PNAS RESEARCH ARTICLE | EVOLUTION

# Language models reveal a complex sequence basis for adaptive convergent evolution of protein functions

Zhenqiu Cao[a,b], Hongjiu Zhang[c], and Zhengting Zou[a,b,1]

Affiliations are included on p. 12.

Edited by David D. Pollock, University of Colorado Health, Aurora, CO; received September 6, 2024; accepted August 13, 2025 by Editorial Board Member Scott V. Edwards

**Convergent evolution, or convergence, refers to repeated, independent emergences of the same trait in two or more lineages of species during evolution, often indicating functional adaptation to specific environmental factors. Many computational methods have been proposed to investigate the genetic basis for organismal functional convergence, as an important way to decode the complex sequence–function map of proteins. These methods mostly focus on the convergence of amino acid states at the level of individual sites in functionally related proteins. However, even without site-level sequence similarity, protein function similarity may also stem from convergence of high-order protein features, which cannot be captured by the conventional methods. To fill this gap, we first derived numerical embeddings from protein sequences by pretrained protein language models (PLM). In four previously reported cases, we found that functionally convergent proteins have similar embeddings despite no site-level convergence, indicating that PLM embeddings can reflect convergence of high-order protein features. We then designed a pipeline to detect Adaptive Convergence by Embedding of Protein (ACEP). ACEP tests were significant on known and additional candidate genes with putative adaptive convergence like echolocation and crassulacean acid metabolism. Genome-wide application showed that the ACEP framework can effectively enrich such candidates. Relations between convergences of PLM embeddings and specific protein physicochemical features were further examined. In conclusion, PLM embeddings can indicate adaptive convergence of high-order protein features beyond site identities, demonstrating the power of deep learning tools for investigating the complex mapping between molecular sequences and functions.**

convergent evolution | protein evolution | language model | adaptation | deep learning

## Significance

In biology, repeated emergence of the same functional trait in evolution is important as it provides opportunity to decode the relations between genome or protein sequences to specific functions. Such functional convergence has been largely linked to sequence convergence at the level of single sites, because conventional methods cannot measure similarity of high-order features of sequences. This study reveals that the recent protein language models can extract embeddings from protein sequences reflecting high-order features, and develops statistical tests to evaluate the adaptive convergence of such features. The findings emphasize an underrated sequence basis for functional trait convergence in evolution, provide corresponding detection framework, and demonstrate potential power of deep learning in investigating the complex sequence–function mapping in evolutionary biology.

Convergent evolution, or convergence, refers to the biological phenomenon that the identical state of a trait emerges independently in two or more lineages of species during evolution. For example, some bat species and all toothed whales (TW) are capable of emitting and perceiving ultrasound (1), while this ability is absent in the common ancestors of these two evolutionarily distant lineages, rendering convergent emergence of echolocation in both lineages a parsimonious explanation. Because the probability of coincidentally arriving at the same state during evolution is low, convergence of organismal traits or biomolecule functions has been considered to be driven by adaptation to similar environmental factors or lifestyles (2), thus becoming a topic of interest in evolutionary biology. In the echolocation case, the convergently evolved ability helps bats and TW to forage in dim-light environments (1).

The genotype–phenotype mapping (GPM), or sequence–function mapping, is a central concept in biology for the understanding of how functions emerge and change by evolution (3). Given a phenotype or function convergence, it is intriguing to investigate whether its genetic basis is also convergent evolution at the molecular sequence level (4–15). For instance, Li et al found that phylogeny reconstruction based on the amino acid sequences of the Prestin protein unites echolocating bats (EB) and the bottlenose dolphin together, indicating high sequence similarity between the two lineages unexpected under the nonadaptive neutral evolution, thus suggesting adaptive sequence convergence as a basis for the functional convergence of echolocation. Specifically, a convergent asparagine-to-threonine substitution at site 7 (N7T) of Prestin in both lineages was identified and later confirmed by experimental assay as functionally related to echolocation (9, 14). Many molecular evolution strategies have been developed to detect such site-level sequence convergence underlying functional convergence, based on site-specific likelihood support for phylogenetic convergence ($\Delta$SSLS) (4, 11), ratio between convergence and divergence (4, 12), ratio between observed and expected convergence (16, 17), ratio between nonsynonymous and synonymous convergence (Csubst) (7), convergence at conservative sites (CCS) (9, 15), amino acid profile change

Author contributions: Z.C., H.Z., and Z.Z. designed research; Z.C. performed research; Z.C. analyzed data; and Z.C., H.Z., and Z.Z. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.D.P. is a guest editor invited by the Editorial Board.

Copyright © 2025 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1] To whom correspondence may be addressed. Email: zouzhengting@ioz.ac.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2418254122/-/DCSupplemental.

Published September 23, 2025.

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

with at least one site-level change Profile Change with One Change (PCOC) (18), correlation between amino acid state and quantitative traits Convergent Amino Acid Substitutions (CAAS) (19), etc.

Despite many reported cases of function-related site-level convergence in proteins, these existing methodologies have a major caveat. Effectively, all strategies focus on site-level sequence changes. However, it is known that GPM is complex with extensive interactions between individual sites, and different sequences may map to similar functions (20). Hence, site-level convergence is not necessary for functional convergence in proteins (21). As an example, having adapted to the hypoxic environment, the hemoglobins (Hbs) of multiple high-altitude waterfowl species have been shown to convergently possess high Hb-O2 affinity. Nevertheless, the respective Hb protein sequences show limited site-level convergence of amino acid states, which are, moreover, largely not responsible for the affinity shift (10). Due to the heterogeneity or lineage specificity of the sequence evolution process, it is likely that organismal functional convergence is achieved through convergence of higher-order features in protein sequences, while exhibiting divergent site-level patterns. Indeed, there are cases of adaptive protein physicochemical or structure convergence without site-level similarity (22–24).

How to detect adaptive convergence of such high-order features in proteins? The current conventional models of protein sequence evolution typically describe dynamics of single amino acid sites, unable to address high-order features like epistatic relations between sites or secondary structures. Hence, we seek the statistical capacity of recent pretrained protein language models (PLM), as they can capture context patterns of sites in the sequences, and these models have been shown to encode high-order protein features for predicting spatial contact, protein structures, and functions (25–28). We trained a neural network encoder on top of the fixed large protein language model ESM-MSA-1b to obtain fixed-length numerical embeddings for any protein sequence, and we demonstrated in multiple known cases that these embeddings reflect high-order feature similarities of proteins with functional convergence, despite the site-level divergence. We further developed an analysis pipeline to detect Adaptive Convergence by Embedding of Protein (ACEP, https://huggingface.co/NEO699700/ACEP), testing for unexpected PLM embedding similarity between proteins of focal species lineages against a simulated null distribution. We applied the ACEP test to specific candidate proteins with convergence for plant crassulacean acid metabolism (CAM), and to a genome-wide set of proteins in echolocating mammals. Alongside significant ACEP results observed for the known candidates, new putatively adaptive convergence genes for echolocation were enriched. We also examined possible relations between convergences of PLM embeddings and specific high-order protein features in multiple cases. Our findings emphasized the prevalent role of high-order protein features in the convergent evolution of organismal functions, provided a computational framework for detecting adaptive protein convergence, and demonstrated the capacity of deep learning methodology to capture evolutionary sequence features and facilitate the understanding of complex GPM.

## Results

**Fixed-Length Embeddings Derived from Pretrained PLM Reflect Evolutionary Relationship between Protein Sequences.** Protein language models are deep neural networks trained on large scale protein sequence datasets by the mask-prediction training strategy widely used in the natural language studies. We focus on the model ESM-MSA-1b, which explicitly harnesses the evolution information in multiple sequence alignments (MSAs) and has been reported to show better performance on tasks like protein structure prediction compared to PLMs with more parameters (29). The pretrained model is composed of 12 MSA Transformer layers with 100 M parameters, and was trained on 26 million MSAs effectively spanning the UniRef database. For each protein sequence of length $L$, ESM-MSA-1b outputs a local embedding $E_l$ of size $L \times 768$, and a global embedding $E_{g0}$ of size 768 can be calculated by averaging the local one across the length ($L$) dimension (*SI Appendix*, Fig. S1). Since it has been reported that averaging is suboptimal for global embedding derivation, we adopted a bottleneck strategy and constructed an encoder–decoder network with $E_l$ as input, trained by the decoder reconstruction loss, while fixing the ESM-MSA-1b backbone (30) (Fig. 1*A*). After training on 37,998 mammal sequences sampled from alignments in the OrthoMaM database (*Materials and Methods*), the encoder output (a vector of size 300) can be used as the global embedding $E_g$.

To capture the molecular basis of functional convergence, embeddings that can reflect unexpected evolutionary similarity between protein sequences is needed. According to neutral theory, the varying levels of similarity between different orthologous sequences largely result from phylogenetic divergence, only occasionally shaped by adaptive convergence in rare cases of a few genes in some taxonomic groups. Hence, embeddings suitable to detect adaptive convergence should also reflect phylogenetic divergence in most genes lacking adaptive convergence. To verify that $E_g$ encodes such evolutionary information, we hypothesized that the distances between $E_g$s of pairs of species highly correlate with their phylogenetic distances, which reflects an evolutionary relationship. Indeed, for all species pairs in 3,000 randomly sampled mammalian MSAs, cosine and Euclidean distances of $E_g$s have high Spearman correlations with phylogenetic distances, up to an average of 0.87, which are significantly higher than the correlations (0.69) between cosine or Euclidean distances of $E_{g0}$s and phylogenetic distances (Wilcoxon test, $P < 1 \times 10^{-300}$) (Fig. 1*B*). Hence, the bottleneck-derived $E_g$ trained by mammal protein sequences seems to carry more evolutionary information for evaluating sequence divergence as well as convergence, and was used as the primary global embedding of a protein sequence in all downstream analyses unless specified.

**PLM Embeddings Reflect Similarity in High-Order Features of Proteins Despite Site-Level Sequence Divergence.** To validate that similarity in high-order features can be reflected by PLM embeddings, we investigated multiple cases in which distantly related proteins converge to serve similar functions or adapt to similar environmental factors. We hypothesized that these highly diverged proteins may appear dissimilar at the site level, but manifest embedding similarity due to convergence of high-order protein sequence features (Fig. 1*C*).

The first case is the convergence of Hbs in jawed and jawless vertebrates. In both cyclostome species (lampreys and hagfish, jawless vertebrates) and gnathostomes (jawed vertebrates), Hbs carry out the function of binding $O_2$ in the blood. However, gene phylogeny reconstructed from amino acid sequences of vertebrates showed that cyclostome Hbs (cHbs) shared most recent common ancestor with the gnathostome cytoglobins (gCygbs) rather than gnathostome Hbs (gHbs) (*Materials and Methods*; *SI Appendix*, Fig. S2*A*; also see figure 1 in ref. 31). This indicates the function of $O_2$ transportation in blood was convergently evolved in cHbs and gHbs, realized by similar oxygenation-linked cooperative changes of quaternary structure (31). To depict the embedding similarity between these proteins, we conducted principal

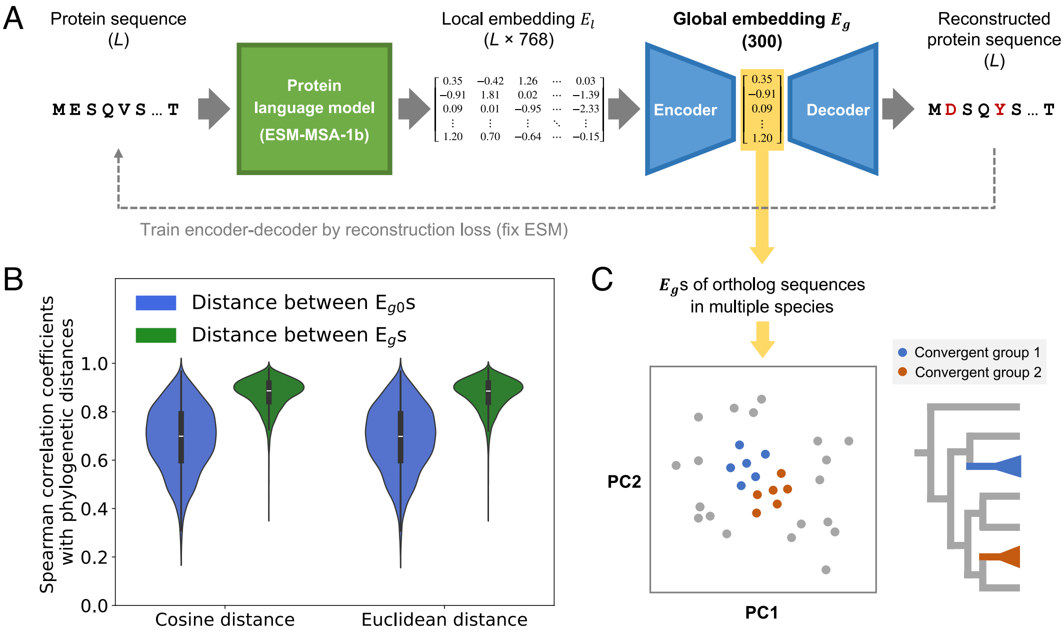

**Fig. 1.** Encoder network trained by bottleneck strategy calculates evolutionarily informative global embedding $E_g$, which may reflect high-order feature similarity of divergent protein sequences. (*A*) The encoder–decoder bottleneck network design and training strategy. Input protein sequences were processed by PLM backbone to get local embedding $E_l$, which were then input into encoder to get global embedding $E_g$ (highlighted in yellow). (*B*) Distributions of Spearman correlation coefficients between embedding distances and phylogenetic distances. Higher coefficient values by using $E_g$ distance than by using $E_{g0}$ indicate the former to be more evolutionarily informative. *X* axis labels indicate whether correlation coefficients are calculated based on cosine or Euclidean embedding distances. For each violin, the upper and lower bounds of the black rectangle represent corresponding quartiles, while the white segment in the middle indicates the median. (*C*) Schematic hypothesis of how PLM embedding may reflect high-order protein feature convergence, in which case orthologous proteins from two divergent lineages (blue and red) in evolution may show similar embeddings.

component analysis (PCA) of the $E_g$ embeddings. Intriguingly, in contrast to the sequence resemblance to gCygbs, cHbs (red dots in Fig. 2*A* and *SI Appendix*, Fig. S2*B*) showed smaller distances to gHbs, particularly gHbβ (blue and green dots in Fig. 2*A* and *SI Appendix*, Fig. S2*B*) in PC2. In the principal component space, the large distances between different cHbs in PC1 were due to deep hagfish-lamprey divergence, and the distances between gCygbs in PC2 was contributed by multiple Cygb paralogs in bony fish species, also likely due to deep divergence after whole-genome duplication events (*SI Appendix*, Fig. S2*C*). The cHb-gHbβ embedding similarity in PC2 thus supported their functional convergence. Since the PLM embeddings may reflect various high-order sequence features of the protein, it is reasonable not to expect all embedding PCs, i.e., all features to show similarity between functionally convergent homologs. However, under neutral evolution without convergence, divergent sets of homologs are not expected to show similarity in any PC axis. Hence, the existence of cHb-gHbβ similarity in PC2 indicated nonneutral sequence convergence. Furthermore, we calculated the cosine $E_g$ embedding distances between different proteins, and found that cHbs have smaller distances with gHbβs than with gCygbs or with gHbαs (Fig. 2*B*, Mann–Whitney *U* test, $P < 2 \times 10^{-37}$). The same was true for Euclidean $E_g$ embedding distance as well (*SI Appendix*, Fig. S2*D*, Mann–Whitney *U* test, $P < 3 \times 10^{-20}$). PCA and distance comparisons based on $E_{g0}$ embeddings exhibited similar patterns, showing higher similarity between cHbs and gHbs than between cHbs and gCygbs (*SI Appendix*, Fig. S2 *E–G*, Mann–Whitney *U* test, $P < 2 \times 10^{-15}$). These observations support that PLM embeddings can reflect the functional agreement of cHbs with gHbs, despite site-level sequence dissimilarity. Specifically, our results suggest high-order feature convergence of cyclostome Hb with gnathostome Hbβ.

Second, we investigated two venom toxins in mammals and reptiles. It has been reported that two kallikrein (KLK)-related serine proteases, i.e., BLTX of the shrew *Blarina brevicauda* and GTX of the lizard *Heloderma horridum*, are toxic due to higher catalytic activity than their nontoxic counterparts. This functional convergence is likely due to similar physicochemical features of the catalytic cleft in these two proteins, both carrying unique insertions in a nearby regulatory loop (22). Correspondingly, in the PCA plots for the $E_g$ embeddings of the two proteins and other KLK-related homologs, we observed that GTX and BLTX locate near each other in PC1 – PC4 (Fig. 2*C* and *SI Appendix*, Fig. S3*A*). Specifically, the Blarinasins were clearly distant from the BLTX and GTX in at least PC2 and PC4, which accounted for ~24% embedding variance. This pattern contradicts the gene phylogeny reconstructed by the same protein sequences based on site-level evolution models, in which the Blarinasins and BLTX were closely related (*SI Appendix*, Fig. S3*B*). This contrast indicated that PLM embeddings may reflect similarity of protein structure and physicochemical features realized by distinct nonneutral sequence changes. PCA based on $E_{g0}$ embeddings exhibited similar patterns, showing higher similarity between GTX and BLTX on PC2 (*SI Appendix*, Fig. S3*C*). Nevertheless, in contrast to the PCA patterns, the cosine and Euclidean $E_g$ embedding distances between GTX and BLTX are not significantly smaller than those between the two toxins and their nontoxic homologs (*SI Appendix*, Fig. S3 *D* and *E*), probably due to the overall structural dissimilarity of BLTX and GTX not captured by major principal components.

In a third case, we focused on the ferrous iron uptake in green plants. It has been reported that the $Fe^{2+}$ uptake protein in green algae (*Chlamydomonas*), the iron-regulated transporter 1 (CrIRT1), is phylogenetically close to Angiosperm Zn transporters while functionally and structurally convergent with the known ferrous iron uptake proteins in Angiosperm (23). Thus $Fe^{2+}$ uptake proteins evolved convergently in the zinc-regulated, iron-regulated transporter-like protein (ZIP) family (*SI Appendix*, Fig. S4*A*, also see figure 3A in ref. 23) Comparing the $E_g$ embedding of CrIRT1 with those of other proteins in the family, we found that the CrIRT1 embedding showed smaller cosine distances with the Angiosperm $Fe^{2+}$ uptake proteins, particularly OsIRT1 in rice, than with the Angiosperm Zn transporters AtZTP29 or OsZIP13, consistent with the known scenario of function and structure convergence (Fig. 2*D*, comparing with structure similarity patterns

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

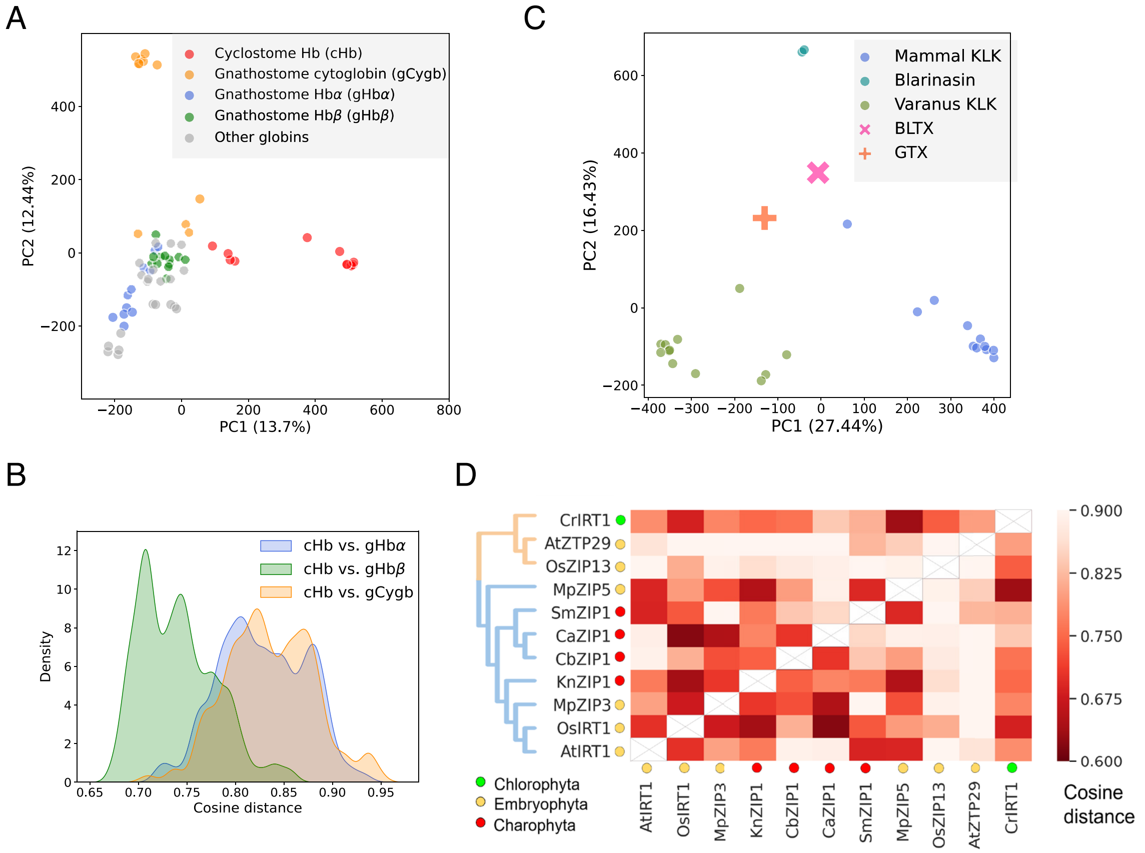

**Fig. 2.** PLM embeddings reflect high-order feature similarity of proteins despite site-level divergence. (*A*) PCA of $E_g$ embeddings reflecting hemoglobin convergence between jawed and jawless vertebrates. Position of each dot indicates the PC1 and PC2 values of the corresponding globin protein embedding. (*B*) Kernel density plot showing the distributions of cosine embedding distances between different globin groups. The kernel density estimation (KDE) parameter bw_adjust was set to 0.5. (*C*) PCA of $E_g$ embeddings reflecting convergence between kallikrein (KLK)-related serine protease toxins in mammal and reptile. Position of each dot indicates the PC1 and PC2 values of the corresponding protein embedding. (*D*) Heatmap showing the pairwise cosine $E_g$ embedding distances between metal ion transporter proteins in green plants. The phylogeny and taxonomic information are adapted from Rodrigues et al. (23).

in figure 5A of ref. 23). The CrIRT1 and OsIRT1 similarity was not obvious when measured by $E_{g0}$ embedding distances (*SI Appendix*, Fig. S4*B*).

Thus, in all three cases above, we observed nontrivial embedding similarity, particularly for the $E_g$ global embeddings, between homologous proteins separated on gene trees but similar in functions. As control, we also evaluated two alternative simple similarity measures, the *p* distance and the BLOSUM62 score (*Materials and Methods*), in the cases of Hbs and proteases. In contrast to the patterns observed on PLM embeddings, we found significantly smaller *p* distances and higher BLOSUM62 scores between cHb and the closely related gCygb than between cHb and the gHbs (Mann–Whitney *U* test, $P < 2 \times 10^{-9}$, $P < 1 \times 10^{-46}$, *SI Appendix*, Fig. S5 *A* and *B*). Hence, simple sequence similarity or physicochemical similarity measures mainly reflected phylogenetic relationship between proteins, while PLM embeddings could reveal functional convergence, putatively according to the high-order features captured.

Additionally, we simulated sets of negative control sequence alignment data for the hemoglobin case, the toxin case and the ferrous iron transporter case (*Materials and Methods*). Simulated with the same gene tree topology, branch lengths, and site-specific evolution rates as inferred from the corresponding real data, these negative control data should contain no adaptive convergence signal. Interestingly, when the gaps in real sequence alignments were directly copied into the simulated alignments, the latter exhibited similar convergent patterns of embedding PCA or distance distributions as the respective real cases (*SI Appendix*, Fig. S5 *C–F*). On the contrary, when simulating without gap copying, embedding PCA or distance distributions of the simulated sequences only followed phylogenetic relationships as expected from the simulated neutral sequence evolution process (*SI Appendix*, Fig. S5 *G–J*). This difference between two simulation strategies strongly indicates that the similarities between sequences with functional convergence were not the result of site-level convergence, but likely resulted from high-order feature convergence partially manifested as gaps in the sequence data.

**Proteins of Thermophilic Bacteria and Archaea Exhibit Convergence of PLM Embeddings Despite Phylogenetic Divergence.** In addition to the above cases of functional convergence in individual genes, we investigated the convergent evolution of proteins in thermophilic Archaea and Bacteria. Evidence has been found supporting that the common ancestor of Archaea and Bacteria are mesophilic (32). Hence for thermophiles in the two domains, although extensive horizontal gene transfer (HGT) has been proposed to facilitate their evolution (32), many proteins may have experienced convergent evolution of high-order features adapting to high-temperature environments (33). To check whether such events can be reflected by PLM embeddings, we collected protein alignments from the COG database (34) with orthologous sequences in 36 prokaryote strains, composed of 10 hyperthermophilic Archaea (AH), 6 hyperthermophilic Bacteria (BH), 10 mesophilic Archaea (AM), and 10 mesophilic Bacteria (BM) (*Materials and Methods* and

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

*SI Appendix*, Table S1). Due to deep divergence, protein alignments of only 27 conserved genes were obtained, and we then excluded the possibility of HGT by requiring monophyly of both Archaea and Bacteria strains separated by an internal branch with bootstrap value higher than 60% in the gene tree. We evaluated possible protein convergence between thermophilic Archaea and Bacteria by testing whether the $E_g$ embedding distances of AH-BH are smaller than those of AH-AM or BH-BM.

Among 27 conserved genes, 15 genes exhibit monophyly of Archaea and Bacteria. For 9 out of these 15 genes, the distances between AH-BH were smaller than those between AH-AM or between BH-BM under both cosine and Euclidean distance (*SI Appendix*, Table S2). These convergent genes include six aminoacyl-transfer ribonucleic acid (tRNA) synthetases, ribonuclease HII, signal recognition particle (SRP) GTPase Ffh and SRP receptor FtsY. As examples, we examined the $E_g$ embeddings of proteins in both COG0013 (alanine-tRNA ligase) and COG0143 (Methionyl-tRNA synthetase). While the thermophilic Archaea and the Bacteria sequences showed reliable divergence on both gene trees (red and orange taxa in Fig. 3 *A* and *B*), the $E_g$ embeddings of AH and BH clearly clustered together in PCAs (red and orange dots in Fig. 3 *C* and *D*). Correspondingly, the embedding distances between AH and BH were significantly smaller than those between AH and AM or between BH and BM (Mann–Whitney *U* test on cosine distance, $P < 6 \times 10^{-5}$, Fig. 3*E*; $P < 2 \times 10^{-11}$, Fig. 3*F*; the same is true for Euclidean distances, *SI Appendix*, Fig. S6 *A* and *B*, $P < 3 \times 10^{-4}$). In addition, we observed the same trends of high thermophile similarities by PCA and distance distribution comparisons in both COG0013 (*SI Appendix*, Fig. S6 *C–E*) and COG0143 (*SI Appendix*, Fig. S6 *F–H*) based on $E_{g0}$ embedding distances. In contrast, the *p* distances and BLOSUM62 scores between AH and BH were not significantly smaller than those between AH and AM or between BH and BM, in both COG0013 and COG0143 (Mann–Whitney *U* test, $P > 0.2$, *SI Appendix*, Fig. S6 *I–L*). As a negative control, we also generated COG0013 and COG0143 sequence alignments by simulation of neutral evolution, as in the previous cases. There was no higher embedding similarity between simulated orthologs of AH and BH than that of AH and AM or that of BH and BM, even when gaps were copied from the real data (*SI Appendix*, Fig. S6 *M–R*). Meanwhile, simulated COG0013 orthologs within Bacteria and within Archaea exhibited higher embedding similarity than between two domains, consistent with corresponding phylogenetic divergence (*SI Appendix*, Fig. S6 *M–O*). This confirmed the significance of the observation in real data, and further suggested that the thermophile convergence of these proteins was not primarily realized by gap-related structural convergence, but mainly by other mechanisms such as similar physicochemical properties. Overall, these findings indicate prevalent high-order physicochemical feature convergence revealed by PLM embeddings during thermophile adaptation in Archaea and Bacteria despite site-level sequence divergence.

**Empirical Tests Based on PLM Embeddings Can Reflect Known Sequence Convergence in Genes Related to Functional Convergence.** Based on the above evidence that PLM embeddings can reflect high-order feature convergence of proteins, we further designed a computational pipeline to detect ACEP. Previous methods detecting site-level sequence convergence usually compared the observed level of convergence events to a neutrally evolved background, such as expectation calculated by sequence substitution models (17) or synonymous convergences (7). In ACEP pipeline, sequences simulated under neutral evolution models served as background, which assumes that the sequences diverge under certain levels of evolutionary conservation, i.e., primarily under purifying selection due to specific functional constraints at each amino acid site, without lineage-specific adaptation. We proposed that, if orthologs of a protein experienced adaptive convergence in two lineages of species with organismal function convergence, the $E_g$ embedding distances between orthologs of these two lineages should be significantly smaller than distances between the simulated backgrounds. Hence, for each gene, we inferred evolution parameters including branch lengths on species tree, site-wise evolution rates, amino acid equilibrium frequencies. Then we simulated sequence evolution by these parameters for 100 times (*Materials and Methods*). Between two focal lineages of species with functional convergence, the mean value of pairwise $E_g$ embedding distances $\overline{d_{real}}$ was calculated and compared with a null distribution of 100 mean distances ($\overline{d_1}$, $\overline{d_2}$, …, $\overline{d_{100}}$) derived from the simulated replicates. An empirical *P*-value can be calculated as the proportion of $\overline{d_i}$ s equal to or smaller than $\overline{d_{real}}$. Accordingly, genes with significantly smaller $\overline{d_{real}}$ than the empirical distribution of $\overline{d_i}$ s were considered candidates of adaptive sequence convergence (Fig. 4*A* and *Materials and Methods*).

We first applied this ACEP pipeline based on $E_g$ embeddings to two known cases of individual genes. One case is the hearing gene *SLC26A5* (Prestin) in echolocating mammals. It is considered to have experienced adaptive convergence in echolocating mammals, with many sites showing convergent patterns between EB and TW (9). Consistent with our expectation, the ACEP analysis was significant (empirical *P*-value < 0.01, based on both cosine and Euclidean $E_g$ embedding distances) for *SLC26A5* when setting EB and TW as focal lineages (Fig. 4 *B* and *C* and *SI Appendix*, Fig. S7*A*). As another existing case, CAM has evolved independently in many lineages of plants. The phosphoenolpyruvate carboxylase (*PEPC*) and its kinase phosphoenolpyruvate carboxylase kinase (*PPCK*) are key genes regulating the periodical fixation of $CO_2$ at night, and they have shown sequence- or expression-level convergence in some but not all of the CAM species (35). The ACEP analysis was conducted for two monocot (*Agave tequilana* and *Ananas comosus*) and two eudicot (*Kalanchoe laxiflora* and *Kalanchoe fedtschenkoi*) CAM species (*SI Appendix*, Fig. S7*B*). Interestingly, two isoforms *PEPC1* and *PEPC2* showed drastically different results. *PEPC1* is highly expressed especially during the light period of the day, and was not significant in the ACEP test ($P > 0.7$, *SI Appendix*, Fig. S7*C*). Expression of *PEPC2* is relatively low, but much higher during the dark period than during the light period. Respectively, *PEPC2* was significantly convergent among CAM plants in the ACEP test ($P < 0.01$, *SI Appendix*, Fig. S7*D*), as well as *PPCK* ($P < 0.01$, *SI Appendix*, Fig. S7*E*). This is consistent with the previous finding that *PEPC2* showed site-level convergence with another CAM species not in our dataset, which may contribute to its increased activity in CAM plants (35).

When substituting the $E_g$ embeddings by the $E_{g0}$ embeddings in the ACEP tests, we only observed significance on Prestin (*SI Appendix*, Fig. S8 *A* and *B*). Thus, the bottleneck-derived $E_g$ embeddings seemed to be more effective in detecting sequence convergence. Besides, our bottleneck encoder–decoder trained on mammalian protein sequence data can derive embeddings that reflect sequence convergence of CAM-related plant proteins. demonstrating its potential flexibility. In turn, we also trained the same bottleneck encoder–decoder network by a comparable amount of protein sequences in vascular plants (*Materials and Methods*). The plant-protein-trained $E_g$ exhibited high phylogenetic informativeness, which was nevertheless slightly lower than the mammal-protein-trained $E_g$ on plant proteins (Wilcoxon tests on Spearman correlation coefficients between embedding distances and phylogenetic distances, $P < 1 \times 10^{-112}$, green and yellow

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

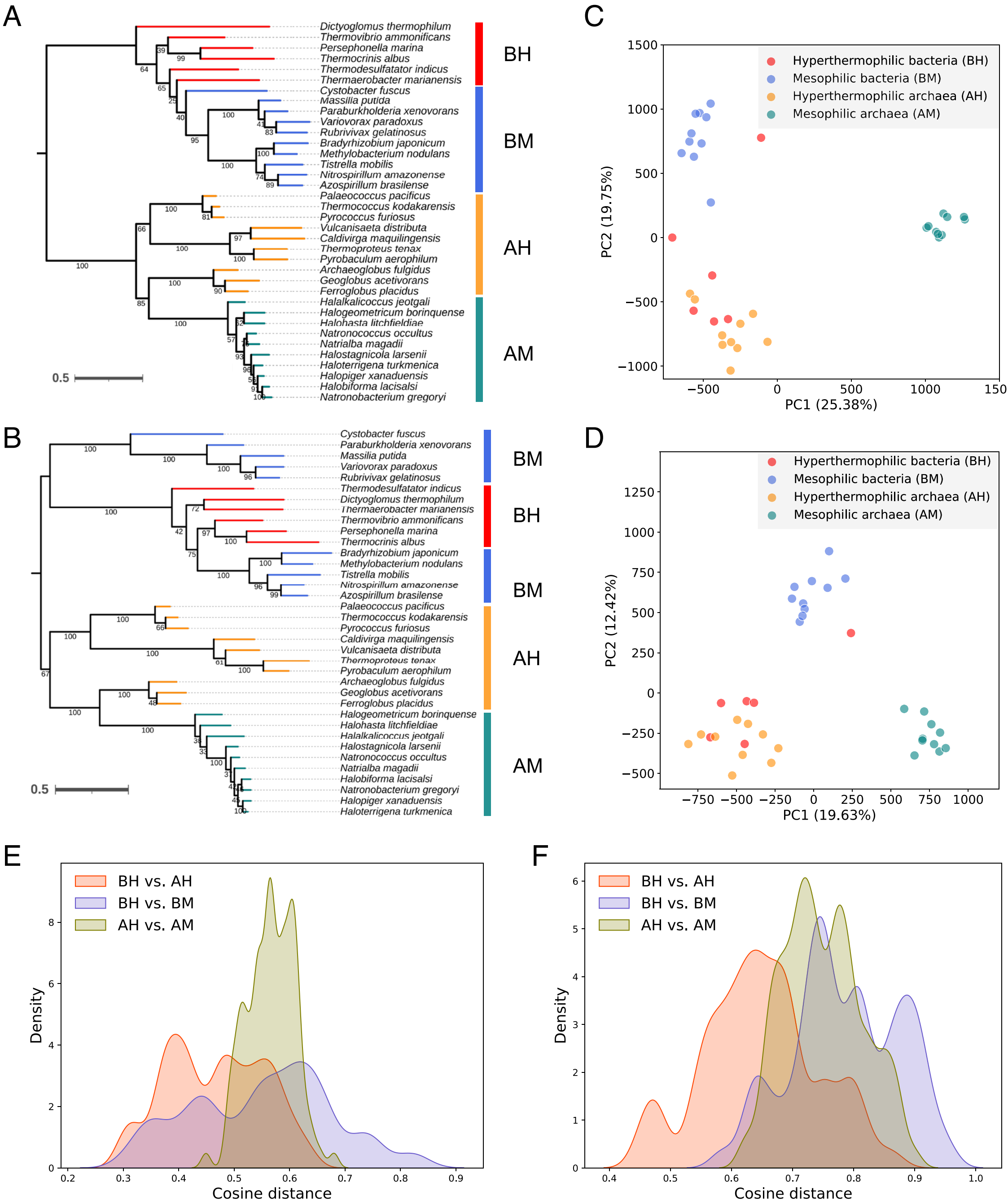

**Fig. 3.** Conserved proteins in thermophilic bacteria and archaea exhibits convergence of PLM embeddings despite phylogenetic divergence. (*A* and *B*) Maximum likelihood gene trees of (*A*) COG0013 and (*B*) COG0143, showing monophyly of bacteria and archaea orthologs. Numbers under each branch represent bootstrap values (N = 100). (*C* and *D*) PCAs of $E_g$ embeddings reflecting convergence between thermophilic bacteria and archaea in (*C*) COG0013 and (*D*) COG0143. Position of each dot indicates the PC1 and PC2 values of the corresponding ortholog $E_g$ embeddings. (*E* and *F*) Kernel density plots showing the distributions of cosine $E_g$ embedding distances between different ortholog groups in (*E*) COG0013 and (*F*) COG0143. The KDE parameter bw_adjust was set to 0.5.

violins in *SI Appendix*, Fig. S8*C*). ACEP tests by plant-protein-trained $E_g$ can only reflect sequence convergence in *PPCK*, showing marginal significance for *PEPC2* and no significance for mammalian Prestin (*SI Appendix*, Fig. S8 *D–G*). Hence, performance of the PLM embeddings may be affected by means of derivation and training data.

In addition to ACEP tests based on PLM embedding distances, we also calculated the *p* distances and BLOSUM62 scores between real sequences and between simulated sequences to derive empirical *P*-values for the above four proteins. Empirical tests by *p* distances were not significant for any of the four cases using the cutoff of $P$ = 0.01, but showed small *P*-values in Prestin and *PEPC2* (*SI Appendix*, Fig. S8*H*), which are the two cases with reported site-level convergence. Meanwhile, empirical tests by BLOSUM62 scores showed significance in *PEPC1* and *PEPC2*, and exhibited small *P*-values for Prestin and *PPCK* (*SI Appendix*, Fig. S8*I*). These results indicated that the site-level physicochemical similarity (BLOSUM62 scores) may reflect additional convergence beyond amino acid states (as measured by *p* distances), however to a lesser extent than the PLM embedding distances.

**The ACEP Tests Identified Known and Additional Candidate Genes Underlying Adaptive Convergent Evolution of Echolocation in a Mammal Genome.** Since the ACEP analysis has demonstrated capability of reflecting adaptive convergence in known cases, it

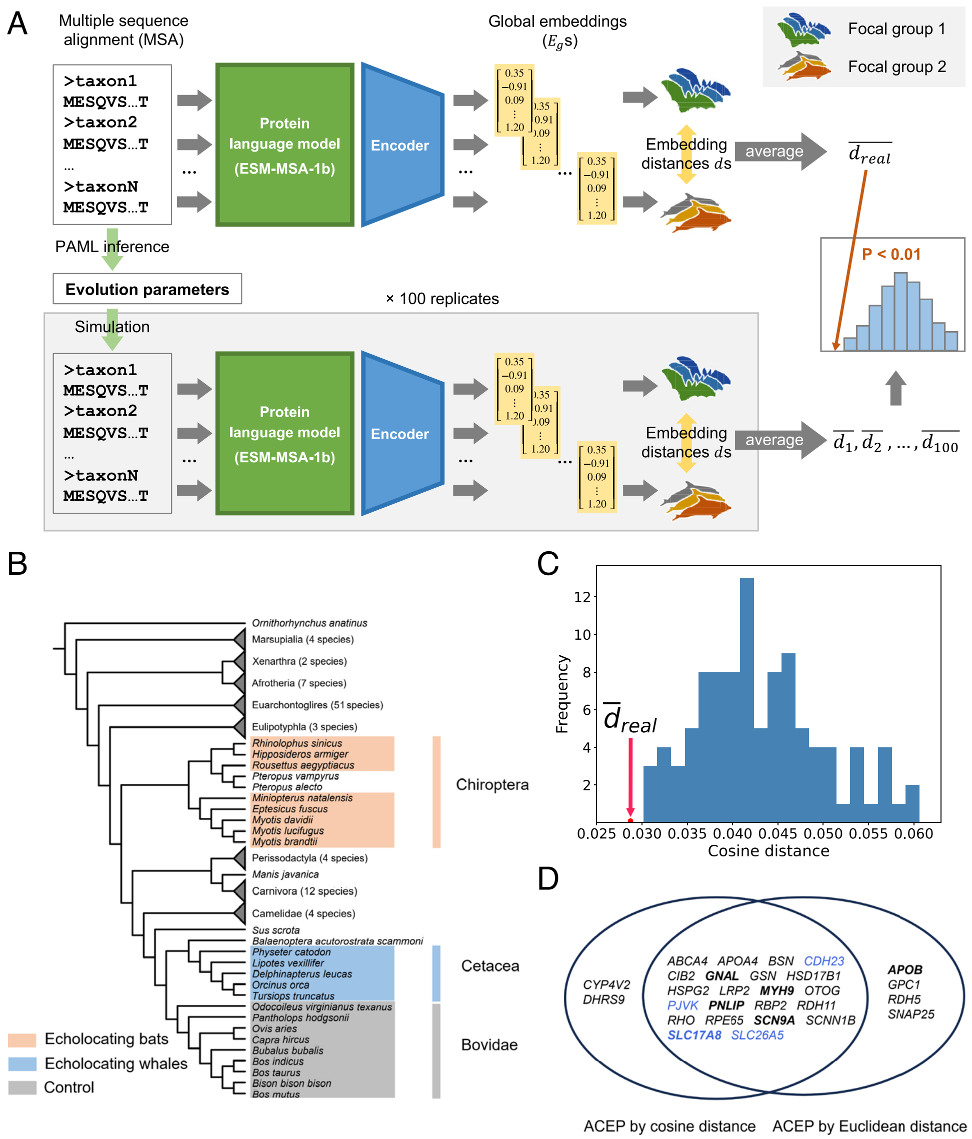

**Fig. 4.** Convergence tests on PLM embeddings identify known and additional candidate genes underlying adaptive convergent evolution of echolocation. (*A*) Diagram of the ACEP test pipeline. Embedding distances between focal species groups are compared with a null distribution formed by 100 neutral evolution simulations. Genes with empirical *P*-values less than 0.01 are considered as significant convergent genes. (*B*) Cladogram of the 115 mammalian species involved in the ACEP test of echolocating mammals, showing major phylogenetic relationships between focal groups, which are highlighted by colors. (*C*) ACEP test result of *SLC26A5* (Prestin) by cosine embedding distance, showing $P < 0.01$. The red arrow points to the $\overline{d_{real}}$ value marked by a red dot on the *X* axis. The blue histogram is the distribution of mean distances derived from each of the 100 simulations. (*D*) Venn diagram showing the two sets of ACEP significant genes based on cosine and Euclidean distances under the Reactome term "Sensory Perception." Previously reported echolocation-related genes are shown in blue. Genes under significant positive selection in echolocating mammals are shown in bold.

was then applied to a genome-wide investigation regarding all available orthologous genes in echolocating mammals. We set the focal lineages with adaptive functional convergence to be EB and TW (Fig. 4*B*), and obtained all protein MSAs in the OrthoMaM v10c database which, after quality control, contain at least one species in both TW and EB and up to 13 echolocating species. Due to the length limit of the PLM input sequence, we segmented long protein sequences, resulting in 12,666 protein fragment MSAs belonging to 11,559 proteins (*Materials and Methods*). For each fragment, the TW-EB mean $E_g$ embedding distance was then tested against simulated null distribution as described above. 769 fragments from 756 genes showed empirical *P*-value < 0.01 when cosine embedding distance was calculated. Functional enrichment analysis showed that the Reactome term Sensory Perception was significantly enriched in the 756 genes, with 24 associated genes (*Q*-value < 0.034, *SI Appendix*, Fig. S9*A*). Similarly, when Euclidean $E_g$ embedding distance was calculated, Sensory Perception was also significantly enriched ($Q < 0.035$, *SI Appendix*, Fig. S9*B*), with 26 genes included in the 849 genes showing $P < 0.01$. The significant genes by cosine distances and by Euclidean distances overlap substantially among all genes (652 genes, $P < 1 \times 10^{-300}$, Hypergeometric test, *SI Appendix*, Fig. S9*C*). These results suggested that the ACEP pipeline can enrich genes with adaptive convergence in echolocating mammals.

We then focused on the genes under the term Sensory Perception. 22 genes were shared between the two gene sets obtained from respective analyses using cosine and Euclidean distances. Among these 22 common genes, there are previously reported echolocation-related genes such as *SLC26A5* (Prestin), *CDH23*, *PJVK*, *SLC17A8* (11, 36–38) (Fig. 4*D*), indicating that the ACEP pipeline results aligned with conventional findings. In addition, the remaining genes in the two sets are also associated with echolocation-related physiological functions. For example, the calcium and integrin-binding protein 2 (CIB2) interacts with the known echolocation-related protein TMC1 (11, 36), and loss of *CIB2* causes profound hearing loss and abolishes mechanoelectrical transduction in mouse auditory hair cells (39, 40). Potential functional relations between the significant genes and echolocation are summarized in *SI Appendix*, Table S3.

As control, we changed the focal lineage of TW to the closely related lineage of bovids (Fig. 4*B*), and repeated the ACEP pipeline looking for convergence between EB and bovids. Although we found more genes with $P < 0.01$ under both cosine and Euclidean distances (*SI Appendix*, Fig. S9 *D* and *E*), the term Sensory Perception was no longer significantly enriched ($Q > 0.13$) in the control analysis. As examples of individual genes, *SLC26A5*, *PJVK*, *CIB2,* and *GSN* all showed $P < 0.01$ in echolocating mammals, while showing $P > 0.24/0.14/0.05/0.14$ in

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

control. Hence, the sensory perception genes found by ACEP between echolocating mammals are likely to be functionally related to echolocation.

To further investigate whether these putatively convergent genes have experienced adaptive evolution in echolocating mammals, we conducted a branch-site test of positive selection for the union of the two sets, totaling 28 genes. All branches in the clades of bats and TW were set as foreground (Fig. 4*B*). Six out of 28 genes (*SLC17A8*, *MYH9*, *SCN9A*, *GNAL*, *PNLIP*, *APOB*) were significant with $Q < 0.02$ (Fig. 4*D*). Among these genes, *SLC17A8*, *MYH9*, *SCN9A* have been reported to cause hearing loss (41–43). *APOB* expression has been observed in the inner ear perilymph (44). *GNAL* encodes a stimulatory G protein subunit $G\alpha_{olf}$, expressed in striatal medium spiny neurons which is related to motor control and Parkinson's disease. Mutations of *GNAL* cause primary torsion dystonia of muscle (45). Previous studies have mentioned the potential adaptation of fast-twitch muscle in echolocating mammals for ultrasound emission (46, 47). These test results and functional relevancies suggest that the genes found by ACEP may have experienced adaptive convergence for the ultrasound emission and perception function in echolocating mammals.

A number of different PLMs have been published during the progress of this study. For example, the ESM-2 model was pretrained on approximately 65 million unique protein sequences and outperforms earlier models (e.g., ESM-1b) at the same parameter scale in protein structure prediction. To explore the impact of different pretrained PLM backbones on the ACEP framework, we substituted the ESM-MSA-1b backbone by ESM-2, and trained the encoder–decoder network by mammalian sequence data likewise (*Materials and Methods*). Interestingly, while the ESM-2 $E_{g0}$ embeddings exhibited a similar level of phylogenetic informativeness as the original ESM-MSA-1b, the ESM-2 $E_g$ showed poorer phylogenetic informativeness than the ESM-MSA-1b $E_g$ (*SI Appendix*, Fig. S10*A*). We then conducted ACEP tests by ESM-2 $E_g$. Under cutoff of $P < 0.01$, we found that Prestin was not significant, with marginal empirical *P*-value (*SI Appendix*, Fig. S10*B*). Among the 1,305 significant genes based on cosine distances, the GO term "Sensory Perception of mechanical stimulus" was significantly enriched (*SI Appendix*, Fig. S10*C*), and the same is true among the 798 genes significant upon ACEP tests by Euclidean distances (*SI Appendix*, Fig. S10*D*). Among the significant genes under this term, there existed genes found by the original ESM-MSA-1b ACEP tests such as *LRP2*, *SCN9A,* and *CDH23*. Besides, many genes known to be related to echolocation (e.g., *OTOF*) (37), ultrasonic hearing (e.g., *PIEZO2*) (48) or hearing loss (e.g., *COCH*, *MINAR2*, *CLRN2*) (49–51) were included (*SI Appendix*, Fig. S10*E*). Thus, unsurprisingly, different PLM backbones may affect the ACEP test results, and further exploration in this direction is warranted.

Given that our ACEP tests reported multiple candidate genes in echolocating mammals, a natural question is whether existing methods can detect the same convergence events. As such a benchmark, we applied analyses according to the CCS method (15), searching for CCSs in TW and two lineages of EB, with nonecholocating bats and bovids as control groups (*Materials and Methods* and *SI Appendix*, Fig. S11*A*). Among the 28 sensory perception genes found by ACEP tests, CCSs were found in 14 genes (*SI Appendix*, Fig. S11*B*). Furthermore, since the ACEP test focuses on high-order features of proteins rather than site-level amino acid state convergence, we also recoded the 20 amino acids into groups with similar physicochemical states, and searched for CCSs with physicochemical state convergence (*Materials and Methods*). Under two different physicochemical group delimitations, we respectively observed 9 and 14 genes with CCSs among the 28 ACEP-identified sensory perception genes (*SI Appendix*, Fig. S11 *C* and *D*). Hence, current methods, even when considering site-level physicochemical features instead of individual amino acid states, can support some of the candidate genes found by ACEP test, but may not fully capture the potential convergence pattern of PLM-derived high-order protein features.

The ACEP test results rely on accurate simulation of neutral sequence evolution that matches the real evolution process only without adaptive convergence. Although simulation parameters including site-wise evolution rates were inferred from real data, Markov process simulation of individual sites may not fully generate the background stochastic convergence of high-order protein features as in the real data. As an alternative strategy, we designed a phylogenetic permulation (phylogeny-aware permutation) test inspired by the recently developed RERconverge expansion (52, 53) (*Materials and Methods*). In this test, we permutated the focal lineage label according to trait value simulation, and derived the empirical *P*-value of a relative distance (*RD*, embedding distance normalized by phylogenetic distance) from a null distribution formed by 1,000 permulations. The *RD* between two focal lineages was the minimum or mean of *RD*s between pairs of species in the two lineages (*SI Appendix*, Fig. S11*E*). Due to the substantial time consumption of the permulation step, we tested the 30 fragments from the 28 genes significant in ACEP tests. Ten (by minimum *RD*) or four (by mean *RD*) genes were significant by at least one type of embedding distance, including *SLC26A5*, *CIB2*, *CDH23*, *MYH9* (*SI Appendix*, Fig. S11*F*). Considering the small number of taxa with echolocation trait, the rate of binary trait change may not be estimated precisely, and the available combinations of permulated label assignment with matching topology might be few, potentially leading to low power of this test. Thus, our finding of significant genes by the permulation test provided conserved evidence supporting the previous ACEP results.

**ACEP Corresponds to Convergence of High-Order Features of Proteins.** Significant results by the ACEP test may reflect convergence of high-order function-related features of proteins. Nonetheless, information about these features is only implicitly encoded in the PLM embeddings. To explore the possible convergent features of the proteins found in the above analyses, we first investigated *SLC17A8* as an example. A previous study reported two site-level convergence events (V109I and R309K) among five echolocating mammal species in *SLC17A8* (38). First, we tested whether the convergence signal from ACEP was completely caused by site-level convergences, by masking the two convergent sites from the *SLC17A8* MSA when calculating embeddings, and then rerunning the ACEP test. Consequently, regardless of using all 13 available echolocating species in our dataset or using the five species as in the original study, ACEP tests were still significant ($P < 0.01$) for *SLC17A8* without the two sites (Fig. 5*A* and *SI Appendix*, Fig. S12 *A–C*). This indicates that protein features other than site-level convergence contribute to the putative functional convergence of *SLC17A8*.

Next, we continued to focus on 43 protein physicochemical features, such as proportions of each amino acid, hydrophobicity, isoelectric point, and net charge density (NCD) mentioned in previous studies (*SI Appendix*, Table S4) (54–56). Values of each feature in the *SLC17A8* orthologs of TW, EB, and bovid species were calculated. If there is no significant difference between the TW values and the EB values, and meanwhile both TW and EB values are significantly different from the bovid values, such features are considered to be convergent in echolocating mammals. For *SLC17A8*, we found six convergent features

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

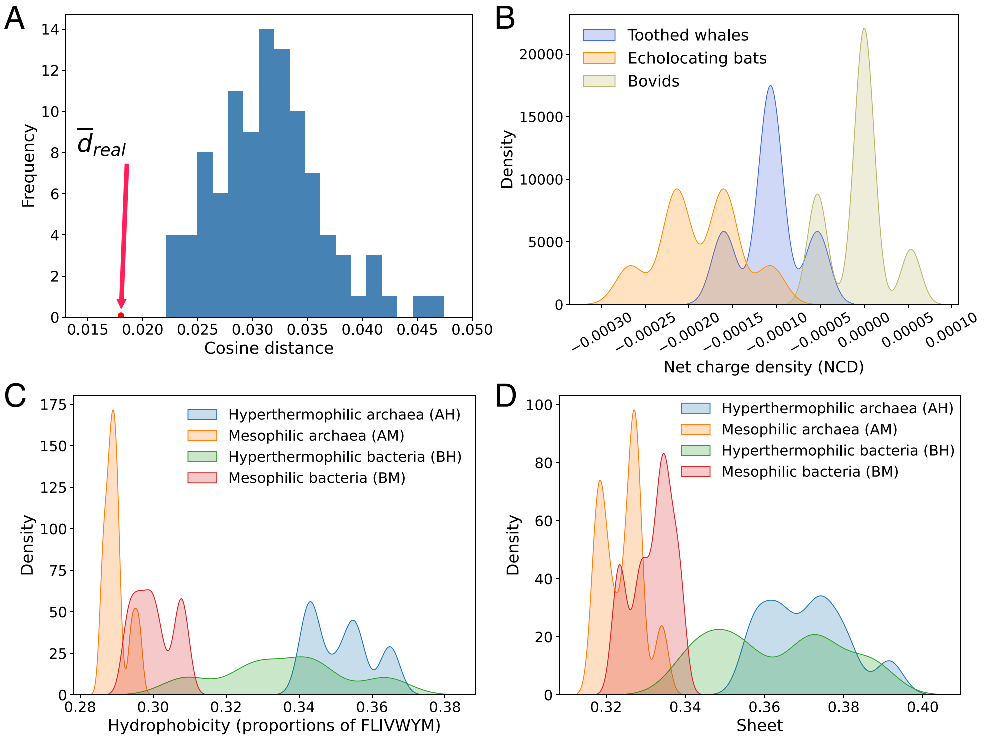

**Fig. 5.** Genes with PLM embedding convergence also show high-order physicochemical feature convergence. (*A*) ACEP test result of *SLC17A8* by cosine embedding distances masking two previously reported convergent sites (V109I and R309K) between five echolocating TW and eight EB, showing $P < 0.01$. The red arrow points to the $\overline{d_{real}}$ value marked by a red dot on the *X* axis, and the blue histogram is the distribution of mean distances derived from each of the 100 simulations. (*B*) Kernel density plots of *SLC17A8* protein net charge density distribution in three focal mammal groups involved in ACEP tests. (*C*) Kernel density plots of the COG0013 physicochemical feature "hydrophobicity (proportion of FLIVWYM)" distribution in four types of prokaryotes. (*D*) Kernel density plots of COG0013 physicochemical feature "sheet" distribution in four types of prokaryotes. In *B*–*D*, The KDE parameter bw_adjust was set to 0.5.

(Mann–Whitney *U* test, using $Q = 0.05$ as significance cutoff), including proportion of arginine (R), helix (proportion of VIYFWL), two hydrophobicity measurements (proportions of FLIV or FLIVWYM), net charge (proportion of K + R - D - E) and NCD (Fig. 5*B* and *SI Appendix*, Fig. S12*D*). Specifically, NCD has been proposed to affect protein mobility and dynamic changes by determining interactions with other cellular components (54, 57). Among the six convergent features, four remained significant after excluding the two convergent sites mentioned above (*Materials and Methods*), including NCD, net charge and proportion of arginine (Mann–Whitney *U* test, using $Q = 0.05$ as significance cutoff). The two hydrophobic features were not significant mainly due to subtle *Q*-value changes during multiple test correction, which was in turn caused by low power of the nonparametric *U* tests on a small number of species in each group. Actually, two other features (solvent accessibility surface area SASA, molecular weight) turned significant upon site exclusion. Meanwhile, the two convergent sites may indeed contribute significantly to the physicochemical convergence of the protein, which does not contradict effects of other sites reflected by our PLM embedding.

As a control, we repeated the ACEP test and feature convergence test in two paralogs of *SLC17A8*, i.e., *SLC17A6* and *SLC17A7*. All three genes encode vesicular glutamate transporters with complementary expression patterns in central neural system cells, and had similar lengths before (560 to 589 amino acids) and after (490 to 537 a.a.) sequence quality control. Conducting an ACEP test in *SLC17A6* and *SLC17A7* among 11 available echolocating species, we found neither paralog significant ($P >$ 0.02, *SI Appendix*, Fig. S12 *E* and *F*). Meanwhile, none of the six convergent features in *SLC17A8* showed a convergent pattern in control, except for hydrophobicity as proportions of FLIV in *SLC17A7*. These results consistently supported the relation between the significance of the ACEP tests and the convergence of high-order function-related features of proteins.

Finally, for the nine genes with significant convergent embedding patterns in the thermophilic prokaryotes case, we also calculated and compared the values of 43 physicochemical features in the AH, BH, AM, and BM groups. Taking COG0013 as an example, the feature values of hydrophobicity as proportions of FLIVWYM showed no significant difference between AH and BH (Mann–Whitney *U* test, $Q > 0.15$), while the AH values were higher than the AM values and BH values were higher than the BM values ($Q < 0.0007$, Fig. 5*C*). The same convergent patterns were obtained for the feature sheet in COG0013 ($Q > 0.65$ for AH-BH comparison, $Q < 0.0007$ for AH-AM and BH-BM, Fig. 5*D*). These two features showed the same patterns in COG0143 (*SI Appendix*, Fig. S12 *G* and *H*) and in most of the other genes with embedding convergence, together with some other physicochemical features (*SI Appendix*, Fig. S12 *I* and *J*). Intriguingly, different features tend to show significant AH-BH convergence in either most of the nine genes or few of them, suggesting some features contribute prevalently to the high-temperature adaptation of many thermophile proteins. These results agreed well with previous experimental findings that hydrophobic interactions contribute to protein stability and thermostable proteins enrich β-sheet structures (58, 59). Furthermore, almost all significant AH-BH feature convergences were observed when the features were simultaneously associated with principal components of the embeddings ($P < 1 \times 10^{-10}$, Hypergeometric test, *SI Appendix*, Fig. S12 *I* and *J*), suggesting that PLM embeddings were able to reflect the convergence of bona fide protein high-order features.

## Discussion

In this study, we harness the capacity of recently developed PLMs, to investigate adaptive convergence of high-order sequence features underlying convergent evolution of protein or organismal functions. We first trained a bottleneck encoder–decoder neural

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

network to derive fixed-length embeddings $E_g$ from PLM outputs of variable lengths, and we verified that the PLM-derived embeddings contain information of evolutionary relationship between sequences. Then, in three existing cases of individual proteins with functional convergence and another case of multiple conserved proteins in thermophilic bacteria and archaea, we confirmed high embedding similarity corresponds to functional similarity, regardless of sequence divergence at the site level. This implied the feasibility of quantifying the adaptive convergence of high-order protein features by PLM embeddings. We then developed an analysis pipeline, ACEP, to test whether gene PLM embeddings are more similar between focal taxa with functional convergence than simulated backgrounds. ACEP tests were significant on known adaptive convergent genes in echolocating mammals and CAM plants, demonstrating its power. Applying ACEP to a genome-wide search of adaptive convergent genes in echolocating mammals, we found significantly enriched functions such as Sensory Perception and additional candidate genes, such as *GSN* and *SCN9A*, with functional relevance and orthogonal support by positive selection tests. We further investigated possible high-order physicochemical features with adaptive convergence in candidate genes found by PLM embeddings, and were able to observe ACEP significance even when site-level convergence was excluded from the data. Overall, our analyses demonstrated that convergence of PLM embeddings can indicate adaptive convergence of high-order protein features.

While our ACEP pipeline showed promising results, the overall strategy warrants further exploration. Despite the many methods developed recently, the detection of adaptive convergence is challenging in that convergence can happen stochastically during neutral sequence evolution. In our ACEP tests on echolocating mammals, many genes unrelated to the focal convergent function were found significant. Moreover, when the sister lineage (i.e., bovids) was used as control, the number of significant genes was even higher than that in the focal pair of echolocating mammals. This is also observed in previous studies focusing on site-level convergence (12, 60), possibly due to more stochastic convergence as a result of faster sequence evolution rate in the bovids than in the TW. In the ACEP pipeline, the control sequences used for generating the null distance distribution were simulated according to site-wise rates estimated from real data, but still may not fully account for the random convergence of high-order protein features. Thus, the ACEP pipeline may pick up genes with stochastic feature convergence. As we propose the current pipeline as a proof of concept, we caution against asserting the ACEP results without further validation. In this study, we used functional enrichment tests to further validate the existence of candidate adaptive convergent genes among all ACEP significant genes. Meanwhile, we have also conducted the conservative permulation test, which is free of sequence simulation and account for phylogenetic dependency between lineages, and we confirmed the significant embedding similarity of multiple candidate genes.

Technically, the various existing PLMs differ in detailed training strategies and sizes of training datasets, which may lead to their different capacity on extracting evolutionary information. Besides, the method for deriving the fixed-length global embeddings may also affect pipeline performance. For example, the mean global embedding $E_{g0}$ seemed to capture sequence similarity between functionally convergent proteins in some cases, but failed in others, indicating the bottleneck-derived embedding $E_g$ was superior in capturing evolutionary patterns such as convergence. We also observed different flexibility of the bottleneck-derived embedding $E_g$ when trained by mammalian or plant protein sequences, and it seems that training on sequences of the corresponding taxa did not always lead to the most significant results. As the original PLM backbone was pretrained on very diverse proteins, to resolve specific sequence similarity within a relatively small taxonomic group, the bottleneck encoder–decoder may need to be trained specifically on capturing more subtle divergence between orthologous sequences. In this sense, the more closely related mammalian proteins may indeed serve as a better training data than plant proteins, which might explain why the mammal-sequence-trained $E_g$ can capture plant protein convergence while the reverse is not true. We have also observed that using another backbone PLM (ESM-2) for ACEP tests obtained relevant but not identical candidate genes as the original backbone ESM-MSA-1b, also identifying functionally related genes.

Thus, substituting the backbone PLM and adjusting the global embedding derivation process may further improve the efficiency of this strategy. For example, some existing PLMs were pretrained by explicitly combining sequence data with higher-order feature data such as protein structures (61–63). Since protein structure may be a more direct indicator of the function than sequences, using such structure-aware models as backbone may improve the power of the ACEP framework. Moreover, considering the high efficiency of protein structure prediction by deep learning methods (26, 64), directly evaluating structure convergence may also be a viable approach, although protein structure information cannot reflect all high-order features. With the PLM embeddings obtained, different distance metrics could also lead to somewhat different ACEP test results, as shown by the cosine and Euclidean distances in this study. Different distance metrics may reflect different aspects of protein feature similarities in the embedding space, which may warrant further investigation. Currently, we propose to explore results based on both distances. Biologically, the ACEP results do not explicitly indicate the mechanistic features with putative adaptive convergence. Hence, interpretable machine learning methods may be applied in the future, to locate the exact protein features underlying adaptive functional convergence. Another limitation of the current ACEP pipeline is that, due to distance-based test scheme, extending the pipeline to cases with more than two focal taxa would be challenging. Future methodological development without distance comparison, e.g. directly evaluating the association between the embeddings and the phenotypes, may improve the flexibility.

Due to complexity of biological GPM, the macroscopic or molecular functional convergence between individual lineages of species may well be caused by convergent changes at different facets of sequence features, i.e., at the same site, at different sites in the same gene, or even in different genes of the same pathway (65). In recent years, efforts have been made to characterize molecular convergence beyond site-level, including gene paralog retention (66), regulation of gene expression (67, 68), or sequence evolution rate shift (5, 69). Defining molecular convergence beyond site-level identity relies on effective extraction of high-order sequence features. Our results indicated that the simple sequence similarity measures $p$ distances reflected less functional convergence than the physicochemically informative BLOSUM62 scores, which in turn were less informative than the PLM embeddings. Thus, our application of PLMs for this purpose substantially broadens the scope of detecting adaptive molecular convergence, by enabling high-throughput extraction of high-order protein features, thus contributing to more comprehensive understanding of the molecular mechanism underlying organismal functional convergence.

Moreover, resolving the high-dimensional, complex GPM has long been a major challenge in molecular evolution studies. Our findings that PLM-extracted high-order features reflect functional

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

similarity with no site-level convergence suggest that the mapping between these PLM features to functions may be simpler, with less complexity such as epistasis. Hence, PLM-derived sequence features may serve as an intermediate layer in the GPM, facilitating its theoretical understanding and practical application such as protein design. With the promising capacity of deep learning models, innovative strategies may be developed to help us elucidate the genetic basis of phenotypic and functional evolution.

## Materials and Methods

**Protein Sequence Data.** For the example of echolocation in mammals, orthologous protein sequence alignments for 14,509 genes were downloaded from OrthoMaM v10c (70). The corresponding tree topology was derived from TimeTree 5 (71). The protein alignments were first filtered by containing less than 5% gaps and containing at least one species from each of the two foreground echolocating mammal lineages. The alignments with more than 100 sites after removing gapped positions were retained and split into segments shorter than 1,024 according to PLM input limit. For the example of CAM plants, we sampled 75 species from JGI Phytozome v13 (72) including four CAM species. All transcripts were downloaded from available versions of genomes for the 75 species. The phylogeny was derived from the R package V.PhyloMaker (73). OrthoFinder v2.5.5 (74) was used to derive orthogroups of genes. The orthologs of focal genes were identified by corresponding anchor sequences and then aligned by MAFFT v7.505 (75). For the examples of vertebrate Hbs, serine protease toxins, acylation in proteases, and ferrous iron transporters, the sequences were derived following the source information in the corresponding original studies. For the example of proteins in thermophilic prokaryotes, orthologous sequence alignments were derived from the COG database (76). The prokaryote species were chosen according to their growth temperature information in the BacDive database (77). See the *SI Appendix, Supporting Text* for more details of each dataset.

**Protein Language Model and Bottleneck Encoder Training.** The local embedding $E_l$ of each protein sequence with size $L \times 768$ was derived by extracting the output representation from the 12th Transformer layer of ESM-MSA-1b. The mean global embedding $E_{g0}$ of size 768 was calculated by averaging the $E_l$ across the length ($L$) dimension. The global embedding $E_g$ was the encoder output of an encoder–decoder network transforming the padded $1{,}024 \times 768$ $E_l$ to size 300 and then expanding it back to a $1{,}024 \times 768$ tensor, which was decoded as predicted protein sequence logits by an MLMHead module. The network was trained by cross entropy loss on 37,998 mammal protein sequences with Adam optimizer. The backbone model ESM-MSA-1b was substituted by ESM-2, or the training data were switched to 39,519 plant protein sequences to train the other encoder–decoder networks mentioned in the results. See the *SI Appendix, Supporting Text* for more details of the model and training processes.

**Sequence Evolution Inference, Simulation, and Positive Selection Detection.** Phylogeny reconstructions in the three cases of individual genes and thermophilic prokaryotes were conducted by IQ-TREE 2.2.5 (78) under default settings and validated by Bayesian inference in the hemoglobin case. Negative control simulation was realized by alisim following IQ-TREE parameter inference. In the ACEP test, simulation parameters and ancestral sequences were inferred by PAML 4.9j (79) from each real MSA, and sequences were simulated by the Evosimz package in Zou et al. (80). Positive selection detection was conducted using the BUSTED program in Hyphy 2.5.8 (81). See the *SI Appendix, Supporting Text* for more details of the simulation and positive selection detection.

**Embedding Distances, Other Sequence Similarities, and ACEP Test.** The cosine embedding distance between two global embedding $\boldsymbol{x}$ and $\boldsymbol{y}$ was calculated as $1 - (x \bullet y)/(|x| \bullet |y|)$. The Euclidean embedding distance was calculated as $\sqrt[2]{(x-y)^T(x-y)}$. For a real or simulated MSA, the mean embedding distance $\overline{d}$ was calculated as the mean of all pairwise distance between embeddings of two groups of focal species (e.g., EB and TW). For each gene, the real mean embedding distance $\overline{d_{real}}$ and 100 simulated mean distances ($\overline{d_1}, \overline{d_2}, \ldots, \overline{d_{100}}$) derived from 100 replicate simulations were calculated. An empirical $P$-value for the gene can be calculated as the proportion of $\overline{d_i}$ s equal to or smaller than $\overline{d_{real}}$. In our analysis, genes with empirical $P$-value smaller than 0.01, i.e., $\overline{d_{real}}$ is smaller than any simulated $\overline{d_i}$, were considered candidates of adaptive sequence convergence.

To calculate $p$ distances and BLOSUM62 scores between a group of orthologous sequences, all orthologs were first aligned by MAFFT v7.505 (75) using the linsi algorithm. For $p$ distance calculation between a pair of sequences, sites with gaps in both sequences were removed, and then the $p$ distance was calculated as number of sites with different states divided by the total number of aligned sites. The BLOSUM62 scores were calculated as in the default blastp (82) raw score calculation based on BLOSUM62 matrix, with −11 for gap opening and −1 for gap extension.

**Functional Enrichment Test.** Functional enrichment tests on sets of significant genes in ACEP tests were conducted by using the online tool MetaScape version 3.5.20240101 (https://metascape.org/gp/index.html). The whole sets of genes used to conduct ACEP tests after filtering were used as background in functional enrichment tests, with gene set size 11,559 for echolocating mammals and 11,479 for control. For the "Analysis as species" setting, we used the default "*H. sapiens*."

**Application of the CCS Test and Alternative Permulation Test.** Analogous to the original CCS method, we first assigned the two lineages of EB and the TW as foreground groups, and assigned the nonecholocating flying foxes, the bovids plus the white-tailed deer, the guinea pig, and human as background, i.e., outgroups (*SI Appendix*, Fig. S11*A*). An amino acid site in an alignment is defined as CCS if: (1) the outgroups share the same amino acid state T; (2) more than 2/3 of the foreground species share the same amino acid state T′ different from T; (3) the state T′ can be observed in both EB and TW. For the modified CCS method regarding site-level physicochemical similarity, we assigned amino acids into physicochemically similar groups, and change the definition of T and T′ in the above criteria from specific amino acid state to specific physicochemical group. The two different physicochemical group delimitations (green tables in *SI Appendix*, Fig. S11 *C* and *D*) were according to previous studies (83, 84).

In the alternative test based on trait value permulation, the permulations were conducted using scripts in the RERconverge GitHub repository (https://github.com/nclark-lab/RERconverge). For each gene, branch lengths of a gene tree were inferred by PAML as previously described, with its topology fixed as the species tree. This tree and the foreground echolocating species labels were then input into the getPermsBinary() function of PermulationFuncs.R to conduct permulations. For each of the 1,000 permulations, the two monophyletic groups in the output were used as the permulated "foreground species." An *RD* was calculated for each pair of species from the two groups as the cosine or Euclidean $E_g$ embedding distance divided by the phylogenetic distance on the gene tree. The minimum or mean *RD* was retained, and the 1,000 minimum or mean *RD*s were lumped to form the permulated null distribution. The real minimum or mean *RD* between two echolocating groups were compared to the null distribution to get an empirical $P$-value (*SI Appendix*, Fig. S11*E*).

**Excluding Convergent Sites in *SLC17A8*.** In the ACEP test, the two convergent sites in *SLC17A8* were masked by the <mask> token when calculating $E_g$ embeddings. When calculating values of the 43 physicochemical features, to keep the protein sequence length unchanged, we changed the amino acid states at position 109 and 309 to X, except for the following cases. When calculating "instability index" and "gravy," only 20 amino acid states are allowed in the sequence. When calculating "molecular weight," changing amino acid to X is equivalent to removing the site. Thus, when calculating the above three features and the "SASA," "NCD" features which depend on "molecular weight," site 109 and 309 were directly removed from the sequence.

**Data, Materials, and Software Availability.** Source codes, model weights, and examples of the ACEP pipeline are available at https://huggingface.co/NEO699700/ACEP. Previously published data were used for this work (70, 72). All other data are included in the article and/or *SI Appendix*.

**ACKNOWLEDGMENTS.** We thank Dr. Jianzhi Zhang, Dr. Xionglei He, Dr. Weiwei Zhai, Shanshan Chen, Han Liu, Tianhan Wang, and three anonymous reviewers for valuable comments. This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDA0460405),

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.

the National Natural Science Foundation of China (grant no. 92370134) and institutional grant to Z.Z. from the Key Laboratory of Zoological Systematics and Evolution and the Institute of Zoology, Chinese Academy of Sciences.

Author affiliations: [a]State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [b]University of Chinese Academy of Sciences, Beijing 100049, China; and [c]Microsoft Canada Development Centre, Vancouver, BC V5C 1G1, Canada

1. P. T. Madsen, A. Surlykke, Functional convergence in bat and toothed whale biosonars. *Physiology (Bethesda)* **28**, 276–283 (2013).
2. J. F. Storz, Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016).
3. S. F. Greenbury, A. A. Louis, S. E. Ahnert, The structure of genotype-phenotype maps makes fitness landscapes navigable. *Nat. Ecol. Evol.* **6**, 1742–1752 (2022).
4. T. A. Castoe *et al.*, Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8986–8991 (2009).
5. M. Chikina, J. D. Robinson, N. L. Clark, Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol. Biol. Evol.* **33**, 2182–2192 (2016).
6. A. D. Foote *et al.*, Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
7. K. Fukushima, D. D. Pollock, Detecting macroevolutionary genotype–phenotype associations using error-corrected rates of protein convergence. *Nat. Ecol. Evol.* **7**, 155–170 (2023).
8. Z. He *et al.*, Convergent adaptation of the genomes of woody plants at the land–sea interface. *Natl. Sci. Rev.* **7**, 978–993 (2020).
9. Y. Li, Z. Liu, P. Shi, J. Zhang, The hearing gene *Prestin* unites echolocating bats and whales. *Curr. Biol.* **20**, R55–R56 (2010).
10. C. Natarajan *et al.*, Convergent evolution of hemoglobin function in high-altitude Andean waterfowl involves limited parallelism at the molecular sequence level. *PLoS Genet.* **11**, e1005681 (2015).
11. J. Parker *et al.*, Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).
12. G. W. Thomas, M. W. Hahn, Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. *Mol. Biol. Evol.* **32**, 1232–1236 (2015).
13. J. Zhang, Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat. Genet.* **38**, 819–823 (2006).
14. Z. Liu, F. Y. Qi, X. Zhou, H. Q. Ren, P. Shi, Parallel sites implicate functional convergence of the hearing gene *prestin* among echolocating mammals. *Mol. Biol. Evol.* **31**, 2415–2424 (2014).
15. S. Xu *et al.*, Genome-wide convergence during evolution of mangroves from woody plants. *Mol. Biol. Evol.* **34**, 1008–1015 (2017).
16. J. Zhang, S. Kumar, Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* **14**, 527–536 (1997).
17. Z. Zou, J. Zhang, Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol. Biol. Evol.* **32**, 2085–2096 (2015).
18. C. Rey, L. Guéguen, M. Sémon, B. Boussau, Accurate detection of convergent amino-acid evolution with PCOC. *Mol. Biol. Evol.* **35**, 2296–2306 (2018).
19. X. Farre *et al.*, Comparative analysis of mammal genomes unveils key genomic variability for human life span. *Mol. Biol. Evol.* **38**, 4948–4961 (2021).
20. E. B. Rosenblum, C. E. Parent, E. E. Brandt, The molecular basis of phenotypic convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**, 203–226 (2014).
21. C. Natarajan *et al.*, Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354**, 336–339 (2016).
22. Y. T. Aminetzach, J. R. Srouji, C. Y. Kong, H. E. Hoekstra, Convergent evolution of novel protein function in shrew and lizard venom. *Curr. Biol.* **19**, 1925–1931 (2009).
23. W. F. C. Rodrigues, A. B. P. Lisboa, J. E. Lima, F. K. Ricachenevsky, L. E. Del-Bem, Ferrous iron uptake via IRT1/ZIP evolved at least twice in green plants. *New Phytol.* **237**, 1951–1961 (2023).
24. A. R. Buller, C. A. Townsend, Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E653–E661 (2013).
25. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
26. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
27. E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, A. Madani, ProGen2: Exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978.e63 (2023).
28. N. Ferruz, S. Schmidt, B. Hocker, ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
29. R. Rao *et al.*, MSA transformer. bioXriv [Preprint] (2021). https://doi.org/10.1101/2021.02.12.430858 (Accessed 21 March 2023).
30. N. S. Detlefsen, S. Hauberg, W. Boomsma, Learning meaningful representations of protein sequences. *Nat. Commun.* **13**, 1914 (2022).
31. F. G. Hoffmann, J. C. Opazo, J. F. Storz, Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14274–14279 (2010).
32. P. Lopez-Garcia, Y. Zivanovic, P. Deschamps, D. Moreira, Bacterial gene import and mesophilic adaptation in archaea. *Nat. Rev. Microbiol.* **13**, 447–456 (2015).
33. S. V. Venev, K. B. Zeldovich, Thermophilic adaptation in prokaryotes is constrained by metabolic costs of proteostasis. *Mol. Biol. Evol.* **35**, 211–224 (2018).
34. M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2015).
35. X. Yang *et al.*, The *Kalanchoë* genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nat. Commun.* **8**, 1899 (2017).
36. K. T. Davies, J. A. Cotton, J. D. Kirwan, E. C. Teeling, S. J. Rossiter, Parallel signatures of sequence evolution among hearing genes in echolocating mammals: An emerging model of genetic convergence. *Heredity (Edinb)* **108**, 480–489 (2012).
37. Y. Y. Shen, L. Liang, G. S. Li, R. W. Murphy, Y. P. Zhang, Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genet.* **8**, e1002788 (2012).
38. A. Marcovitz *et al.*, A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21094–21103 (2019).
39. A. P. J. Giese *et al.*, CIB2 interacts with TMC1 and TMC2 and is essential for mechanotransduction in auditory hair cells. *Nat. Commun.* **8**, 43 (2017).
40. Y. Wang *et al.*, Loss of CIB2 causes profound hearing loss and abolishes mechanoelectrical transduction in mice. *Front. Mol. Neurosci.* **10**, 401 (2017).
41. A. K. Lalwani *et al.*, Human nonsyndromic hereditary deafness DFNA17 is due to a mutation in nonmuscle myosin *MYH9*. *Am. J. Hum. Genet.* **67**, 1121–1128 (2000).
42. J. Ruel *et al.*, Impairment of *SLC17A8* encoding vesicular glutamate transporter-3, VGLUT3, underlies nonsyndromic deafness DFNA25 and inner hair cell dysfunction in null mice. *Am. J. Hum. Genet.* **83**, 278–292 (2008).
43. J. Yuan *et al.*, Hereditary sensory and autonomic neuropathy type IID caused by an *SCN9A* mutation. *Neurology* **80**, 1641–1649 (2013).
44. H. A. Schmitt *et al.*, Personalized proteomics for precision diagnostics in hearing loss: Disease-specific analysis of human perilymph by mass spectrometry. *ACS Omega* **6**, 21241–21254 (2021).
45. T. Fuchs *et al.*, Mutations in *GNAL* cause primary torsion dystonia. *Nat. Genet.* **45**, 88–92 (2013).
46. J. H. Lee *et al.*, Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. *Sci. Adv.* **4**, eaat9660 (2018).
47. C. P. Elemans, A. F. Mead, L. Jakobsen, J. M. Ratcliffe, Superfast muscles set maximum call rate in echolocating bats. *Science* **333**, 1885–1888 (2011).
48. J. Li *et al.*, PIEZO2 mediates ultrasonic hearing via cochlear outer hair cells in mice. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101207118 (2021).
49. N. Danial-Farran *et al.*, Homozygote loss-of-function variants in the human *COCH* gene underlie hearing loss. *Eur. J. Hum. Genet.* **29**, 338–342 (2021).
50. G. Bademci *et al.*, Mutations in *MINAR2* encoding membrane integral NOTCH2-associated receptor 2 cause deafness in humans and mice. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204084119 (2022).
51. F. Ahmad *et al.*, A novel *CLRN2* variant: Expanding the mutation spectrum and its critical role in isolated hearing impairment. *Genes Genom.* **47**, 417–423 (2024).
52. E. Saputra, A. Kowalczyk, L. Cusick, N. Clark, M. Chikina, Phylogenetic permulations: A statistically rigorous approach to measure confidence in associations in a phylogenetic context. *Mol. Biol. Evol.* **38**, 3004–3021 (2021).
53. R. Redlich *et al.*, RERconverge expansion: Using relative evolutionary rates to study complex categorical trait evolution. *Mol. Biol. Evol.* **41**, msae210 (2024).
54. E. V. Estrada, M. Oliveberg, Physicochemical classification of organisms. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2122957119 (2022).
55. D. Repecka *et al.*, Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
56. J. R. Lobry, C. Gautier, Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* **22**, 3174–3180 (1994).
57. X. Mu *et al.*, Physicochemical code for quinary protein interactions in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4556–E4563 (2017).
58. P. Leuenberger *et al.*, Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* **355**, eaai7825 (2017).
59. C. N. Pace *et al.*, Contribution of hydrophobic interactions to protein stability. *J. Mol. Biol.* **408**, 514–528 (2011).
60. Z. Zou, J. Zhang, No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241 (2015).
61. A. X. Lu *et al.*, Tokenized and continuous embedding compressions of protein sequence and structure. bioXriv [Preprint] (2024). https://doi.org/10.1101/2024.08.06.606920 (Accessed 30 January 2025).
62. J. Su *et al.*, SaProt: Protein language modeling with structure-aware vocabulary. bioXriv [Preprint] (2024). https://doi.org/10.1101/2023.10.01.560349 (Accessed 5 June 2024).
63. R. Ma *et al.*, SCOP: A sequence-structure contrast-aware framework for protein function prediction. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2411.11366 (Accessed 29 January 2025).
64. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
65. T. B. Sackton, N. Clark, Convergent evolution in the genomics era: New insights and directions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **374**, 20190102 (2019).
66. L. G. Nagy *et al.*, Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat. Commun.* **5**, 4471 (2014).
67. T. B. Sackton *et al.*, Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**, 74–78 (2019).
68. J. R. Gallant *et al.*, Genomic basis for the convergent evolution of electric organs. *Science* **344**, 1522–1525 (2014).
69. Z. Hu, T. B. Sackton, S. V. Edwards, J. S. Liu, Bayesian detection of convergent rate changes of conserved noncoding elements on phylogenetic trees. *Mol. Biol. Evol.* **36**, 1086–1100 (2019).
70. C. Scornavacca *et al.*, OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol. Biol. Evol.* **36**, 861–862 (2019).
71. S. Kumar *et al.*, TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
72. D. M. Goodstein *et al.*, Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
73. Y. Jin, H. Qian, V.PhyloMaker: An r package that can generate very large phylogenies for vascular plants. *Ecography* **42**, 1353–1359 (2019).
74. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
75. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
76. M. Y. Galperin *et al.*, COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
77. L. C. Reimer *et al.*, BacDive in 2022: The knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.* **50**, D741–D746 (2022).
78. B. Q. Minh *et al.*, Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
79. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
80. Z. Zou, H. Zhang, Y. Guan, J. Zhang, Deep residual neural networks resolve quartet molecular phylogenies. *Mol. Biol. Evol.* **37**, 1495–1507 (2020).
81. B. Murrell *et al.*, Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–1371 (2015).
82. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
83. L. Yang, J. F. Xia, J. Gui, Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept. Lett.* **17**, 1085–1090 (2010).
84. S. Chaurasia, J. Y. Dutheil, The structural determinants of intra-protein compensatory substitutions. *Mol. Biol. Evol.* **39**, msac063 (2022).

Downloaded from https://www.pnas.org by Zhengting Zou on September 23, 2025 from IP address 124.16.149.119.