nature methods

Article

https://doi.org/10.1038/s41592-026-03028-7

Quantifying uncertainty in protein representations across models and tasks

Received: 15 May 2025

Accepted: 11 February 2026

Published online: 1 April 2026

Check for updates

R. Prabakaran [1] & Yana Bromberg [1,2]

Biomolecular embeddings serve as efficient representations of sequence and structure, enabling tasks such as similarity searches, structure and function prediction and estimation of biophysical properties. However, relying on embeddings without assessing their ability to accurately represent biomolecules is a critical flaw—akin to using a scalpel in surgery without verifying its sharpness. Here we propose a means to evaluate the capacity of protein language models to encode biologically meaningful information. For each protein, representation uncertainty is scored as the fraction of non-biological 'synthetic' sequences among its nearest neighbors in latent space. Our analysis reveals that low-quality embeddings often fail to capture meaningful biology, displaying vector properties indistinguishable from those of randomly generated sequences. Our model-agnostic scoring framework is, to our knowledge, the first to quantify protein sequence embedding reliability. It enables embedding screening prior to downstream applications and inferences, significantly improving their reliability. We propose that embedding evaluation should be undertaken for other uses of language models in science as well.

Language models (LMs), originally developed for natural language processing (NLP)[1], are increasingly accepted as the preferred in silico representation of the primary and higher-order structures of protein, DNA and RNA[2–5]. Their ability to learn an encoding that captures many aspects of a given biomolecule from simple amino or nucleic acid sequence has made them a promising tool for deriving biological insights[6–11]. LMs encode a biomolecule as an embedding—that is, a sequence of numbers representing a point in a multidimensional latent space. Embeddings serve as powerful computational proxies for facilitating a range of downstream tasks, such as similarity searches, structural and functional annotations and prediction of biomolecule properties[12–15]. For instance, embeddings from protein language models (pLMs) have been used to predict protein function, mutation effect and subcellular localization, achieving performance that rivals or surpasses traditional methods[13,14,16–19]. Additionally, fine-tuning pretrained pLMs has been shown to enhance predictions across multiple additional biological tasks, underscoring the versatility of these models[2,20,21].

Despite the advantages of embeddings as biomolecular representations, the reliability or confidence of an embedding remains largely unquestioned. Unlike most machine-learning-based predictions that have a corresponding prediction probability/reliability score, a given embedding is not questioned as a representation of a protein any more than a protein sequence would be.

Embeddings are low-dimensional representation of biomolecules in the latent space of the LM, with each vector element serving as a coordinate in the map of this space. Coordinates are learned to encode the training data while minimizing the loss associated with the training tasks[22]. The model's uncertainty or confidence in an embedding originates from the same sources as any of its predictions—the LM's training process, optimized to reach a computationally feasible solution that balances task performance within cost and time constraints, rather than achieving complete learning or a globally optimal representation. Put simply, the latent space of a model is just one of many possible optimal mappings for the given training dataset and the training objective. Moreover, datasets may not comprehensively capture the full sequence space—a limitation that is, arguably, even more obvious for protein sequences than for human languages. As a result, each protein's projection into the latent space carries an inherent uncertainty,

[1]Department of Biology, Emory University, Atlanta, GA, USA. [2]Department of Computer Science, Emory University, Atlanta, GA, USA.
e-mail: prabakaran@emory.edu; yana.bromberg@emory.edu

https://doi.org/10.1038/s41592-026-03028-7

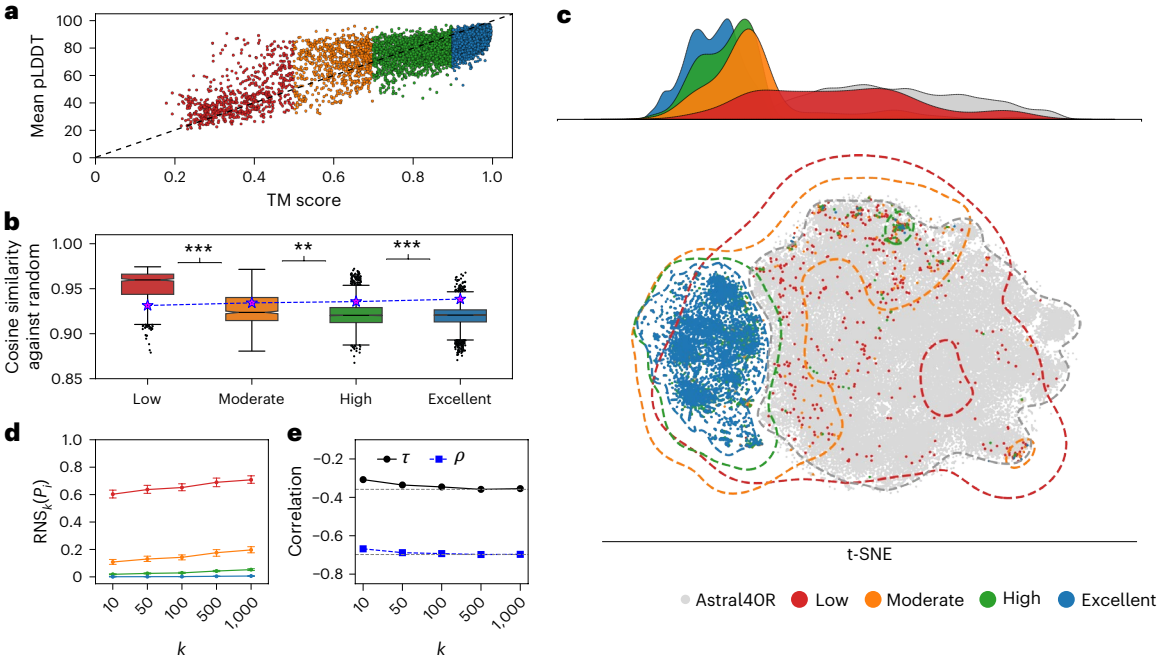


Fig. 1 | Protein structure prediction quality as a function of embedding certainty. Across **a**–**d**, color indicates the prediction confidence levels, judged by TM scores of ESM-2-predicted versus experimentally determined structures of Astral40 domains (n = 14,711): excellent (blue, TM > 0.9, n = 10,251), high (green, 0.7 < TM ≤ 0.9, n = 3,040), moderate (orange, 0.5 < TM ≤ 0.7, n = 768) and low (red, TM ≤ 0.5, n = 652). **a**, ESM-2-predicted structure quality (mean pLDDT, y axis) is, as expected, correlated with TM scores (x axis) of alignments of predicted versus experimental structures across Astral40 domains. **b**, Average cosine similarity (y axis) of Astral40's ESM-2 embeddings to a set of randomly generated, biologically irrelevant sequences in Astral40R (n = 73,555) differs across structure prediction confidence levels (x axis)—that is, proteins with poor prediction confidence (red box) tend to have more similarity to random embeddings than high-confidence predictions (green and blue). Blue circles are outliers beyond 1.5× IQR from the first and third quartiles of distribution. Statistical significance—$^{***}P \le 1 \times 10^{-3}$, $^{**}P \le 1 \times 10^{-2}$ and 'NS' otherwise, between low versus moderate (6.55×10^{-5}), moderate versus high (4.29×10^{-3}) and high versus excellent (9.35×10^{-50})—was assessed using two-sided Mann–Whitney U-test. Purple stars denote the mean cosine similarity of the Astral40R embeddings within each set (box), further highlighting the distinction between low-quality and high-quality embeddings versus random. **c**, Two-dimensional t-SNE projections of ESM-2 embeddings for Astral40 and Astral40R (gray) illustrate the specifics of the overlap of the two sets. That is, low-scoring protein embeddings (red dots and density lines) fall into the latent space also covered by random sequences (gray). At the same time, the latent space of excellent-scoring embeddings (blue) is nearly disjoint from the random space (gray). **d**, RNS (y axis), computed across varying values of k (k nearest neighbors, x axis), effectively discriminates ESM-2 embeddings corresponding to low-confidence structures (TM ≤ 0.5, red line) from higher-quality ones (all others). Error bars indicate the 95% confidence interval of mean RNS, derived through 100× bootstrapping. **e**, RNS is moderately inversely correlated with the TM score, reaching peak correlation (Kendall's $\tau = -0.36$, black line; Pearson's $\rho = -0.70$, blue line, at k = 500) for higher 'k'. NS, not significant.

raising the question of whether embeddings consistently place every protein in a biologically meaningful position—for example, one where functionally or evolutionarily related sequences remain nearby.

The uncertainty of protein embedding propagates to downstream tasks. Consider a pLM that can generate embeddings to serve as representations of protein sequences to enable downstream tasks such as predicting protein subcellular localization, T_s, or transferring functional annotations between proteins based on their embedding similarity, T_f[14,20,21]. Most pLMs do not provide an explicit measure of a model's confidence or uncertainty, $U(P_1)$, associated with an embedding, E_1, for a given protein P_1. Devoid of a score to indicate the biological information content in an embedding, it is difficult to prescreen embeddings or diagnose sources of error in downstream tasks like T_s and T_f.

Traditional sequence alignment-based approaches offer a biologically interpretable foundation for inference placed on top of their captured information, leveraging the level of evolutionary conservation as a well-established measure of confidence. pLM embeddings, by contrast, lack the standardized framework for evaluating their biological relevance and robustness across the complete range of sequences, seen or yet unseen. This issue is particularly clear when embeddings are derived from heterogeneous sequence distributions, as model biases or incomplete training data may lead to representations that fail to encode meaningful biological information. Without a mechanism to recognize low-confidence embeddings, erroneous predictions propagate across applications, misleading biological insights.

Establishing a systematic approach to assess embedding quality, independent of model architecture, training paradigms or downstream applications, would enable the prescreening of embeddings, improve downstream task performance and, more importantly, enhance the reliability and interpretability of pLM-based predictions[23]. Additionally, such an approach would provide a means to identify underrepresented and poorly learned regions of the protein space, guiding further experimental and computational exploration. Outside of biology—that is, in the general LM space—there have been a few efforts in this direction. For example, Tsitsulin et al.[24] proposed four complementary metrics to capture different aspects of unsupervised learning-generated embedding quality. Although no single metric was found to be universally optimal, coherence—that is, a measure of alignment between embedding dimensions to canonical basis vectors—and stable rank—that is, a soft rank proxy for the effective dimensionality of a set of representations—were shown to correlate strongly with downstream shallow model performance. Similarly, May et al.[25] proposed the eigenspace overlap score as a selection criterion for evaluating the quality of compressed (lower-dimensional) word embeddings. However, assessment of the uncertainty of learned protein representations in a biologically meaningful way has not yet been performed.

Table 1 | List of protein sequence embedders evaluated

Embedder	Model size	Architecture	Embedding dimension	Trained dataset	Reference
ESM-2	3B	Transformer	2,560	UniRef50	6
ProtT5	3B	T5 Transformer	1,024	UniRef50	34
ProtTrans T5-BFD[a]	3B	T5 Transformer	1,024	BFD	34
ESM-2 (650M)[b]	650M	Transformer	1,280	UniRef50	6
ESM-1v[b]	650M	Transformer	1,280	UniRef50	8
ESM-1b[a]	650M	Transformer	1,280	UniRef50	8
ESM-1[a]	650M	Transformer	1,280	UniRef50	8
Bepler & Berger[a]	93M	BiLSTM	121	SCOPe	45
PLUS-RNN[a]	59M	RNN	1,024	Pfam	46
Word2Vec[a]	6.4M	Skip-gram	512		47,48
FastText[a]	5.5M	Skip-gram	512		49
GloVe[a]	6.4M	Matrix Factorization	512		50

[a] Bio Embeddings was used to run these models[13]. [b] These models were used for the analysis of the human variant effect prediction.

In the present work, we illustrate a model-agnostic, empirical approach to measure the uncertainty associated with protein embeddings and to assess the biological relevance of said embeddings. By proposing methods for pLM embedding evaluation, we seek to standardize, expand and improve the usefulness of these models in moving biology forward.

Results

In this study, we systematically evaluate the quality of pLM-derived embeddings using the Astral40 dataset[26], a curated collection of protein structures. Our analysis reveals that inconsistencies in the generated embeddings stem from biased learning processes, where non-uniform representation of sequence space in the training set leads to suboptimal encoding. To address this challenge, we propose a rigorous screening framework to assess the biological relevance of embeddings prior to their application in computational analyses. By establishing quality control measures, we aim to enhance the reliability of pLMs for biological inference, ensuring their robustness in real-world applications.

Embeddings share the same defect as predictions

Few pLM designs provide a confidence score to assess the quality of their learning. One such model is Evolutionary Scale Modeling 2 (ESM-2)[6], which provides per-residue predicted local distance difference test (pLDDT) scores that describe how well the predicted structure is expected to align with a corresponding real protein structure[6]; we, thus, chose ESM-2 to illustrate embedding uncertainty. We collected embeddings, predicted structures and corresponding pLDDT scores by running ESM-2 on the Astral40 protein dataset, which consists of 14,711 unique Structural Classification of Proteins–extended (SCOPe) domains (Methods and Supplementary Table 1)[26–29]. We aligned[30] ESM2-predicted structures with respective Protein Data Bank (PDB) structures[27,31] to compute template modeling (TM) scores[30,32], a standard measure of goodness of structural alignment (Methods). We labeled the quality of predicted structures as very high, high, moderate and low quality if the TM scores fell within (0.9, 1), (0.7, 0.9), (0.5, 0.7) and (0, 0.5) bins, respectively (Fig. 1a). ESM-2's mean pLDDT scores for the predicted structures correlated well with TM scores, as expected of different measures of structural similarity (Pearson's $\rho = 0.84$, Kendall's $\tau = 0.56$).

We hypothesized that low-quality structures could derive from information-poor or ambiguous embeddings. To illustrate this, we created the Astral40R dataset that consists of five randomly shuffled sequences for every one of 14,711 unique domains in Astral40 (Methods and Supplementary Table 1). Note that, among the various methods for generating random sequence sets, we chose residue shuffling because we hypothesize that one of the advantages of pLMs over traditional statistical models arises from their ability to learn more than merely the amino acid composition of a sequence. Sequences in Astral40R are, thus, unique, randomly generated and exhibit no appreciable similarity to biological sequences in UniRef50 (<30% identity[33]).

We then computed the cosine and Euclidean similarities (equations (1) and (2)) of every domain/sequence in Astral40 (biological) against Astral40R (non-biological) datasets (Fig. 1b). Indeed, domain embeddings whose structures were accurately predicted were the least similar to embeddings of random sequences in Astral40R. As the quality of structures dropped, the similarity of domain embeddings to those of Astral40R increased; in other words, the embeddings corresponding to poorly predicted structures were less meaningful. This observation is neither unexpected nor unique to the ESM-2 model. We repeated our experiment with ProtT5 (ref. 34) and other LMs (Table 1) and observed similar patterns (Supplementary Figs. 1 and 2)—that is, Astral40 sequences with low-quality ESM-2-predicted structures were not meaningfully represented by ProtT5 either.

LMs are deep learning models, and the embeddings are as much model predictions as are the predicted structures. Predictions, however, are only as good as the model's understanding of the input; in other words, if an input does not fall within the model's scope, the prediction mirrors random inputs. We think that this observation is common among all LMs, but, unlike ESM-2, they are not trained to report a certainty (pLDDT-like) score that can be used to screen embeddings for future use.

Is there a 'junkyard' of embeddings?

We asked: Is there a subspace for low-quality embeddings in the latent spaces of pLMs where the low-quality, underlearned, noisy or less biomeaningful embeddings go? We used t-distributed stochastic neighbor embedding (t-SNE) to explore model embeddings of Astral40 and Astral40R (Methods). Indeed, we observed a distinct subspace for the random sequences of Astral40R in both ESM-2 and ProtT5 latent spaces (Fig. 1c and Supplementary Figs. 1 and 2).

We repeated this analysis for a selected set of models covering diverse architectures (Table 1). We observed similar patterns for ProtT5 and ESM family pLMs but not for Bepler and PLUS-RNN (Supplementary Figs. 1 and 2). The training tasks used for these latter two models differ from mask language modeling (MLM), the standard training protocol for most LMs. PLUS-RNN is a bidirectional recurrent neural network (RNN) model, trained with a combination of MLM and contrastive learning loss, scoring similar representation of protein

pairs from same Pfam families. Bepler is a bidirectional long short-term memory (LSTM) model trained on contrastive learning of global and local protein structural similarity. One could hypothesize the incorporation of contrastive loss configures a model's latent space in a manner different from that of LMs trained solely via MLM.

Traditional NLP algorithms Word2Vec, GloVe and FastText, which learn through the co-occurrence words or *k*-mers, failed to distance Astral40R embeddings from Astral40. Additionally, ESM-2 embeddings of Astral40 biological sequences with high-quality predicted structures (TM ≥ 0.9; Methods) occupied a clearly distinct space within the Astral40 range, whereas sequence embeddings with low-quality structures spread out into the Astral40R random sequence space.

What is a 'good' embedding?

Based on these observations, we hypothesized that, if the model failed to adequately learn its representation during training, a protein would be positioned close to biologically irrelevant sequences, such as Astral40R. As above, we define this region of poorly learned or uncharacterized protein embeddings as a 'junkyard'.

We further propose that the degree of overlap in latent space between a protein's nearest neighbors and the embeddings of non-biological sequences is inversely correlated with the model's confidence in the embedding—that is, the embedding quality. We formulate this relationship as random neighbor score (RNS) (RNS_k (P_1)), reflecting the number of non-biological sequence neighbors of a given protein P_1 in a given pLM's latent space (equation (5)). RNS is computed by measuring the fraction of randomly generated sequences among *k* nearest neighbors of a given embedding (equations (5) and (6)).

RNS is a measurable quantity reflecting the confidence of an embedding in a model's latent space. We tested this new measure on the Astral40 dataset in ESM-2's latent space (Fig. 1d). For ESM-2-predicted structures, RNS inversely correlated with the TM score of alignment to the PDB structures, with a maximum Pearson's correlation of −0.70 (Kendall's $\tau = -0.36$) across different values of *k*; for comparison, the TM score to ESM's pLDDT correlation was 0.84 ($\tau = 0.56$). Unlike the pLDDT score, however, which is an output of supervised learning, coupled with ESM-2's structure prediction architecture, RNS is model agnostic and (1) can be computed for any language model, (2) does not depend on the embedding downstream use and (3) can be further refined for the purposes of the data and tasks at hand. Note that, although computation of RNS is independent of downstream tasks, it can be applied for active learning of fine-tuning tasks[23].

RNS depends on the selection of the parameter *k*—that is, the number of nearest neighbors to consider. Higher *k* values are preferred, ensuring enhanced score stability in light of perturbation and outliers; however, *k* should be substantially smaller than the dataset size to ensure that the theoretical ideal RNS score equals zero. For example, we observed a $k > 100$ range to be appropriate for our work with datasets sized from 930 to 14,711 sequences (Fig. 1e and Supplementary Table 1).

Because RNS is computed as a fraction of random nearest neighbors, it also assumes that the biological—that is, non-random—protein set is sufficiently diverse in its role as a secondary anchor for filtering out uncertain embeddings. We had initially selected Astral40R for its assumed sufficient diversity, but, to further evaluate the impact of the random set selection, we substituted Astral40R in its position as a junkyard anchor by another randomized set, Proteome4R (comprising one shuffled sequence per protein from four complete proteomes; Methods). This substitution yielded similar results across analyses in this study; for instance, the correlation with ESM-2's TM score (−0.70) remained unchanged.

Embedding uncertainty and downstream performance

In its reliability scoring role, RNS is similar to perplexity (equations (7) and (8)) as a lower score indicates higher model confidence, whereas a higher score reflects uncertainty. However, unlike perplexity, which quantifies the probability predicted by the decoder and is, thus, tied to decoder design and training paradigm, RNS is independent of the decoder, deriving uncertainty directly from encoder embeddings. For example, ESM-2 is an encoder-only model with a shallow multilayer perceptron decoder, whereas ProtT5 adopts a full encoder–decoder architecture with two transformer modules. Consistent with this difference, the correlation coefficient between $RNS_{k=1,000}$ and pseudo-perplexity (equation (7)) computed for the Astral40 dataset was higher for ESM-2 ($\tau = 0.59$) than for ProtT5 ($\tau = 0.44$) (Fig. 2a,b). We saw similar results across different datasets (Supplementary Tables 2 and 3). These observations highlight the degree to which decoder complexity decouples perplexity and RNS.

To evaluate other possible measures of embedding quality, we compared RNS to Jensen–Shannon divergence, computed between the distribution of amino acid dimers or trimers, extracted from real protein sequence, versus a random uniform distribution (Methods and equation (9)). We found weak or no correlation of the two measures across all datasets used in this study (Fig. 2a,b and Supplementary Fig. 3). On the other hand, across datasets, RNS inversely correlated with the number of alignments matches in UniRef50 and UniRef90 (Supplementary Fig. 3), indicating that models learn protein families better than orphan sequences[35].

Model uncertainty is known to influence predictive performance of foundation and fine-tuned pLMs[6,23]. The magnitude of these effects depends on architecture-specific and task-specific biases that influence how encoder-derived embeddings are used by the classifier module. As a result, we do not expect a linear relationship between RNS and overall prediction performance. However, as representations of sequences with higher pretraining RNS (low quality) are unlikely to improve upon fine-tuning, our score could potentially serve as an indicator of a task-specific upper limit of prediction performance.

To assess whether RNS that quantifies embedding uncertainty, can serve as an indicator of predictive performance, we systematically compared RNS ($RNS_{k=1,000}$) with task accuracy across two representative fine-tuned task scenarios: (1) residue–residue contact prediction using ESM-2 on Astral40 and PDB23to24—a collection of recent, experimentally determined structures (Methods)—and (2) secondary structure prediction with ProtT5 on the same datasets.

Note that these two tasks differ considerably in design: ESM-2's residue–residue contact prediction is unsupervised—that is, derived from attention maps—whereas ProtT5's secondary structure prediction is supervised, with a fine-tuned classifier transforming residue-level embeddings into three-class (helix, extended strand and other) predictions. In all evaluations, proteins with higher RNS consistently exhibited lower predictive accuracy (Fig. 2c–f and Supplementary Figs. 4–6) irrespective of their sequence similarity to the PDB (Supplementary Fig. 7).

For ESM-2 residue–residue contact predictions for both Astral40 and PDB23to24, precision (equation (12)) decreased sharply (approximately 40%) for proteins with high RNS (Fig. 2c,d and Supplementary Figs. 4 and 5). The decline was most pronounced (>60%) for long-range contacts (residues ≥24 sequence positions apart), highlighting a stronger dependence of non-local over local interaction predictions on embedding quality.

For protein residues from both datasets, ProtT5's three-class secondary structure prediction quality (equation (16)) degraded in high-RNS proteins, with helices and extended states disproportionately affected relative to the 'other' class (Fig. 2e,f and Supplementary Fig. 6). The overall effect, however, was less severe than for contact prediction, possibly reflecting that (1) secondary structure assignment is primarily determined by local interactions, which are less sensitive to embedding uncertainty, and/or (2) RNS, derived from protein embeddings, mean-pooled from residues, may underrepresent fine-grained residue-level information.

Together, these observations reinforce the utility of RNS as a robust, architecture-agnostic embedding uncertainty measure,

https://doi.org/10.1038/s41592-026-03028-7

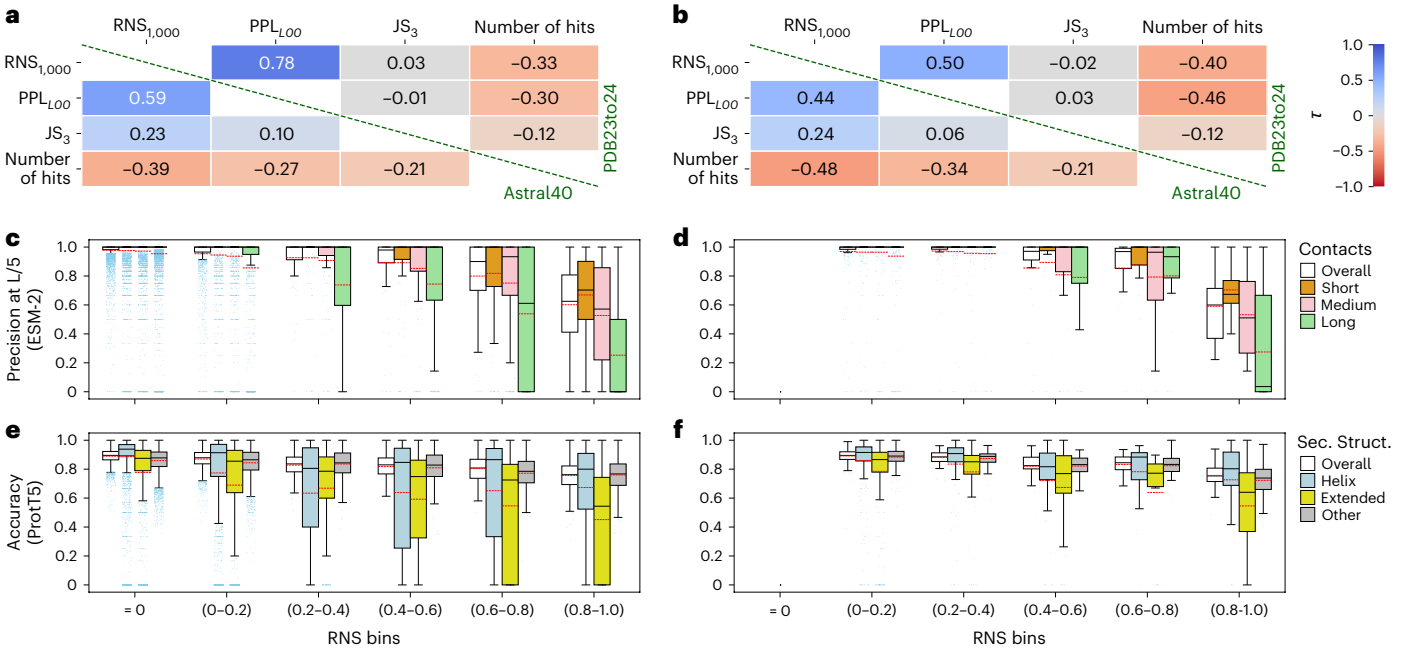


Fig. 2 | Higher RNS is associated with greater uncertainty and lower downstream prediction accuracy. For ESM-2 (**a**) and ProtT5 (**b**), Kendall's rank correlations (τ) between RNS and other uncertainty indicators (leave-one-out pseudo-perplexity, Jensen–Shannon divergence and number of sequence matches in UniRef; Methods) are shown for the embeddings of Astral40 (below diagonal, n = 14,711) and PDB23to24 (above diagonal, n = 663) protein sets. For example, RNS and pseudo-perplexity are somewhat correlated (medium blue), whereas the RNS and the number of alignment hits are somewhat inversely correlated (medium red). Prediction performance for contact prediction with ESM-2 (**c**,**d**) is reported as precision at L/5 (y axis). Colored bars represent performance across all contact classes combined (white, n = 19 M and 1.4 M) and separately for short-range (orange, n = 3.7 M and 251 k), medium-range (pink, n = 3.27 M and 2.21 k) and long-range (light green, n = 12.5 M and 935 k) contacts. Performance for secondary structure prediction with ProtT5 (**e**,**f**) is reported as Q3 accuracy (y axis). Colored bars represent performance across all residues' (n = 2.7 M and 192 k) secondary structure classes combined (white) and separately for helix (light blue), extended strand (yellow) and other (gray) classes. Results are for Astral40 (**c**,**e**) and PDB23to24 (**d**,**f**) proteins binned by RNS intervals (x axis). In **c**–**f**, the red dashed lines in box plots indicate the means of distributions; black lines are medians; and blue dots represent statistical outliers beyond 1.5× IQR from the first and third quartiles of distribution. Sec. Struct., secondary structure.

independent of downstream analyses. It, thus, holds promise as a general diagnostic tool for protein representation learning.

Discussion

Choosing a model for a protein set

Models trained on the same sequence sets but with different architectures and/or training regimes end up being distinct. Do these learning differences make certain models better at representing specific test sets? We applied RNS to evaluate the appropriateness of different pretrained pLMs for diverse protein sequence datasets. We compared the average RNS score (equation (6)) of different embedders (Table 1) on six datasets: Astral40 domains, a composite set of four complete proteomes (Proteome4), intrinsically disordered proteins (IDP) and intrinsically disordered regions (IDR)[36], novel metagenomic sequences[35] and novel 'hallucinated' sequences[37] (Methods, Fig. 3a and Supplementary Tables 1 and 4). Note that, because RNS was formulated to quantify uncertainty of individual protein embeddings, averages across a dataset should be viewed as indicative, rather than explicit, measures of dataset-level uncertainty.

Among the tested models, ProtT5 and ESM-2 consistently yielded the lowest RNS scores (<0.1) across all datasets (Fig. 4a,b, Supplementary Fig. 8 and Supplementary Table 4), suggesting higher confidence in their learned representations. These two models outperformed all other pLMs, including variants of similar architecture—for example, ProtTrans T5-BFD and ESM-1.

Interestingly, all pLMs, including ProtT5 and ESM-2, assigned higher RNS scores ($RNS_{k=1,000}$ = 0.23 and 0.25, respectively) to IDRs compared to the structured Astral40 dataset ($P < 1 \times 10^{-6}$, two-sided Mann–Whitney U-test), placing IDRs closer in representational uncertainty to their junkyards (Fig. 4). Notably, for the IDR set, ESM-1 showed the most uncertainty among pLMs—a possible consequence of training bias toward structured proteins, which could limit its ability to represent the disordered sequence space. As expected, for all models, the RNS of IDPs, which contain both structured and disordered regions, fell between RNSs of Astral40 and IDR datasets.

NLP-based models, although generally underperforming in other datasets, showed improved performance on IDPs and IDRs. This may reflect the presence of low-complexity regions in IDRs, producing distinctive k-mer signatures diverging from both Astral40 and Astral40R.

We also evaluated RNS on two novel sequence sets: metagenomic proteins, translated from predicted genes in metagenome-assembled genomes (MAGs), and hallucinated sequences, generated de novo using trRosetta. Note that the term 'novel' in this context refers to sequences with less than 30% sequence identity to any protein in UniProt.

Most pLMs, except ESM-1, PLUS-RNN and Bepler, attained low RNS scores for novel metagenomic sequences, suggesting recognition of coherent biological patterns despite their sequence divergence from training data (Fig. 4). A more striking pattern was observed for the trRosetta hallucinated sequences. The embeddings for all models, including NLP-based ones, for these sequences attained low RNS (Fig. 4). This uniformity suggests a shared recognition of synthetic sequences that are biologically plausible.

RNS captures vector properties of unlearned embeddings

To examine the information content of un(der)learned embeddings, we compared the information richness of the Astral40 embeddings with those of their randomly shuffled counterparts (Astral40R). The informational richness of each embedding was quantified using four

https://doi.org/10.1038/s41592-026-03028-7

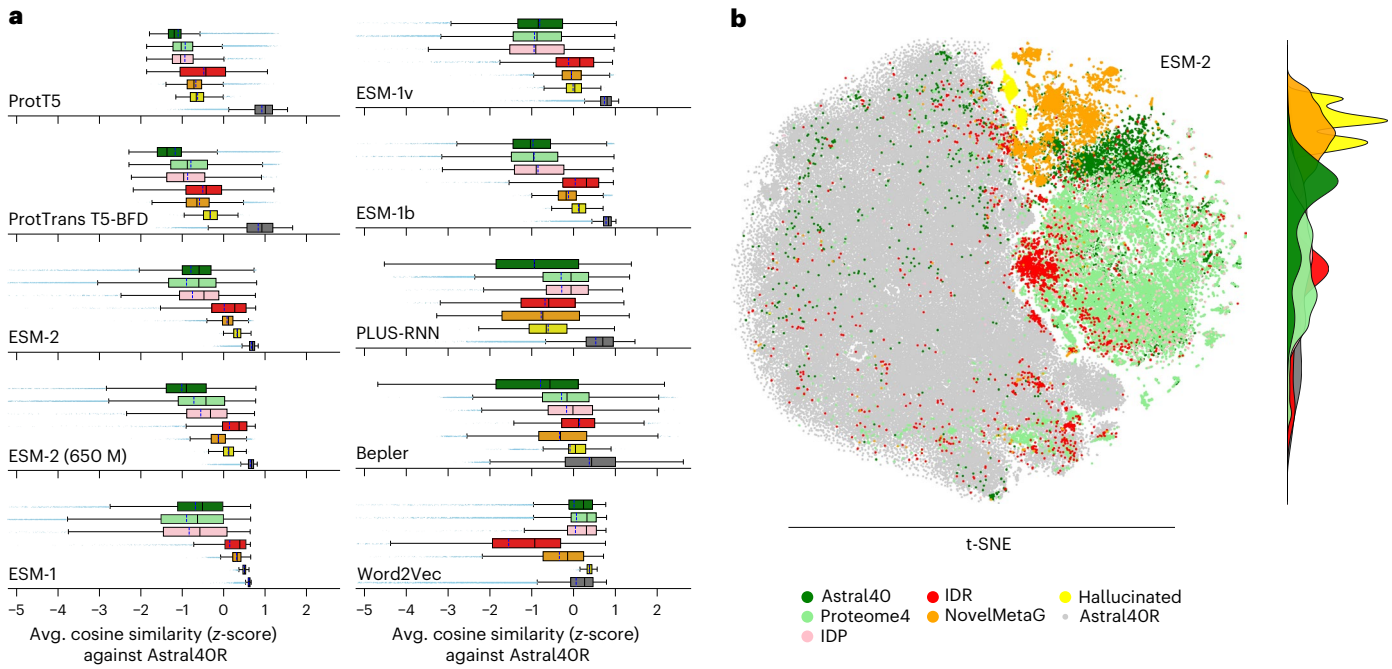


Fig. 3 | Uncertainty in pLM latent representations varies across protein sets. Our RNS quantifies how similar a protein embedding is to the embeddings of random sequences in a given pLM's latent space. Embedding of Astral40 protein domains (dark green, $n = 14{,}711$), four complete proteomes (Proteome4, light green, $n = 42{,}471$), IDPs (pink, $n = 2{,}028$), IDRs (red, $n = 3{,}025$) within IDPs, novel sequences from MAGs (orange, $n = 11{,}369$) and de novo hallucinated proteins (yellow, $n = 2{,}000$) was computed through different models (Methods and Supplementary Table 1). **a**, For the same model, the cosine similarity between protein and random sequence (Astral40R, gray) embeddings differs across protein sets (colors), reflecting the model's learning bias. The distributions of average cosine similarities with random embeddings for each sequence set (each sequence versus all random) are presented as *z*-scores (*x* axis, computed across all sequence sets for each model); outliers are omitted for clarity. The dashed blue lines in each box plot represent the mean of the distribution; black lines are medians; and blue dots represent statistical outliers beyond 1.5× IQR from the first and third quartiles of distribution. **b**, A two-dimensional t-SNE projection of ESM-2 embeddings shows distinct spatial distributions for diverse sequence sets. The extent of overlap with non-biological/random Astral40R embeddings (gray) illustrates uncertainty in ESM-2's latent space (Supplementary Table 1). NovelMetaG, novel metagenomic sequences.

metrics (equations (17–22)): (1) L2 norm of the full sequence embedding—that is, the measure of magnitude or length of the embeddings; (2) covariance of residue embeddings across the full sequence (COV_{prot}); (3) mean covariance among residue embeddings within all possible 15-residue fragments of the sequence ($\overline{COV}_{\text{frag}}$); and (4) average cosine similarity among consecutive residues in all possible 15-residue fragments ($\bar{S}_{\text{frag}}$). Residue covariance and residue similarity within a sequence indicate mutual agreement or disagreement between residues in vector space. Unlike randomly generated sequences, biological ones have evolved to perform specific functions. This constraint enforces synergy among residues, enabling protein folding, dynamic conformational changes, allosteric regulation and carrying out functions like catalysis.

We evaluated the difference in these metrics between embeddings of Astral40 and Astral40R datasets using standardized mean difference (Cohen's *d*, equations (23) and (24); Extended Data Fig. 1a and Supplementary Table 6). Although the L2 norm significantly distinguished Astral40 from Astral40R for the ESM family of models, this distinction was absent in other pLMs (Supplementary Table 6 and Extended Data Fig. 1a). By contrast, $\overline{COV}_{\text{frag}}$ showed more consistent differentiation for the most evaluated pLMs—ESM family, PLUS-RNN and Bepler—but not the ProtT5 family. Interestingly, $\bar{S}_{\text{frag}}$ was significantly different between Astral40 and Astral40R for ESM and ProtT5 families but not for PLUS-RNN and Bepler. These findings suggest that the different metrics capture distinct aspects of representational learning across architectures. To convey the signals of these diverse characteristics into a single uncertainty metric, we devised the RNS, which ranks each protein embedding's position—relative to both randomized sequence and other protein embeddings—within the latent space of a pLM (Extended Data Fig. 1b).

We note that embedding representations of a protein sequence are derived from residue embeddings as either an average of all token (residue) embeddings or a most representative token embedding. In either case, the complete set of token embeddings, an array of vectors, would be more informative. Indeed, in our analysis, COV_{prot} was less distinct than $\overline{COV}_{\text{frag}}$, confirming the information loss incurred from averaging over the entire sequence. Nevertheless, protein-level embeddings remain widely used due to their convenience and lower computational costs. For this reason, we used protein-level embeddings to illustrate the utility of RNS in this study, whereas a more comprehensive residue-level analysis of RNS and related features may have been more informative.

The un(der)learned portion of the human proteome informs predictive performance

We found that 19.1% (3,450) and 46.2% (8,372) of the human proteome (18,100 proteins of length ≤1,022) is un(der)learned ($\text{RNS}_{k=1,000} > 0$) by ProtT5 and ESM-2 (3 billion (3 B)) models, respectively. Note that when all isoforms ($n = 37{,}477$) are considered, these proportions increase to 19.7% (7,386) and 49.8% (18,677). Unexpectedly, ESM-2—a model with 3 B parameters and also the backbone of ESMFold—exhibited substantially higher uncertainty than ProtT5. We hypothesize that this could stem from the specialization of ESM-2 for structure predictions, potentially at the expense of sequence-level generalization.

We extended our analysis to include other ESM variants (Table 1), namely ESM-1v and ESM-2 (with 650 million (650 M) parameters). These demonstrated lower uncertainty of 27.5% and 15.2% of the human proteome, respectively. These results further underscore that larger models do not universally outperform smaller ones across tasks and datasets. Note that, for this computation, proteins exceeding 1,022

https://doi.org/10.1038/s41592-026-03028-7

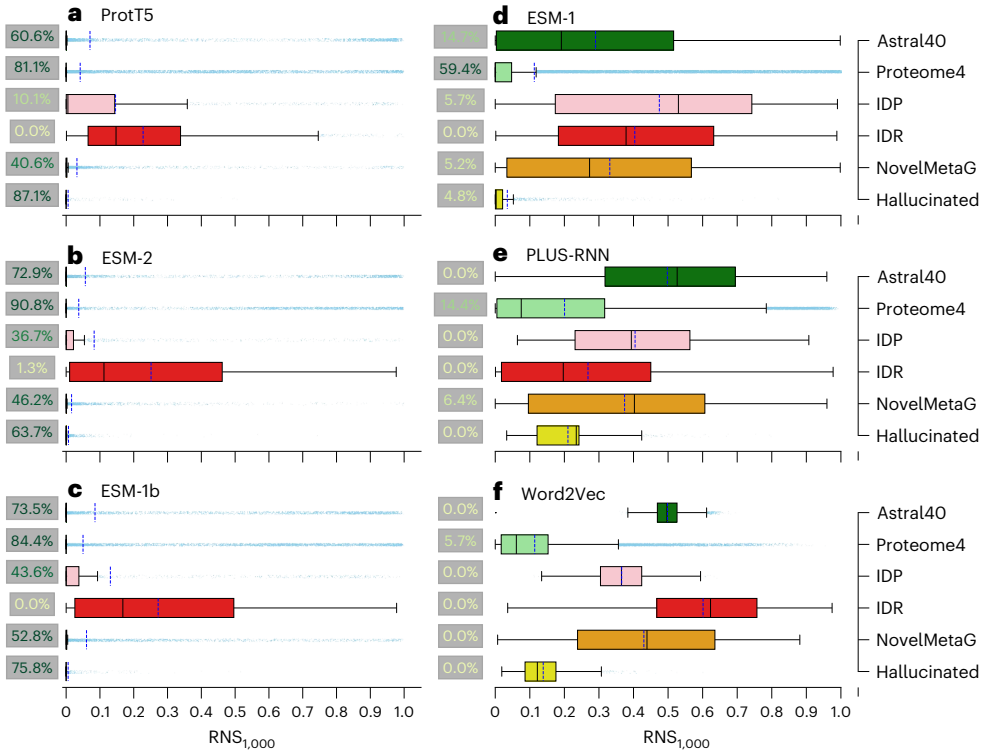


Fig. 4 | RNS measures embedding uncertainty across a range of protein sets and pLMs. The figure shows the distribution of $RNS_{k=1,000}$ values (*x*axis), computed for protein embeddings from multiple pLMs (**a**–**f**) across diverse sequence sets (*y*axis, colors), including protein domains from the Astral40 dataset ($n = 14,711$), proteomes from four model organisms (Proteome4, $n = 42,471$), IDPs ($n = 2,028$), IDRs ($n = 3,025$) of IDPs, novel protein sequences derived from MAGs ($n = 11,369$) and novel de novo hallucinated proteins ($n = 2,000$) (Methods). The uncertainty indicated by RNS highlights the biases of pLMs in representation of certain sequence sets in latent space. For example, IDR representations are consistently uncertain (close to random sequences in Astral40R) across all models, whereas Astral40 representations are mostly certain for models in **a**–**c**. The blue dashed line in box plots indicates the means of distributions; black lines are medians; and blue dots represent statistical outliers beyond 1.5× IQR from the first and third quartiles of distribution. The fraction of completely reliable representations (RNS = 0) is displayed for each set/model combination in gray labels. NovelMetaG, novel metagenomic sequences.

residues were excluded due to the sequence length limitations imposed by ESM models.

With a more relaxed RNS threshold ($RNS_{k=1,000} > 0.1$), illustrating the performance of a potentially more tolerant downstream application, the fraction of uncertain embeddings decreased, to 9.7% for ProtT5, to 10.4% for ESM-2 (3 B), to 8.8% for ESM-2 (650 M) and to 16.1% for ESM-1v, but did not disappear. Given that the human proteome is exceedingly well studied[38]—for example, there are many human protein structures in the PDB[39]—the expectation for other proteomes would be a higher number of underlearned protein representations[40]. Thus, we expect that application of embedding screening prior to downstream applications could increase the precision of the latter.

To evaluate the above conjecture, here we demonstrate the utility of our RNS-based screening in improving the prediction of the effects of protein variants. We evaluated three pLMs—ProtT5, ESM-1v and the 650 M-parameter version of ESM-2 (Table 1)—on two different datasets of human variants, as described in ref. 16. In brief, these are variants (single-nucleotide polymorphisms (SNPs)) (1) annotated by the Protein Mutant Database (PMD)[41] as knockout, effect or neutral with regard to the corresponding protein function changes ('Function') or (2) labeled by ClinVar[42] as pathogenic or likely pathogenic versus identified as common or rare[43] in the human population (Methods and Supplementary Table 1). For each model, we binned the proteins based on the RNS score of the corresponding protein and compared the log-likelihood score of variants, inferred from embeddings[16], to the variant classes. Note that, although RNS can be computed at the residue level, this analysis focused on protein-level RNS for illustration purposes.

We evaluated the performance (equations (10–15)) of these three pLMs in distinguishing SNPs as all combinations of (knockout or effect) versus neutral and (pathogenic or likely pathogenic) versus (common or rare). The models attained the highest performance (area under the receiver operating characterisitc (AUROC) > 0.8; Methods) for variants from proteins with zero embedding uncertainty ($RNS_{k=1,000} = 0$; Fig. 5); AUROC dropped to approximately 0.5 for proteins with $RNS_{k=1,000} > 0.8$.

Given that RNS quantifies the model's uncertainty in its representation of a protein, we hypothesized an inverse relationship between RNS and the ability of the model to correctly identify functionally impactful variants. Across all three models, effects of variants in proteins with low RNS values (<0.5)—that is, more certain embeddings—were, indeed, more readily distinguishable, whereas, for proteins with higher RNS values, variant impact was indistinguishable from wild-type. We also observed this trend when comparing predictions of ClinVar pathogenicity versus common SNPs (Supplementary Fig. 9).

These results demonstrate diminished predictive capacity in downstream tasks when using embeddings of greater representational uncertainty. In other words, prescreening for embeddings of high uncertainty could improve model performance. Beyond improving existing model performance, we suggest that the ability to select well-represented samples could also lead to better training sets for newly built downstream models. After all, better and more suitable-to-the-task data are arguably the best source of model improvement.

'All models are wrong, but some are useful'

George Box's famous maxim applies to pLMs just as well as any other model[44]. Here we demonstrated the limitations of using pLM-produced

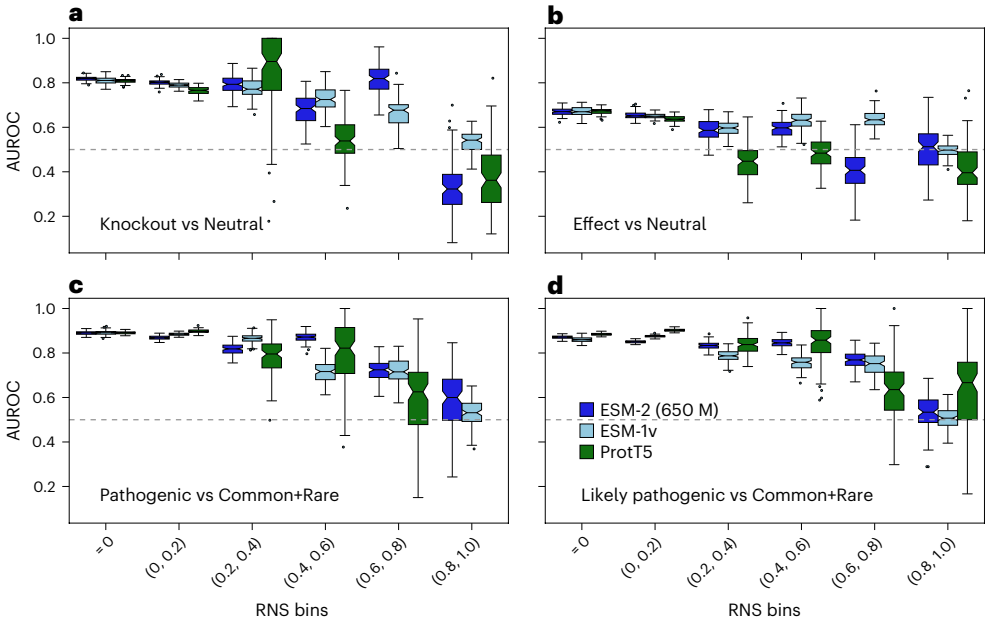


Fig. 5 | Variant classification tasks are improved with RNS-based screening. AUROC performance of three pLMs—ESM-2 (650 M parameters; dark blue), ESM-1v (light blue) and ProtT5 (green)—across binned RNSs ($RNS_{k=1,000}$) for four classification tasks: knockout ($n = 1,584$) versus neutral ($n = 1,777$) (**a**), effect ($n = 3,818$) versus neutral ($n = 1,777$) (**b**), pathogenic ($n = 2,499$) versus common+rare ($n = 4,960$) (**c**) and likely pathogenic ($n = 4,804$) versus common+rare ($n = 4,960$) (**d**). Each RNS bin (*x*axis) left to right reflects increasing embedding uncertainty (higher values indicate greater uncertainty). Variant impact prediction performance declines at higher RNS ($k = 1,000$), particularly beyond the 0.6 threshold, underscoring the relationship between embedding certainty and task performance. AUROC distribution for each protein's RNS bins was derived by 100 iterations of undersampling to balance mutation types. Horizontal dashed lines indicate random classification (AUROC = 0.5); blue circles represent statistical outliers beyond 1.5× IQR from the first and third quartiles of distribution; and the black line at the center of the box represents the median of the distribution.

protein embeddings as protein representations. We advocate for assigning confidence measures to these embeddings prior to any downstream applications. To this end, we propose several measures of embedding uncertainty estimates, reflecting incomplete learning or representation gaps. One of these, the RNS, is a model-agnostic metric that can be readily applied to outputs of existing foundational models with minimal implementation effort. We argue that incorporating such screening strategies will enhance the reliability and interpretability of deep learning models in protein science. RNS can be integrated with pLM training workflows to pinpoint blindspots in the protein representation space and dynamically steer model training.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03028-7.

References

1. Qin, L. et al. A survey of multilingual large language models. *Patterns (N Y)* **6**, 101118 (2025).
2. Weissenow, K. & Rost, B. Are protein language models the new universal key? *Curr. Opin. Struct. Biol.* **91**, 102997 (2025).
3. Consens, M. E. et al. Transformers and genome language models. *Nat. Mach. Intell.***7**, 346–362 (2025).
4. Zablocki, L. I. et al. Comprehensive benchmarking of large language models for RNA secondary structure prediction. *Brief. Bioinform.* **26**, bbaf137 (2025).
5. Wang, Z. et al. Large language models in bioinformatics: a survey. *Findings of the Association for Computational Linguistics: ACL 2025* (eds Wanxiang, C. et al.) 3602–3615 (Association for Computational Linguistics, 2025).
6. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
7. Marquet, C. et al. Embeddings from protein language models predict conservation and variant effects. *Hum. Genet.* **141**, 1629–1647 (2022).
8. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
9. Hoarfrost, A., Aptekmann, A., Farfanuk, G. & Bromberg, Y. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat. Commun.* **13**, 2606 (2022).
10. Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
11. Prabakaran, R. & Bromberg, Y. Deciphering enzymatic potential in metagenomic reads through DNA language models. *Nucleic Acids Res.* **53**, gkaf836 (2025).
12. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
13. Dallago, C. et al. Learned embeddings from deep learning to visualize and predict protein sets. *Curr. Protoc.* **1**, e113 (2021).
14. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.* **11**, 1160 (2021).
15. Thumuluri, V. et al. NetSolP: predicting protein solubility in *Escherichia coli*using language models. *Bioinformatics* **38**, 941–946 (2022).
16. Bromberg, Y., Prabakaran, R., Kabir, A. & Shehu, A. Variant effect prediction in the age of machine learning. *Cold Spring Harb. Perspect. Biol.* **16**, a041467 (2024).

17. Tran, C., Khadkikar, S. & Porollo, A. Survey of protein sequence embedding models. *Int. J. Mol. Sci.* **24**, 3775 (2023).
18. Fenoy, E., Edera, A. A. & Stegmayer, G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Brief. Bioinform.* **23**, bbac232 (2022).
19. Chen, B. et al. xTrimoPGLM: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nat. Methods* **22**, 1028–1039 (2025).
20. Schmirler, R., Heinzinger, M. & Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **15**, 7407 (2024).
21. Dickson, A. & Mofrad, M. R. K. Fine-tuning protein embeddings for functional similarity evaluation. *Bioinformatics* **40**, btae445 (2024).
22. Ciernik, L. et al. Objective drives the consistency of representational similarity across datasets. In *Proc. 42nd International Conference on Machine Learning* (eds Singh, A. et al.) **267**, 10920–10948 (PMLR, 2025).
23. Greenman, K. P., Amini, A. P. & Yang, K. K. Benchmarking uncertainty quantification for protein engineering. *PLoS Comput. Biol.* **21**, e1012639 (2025).
24. Tsitsulin, A., Munkhoeva, M. & Perozzi, B. Unsupervised embedding quality evaluation. In *Proc. 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)* (eds Doster, T. et al.) 169–188 (PMLR, 2023).
25. May, A., Zhang, J., Dao, T. & Ré, C. On the downstream performance of compressed word embeddings. *Adv. Neural Inf. Process Syst.* **32**, 11782–11793 (2019).
26. Fox, N. K., Brenner, S. E. & Chandonia, J. M. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
27. Brenner, S. E., Koehl, P. & Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28**, 254–256 (2000).
28. Chandonia, J. M. et al. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* **32**, D189–D192 (2004).
29. Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol.* **266**, 635–643 (1996).
30. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).
31. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
32. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
33. Hauser, M., Steinegger, M. & Soding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).
34. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 7112–7127 (IEEE, 2022).
35. Prabakaran, R. & Bromberg, Y. Functional profiling of the sequence stockpile: a protein pair-based assessment of in silico prediction tools. *Bioinformatics* **41**, btaf035 (2025).
36. Aspromonte, M. C. et al. DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res.* **52**, D434–D441 (2024).
37. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
38. Li, Z. L. & Buck, M. Beyond history and 'on a roll': the list of the most well-studied human protein structures and overall trends in the protein data bank. *Protein Sci.* **30**, 745–760 (2021).
39. Burley, S. K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).
40. Ding, F. & Steinhardt, J. Protein language models are biased by unequal sequence sampling across the tree of life. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design* (eds Bouamor, H. et al.) (ICLR, 2024).
41. Kawabata, T., Ota, M. & Nishikawa, K. The Protein Mutant Database. *Nucleic Acids Res.* **27**, 355–357 (1999).
42. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
43. Phan, L. et al. ALFA: allele frequency aggregator. http://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/ (National Center for Biotechnology Information, US National Library of Medicine, 2020).
44. George, E. P. B. Science and statistics. *J. Am. Stat. Assoc.* **71**, 791–799 (1976).
45. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
46. Min, S. et al. Pre-training of deep bidirectional protein sequence representations with structural information. *IEEE Access* **9**, 123912–123926 (2021).
47. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1301.3781 (2013).
48. Asgari, E. & Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* **10**, e0141287 (2015).
49. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics* **5**, 135–146 (MIT Press, 2017).
50. Pennington, J., Socher, R. & Manning, C. GloVe: global vectors for word representation. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014).

Methods

Protein datasets

We collected 14,711 sequences and structures of SCOPe domains of 30–1,000 residues in length from the Astral40 dataset[26,28]. Astral40 consists of non-redundant cluster representatives, clustered at 40% sequence identity. For every Astral40 sequence, we generated five corresponding 'synthetic' sequences by randomly rearranging the residues. Thus, we generated Astral40R, comprising 73,555 random sequences with the same amino acid composition as the Astral40 set.

In addition to Astral40 and Astral40R, we also collected six other sequence datasets: IDP, IDR, novel metagenomic proteins, novel hallucinated proteins, PDB23to24 and Proteome4 (Supplementary Table 1). IDP and IDR consist of 2,028 and 3,025 intrinsically disordered proteins and regions, respectively, extracted from DisProt[36]. The novel metagenomic protein set consists of 11,369 proteins translated from genes of MAGs of less than 30% sequence identity to UniRef100 (refs. 35,51). The novel hallucinated protein set includes 2,000 synthetic proteins of length 100 residues, generated using the trRosetta deep neural network[37].

The PDB23to24 set consists of 663 protein chains (non-redundant at 30% sequence identity[33]) representing 930 monomeric and homo-multimeric structures with resolution <2.5 Å, deposited into the PDB[31] between 2023 and 2024—that is, after the release of the most recent pLM considered in this study. The Proteome4 set comprises complete proteomes from four model organisms: *Drosophila melanogaster* (fruit fly), *Saccharomyces cerevisiae* (baker's yeast), *Mus musculus* (mouse) and *Homo sapiens* (human), contributing, 6,176, 3,253, 15,104 and 17,938 protein sequences, respectively[52]. To ensure uninterrupted processing of the sequences through pLMs, we restricted sequence lengths to 30–1,000 residues and removed any sequences containing non-standard amino acid residues.

Variant datasets

We used two human protein variant sets: (1) the functional set consists of 1,584, 3,818 and 1,777 mutants from 2,056 human proteins, classified as knockout, effect and neutral, respectively, based on the effect of these variants on protein function[16,41]; (2) the pathogenic set comprises 2,499 pathogenic and 4,804 likely pathogenic genomic variants (SNPs) from ClinVar, combined with 1,887 common (minor allele frequency (MAF) ≥ 0.01) and 3,073 rare (0.01 > MAF ≥ 0.001) variants from 1,430 human genes/proteins[16,42,53]. For each of the variants, we extracted the log-likelihood scores[16] from pLMs: ESM-2 (650 M), ESM-1v and ProtT5 (Table 1).

Dataset of non-biological sequences

To capture the latent space occupied by biologically meaningless sequences, we derived a negative control set of random sequences termed Astral40R. Astral40R consists of five randomly shuffled sequences for every one of 14,711 sequences in Astral40 (Supplementary Table 1). The random shuffling rearranges the residues in a given sequence to new positions without changing the overall sequence amino acid composition. This removes any biologically meaningful residue–residue interactions. We also repeated the experiments in this study using another version of the random sequence set, where Proteome4R contains one sequence generated by random shuffling of residues for every one of 42,471 proteins in Proteome4.

Extracting embeddings from pLMs

We analyzed protein representations in the form of embeddings derived from a range of pretrained pLMs. These included models from the ESM family—ESM-2 (3 B and 650 M), ESM-1, ESM-1b and ESM-1v (refs. 6,8)—and the ProtTrans T5 family, comprising ProtT5 (ProtT5-XL-U50) and ProtTrans T5-BFD[34]. For comparison, we also included PLUS-RNN[46], the Bepler & Berger model[45] and three classical natural language models—Word2Vec[47], FastText[49] and GloVe[50]—adapted for amino acid sequences.

Protein embeddings and the corresponding perplexity scores from ESM-2, ESM-1v and ProtT5 models were retrieved via the Hugging Face Transformers interface. Protein embeddings from the three classical natural language models and the Bepler & Berger model[45] were generated using the Bio Embeddings framework[13]. For PLUS-RNN[46], we used Bio Embeddings for extracting embeddings and the author's GitHub repository for perplexity calculation (ref. 46; https://github.com/mswzeus/PLUS). For all models and datasets, protein-level embeddings were computed by averaging residue-level embeddings across the full length of each sequence.

Embedding similarity and distances

To quantify the similarity between two embeddings (e_1 and e_2), both cosine and Euclidean similarity metrics were employed (equations (1) and (2)). We chose a scaling factor of 0.5 over 1 in equation (2) to adjust the distribution's spread and centering. This value, used consistently across our analyses, is arbitrary and chosen to aid visualization of pLM embedding similarity.

$$\text{Cosine similarity}(e_1, e_2) = \frac{e_1 \cdot e_2}{|e_1|_2 . |e_2|_2} \tag{1}$$

$$\text{Euclidean similarity}\,(e_1, e_2) = \frac{0.5}{0.5 + |e_1 - e_2|_2} \tag{2}$$

Similarly, we had used cosine and Euclidean distances to measure distances between embeddings (equations (3) and (4)). Although all analyses were conducted using both distances/similarities, only the results derived from cosine-based measures are reported in this paper, given the consistency of the inferences obtained. Note that we also repeated all analyses using length-normalized embeddings (unit vectors), observing results consistent with those obtained without normalization.

$$\text{Cosine distance}\,(e_1, e_2) = 1 - \frac{e_1 \cdot e_2}{|e_1|_2 . |e_2|_2} \tag{3}$$

$$\text{Euclidean distance}\,(e_1, e_2) = |e_1 - e_2|_2 \tag{4}$$

Embedding visualization

We used t-SNE for dimensionality reduction of the pLM latent space. t-SNE two-dimensional projections were computed using ref. 54 with perplexity of 30 for 300 iterations, followed by 200 iterations of exaggeration of 4, using cosine distance as metric.

RNS

We propose a new measure to quantify uncertainty of an embedding generated by a pLM, as a protein representation. RNS of a protein P_1 ($\text{RNS}_k(P_1)$) is the fraction of non-biological neighbors in a pLM's latent space (equation (5)). In the study, we employ embeddings of Astral40R sequence set—generated sequences with random amino acid residue permutations (defined earlier) as representation of non-biological space. RNS is computed by measuring the fraction of these random sequences among 'k' nearest neighbors of a given protein embedding. RNS can further be extended to estimate uncertainty of a pLM for representation of a set of proteins (equation (6)).

$$\text{RNS}_k\,(P_1) = \frac{1}{k}\sum_{i=1}^{k}\delta\,(i)\,,\ \text{where}\ \delta\,(i) = 1\,\text{if}\,i \in R,\ \text{random set} \tag{5}$$

$$\text{RNS}_k^{\text{dataset}}\,(D_1) = \frac{1}{n}\sum_{i=1}^{n}\text{RNS}_k\,(P_i) \tag{6}$$

In this work, for a given set of proteins, the neighbors were identified from a combined pool comprising set protein embeddings and those of Astral40R, described above, that is a reference set of randomly generated non-biological sequences. To mitigate potential biases due to dataset imbalance, we applied undersampling to ensure an equal number of embeddings from each sequence set in each of the 10 iterations. RNS values reported in this study are averaged over 100 iterations. The neighborhood size (k) was varied from 10 to 1,000 to assess the sensitivity of RNS across scales.

Uncertainty measures

We compared RNS against existing uncertainty measures. We computed two variants of pseudo-perplexity (PPL): PPL_{LOO} (leave-one-out) and PPL_{unmasked}[55].

For a protein sequence of length 'l' with tokens $(x_1, x_2, \ldots, xl)$, PPL_{LOO} was computed by masking one residue at a time and re-parsing the sequence through the pLM, requiring 'l' forward pass in total. At each position i, the model predicts the probability $P(x_i \mid x_i^M)$, where x_i^M denotes the sequence with position i masked. The pseudo-perplexity is defined as:

$$PPL_{\mathrm{LOO}} = \exp\left(-\frac{1}{l}\sum_{i=1}^{l}\log P\left(x_i \mid x_i^M\right)\right) \tag{7}$$

By contrast, PPL_{unmasked} was computed using a single forward pass of the unmasked sequence. The logits at each position were used to estimate the token probabilities, and the pseudo-perplexity was calculated as:

$$PPL_{\mathrm{unmasked}} = \exp\left(-\frac{1}{l}\sum_{i=1}^{l}\log P\left(x_i \mid x\right)\right) \tag{8}$$

PPL_{LOO} provides a more faithful estimate of sequence-level uncertainty but at substantially higher computational cost, whereas PPL_{unmasked} offers a scalable approximation suitable for large datasets and models. We computed PPL_{LOO} on representative subsets of Astral40 and PDB23to24 using ProtT5 and ESM-2, enabling direct comparison to RNS under controlled conditions. For broader coverage, we computed PPL_{unmasked} across all datasets using both ESM-2 and PLUS-RNN.

Jensen–Shannon divergence analysis

To further quantify sequence-level uncertainty, we computed the Jensen–Shannon divergence between empirical and theoretical k-mer distributions. For a given protein sequence, we extracted contiguous k-mers ($k = 2,3$) and computed their empirical frequency distribution P. The theoretical k-mer distribution Q was obtained assuming uniform distribution. Because sparse k-mer distributions can lead to zero probabilities, we applied additive smoothing with parameter α, adjusting raw counts c_i for each k-mer i.

$$p_i = \frac{c_i + \alpha}{\sum_{j=1}^{V} c_j + \alpha}$$

where V is the total number of possible k-mers (20^k). The Jensen–Shannon (JS) divergence between smoothed distributions P and Q is defined as:

$$\mathrm{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{9}$$

where $M = \frac{1}{2}(P+Q)$ and $D_{KL}(P||Q) = \sum_{i=1}^{V} p_i \log\frac{p_i}{q_i}$

We computed Jensen–Shannon divergence for $k = 2$ and $k = 3$ across multiple values of the smoothing parameter α ($\alpha \in [0, 0.1, 0.01, 0.001]$), assessing the sensitivity of divergence to rare k-mer frequencies. The values reported in this paper correspond to $\alpha = 0$.

Alignment hits

We ran every protein sequence against UniRef50 and UniRef90 (release 2022_01) using MMseqs[33] with sensitivity parameter (s) of 7.5 and minimum sequence identity of 30% (–cov-mode 2 –realign 1 –alignment-mode 3 –max-seqs 100,000) to estimate the number of possible matches found during the model's training. We compared the estimated number of matches to RNS.

Evaluation of prediction performance

We measured the impact of RNS-based prescreening of protein embeddings on various prediction performance using standard metrics (equations (10–15)) and including AUROC and area under the precision-recall curve (AUPRC)[56].

$$\text{Accuracy} = \frac{\text{Correctly predicted positives and negatives}}{\text{Total number of positives and negatives}} \tag{10}$$

$$\text{Recall} = \frac{\text{Correctly predicted positives}}{\text{Total number of positives}} \tag{11}$$

$$\text{Precision} = \frac{\text{Correctly predicted positives}}{\text{Total number of predicted positives}} \tag{12}$$

$$\text{Specificity} = \frac{\text{Correctly predicted negatives}}{\text{Total number of negatives}} \tag{13}$$

$$F_1\ \text{score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{14}$$

$$\text{Balanced accuracy} = \frac{\text{Recall} + \text{Specificity}}{2} \tag{15}$$

For contact prediction, we focused on the precision of top L/k predictions, where L denotes the protein sequence length; $k \in [1, 2, 5, 10]$; and contacts are defined as the number of unique residue pairs within 8-Å distance between heavy atoms. Furthermore, we categorized the contacts into short-range (6–11 residues apart in the sequence), medium-range (12–23) and long-range (≥24) contacts. Similarly, for three-class secondary structure prediction for protein residues (helix, extended and other), we used Q3 accuracy (equation (16)).

$$\text{Accuracy, Q3} = \frac{\text{Number of correctly predicted residues}}{\text{Total number of residues}} \tag{16}$$

Features derived from embeddings

To quantify the information content of protein embeddings, we computed four different measures:

(1) L2 norm of the full sequence embedding, representing the embedding magnitude.

$$|e|_2 = \sqrt{\sum_{i=1}^{d} e_i^2} \tag{17}$$

(2) Covariance of residue embeddings across the full sequence (COV_{prot})

$$COV_{\mathrm{prot}} = sign\left(\det(\Sigma)\right) \cdot \log\left|\det(\Sigma)\right|, \tag{18}$$

where $\Sigma = \frac{1}{n}\sum_{j=1}^{n}(e_j - \bar{e})(e_j - \bar{e})^T$ and n is the length of the protein and the number of residue embeddings.

(3) Mean covariance among residue embeddings within all possible 15-residue fragments of the sequence ($\overline{COV}_{\text{frag}}$)

$$\overline{COV}_{\text{frag}}(w) = \frac{1}{n-w+1}\sum_{i=1}^{n-w+1} COV_{\text{frag}\,i}, \text{ where } w \text{ is set as 15} \quad (19)$$

(4) Average cosine similarity among consecutive residues in all possible 15-residue fragments ($\bar{S}_{\text{frag}}$; equations (20) and (21)). We introduce this successive residue cosine similarity (SRCS) to measure the change in similarity between proximal residues along a protein sequence stretch (equation (20)). SRCS is a vector of similarity values, described as:

$$\text{SRCS}_{\text{frag}}(g,w) = \left[S(e_i, e_{i+g}) \mid 1 \le i \le w-g\right] \quad (20)$$

$$\overline{S_{\text{frag}}}(g,w) = \frac{1}{w-g}\sum_{i=1}^{w-g} S(e_i, e_{i+g}) \quad (21)$$

where $S(e_i, e_j)$ represents the embedding similarity between residue i and j (equations (1) and (2)); w is the fragment length; and g denotes the sequence distance between residues. Unless otherwise specified, default values for w and g are 15 and 1, respectively.

For a given protein, we consolidate SRCS as a single score by averaging over all fragments of length w (equation (22)).

$$\overline{S_{\text{prot}}}(g,w) = \frac{1}{l-w+1}\sum_{i=1}^{l-w+1} \overline{S_{\text{frag i}}}(g,w) \quad (22)$$

Standardized mean difference

To quantify the size of the effect on embedding similarity of protein domains in real proteins versus randomly generated non-biological sequences, we computed standardized mean difference (SMD, Cohen's d) for all embedding-derived metrics (equations (17–19) and (22)). For each metric, Cohen's d was calculated as:

$$d = (M_1 - M_2)/s_{\text{pooled}} \quad (23)$$

where M_1 and M_2 represent the mean values of the metric for the Astral40 and Astral40R sets, respectively, and pooled standard deviation (s_{pooled}) is defined as:

$$s_{\text{pooled}} = sqrt\left(\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)/(n_1+n_2-2)\right) \quad (24)$$

with s_1 and s_2 denoting the standard deviations and n_1 and n_2 the sample sizes for each Astral40 and Astral40R.

Positive d values indicate that the metric is higher in Astral40 than in Astral40R, whereas negative values indicate the reverse. Effect sizes were interpreted using standard thresholds (small ≥ 0.2, medium ≥ 0.5, large ≥ 0.8)[57]; relative trends across pLMs were emphasized for comparative purposes.

Statistical tests

Statistical significance was assessed using two-sided Mann–Whitney U-tests, Student's t-tests, Pearson's correlation and Kendall's rank correlation, all implemented in SciPy[58]. Confidence intervals and standard errors were estimated by bootstrapping (100 iterations, 90% subsampling without replacement). Box plots were generated in Matplotlib, with boxes spanning the interquartile range (IQR); central lines indicate medians; whiskers extend to the furthest points within 1.5× IQR; and points beyond are defined as outliers[59].

Code runtime

We evaluated time requirements for RNS computation (Supplementary Fig. 10). For a dataset of 10,000 proteins with maximum k set at 1,000 neighbors, the code for RNS computation with 100 iterations of undersampling typically takes approximately 2 minutes on GPU (16 cores, 64 GB RAM; NVIDIA RTX 6000 Ada) and approximately 14 minutes on CPU (16 cores, 64 GB RAM).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Protein sequence datasets used in this study and their computed RNSs are available at https://doi.org/10.6084/m9.figshare.29080301 (ref. 60). Source data are provided with this paper.

Code availability

A Python package to compute RNS is available at the figshare link above[60] and at https://bitbucket.org/bromberglab/rns/src/main/.

References

51. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
52. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
53. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41**, D8–D20 (2013).
54. Policar, P. G., Strazar, M. & Zupan, B. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *J. Stat. Softw.* **109** https://doi.org/10.18637/jss.v109.i03 (2024).
55. Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. Masked language model scoring. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics* 2699–2712 (Association for Computational Linguistics, 2020).
56. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
57. Sullivan, G. M. & Feinn, R. Using effect size—or why the *P* value is not enough. *J. Grad. Med. Educ.* **4**, 279–282 (2012).
58. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
59. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
60. Prabakaran, R. & Bromberg, Y. Consolidated dataset from the RNS study: protein embedding RNS scores, Jensen–Shannon divergence, and sequence alignment matches. https://doi.org/10.6084/m9.figshare.29080301 (2025).

Acknowledgements

The work of R.P. and Y.B. was supported by National Science Foundation award number 2310114 to Y.B. The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank I. Nemenman and K. M. Martini (both Emory University) for insightful discussions on the utility of protein embeddings and F. D. M. Guintu (Emory University) for valuable input on the similarity of random embeddings. We sincerely thank A. Maloney-Bertelli (Emory University) for helping us improve our code repository and for assistance with method benchmarking and M. Heinzinger (Technical University of Munich) for guidance on computing perplexity for ProtT5.

Author contributions

Conceptualization and design: Y.B. and R.P. Data acquisition and analysis: R.P. Interpretation, writing and revision: R.P. and Y.B. Funding acquisition: Y.B.

Competing interests
The authors declare no competing interests.

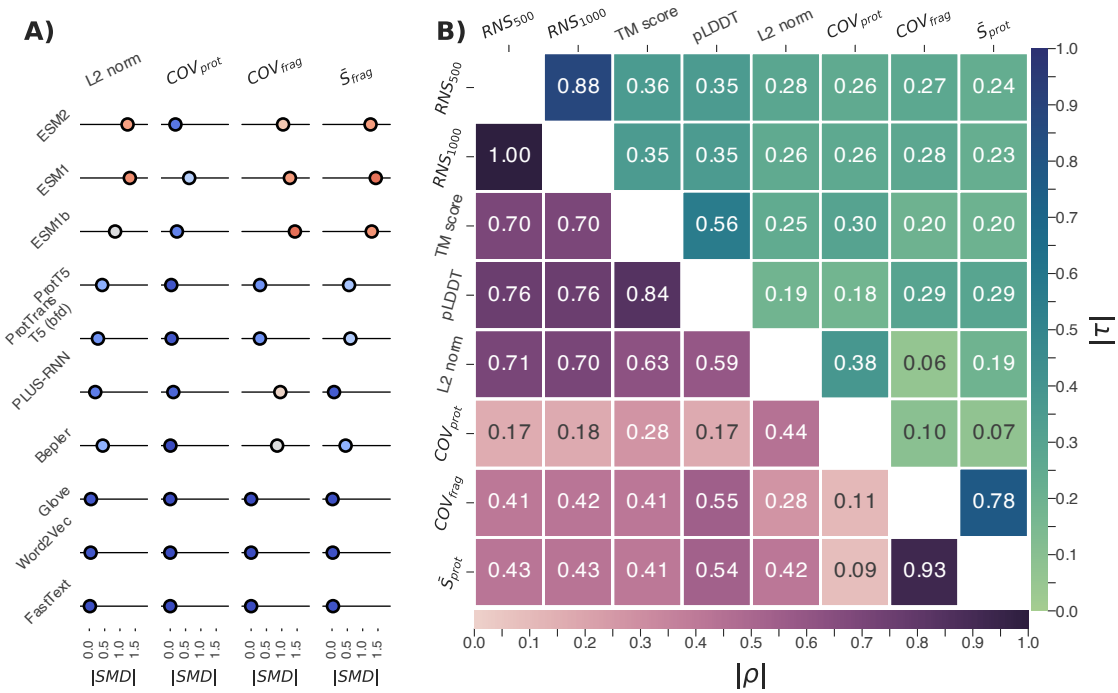
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41592-026-03028-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-026-03028-7.

Correspondence and requests for materials should be addressed to R. Prabakaran or Yana Bromberg.

Peer review information *Nature Methods* thanks Bo Chen, Krzysztof Kotowski and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-026-03028-7



Extended Data Fig. 1 | Information-theoretic measures as indicators of protein uncertainty. (**a**) The absolute Standardized mean differences (Cohen's *d*, circle location on |SMD| X-axes) were calculated to quantify the distinction between embeddings of Astral40 domains (n = 14,711) and Astral40R random sequences (n = 73,555) across multiple information-content-based measures, revealing differences in sensitivities across pLMs. The color of each circle ranges from blue (low SMD) to red (high SMD). (**b**) Correlation matrix presents the relationships between these measures and key confidence indicators, including Random Neighbor Score (RNS), ESM-2's pLDDT scores, and the TM-scores comparing ESM2-predicted structures with experimental structures. Pearson's correlation coefficients are below the matrix diagonal, while Kendall rank correlations are displayed above the diagonal, highlighting both linear and rank-based associations.

nature portfolio

Corresponding author(s): R. Prabakaran and Yana Bromberg

Last updated by author(s): Nov 26, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Protein embeddings were collected through execution of following open source repositories: a) https://github.com/sacdallago/bio_embeddings b) https://github.com/facebookresearch/esm c) https://huggingface.co/Rostlab/prot_t5_xl_uniref50
Data analysis	The python module used to compute Randomized Neighbor Score, a new measure introduced in this manuscript, is deposited at https://bitbucket.org/bromberglab/rns/src/df86008/. The computed data is available at https://doi.org/10.6084/m9.figshare.29080301.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

A python package to compute RNS is available at https://bitbucket.org/bromberglab/rns/src/df86008/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.