Predicting the Evolutionary and Functional Landscapes of Viruses with a Unified Nucleotide-Protein Language Model: LucaVirus

Yuan-Fei Pan[1,2,3]*#, Yong He[4]*, Yu-Qi Liu[5,6,7,8], Yong-Tao Shan[3,8], Shu-Ning Liu[9], Xue Liu[2], Xiaoyun Pan[1], Yinqi Bai[7], Zan Xu[4], Zheng Wang[4], Jieping Ye[4], Edward C. Holmes[10], Bo Li[1,11]#, Yao-Qing Chen[9]#, Zhao-Rong Li[4,8]#, Mang Shi[2,3,12]#

1. State Key Laboratory of Wetland Conservation and Restoration, National Observations and Research Station for Wetland Ecosystems of the Yangtze Estuary, Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, and Institute of Eco-Chongming, School of Life Sciences, Fudan University, Shanghai, China
2. National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.
3. State Key Laboratory for Biocontrol, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.
4. Apsara Lab, Alibaba Cloud Intelligence, Alibaba Group, Hangzhou, China
5. Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China.
6. HIM-BGI Omics Center, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, China.
7. BGI Research, Hangzhou, China.
8. Zhejiang Lab, Hangzhou, China.
9. School of Public Health (Shenzhen), Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China
10. School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia
11. Yunnan Key Laboratory of Plant Reproductive Adaptation and Evolutionary Ecology and Centre for Invasion Biology, School of Ecology and Environmental Science, Yunnan University, Kunming, Yunnan, China
12. Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China.

*These authors contributed equally.

#Corresponding authors

Abstract

Predicting the evolution and function of viruses is a fundamental biological challenge, largely due to high levels of sequence divergence and the limited knowledge available in comparison to cellular organisms. To address this, we present LucaVirus, a unified, multi-modal foundation model specifically designed for viruses. Trained on 25.4 billion nucleotide and amino acid tokens encompassing nearly all known viruses, LucaVirus learns biologically meaningful representations that capture the relationships between nucleotide and amino acid sequences, protein/gene homology, and evolutionary divergence. Building on these interpretable embeddings, we developed specialized downstream models to address key challenges in virology: (i) identify viruses hidden within genomic "dark matter", (ii) characterize enzymatic activities of unknown proteins, (iii) predict viral evolvability, and (iv) discover antibody drugs for emerging viruses. LucaVirus achieves state-of-the-art performance in tasks (i), (iii), and (iv), and matches the leading models in task (ii) with one-third the parameter size. These findings demonstrate the power of a unified foundation model to comprehensively decode the viral world. LucaVirus is a new tool in AI-driven virology, offering an efficient and versatile platform for board applications from virus discovery to functional predictions.

Introduction

The global expansion of metagenomic and metatranscriptomic sequencing has revealed a rapidly growing diversity of viruses across animals (*1*) and diverse environments (*2, 3*). Despite their often very small size, viral genomes and proteomes encode nearly all biologically relevant information required for their host interaction (*4*), immune evasion (*5*), and transmission (*6*)—making them a compact, yet comprehensive, blueprint of life. However, translating this genomic wealth into biological insight remains challenging (*7*), mainly due to rapid evolutionary rates (*8*) which leads to high levels of sequence divergence (*9*), particularly in viruses with RNA genomes. Conventional bioinformatic methods, reliant on sequence homology and conserved domain detection, often fall short when it comes to viruses, creating a major bottleneck between viral discovery, functional understanding and public health applications.

Recent advances in artificial intelligence, especially in large language models (LLMs), have created new opportunities to learn complex biological representations directly from sequences. Models such as the ESM series (*10, 11*) and the Evo series (*12, 13*) have shown impressive performance in protein structure and function prediction across cellular organisms. However, their applicability in virology remains limited due to two critical factors. First, most existing models are unimodal, focusing exclusively on either protein or nucleotide sequences. Viral genomes, however, encode rich evolutionary information (*14, 15*), such as codon bias, which may indicate host range and is not present in protein sequences. Conversely, because nucleotide sequences evolve rapidly, protein sequences, which are more conserved, are better suited for analyses of gene homology and function. Second, viral sequences exhibit distinct evolutionary constraints and composition biases that differ fundamentally from those in cellular life (*16*). For instance, viral genomes are typically highly compact, with greater coding density than cellular organisms. Many RNA viruses have also evolved mechanisms such as polyprotein and programmed ribosomal frameshifting to encode multiple distinct proteins from a single mRNA, which are rarely found in cellular organisms (*17, 18*). The viral "language" is therefore underrepresented, and often misrepresented, in foundation models trained on prokaryotic and

eukaryotic data sets.

Given these challenges, there is a pressing need for a domain-specific foundation model tailored to the unique evolutionary and functional landscape of viruses. Viral genomes exhibit exceptionally strong genotype-phenotype coupling, where even single-nucleotide or amino acid changes can dramatically alter host specificity, immune evasion, and replication fitness. This makes them particularly well-suited for representation learning approaches that extract nuanced functional signals directly from sequence data. Furthermore, the lack of structural and functional annotations for viral proteins demands the use of models that are capable of generalizing across highly divergent, sparsely annotated sequences, while preserving biological interpretability.

To address these challenges, we present "LucaVirus", a unified genome-protein language model designed for viruses. By jointly modelling nucleotide and amino acid sequences and pre-training on one of the most comprehensive viral sequences created to date, LucaVirus captures the distinct "grammatical" and "semantic" features of viral evolution. This foundation model supports a wide array of downstream tasks, from identifying divergent viral sequences and inferring protein function to modeling fitness landscapes and predicting antibody-antigen interactions. Ultimately, LucaVirus offers a generalizable and interpretable AI framework that connects viral sequences with biological functional, with the aim of advancing both basic virological research and pandemic preparedness.

Results

LucaVirus: a unified model for viral genomes and proteins

LucaVirus is a unified genome-protein language model with one billion parameters, designed to interpret viral sequence at single-nucleotide or single-amino acid resolution (Fig. 1A). As knowledge of viral diversity continues to expand, a comprehensive and scalable model is urgently needed for capturing their genetic and functional characteristics. LucaVirus addresses this need by integrating nearly all known viral sequences into a single, cohesive framework

106 (Fig. 1B-D).

107 To support such broad representation, we assembled a large and diverse training data set
108 containing 25.4 billion tokens—comprising 23.7 billion nucleotide tokens and 1.6 billion
109 amino acid tokens—from 15.7 million sequences (10.4 million nucleotide and 5.2 million
110 protein). This data set is nearly twice the size of the NCBI Virus database and 30% larger than
111 UniProt (*19*) in terms of base or residue counts of non-redundant sequences (Fig. 1B).
112 Importantly, the data set incorporates recent discoveries from large-scale viral discovery studies
113 (*9, 20-24*), including 1,663,163 novel "species-level" taxa (defined by 80% ANI) not yet
114 cataloged in reference databases (Table S1). Collectively, the data set comprises a wide
115 spectrum of viral taxa, with the majority belonging to the double-strand (ds) DNA viruses
116 (kingdom *Heunggongvirae*, 27% count by tokens), RNA viruses (*Orthornavirae*, 26%),
117 retroviruses (*Pararnavirae*, 20%), others (single-strand (ss) DNA viruses and others, 6%) and
118 unclassified viruses (21%) (Fig. 1C). The data set also spans a remarkably broad host range
119 (Fig. 1D), covering all six cellular kingdoms – animals (48%), bacteria (25%), plants (12%),
120 protists (2%), fungi (2%), archaea (1%), and unidentified hosts (22%).

121 We designed a tailored model architecture for LucaVirus to effectively learn from such a
122 large and diverse viral data set. Building upon our previously developed general-purpose
123 biological language model, LucaOne (*25*), we reduced the number of transformer encoder
124 layers from 20 to 12 and attention heads from 40 to 20, decreasing model size from 1.8 billion
125 to 1 billion parameters. These refinements enhanced computational efficiency and supported
126 an extended context window, increased from 1,280 to 3,072 tokens. This expanded window
127 enables LucaVirus to capture long-range dependencies and richer contextual information,
128 covering the full length of 99% of viral proteins and 92% of coding DNA/RNA sequences in
129 the SwissProt reference database (*19*).

130 Finally, to better model sequence-function association, LucaVirus was pre-trained using a
131 self- and semi-supervised strategy that extends beyond standard masked language modeling
132 (MLM) (*26*). This approach, which explicitly promotes learning sequence-function
133 relationships, was validated in several previous studies (*25, 27, 28*) and is particularly effective

134 at decoding the functional landscape of viral genomes (see Methods).

135

136 LucaVirus learns interpretable representations of viral sequences

137 As an encoder model, LucaVirus extracts biologically meaningful features from viral
138 sequences and embeds them into a high-dimensional space. To demonstrate the interpretability
139 of LucaVirus embeddings, we correlated the model's outputs with biologically relevant and
140 human-interpretable sequence features.

141 Our results suggest that LucaVirus nucleotide embeddings capture key aspects of the
142 relationship between nucleotide and amino acid sequences. As shown by principal components
143 analysis (PCA), LucaVirus embeddings reflect multiple layers of genomic organization,
144 including differentiation between coding and non-coding regions, codon phase, and
145 synonymous codons. For example, the synonymous codons ATT and ATC (that both encode
146 isoleucine and differ by a T-to-C transition) formed tight clusters, yet were distinctly separated
147 from ATA (a T/C-to-A transversion that also encodes isoleucine). These patterns indicate that
148 LucaVirus embeddings are sensitive to both functional coding features and mutational
149 dynamics, implicitly capturing the relationship between nucleotide and protein sequences
150 without explicit supervision.

151 For protein embeddings, PCA visualization demonstrated that LucaVirus embeddings
152 distinguish protein families with single-amino-acid resolution (Fig. S2) and outperforms other
153 models, including LucaOne (*25*), ESM2-3B (*10*), and ESMC-600M (*11*), in separating
154 sequences by protein families (e.g., RNA vs. DNA viruses).

155 Our data also showed that LucaVirus embedding is evolutionary informative. By
156 correlating embedding distances with genetic distances (e.g., p-distance, nucleotide
157 substitutions per site) across 5,566 reference protein families and 938 "dark matter" protein
158 families not included in pre-training (Fig. 2, Fig. S3; Table S3-S4), we observed a significant
159 and robust association, supporting evolutionary interpretability. In the reference protein data
160 set, LucaVirus achieved a Pearson correlation of 0.73, comparable to ESM (0.75) and ESMC
161 (0.77), although lower than LucaOne (0.86). However, in the unseen dark matter data set,

LucaVirus outperformed all models with a correlation of 0.71, surpassing LucaOne (0.67), ESM2-3B (0.46), and ESMC (0.46), demonstrating superior generalization (Fig. 3A). While ESM and ESMC captured broad sequence divergence, their embeddings showed limited ability to distinguish between aligned and unaligned positions. In contrast, LucaVirus achieved high correlations for nucleotide sequences—0.90 (RefSeq) and 0.76 (dark matter)—outperforming LucaOne (0.77 and 0.64, respectively). These results inspired a novel sequence alignment algorithm based on LucaVirus embeddings, which replaces fixed scoring matrices (e.g., BLOSUM62) with learned similarity metrics. This approach yields local alignments comparable to the Smith-Waterman algorithm, further validating the evolutionary relevance of the learned embedding space (Fig. 2B, Fig. S4).

Together, LucaVirus encodes rich, interpretable representations of viral sequences at both the nucleotide and amino acid levels. Its ability to recapitulate biological structures, resolve synonymous variation, reflect mutational dynamics, and generalize to evolutionary divergent data provides a robust foundation for diverse downstream applications.

LucaVirus helps resolves key challenges in virology

Viral genomics presents persistent challenges, such as detecting highly divergent sequences and annotating uncharacterized proteins. To showcase the practical utility of LucaVirus, we systematically assessed its performance in four representative tasks that address these key challenges: (i) revealing the diversity of viruses hidden within “dark matter”, (ii) characterizing unknown proteins, (iii) forecasting virus evolvability, and (iv) predicting viral antigen-antibody binding affinity.

For each task, we trained lightweight, task-specific downstream networks that take as input the embedding matrix generated by pre-trained language models (see Methods). The downstream network adopts a streamlined architecture comprising a single attention-based pooling layer, a fully connected layer, and a task-specific prediction head (Fig. S4). Notably, the upstream language model remains frozen during downstream training, substantially reducing the number of trainable parameters. This design improves efficiency and mitigates

190 the need for large-scale labeled data sets—a major advantage given the scarcity of
191 experimentally validated biological data.
192
193 ***Revealing viral dark matter***
194 Building on our previous work (*9*), we targeted two hallmark viral proteins—the RNA-
195 dependent RNA polymerase (RdRP) and viral capsid protein—to uncover currently
196 unclassified RNA virus sequences present within genomic dark matter. The RdRP is a highly
197 conserved enzyme critical for RNA virus replication and rarely found in cellular organisms
198 (*29*), making it a reliable marker for RNA virus discovery (*30*). In contrast, capsid proteins,
199 essential for virion structure, are often highly divergent, posing challenges for conventional
200 homology-based detection (*31*).
201 Leveraging these two hallmark proteins, the Lucavirus model reveals the viral composition
202 of dark matter by classifying query proteins as viral RdRP or capsids (Fig. 3A; Table S5-S6).
203 On a validation data set of 23,000 RdRP and 48,000 capsid proteins, the LucaVirus-based
204 model achieved remarkable accuracies of 99.98% for RdRP and 99.83% for capsids (Fig. 3B).
205 In terms of precision and recall, the model recorded 99.44% and 99.81%, respectively.
206 Importantly, it matches current state-of-the-art accuracy recorded by the ESM2-3B-based
207 model while utilizing only one-third the number of parameters, reducing computational
208 complexity from 17 TFLOPs to 5 TFLOPs.
209 To evaluate LucaVirus in a real-world dark matter mining scenario, we benchmarked its
210 sensitivity in viral capsid prediction using putative proteins translated from recently reported
211 novel RNA viruses (*9*). A total of 68 structurally validated novel capsid proteins (see Methods)
212 were tested. The LucaVirus-based model accurately identified all the 68 proteins, achieving a
213 100% recall rate. In comparison, the ESM2-3B-based model recalled 63%, while the
214 conventional homology-based BLAST (*32*) search recovered only 2.9% (Fig. 3C). Notably,
215 LucaVirus detected all types of structurally divergent capsid proteins, including nucleocapsids,
216 single jelly roll (SJR), and double jelly roll (DJR) capsids (Fig. 3D). These results demonstrate
217 that although LucaVirus performs similarly to other models on curated data sets, it generalizes

218 better to real-world viral dark matter, highlighting its robustness and practical utility in virus
219 discovery.
220
221 ***Characterization of unknown proteins***
222 Although metagenomics has accelerated the discovery of novel viruses, deciphering their gene
223 and protein functions lags far behind. The high sequence divergence of viral proteins renders
224 conventional homology-based methods largely ineffective, severely limiting our ability to
225 effectively interrogate sequence data.
226 LucaVirus demonstrated exceptional performance in predicting Enzyme Commission (EC)
227 numbers from viral coding sequences (Fig. 3E; Table S7), achieving 99.97% accuracy, with an
228 F1 score of 0.9889. Notably, it matched the accuracy performance of ESM2-3B while operating
229 with a much smaller parameter size (Fig. 3F).
230 To access LucaVirus's utility for functional protein discovery, we applied it to identify
231 putative novel proteases (EC 3.4.-.-) from novel RNA virus genomes recently uncovered by an
232 AI-based virus discovery study (*9*). LucaVirus detected 8,074 proteins with putative protease
233 activity, from which 46 clusters (each with > 10 members) were selected for structural
234 validation using ColabFold (*33*). Eight clusters displayed moderate pLDDT scores and
235 significant structural similarity to known proteases (Foldseek (*34*), E-value < 1e-5), despite
236 sharing only ~12% average sequence identity with their closest known sequences. To confirm
237 these predictions, two predicted proteases were expressed and purified using an *E. coli*
238 expression system, and their protease activity was confirmed through *in vitro* assays (Figure
239 SX). Remarkably, LucaVirus accomplished this solely from sequence data, without relying on
240 structural inputs. This demonstrates its ability to infer sequence-to-function relationships and
241 underscores its advantage over slower, less reliable structure-based methods for discovering
242 novel functional proteins.
243
244 ***Forecasting virus evolvability***
245 Characterizing the fitness landscape of proteins or genomes is fundamental to evolutionary

246 biology and essential for predicting viral evolvability. Unlike tasks that generalize protein
247 function across viral species, modeling fitness landscapes requires capturing the subtle
248 biological effects of intra-species variation. To evaluate LucaVirus's capability in this context,
249 we utilized deep mutational scanning (DMS) data, which systematically introduce all possible
250 single mutations into protein-coding sequences and experimentally quantifies their impact on
251 fitness traits such as binding affinity (*35*).

252 In a DMS data set evaluating the binding affinity of SARS-CoV-2 variants (*36*), the
253 LucaVirus-based prediction model outperformed other leading language models (Fig. 4; Table
254 S8-S9), achieving a Spearman correlation coefficient of 0.92, compared to 0.87 for LucaOne,
255 0.79 for ESM2-3B, and 0.60 for ESMC-600M. Strikingly, the LucaVirus-based model
256 maintained strong performance even when using nucleotide sequences as input, with a
257 Spearman coefficient of 0.88. This robustness likely stems from LucaVirus's interpretable
258 representations. Visualization of mean-pooled embeddings revealed a distinct clustering of
259 variants and single amino acid substitutions, suggesting that LucaVirus effectively captures
260 underlying phylogenetic relationships. This was further supported by a correlation analysis
261 between embedding similarity and sequence divergence (Fig. S6).

262 Notably, the above models were trained on experimental data from three early SARS-CoV-
263 2 variants and tested on two later, unseen variants, demonstrating robust generalization. This
264 highlights how integrating high-throughput assays like DMS with LucaVirus allows
265 researchers to experimentally probe a limited set of strains while accurately extrapolating to
266 emerging variants. Such capabilities are invaluable for anticipating viral evolution and
267 improving preparedness against emerging viruses.

268

269 ***Prediction of Ag-Ab binding affinity***

270 LucaVirus also demonstrated excellent performance in identifying antibodies from repertoires
271 targeting emerging viruses. On a publicly available coronavirus antibody database, LucaVirus
272 achieved a 93% accuracy in predicting the binding potential of SARS-CoV-2 spike protein to
273 antibodies in the validation data set.

To assess its practical applicability, we simulated antibody discovery from human B-cell receptor repertoires (Fig. S7; Table S13). From single B-cell sequencing data of patients recovered from SARS-CoV-2 infection, 80 antibodies were randomly selected and evaluated for binding to the Wuhan-Hu-1 (i.e., the reference SARS-CoV-2 sequence) and Omicron spike proteins. For the Wuhan-Hu-1 strain, the model achieved an accuracy of 84.6%, a recall of 68.4%, and a precision of 68.4%. For the Omicron variant, it attained an accuracy of 89.8%, a recall of 70.0%, and a precision of 63.6%. This performance surpassed existing protein language models and structure-based prediction methods such as AlphaFold3 (*37*) combined with CSM-Ab (*38*) and DG-affinity (*39*) in determining the binding capability of specific antibody sequences to viral proteins (Fig. 5A-B; Table S10). Interestingly, visualization of attention weights from the model's pooling layers revealed a correlation between high-attention regions and known spike protein epitopes, further suggesting its ability to identify epitopes (Fig. 5F).

To evaluate LucaVirus's generalizability, the model was retrained using alternative input types—spike-coding CDS nucleotide sequences and full viral genomes—instead of spike protein sequences (Table S11-S12). The CDS-based model achieved an accuracy of 84.3%, while the whole-genome-based model yielded a comparable accuracy of 83.6%. These results demonstrate that LucaVirus effectively generalizes across diverse sequence types.

Together, LucaVirus represents a transformative tool for strain-specific antibody discovery, offering high accuracy, robust generalizability, and the ability to extract critical epitope information from diverse genomic inputs. By streamlining the identification of effective antibodies, LucaVirus holds potential to expedite therapeutic development and enhance global preparedness for emerging viral threats.

Computational efficiency of LucaVirus enables LLMs for all

To assess the feasibility of developing a more computationally efficient model for resource-limited applications, we investigated how model performance scales with parameter size. Specifically, we trained two smaller variants—referred to as the "base" and "small" models—

alongside the default "large" model (Fig. S7; Table S2). The base model comprises 112 million parameters, while the small model has 62 million. These reductions were achieved by scaling down the embedding dimension from 2,560 (in the large model) to 256 and 128 for the base and small models, respectively, while keeping all other hyperparameters constant. This downscaling reduces the size of the output embedding matrix and significantly lowers the overall computational complexity.

We evaluated the performance of the base and small models on receptor-binding domain (RBD) binding affinity prediction (Table S14). The small model (62M parameters) achieved a Spearman correlation coefficient of 0.77, closely matching the ESM2-3B model correlation coefficient of 0.78. The base model (112M parameters) attained a Spearman correlation coefficient of 0.90, comparable to the LucaOne-1.8B model (0.89). Although performance generally improves with larger parameter sizes, the 62M-parameter small model achieved results on par with a 3B-parameter model. This highlights the remarkable efficiency of the LucaVirus architecture and its promise for use in computationally limited settings. Notably, the large model ran efficiently on a gamer-grade GPU with 24GB of memory, while the small and base models operate seamlessly on an 8GB GPU, making LucaVirus accessible to all biologists at low cost.

Discussion

Modeling the evolutionary and functional landscape of viruses remains a fundamental challenge, largely due to their extensive sequence diversity and the limited biological understanding relative to cellular organisms. To address this, we present LucaVirus—the first unified, multi-modal foundation model specifically tailored for viruses. By integrating nearly all available viral sequence data across both nucleotide and protein modalities, LucaVirus establishes a comprehensive framework for interpreting viral biology at multiple levels.

Our findings demonstrate that LucaVirus learns biologically meaningful and interpretable representations of viral sequences, enabling robust performance across a wide range of critical virological tasks. The model's embeddings capture rich biological features—including codon

structure, synonymous mutations, and phylogenetic relationships—providing a solid foundation for understanding sequence-function relationships. This ability likely stems from semi-supervised pretraining, which allows the model to extract functional insights from large-scale unlabeled data. While previous studies have shown that training with explicit sequence-function mappings significantly enhances performance (*25, 27, 28*), such labeled data are limited compared to the abundance of raw sequences. Unlike natural language, which is self-descriptive (*40*), biological sequences do not inherently reveal their functional roles. This fundamental difference may explain why auto-regression strategies like MLM are less efficient in capturing biological function, underscoring the importance of explicit functional supervision in modeling biological sequences. In this context, LucaVirus stands out by combining large-scale pre-training with biologically informed representations, enabling it to match or exceed the performance of much larger models while maintaining greater parameter efficiency and broader applicability in biological domains.

LucaVirus can be used for a new approach in computational virology, where foundation models serve not only as predictive tools but also as hypothesis-generating platforms. Analogous to organoids in experimental biology, LucaVirus serves as an "*in silico* virology lab," enabling rapid, large-scale virtual experimentation to investigate viral evolution, functional dynamics, and host interactions. Researchers can introduce mutations into viral sequences and observe the model-predicted outcomes, facilitating both mechanistic insights and hypothesis testing. The use of AI to simulate biological systems is well-established in fields such as neuroscience, where neural models replicate mouse behavior to decode underlying mechanisms (*41*). Similarly, LucaVirus provides a versatile platform for virology, capable of accelerate discoveries, enhancing mechanistic understanding, and reducing reliance on time- and resource-intensive experimental procedures.

Finally, as an encoder model, LucaVirus extracts feature matrices from input sequences and does not possess the capability to generate viral sequences, thereby mitigating biosecurity risks associated with generative models. This encoder architecture might also enhance interpretability, allowing us to directly associate feature matrices with biological hypotheses

under investigation, facilitating straightforward validation and analysis. In contrast, generative models often exhibit more complex interpretability challenges and pose greater biosecurity risks due to their sequence-generating potential.

In summary, LucaVirus demonstrates the value of domain-specific foundation models in virology. By integrating diverse viral sequence data into unified, interpretable representations, it addresses key challenges that span viral diversity, functional roles, evolutionary patterns and processes, and disease associations. As such, LucaVirus contributes to our fundamental understanding of the virosphere and offers practical tools to accelerate diagnostics and therapeutic development in the face of future viral outbreaks.

Materials and Methods

Model architecture

The LucaVirus architecture builds upon our general-purpose biological language model, LucaOne (*25*), with key hyperparameter modifications tailored to the unique properties of viral sequences. Briefly, LucaOne is a 1.8-billion-parameter transformer encoder-only model pre-trained on nucleic acid and protein sequences from 169,861 species spanning all domains of life, using 10 biologically informed semi-supervised tasks to capture unified representations of protein and genome "language". We opted to train an encoder model rather than a generative model based on two key considerations. First, an encoder model extracts learned features into a representation matrix, which facilitates interpretability by enabling straightforward analysis of the matrix. Second, we are aware of the risks associated with generating viral sequences that could be misused. An encoder model lacks the capability to generate sequences, thereby mitigating this concern.

To construct LucaVirus, we first determined the optimal model size based on the available pretraining tokens (see "Pre-training Data set" section), adopting a parameter-to-token ratio of 1:25. This yielded a model with one billion parameters, a scale well supported by the pre-training corpus—which contains twice the non-redundant tokens of the NCBI Virus database and 30% more than UniProtKB, thereby encompassing nearly all known viral sequence

information to date. This extensive data set supports the adequacy of the selected parameter size.

To preserve representational power while reducing the parameter count from 1.8B (LucaOne) to 1B (LucaVirus), we retained the embedding dimension of 2,560 but reduced the number of transformer layers from 20 to 12 and the prediction head size from 40 to 20. Additionally, to better accommodate viral sequence length, we extended the model's maximum context length from 1,280 tokens (LucaOne) to 3,072 tokens (LucaVirus)—optimized for NVIDIA A100 (80GB) hardware. This expanded context length covers the full-length sequence of 99.1% of viral proteins in UniProtKB.

Pre-training data set

We curated OpenVirus, a comprehensive, large-scale data set of viral sequences used to train the LucaVirus model. This data set comprises 15.7 million viral sequences, totaling 25.4 billion tokens—including 23.7 billion nucleotide tokens from 10.4 million sequences and 1.6 billion amino acid tokens from 5.2 million protein sequences. Nucleotide sequences were primarily sourced from the NCBI Virus database and seven independent viral diversity studies (*9, 20-24*), ensuring inclusion of sequences not available in NCBI. Protein sequences were obtained from the UniProtKB (*19*) and MGnify (*42*) databases.

The OpenVirus data set covers all known viral taxa. The major groups include: double-strand (ds) DNA viruses (27% of sequences), RNA viruses (26%), reverse-transcribing viruses (20%), single-strand (ss) DNA viruses and others (6%), and unclassified viruses (21%). These four groups collectively account for 94% of the total sequence count. The data set includes viruses infecting all three domains and six kingdoms of cellular life, including animals (48%), bacteria (25%), plants (12%), protists (2%), fungi (2%), archaea (1%), and unknown hosts (22%).

In addition to sequence data, the training data set integrates corresponding annotation information. For nucleotide sequences, annotations were primarily derived from GenBank GBFF files, from which coding region start sites and order-level taxonomic labels were

extracted. Sequences from other sources that lacked annotations were retained with missing fields left blank. Protein sequence annotations were obtained from UniProt and InterPro, with InterPro contributing details on functional sites, conserved domains, and homologous superfamilies. Taxonomic classification was based on 63 order-level labels, with missing entries assigned a default label of -100. Functional annotations include: (i) 946 functional site categories (e.g. active sites, binding sites, conserved sites, and post-translational modifications), (ii) 3,460 homologous superfamilies, (iii) 1,659 conserved domains, and (iv) 603 functional keyword groups.

Pre-training procedures

LucaVirus employs a semi-supervised pre-training strategy, building on the framework established by LucaOne (*25*). The model initializes its corresponding layers with weights derived from LucaOne's latest training checkpoint at step 1,760,000. The pre-training process integrates self-supervised masked language modeling (MLM) with seven biologically relevant supervised tasks to enhance the model's ability to capture diverse biological features. These tasks are categorized as follows:

Sequence-level classification tasks: (i) Order taxonomy prediction for nucleotide sequences; (ii) Order taxonomy prediction for protein sequences; and (iii) UniProt functional keyword prediction for protein sequences.

Token-level classification tasks: (i) gene prediction for nucleotide sequences; (ii) protein homologous superfamily annotation; (iii) protein conserved domain annotation; and (iv) protein active site prediction.

The pre-training parameter configuration aligns with that of LucaOne, with modifications to accommodate differences in data scale and computational resources. Task weights are assigned as follows: protein MLM, nucleic acid MLM, and protein keyword prediction tasks are weighted at 1.0, while the remaining supervised tasks are weighted at 0.2. Training is conducted with a learning rate of 2×10^{-4}, incorporating a warm-up phase, a batch size of 8, and gradient accumulation over 32 steps. We trained LucaVirus using eight A100 GPUs on Alibaba

Cloud for a total of 3,800,000 steps (approximately 1.9 epochs) over a duration of 70 days. total of 49,194,607,334 tokens was trained at 3,800,000 steps. Average loss was saved every 4,000 steps to evaluate the training progress.

Scaling experiments

To evaluate how model performance scales with parameter size, we conducted a series of scaling experiments. In addition to the default "large" model, we trained two smaller variants: a "base" model with 112 million parameters (112M) and a "small" model with 63 million parameters (63M). Parameter reduction was achieved by decreasing the embedding dimension from 2,560 in the large model to 512 for the base model and 256 for the small model, while keeping all the other architectural hyperparameters unchanged. We specifically chose to reduce the embedding dimension, as it lowers computational complexity and reduces the size of the output embedding matrix, making it more resource-efficient for users with limited computational and storage resources.

The pre-training procedures for the base and small models followed the same protocol as the large model. Due to resource constraints, the base and small models were trained on four GPUs in parallel. The base model was trained for one epoch, while the small model for two epochs. To ensure fair comparison, we evaluated all three models—the small, base, and large variants—using their respective checkpoints from the end of first epoch. Performance was assessed on downstream tasks involving RBD binding affinity prediction. Details regarding the dataset and training procedure can be found in the "Prediction of virus evolvability" methods section.

Embedding analyses

Nucleotides: To evaluate LucaVirus's representation of nucleotide sequences, 30 annotated reference genome sequences of *Enterovirus* spp. were downloaded from NCBI/GenBank. These 30 sequences were input into the LucaVirus model to generate feature matrices, where each column represents the vector representation (i.e., feature point) of an individual nucleotide within the model's learned feature space. Principal component analysis

468 (PCA) was then applied to reduce the dimensionality of these nucleotide-level embeddings for
469 visualization. To explore the biological relevance of the learned feature, the feature points were
470 colored based on the four attributes: (i) nucleotide identity (A, T, C, G), (ii) codon position (1st,
471 2nd, or 3rd base), (iii) the amino acid encoded by the corresponding codon, and (iv) identity of
472 synonymous codon. This analysis will assess how well the model's embedding space captures
473 biologically meaningful patterns underlying the relationships between nucleic acids and
474 proteins, codon structure, and protein translation.

475 ***Proteins:*** To further evaluated the model's capacity to represent protein sequences, we
476 compared LucaVirus to mainstream models that contain eukaryotic virus sequences in their
477 training data set, such as LucaOne (ref), ESM2 (*10*) and ESMC (*11*). Specifically, all viral
478 reference protein sequences were obtained from NCBI RefSeq, and grouped into protein
479 families using an unsupervised clustering approach. Pairwise sequence similarities were
480 computed using DIAMOND software (version 2.1.4) (*32*), followed by clustering via the
481 Markov clustering algorithm (*43*), as implemented in the VOGDB pipeline (*44*). From the
482 resulting clusters, we selected protein families containing more than 20 sequences, and
483 randomly sampled 10 families for further analysis. Within each family, 10 protein sequences
484 were chosen randomly, yielding a total of 100 protein sequences. These were input into the
485 large model to generate per-residue feature embeddings (i.e., column vectors in the feature
486 matrix). PCA was then applied to reduce the dimensionality of these embeddings for
487 visualization. This analysis was used to examine the model's ability to distinguish between
488 different protein families based on its learned representation space.

489 ***Correlation analyses:*** To quantitatively assess whether the embedding matrix learned by
490 LucaVirus capture meaningful evolutionary information, we conducted correlation analyses
491 between embedding distances (cosine and Euclidean distance) and metrics of genetic
492 divergence (p-distance and maximum likelihood estimates of the number of substitutions). We
493 used 5,566 viral orthologous protein families and their corresponding coding DNA sequences
494 (CDS) from RefSeq, alongside 938 protein families from novel RNA viruses discovered in our
495 previous AI-based virus mining study (*9*) that was not pre-trained (i.e., the "dark matter" data

set, no significant BLAST hit to pre-training proteins). Protein families were defined by *de novo* clustering following the pipeline described by VOGDB study (*44*). Mean cosine and Euclidean distances were computed for pairs of homologous proteins and compared with traditional multiple-alignment results.

LLM-based sequence alignment experiments: To validate that the sequence representations learnt by LucaVirus capture meaningful evolutionary signals, we developed LMAlign—a novel pairwise alignment algorithm that achieves comparable performance to traditional fixed scoring matrix methods through the use of embedding-based similarity measures. Specifically, the two input sequences were first processed by LucaVirus to generate their respective embedding matrices, with dimensions L1 * 2,560 and L2 * 2,560, where L1 and L2 denote the sequence lengths and 2,560 is the embedding dimension. Each column vector in these matrices represents the contextual embedding of an amino acid or nucleotide token.

Next, we computed a similarity matrix of shape L1 * L2 by calculating the cosine similarity between all pairs of column vectors from the two embedding matrices. This dynamic similarity map replaces the traditional fixed scoring matrix (e.g. transition-transversion scoring matrix for nucleotide and BLOSUM62 for protein alignments). The remainder of the alignment procedure follows the framework of the classical Smith-Waterman algorithm, applying dynamic programming to identify the optimal local alignment guided by the learned similarity scores.

A general modeling framework for downstream tasks

The framework includes two universal modeling strategies designed for different downstream prediction tasks: (i) Single-sequence input tasks, which include the tasks related to "revealing viral dark matter", "viral protein function prediction", and "viral fitness prediction", and (ii) multi-sequence input tasks such as the "virus antigen–antibody binding prediction" also covered here. All four predictive models discussed in this here are built upon this general-purpose architecture.

Universal Model Architecture for Single-Sequence Input: This model accepts a single

524 nucleotide or protein sequence as input and processes it through two parallel channels: (i) a
525 language model embedding channel, and (ii) an optional sequence modeling channel. The
526 embedding channel utilizes a pre-trained large language model (e.g., LucaVirus, ESM2,
527 amongst others) to generate a sequence embedding matrix, which is then pooled—using mean
528 pooling, max pooling, or value-level attention pooling (VALP) (*45*)—into a feature vector.
529 During training, the large language model's parameters remain frozen. In parallel, the sequence
530 modeling channel applies a trainable Transformer encoder to extract an additional feature
531 vector from the input. The final prediction is generated by concatenating the pooled outputs
532 from both channels and feeding them through a fully connected (FC) layer to generate the final
533 prediction.

534 *Universal Model Architecture for Multi-Sequence Input*: This model is designed for
535 prediction tasks involving two or more input sequences, such as protein–protein interactions.
536 Each input sequence is independently processed by a pre-trained large language model to
537 generate an embedding matrix, which is then pooled into a feature vector. The resulting vectors
538 are concatenated and passed through a FC layer to generate the final prediction. Similar to the
539 single-sequence model, the language model parameters remain frozen during training. Only the
540 pooling layers (if VALP is used) and the FC layer are trainable.

541 The outputs of both modeling frameworks support four different types of prediction tasks:
542 binary classification, multi-class classification, multi-label classification, and regression.
543 Customizing the model for a specific task requires only adjusting the dimension and activation
544 function of the FC layer.

545

546 Identification of viral "dark matter"

547 ***Data set:*** The data set for predicting the identity of viral dark matter comprises two binary
548 classification tasks targeting the RNA-dependent RNA polymerase (RdRp) and viral capsid
549 proteins. For RdRp prediction, the data set includes 5,979 positive RdRp protein sequences and
550 229,434 negative non-RdRp protein sequences, sourced from our previous study (*9*). For capsid
551 protein prediction, positive samples comprise 216,742 protein sequences annotated with the

"Capsid" keyword (UniProt Keyword: KW2013) from UniProtKB. Negative samples include 83,461 “hard” negative sequences (e-value < 10^{-5} via DIAMOND alignment against UniProtKB and NCBI nr databases) and 133,281 “easy” negative sequences randomly sampled from non-capsid viral proteins on SwissProt. The data is split into 70% training and 30% testing sets.

Model architecture: The model employs a general single-sequence input framework. A protein sequence is processed by a pre-trained large language model (e.g., LucaVirus, ESM2-3B with embedding dimensions of 2,560, or ESMC-600M with 1,152) to generate an L×E feature matrix. This matrix undergoes value-level attention pooling (VALP) to produce an E-dimensional feature vector. The vector is passed through two fully connected (FC) layers: the first with GELU activation maps to a 128-dimensional latent space, and the second with sigmoid activation outputs a binary classification probability. Pre-trained model parameters remain frozen, with only the VALP layer and FC layers trainable.

Training procedure: The model was trained using the AdamW optimizer with a learning rate of 2×10^{-4} (including a warmup phase), a batch size of 16, and 50 epochs on NVIDIA A100 GPUs. Model performance is evaluated on the validation set after each epoch, selecting the model with the highest accuracy for inference. Evaluation metrics include accuracy, precision, recall, F1 score, AUC, and PR-AUC.

Real-world validation: To assess the model's ability to detect distant viral capsid proteins, sequences from LucaProt (*9*) that did not match known proteins in NCBI nr were screened. Viruses that were positively identified were clustered using DIAMOND blastp and mcl software, with representative sequences subjected to structure prediction via ESMFold (*10*) and ColabFold (*33*). FoldSeek (*34*) searches of the PDB, UniRef50, and SwissProt databases were used to confirm the true capsid proteins with e-value < 10^{-5}. Selected positive virus identifications were further validated through structure prediction with AlphaFold3 (*37*).

Characterization of unknown proteins

Data set: The data set was derived from UniProt, containing 10,715 viral proteins annotated

with EC (enzyme commission) annotations. EC numbers with fewer than 100 associated proteins were filtered out. The data set was split into 70% training and 30% testing sets.

Model architecture: A multi-label classification model was built using the single-sequence input framework. A protein sequence is processed by a pre-trained language model (e.g., LucaVirus with 2,560 embedding dimensions or ESMC with 1,152) to generate an L×E feature matrix. VALP pooling produces an E-dimensional vector, which is passed through two FC layers: the first with GELU activation maps to a 128-dimensional latent space, and the second with sigmoid activation outputs an N-dimensional probability vector. Only the VALP layer and FC layers are trainable, with pretrained model parameters frozen.

Training procedure: Training uses the AdamW optimizer with a learning rate of 2×10^{-4} (with warmup), a batch size of 16, and 50 epochs on NVIDIA A100 GPUs. The model with the highest validation accuracy is selected. Evaluation metrics include accuracy, precision, recall, F1 score, and Jaccard distance.

Real-world validation: The model's performance was validated by comparing predictions against known viral protein functions in UniProt, focusing on the ability to correctly assign multiple GO terms to proteins with complex functional profiles.

Forecasting virus evolvability

Data set: The data set is sourced from deep mutational scanning (DMS) data of the SARS-CoV-2 receptor-binding domain (RBD) (*36*). It includes 4,020 mutant sequences (201 amino acid positions × 20 substitutions), each with three technical replicate measurements of receptor (ACE2) binding affinity. The training set (34,055 sequences) comprises DMS data from early variants (Wuhan-Hu-1, Alpha, Beta), while the validation set (19,274 sequences) includes later variants (Iota, Delta), split by variant reporting time to simulate real-world generalization.

Model architecture: A regression model was built using the single-sequence input framework. The RBD sequence was processed by a pre-trained language model (e.g., LucaVirus with 2,560 dimensions or ESMC with 1,152) to generate an L×E feature matrix. VALP pooling yields an E-dimensional vector, passed through two FC layers: the first with

GELU activation maps to a 128-dimensional latent space, and the second with no activation outputs a scalar binding affinity value. Only the VALP layer and FC layers are trainable.

Training procedure: Training employed the AdamW optimizer with a learning rate of 2×10^{-4} (with warmup), a batch size of 16, and 50 epochs on NVIDIA A100 GPUs. The model with the lowest mean squared error (MSE) on the validation set is selected. Evaluation metrics include MSE, mean absolute error (MAE), and Pearson correlation coefficient.

Prediction of Ag-Ab binding affinity

Data set: Data is sourced from the CoVAbDab database (*46*), with positive samples restricted to human-derived antibodies binding to SARS-CoV-2 S protein, filtered for strong binding affinity. Negative samples were augmented with pre-pandemic human B-cell receptor (BCR) sequences from single B-cell sequencing, assumed to have negligible binding to SARS-CoV-2 due to lack of exposure. Negative samples were expanded to 10 times the positive samples to ensure balance. The data set was split into 70% training and 30% testing sets.

Model architecture: A binary classification model used the multi-sequence input framework. Antibody light chain, heavy chain, and SARS-CoV-2 antigen (S protein) sequences were individually processed by a pretrained language model (e.g., LucaVirus with 2,560 dimensions or ESM2-3B with 2,560) to generate L×E feature matrices. Each matrix undergoes VALP pooling to produce an E-dimensional vector. These vectors are passed through FC layers with GELU activation to map to 128-dimensional spaces, concatenated into a 128×3 vector, and fed into a final FC layer with sigmoid activation to output binding probability. Only the VALP layer, FC layers, and classification head are trainable.

Training procedure: Training used the AdamW optimizer with a learning rate of 2×10^{-4} (with warmup), a batch size of 16, and 50 epochs on NVIDIA A100 GPUs. The model with the highest validation accuracy is selected. Evaluation metrics include accuracy, precision, recall, F1 score, AUC, and PR-AUC.

Real-world validation: Single cell V(D)J sequencing data from SARS-CoV-2 convalescent and vaccinated volunteers were assembled using Cell Ranger (10× Genomics,

636 version 5.0.0) and reannotated with igblastn (version 1.18.0) as previously described (PMID:
637 35899581). V(D)J amino acid sequences containing only one heavy chain and one light chain
638 were selected to be imported into the model. The V(D)J fragments predicted to bind to the
639 SARS-CoV-2 Spike were synthesized and cloned into IgG1 expression plasmid. The IgG1
640 heavy- and light-chain plasmids were co-transfected into HEK293T cells, and the supernatant
641 was collected and purified with protein-A agarose beads to obtain monoclonal antibodies
642 (mAbs). ELISA (Enzyme-linked immunosorbent assay) was performed to verify the prediction.
643 Briefly, high protein binding microtiter plates (Costar) were coated with 2 μg/ml each of SARS-
644 CoV-2 WT Spike, WT RBD, and Omicron Spike recombinant proteins in PBS overnight at 4°C.
645 After blocking with 3% BSA, purified mAbs (10 μg/ml) serially diluted in a 1:3 ratio were
646 added to the plates and incubated for 1 hour at 37°C. Bound mAbs were detected with HRP
647 (horseradish peroxidase)-conjugated goat anti-human IgG antibody (Jackson, 1:2000).
648 Absorbance was measured at 405 nm using a microplate spectrophotometer (SpectraMax ABS
649 Plus). Antibodies with absorbance values exceeding twice that of the PBS blank control were
650 considered as antigen-binding. The area under the curve (AUC) was calculated and visualized
651 as a heatmap using GraphPad Prism 8.0 (GraphPad Software Inc.) to show binding strength of
652 positive mAbs.

Acknowledgements

Funding: This work is supported by National Natural Science Foundation of China (NSFC) Special Fund for Human Virome Project (Grant No. 82341118, awarded to M.S. and Z.-R.L.) and Basic Research Project for Doctoral Students (Grant No. 323B2018, awarded to Y.-F.P.), Shenzhen Science and Technology Program (Grant No. KQTD20200820145822023, awarded to M.S. and Y.-Q.C.), Major Project of Guangzhou National Laboratory (Grant No. GZNL2023A01001, awarded to M.S.), NHMRC (Australia) Investigator Award (Grant No. GNT2017197, awarded to E.C.H), and AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong S.A.R (Awarded to E.C.H).

Author contributions: Y.-F.P., Y.H., Z.-R.L., and M.S conceptualized the project. B.L., Y.-Q.C., Z.-R.L., and M.S. supervised the project. Y.-F.P. and Y.H. designed the model architecture. Y.-F.P. and Y.H. curated and processed the pretraining and downstream task datasets. Y.-F.P. and H.Y. developed the optimized training and inference code, with contributions from Y.-Q.L. to the training and inference code. Y.-F.P., Y.H., and Y.-Q.L. established and managed the hardware infrastructure and training environment. Y.-F.P. and Y.H. conducted pretraining, while Y.-F.P., Y.H., and Y.-Q.L. performed downstream task training. S.-N.L. carried out antibody validation under the supervision of Y.-Q.C. Y.-F.P. drafted the initial manuscript, with all authors contributing to revisions.

We thank J.-S. Eden and Ziwang for helpful discussions and assistance with manuscript preparation. The computations in this research were also supported by the CFFF platform of Fudan University.

Competing interests: All authors declare no competing interests.

Data availability: We confirm that the trained model weights, code, and datasets required to replicate our results will be publicly accessible upon publication.

Code availability: The source code for modeling and pretraining is available at https://github.com/LucaOne/LucaVirus. The code for downstream task modeling can be found at https://github.com/LucaOne/LucaVirusTasks.

References

1. E. Harvey, E. C. Holmes, Diversity and evolution of the animal virome. *Nature Reviews Microbiology*, (2022).
2. J. K. Jansson, R. Wu, Soil viral diversity, ecology and climate change. *Nature Reviews Microbiology*, (2022).
3. J. R. Brum *et al.*, Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
4. S. A. Babayan, R. J. Orton, D. G. Streicker, Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* **362**, 577-580 (2018).
5. N. N. Thadani *et al.*, Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818-825 (2023).
6. C. C. S. Tan, L. van Dorp, F. Balloux, The evolutionary drivers and correlates of viral host jumps. *Nature Ecology & Evolution* **8**, 960-971 (2024).
7. J. Nomburg *et al.*, Birth of protein folds and functions in the virome. *Nature* **633**, 710-717 (2024).
8. S. Duffy, L. A. Shackelton, E. C. Holmes, Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**, 267-276 (2008).
9. X. Hou *et al.*, Using artificial intelligence to document the hidden RNA virosphere. *Cell* **187**, 6929-6942.e6916 (2024).
10. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
11. T. Hayes *et al.*, Simulating 500 million years of evolution with a language model. *Science* **387**, 850-858 (2025).
12. G. Brixi *et al.*, Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025.2002.2018.638918 (2025).
13. E. Nguyen *et al.*, Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
14. N. Mollentze, D. G. Streicker, Predicting zoonotic potential of viruses: where are we? *Current Opinion in Virology* **61**, 101346 (2023).
15. N. Mollentze, S. A. Babayan, D. G. Streicker, Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLOS Biology* **19**, e3001390 (2021).
16. M. Holmudden, J. Gustafsson, Y. J. K. Bertrand, A. Schliep, P. Norberg, Evolution shapes and conserves genomic signatures in viruses. *Communications Biology* **7**, 1412 (2024).
17. R. Ketteler, On programmed ribosomal frameshifting: the alternative proteomes. *Frontiers in Genetics* **Volume 3 - 2012**, (2012).
18. S. A. Yost, J. Marcotrigiano, Viral precursor polyproteins: keys of regulation from replication to maturation. *Current Opinion in Virology* **3**, 137-142 (2013).
19. C. The UniProt, UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids*

Research **53**, D609-D617 (2025).
20. U. Neri *et al.*, Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037.e4018 (2022).
21. L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, T. D. Lawley, Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e1099 (2021).
22. Y. I. Wolf *et al.*, Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology* **5**, 1262-1270 (2020).
23. A. A. Zayed *et al.*, Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156-162 (2022).
24. A. C. Gregory *et al.*, Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114 (2019).
25. Y. He *et al.*, LucaOne: generalized biological foundation model with unified nucleic acid and protein language. *bioRxiv*, 2024.2005. 2010.592927 (2024).
26. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* **arXiv:1810.04805**, (2018).
27. W. Xiang, X. Han, X. Chai, J. Bai, BSM: Small but Powerful Biological Sequence Model for Genes and Proteins. *arXiv preprint arXiv:2410.11499*, (2024).
28. Y. Ren *et al.*, COMET: Benchmark for Comprehensive Biological Multi-omics Evaluation Tasks and Language Models. *arXiv preprint arXiv:2412.10347*, (2024).
29. Y. I. Wolf *et al.*, Origins and Evolution of the Global RNA Virome. *mBio* **9**, e02329-02318 (2018).
30. V. Koonin Eugene *et al.*, Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews* **84**, 10.1128/mmbr.00061-00019 (2020).
31. M. Krupovic, E. V. Koonin, Multiple origins of viral capsid proteins from cellular ancestors. *Proceedings of the National Academy of Sciences* **114**, E2401-E2410 (2017).
32. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366-368 (2021).
33. M. Mirdita *et al.*, ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679-682 (2022).
34. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243-246 (2024).
35. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nature Methods* **11**, 801-807 (2014).
36. T. N. Starr *et al.*, Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420-424 (2022).
37. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
38. Y. Myung, D. E. V. Pires, D. B. Ascher, CSM-AB: graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics* **38**, 1141-1143 (2022).

39. Y. Yuan, Q. Chen, J. Mao, G. Li, X. Pan, DG-Affinity: predicting antigen–antibody affinity with language models from sequences. *BMC Bioinformatics* **24**, 430 (2023).
40. R. T. McCoy, S. Yao, D. Friedman, M. D. Hardy, T. L. Griffiths, Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences* **121**, e2322420121 (2024).
41. D. Aldarondo *et al.*, A virtual rodent predicts the structure of neural activity across behaviours. *Nature* **632**, 594-602 (2024).
42. L. Richardson *et al.*, MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **51**, D753-D759 (2023).
43. S. Van Dongen, Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications* **30**, 121-141 (2008).
44. L. Trgovec-Greif *et al.*, VOGDB—Database of Virus Orthologous Groups. *Viruses* **16**, 1191 (2024).
45. Y. He *et al.*, KG-MTT-BERT: Knowledge graph enhanced BERT for multi-type medical text classification. *arXiv preprint* **arXiv:2210.03970**, (2022).
46. M. I. J. Raybould, A. Kovaltsuk, C. Marks, C. M. Deane, CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734-735 (2021).

Figures

Figure 1. A unified language model trained on extensive viral genomic and protein data sets. (**A**) Schematic of the LucaVirus model architecture and pre-training tasks. (**B**) Total tokens of non-redundant protein and nucleotide sequences in our pre-training data set compared with mainstream public databases – UniProtKB and NCBI Virus. (**C**) Taxonomic composition of viruses in the data set (measured by sequence count). (**D**) Taxonomic composition of putative hosts in the data set (measured by sequence count). (**E-H**) The four downstream tasks.

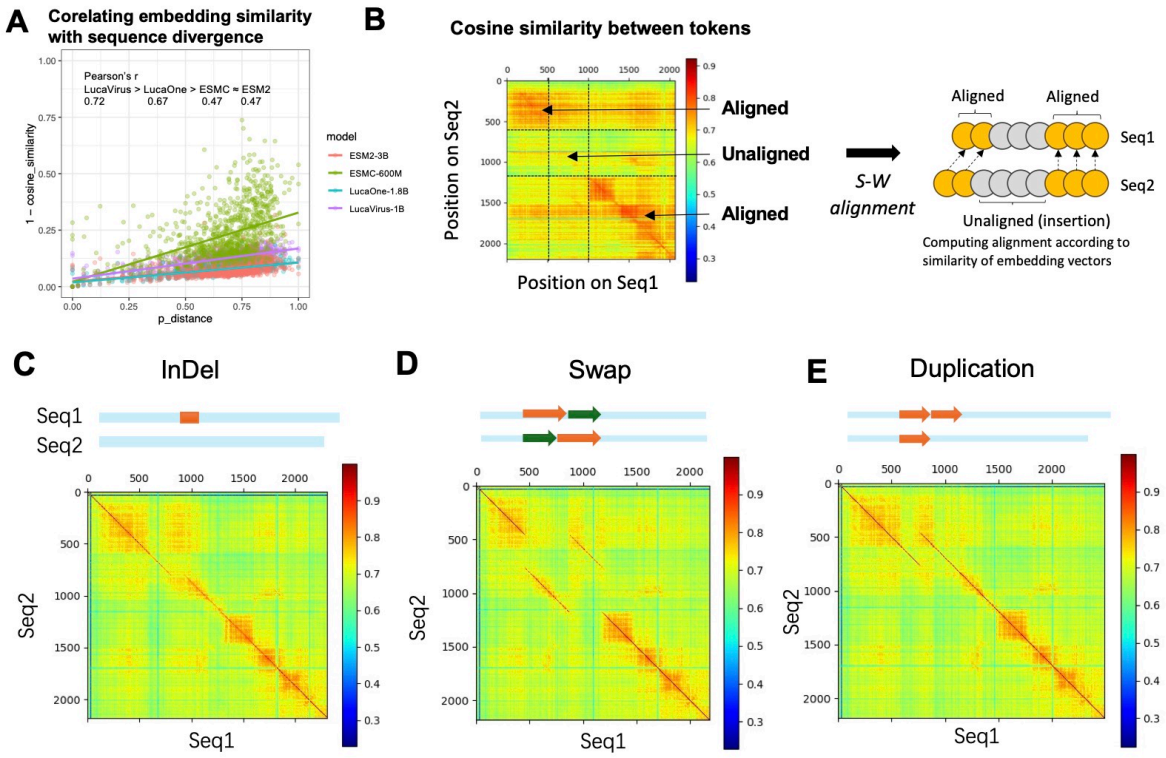

Figure 2. LucaVirus learns interpretable representation of viral sequences that reflect genetic divergence. (**A**) This panel shows the correlation between embedding dissimilarity and p-distance for pairs of unseen protein sequences. LucaVirus exhibits the highest correlation among four models (LucaVirus, LucaOne, ESM2-3B, ESMC-600M). (**B**) This heatmap illustrates the similarity between token pairs from two sequences, reflecting their homology relationship. (**C-E**) Similarity map between pairs of protein sequences can also reflect structural variations.

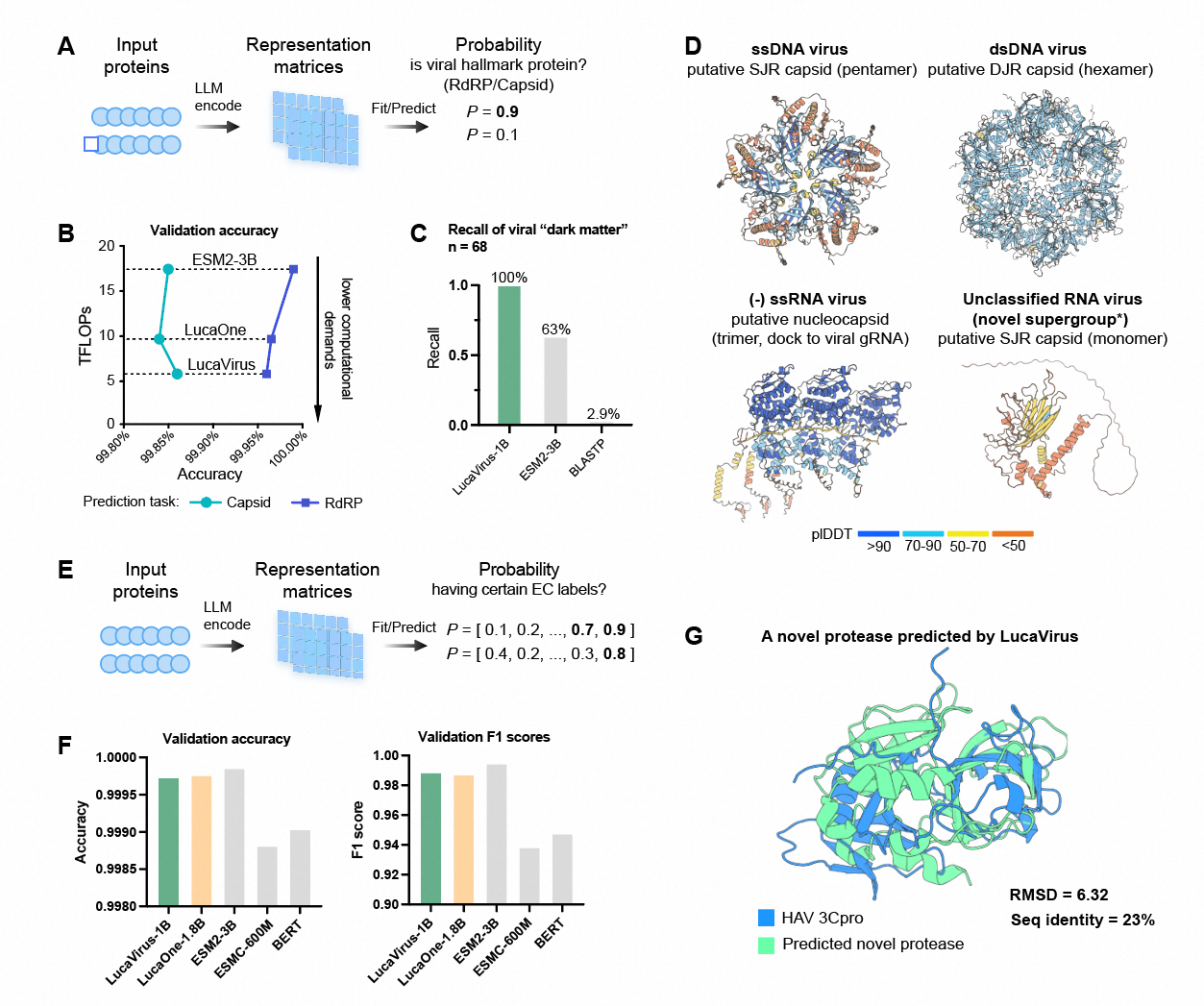

Figure 3. Exploring the hidden diversity and functional proteins of viruses. (**A**) Schematic illustration of predicting viral diversity using LucaVirus. (**B**) Comparison of the accuracy and computational complexity (floating point operations per inference, TFLOPs) among three large language models in RNA-dependent RNA polymerase (RdRp) and capsid protein prediction. (**C**) Performance comparison between LucaVirus and ESM2-3B in predicting distantly related viral capsid proteins. (**D**) Distantly related viral capsid proteins discovered by LucaVirus, including different capsid types (SJR, DJR, nucleocapsid). (**E**) Schematic illustration of the viral protein function prediction task. (**F**) Comparison of the model performance. (**G**) A novel viral protease identified by the LucaVirus model that exhibits no BLAST hits to known proteins. The closest known viral protease shares only 23% sequence identity and displays significant structural divergence.

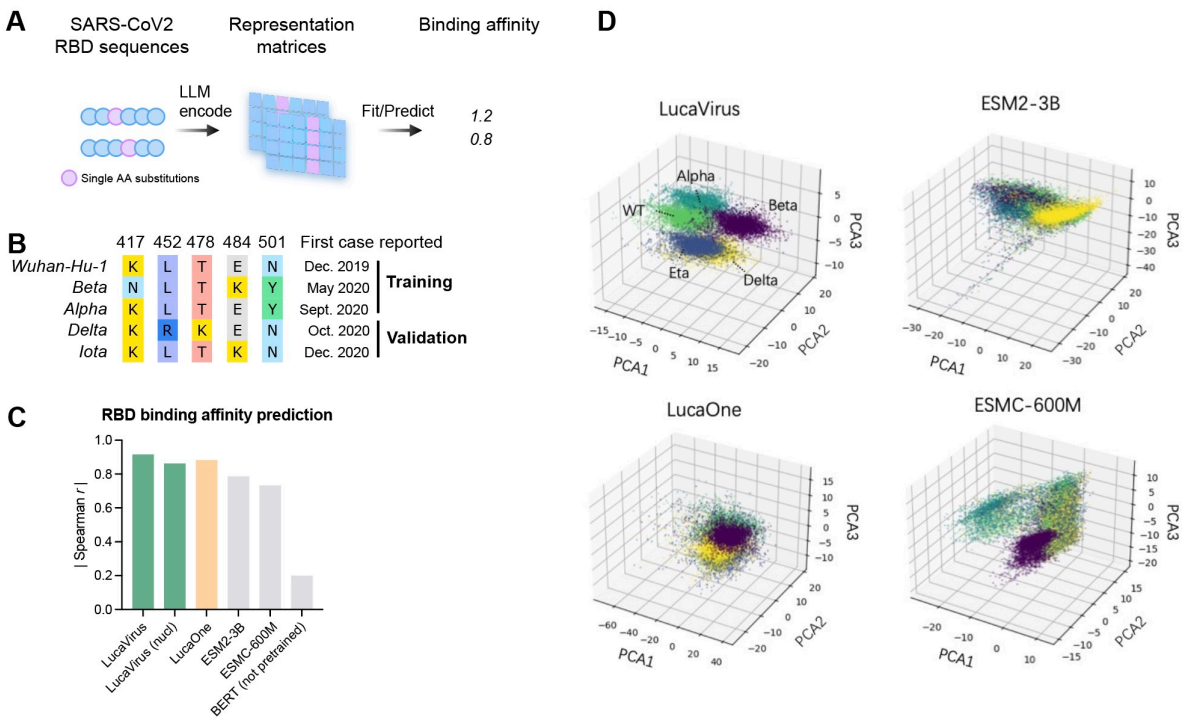

Figure 4. Fitting and predicting the fitness landscapes of a viral protein. (**A**) Schematic illustration of viral protein fitness (receptor affinity) prediction. (**B**) Amino acid mutations in the RBD region of five SARS-CoV-2 variants in the training data set. (**C**) Comparison of prediction accuracy between LucaVirus and other mainstream large language models. (**D**) PCA-based visualization of sequence embedding matrices from four different large language models, where each point represents the feature vector of a single amino acid residue generated by the models.

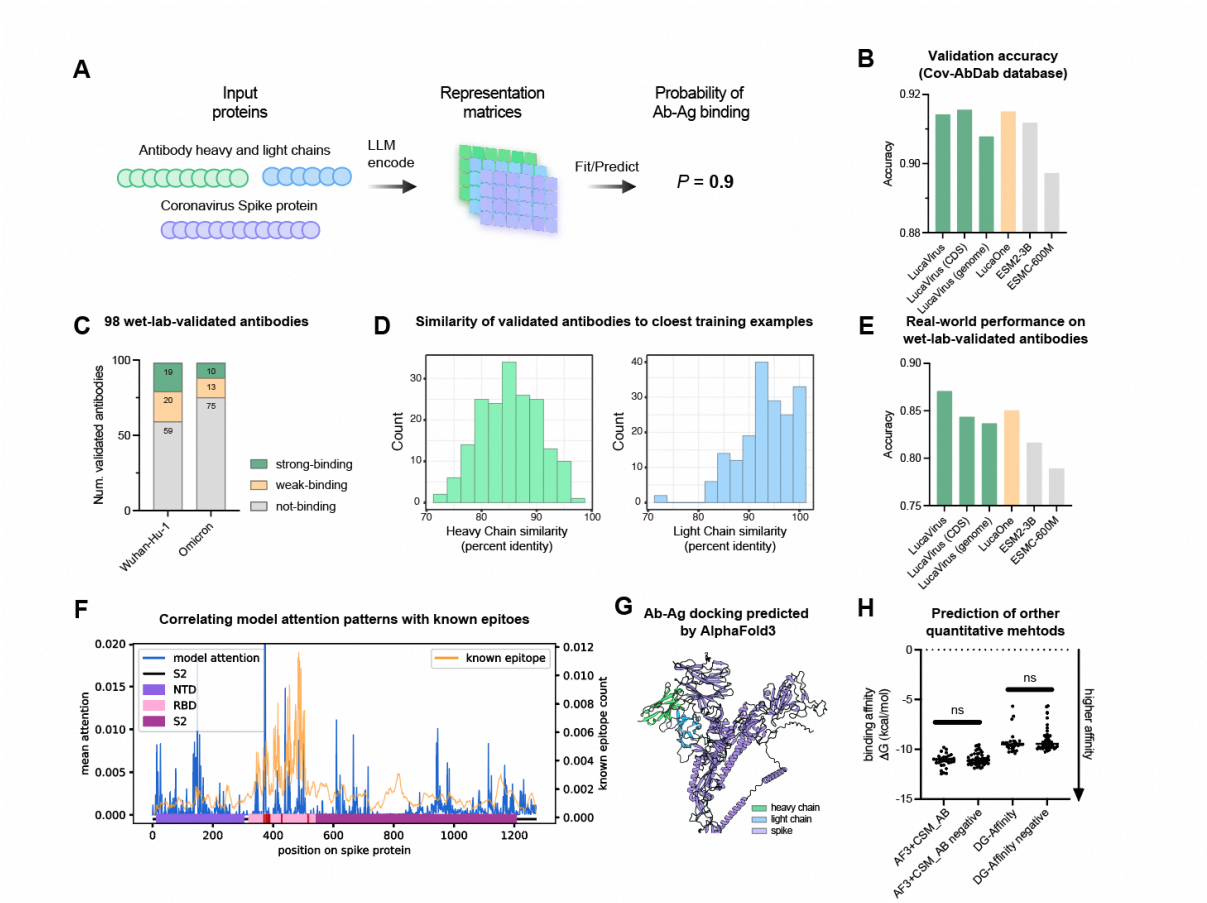

Figure 5. Performance of LucaVirus in antibody-antigen binding prediction. (**A**) Schematic diagram of the antibody-antigen binding prediction task. (**B**) Comparison of LucaVirus model's prediction accuracy under different input types. (**C**) Experimental validation using real-world antibody test data, illustrating the distribution of antibodies binding to the S protein of Wuhan-Hu-1 and Omicron strains in this data set (strong binding, weak binding, no binding). (**D**) Taxonomic composition of viruses in the data set (measured by sequence count). (**E**) Comparison of LucaVirus prediction accuracy on real-world data sets. (**F**) Distribution of attention weights in the LucaVirus model and known antigen epitopes. (**G**) Performance comparison of mainstream antibody prediction models (AlphaFold3+CSMab and DG-Affinity) on real-world data sets. (**H**) Comparison of LucaVirus prediction accuracy on real-world data sets.