nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01044-4

Generalized biological foundation model with unified nucleic acid and protein language

Received: 9 September 2024

Accepted: 29 April 2025

Published online: 18 June 2025

Check for updates

Yong He[1] ✉, Pan Fang[1], Yongtao Shan[2], Yuanfei Pan[3], Yanhong Wei[4], Yichang Chen[5], Yihao Chen[6], Yi Liu[1], Zhenyu Zeng[1], Zhan Zhou[5], Feng Zhu[7], Edward C. Holmes[8], Jieping Ye[1], Jun Li[9], Yuelong Shu[6,10,11], Mang Shi[12] ✉ & Zhaorong Li[1] ✉

The language of biology, encoded in DNA, RNA and proteins, forms the foundation of life but remains challenging to decode owing to its complexity. Traditional computational methods often struggle to integrate information across these molecules, limiting a comprehensive understanding of biological systems. Advances in natural language processing with pre-trained models offer possibilities for interpreting biological language. Here we introduce LucaOne, a pre-trained foundation model trained on nucleic acid and protein sequences from 169,861 species. Through large-scale data integration and semi-supervised learning, LucaOne shows an understanding of key biological principles, such as DNA–protein translation. Using few-shot learning, it effectively comprehends the central dogma of molecular biology and performs competitively on tasks involving DNA, RNA or protein inputs. Our results highlight the potential of unified foundation models to address complex biological questions, providing an adaptable framework for bioinformatics research and enhancing the interpretation of life's complexity.

From the discovery of DNA to the sequencing of every living form, the faithful rule-based flow of biological sequence information from DNA to RNA and protein has been the central tenet of life science. These three major information-bearing biopolymers carry out most of the work in the cell and then determine the structure, function and regulation of diverse living organisms[1,2].

The basic information in these three biopolymers is presented in a linear order of letters: 4 nucleotides for DNA or RNA and 20 standard and several non-standard amino acids for proteins. Their secondary or higher structure also contains information attributed to biological functions and phenotypes. This genetic principle resembles the human linguistic system. Darwin wrote in his *The Descent of Man*: "The formation of different languages and distinct species, and the proofs that both have been developed through a gradual process, are curiously the same[3]." Various studies have testified to these parallels ever since, promoting the understanding and decoding of biological language[4–6].

Given the rapid advancements in machine learning technologies for human language processing, our efforts to decode biological language are bound to accelerate by leveraging insights from the former. The recent development of transformer architecture showed the superior capability of generalizing massive sequence-based knowledge from large-scale labelled and unlabelled data, which empowered language models and achieved unprecedented success in natural language processing tasks. By pre-training on large datasets, foundational models learn the general characteristics of biological sequences. These models compute the input sequence into an embedding, a numerical representation that succinctly captures its semantic or functional properties. On this basis, various biological computation problems

A full list of affiliations appears at the end of the paper. ✉e-mail: sanyuan.hy@alibaba-inc.com; shim23@mail.sysu.edu.cn; lzr098@gmail.com

can be addressed through direct prediction, embedding analysis or transfer learning[7]. In life science, substantial efforts have been put into adopting such language models, especially in protein tasks (ProTrans[8], ProteinBERT[9], ESM2[10], Ankh[11]), such as structure prediction[10,12] and function annotation[13,14]. In the realm of nucleic acid-focused tasks, several models have been introduced within niche areas (DNABert2[15], HyenaDNA[16], ScBert[17]). However, a broadly applicable, foundational model for nucleic acids remains elusive in widespread adoption across various disciplines.

Therefore, we have opted for a more fundamental and universal approach and developed a pre-trained, biological language semi-supervised foundation model, designated as 'LucaOne', which integrates nucleic acid (DNA and RNA) and protein sequences for concurrent training. This methodology allows the model to process and analyse data from nucleic acids and proteins simultaneously, facilitating the extraction of complex patterns and relationships inherent in the processes of gene transcription and protein translation[18,19].

We further examine that LucaOne shows an emergent understanding of the central dogma in molecular biology: the correlation between DNA sequences and their corresponding amino acid sequences, supporting the notion that the concurrent training of nucleic acid and protein sequences together yields valuable insights[20]. To illustrate LucaOne's practical effectiveness, we present seven distinct bioinformatics computational scenarios. These examples highlight LucaOne's ease of use in real-world applications and demonstrate its superior performance compared with state-of-the-art models and other existing pre-trained models.

Results

LucaOne as a unified nucleic acid and protein foundation model

LucaOne was designed as a biological language foundation model through extensive pre-training on massive datasets, enabling the extraction of generalizable features for effective adaptation to various downstream tasks, therefore allowing researchers to efficiently employ pre-trained embeddings from LucaOne for a diverse range of bioinformatics analysis, even when there is limited training data, thereby substantially enhancing their performance. This model leverages a multifaceted computational training strategy that simultaneously processes nucleic acids (DNA and RNA) and protein data from 169,861 species (only those with a minimum of 10 sequences within the training dataset are counted). Consequently, LucaOne has the capability to interpret biological signals and, as a foundation model, can be guided through input data prompts to perform a wide array of specialized tasks in biological computation.

Figure 1 depicts the LucaOne framework, which adopts and enhances the transformer encoder[21] ('Model architecture' in Methods). LucaOne's vocabulary comprises 39 unique tokens representing nucleotides and amino acids ('Vocabulary' in Methods). We used pre-layer normalization to supersede post-layer normalization to make deep networks easier to train. Rotary position embedding replaces traditional absolute positional encoding for inferring longer sequences. In addition, the mixed-training model distinguishes nucleotides and amino acids by utilizing token-type encoding, assigning 0 to nucleotides and 1 to amino acids.

To comprehensively assimilate the patterns and structures pervasive in universal biological language and the inherent knowledge these patterns convey, we have compiled an extensive collection of nucleic acid and protein datasets as the foundational pre-training material. RefSeq provided nucleic acid sequences, including DNA and RNA, and annotations for eight selected genome region types and their order-level taxonomy. Protein data included sequences (from UniProt and ColabFoldDB), annotations (from InterPro, UniProt and ColabFoldDB) and tertiary structures (from RCSB-PDB and AlphaFold2; Fig. 2a, Extended Data Figs. 1 and 2, and Supplementary Fig. 1). A semi-supervised learning[19] approach was employed to enhance its applicability in biological language modelling. Therefore, our pre-training tasks have been augmented with eight foundational sequence-based annotation categories. These annotations complement the fundamental self-supervised masking tasks, facilitating more effective learning for improved performance in downstream applications (Fig. 2b and Supplementary Fig. 3). Overall, LucaOne comprised 20 transformer-encoder blocks with an embedding dimension of 2,560 and a total of 1.8 billion parameters. The downstream task utilized a model checkpoint at 5.6 million ('Pre-training information' in Methods). To illustrate the benefits of mixed training for nucleic acids and proteins, we trained the two additional models (LucaOne-Gene and LucaOne-Prot) separately using nucleic acids and proteins individually, and made a comparison using the same checkpoint in the central dogma of molecular biology task. Details of the pre-training data, pre-training tasks and pre-training details refer to Pre-training data details', 'Pre-training tasks details' and 'Pre-training information' in Methods, respectively.

We utilized *t*-distributed stochastic neighbour embedding (t-SNE) to visualize the embeddings from three distinct datasets: a nucleic acid dataset (S1), comprising sequences from 12 marine species, a protein dataset (S2), consisting of sequences from 12 clans (Pfam clans are groups of protein families that are evolutionarily related and share similar structures and functions), and another protein dataset (S3), organizing recently updated sequences from the top 12 most prevalent Gene Ontology (GO) terms, biological processes subset. This visualization was compared with the results obtained using the MultiHot, DNABert2[15] and ESM2-3B[10] embedding approaches. The outcomes, as illustrated in Fig. 2c–j, revealed that the embeddings produced by LucaOne were more densely clustered, indicating that this method may encapsulate additional contextual information beyond the primary sequence data (dataset S1, S2 and S3 details are in 'LucaOne embeddings level analysis' in Methods, and the embedding clustering metrics are in Extended Data Table 1). In addition, we examined the correlation between nucleic acid sequences and protein sequences of the same genes based on embeddings. The results demonstrated that, despite the absence of paired data and explicit correspondence relationships during training, the sequences (nucleic acids and proteins) of the same gene exhibited convergence within the LucaOne embedding space. Moreover, this convergence was more pronounced compared with other independently trained pre-trained models and sequence alignment methods (details in 'LucaOne embeddings level analysis' in Methods).

Learning the central dogma of molecular biology

Our additional objective was to account for known gene and protein sequences occupying a minuscule yet biologically active niche within their respective design spaces, with a subset of these sequences exhibiting correspondence based on the central dogma. Consequently, throughout the training phase of the LucaOne model, we refrained from incorporating any explicit representations of the relationships between DNA, RNA and protein sequence, seeking to test whether the model inherently grasped the correlation between the genetic and protein data[22,23].

We designed an experimental task to assess the ability of LucaOne to recognize the inherent link between DNA sequences and their corresponding proteins. We have constructed a dataset comprising DNA and protein matching pairs derived from the National Center for Biotechnology Information (NCBI) RefSeq database, with a proportion of 1:2 between positive and negative samples (Fig. 3a,b and 'Details of central dogma tasks' in Methods). To better test whether the LucaOne model has already learned the correspondence between nucleic acid and protein sequences in the central dogma, few-shot learning was employed for validation. The samples were then randomly allocated across the training, validation and testing sets in a ratio of 4:3:25, respectively (refer to 'original dataset' in the following sections).

https://doi.org/10.1038/s42256-025-01044-4

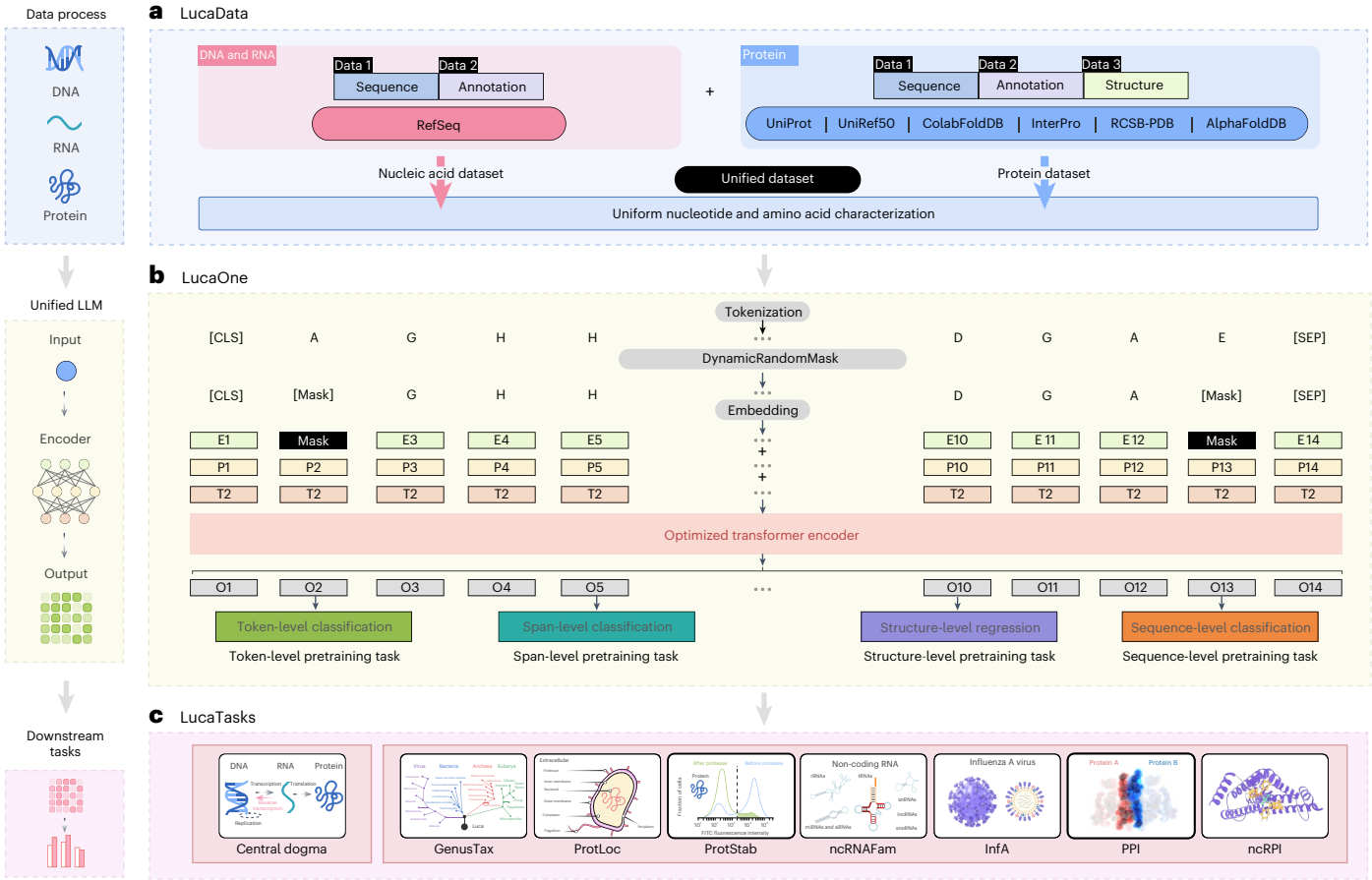


Fig. 1 | The workflow of LucaOne. a, Data source and processing for pre-training. The nucleic acid data are from RefSeq and included sequences and annotations, which consisted of order-level taxonomy and eight selected genome region types. Protein encompasses sequences (from UniRef50, UniProt and the ColabFoldDB metagenomic protein collection (that is, ColabFoldDB), where UniRef50 is clustered set of sequences from the UniProt with at least 50% sequence identity to enhance the learning of these representative sequences), annotations (order-level taxonomy from UniProt and ColabFoldDB, keywords from UniProt, and features such as sites, homologous superfamilies and domains from InterPro) and tertiary structures (experimentally determined structure from RCSB-PDB and predicted structure from AlphaFold2-Swiss-Prot). **b**, Pre-training model architecture and pre-training tasks. The encoder is an improved transformer encoder. Based on two self-supervised mask tasks, an additional eight semi-supervised pre-training tasks were introduced to enhance the model's understanding of the data through annotations in the sequences. **c**, Downstream tasks for validation based on LucaOne embedding. The representational capabilities of LucaOne were verified using eight downstream tasks, whose inputs include DNA, RNA, proteins and their interrelated pairs. [CLS], a special token added at the start of the input sequence to indicate its beginning. [SEP], a special token added at the end of the input sequence to indicate its ending. X (A, G, H, D, E, etc.) represents the input sequence tokens (nucleotides or amino acids). E, embeddings of amino acids or nucleotides; P, positional embeddings; T, the molecular type embedding of the input sequence, where T1 denotes the nucleic acid and T2 denotes the protein. O, the output representation vectors of each token in the input sequence via transformer-encoder. FITC, fluorescein isothiocyanate.

The study employed a simple downstream network to evaluate LucaOne's predictive capacity (Fig. 3c). LucaOne encoded nucleic acid and protein sequences into two distinct fixed embedding matrices (Frozen LucaOne). Then, each matrix was processed through pooling layers (either max pooling or value-level attention pooling[24]) to produce two separate vectors. The vectors were concatenated and passed through a dense layer for classification.

We compared the performance of different modelling approaches, including one-hot with a transformer, a transformer model with the random initialization, nucleic acid embeddings from DNABert2, protein embeddings from ESM2-3B, and two separate versions of the LucaOne foundation model trained independently on nucleic acid and protein sequences (LucaOne-Gene and LucaOne-Prot), and the unified training foundational version of LucaOne (Fig. 3d and Extended Data Table 2). The findings indicated that modelling methods lacking pre-trained elements (one-hot and random initialization; Extended Data Table 2) were unable to acquire the capacity for DNA–protein translation in this dataset. In contrast, LucaOne's embeddings were able to learn this capacity with limited training examples effectively and substantially surpassed both the amalgamation of the other two pre-trained models (DNABert2 + ESM2-3B) and the combined independent nucleic acid and protein LucaOne models using the same dataset, architecture and checkpoint. This suggests that pre-trained foundational models can provide additional information beyond the specific task samples for such biological computation tasks. Moreover, LucaOne's unified training approach for nucleic acids and proteins enabled it to learn within a single framework, thereby capturing the fundamental intrinsic relationships between these two categories of biological macromolecules to some extent.

A CDS–protein dataset using data from the original task was prepared to further evaluate the model's capabilities. Figure 3d shows that models trained exclusively on the CDS–protein dataset demonstrated improvements across multiple performance metrics, including accuracy, F1 score and AUC. When comparing the LucaOne model with the LucaOne-Gene and LucaOne-Prot models and the DNABert2 + ESM2-3B model, the enhancements were more substantial in the latter two model groups compared with LucaOne alone. This suggests that the LucaOne model has marginally enhanced

https://doi.org/10.1038/s42256-025-01044-4

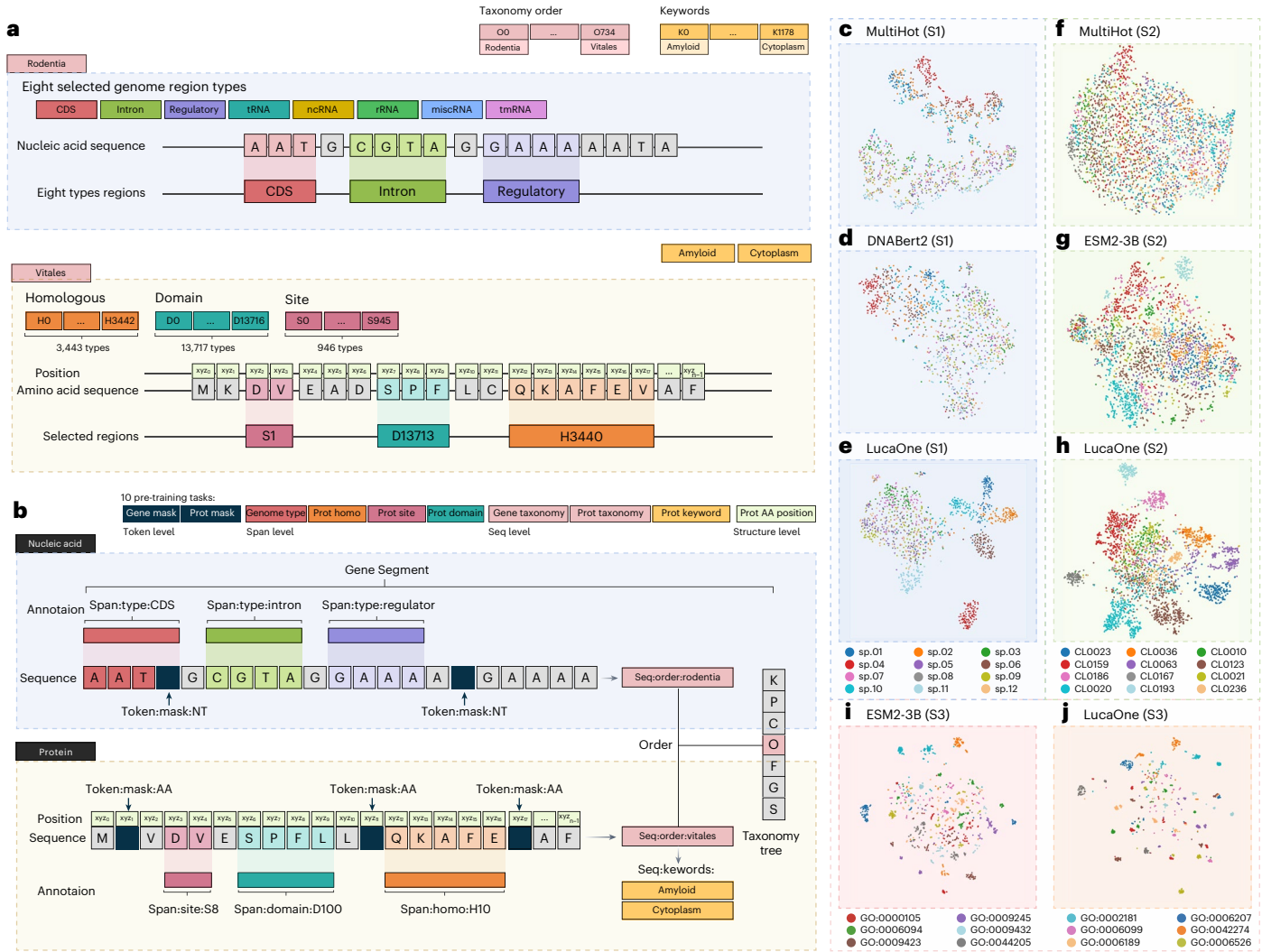


Fig. 2 | The data and tasks for pre-training LucaOne, and t-SNE on four embedding models. a, Details of pre-training data. Nucleic acids included sequence and two kinds of annotation. The protein consisted of sequence, five types of annotation and tertiary structure coordinates. NT, nucleotides; AA, amino acids. **b**, Details of pre-training tasks. The pre-training tasks included two self-supervised mask tasks and eight semi-supervised tasks. **c**–**j**, t-SNEs of the four embedding methods on the S1 nucleic acid contigs with 12 species from the CAMI2 database (**c**,**d**,**e** by MultiHot, DNABert2 and LucaOne, respectively), S2 protein sequences across 12 clan categories from the Pfam database (**f**,**g**,**h** by MultiHot, ESM2-3B and LucaOne, respectively), and S3 protein sequences across the top 12 most prevalent GO terms from the UniProt database (**i**,**j** by ESM2-3B and LucaOne, respectively). The results show that LucaOne's representation has better clustering on these three datasets (nucleic acid sequences of the same species should be clustered because of high sequence similarity, and protein sequences of the same Pfam clan or GO term should be clustered of similar structures and functions). sp.01, unclassified *Pseudomonas* species; sp.02, *Aeromonas salmonicida*; sp.03, unclassified *Vibrio* species; sp.04, *Streptomyces albus*; sp.05, *Aliivibrio salmonicida*; sp.06, unclassified *Brevundimonas* species; sp.07, *Vibrio anguillarum*; sp.08, *Aliivibrio wodanis*; sp.09, *Moritella viscosa*; sp.10, unclassified *Enterobacterales* species; sp.11, unclassified *Tenacibaculum* species; sp.12, unclassified *Aliivibrio* species; GO:0000105, L-histidine biosynthetic process; GO:0009245, lipid A biosynthetic process; GO:0002181, cytoplasmic translation; GO:0006207, 'de novo' pyrimidine nucleobase biosynthetic process; GO:0006094, gluconeogenesis; GO:0009432, SOS response; GO:0006099, tricarboxylic acid cycle; GO:0042274, ribosomal small subunit biogenesis; GO:0009423, chorismate biosynthetic process; GO:0044205, 'de novo' uridine monophosphate (UMP) biosynthetic process; GO:0006189: 'de novo' inosine monophosphate (IMP) biosynthetic process; GO:0006526, L-arginine biosynthetic process).

discriminative capabilities between coding and non-coding regions. However, our experimental results (Supplementary Fig. 9) demonstrate a decline in LucaOne's prediction accuracy as the number of exons within the target sequence region increases. This observed limitation represents a critical area for future model optimization. Furthermore, when evaluating performance across datasets from different species, both models show consistent results, except for a notable decrease in performance with *Ciona intestinalis*. This deviation can largely be attributed to its unique codon usage patterns, which differ significantly from other species in the study (Fig. 3e,f). Given the minimal sample size for this species in the dataset and with only 16% designated for training, it is likely that the models were unable to adequately learn the specific rules of the central dogma under these codon preferences, even though the analysis was conducted under the rule of the standard code. The observed divergence in codon preference suggests that *C. intestinalis* may have more distinctive translation mechanisms from genetic material to proteins, which could be attributed to its unique evolutionary trajectory and selective pressures[25]. Furthermore, a dataset expanded with two urochordate species was utilized for model training and testing. The F1 score of the new model improved significantly for *C. intestinalis*, while the performance for other species remained comparable to that of the original model ('Details of central dogma tasks' in Methods and Extended Data Table 3). Based on this, it is

https://doi.org/10.1038/s42256-025-01044-4

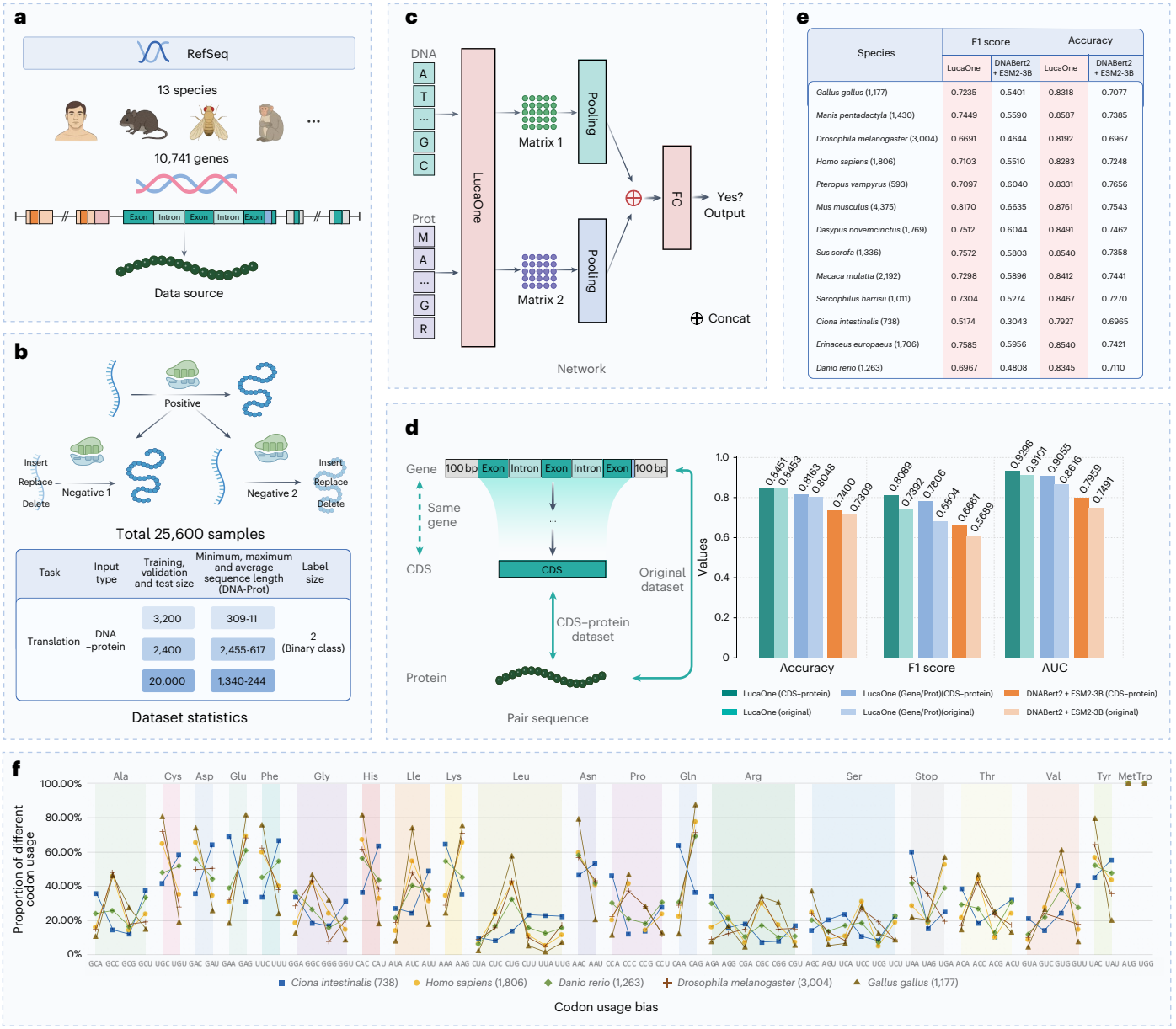


Fig. 3 | The workflow of the central dogma of molecular biology task. **a**, Dataset from 13 species with 10,471 genes in RefSeq. **b**, Prepared 8,533 positive samples and 17,067 negative samples and took a specific sample dividing strategy to test the model performance in this task (training, validation and testing sets in a ratio of 4:3:25). **c**, Based on different embedding methods of DNA–protein pair sequences, a simple downstream network was used for modelling and illustrating their representational ability. **d**, Models performance comparison (validation + testing dataset) on original and CDS–protein datasets. **e**, Comparative performance analysis (validation + testing dataset) of the models across diverse species datasets (sample counts in brackets). FC, fully connected layer. **f**, One species for each class was selected to undergo a codon usage bias analysis, which adheres to the conventions of the standard genetic code; this entails comparing the relative usage frequencies of different codons for each amino acid, ensuring that the total adds up to 100%. The species *C. intestinalis* exhibits a codon usage bias that is markedly distinct from that of other species—overall lower GC content. Details in 'Details of central dogma tasks' in Methods.

inferred that with an expanded training data size encompassing a wider array of central dogma rules, LucaOne has the potential to more thoroughly assimilate the syntactical rules associated with genetic information processing, enabling its application to a more diverse set of scenarios.

LucaOne provides embeddings for diverse biological computational tasks

To ascertain the capacity of the LucaOne model to provide effective embeddings for a variety of downstream tasks, we conducted validation studies across seven distinct downstream tasks, which include single-sequence tasks such as prediction of genus taxon (GenusTax), classification of non-coding RNA (ncRNA) families (ncRNAFam), and the prediction of protein subcellular localization (ProtLoc) as well as the assessment of protein thermostability (ProtStab). For homogeneous paired-sequence tasks, we predicted influenza haemagglutination assays based on a pair of nucleic acid sequences (InfA) and assessed protein–protein interactions (PPI) utilizing pairs of protein sequences. In addition, we forecasted the interactions between ncRNA and proteins (ncRPI) for the heterogeneous sequence task (full task descriptions in 'Downstream tasks details' in Methods and Extended Data Table 4).

For each task, we performed two types of comparative analysis: one against the state-of-the-art results and another using the same downstream network to assess LucaOne embeddings against the widely

https://doi.org/10.1038/s42256-025-01044-4

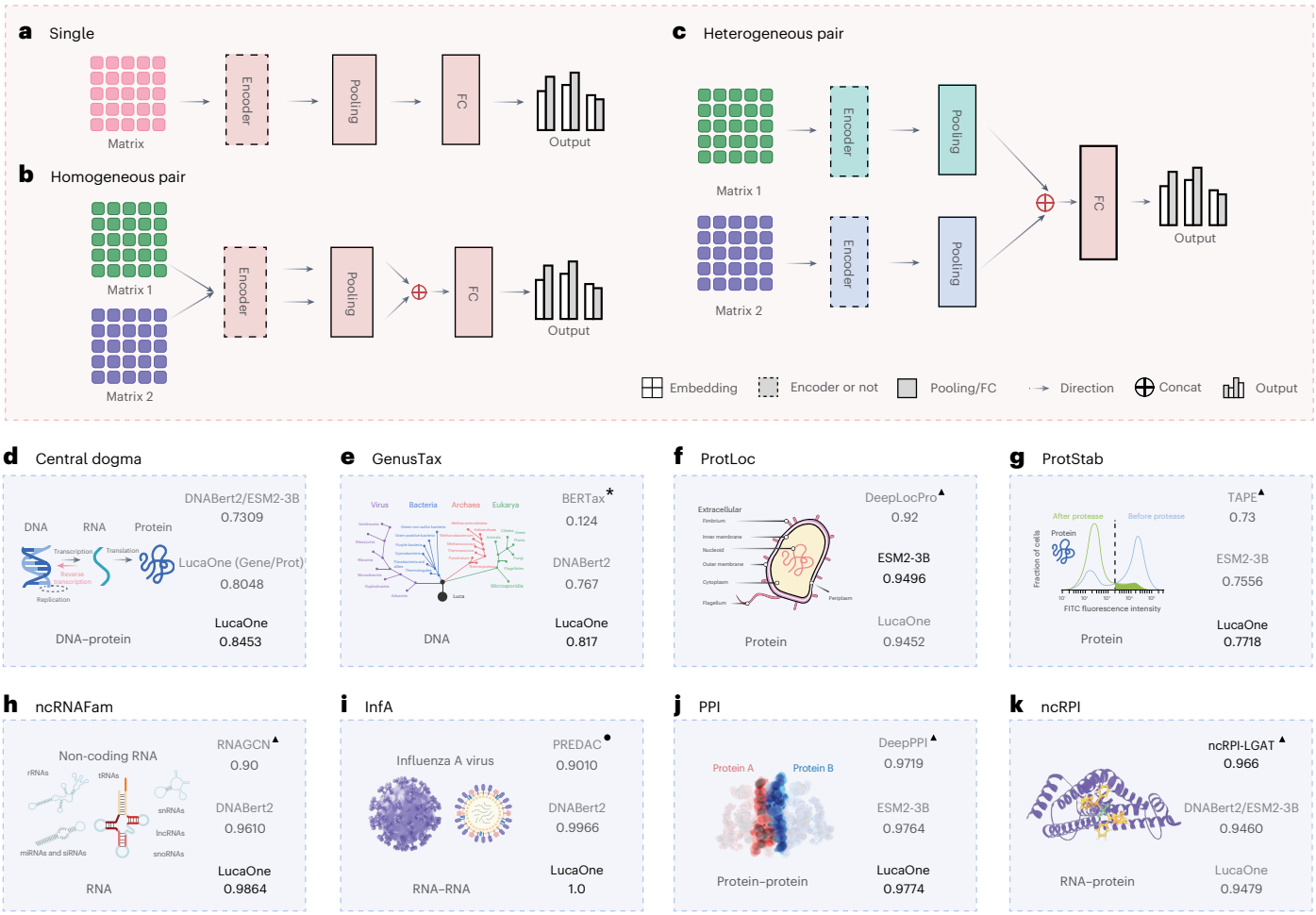


Fig. 4 | Downstream task networks with three input types and results comparison of eight verification tasks. Based on the embedding matrix, three types of input in the downstream task are corresponding networks. **a**, A single sequence, including GenusTax, ncRNAFam, ProtLoc and ProtStab. **b**, Two same-type sequences, including InfA and PPI. **c**, Two heterogeneous sequences: central dogma and ncRPI. **d**–**k**, Comparison results of eight downstream tasks (Central dogma (**d**), Genus taxonomy (**e**), ProtLoc (**f**), ProtStab (**g**), ncRANFam (**h**), InfA (**i**), PPI (**j**), ncRPI (**k**)). The Spearman correlation coefficient was used for the ProtStab regression task and accuracy was used for other tasks. Comparative methods include the state of the art, DNABert2-based (for nucleic acids), ESM2-3B-based (for proteins) and LucaOne-based. The top right asterisk indicates inference using the trained method, the top right triangle indicates direct use of the results in its paper, and the top right circle indicates repetition using its method and is better than the results in the paper.

used nucleic acid and protein pre-trained language models, DNABert2 and ESM2-3B, respectively. These comparative analyses are instrumental in elucidating the incremental contributions of foundation models when addressing related analytical tasks and in evaluating the specific effectiveness of the embeddings generated by LucaOne with DNABert2 and ESM2-3B.

Similarly, we used a simple downstream network to facilitate processing these tasks. We illustrated the capacity of trained and frozen LucaOne to analyse nucleic acid (DNA and RNA) and protein sequences. Figure 4a–c shows the network architectures for three distinct input types. For tasks requiring paired inputs, a concatenation step is necessary to merge the output vectors of the pairs into a single extended vector. Finally, a fully connected layer was utilized for the ultimate output, which could be for classification or regression purposes.

Figure 4d–k shows a comparative analysis of performance on seven distinct biomedical tasks, revealing that LucaOne demonstrates superior representational capabilities over competing models in the GenusTax, ProtStab, ncRNAFam, influenza A antigenic relationship prediction (InfA) and PPI evaluations, and comparable performance on the other two, ProtLoc and ncRPI. Notably, within the nucleic acid-centric GenusTax and ncRNAFam, LucaOne's accuracy has increased by 0.05 and 0.026, respectively, indicating a marked improvement over DNABert2. In the InfA task, LucaOne excelled with an exceptional accuracy of 1.0, reflecting its outstanding ability to represent these task data. For the ProtStab task, it surpassed ESM2-3B with a 0.015 increase in Spearman's rank correlation coefficient and similarly showed a slight improvement in the evaluation of PPI. Compared with DeepLocPro[26] in the task of ProtLoc, LucaOne was competitive with ESM2-3B and showed a 0.025 accuracy improvement. Although LucaOne did not outperform the elaborate network model ncRPI-LGAT[27] in evaluating ncRPI, it still exceeded the combined abilities of DNABert2 and ESM2-3B. LucaOne's effectiveness was particularly notable in processing tasks involving heterogeneous sequences of nucleic acids and proteins; employing a unified representation model is advantageous compared with using separate models. The outcomes of these tasks underscored the robust representational capabilities of LucaOne for both nucleic acid and protein sequences. LucaOne could improve performance across a spectrum of downstream tasks, streamline networks for downstream tasks and reduce computational resource demands (more results of hyperparameter comparison experiments and detailed metrics in 'Comparison result details' in Methods, Extended Data Table 5 and Supplementary Fig. 4).

Discussion

The attempt to build a universal biological language model is to develop a sophisticated cataloguing and retrieval system for 'The Library of Mendel'—the genetic version of 'The Library of Babel'[28,29]. The diversity of genetic variations presents a vast 'design space' that is arguably as rich as the entirety of human literature, if not more so, given the far longer history of life on Earth compared with our record of literature. However, in stark contrast, the proportion of genetic sequences we have successfully identified and catalogued remains considerably smaller than the volume of documented human languages. Moreover, the growth of our understanding and documentation of this 'biological language' is unlikely to occur suddenly or rapidly[30,31]. Our endeavour herein offers a computational model that posits the potential to represent the paradigm of biological language. However, we must temper our expectations regarding this model's rapid and seamless refinement towards an idealized state of perfection.

In developing the LucaOne model, we used deep learning frameworks and techniques from natural language processing. However, we observed systemic discrepancies when applying these models, which were highly successful in natural language contexts, to genomic language[32]. The architecture of BERT-based pre-trained language models focuses on understanding context but may not efficiently capture biological sequences' unique attributes and characteristics[33,34]. Furthermore, the functions and expressions of biological sequences are determined not solely by their genetic sequences but also by the environment in which they are expressed—a factor for which there is at present no practical modelling approach. Standardized methods for processing annotated or phenotypic data are lacking, which can lead to inaccuracies and omissions[35,36]. Moreover, the continual learning and scalability aspects have yet to be fully explored in this study, primarily owing to resource constraints. As a result, the complexities of the model's learning capabilities have not been thoroughly examined at this point, highlighting the primary area of research for the subsequent phase[37]. In terms of application, owing to the diversity of contexts, a robust evaluation system is absent for generalizability and domain adaptability, with small, specialized models occasionally outperforming large pre-trained models in conjunction with downstream tasks in certain areas[32,38].

In light of these considerations, researchers may need to develop specialized pre-trained models tailored to genomic language to improve encoding and comprehension of biological data, ensuring adaptability across a broader spectrum of computational biology tasks. Promising directions include architectural innovations in pre-training models, such as incorporating genetic programming concepts into large language models[39,40]. Another avenue is to harmonize multimodal data, encompassing sequences, feature annotations, experimental results, images and phenotypical information to better understand biological systems beyond unsupervised sequence data learning[41,42]. In addition, employing more transparent algorithms may enhance the interpretability of the model, facilitating better integration with existing biological research frameworks and model development[43,44]. Lastly, given the necessity for pre-trained models to efficiently fine-tune or apply to downstream tasks, paradigms need to expedite model adaptation to new tasks and broader application contexts[32].

To conclude, this paper documents our effort to build a comprehensive large model to represent the intricacies of the biological world. The capabilities demonstrated by LucaOne showed considerable promise and highlighted several areas that necessitate substantial advancements. Such multimodal pre-trained foundational models, grounded in bioinformatics, will prove immensely valuable in accelerating and enhancing our comprehension of biological phenomena.

Methods

Model architecture

Figure 1b illustrates the design of LucaOne, which utilizes the transformer-encoder[21] architecture with the following enhancements:

(1) The vocabulary of LucaOne comprises 39 tokens, including both nucleotide and amino acid symbols (refer to 'Vocabulary').
(2) The model employs pre-layer normalization over post-layer normalization, facilitating the training of deeper networks[45].
(3) Rotary position embedding[46] is implemented instead of absolute positional encoding, enabling the model to handle sequences longer than those seen during training.
(4) It incorporates mixed training of nucleic acid and protein sequences by introducing token-type embeddings, assigning 0 for nucleotides and 1 for amino acids.
(5) Besides the pre-training masking tasks for nucleic acid and protein sequences, eight semi-supervised pre-training tasks have been implemented based on selected annotation information (refer to 'Pre-training tasks details').

Vocabulary

The vocabulary of LucaOne consists of 39 tokens. Owing to the unified training of nucleic acid and protein sequences, the vocabulary includes 4 nucleotides ('A', 'T', 'C' and 'G') of nucleic acid ('U' compiled 'T' in RNA), 'N' for unknown nucleotides, 20 amino acids of protein (20 uppercase letters excluding 'B', 'J', 'O','U', 'X' and 'Z'), 'X' for unknown amino acids, 'O' for pyrrolysine, 'U' for selenocysteine, other 3 letters ('B', 'J' and 'Z') not used by amino acids, 5 special tokens ('[PAD]', '[UNK]','[CLS]','[SEP]' and '[MASK]'), and 3 other ':', '-' and '*'). Owing to the amino acid letters overlapping with the nucleotide letters, the use of '1', '2', '3', '4' and '5' instead of 'A', 'T', 'C', 'G' and 'N', respectively.

Pre-training data details

Nucleic acid. The nucleic acid was collected from the NCBI RefSeq genome database, involving 297,780 assembly accessions. The molecular types included DNA and RNA (Fig. 2a). The DNA sequence, DNA selected annotation, RNA sequence and RNA selected annotation were obtained from the format files 'genomic.fna', 'genomic.gbff', 'rna.gbff' and 'rna.fna', respectively. Among all pre-training sequences, 70% of DNA sequences and 100% of RNA sequences were derived from annotated genomes, while the remaining unannotated sequences were retained to ensure diversity.

DNA reverse strand: the DNA dataset expanded reverse strand sequences with their annotation. A total of 23,095,687 reverse-strand DNA sequences were included.

Genome region types: eight important genome region types in nucleic acids were selected, including 'CDS', 'intron', 'tRNA', 'ncRNA', 'rRNA', 'miscRNA', 'tmRNA' and 'regulatory'. Each nucleotide in the sequence had a label index of 8 categories (0–7) or −100 when it did not belong to these 8 categories.

Order-level taxonomy: the order-level label of the taxonomy tree was selected as the classification label of the nucleic acid sequence. Each sequence had a label index of 735 categories (0–734) or −100 without the order-level taxonomy.

Segmentation: owing to the limited computing resources, each nucleic acid sequence was segmented according to a given maximum length. The fragmentation strategy was presented in Supplementary Fig. 2.

Protein. Protein sequence data were obtained from UniRef50, UniProt and ColabFoldDB. UniRef50 was added to the UniProt database to sample high-quality representative sequences, while ColabFoldDB was incorporated to enhance the diversity of protein sequences. For ColabFoldDB, redundancy within each cluster was minimized by retaining only the ten most diverse sequences. Duplicated sequences between UniProt and ColabFoldDB were excluded. Sequence, taxonomy and keywords were collected from UniProt and ColabFoldDB. The sites, domains and homology regions were extracted from Interpro. The tertiary structure was derived from RCSB-PDB and AlphaFold2-Swiss-Prot.

Sequence: the right truncation strategy was applied when the sequence exceeded the maximum length.

Order-level taxonomy: order-level classification information is used as the protein sequence taxonomy. There were 2,196 categories; each sequence had a label index (0–2,195) or −100 if its order-level information was missing.

Site: four types of site regions ('active site', 'binding site', 'conserved site' and 'PTM') with 946 categories were included. For each amino acid in a sequence, if it was a site location, there was a label index (0–945); otherwise, it was marked with −100.

Homology: a homologous superfamily is a group of proteins that share a common evolutionary origin with a sequence region, reflected by similarity in their structure. There were 3,442 homologous region types; each amino acid in these regions had a label index (0–3,441) corresponding to its type, and the other amino acids were labelled −100.

Domain: domain regions are distinct functional, structural or sequence units that may exist in various biological contexts. A total of 13,717 domain categories were included; each amino acid in these regions had a label index (0–13,716) corresponding to its category, and the other amino acids were marked with −100.

Keyword: keywords are generated based on functional, structural or other protein categories. Each sequence was labelled as a set of label indices with 1,179 keywords or −100 without keywords.

Structure: the spatial coordinates of the C_α atom were used here as the amino acid coordinates. Each amino acid was labelled with a three-dimensional coordinate normalized within the protein chain (preferentially selected the structure from RCSB-PDB). The amino acids at the missing locations were labelled (−100, −100, −100). In total, only about half a million protein sequences had structural information.

Pre-training tasks details

LucaOne has employed a semi-supervised learning approach to enhance its applicability in biological language modelling. Bioinformatics analysis often involves different modalities for input and output data, and most downstream tasks extend from understanding nucleic acid or protein sequences, so our pre-training tasks have been augmented with eight foundational sequence-based annotation categories. These annotations complement the self-supervised masking task, facilitating more effective learning for improved performance in downstream applications. The selection criteria for these annotations focused on universality, lightweight design and high confidence level; consequently, only a subset of the data has such annotations. As listed in Supplementary Fig. 3, there are ten specific pre-training tasks at four levels. All loss functions are presented in Supplementary Note 1.

Token-level tasks: Gene-Mask and Prot-Mask tasks randomly mask nucleotides or amino acids in the sequence following the BERT masking scheme[47] and predict these masked nucleotides or amino acids based on the sequence context in training.

Span-level tasks: the model is trained to recognize some essential regions based on the sequence context. For nucleic acid sequences, eight essential genome region types are learned. For protein sequences, three types of region are labelled: site, homology and domain regions.

Sequence-level tasks: Gene-Taxonomy, Prot-Taxonomy and Prot-Keyword are the order-level taxonomies of nucleic acid, protein and protein-tagged keywords, respectively. They are all sequence-level learning tasks.

Structure-level tasks: as the structure of a protein determines its function, we use a small amount of protein data with a tertiary structure for simple learning in the pre-training phase. Instead of learning the spatial position at the atomic level, the spatial position of amino acids is trained (using the position of the C_α atom as the position of the amino acid).

Pre-training information

On the dimensions of the embedding, the research conducted by Elnaggar et al.[11] demonstrates that the ESM2-3B (embedding dimension 2,560) model outperforms the 650 million (embedding dimension 1,280) version, while the 15 billion (embedding dimension: 5,120) version does not consistently improve performance and substantially increases the computational burden. For the relationship between model size and training data size, Hoffmann et al. suggest that a minimum of 20.2 billion tokens is essential to adequately train a 1 B-sized model[48].

The critical hyperparameters we adopted are as follows: the architecture of LucaOne consists of 20 transformer-encoder blocks with 40 attention heads each, supports a maximal sequence length of 1,280 and features an embedding dimension of 2,560. The model is composed of a total of 1.8 billion parameters. We employed 10 different pretraining tasks, assigning an equal weight of 1.0 to the Gene-Mask, Prot-Mask, Prot-Keyword and Prot-Structure tasks, while assigning a reduced weight of 0.2 to the remaining tasks to equilibrate task complexity (Supplementary Note 1, equation (11)). We used the AdamW optimizer with betas (0.9, 0.98) and a maximum learning rate of 2×10^{-4}, incorporating a linear warm-up schedule throughout the learning-rate updates. For the model training regimen, we utilized a batch size of 8 coupled with a gradient accumulation step of 32. The model underwent training on 8 Nvidia A100 graphics processing units spanning 120 days. A model checkpoint of 5.6 million (5.6 million, trained with 36.95 billion tokens) was selected for the subsequent validation tasks, aligned with ESM2-3B in terms of the volume of data trained for comparison.

To elucidate the advantages of mixed training involving both nucleic acids and proteins, we further conducted experiments with two supplementary models, LucaOne-Gene and LucaOne-Prot, trained exclusively with nucleic acids and proteins, respectively. Their performance in the central dogma of the biology task was evaluated with the same checkpoint (5.6 million) of the two models.

Checkpoint selection criteria: we have released the 5.6 million checkpoint aligned with the ESM2-3B model for a comparable volume of data trained, which was trained with 36.95 billion tokens over 20 times the model's parameters. We also released the 17.6 million checkpoint (trained with 116.62 billion tokens) based on three criteria: (1) the loss curve slowly descended after 17.6 million steps during training (Extended Data Fig. 3a); (2) the losses are relatively stable on the validation and testing set between 15 million and 20 million steps, making 17.6 million optimal (Extended Data Fig. 3b,c); (3) the improvement in the performance of representative downstream tasks is very limited. For example, in the ncRPI task, the accuracy is 94.93% at checkpoint 17.6 million, which is only a marginal improvement compared to an accuracy of 94.78% at checkpoint 5.6 million (Extended Data Fig. 3d).

LucaOne embeddings-level analysis

Details of t-SNE datasets: the S1 dataset was curated from marine data available in CAMI2[49], selecting contigs with lengths ranging from 300 to 1,500 nucleotides. The contigs of each species were redundant by MMSeqs, employing a coverage threshold of 80% and sequence identity of 95%, culminating in a collection of 37,895 nucleic acid contigs from 12 species. We randomly selected 100 samples from each species, totalling 1,200 items for visualization.

The S2 dataset originated from clan data within Pfam, maintaining clan categories with a minimum of 100 Pfam entries, resulting in 189,881 protein sequences across 12 clan categories. For visualization, we randomly selected one sample for each Pfam entry in every clan, amounting to 2,738 samples.

The S3 dataset was selected from the UniProt database from 1 May 2023 to 16 December 2024, which does not overlap with the pre-training data of LucaOne (before 29 May 2022). This dataset focused on the lowest-tier GO annotations within the hierarchical annotation framework of the biological-processes subset, identifying

the 12 most prevalent terms at this foundational level. Each GO term randomly samples 100 sequences between 100 and 2,500 amino acids in length, resulting in 1,200 protein sequences across the 12 GO terms (Supplementary Note 2).

Convergence of nucleic acid and protein sequences for the same gene: we prepared an additional dataset comprising nucleic acid and protein sequences for the same genes and examined their correlations solely on embeddings. The results indicated that, despite nucleic acid and protein sequences not being paired during model training, those corresponding to the same gene demonstrated convergence within the LucaOne Embedding Space. More details in Supplementary Note 6 and Supplementary Fig. 12.

Task on pseudogene correction: we conducted a mask task prediction analysis (zero shot) on the data of the true gene (protein coding) and pseudogene pairs. The higher pseudogene correction rate and the true gene recovery rate demonstrated the model's ability to identify the differences between pseudogenes and functional genes. More details in Supplementary Note 7, and Supplementary Figs. 13 and 14.

Task on codon degeneracy: we designed an additional task based on influenza virus haemagglutinin sequence data to verify whether LucaOne can distinguish between synonymous and non-synonymous mutations in a zero-shot manner (more details in Supplementary Fig. 16).

Details of central dogma tasks

Dataset construction, original dataset: we devised an experimental task to determine whether LucaOne has established the intrinsic association between DNA sequences and their corresponding proteins. A total of 8,533 accurate DNA–protein pairs from 13 species were selected in the NCBI RefSeq database, each DNA sequence extending to include an additional 100 nucleotides in the 5′ and 3′ contexts, preserving intron sequences within the data. In contrast, we generated double the number of negative samples by implementing substitutions, insertions and deletions within the DNA sequences or altering amino acids in the protein sequences to ensure the resultant DNA sequences could not be accurately translated into their respective proteins, resulting in a total of 25,600 samples—DNA–protein pairs. Then the positive and negative samples were each subjected to random shuffles and subsequently divided into 32 equally sized subsets. Then these subsets were assigned to the training, validation and testing sets in a 4:3:25 ratio. For more details, see Extended Data Table 4 and 'Data availability'.

Analysis of misclassified samples: we analyse the misidentified samples from two perspectives—sequence and embedding. The relationship between sequence identity metrics and the prediction accuracy of the LucaOne embedding is presented in Extended Data Fig. 4a,b. Data details are presented in Supplementary Note 3. Extended Data Fig. 4a,b shows that the prediction accuracy of LucaOne for mutated sample pairs improved as sequence similarity decreased. We also evaluated the embedding distance alterations corresponding to modifications in nucleic acid and protein sequences by employing mean pooling to calculate these distances. As illustrated in Extended Data Fig. 4c,d, greater changes in embedding distances were correlated with improved predictive precision.

Dataset construction, two more species of urochordates: we incorporated two species with annotated reference genome urochordates (referred to as tunicate in the NCBI taxonomy) into our dataset: *Styela clava* (ASM1312258v2, GCF_013122585.1) and *Oikopleura dioica* (OKI2018_I68_1.0, GCA_907165135.1). For each of these urochordate species, 480 genes were randomly selected, and positive gene samples, nucleic acid negative samples and protein-negative samples were constructed using the same approach as in the original dataset. The same data shuffling and partitioning principles were applied and integrated with the original dataset to retrain the central dogma model. Data details and model performance are presented in Extended Data Table 3, Extended Data Fig. 5 and 'Data availability'.

Comparative performance analysis: upon integrating two additional urochordate species data, dataset version 2 as the model showed performance comparable to the original dataset models across all species except *C. intestinalis*. In particular, the F1 score for *C. intestinalis* improved significantly, despite the nearly unchanged accuracy. These findings suggest that supplementing the dataset with species that utilize a codon code similar to *C. intestinalis* enhances the model's sensitivity to DNA–protein correlations in these organisms while preserving its sensitivity to DNA–protein correlations in species adhering to the standard codon code. For more details, see Extended Data Table 3 and 'Data availability'.

CDS–protein task: in the current NCBI RefSeq database, genomes with complete intron annotations are limited, and the accuracy of intron predictions from alternative tools may directly impact model performance. To mitigate these challenges, coding sequence (CDS) regions corresponding to genes in the original dataset were extracted as intron-free nucleic acid sequences to perform the same task. See Supplementary Notes 4 for data details and Fig. 3d for analysis.

Task for cross-species homologous gene pairs: we designed an additional task related to the central dogma by modifying the negative samples in the original study. Instead of manually altering the sequences, the negative samples were replaced with homologous genes from closely related species. Please refer to Supplementary Notes 5 for details.

Downstream tasks details

Genus taxonomy annotation (GenusTax): this task is to predict which genus (taxonomy) the nucleic acid fragment may come from, which is very important in metagenomic analysis. A comparative dataset was constructed utilizing NCBI RefSeq, comprising 10,000 nucleic acid sequences, each extending 1,500 nucleotides and annotated with labels corresponding to 157 distinct genera (distributed as 33, 50, 29 and 45 across the four kingdoms: Archaea, Bacteria, Eukaryota and Viruses, respectively). The dataset was randomly segregated into training, validation and testing sets, adhering to an 8:1:1 partitioning ratio. It is important to note that while the LucaOne pre-training task utilized taxonomy annotations at the order level, the current task employs more granular genus-level annotations, thereby preventing label information contamination. This dataset was also employed for two additional analyses: predicting the taxonomy of sequences at the superkingdom and species levels. The details are presented in Extended Data Table 5.

Prokaryotic protein subcellular location (ProtLoc): this task is to predict the subcellular localization of proteins within prokaryotic cells, which has garnered substantial attention in proteomics due to its critical role[50]. It involves classifying proteins into one of six subcellular compartments: the cytoplasm, cytoplasmic membrane, periplasm, outer membrane, cell wall and surface, and extracellular space. Our approach adopted the same dataset partitioning strategy as DeepLocPro[26], a model based on experimentally verified data from the UniProt and PSORTdb databases. For this dataset, we additionally designed a task based on the corresponding nucleic acid embeddings of the proteins. The result showed that embeddings derived from nucleic acid sequences are applicable to the task related to their corresponding protein sequences. The dataset and analytical results are provided in Supplementary Notes 8.

Protein stability (ProtStab): the evaluation of protein stability is paramount for elucidating the structural and functional characteristics of proteins, which aids in revealing the mechanisms through which proteins maintain their functionality in vivo and the circumstances that predispose them to denaturation or deleterious aggregation. We utilized the same dataset from TAPE[51], which includes a range of de novo-designed proteins, natural proteins, mutants and their respective stability measurements. It is a regression task; each protein input (x) correlates with a numerical label ($y \in \mathcal{R}$), quantifying the protein's intrinsic stability.

Non-coding RNA family (ncRNAFam): ncRNA represents gene sequences that do not code for proteins but have essential functional and biological roles. The objective is to assign ncRNA sequences to their respective families based on their characteristics. For this purpose, we utilize the dataset from the nRC[52], which is consistent with the data employed in the RNAGCN[53] study. Our methodology adheres to the same data partitioning into training, validation and testing sets as done in these previous studies, enabling direct comparison of results. This project involves a multi-class classification challenge that encompasses 88 distinct categories.

Influenza A antigenic relationship prediction (InfA): one of the foremost tasks in influenza vaccine strain selection is monitoring haemagglutinin variant emergence, which induces changes in the virus's antigenicity. Precisely predicting antigenic responses to novel influenza strains is crucial for developing effective vaccines and preventing outbreaks. The study utilizes data from the PREDAC[54] project to inform vaccine strain recommendations. Each data pair in this study comprises two RNA sequences of the haemagglutinin fragment from distinct influenza strains, accompanied by corresponding antigenic relationship data. The objective is framed as a binary classification task identifying the antigenic similarity or difference between virus pairs.

Protein–protein interaction (PPI): the forecasting of PPI networks represents a significant area of research interest. Our study utilized the DeepPPI[55] database, whose positive dataset samples were sourced from the Human Protein Reference Database after excluding redundant interactions, leaving 36,630 unique pairs. This dataset was randomly partitioned into three subsets: training (80%), validation (10%) and testing (10%). The primary objective of this research is to perform binary classification of PPI sequences.

ncRNA–protein interactions (ncRPIs): an increasing number of functional ncRNAs, such as snRNAs, snoRNAs, miRNAs and lncRNAs, have been discovered. ncRNAs have a crucial role in many biological processes. Experimentally identifying ncRPIs is typically expensive and time-consuming. Consequently, numerous computational methods have been developed as alternative approaches. For comparison, we have utilized the same dataset as the currently best-performing study, ncRPI-LGAT[27]. It is a binary classification task involving pairs of sequences.

Comparison result details

We conducted a series of comparative experiments. According to Fig. 4, for all embedding methods, we compare whether the transformer encoder and two pooling strategies (max pooling and value-level attention pooling) were used on the model. At the hyperparameter level, we compared the number of encoder layers with the number of heads (4 layers with 8 heads and 2 layers with 4 heads), the peak learning rate of the Warmup strategy (1×10^{-4} and 2×10^{-4}), and the batch size (8 and 16). Extended Data Table 5 shows the result of comparing whether the encoder was used and which pooling method was used accordingly, and Supplementary Fig. 4 shows more metrics on comparison experiments.

In the ProtLoc task, LucaOne's accuracy is very close to that of the ESM2-3B. In the ncRPI task, the accuracy of the simple network with LucaOne's embedding matrix is less than that of ncRPI-LGAT[27] but higher than that of DNABert2 + ESM2-3B. In the other five tasks, we achieved the best results. It is better not to use an encoder for ProtLoc, InfA, PPI and ncRPI tasks. Using the max pooling strategy straightforwardly for the ncRNAFam and GenusTax tasks can obtain better results. We extended 2 tasks, 4 superkingdoms and 180 species prediction tasks for the genus classification task with the same sequence data. LucaOne's accuracy improved by 0.1 and 0.054, respectively. In particular, LucaOne is more effective than other large models in embedding sequences without an encoder.

Computational resources

The data processing and training operations for LucaOne were carried out on Alibaba Cloud Computing. In addition, several tasks related to further processing or downstream computing were performed on alternative computing platforms, including Yunqi Academy of Engineering (Hangzhou, China) and Zhejiang Lab (Hangzhou, China).

Reporting summary

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

Data availability

The pre-training dataset of LucaOne has been deposited into CNGB Sequence Archive (CNSA)[56] with accession number CNP0007266. The datasets of all downstream tasks and other supplementary materials are available at https://doi.org/10.5281/zenodo.15171943 (ref. 57).

Code availability

The model code of LucaOne is available at https://github.com/LucaOne/LucaOne. The embedding inference code is available at https://github.com/LucaOne/LucaOneApp and the downstream tasks are available at https://github.com/LucaOne/LucaOneTasks. The trained checkpoint files and an archived version of the above mentioned code repositories are available at https://doi.org/10.5281/zenodo.15171943 (ref. 57).

References

1. Crick, F. et al. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
2. Searls, D. B. The language of genes. *Nature* **420**, 211–217 (2002).
3. Darwin, C. *The Descent of Man, and Selection in Relation to Sex* Vol. 1. (John Murray, 1871).
4. Gimona, M. Protein linguistics—a grammar for modular protein assembly? *Nat. Rev. Mol. Cell Biol.* **7**, 68–73 (2006).
5. Barbieri, M. *The Organic Codes An Introduction to Semantic Biology* (Cambridge Univ. Press, 2002).
6. Pinker, S. *The Language Instinct* (William Morrow, 1994).
7. Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer. *Nat. Methods* **21**, 1422–1429 (2024).
8. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2021).
9. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
10. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
11. Elnaggar, A. et al. Ankh: optimized protein language model unlocks general-purpose modelling. Preprint at https://arxiv.org/abs/2301.06568 (2023).
12. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
13. Hou, X. et al. Using artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929–6942 (2024).
14. Yu, T. et al. Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
15. Zhou, Z. et al. DNABert-2: efficient foundation model and benchmark for multi-species genome. In *12th International Conference on Learning Representations* (ICLR, 2024).
16. Nguyen, E. et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. *Adv. Neural. Inf. Process. Syst.* **36**, 43177–43201 (2023).
17. Yang, F. et al. ScBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
18. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with evo. *Science* **386**, eado9336 (2024).

19. Li, Q. et al. Progress and opportunities of foundation models in bioinformatics. *Brief. Bioinform.* **25**, bbae548 (2024).
20. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
21. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**, 6000–6010 (2017).
22. Koonin, E. V. Why the central dogma: on the nature of the great biological exclusion principle. *Biol. Direct* **10**, 1–5 (2015).
23. Yockey, H. P. Information theory, evolution and the origin of life. *Inf. Sci.* **141**, 219–225 (2002).
24. He, Y. et al. KG-MTT-BERT: knowledge graph enhanced bert for multi-type medical text classification. Preprint at https://arxiv.org/abs/2210.03970 (2022).
25. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006).
26. Moreno, J., Nielsen, H., Winther, O. & Teufel, F. Predicting the subcellular location of prokaryotic proteins with DeepLocPro. *Bioinformatics* **40**, btae677 (2024).
27. Han, Y. & Zhang, S.-W. ncRPI-LGAT: prediction of ncRNA–protein interactions with line graph attention network framework. *Comput. Struct. Biotechnol. J.* **21**, 2286–2295 (2023).
28. Robbins, J. W. *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (JSTOR, 1996).
29. Chomsky, N. Three factors in language design. *Linguist. Inq.* **36**, 1–22 (2005).
30. Touvron, H. et al. Llama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
31. Liu, J. et al. Large language models in bioinformatics: applications and perspectives. Preprint at https://arxiv.org/abs/2401.04155v1 (2024).
32. Sapoval, N. et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **13**, 1728 (2022).
33. Vig, J. et al. BERTology meets biology: interpreting attention in protein language models. Preprint at https://arxiv.org/abs/2006.15222 (2020)
34. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
35. Nakano, F. K., Lietaert, M. & Vens, C. Machine learning for discovering missing or wrong protein function annotations: a comparison using updated benchmark datasets. *BMC Bioinform.* **20**, 485 (2019).
36. Alharbi, W. S. & Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum. Genomics* **16**, 26 (2022).
37. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
38. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).
39. Banzhaf, W., Machado, P. & Zhang, M. (eds) *Handbook of Evolutionary Machine Learning* Genetic and Evolutionary Computation (Springer, 2024).
40. Blanchard, A. E. et al. Automating genetic algorithm mutations for molecules using a masked language model. *IEEE Trans. Evol. Comput.* **26**, 793–799 (2022).
41. Ebrahim, A. et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.* **7**, 13091 (2016).
42. Vahabi, N. & Michailidis, G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front. Genet.* **13**, 854752 (2022).
43. Han, H. & Liu, X. The challenges of explainable AI in biomedical data science. *BMC Bioinform.* **22**, 443 (2022).
44. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9**, 1312 (2019).
45. Xiong, R. et al. On layer normalization in the transformer architecture. In *International Conference on Machine Learning* 10524–10533 (PMLR, 2020).
46. Su, J. et al. RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
47. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, L. K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Advances in North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186 (Association for Computational Linguistics, 2019).
48. Hoffmann, J. et al. Training compute-optimal large language models. In *Proc. 36th International Conference on Neural Information Processing Systems*, 30016–30030 (NeurIPS, 2022).
49. Meyer, F. et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
50. Xu, Q., Hu, D. H., Xue, H., Yu, W. & Yang, Q. Semi-supervised protein subcellular localization. *BMC Bioinform.* **10**, S47 (2009).
51. Rao, R. et al. Evaluating protein transfer learning with tape. *Adv. Neural. Inf. Process. Syst.* **32**, 9689–9701 (2019).
52. Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R. & Urso, A. nRC: non-coding RNA classifier based on structural features. *Biodata Min.* **10**, 27 (2017).
53. Rossi, E., Monti, F., Bronstein, M. & Liò, P. ncRNA classification with graph convolutional networks. In *Proc. 1st International Workshop on Deep Learning on Graphs: Methods and Applications* (DLG, 2019).
54. Du, X. et al. Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. *Nat. Commun.* **3**, 709 (2012).
55. Sun, T., Zhou, B., Lai, L. & Pei, J. Sequence-based prediction of protein-protein interaction using a deep-learning algorithm. *BMC Bioinform.* **18**, 277 (2017).
56. Wang, W. et al. The China National Genebank Sequence Archive (CNSA) 2024 update. *Hortic. Res.* **12**, 036 (2025).
57. He, Y. Generalized biological foundation model with unified nucleic acid and protein language. *Zenodo* 10.5281/zenodo.15171943 (2025).
58. Mock, F., Kretschmer, F., Kriese, A., Böcker, S. & Marz, M. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proc. Natl Acad. Sci. USA* **119**, e2122636119 (2022).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82341118). M.S. is funded by the Shenzhen Science and Technology Program (KQTD20200820145822023), the Guangdong Province 'Pearl River Talent Plan' Innovation and Entrepreneurship Team project (2019ZT08Y464), and the Guangzhou National Laboratory Major Project (GZNL2023A01001). Y.P. is funded by the National Natural Science Foundation of China (NSFC) Basic Research Project for Doctoral Students (grant number 323B2018). We thank J. Wang, D.-C. Ma and D.-Z. Shi for computational resource coordination. We thank H.-W. Zhang for maintaining computational resources and optimizing specific computing tasks at Yunqi Academy of Engineering (Hangzhou). We thank Y.-Q. Liu and M. Zhou for their participation in the subsequent technical validation in conjunction with this research. We thank X.-J. Du, W.-C. Wu, J.-Y. Yang and S.-Q. Mei from Sun Yat-sen University (Shenzhen) for a valuable conversation on the development of the method, especially on understanding the downstream tasks. We thank C. Darwin, R. Dawkins, S. Pinker and D. Dennett for their

profound insights that led to the early conceptual foundations of this study and guided its development pathway.

Author contributions

Conceptualization: Y.H., Z.L. and M. S. Model development and data preparation for LucaOne: Y.H., Y.W., Y.P. and Yichang Chen. Downstream tasks understanding and models training: Y.H., P.F., Y. Shan and Yihao Chen. Original draft: Y.H., Z.L. and P.F. Writing—review and editing: all authors. Graphic presentation design: Y. L. and Y.H. Engineering leadership and resource acquisition: Z. Zeng and J.Y. Science leadership and resource acquisition: J.L., E.C.H., Z. Zhou, F.Z. and Y. Shu. Supervision: Y.H., M.S. and Z.-L.

Competing interests

Y.H., Z.L., P. F. and J.Y. have filed an application for a patent covering the work presented. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-025-01044-4.

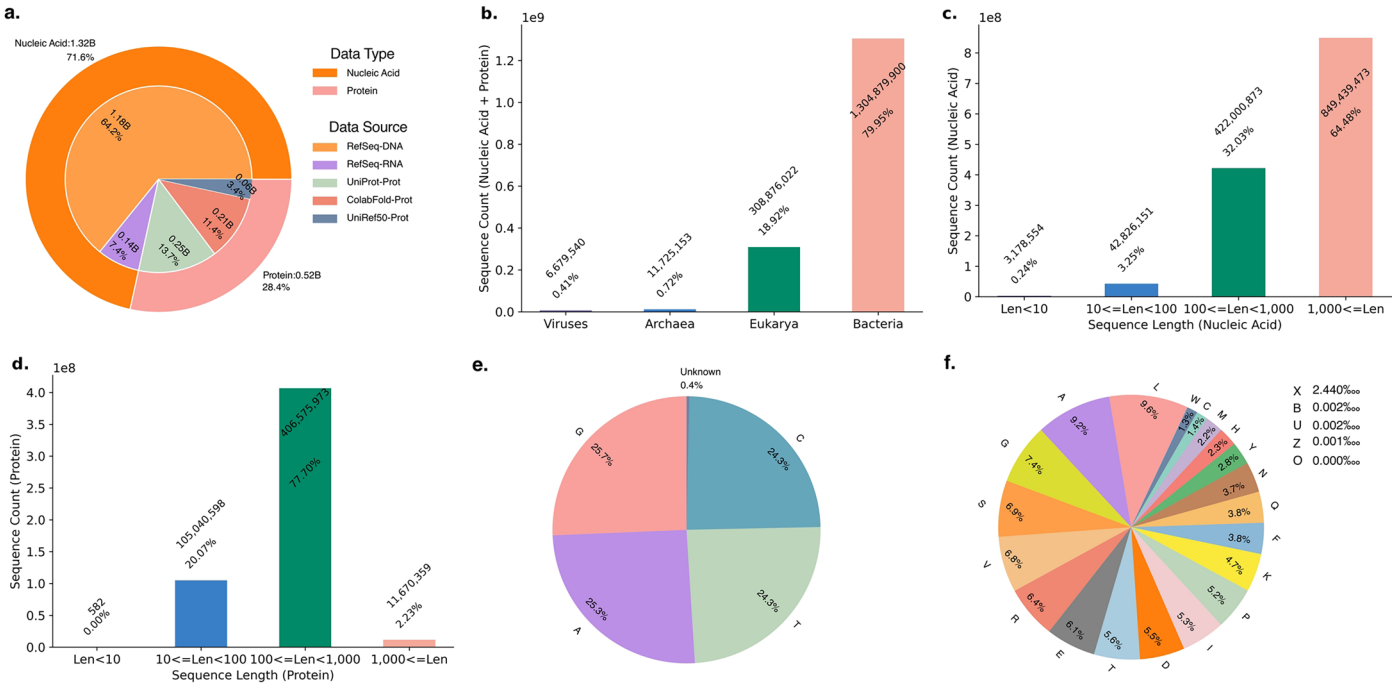
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01044-4.

Correspondence and requests for materials should be addressed to Yong He, Mang Shi or Zhaorong Li.

Peer review information *Nature Machine Intelligence* thanks anonymous reviewer(s) for their contribution to the peer review of this work.

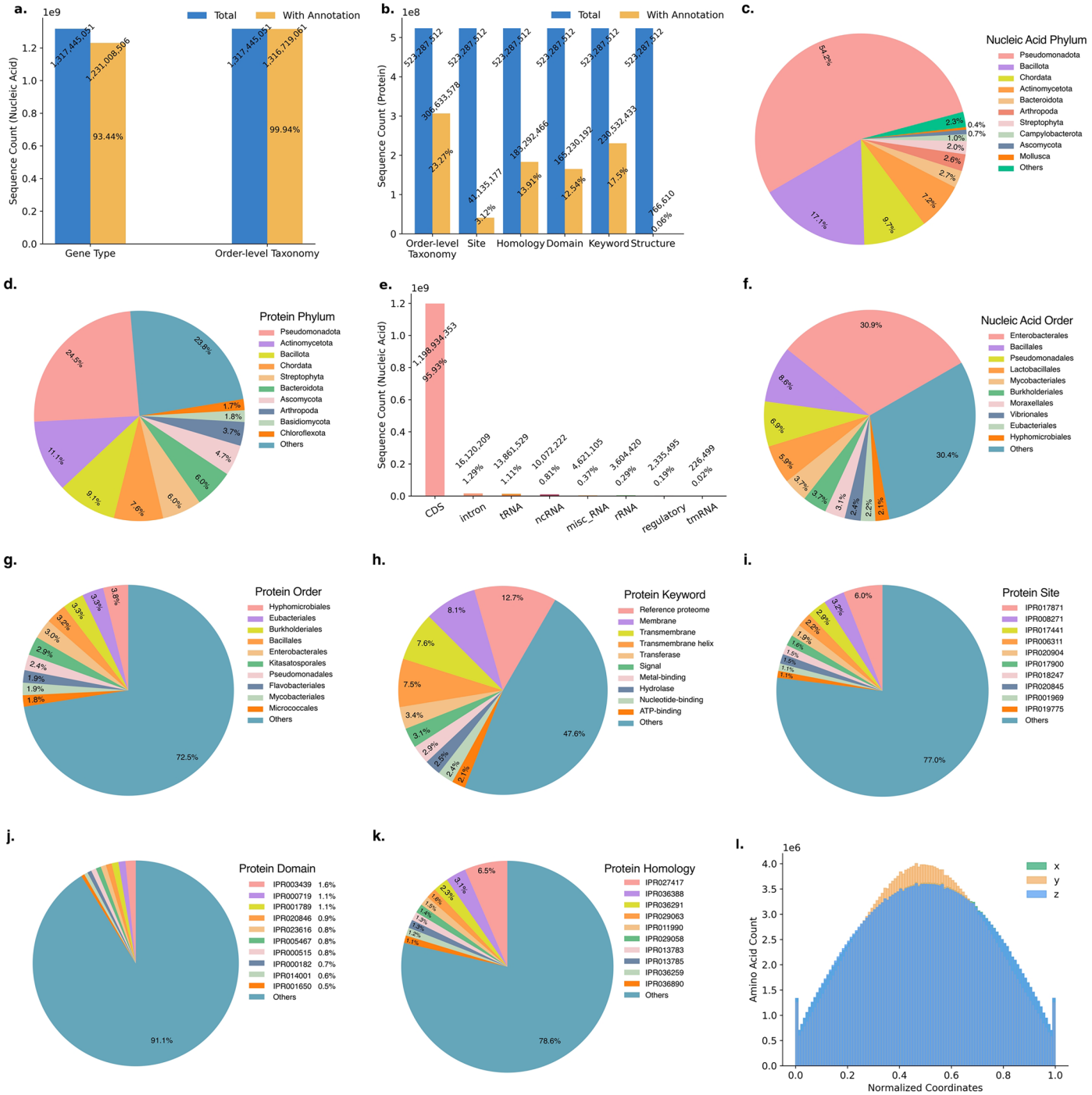
[1]Apsara Lab, Alibaba Cloud Intelligence, Alibaba Group, Hangzhou, China. [2]Centre for Infection and Immunity Study (CIIS), School of Medicine (Shenzhen), Sun Yat-sen University, Shenzhen, China. [3]Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai, China. [4]Alibaba Health Information Technology, Alibaba Group, Hangzhou, China. [5]State Key Laboratory of Advanced Drug Delivery and Release Systems and Innovation Institute for AI in Medicine, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China. [6]Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [7]College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China. [8]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, New South Wales, Australia. [9]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary 33 Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China. [10]Key Laboratory of Pathogen Infection Prevention and Control, Peking Union Medical College, Ministry of Education, Beijing, China. [11]School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, China. [12]State Key Laboratory for Biocontrol, Centre for Infection and Immunity Studies, School of Medicine, Sun Yat-sen University, Shenzhen, China. ✉e-mail: sanyuan.hy@alibaba-inc.com; shim23@mail.sysu.edu.cn; lzr098@gmail.com

https://doi.org/10.1038/s42256-025-01044-4



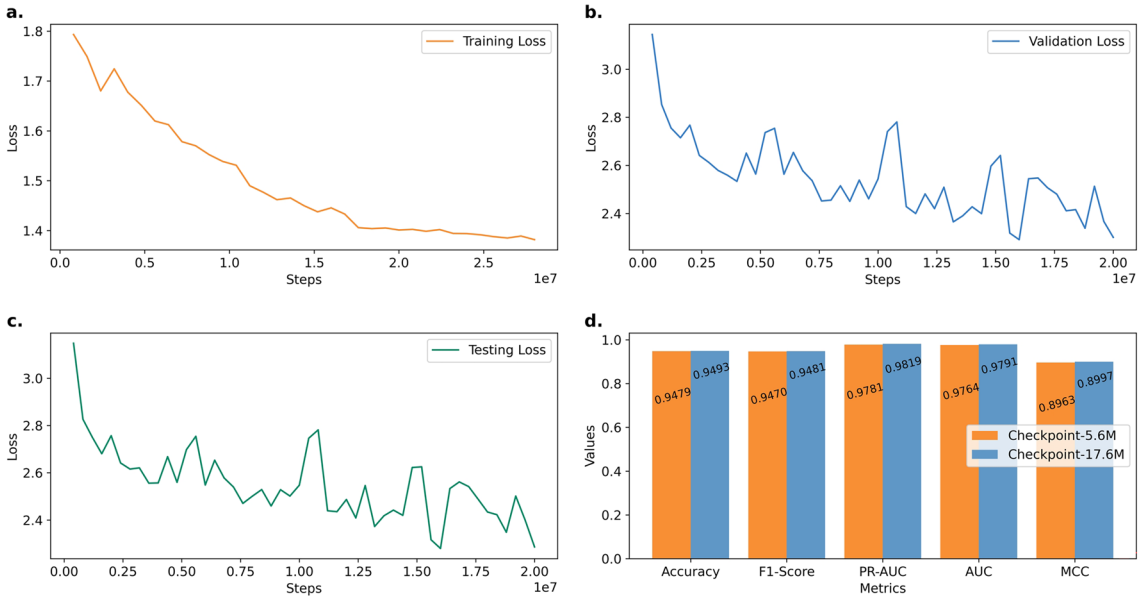
Extended Data Fig. 1 | Overall statistics on pre-training data of LucaOne. **a**. Sequences (DNA, RNA, and proteins) were derived from RefSeq, UniProt, ColabFoldDB, and UniRef50. **b**. The data (nucleic acids and proteins) involved four superkingdom types: Viruses, Archaea, Eukarya, and Bacteria, of which Bacteria accounted for the most. **c**. The sequence length distribution of nucleic acids, with the most being more than 1,000. **d**. The sequence length distribution of proteins, with the maximum length ratio between 100 and 1,000. **e**. The proportion of five nucleotides ('A', 'T', 'C', 'G', and 'Unknown') in nucleic acid sequences ('U' compiled with 'T' in RNA) and the four identified nucleotides were close in proportion. **f**. The proportion of the 20 standard amino acid letters and five other letters (including four non-standard amino acids and 'X' for unknown amino acid) in the protein sequence, and Leucine has the highest proportion.

https://doi.org/10.1038/s42256-025-01044-4



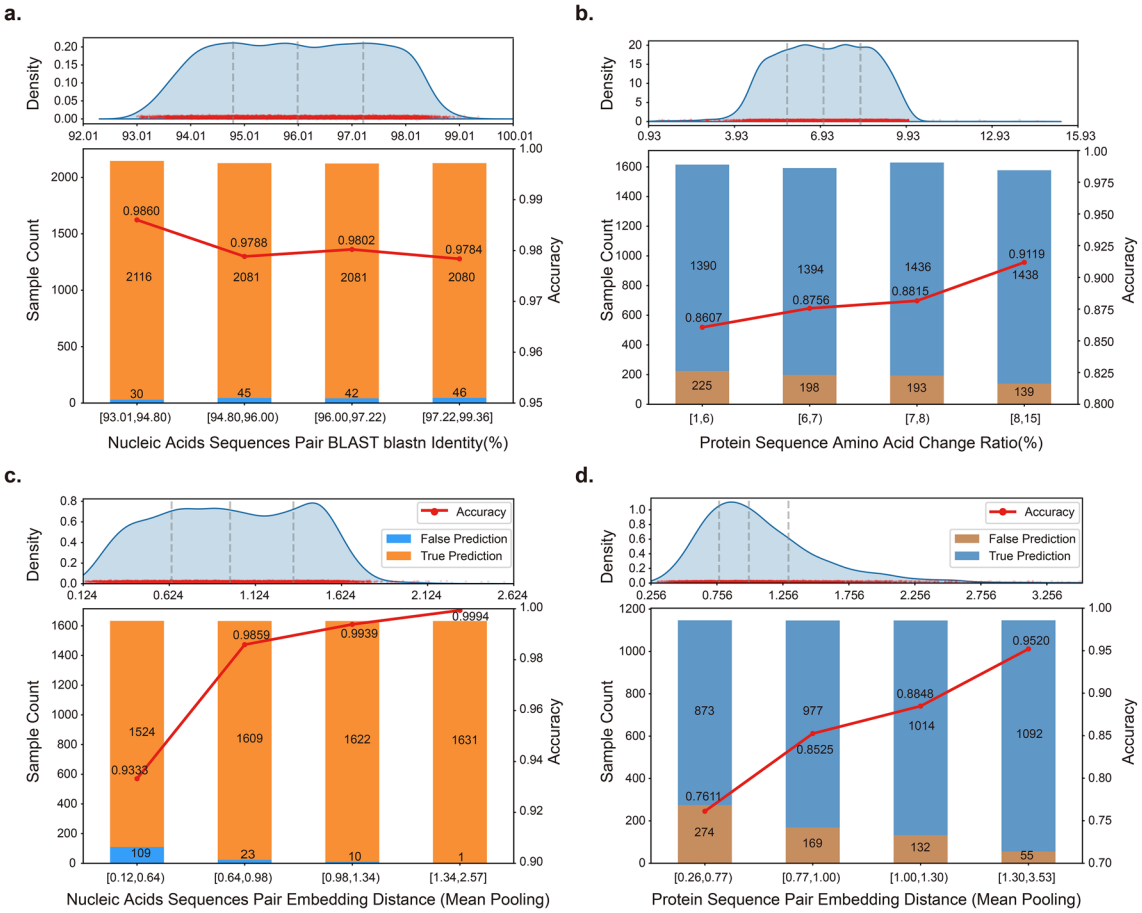
Extended Data Fig. 2 | Annotation statistics on pre-training data of LucaOne. **a**. The proportion of genome region types and order-level taxonomy in nucleic acid. Most sequences have both types of annotation information. **b**. The proportion of the count of sequences with each of the selected six annotations, including order-level taxonomy, keyword, site, domain, homology, and tertiary structure, of which the proportion of sequence count with tertiary structure is tiny. **c.** and **d**. The proportion of sequence counts in the top 10 phylum-level taxonomy of nucleic acids and proteins, respectively. **e**. The distribution of eight selected genome region types in nucleic acids, of which the CDS region is the most. **f.** and **g**. The proportion of sequence counts in the top 10 order-level taxonomy (total 2,196 categories) of nucleic acids and proteins, respectively. **h**–**k**. The proportion of protein sequence counts in the top 10 keywords (total 1,179 categories), the top 10 site types (total 946 categories), the top 10 domain types (total 13,717 categories), and the top 10 homology types (total 3,442 categories), respectively. **l**. The *coord-(x, y, z)* distribution of C_α-*atom* position (local normalization within a protein chain). It is very similar to the normal distribution. The distribution has a long tail in **c**–**f**. The distribution is ladder decreasing in **g**–**k**.

https://doi.org/10.1038/s42256-025-01044-4



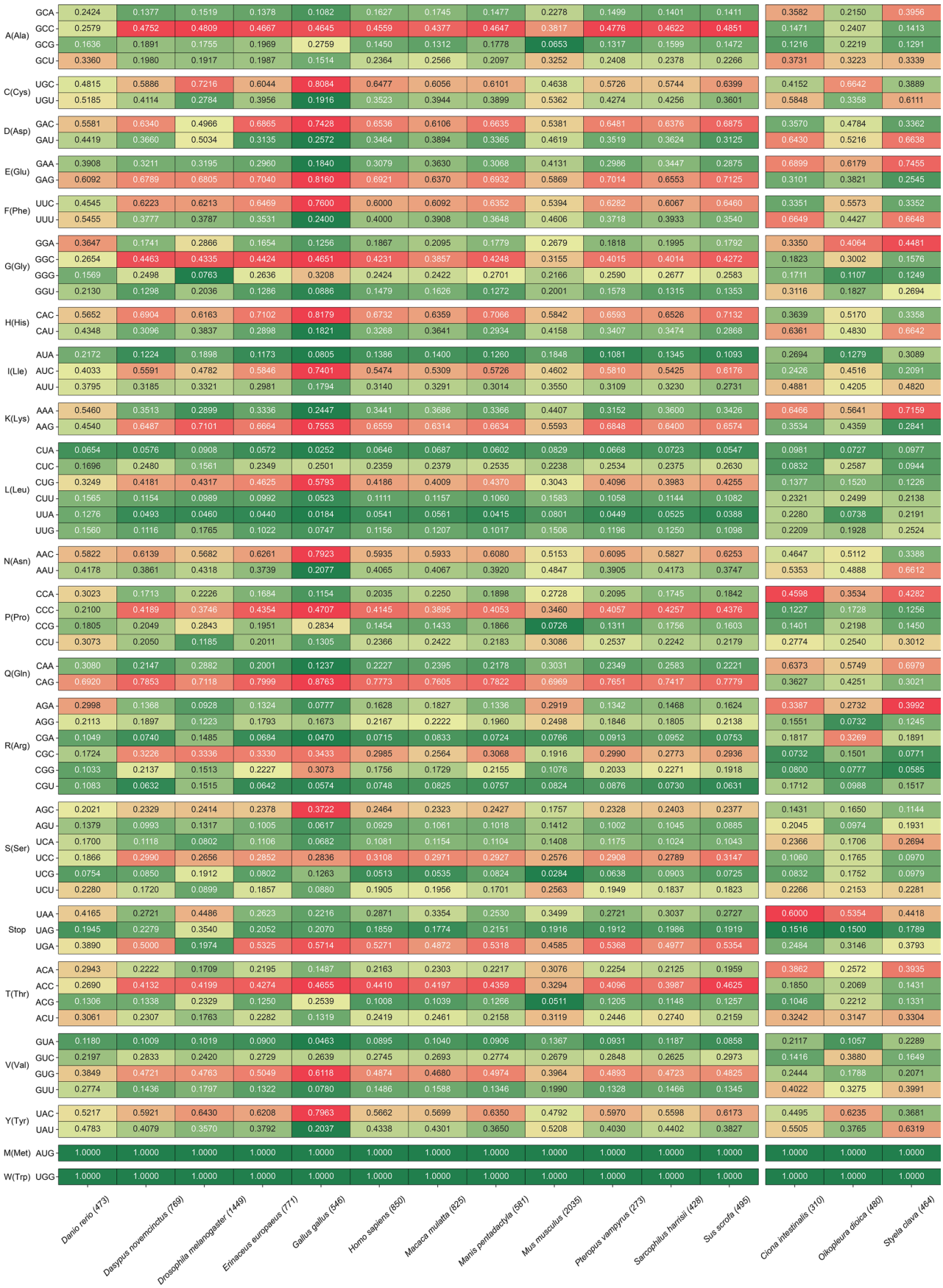
Extended Data Fig. 3 | LucaOne 17.6M checkpoint selection criteria. a. The loss trend during training. **b.** and **c**. The trend of loss on the validation and testing sets. **d**. Performance evaluation between 5.6M and 17.6M checkpoints on downstream tasks - ncRPI.

https://doi.org/10.1038/s42256-025-01044-4



Extended Data Fig. 4 | Identity between positive and negative samples and prediction accuracy on the Central Dogma task. a. and **b**. The relationship between sequence identity metrics and LucaOne model prediction accuracy: NCBI blastn sequence identity for nucleic acid and protein sequences before and after mutation. **c.** and **d**. Embedding Euclidean distances based on mean pooling and their prediction accuracy in LucaOne for nucleic acid and protein sequences before and after mutation. Upper panels: Sample distributions across sequence similarity, change ratio, or embedding Euclidean distance ranges. Lower panels: Prediction counts and accuracy of the LucaOne embedding within each respective range. Note: Data for **a.** and **b**. includes all nucleic acid and protein-negative samples from the validation and testing sets. Data for **c.** and **d**. includes only positive-negative sample pairs that are both present in the combined validation and testing datasets. Divide the statistical intervals of the metrics into quartiles according to the data distribution.

https://doi.org/10.1038/s42256-025-01044-4



Extended Data Fig. 5 | Codon usage heatmap. Based on the dataset for 15 species - Original Dataset plus two urochordate species. The distribution of different codons for a single amino acid totals 100%. Coloured representations indicate the relative proportions, where red signifies higher proportions and green signifies lower proportions.

https://doi.org/10.1038/s42256-025-01044-4

Extended Data Table 1 | Embedding-based clustering metrics on three datasets

DataSet	Embedding	ARI	AMI	HS	CS	V-measure	FMI
12-Marine Species (S1)	Multi-OneHot	0.1258	0.2782	0.2918	0.2949	0.2934	0.1998
	DNABert2	0.1313	0.2556	0.2686	0.2743	0.2714	0.2063
	LucaOne	**0.3410**	**0.5524**	**0.5488**	**0.5760**	**0.5621**	**0.4040**
12-Clan Pfam (S2)	Multi-OneHot	0.1062	0.1860	0.1763	0.2167	0.1944	0.2341
	ESM2-3B	0.0731	0.1526	0.1617	0.1595	0.1606	0.1644
	LucaOne	**0.2379**	**0.4179**	**0.4228**	**0.4240**	**0.4234**	**0.3163**
12-Go Terms (S3)	Multi-OneHot	0.0942	0.1920	0.1934	0.2315	0.2107	0.2004
	ESM2-3B	0.1079	0.2582	0.2596	0.2928	0.2752	0.2027
	LucaOne	**0.1535**	**0.3779**	**0.3707**	**0.4159**	**0.3920**	**0.2474**

The clustering scores (using K-Means++) of the four embedding methods on the S1, S2, and S3 datasets.

https://doi.org/10.1038/s42256-025-01044-4

Extended Data Table 2 | Performance comparison for LucaOne and other embedding tools on the Central Dogma task

Embedding	Network	Accuracy		F1-Score		AUC	
		Original	CDS-Protein	Original	CDS-Protein	Original	CDS-Protein
One-Hot	Transformer + Pooling + FC	0.6667	-	0.0000	-	0.4943	-
Random Init	Transformer + Pooling + FC	0.6662	-	0.0015	-	0.5085	-
DNABert2 + ESM2-3B	Pooling + FC	0.7309	0.7400	0.5689	0.6661	0.7491	0.7959
LucaOne (Gene/Prot)	Pooling + FC	0.8048	0.8163	0.6804	0.7806	0.8616	0.9055
LucaOne	Pooling + FC	**0.8453**	**0.8451**	**0.7392**	**0.8089**	**0.9101**	**0.9298**

LucaOne was not only compared with several existing embedding methods but also with itself, which was trained using nucleic acids and proteins separately (LucaOne-Gene/LucaOne-Prot). LucaOne, with mixing training, obtained the best performance for both the original dataset and the CDS-Protein dataset.

https://doi.org/10.1038/s42256-025-01044-4

Extended Data Table 3 | Comparative performance analysis on the Central Dogma task

Species (Samples count)	Accuracy				F1-Score			
	LucaOne		DNABert2 + ESM2-3B		LucaOne		DNABert2 + ESM2-3B	
	Original	New with two more urochordates species	Original	New with two more urochordates species	Original	New with two more urochordates species	Original	New with two more urochordates species
Gallus gallus (1177)	**0.8318**	0.8139	0.7077	0.7128	**0.7235**	0.7060	0.5401	0.5541
Manis pentadactyla (1430)	**0.8587**	0.8517	0.7385	0.7329	**0.7449**	0.7402	0.5590	0.5431
Drosophila melanogaster (3004)	0.8192	**0.8517**	0.6967	0.6974	0.6691	**0.6912**	0.4644	0.4706
Homo sapiens (1806)	**0.8283**	0.8267	0.7248	0.7143	0.7103	**0.7295**	0.5510	0.5335
Pteropus vampyrus (593)	**0.8331**	0.8314	0.7656	0.7622	0.7097	**0.7207**	0.6040	0.5960
Mus musculus (4375)	0.8761	**0.8805**	0.7543	0.7586	0.8170	**0.8325**	0.6635	0.6583
Dasypus novemcinctus (1769)	**0.8491**	0.8428	0.7462	0.7507	**0.7512**	0.7477	0.6044	0.5995
Sus scrofa (1336)	0.8540	**0.8570**	0.7358	0.7388	0.7572	**0.7745**	0.5803	0.5770
Macaca mulatta (2192)	**0.8412**	0.8335	0.7441	0.7336	0.7298	**0.7330**	0.5896	0.5780
Sarcophilus harrisii (1011)	**0.8467**	0.8328	0.7270	0.7211	**0.7304**	0.7216	0.5274	0.5284
Erinaceus europaeus (1706)	**0.8540**	0.8429	0.7421	0.7345	**0.7585**	0.7500	0.5956	0.5878
Danio rerio (1263)	**0.8345**	0.8314	0.7110	0.6896	0.6967	**0.7230**	0.4808	0.4368
Ciona intestinalis (738)	**0.7927**	0.7873	0.6965	0.6978	0.5174	**0.6391**	0.3043	0.3957
Styela clava (1260)	**0.6817***	0.6627	0.6817*	0.6159	0.3725*	**0.4311**	0.2533*	0.2461
Oikopleura dioica (1260)	**0.6747***	0.6476	0.6747*	0.6373	0.3315*	**0.3951**	0.2293*	0.3462
Average (All samples)	**0.8451**	0.8212	0.7318	0.7188	**0.7391**	0.7172	0.5706	0.5411

Comparative performance analysis (validation and testing set) of the models across diverse species datasets (Sample counts in brackets). The original dataset includes 13 species, and the new dataset adds two more urochordate species data. F1-score and accuracy are calculated and presented. The top right * indicates the predictive performances of the model trained by the original version of the Central Dogma dataset (w/o *Oikopleura Dioica* and *Styela Clava* data). More details in Data Availability.

https://doi.org/10.1038/s42256-025-01044-4

Extended Data Table 4 | Details on downstream validation tasks

Task	Task Type	Input Type	Train/Valid/ Test Size	Seq Length (Max/Min/Median)
Central Dogma	Binary-Class(2)	DNA-Protein	3,200/2,400/20,000	2,455-617/ 309-11/ 1,273-260
SpeciesTax	Multi-Class(180)	DNA	8,000/1,000/1,000	1,500/1,500/1,500
GenusTax	Multi-Class(157)	DNA	8,000/1,000/1,000	1,500/1,500/1,500
SupKTax	Multi-Class(4)	DNA	8,000/1,000/1,000	1,500/1,500/1,500
ProtLoc	Multi-Class(6)	Protein	9,915/1,991/1,131	5,627/8/396
ProtStab	Regression	Protein	53,614/2,512/12,851	50/43/43
ncRNAFam	Multi-Class(88)	RNA	105,864/17,324/25,342	200/24/114
InfA	Binary-Class(2)	RNA-RNA	4,645/581/581	1,690-1,690/ 984-984/ 1,095-1,095
PPI	Binary-Class(2)	Protein-Protein	59,766/7,430/7,425	33,423-33,423/ 24-24/ 465-437
ncRPI	Binary-Class(2)	RNA-Protein	16,660/-/4,164	3,999-3,678/ 52-49/ 1,858-414

Details of 10 downstream tasks, including task name, task type, input type of task, sample number of training set, validation set, test set, and sequence length statistics of each task.

https://doi.org/10.1038/s42256-025-01044-4

Extended Data Table 5 | Detailed results of the testing set on downstream validation tasks (results of the better pooling method for each task with or without encoder)

Task	Input	Method	Encoder	Better Pooling	Acc/SRCC
CentralDogma	DNA-Protein	DNABert2 + ESM2-3B	W/O Encoder	Attention	0.7309
			Encoder	Attention	0.7192
		LucaOne(Gene/Prot)	W/O Encoder	Attention	0.8048
			Encoder	Attention	0.7938
		LucaOne	W/O Encoder	Attention	**0.8453**
			Encoder	Attention	0.8125
SpeciesTax	DNA	BERTax★ [58]	-	-	-
		DNABert2	W/O Encoder	Attention	0.519
			Encoder	Max	0.696
		LucaOne	W/O Encoder	Attention	0.713
			Encoder	Attention	**0.750**
GenusTax	DNA	BERTax★ [58]	-	-	0.124
		DNABert2	W/O Encoder	Attention	0.551
			Encoder	Max	0.767
		LucaOne	W/O Encoder	Attention	0.765
			Encoder	Max	**0.817**
SupKTax	DNA	BERTax★ [58]	-	-	0.816
		DNABert2	W/O Encoder	Attention	0.805
			Encoder	Attention	0.848
		LucaOne	W/O Encoder	Attention	0.940
			Encoder	Attention	**0.947**
ProtLoc	Protein	DeepLocPro▲ [26]	-	-	0.92
		ESM2-3B	W/O Encoder	Attention	**0.9496**
			Encoder	Max	0.9408
		LucaOne	W/O Encoder	Attention	0.9452
			Encoder	Max	0.9310
ProtStab	Protein	TAPE▲ [51]	-	-	0.73
		ESM2-3B	W/O Encoder	Attention	0.7556
			Encoder	Attention	0.7102
		LucaOne	W/O Encoder	Attention	0.7512
			Encoder	Attention	**0.7718**
ncRNAFam	RNA	RNAGCN▲ [53]	-	-	0.90
		DNABert2	W/O Encoder	Attention	0.9036
			Encoder	Max	0.9610
		LucaOne	W/O Encoder	Attention	0.9743
			Encoder	Max	**0.9864**
InfA	RNA-RNA	PREDAC• [54]	-	-	0.9010
		DNABert2	W/O Encoder	Attention	0.9966
			Encoder	Attention	0.9966
		LucaOne	W/O Encoder	Attention	**1.0**
			Encoder	Attention	0.9983
PPI	Protein-Protein	DeepPPI▲ [55]	-	-	0.9719
		ESM2-3B	W/O Encoder	Attention	0.9764
			Encoder	Attention	0.9745
		LucaOne	W/O Encoder	Attention	**0.9774**
			Encoder	Attention	0.9751
ncRPI	RNA-Protein	ncRPI-LGAT▲ [27]	-	-	**0.966**
		DNABert2 + ESM2-3B	W/O Encoder	Attention	0.9460
			Encoder	Attention	0.9332
		LucaOne	W/O Encoder	Attention	0.9479
			Encoder	Attention	0.9380

The top right ⋆ indicates inference using the trained method, the top right ▲ indicates direct use of the results in its paper, and the top right • indicates repetition using its method and higher than the results in the paper. BERTax from ref. 58.

nature portfolio

Corresponding author(s): Zhaorong Li; Mang Shi; Yong He

Last updated by author(s): Apr 4, 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	The data for model training are all from public data depository, as illustrated clearly in the manuscript and also at our Zenodo: https://doi.org/10.5281/zenodo.15171943 and Github: https://github.com/LucaOne/LucaOne
Data analysis	The source code and compiled standalone version of the software are available at https://github.com/LucaOne/LucaOne Other softwares used in this study include:Biopython(1.80), Scikit-learn(1.2.1), Torch(1.13.1), Transformes(4.26.0),

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

All the data for training and testing are illustrated clearly both in the manuscript and at Zenodo: https://doi.org/10.5281/zenodo.15171943

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

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 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.*
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.*

Behavioural & social sciences study design

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

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were 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. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*

Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

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

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were 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 collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
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.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☒ No

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		Methods	
n/a	Involved in the study	n/a	Involved in the study
☐	☐ Antibodies	☐	☐ ChIP-seq
☐	☐ Eukaryotic cell lines	☐	☐ Flow cytometry
☐	☐ Palaeontology and archaeology	☐	☐ MRI-based neuroimaging
☐	☐ Animals and other organisms		
☐	☐ Clinical data		
☐	☐ Dual use research of concern		

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s) — *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication — *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination — *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) — *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

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

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals — *For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Reporting on sex — *Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*

Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates	Describe the experimental replicates, specifying number, type and replicate agreement.
Sequencing depth	Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.
Antibodies	Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.
Peak calling parameters	Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.
Data quality	Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.
Software	Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.
Instrument	Identify the instrument used for data collection, specifying make and model number.
Software	Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.
Cell population abundance	Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.
Gating strategy	Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	Indicate task or resting state; event-related or block design.
Design specifications	Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.
Behavioral performance measures	State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*

Field strength *Specify in Tesla*

Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*