independent phylogenetic and population-genetic benchmarks?

7. How general are evolutionary patterns across different pLMs?—To what extent do evolutionary signals detected in the embedding space of one model architecture, training objective, layer, or dataset generalize to other models, and what does model-to-model variation reveal about which aspects of protein evolution are robustly encoded versus contingent on training choices?

as FANTASIA further demonstrate that embedding-based similarity can be exploited genome-wide to annotate the 'dark proteome' across the animal tree of life, offering a practical glimpse of how these spaces can be used to study gene and genome evolution at scale [4,5]. Embracing embeddings alongside traditional approaches can both shed new light on how proteins explore evolutionary space and help alleviate long-standing limitations of classical methods.

A learned, context-dependent coordinate system for protein evolution

Although pLM embeddings were initially difficult to interpret, they are now increasingly amenable to explanation and offer a way to overcome many of the constraints of the traditional MSA-based framework used to study molecular evolution. MSAs remain essential for computing sequence profiles, building evolutionary models, and informing 3D structure prediction, but the parametric models traditionally applied to them rely on simplifying assumptions (such as site independence and context-free substitution) that may fail to capture higher-order dependencies present in protein evolution [6]. Alongside the direct use of embeddings, a complementary route is to exploit pLMs as structure predictors (e.g., tools such as ESMFold can generate atomic-resolution models at genomic scale without MSAs [7]), opening the door to structure-based evolutionary analyses, such as structural phylogenetics, that were previously limited to proteins with experimentally determined structures. In an alignment-free context as provided by pLMs, nearby points in embedding space tend to share physicochemical properties, secondary and tertiary structure, and often broad functional roles, even when sequence identity is low, such as in many intrinsically disordered proteins. This reflects a key difference from classical parametric models: instead of assuming a single context-free substitution process, the model learns how a mutation's fate depends on its structural and functional environment.

For evolutionary biologists, a useful mental model is that embeddings locate sequences within a learned, context-dependent latent space, which could eventually be turned into an explicit coordinate system once appropriate quantitative axes and reference points are defined. Importantly, proximity in embedding space primarily reflects shared structural and functional patterns rather than evolutionary relatedness *per se*; pLMs have no mechanism for tracking the historical sequence of mutations that produced current proteins [1]. This means that embedding similarity and phylogenetic distance can diverge, especially at fine scales where even single amino acid changes can substantially alter representations [1,2]. Recognizing this distinction (that embeddings capture the consequences of evolution rather than its process) is essential if we are to use them rigorously for evolutionary inference. In this view, evolutionary change becomes motion: lineages trace trajectories as sequences diverge, duplicate, and specialize (Figure 1B). This idea has been partially realized by the 'evolutionary velocity' framework [8], which uses pLM predictions of local mutational preferences to construct a vector field of protein evolution, recovering directional trajectories across timescales from viral immune escape to the diversification of eukaryotic

protein families. Gene duplication followed by neofunctionalization may appear as a branch peeling away from an ancestral cluster, and convergent evolution as distant branches that nonetheless bend toward the same functional region of the space (Figure 1B). Importantly, these coordinates are learned directly from the distribution of natural sequences, not imposed by alignment-based constraints. That makes embeddings both powerful and imperfect: because they can identify very complex patterns, they can recover relationships that are invisible to constrained, context-free alignment models, especially in the 'twilight zone' or across deeply diverged lineages. Yet, they also inherit the sampling biases and blind spots of current sequence databases. pLMs are known to be biased by unequal species representation in their training data [9], and phylogenetic imbalances can distort the geometry of the very embedding spaces we propose to study [1,10]. It is also important to recognize that protein embeddings are not data: they are constructions derived from data, shaped by the model architecture, training objective, and composition of the training set. Different pLMs, trained with masked language modeling, autoregressive, or contrastive objectives, and on datasets of varying taxonomic breadth, can produce embedding spaces with meaningfully different geometries. This raises a critical question of generality, as evolutionary patterns detected in the embedding space of one model may not hold across architectures, and cross-model comparisons will be essential before drawing robust evolutionary conclusions from any single model. Recognizing and correcting these biases is therefore a prerequisite for any systematic evolutionary analysis of embedding landscapes.

From black box to evolutionary map

If embeddings provide us with a powerful, context-dependent latent space with the potential to be converted into a coordinate system, they also confront us with an important limitation: we do not yet know how to read the axes. We see that similar proteins cluster and that evolutionary relationships leave recognizable patterns, but we rarely know which directions correspond to specific biophysical, structural, or functional changes [2,11]. In that sense, pLMs are not just black boxes that make predictions; they are black boxes that have already summarized protein evolution for us, and we do not yet understand the summary [12,13].

We argue that this is not a fatal flaw but rather the central opportunity for the coming years. Rather than treating embeddings as magic features that improve similarity search, we should treat them as empirical objects to be interrogated with the full toolkit of molecular evolution and quantitative genetics. At the level of individual proteins, this task means asking which geometric and statistical properties of embedding space correspond to known constraints on protein-coding genes. At the level of whole proteomes, it suggests treating sets of embeddings as high-dimensional traits that describe a species' protein repertoire and can themselves evolve, covary across lineages, and come under selection.

We see this as a concrete research agenda for evolutionary embedding analysis. First, we need to map the landscape by systematically probing embedding spaces across different model architectures, training objectives, and datasets to identify which biochemical, structural, and evolutionary properties are encoded where, focusing on increasing explainability [14]. Second, we should study evolutionary paths in this space: sequences of embeddings that connect related proteins or proteomes, and ask which types of mutational changes and selection pressures could generate the observed changes along these paths [8] (Figure 1C). Third, we must bridge back to parameters by translating the pattern of nearby points around a sequence in embedding space into context-dependent propensities for mutations to occur and persist, so that embeddings inform and refine, rather than replace, interpretable evolutionary models [2,6,11]. If embedding spaces do encode intrinsic evolutionary information, then analyzing them as evolving traits in their own right, rather than as a mere aid to sequence search, should become a central goal of molecular evolution.

A new synthesis, not just a new tool

pLMs do not make classical molecular evolution obsolete, but they change what is possible when the two are combined. A phylogeneticist armed only with a pLM is like a geographer armed only with a detailed but unlabeled satellite image: the view is rich and continuous, but without the proper knowledge to read it, it is difficult to know what any given feature means. Embeddings give us a new picture of protein evolution, yet they still need the interpretability and hypothesis-testing framework that decades of molecular evolution methods provide.

The most exciting future lies in treating this relationship as a two-way street. On one hand, embeddings can do more than quietly assist existing pipelines; they can act as informative, context-aware priors that reshape how we set up evolutionary models, from which substitutions we consider plausible to which regions of sequence space we even bother to explore. On the other hand, traditional tools should not be relegated to post hoc validation but used extensively to interrogate what pLMs have learned: signals of positive selection, shifts in site-specific rates, or changes in constraint can be projected onto the latent landscape to reveal where, and perhaps why, the model's view of evolution aligns with or departs from our theories. This two-way exchange extends naturally to generative protein design, where models such as

ESM-C can produce functional sequences far from any natural homolog [15]; if embeddings encode evolutionary constraints, then the proteins these models generate should respect them, making generative design itself a test of the evolutionary structure pLMs have learned. Notably, the explicit incorporation of MSA-derived evolutionary information into modern pLMs does not generally improve and can even decrease performance on standard prediction tasks, suggesting that these models have already absorbed much of the signal that MSAs capture [6]. However, this finding does not make classical evolutionary tools redundant. MSAs and parametric models miss important aspects of evolution, such as epistatic interactions between sequence-distant positions, domain shuffling, and saturation, while pLMs encode these effects only implicitly and with unknown bias. The question is therefore not whether pLMs 'need' MSAs, but how explicit evolutionary models and comparative genomics can be used to test, interpret, and correct the evolutionary structure that pLMs have learned. It is worth noting that standard pLMs are trained with objectives, such as masked language modeling, that do not explicitly optimize for evolutionary signal, meaning that any evolutionary information retained in the embeddings is an emergent byproduct rather than a design goal, and some evolutionarily informative variation may have been lost or distorted in the process. This motivates two complementary strategies. The first, which we advocate throughout this piece, is to interrogate existing pLMs with evolutionary tools to determine how much evolutionary signal has been retained and where it resides. The second, potentially more powerful in the long run, is to design new pLMs whose training objectives explicitly incorporate evolutionary structure, for example, by training on phylogenetically balanced datasets, using site-specific evolutionary rates as auxiliary supervision, or incorporating comparative genomic signals. Fine-tuning existing models on evolutionarily structured tasks represents a practical intermediate path. These approaches are not mutually exclusive, as probing existing models can reveal which aspects of evolutionary information are systematically lost, directly informing the design of evolutionarily aware architectures.

Seen this way, pLMs offer a new lens on protein evolution that is richer than what we can encode in a single rate matrix, but they are not a replacement for explicit models or careful inference. Our community's task is not to stand before these learned landscapes in a mix of admiration and apprehension but to develop the conceptual and methodological tools needed to explore and map them and to integrate their insights with everything we have learned from half a century of molecular evolution. The journey to understand how protein sequences evolve has not ended; it has acquired a new compass.

Acknowledgments

R.F. acknowledges support from the European Research Council (grant agreement no. 948281) and the Spanish Agency of Research (grant agreement PID2024-161173NB-I00, funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU). R.F., A.B., and A.M.R. acknowledge support from the OSCARS project (European Commission's Horizon Europe Research and Innovation programme, grant agreement no. 101129751). S.V. is supported by the Spanish Ministry of Science and Innovation (MICIU) through the State Research Agency (AEI), grant PID2024-162055NB-I00, funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU. R.F. and S.V. acknowledge the support of the Departament de Recerca i Universitats de la Generalitat de Catalunya (2021 SCR 00420). I.C. and A.M.R. acknowledge support from the Spanish Agency of Research (grant agreement PID2024-162736OB-I00, funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU). S.A.H. was supported by institutional funds from Washington University School of Medicine.

Declaration of interests

The authors declare no competing interests.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Perplexity in order to improve clarity and grammar. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

[1]Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain
[2]Institute of Molecular and Cellular Plant Biology (CSIC-UV), Valencia, Spain
[3]Andalusian Center for Developmental Biology (CSIC-UPO), Seville, Spain
[4]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA

*Correspondence:
rosa.fernandez@ibe.upf-csic.es (R. Fernández) and ana.rojas.m@csic.es (A.M. Rojas).

https://doi.org/10.1016/j.tig.2026.05.014

References

1. Shaw, R. *et al.* (2025) Evaluating pretrained protein language model embeddings as proxies for functional similarity. *J. Mol. Evol.* 93, 765–776
2. Tule, S. *et al.* (2024) Do protein language models learn phylogeny? *Brief. Bioinform.* 26, bbaf047
3. Cao, Z. *et al.* (2025) Language models reveal a complex sequence basis for adaptive convergent evolution of protein functions. *Proc. Natl. Acad. Sci. U. S. A.* 122, e2418254122
4. Barrios-Núñez, I. *et al.* (2024) Decoding functional proteome information in model organisms using protein language models. *NAR Genom. Bioinform.* 6, lqae078
5. Martínez-Redondo, G.I. *et al.* (2025) FANTASIA leverages language models to decode the functional dark proteome across the animal tree of life. *Commun. Biol.* 8, 1227
6. Erckert, K. and Rost, B. (2024) Assessing the role of evolutionary information for enhancing protein language model embeddings. *Sci. Rep.* 14, 20692
7. Lin, Z. *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130
8. Hie, B.L. *et al.* (2022) Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* 13, 274–285.e6
9. Sawhney, R. *et al.* (2025) Fine-tuning protein language models unlocks the potential of underrepresented viral proteomes. *PeerJ* 13, e19919
10. Nijkamp, E. *et al.* (2023) ProGen2: exploring the boundaries of protein language models. *Cell Syst.* 14, 968–978.e3
11. Ektefaie, Y. *et al.* (2025) Evolutionary reasoning does not arise in standard usage of protein language models. *bioRxiv* https://doi.org/10.1101/2025.01.17.633626
12. Zhang, Z. *et al.* (2024) Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2406285121
13. Gujral, O. *et al.* (2025) Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proc. Natl. Acad. Sci. U. S. A.* 122, e2506316122
14. Hunklinger, A. and Ferruz, N. (2026) Towards the explainability of protein language models. *Nat. Mach. Intell.* 8, 649–662
15. Hayes, T. *et al.* (2025) Simulating 500 million years of evolution with a language model. *Science* 387, 850–858