Nucleic Acids Research, 2026, **54**, gkaf1448
https://doi.org/10.1093/nar/gkaf1448
Data Resources and Analyses

Protein structure-informed bacteriophage genome annotation with Phold

George Bouras [1,2,*], Susanna R. Grigson [3], Milot Mirdita[4], Michael Heinzinger[5,6], Bhavya Papudeshi [3], Vijini Mallawaarachchi[3], Renee Green[3], Rachel Seongeun Kim[4,7], Victor Mihalia[4,7], Alkis James Psaltis[1,2], Peter-John Wormald[1,2], Sarah Vreugde[1,2], Martin Steinegger [4,7,8,9], Robert A. Edwards [3,*]

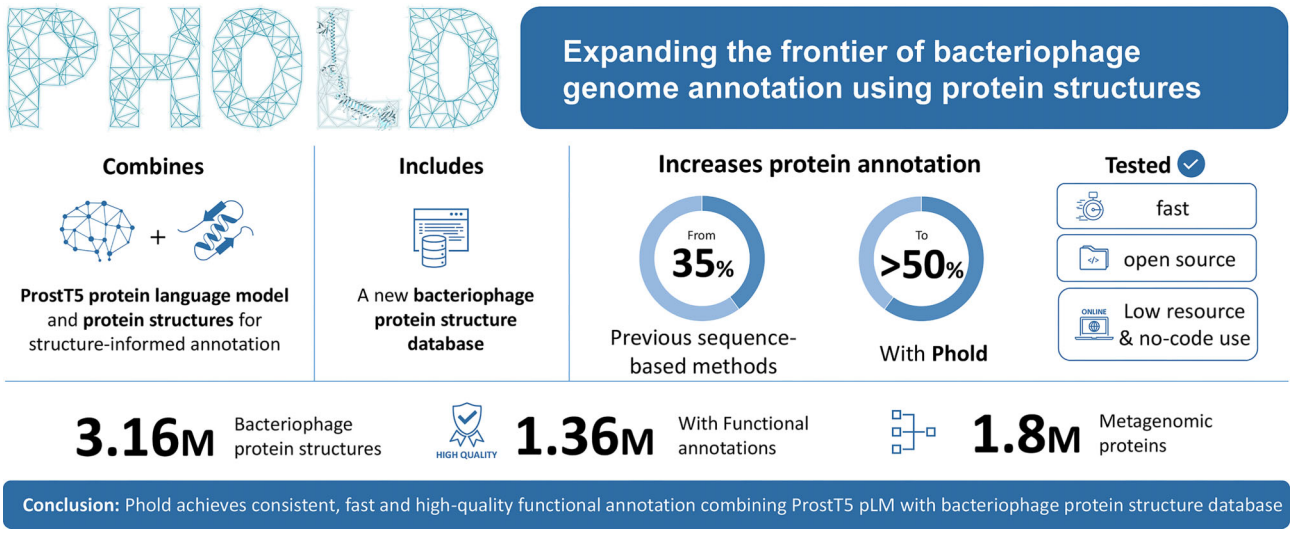
[1]Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, 5005, Australia
[2]The Department of Surgery—Otolaryngology Head and Neck Surgery, Central Adelaide Local Health Network, Adelaide, 5000, Australia
[3]College of Science and Engineering, Flinders University, Bedford Park, 5042, Australia
[4]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea
[5]Institute of Computational Biology, Helmholtz Center, Munich, 85764, Germany
[6]TUM School of Computation, Information and Technology, Technical University of Munich, Munich, 85748, Germany
[7]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[8]Artificial Intelligence Institute, Seoul National University, Seoul, Republic of Korea
[9]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Republic of Korea

*To whom correspondence should be addressed. Email: george.bouras@adelaide.edu.au
Correspondence may also be addressed to Robert A. Edwards. Email: robert.edwards@flinders.edu.au

Abstract

Bacteriophage (phage) genome annotation is essential for understanding their functional potential and suitability for use as therapeutic agents. Here, we introduce Phold, an annotation framework utilizing protein structural information that combines the ProstT5 protein language model and structural alignment tool Foldseek. Phold assigns annotations using a database of over 1.36 million predicted phage protein structures with high-quality functional labels. Benchmarking reveals that Phold outperforms existing sequence-based homology approaches in functional annotation sensitivity whilst maintaining speed, consistency, and scalability. Applying Phold to diverse cultured and metagenomic phage genomes shows it consistently annotates over 50% of genes on an average phage and 40% on an average archaeal virus. Comparisons of phage protein structures to other protein structures across the tree of life reveal that phage proteins commonly have structural homology to proteins shared across the tree of life, particularly those that have nucleic acid metabolism and enzymatic functions. Phold is available as free and open-source software at https://github.com/gbouras13/phold.

Graphical abstract



Received: September 17, 2025. **Revised:** November 12, 2025. **Accepted:** November 13, 2025

Introduction

Bacteriophages (phages), viruses that infect bacteria, are the most prevalent biological entities on Earth [1]. The persistent growth of metagenomic sequencing data [2] facilitates the discovery of numerous novel uncultured phages with little or no sequence similarity to known phages [3–7]. Further, increased interest in wet-lab-based studies of cultured phages, driven by the resurgence of phage therapy [8, 9]—i.e. using phages to target pathogens, particularly those that are anti-microbial resistant—adds to our arsenal of relatively well-characterized phages [10].

Despite the increasing number of available phage genomes, phage genome annotation remains extremely challenging. Traditionally, phage genome annotation is conducted using sequence-based homology approaches, consisting of alignment-based methods like MMseqs2 [11] and Diamond [12], or profile Hidden Markov Model (HMM)-based methods like HHblits [13] and HMMER [14, 15]. These algorithms are used by phage genome annotation pipelines [16–20] to search against general sequence databases like UniProt [21] or more targeted phage-specific databases, like PHROGs [22] or VOGDB [23], transferring functional labels to query proteins for strong, likely homologous hits.

However, the divergence and fast evolution of phage genomes limit the ability of sequence-based homology annotation to annotate most phage proteins; over 65% of phage proteins cannot be functionally annotated using sequence-based homology [8]. While protein structure is far more conserved than sequence [24], until recently it was not possible to utilize structural information for gene annotation due to the small number of solved structures. Driven by the breakthrough of AlphaFold2 [25] in predicting protein structure from sequence, combined with the development of Foldseek [26] as a rapid protein structural alignment tool fast enough to search through databases containing millions of protein structures to find homology in and below the 'twilight zone' of protein sequence alignment (especially below 25% sequence identity) [27], large-scale structural annotation of phages has recently become possible [28, 29].

This approach has two main bottlenecks. The first is that, despite the advent of tools like ColabFold [30] and ESMFold [31] that democratize structure prediction, predicting protein structures requires large computational resources even for a single phage. For example, running ColabFold batch with default AlphaFold2 model and parameters for all 66 coding sequences (CDSs) of *Escherichia* phage Lambda (GenBank accession: J02459) takes ∼100 min on a NVIDIA A100 GPU, not including multiple sequence alignment (MSA) generation. The second is that while there exist viral sequence databases that cover the general [29, 32] and specific eukaryotic [33, 34] viral spaces, there is no existing database of phage protein structures with high-quality curated functional labels.

Other approaches recently used for genome annotation beyond sequence-based homology are tools that leverage neural networks, particularly protein language models (pLMs). These tools include EAT [35] and Gaia [36] for general annotation, and Phynteny [37], PHANNs [38], VPF-PLM [39], and Empathi [40] for viruses. pLMs are mostly transformer-based language models trained on millions or billions of protein sequences [31, 41], whose resulting embeddings contain a high-dimensional understanding of a given protein, including implicitly elements of its structure [42]. However, despite some recent advances [43], a substantial drawback of using pLM-based methods for gene annotation is the lack of interpretability compared to traditional sequence-based methods, which output sequence alignments and statistical measures of confidence, like *E*-values.

Here, we present Phold, a tool for rapid, consistent, and accurate annotation of phages using protein structural information. We demonstrate that Phold's framework, which combines the ProstT5 pLM with Foldseek to search against our new, curated database of over 1.36 million mostly phage-derived protein structures, substantially outperforms sequence-based homology methods for phage genome annotation while maintaining similarly interpretable output. Applying Phold across the diversity of known cultured phages and uncultured metagenomic viral genomes revealed that over 50% of proteins on an average phage genome can be reliably annotated. We show that Phold using ProstT5 is almost as sensitive as using predicted protein structures but orders of magnitude more resource-efficient, allowing scalability to large metagenome datasets. Finally, by comparing Phold database and other cultivated phage protein structures to protein structures across the tree of life, we show many phage proteins have strong structural similarity to proteins shared across the tree of life, including in *Homo sapiens*.

Materials and methods

All benchmarking in this manuscript was conducted on a single high-performance computing (HPC) node with a single NVIDIA A100-40G GPU, along with 32-threads CPU and 128 GB RAM of 2× Intel Xeon Platinum 3460Y CPU system with 36 cores @ 2.4GHz. ProstT5 was run with GPU unless otherwise specified.

Phold database construction

PHROG

The core of the Phold database is the Prokaryotic Virus Remote Homologous Groups (PHROG) database [22]. All 868 340 PHROG proteins from the core 38 880 PHROGs containing at least two proteins were downloaded and deduplicated to retain 440 550 non-redundant proteins. All proteins over 3000 amino acids (AA) in length were then split into equal sizes below 3000 AA (e.g. a 5000 AA protein was split into two 2500 AA parts, an 8100 AA protein was split into three 2700 AA parts) due to this being approximately the memory limit of our available GPUs for structure prediction (NVIDIA A100 40 GB on the University of Adelaide Phoenix HPC and AMD MI250x on Setonix at the Pawsey Supercomputing Research Centre), yielding 441 177 unique protein sequences (including fragments) overall. Structure prediction was then conducted for all proteins using two methods. The first was using ColabFold v1.5.3 [30] implementing AlphaFold2 [25], conducted on Phoenix. Specifically, MSAs were created using both the uniref2302_30 and colabfold_envdb_202 108_db databases using MMseqs2 [11] v71dd32, and AlphaFold2 was run in batch mode with three models ('–num-models 3') and the default three recycles, without AMBER relaxation. The model with the highest predicted local distance difference test (pLDDT) was chosen as the best ColabFold model. Additionally, structures for all PHROG proteins that fit in memory (less than ∼900 AA) were then

predicted using ESMFold [31]. The structure prediction with the highest pLDDT between ESMFold and ColabFold models was then selected as the best model to be included in the Phold DB 3.16M.

EFAM

To diversify existing PHROG groups, we downloaded all 402 958 extremely conservative efam viral proteins taken from the Global Ocean Virome 2.0 dataset [44]. 392 139 proteins remained after filtering out all proteins with unclassified amino acids, of which 392 079 were lower than 3000 AA and kept for inclusion in the Phold DB 3.16M. Protein structure prediction was then conducted for these using ColabFold and ESMFold in an identical fashion as the PHROGs.

To assign efam proteins to a PHROG, the top-ranking structure for each efam protein was then compared against the PHROG structures in an all-versus-all search using Foldseek v 6cfb880 [26] using the parameters '-c 0.7 –num-iterations 3' to retain only hits with a bidirectional coverage of 70% (i.e. across the majority of both the query efam and target PHROG protein). All hits were then filtered using a minimum alignment TM-score value (computed by Foldseek) of 0.6, indicating that both the query and target are very likely to share the same fold [45]. For alignments under 75 AA, a Foldseek *E*-value cutoff of 0.01 was also implemented along with the TM-score filter to attempt to avoid spurious protein structure prediction alignments [46]. The top-hit PHROG was then assigned for each of the 233 181 efam proteins that passed these criteria, which were included in the Phold Search DB 1.36M. The remaining 158 898 proteins without PHROG hits were not included in the Phold Search DB 1.36M but remain available in the overall Phold DB 3.16M.

ENVHOG

To further diversify existing PHROG groups and provide a more general representation of the viral protein space, we downloaded MMseqs2 position-specific scoring matrix (PSSM) profiles for all of the Environmental Viral Homologous Groups (enVhogs) accessed on 31 January 2025 from http://envhog.u-ga.fr/envhog/ [47]. Consensus sequences for each profile were extracted for each of the 2 203 457 enVhogs using the 'mmseqs profile2consensus' command using MMSeqs2 v16-747e6 [11]. All enVhogs over 3000 AA were fragmented in the same fashion as PHROGs before structure prediction, yielding 2 205 969 sequences overall.

Protein structure predictions were then generated for each consensus enVhog using the same methodology as for PHROGs, using ColabFold and ESMFold, with two differences. The first difference is that Colabfold v1.5.5 was used with the AMD MI250x GPUs on Setonix in a Singularity container available from https://quay.io/repository/sarahbeecroft9/colabfold/rocm6.0.0_cpuTF, but otherwise identical parameters were used for protein structure generation. The second difference was that MSA generation was done with MMseqs2 v15.6f452 [11] with a third database added to enrich the MSA generation pipeline using beyond ColabFold's default uniref2302_30 and colabfold_envdb_202 108_db databases to generate deeper MSAs for these proteins. Implementation to search using this third 'colabfoldv' database is available at https://github.com/gbouras13/colabfoldv and contains 129 944 764 non-redundant viral (predominantly phage) proteins. Structure prediction was also conducted for these using ESMFold, with the top-ranking protein chosen in an identical fashion as for the PHROGs.

Assignment of enVhog proteins to PHROG annotations was done using two methods. First, the top-ranking predicted structure for each enVhog protein was then compared against the PHROG structures in an all-versus-all search using Foldseek v6cfb88 [26], using the parameters '-c 0.7 –num-iterations 3' to retain only hits with a bidirectional coverage of 70% (i.e. across the majority of both the query enVhog and target PHROG protein). All hits were then filtered with a minimum alignment TM-score value computed by Foldseek of 0.6, indicating that both the query and target are very likely to share the same fold [45]. For alignments under 75 AA, a Foldseek *E*-value cutoff of 0.01 was also implemented to avoid spurious protein structure prediction alignments, as for efam [44]. Second, a HMM–HMM comparison approach using HHblits (via hhsuite v3.3.0) [48] implemented using the same methods as in the enVhog manuscript was conducted using an *E*-value of 0.001. However, to increase the confidence in the functional PHROG assignment by considering coverage across the entire protein, we added a filter to HMM–HMM hits, ensuring at least 70% coverage of both the query HMM and target HMM consensus sequences.

In the case that an enVhog had both Foldseek and HMM–HMM hits to different PHROGs, the hit with lowest *E*-value was taken as the best hit for that enVhog. Overall, 562 369 enVhog consensus proteins with either Foldseek or HMM–HMM hits to PHROG groups with known function were assigned. Of these, 383 648 were unique to Foldseek, 48 570 unique to HMM–HMM comparions, and 130 151 had hits with both methods. Of these 130 151 with hits to both methods, 94 526 had hits to identical PHROGs, 9951 had hits to different PHROGs with identical annotations, 24 035 had hits to different PHROGs with the same PHROG category but different specific annotations, 8527 had hits to different PHROG categories where one hit was to an unknown function PHROG category (in which case the known-function hit was taken), while 3063 had conflicting hits to different PHROG categories where both had known annotations. The remaining 1 638 667 enVhog proteins without PHROG hits were not included in the Phold Search DB 1.36M but remain available in the overall Phold DB 3.16M.

PHROG singletons

To improve the functional understanding of singleton PHROGs, we took all PHROG singleton proteins (i.e. from PHROG 38 881 to 109 404) and initially generated predicted protein structures using ColabFold and ESMFold in the same way as described for PHROGs above. Due to the likelihood for these singletons to have sparser MSAs and lower predicted pLDDT, we also generated more predicted structures using enriched MSAs to increase the chances of generating a high confidence structure. Specifically, we generated MSAs using both the standard ColabFold databases (i.e. uniref2302_30 and colabfold_envdb_202 108_db) and the standard ColabFold databases augmented with the 'colabfoldv' database in the same way as for the enVhogs using MMseqs2 v15.6f452 [11]. The protein structure with the highest pLDDT between the six ColabFold models (three with default MSA, three with enriched MSA) and the ESMFold model was chosen as the best structure.

Anti-CRISPRdb

All 3693 AlphaFold2 predictions from Sahakyan *et al.* [49] of anti-CRISPRdb proteins were downloaded from https://zenodo.org/records/7747008. A total of 3652 predicted anti-CRISPR structures with average pLDDT over 70 (from the metadata as provided by Sahakyan *et al.*) were kept for inclusion in Phold. Predicted structures for these were also generated with ESMfold. The protein structure with the highest average pLDDT was then included in Phold's databases.

CARD

The CARD Data v3.2.8 release (2 October 2023) was downloaded from https://card.mcmaster.ca/download [50]. A total of 4804 protein homolog models were considered for inclusion in Phold. 2893 of these proteins had structure predictions already in the AlphaFold database v4 [51], which were downloaded. Structure predictions for the remaining 1911 proteins were generated on Phoenix in the same way as the PHROGs. Predicted structures for these were also generated with ESMFold. The protein structure with the highest average pLDDT was then included in Phold's databases.

DefenseFinder

All 461 monomer DefenseFinder [52] protein structures generated using AlphaFold2, available as of 31 January 2024, were downloaded from https://defensefinder.mdmlab.fr/. Four hundred eight of these with average pLDDT above 70 were kept. Predicted structures for these were also generated with ESMfold. The protein structure with the highest average pLDDT was then included in Phold's databases.

Diversity-generating retroelement reverse transcriptases

To improve the diversity of diversity-generating retroelements (DGRs) in Phold's databases, we took 12 760 putative DGR reverse transcriptase sequences from Roux *et al.* [53], representing all OTUs (≥95% AAI) that corresponded to DGR references. We generated predicted protein structures for all 12 760 in the same way as for PHROGs on Phoenix. We then ran Foldseek easy-search v9.427df8a [26] to compare the best predicted structure based on pLDDT against all PHROG 1423 (annotated as reverse transcriptase) using ColabFold predicted protein structures with an *E*-value threshold of 1E-05. 12 683 out of 12 670 putative DGRs with hits were then kept for inclusion in Phold. Predicted structures for these were also generated with ESMFold. Finally, the protein structure with the highest average pLDDT was then included in Phold's database.

NetFlax

We used the NetFlax [54] toxin–antitoxin proteins to improve the diversity of toxin–antitoxin system proteins in the Phold database. We took all 7152 toxin and antitoxin protein structures predicted with AlphaFold2 from NetFlax, as provided by the authors. Predicted structures for these were also generated with ESMFold. Finally, the protein structure with the highest average pLDDT was then included in Phold's databases.

Virulence Factor Database

We generated protein structure predictions for the Virulence Factor Database (VFDB) [55] (downloaded on 19 January 2024 from http://www.mgc.ac.cn/VFs/download.htm) for 27 823 VFDB virulence factor proteins using the same methods as PHROGs. The protein structure with the highest average pLDDT between ColabFold and ESMFold methods was then included in Phold's databases.

Phold benchmarking datasets

All per genome and per protein structure information for the benchmarking datasets can be found in Supplementary Tables S14, S20, and S21.

We used five different datasets for benchmarking Phold:

1. 182 INPHARED [10] viral genomes with a total of 16 460 CDS ('INPHARD 182'). Specifically, these were comprised of one representative of all unique genera added to INPHARED after the 1 October 2021 release, as of the 1 November 2023 release. We chose 1 October 2021 as a timepoint cutoff, as theoretically phages added after 1 October 2021 were not part of the original PHROGs database, though we note that we do not guarantee this, due to the frequent taxonomic renaming of phages and constant updating of INPHARED. We used this dataset to broadly capture the known diversity of reasonably well-characterized, mostly isolated phages and archaeal viruses in place of the 1419 INPHARED dataset for all Phold performance ablations, due to its smaller size.
2. 1419 INPHARED viral genomes comprising one representative for every unique viral genus in INPHARED as of the 14 April 2025 release, containing 148 196 CDS ('INPHARED 1419').
3. 249 *Crassvirales* phage genomes from Edwards *et al.* [6] with a total of 22 127 CDS ('Crass').
4. 63 metagenomic assembled phage genomes from Cook *et al.* [3] with 3 699 CDS ('Cook').
5. 45 metagenomic assembled phage genomes from Tara Oceans [56] with 2 112 CDS ('Tara').

For each benchmarking dataset, all phage genomes were first run with Pharokka v1.7.4 or v1.7.5 [16] using pyrodigal-gv [18, 57, 58] as the gene caller with '-m' (meta mode) with and without '–meta_hmm' that implements PyHMMER [15] based annotation in addition to MMSeqs2 profile-based annotation. Protein structure predictions for all predicted CDSs were folded using ColabFold v1.5.3 and ESMFold using the same parameters as the PHROGs for database construction, though in this case, the top-ranking pLDDT model was not chosen, as ESMFold and ColabFold were considered differently in benchmarking (i.e. the ColabFold model was used as the pseudo ground truth). A small number of large CDS over 3000 AA in each dataset due to computational limitations, and these were ignored in benchmarking. Phold parameters '–structures –structure_dir' with ColabFold-generated structures, along with the default '–max_seqs 1000 –masking_threshold 25' were used with the Phold Search DB 1.36M and an *E*-value threshold of 0.01 were used to generate the baseline pseudo 'ground-truth' for all sensitivity analyses. For the purposes of benchmarking, proteins with ground truth non-hypothetical functional annotations were considered as true positives. Of the metrics presented, the binary task indicates whether the 3Di inference method could generate any non-hypothetical annotation for the proteins, while multiclass metrics indicate that the

method generated the correct specific annotation (i.e. at either PHROG group, product, or category level). We also annotated datasets 2–5 (INPHARED 1419, Cook, Crass and Tara) with Phynteny [37] v0.1.3 using the Phold with ColabFold structures annotations as the input for Phynteny. We used the default minimum Phynteny confidence threshold of 0.8 to assign annotations to hypothetical proteins.

More information on each of the metagenomic phage datasets are as follows:

Crassvirales phages ('Crass')

Two hundred forty-nine *Crassvirales* genomes were taken from Edwards *et al.* [6]. Protein structure predictionx were generated for all proteins 2000 amino acids and below, with a total of 21 740 protein structure predictions were generated.

Cook *et al.* phages ('Cook')

All putative phage genomes assembled with Phables v0.2.0 [59] and MEGAHIT v1.2.9 [60] in the file BINS_sample_01_phables.fa.gz from Cook *et al.* were downloaded. CheckV v1.0.3 [61] was run to assess the completeness of these genomes. Sixty-four genomes were assessed to be complete. One genome (phage_comp_52_cycle_2) was excluded as being very similar to another genome (phage_comp_52_cycle_1) based on the Phables assembly method from the manuscript. Of 63 remaining genomes, predicted structures were generated for 3695 out of 3699 proteins.

Tara Oceans phages ('Tara')

Reads from Sequence Read Archive accessions ERR2750826, ERR2750828, and ERR2750829 were downloaded and assembled with Phables[84] v1.3.0, and metaSPAdes v3.15.5 [62] yielding 45 phage genomes. Predicted structures were generated for 2110 out of 2112 predicted phage genomes.

Phold search parameter optimization and comparison with Foldseek-GPU

To understand the effect of differing Foldseek [26] search parameters on search sensitivity and resource consumption, we ablated the '–max-seqs' parameter (controlling search sensitivity), choosing 1000 (default in Foldseek), 5000, 10 000, 20 000, 50 000, and 250 000. We chose to do this using both Foldseek CPU and GPU, as it is likely that the many Phold users would use either implementation depending on their hardware. We also clustered the Phold search database at different values of '–seq-id' (0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) and searched the clustered database using the '–cluster-search 1' parameters and compared it to a full 'Foldseek search' of the entire Phold search database. We performed the same ablations against both the Phold Search DB 1.36M (i.e excluding enVhogs and efam proteins with no PHROG hits) consisting of 1 363 704 proteins and the full Phold DB 3.16M (i.e. including efam and enVhog proteins with no PHROG hits) consisting of 3 166 202 proteins. The query dataset used for this was INPHARED 182. Phold with ProstT5 with the commit '59d356e' (with customized clustered databases as described earlier) was used. For Foldseek GPU searches, we specified the Phold parameter '–foldseek_gpu', which passes '–gpu 1 –prefilter-mode 1' to Foldseek. The metrics compared were total wall-clock time taken, total CDS with functional hits found (i.e. hits to proteins with a function that was not unknown function) and total CDS with all hits (i.e. including hits to proteins with unknown function). Foldseek v10.941cd33 with an *E*-value of 0.001 was used for Phold runs.

Phold ProstT5 confidence masking benchmarking

We define the per-residue ProstT5 confidence as a score out of 100. It is calculated as 100 multiplied by the softmax of the logits resulting from the ProstT5-CNN head for the 3Di token with the highest value that is chosen at that residue. The per-protein ProstT5 confidence metric is accordingly defined as the mean of each per-reside confidence score across the protein. It reflects a measure of the model's confidence in the 3Di prediction at each residue. In order to understand the effect of masking low confidence ProstT5 predicted residues on Phold's search sensitivity, we conducted an ablation study with the '–mask_threshold' parameter. The '–mask_threshold' masks 3Di tokens in the Foldseek search for residues where the ProstT5 confidence is below this threshold. We tested values of 0 (i.e. no masking), 10, 20, 30, 40, 50, 60, 70, and 80, and compared the number of CDS annotated compared to running Phold with the ColabFold generated structures against the search database, using an *E*-value of 0.001 and max-seqs 1000.

ProstT5 confidence analysis

For the 3 166 202 protein structure predictions in Phold DB 3.16M, we ran 'phold predict' with ProstT5-CNN to predict 3Di tokens. For each protein, the mean ProstT5 3Di confidence, calculated by Phold as the mean of all per-residue ProstT5 confidence scores across the protein, was taken. The actual 3Di similarity was calculated by computing the edit distance between ProstT5-generated 3Di and the 3Di tokens generated by the Foldseek VQ-VAE from the 3 166 202 protein structure predictions in Phold DB 3.16M. Pearson's r and Spearman ρ were then calculated for between the ProstT5 confidence and actual similarities.

ProstT5 LoRA finetuning and CNN training

To finetune ProstT5, 2 205 504 predicted protein structures with a mean pLDDT of at least 70 were taken from Phold DB 3.16M protein structures. They were clustered with Foldseek v10.941cd33 using the 'Foldseek cluster' command using the default bi-directional coverage of 80%, yielding 149 321 non-singleton and 625 617 singleton clusters. To avoid oversampling large clusters in the training dataset, only the 20 most diverse sequences for non-singleton clusters with at least 20 members were chosen. For training, proteins larger than 512 amino acids were truncated to fit in GPU VRAM, keeping only the first 512 residues. We held out 100 random non-singleton clusters with 719 proteins and 625 random singletons as the validation set, with the remaining 1 323 937 proteins comprising the training set. We also used 500 arbitrary Swiss-Prot proteins as a non-phage validation set. We applied low order rank adaptation (LoRA) fine-tuning [63] on the full ProstT5 encoder-decoder with weights accessed via HuggingFace. Specifically, our configuration was the same as suggested by Schmirler *et al.* [64]: rank 4, alpha 1, applied to query, key, value, and output of the attention layers. Overall, out of 2 818 852 864 (2.8B) ProstT5 encoder-decoder parameters, 5 900 288 (5.9M) were available for fine-tuning. Our fine-tuned model was trained with Pytorch v2.5.1 using the

default HuggingFace trainer on a single GPU node containing 4× NVIDIA A100 40GB GPUs. We trained our model for four epochs with an effective batch size of 32 (per GPU minibatch size of 1, gradient accumulation steps of 8), using mixed precision, a learning rate of 3e-4 and no warmup. The fine-tuned model is freely available at https://huggingface.co/gbouras13/ProstT5Phold.

We then trained two-layer convolutional neural networks (CNNs) on top of the LoRA-finetuned ProstT5 encoder embeddings using biotrainer [41]. We used identical hyperparameters to the CNN trained and provided with the original ProstT5 (two layers, bottleneck dimension 32, no drop out, batch size 8, 5 epochs, and a learning rate of 1e-3) and trained the CNNs on a single NVIDIA RTX4090 GPU. We trained two CNNs: one using the "CASP14" dataset from ProstT5, containing proteins from the Protein Data Bank (PDB) along with some CASP14 proteins named 'vanilla CNN' in this manuscript; and one using 50 000 randomly selected Phold DB 3.16M proteins taken from the 1 323 937 proteins in the LoRA finetuning training dataset named 'finetune CNN' in this manuscript. These CNNs are available in the Phold GitHub repository. We ultimately chose to keep the original ProstT5 with CNN for use with Phold (see Supplementary Note 2).

Phold DB database curation and unknown function PHROG annotation propagation

Original PHROGs with two or more members (1–38 880)

For all PHROG proteins that had at least two members in the original PHROGs v4 (i.e. from PHROG 1 to 38 880) which were annotated as 'unknown function', a semi-manual curation approach was as follows. Overall, 668 199 proteins (i.e. original PHROGs expanded with efam and enVhog proteins) belonged to the 33 553 PHROGs that were 'unknown function' in v4, while 651 666 proteins belonged to the 5327 PHROGs that had an annotated function.

We ran Foldseek (CPU) easy-search with the parameters '-c 0.7 –max-seqs 1000 –num-iterations 3' against the following three databases:

1. All BFVD structures.
2. All AFDB structures (searched using the Foldseek AFDB50 database with –cluster-search 1)
3. All 651 666 proteins belong to the 5 327 known function PHROGs.

We then post-filtered all the Foldseek results to keep all hits that had *E*-value <0.01 and a TM-score of at least 0.6, indicating that the query and target likely shared the same fold [45]. Annotation transfer for PHROGs that were annotated as 'unknown function' in v4 was then conducted manually using the following heuristics:

- For each PHROG group, number and strength [in terms of *E*-value TM-score and local distance difference test (LDDT)] of the top hit target against (i) the BFVD and (ii) the known PHROGs were considered as the best source of information, as these databases most tailored for viral and phage proteins. For any annotation transfer to take place, at a minimum, at least 10% of query proteins in the PHROG must have had an identical or very similar annotation to considered (though generally a higher amount was needed to transfer annotation).
- Generally, vague annotations (e.g. those with 'DUF', 'Uncharacterized', 'Hypothetical', 'Putative') and hits near *E*-value (0.01), TM-score (0.6), and percentage (10%) thresholds were not propagated. Vague annotations (e.g. uncharacterized) were generally not counted as informative (i.e. not penalized as divergent) if the rest of the annotations were informative and consistent.
- PHROGs with queries that had multiple divergent annotations were not propagated.
- Other PHROG-level information computed in the original PHROG manuscript and available for download and on the webserver was also considered as informative. These particularly include the annotation of similar PHROGs computed using HMM–HMM searches and similar PFAM, KEGG, and GO hits.
- If a PHROG had only 1 or 2 members, then a strict 100% criterion was enforced, and if three members, a 2/3 criterion was applied.
- Putative integrases were specifically scrutinized in more detail, including visualizing and comparing hits manually, given the therapeutic interest in identifying them.
- BFVD and similar PHROG information were supplemented by AFDB top-hit counts and information. This was particularly used for PHROG queries that had no hits to BFVD proteins and for PHROGs likely to belong to non-structural categories (i.e. other, moron, auxiliary metabolic gene and host takeover, DNA, RNA, and nucleotide metabolism, and transposases).

Original singleton PHROGs (38 881–109 399)

For PHROG singleton proteins (i.e. PHROG 38 881–109 399), the semi-manual functional curation approach followed the same Foldseek easy-search parameters as for PHROGs 1–38 880. The heuristics followed for annotation transfer for these PHROGs were as follows:

- Consider the number and strength of hits of the annotation against the PHROGs 1–38 880 as the post important datapoint. Singletons with only few hits and hits near the *E*-value and TM-score thresholds did not have their annotation propagated.
- Annotation transfer was also not conducted if there were many hits for the same singleton with divergent annotations.
- Also, consider the AFDB and BFVD top-hit annotation for each singleton, if they exist. If they are strongly divergent, then do not proceed with annotation transfer.
- Manually view the predicted structure to see if it visually concords with the predicted function.

Low, medium, and high phold annotation confidence heuristics

To make the output of Phold more interpretable for users, particularly those without understanding of protein structural alignment methods, we assign each CDS annotation as either 'high', 'medium', or 'low' confidence based on the following heuristics. High-confidence hits are where both the query and target proteins have at least 80% reciprocal alignment coverage, along with either (i) >30% amino acid sequence identity (suggesting the hit is in the light zone of sequence homology[27]), or (ii) query mean ProstT5 confidence of at least 60% (suggesting a very good-quality ProstT5 3Di prediction), or (iii) an alignment *E*-value < 1e-10. Medium-confidence hits

are where either the query or target protein hit has at least 80% coverage, along with either (i) >30% amino acid sequence identity, or (ii) ProstT5 confidence between 45% and 60% (suggesting a good-quality ProstT5 3Di prediction), and (iii) an *E*-value < 1-e05. Low-confidence hits are all other hits below the specified *E*-value that do not fit those thresholds (i.e. hits with low coverage, low amino acid sequence identity, low ProstT5 confidence, or hits near the *E*-value threshold of 0.001). If Phold is run with user-provided input structures instead of ProstT5, the heuristics are identical other than the ProstT5 confidence criteria. In this case, Phold will also output alignment template modeling score (TM-score) [65] and LDDT [66] values from Foldseek, which may also guide the user in assessing annotation quality.

NEFF and MSA analysis

The number of homologs was calculated as the number of MMSeqs2 hits found by ColabFold against both the Uniref50 and environmental databases. The NEFF for each MSA was calculated using NEFFy [67].

Phold DB and INPHARED tree of life homology analyses

To compare protein structures across the tree of life, we took all Phold DB 3.16M and 147 946 INPHARED protein structures and ran Foldseek (CPU) easy-search with the parameters '-c 0.7 –max-seqs 1000 –num-iterations 3' against the following two databases:

1. AlphaFold Database [51] with over 214 million proteins (searched using the Foldseek AFDB50 database with –cluster-search 1).
2. AlphaFold Proteome database consisting of 48 high-quality reference proteomes (16 from *Bacteria*, 31 from *Eukayota*, and a single archaeon, *Methanocaldococcus jannaschii*) with 564 446 proteins [68].

We then post-filtered all the Foldseek results to keep all hits that had *E*-value <0.01 and a TM-score of at least 0.6, indicating that the query and target likely shared the same fold [70]. Taxonomic labels for all hits were taken from the Foldseek output. *Homo sapiens* hits were post-filtered and taken from the Proteome searches, as the smaller target database size led to substantially more hits to the specific taxon (i.e. 17 041 Phold DB proteins had *Homo Sapiens* hits when post-filtering AFDB results compared to 54 406 when post-filtering the Proteome search). We also kept the Foldseek-computed LDDT score for every alignment.

Results

A comprehensive database of high-quality and well-annotated predicted bacteriophage protein structures

As a prerequisite for structure-based phage annotation, we first curated and constructed a database of well-annotated phage protein structures. Overall, we predicted structures for 3 166 202 proteins from various sources (hereafter 'Phold DB 3.16M'). For the final Phold search database (hereafter 'Phold Search DB 1.36M'), we retained a subset of 1 363 704 proteins with high-quality functional annotations (Supplementary Table S1). These include: 441 177 deduplicated PHROG [22] proteins from each PHROG cluster containing at least two proteins (i.e. PHROG clusters 1–38 880); 70 455 singleton PHROG proteins (i.e. PHROG clusters 38881–109 404); 2 205 969 enVhog proteins [47] (consisting of one representative per enVhog group), of which 562 369 were assigned a PHROG group and kept in Phold Search DB; 392 080 efam [44] proteins, of which 233 181 were assigned a PHROG group and kept in Phold Search DB; 12 683 DGR reverse transcriptases [53] assigned to PHROG 1423; 3652 anti-CRISPRs (ACRs); 4 804 CARD [50] antimicrobial resistance proteins; 408 DefenseFinder [52] proteins; 27 823 VFDB [55] proteins; and 7 152 NetFlax [54] toxin–antitoxin system proteins. All specialized databases (ACRs, CARD, DefenseFinder, VFDB, NetFlax) were assigned the PHROG category moron, auxiliary metabolic gene, and host takeover. All proteins in the Phold Search DB are assigned 1 of 10 general PHROG category labels, along with a more detailed PHROG product label.

Overall, the predicted protein structures were generally of good quality. The mean pLDDT for all proteins in Phold DB 3.16M was 75.69, with mean length of 211 AA, while the mean pLDDT of the Phold Search DB 1.36M was 82.09, with mean length of 219 AA. The remaining 1.8M structures, excluded from the Phold Search DB due to having no assignable PHROG, had mean pLDDT of 70.85 and mean length of 207 AA. 86.9% of structures included in the Phold Search DB 1.36M had mean pLDDT of at least 70 (Fig. 1A). PHROG singleton and enVhog proteins were on average shorter and had lower pLDDT than the other constituents of the Phold Databases (Supplementary Fig. S1B–H).

We predicted structures for all structures using both ColabFold and ESMFold, selecting the structure with the highest pLDDT score (Supplementary Note 1). ColabFold outperformed ESMFold for 86.5% of proteins, with an overall higher mean pLDDT (74.8 versus 61.2). We observed that while structure quality was correlated with MSA depth for both methods (Supplementary Fig. S2A), for ESMFold proteins with fewer than 100 homologs in the ColabFold MSA, there was no correlation between ESMFold pLDDT and number of homologs with generally low-quality predictions (Spearman's ρ: -0.01, median pLDDT: 54.6) (Supplementary Fig. S2B). For those with at least 100 homologs, there was a positive correlation and more confidently predicted structures (Spearman's ρ: 0.30, median pLDDT: 67.6) (Fig. 1B and Supplementary Fig. S2C).

For PHROG singletons and enVhogs, we enriched the MSAs by extending ColabFold's default search databases to include our collection of 129 944 764 viral proteins. For PHROG singletons, we produced and compared ColabFold predictions with and without enriched MSAs. Enrichment increased the number of homologs from a median of 19–53, improving pLDDT for 41 802/70 524 (59.3%) singletons. The median ColabFold pLDDT increased from 74.4 (base MSA) to 76.4 (enriched), yielding an overall median ColabFold pLDDT of 77.3 when we took the highest pLDDT from the two methods. The number of high-quality (pLDDT > 70) predictions increased by 4 406/70 524 (6.2%) (Supplementary Fig. S3A). Proteins with base MSAs that were neither completely empty nor very deep and high quality had the largest improvement in pLDDT with the additional sequences (base MSA homologs from 11–20 improved by median 1.22 pLDDT; Supplementary Fig. S3B and C).

Overall, 75.0% of PHROG groups containing 89.4% of all proteins had mean pLDDT of at least 70. The mean pLDDT

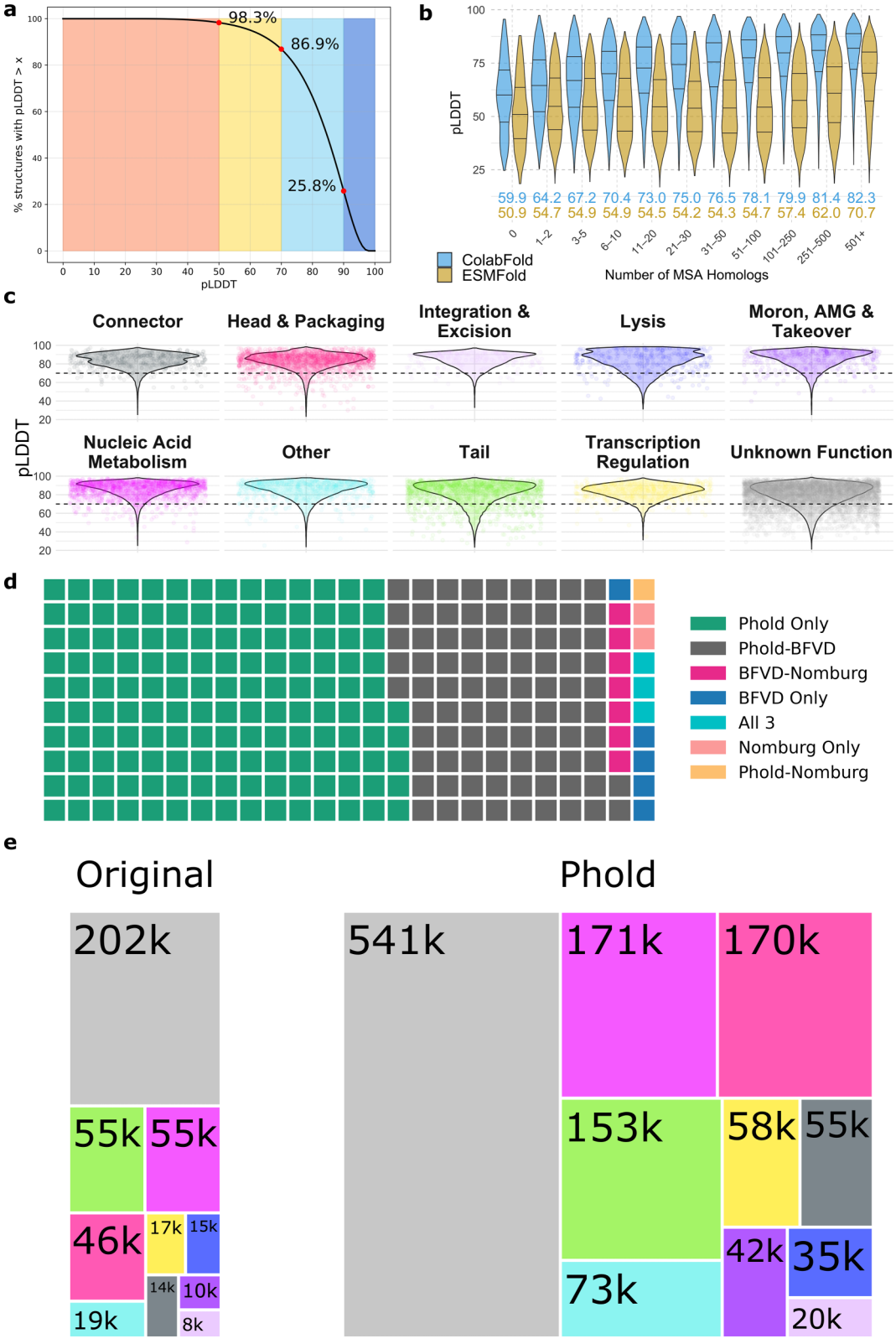


Figure 1. Phold Databases Predicted Quality, Clustering, and Annotation Transfer Overview. (**a**) The proportion of protein structures above a given mean pLDDT value in the Phold Search Database 1.36M (*y*-axis) plotted against the pLDDT value (*x*-axis). Colours follow the AlphaFold2 pLDDT colour scheme as implemented in ColabFold's visualization. (**b**) Violin plots showing pLDDT for ColabFold (blue) and ESMFold (gold) (*y*-axis) against the number of homologs in the protein's ColabFold MSA (*x*-axis) in bins for all PHROG, enVhog, efam, and PHROG singletons proteins ($n = 3\,099\,094$). (**c**) Violin plots showing the distribution of mean pLDDT values for proteins in each of the 10 PHROG functional categories. Each protein is a separate point. A subsample of 1% of all proteins were taken for visualization clarity. (**d**) Waffle plot showing the number of Foldseek clusters containing exclusively viral proteins (i.e. no AFDB50 members) shared between Phold DB 3.16M, BFVD, and Nomburg. One square represents ~500 clusters. (**e**) The number of proteins in each of the ten PHROG functional categories in the original PHROGs v4 database (left) compared to the Phold Search DB (right). The colour scheme matches (b), with the largest grey boxes constituting unknown-function proteins belonging to a PHROG group.

for all nine functional PHROG categories was high, ranging from 81.1 for tail proteins up to 85.6 for nucleic acid metabolism proteins (Supplementary Table S2). Poor predictions were more commonly generated for tail and lysis category proteins compared to the other categories (Fig. 1C), with 10th percentile pLDDT means of 63.5 and 66.4, respectively, compared to a minimum of 71.4 for any other functional category (Supplementary Table S2). The quality of predicted structures for PHROGs with unknown function was high (mean pLDDT 79.8) and much higher than enVhog and efam proteins with no assignable PHROG (mean pLDDT 70.2) (Supplementary Fig. S1). 815/856 unique PHROG product descriptions had mean pLDDT of at least 70 (Supplementary Table S3).

Phold's database expands the known phage protein structure space

To understand the novelty of Phold's predicted viral protein structures, we co-clustered Phold DB 3.16M with AlphaFold Database's AFDB50 representatives [51, 69] ($n = 53\,665\,860$) and two existing viral protein-structure databases—BFVD [29] ($n = 351\,242$) and Nomburg [33] ($n = 74\,129$)—with Foldseek using a 70% coverage cutoff. 57 257 433 proteins yielded 16 590 479 clusters, 14 395 143 of which were singletons (25.1%), while 42 862 290 (74.9%) belonged to 2 195 336 clusters with at least two members. Remarkably, 441 724 clusters summarizing 23 633 190 proteins contained at least one member from any viral database (i.e. Phold DB 3.16M, BFVD, or Nomburg; 'viral clusters') (Supplementary Table S4). 250 319 (56.7%) of these viral clusters contained both AFDB50 and Phold DB proteins only (Supplementary Fig. S4 and Supplementary Tables S4–S5). Phold DB proteins were present in 95.2% of non-singleton viral clusters containing 97.2% of non-singleton viral cluster members, while AFDB50 proteins were present in 71.3% of non-singleton viral clusters containing 98.2% of such members. 73 525 clusters containing 215 138 members (16.7% of clusters containing 0.9% of members) were unique to Phold DB (Supplementary Fig. S4 and Supplementary Tables S4–S5), suggesting Phold DB primarily expands the diversity of less common viral structures and folds.

Of the 441 742 viral clusters, 126 753 clusters containing 422 630 proteins did not contain any AFDB50 member. Of these 126 753 clusters, Phold DB structures were present in 119 875 (94.6%), with 73 525 containing only Phold DB proteins (58.0%) (Fig. 1D). There was substantial overlap between Phold DB and BFVD, with 44 278 (35.0%) clusters containing only Phold DB and BFVD proteins. This suggests that despite its relatively small size (i.e. almost 10-fold smaller than Phold DB), BFVD covers a substantial portion of viral protein structure space outside of the AFDB.

Overall, 54.5% (1 957 655/3 591 573) of viral database proteins could be structurally clustered with AFDB50. This suggests that despite explicitly omitting viral proteins, the AFDB implicitly includes proteins with homology to many viral and phage protein structures, covering a substantial portion of the known viral protein structure space. This could be due to the presence of proteins found on both phages and across the tree of life, and the inclusion of provirus protein structures present in AFDB, particularly those derived from prophages [70]. Additionally, it may explain why our efforts to fine-tune ProstT5 using Phold DB protein structures showed only modest improvements compared to using the standard ProstT5 (Supplementary Note 2, Supplementary Fig. S5, and Supplementary Table S6).

Structural similarity expands annotation labels in Phold's database

The major impact of the original PHROG database was the manual high-quality expert-curated annotations provided for 5 133/38 880 PHROGs (13.2%) containing 239 059/440 550 unique proteins (54.3%), along with 165/70 234 PHROG singletons (0.2%). Given the structural similarities between Phold DB proteins with the AFDB and BFVD, we used structural similarity searches to guide manual curation (see 'Materials and methods' section) to transfer functional annotations to an additional 2 798 PHROGs containing 30 490/440 550 (6.9%) PHROG proteins along with 3 019 singletons (Supplementary Fig. S6 and Supplementary Table S7). Combining proteins enVhog and efam proteins assigned PHROGs and specialized databases; 822 262 unique proteins in the Phold Search DB are assigned functional annotations, while a further 541 442 are included with a PHROG but with yet unknown function (Fig. 1E).

Phold's framework for phage protein annotation

Phold combines the structure-informed pLM ProstT5 [71] with the structure-comparison tool Foldseek [26] to search against the Phold Search DB 1.36M containing 1 363 704 protein structure predictions. Phold's workflow consists of four stages: (a) initial gene calling and genome feature annotation; (b) Foldseek 3Di token prediction using ProstT5 or 3Di token extraction from pre-computed PDB/mmcif; (c) Foldseek structure comparison and annotation transfer; and (d) plotting and summary generation (Fig. 2).

Phold accepts Pharokka [16], Bakta [72], or NCBI GenBank formatted genomes as the desired input. Alternatively, Phold can take nucleotide FASTA as input and utilizes Pyrodigal-gv [18, 57, 58] as a gene caller, as it has been shown to improve the annotation of phages that use alternative genetic codes [73]. Phold extracts all predicted CDSs, generates 1024-dimensional embeddings for each CDS residue using the 1 billion-parameter ProstT5 pLM encoder. These embeddings are then fed into a two-layer CNN trained to predict the Foldseek 3Di token at each residue. Alternatively, 3Di tokens are extracted from predicted protein structures in the PDB or mmCIF format, if available. Phold then uses Foldseek to search against Phold's database, which consists of predicted phage protein structures based on the PHROGs database, along with other specialized sub-databases. The top-ranking hit below the *E*-value with an assigned function in the database is finally transferred to annotate each query protein.

ProstT5 confidence is well-calibrated and positively correlated with protein structure quality

We next analysed reliability of generating ProstT5 3Di tokens compared to established methods (i.e. from ColabFold/Alphafold2 structures) and reliability metrics (i.e. pLDDT) using the ProstT5 confidence metric (Supplementary Note 3 and Supplementary Figs S7 and S8). Across all proteins in Phold DB 3.16M, we found that the empirical similarity between ProstT5 3Di predictions and 3Di tokens generated from predicted protein structures was strongly positively correlated (Spearman's $\rho = 0.87$) (Fig. 3A). ProstT5 confidence was also

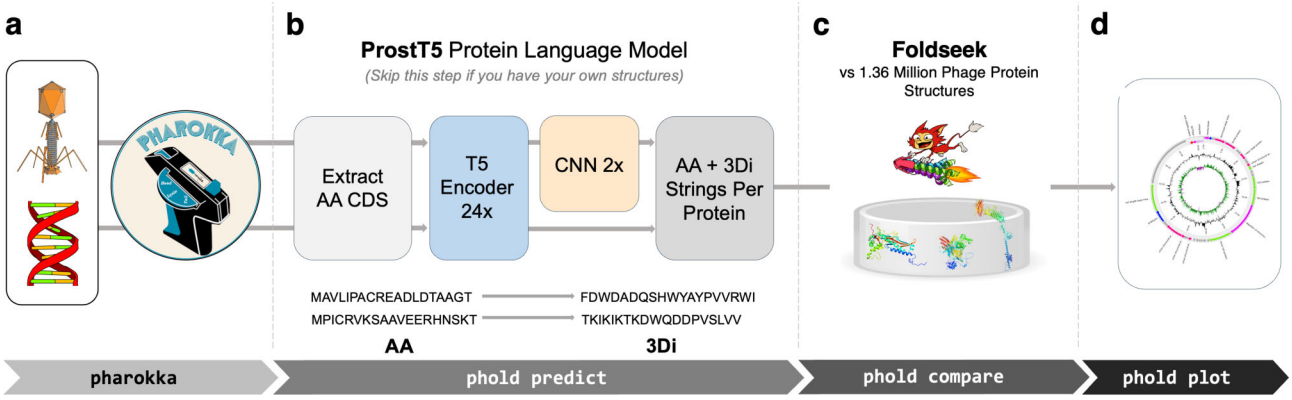


Figure 2. Phold's workflow for annotating bacteriophage genomes. (**a**) Phold begins with assembled phage genomes. Pharokka is run first to predict CDSs and other genomic features such as transfer RNA and transfer-messenger RNAs and to annotate CDS using sequence-based homology-detection methods. (**b**) The 'phold predict' module predicts Foldseek 3Di tokens for each residue of each CDS using the ProstT5 pLM encoder and CNN. Alternatively, predicted protein structures can be used instead of this module if available. (**c**) The 'phold compare' module then compares every CDS using the amino acid and ProstT5-predicted 3Di tokens against the Phold Search DB 1.36M containing 1 363 704 protein structures. The top-ranking database hit (based on Foldseek *E*-value) with a non-hypothetical function is assigned to each CDS. (**d**) Summaries and genomic plots can then be created with the 'phold plot' module.

positively correlated with mean pLDDT (Spearman's ρ 0.40) (Fig. 3B). ProstT5 confidence, though having the same range of possible values as pLDDT (i.e. 0–100), was generally substantially lower overall (mean 58.96 compared to 75.69 for pLDDT). ProstT5 confidence was higher for the Phold Search DB 1.36M proteins (mean 61.33) compared to the proteins without PHROG group assignment (mean 57.17). Although correlated, we caution interpreting ProstT5 confidence exactly like pLDDT, particularly in terms of any hard cutoffs (i.e. the widely used pLDDT 70 threshold does not translate to a specific ProstT5 confidence score).

Phold with ProstT5 is orders of magnitude more resource efficient than with ESMFold or AlphaFold2 while yielding similar annotation performance

We next benchmarked the annotation performance of Phold using ProstT5-based 3Di predictions with Foldseek. We chose four benchmarking datasets, ranging from relatively well-studied isolated and taxonomically classified phages (INPHARED [10] 182) to metagenomically assembled phage genomes from various sources (Tara [56], Cook [3], and Crass [6]). The benchmarked task was whether Phold with ProstT5 could annotate the PHROG category (general, practically useful e.g. 'tail'), product (specific, practically useful e.g. 'tail fiber protein'), or PHROG group (very specific, less practically useful e.g. 'PHROG 295') for each protein compared to using the pseudo ground truth of annotation using Foldseek with ColabFold (AlphaFold2) structures top hits at an E-value cut-off of 0.001.

We first benchmarked the impact of masking residues with low ProstT5 3Di prediction confidence at different thresholds to see whether masking may be used as an estimate of 3Di prediction quality akin to pLDDT for protein structure prediction. Masking residues below 15, 20, 25, and 30 ProstT5 confidence led to improvement in recall with minimal impact on precision on INPHARED 182, with the binary annotation F1 increasing from 0.927 to 0.937 and multiclass micro-averaged F1 from 0.921 to 0.930 compared to no masking (Fig. 3C). For the specific PHROG product task, masking residues below 25 increased multiclass micro-averaged F1 from 0.916 to 0.924 (Supplementary Fig. S9A), with similar results for the general PHROG category prediction task (multiclass micro F1: 0.862 versus 0.878) (Supplementary Fig. S9B). When masking residues above ProstT5 confidence of 30, performance decreased, driven by collapsing recall (Fig. 3C, Supplementary Fig. S9, and Supplementary Tables S8 and S9) as increasingly more 3Di residues were masked. Similar results were observed for the three metagenomic phage datasets (Supplementary Fig. S10). We chose to implement a default ProstT5 confidence masking threshold of 25 in Phold.

We then benchmarked Phold using ProstT5 versus ESMFold-derived structures across varying Foldseek *E*-value thresholds, using pseudo ground truth from ColabFold ($E = 0.001$). ProstT5 enabled ∼50–100 × faster 3Di token generation than ESMFold (Supplementary Table S10), with only marginal reductions in annotation performance across all four datasets (Fig. 3D–F, Supplementary Figs S11 and S12, and Supplementary Tables S11 and S12). On the INPHARED 182 dataset of relatively well-characterized phages, both methods performed similarly across PHROG group, product, and category levels, with optimal precision-recall trade-offs at $E = 0.01$–0.001. Phold with ProstT5 showed strongest performance on the more practically useful broader annotation tasks (F1 = 0.939 for category; F1 = 0.924 for product), with reduced accuracy at the group level (F1 = 0.855 at $E = 0.001$). Performance was robust across all nine PHROG categories on INPHARED 182 (F1 = 0.889–0.965; Fig. 3G), and generally consistent for Cook and Tara datasets, but lower on Crass, particularly for the uncommon connector category (F1 = 0; $n = 11$; Supplementary Fig. S12B), indicating the difficulty of this dataset. Notably, the therapeutically relevant integration and excision category showed high performance across datasets (F1 = 0.872–0.965).

As Foldseek search sensitivity can be varied and run with either CPU- or GPU-acceleration [74] applying different prefilter algorithms, we next ran a series of ablations to test the impact of varying hardware and sensitivity and database pre-clustering on Phold's performance and runtime (Supplementary Table S13). Using INPHARED 182, Phold using ProstT5 with Foldseek-GPU was faster than Foldseek-CPU at default sensitivity. Annotation performance increased marginally with sensitivity (Fig. 3H), at the cost of substantial runtime increases (Fig. 3I). Phold with

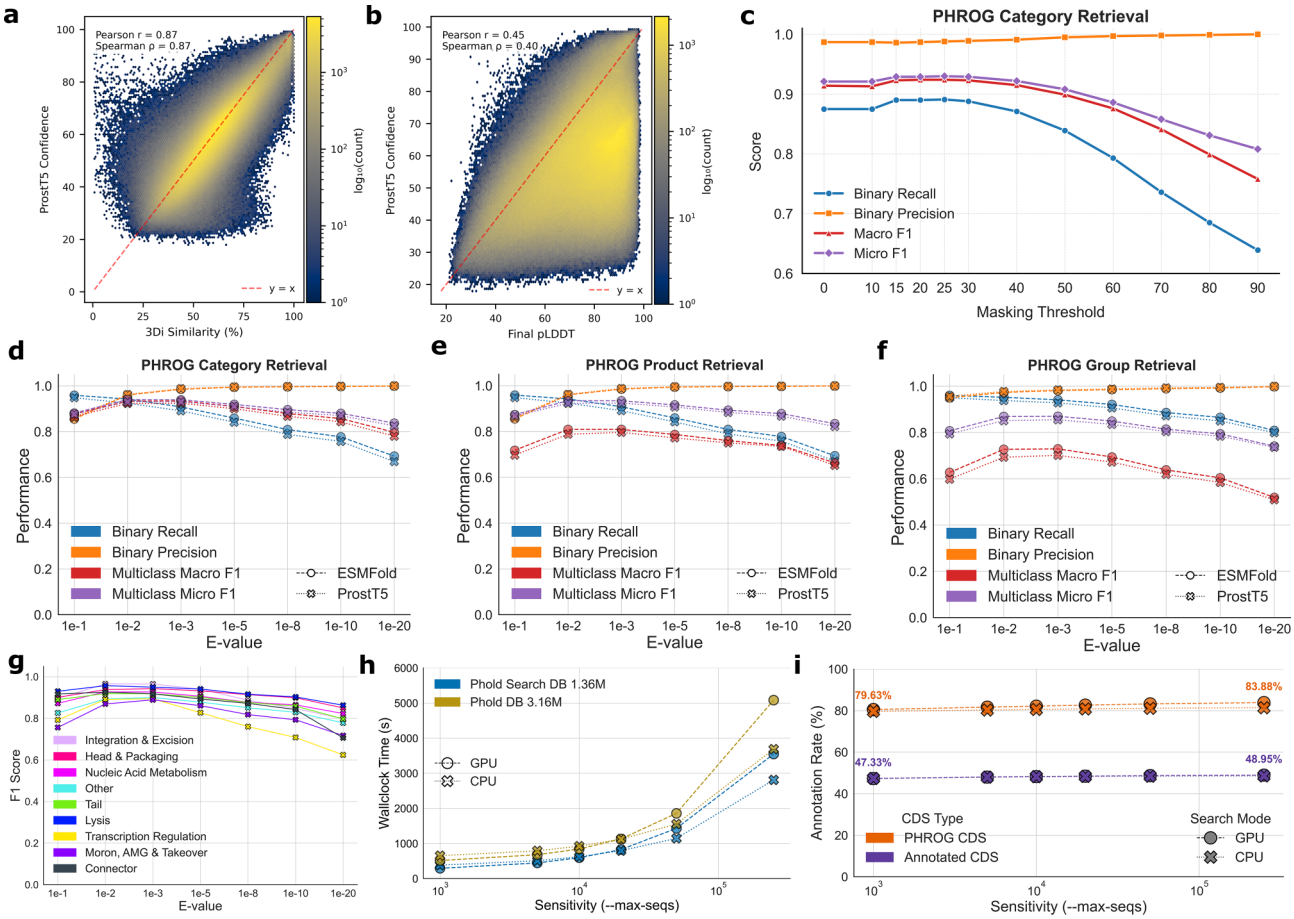


Figure 3. Phold with ProstT5 annotation performance. (**a**) Empirical sequence similarity between ProstT5-CNN generated 3Di tokens and the Foldseek VQ-VAE derived 3Di tokens (*x*-axis), and ProstT5 confidence (*y*-axis) for all Phold DB 3.16M protein structures. (**b**) Final pLDDT (i.e. the higher of ColabFold or ESMFold pLDDT) (*x*-axis), and ProstT5 confidence (*y*-axis) for all Phold DB 3.16M protein structures. (**c**) PHROG category annotation performance of Phold at *E*-value = 0.001 when masking low ProstT5-confidence residues below threshold (*x*-axis) on the INPHARED 182 dataset, compared to the pseudo ground truth of annotations with ColabFold-generated structures at *E*-value 0.001. Binary recall and precision measure whether ProstT5 with masking can annotation any non-hypothetical annotation for annotated ground-truth proteins, while multiclass F1 scores are measured across all nine PHROG categories. (**d**) PHROG category annotation performance of Phold using ProstT5 (crosses) and ESMFold (circles) at a variety of *E*-values on the INPHARED 182, compared to the pseudo ground truth of annotations with ColabFold-generated structures at *E*-value 0.001. The same but for the PHROG product annotation task (i.e. more specific annotations) (**e**) and PHROG group assignment task (**f**). (**g**) Individual PHROG category annotation performance of Phold using ProstT5 at a variety of *E*-values on the INPHARED 182, compared to the pseudo ground truth of annotations with ColabFold generated structures at *E*-value 0.001. (**h**) Runtime of 'phold compare' on the INPHARED 182 dataset (*y*-axis) for different sensitivity values (i.e. Foldseek '–max-seqs' parameter that dictates the number of hits that pass through the prefilter) (*x*-axis), using both Foldseek-CPU and Foldseek-GPU and Phold Search DB 1.36M and Phold DB 3.16M. (**i**) Functional (purple) and PHROG (orange, i.e. also including unknown function PHROG hits) annotation rates with Phold using ProstT5 across the 16 460 CDS in INPHARED 182 (*y*-axis) using the Phold Search DB 1.36M for different sensitivity (i.e. Foldseek '–max-seqs') values (*x*-axis).

Foldseek-GPU was able to detect more CDS hits to any PHROG (i.e. including PHROGs with unknown function) compared to Foldseek-CPU regardless of sensitivity (Fig. 3I), but there were minimal differences in terms of functionally annotated proteins (Fig. 3I). This trend is observed because Foldseek-GPU performs full alignment of all sequences, while its CPU counterpart performs *k*-mer-based pre-filtering first and only aligns high-quality hits. Pre-clustering the Phold Search DB did not substantially impact annotation rate or runtime (Supplementary Note 4 and Supplementary Fig. S13).

Phold enables consistent annotation of over 50% of genes for an average phage genome

We next compared Phold's annotation performance with that of existing sequence homology-based phage annotation methods. We compared Phold to Pharokka for several reasons: Pharokka is widely used in the phage genome annotation field; it has options to search using either PSSM profiles (MMseqs2) or with HMM (PyHMMER)-based homology inference; and it utilizes the same hierarchy of PHROG categories and product annotation labels, making it easy to compare functional labels across different homology detection methods.

We tested the annotation performance on 1 419 INPHARED viruses from distinct genera (see Supplementary Table S14 for information on all viruses included). We compared four methods: Pharokka using MMSeqs2, Pharokka using PyHMMER, Phold with Foldseek and ProstT5 for 3Di, and Phold with Foldseek and 3Di from ColabFold-predicted structures. The 1419 viruses contained 148 194 CDS. We generated ColabFold structure predictions for 147 946 proteins, consisting of all but the longest proteins (over 3000 AA) (mean pLDDT of 77.05). This dataset represents the largest

consistently annotated set of protein structural predictions across the diversity of known bacteriophage and archaeal virus genomes.

Phold with ProstT5 annotated 73 268 (49.4%) CDS, increasing to 76 322 (51.5%) using ColabFold-generated structures, a marked increase from Pharokka's 51 399 (34.7%) with MMseqs2 and 55 923 (37.7%) with PyHMMER. Phold was able to improve annotations for all nine PHROG categories (Fig. 4A), with transcriptional regulation showing the highest proportional increase (1742 with Pharokka with MMseqs2 versus 3323 Phold with structures) and lysis the lowest (2879 versus 3847). The annotation improvements across PHROG categories for the Cook, Tara, and Crass metagenomic datasets were similar (Supplementary Fig. S14A–C).

The median per-phage annotation rate for Phold was 54.3% with ProstT5 and 56.3% with structures, similar to previously reported annotation rates for phages using Foldseek with ColabFold structures searching against BFVD [29] or AFDB [51] (Fig. 4B). Phold annotation rates were similar for Cook (Fig. 4C) and Tara (Fig. 4D), despite having lower Pharokka annotation rates. While Phold annotation rates were lower for Crass, they were still more than double Pharokka's (Fig. 4E). These results indicate that Phold is especially useful compared to sequence-based methods for annotating more distant, difficult-to-annotate phages, with computation efficiency allowing scaling to large metagenomic-scale datasets.

Phold's annotation rate strongly depended on protein length (Supplementary Fig. S15, Supplementary Table S15, and Supplementary Note 5). Short proteins (<100 AA) were rarely annotated (17.7% Phold annotation rate on INPHARED 1419), whereas long proteins (>250 AA) were almost always annotated (89% on INPHARED 1419). The annotation rate, especially for short proteins, improved further when using complementary tools like Phynteny [37], which leverages gene-order information to annotate remaining hypotheticals after Phold. For example, combining Phynteny with Phold increased the median per-phage annotation rate on INPHARED 1419 from 56.3% to 64%, with similar gains across other the other datasets (Supplementary Fig. S16 and Supplementary Note 6). The largest gain was for short proteins (<100 AA), where the annotation rate increased to 26.4% on INPHARED 1419. While any approach will always find annotating shorter proteins more difficult (as their sequence contains less information than longer ones), combined with the fact that short proteins tend to be difficult for structure prediction methods [46], Phold's relatively poor performance on them suggests different approaches using orthogonal information may be required.

On a single NVIDIA A100 40GB GPU, ProstT5 3Di inference for all 1419 INPHARED viruses with Phold took 3942 seconds, whereas generating large-scale protein structure predictions with ColabFold/AlphaFold2 involves generating MSAs, requires large-scale computational infrastructure, and is three to four orders of magnitude more resource intensive [71].

1393/1419 (98.2%) and 943/1419 (66.5%) of the INPHARED viruses had at least 30% and 50% of their CDS annotated, respectively. Archaeal viruses had a lower annotation rate (median 40%, $n = 52$) compared to bacterial viruses (median 57.1%, $n = 1319$) and with unknown host (median 47.5%, $n = 48$) (Fig. 4F). Of the host genera with at least 10 infecting viruses, the median annotation rate ranged from 38.3% (*Halocarcula*) to 69.3% (*Burkholderia*) (Supplementary Fig. S17A). Phages for *Staphylococcus* and *Escherichia* had among the highest annotation rates (67.9% and 63.2%). The percentage of Phold annotated CDS was strongly correlated with the mean ColabFold pLDDT for all CDS in the virus, regardless of whether ProstT5 or ColabFold structures were used (Phold with ColabFold structures: Spearman ρ 0.60 Fig. 4G: Phold with ProstT5: Spearman ρ 0.584 Supplementary 15B). This is unsurprising given the dependence of both annotation and structure prediction quality on the existence of homologous proteins.

Phold allows for large-scale hypothesis generation of unknown phage protein functions

Detailed analysis of the 1419 INPHARED genome annotations (https://doi.org/10.5281/zenodo.17575988) revealed an interesting wealth of annotations to the specialized databases employed by Phold. 5/1419 phages contained six proteins with homology to CARD antimicrobial resistance proteins, five of which have sequence similarity below 30%. For instance, *Rhizobium* phage RHph_Y65 (GenBank accession: NC_070 967) possesses a CDS annotated by Phold with high structural similarity (TM-score 0.93, LDDT 0.76) to the CARD trimethoprim-resistant dihydrofolate reductase gene DfrA37 with sequence similarity of only 26% (Fig. 5A), while *Vibrio* phage VAP7 (GenBank accession: NC_048 765) also possesses a CDS with high structural similarity (TM-score 0.91, LDDT 0.77) to the related CARD trimethoprim-resistant dihydrofolate reductase gene dfrA26 (Supplementary Fig. S18A). Both hits were also found with Phold ProstT5 (Foldseek E-values $1.8E^{-17}$, $5.3E^{-18}$), but missed with Pharokka using MMseqs2-based sequence homology searches against CARD. The presence of phage-encoded dihydrofolate reductases does not always confer trimethoprim resistance [75], and structural similarity is no guarantee of function, but it has been shown to occur in some prophages [76] and has been implicated in resistance phenotypes.

Proteins with hits to Anti-CRISPRs were found in 82/1419 INPHARED viruses, containing 112 predicted Anti-CRISPRs, while 117 proteins with structural similarity to virulence factors were found in 85/1419 viruses. Sixty-two proteins with structural similarity to the DefenseFinder database of bacterial anti-phage proteins were found in 52/1419 viruses. For example, two *Clostridioides* prophages (phiC2, GenBank accession: NC_009231 and phiCD27, GenBank accession: NC_011 398) encode a CDS with only 20% sequence identify to the Pycsar defense system [77] cyclase, despite extremely high structural similarity (TM-score 0.87, LDDT 0.77) (Fig. 5B), while six *Synechococcus* phages encode CDS with very low sequence identity (12.7%–17.1%) but high structural similarity to the DruA protein from the Druantia Type I defense system [78] (TM-score 0.73, LDDT 0.56) (Fig. 5C). Additional examples include a CDS from *Klebsiella* prophage ST437-OXA245phi4.1 (GenBank accession: NC_049 448) with similarity to the GAPS1 system [79] (TM-score 0.87, LDDT 0.66) (Fig. 5D); AriaA homologs from both subsystems of the PARIS defense system [80] on *Fusobacterium* phage vB_FnuS_FNU3 (Genbank accession: OQ808965) and *Yersinia* phage vB_YenM_31.17, respectively (Supplementary Fig. S18B and C); and a letS protease homolog to the Gao_let defense system [81] on

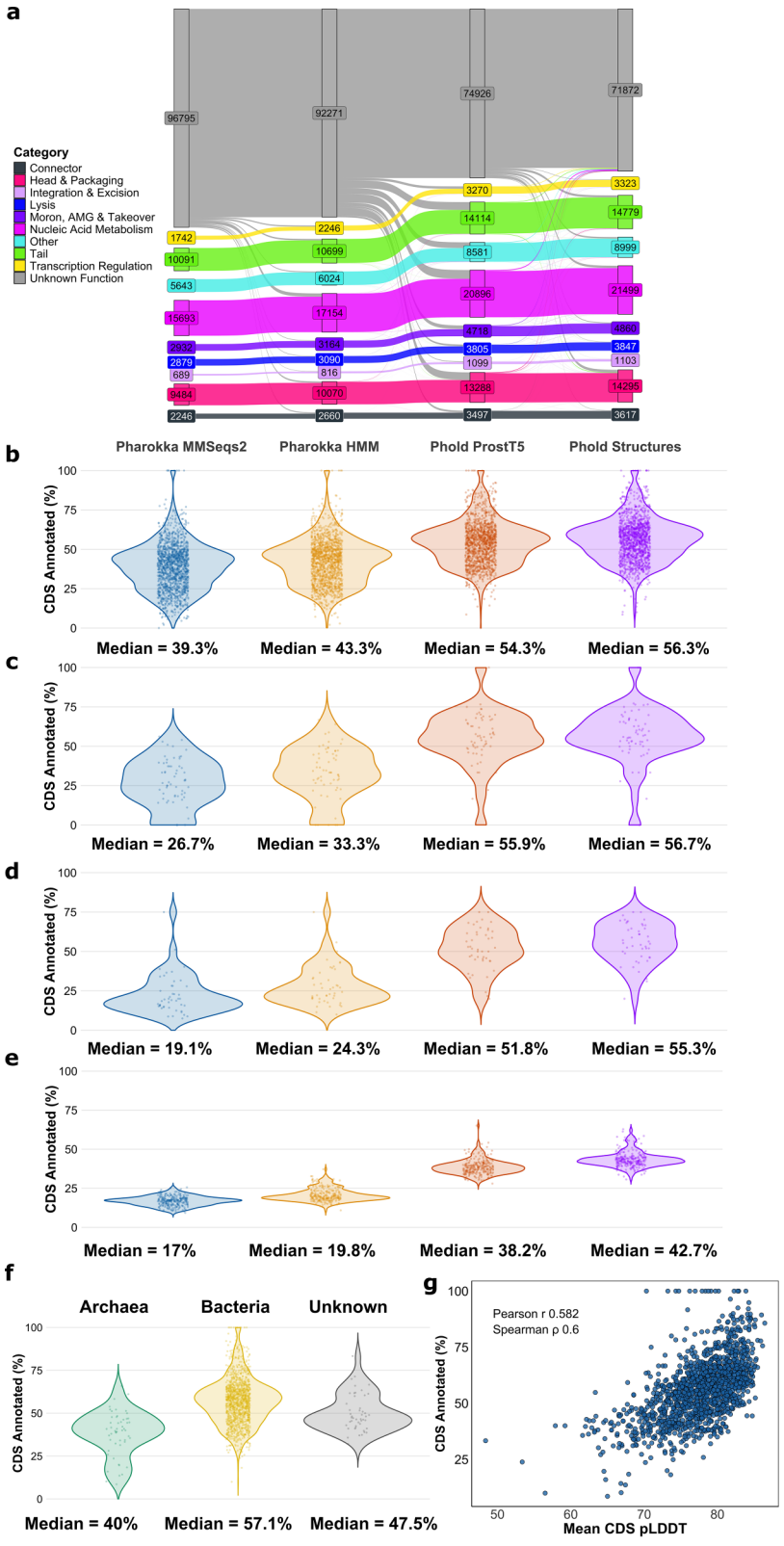


Figure 4. Phold annotates over 50% of CDSs in the average phage genome. (**a**) Flow of annotations for all 10 PHROG categories (including unknown function) between the four annotation methods: Pharokka with MMSeqs2 (left); Pharokka with HMMs (PyHMMER) (centre-left); Phold with ProstT5 (centre-right); and Phold with ColabFold structures (right). Percentage of CDS annotated (*y*-axis) for every virus in the (**b**) INPHARED 1419, (**c**) Cook, (**d**) Tara, and (**e**) Crass datasets for each of four annotation methods. Each point represents a phage, with the median indicated at the bottom of each plot. (**f**) Percentage of CDS annotated (*y*-axis) for every virus in the INPHARED 1419, grouped by the domain of the host taxa. 'Unknown' indicates that the virus was metagenomically assembled, or the host was not available in the metadata associated with the genome. (**g**) Mean CDS ColabFold pLDDT (*x*-axis) against the percentage of CDS annotated (*y*-axis) for every virus in INPHARED 1419. Each point represents the mean CDS pLDDT for one virus.

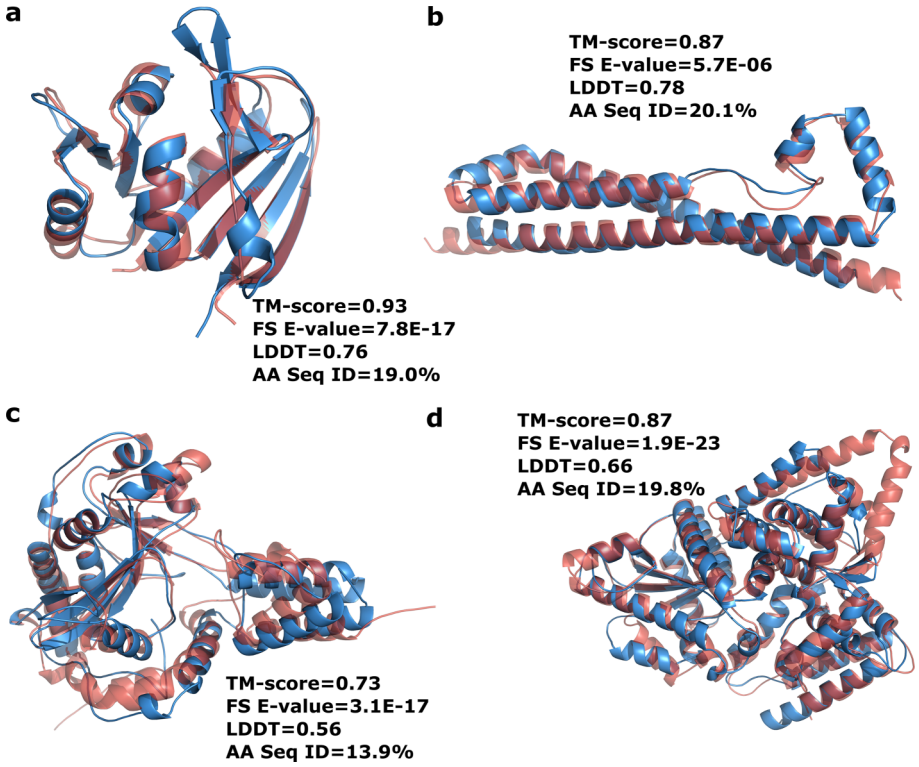


Figure 5. Phold allows for the discovery of phage proteins with high structural similarity to anti-microbial resistance and anti-phage defense proteins. (**a**) ColabFold-predicted structures of a predicted dihydrofolate reductase from *Rhizobium* phage RHph_Y65 (GenBank accession: NC_070967) (blue) and CARD trimethoprim-resistant dihydrofolate reductase gene DfrA37 (red). (**b**) Predicted cyclase from *Clostridioides* phage phiCD27 (GenBank accession: NC_011398) (blue) and the Pycsar defense system cyclase (red). (**c**) Predicted DruA-like protein from *Synechococcus* phage S-PRM1 (GenBank accession: NC_055761) (blue) and DruA protein from the Druantia Type I defense system (red) (**d**) Predicted GAPS1-like protein from *Klebsiella* prophage ST437-OXA245phi4.1 (blue) and GAPS1 defense system protein (red). 'FS E-value' = Foldseek *E*-value. 'AA Seq ID' = amino acid sequence identity.

Enterocbacter phage phiT5282H (GenBank accession: NC_049 429) (Supplementary Fig. S18D).

While structural homology is no guarantee of functional similarity, the prevalence of homologs to bacterial anti-phage defense system proteins in the 1419 INPHARED viruses emphasizes that phages may serve as a reservoir of anti-phage defence systems that are prophage encoded [82], a concern that supports the exclusion of using prophages in therapeutic contexts [83]. It also suggests that phages may use similar mechanisms to defeat or mimic known bacterial defense systems in the phage-host arms race [84].

Many phage proteins have structural similarity to proteins across the tree of life

The scale and quality of protein structure predictions and annotations of the Phold DB 3.16M and the 1419 INPHARED viruses enable broad-scale comparison to proteins belonging to vastly different taxa [85, 86]. Using thresholds of 70% bidirectional coverage, a maximum Foldseek *E*-value of 0.01, and minimum TM-score of 0.6 to ensure a high likelihood of a shared fold across the protein [45], we found 346 495 (10.9%) of Phold DB 3.16M proteins had AFDB hits to *Eukaryota* proteins, 632 793 (20.0%) to *Archaea* proteins, and 1 176 208 (37.1%) to *Bacteria* proteins. 296 912 (9.4%) of proteins had hits to all three domains of life (Supplementary Table S16). For INPHARED 1419, the numbers were proportionally similar between the three domains and higher overall [24 654 (16.7%) for *Eukaryota*, 43 155 (29.2%) for *Archaea* and 74 103 (50.1%) for *Bacteria* and 22 129 (15.0%) with hits to all three domains] (Supplementary Table S17). Proteins with metabolic or enzymatic functions belonging to nucleic acid metabolism, other, and moron, auxiliary metabolic gene and host takeover PHROG categories had the highest proportion of hits (Fig. 6A and B). Transcription regulation and unknown function categories had the lowest proportions of hits (Fig. 6A and B), indicating that finding structural similarity to proteins in these PHROG categories is potentially more difficult due to their shorter length [46] (Supplementary Table S1) and that proteins in these categories (especially unknown function) are more likely to be unique to phages.

The Phold DB 3.16M PHROG with the most hits to all three domains was PHROG 1423 (12 673 hits) (Supplementary Tables S16 and S17), containing DGR reverse transcriptases, unsurprising given the prevalence of retrotransposons and retroviruses in *Eukaryota* and prokaryotes [87]. For example, homologs to the reverse transcriptase present in *Faecalibacterium* phage Taranis (GenBank accession: NC_047 914) were found in the agaric fungus *Mycena chlorophos* and an uncultured archaeon with TM-scores of 0.87 and 0.93 despite sequence identities of only 19.5% and 23.0% (Fig. 6C).

To focus our analysis on a subset of well-annotated genomes, we repeated the same analysis for the 48 high-quality reference proteomes (16 from *Bacteria*, 31 from *Eukayota*, and a single archaeon, *Methanocaldococcus jannaschii*) with 564 446 proteins that make up the AlphaFold-Proteome database [68]. Hits to enzymatic and metabolic PHROG categories again dominated, with fewer hits to phage

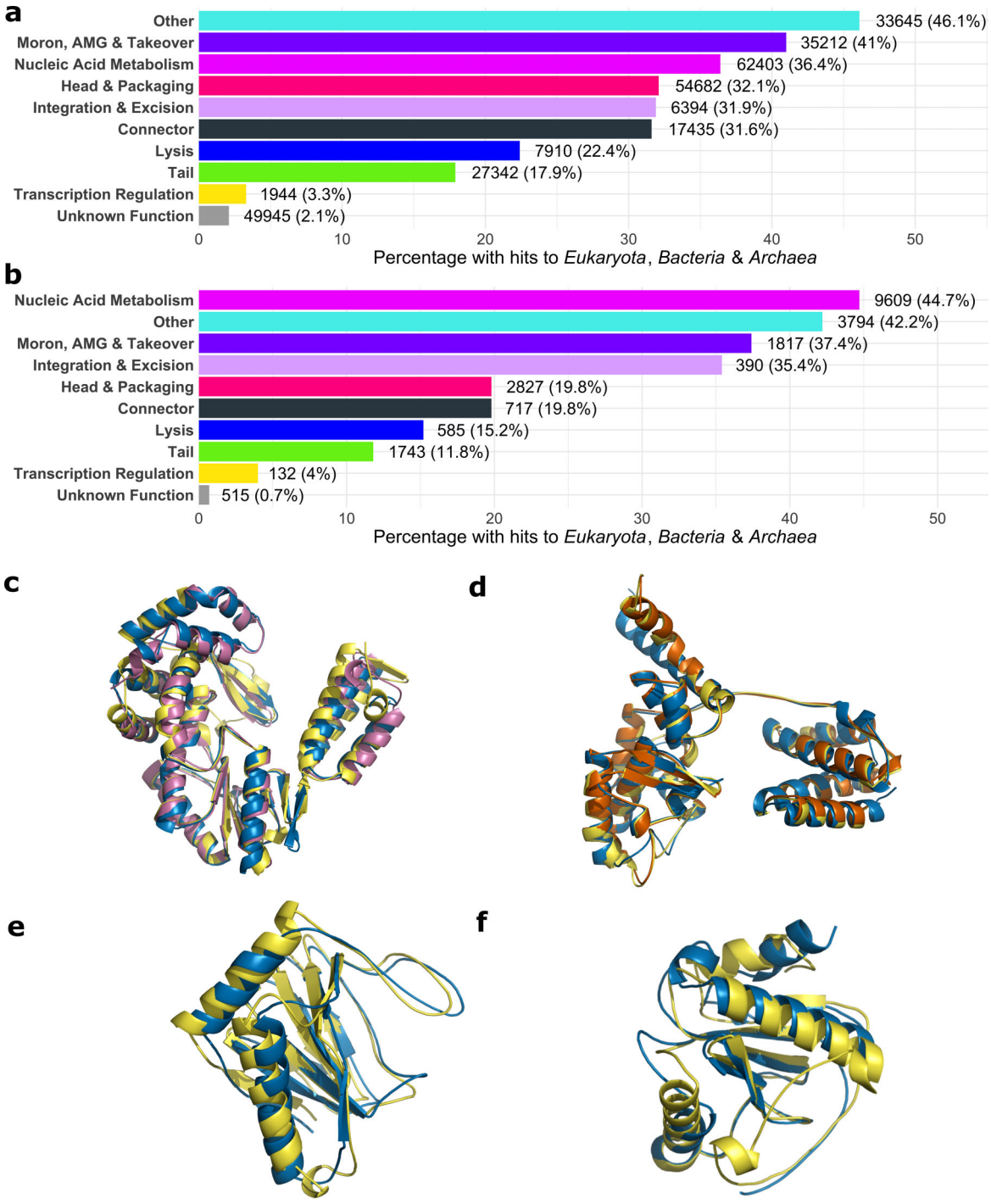


Figure 6. Structural similarity of phage proteins across the tree of life. Percentage of proteins belonging to each PHROG category that have strong structural similarity (Foldseek *E*-value < 0.01 and alignment TM-score > 0.6) to all three domains of life (*Eukaryota*, *Bacteria*, and *Archaea)* from AlphaFold Database proteins for (**a**) Phold DB 3.16M proteins and (**b**) INPHARED 1419 proteins. The raw number and percentage per category are indicated to the left of each bar. (**c**) Shows the predicted protein structures for reverse transcriptases from *Faecalibacterium prausnitzii* infecting phage Taranis (GenBank accession: NC_047914) in blue, *Mycena chlorophos* (AFDB accession AF-A0A146HGT5-F1-v4, sequence identity 19.5%, TM-score 0.87) in yellow, and from an uncultured archaeon (AFDB accession AF-A0A1S6HH33-F1-v4, sequence identity 23%, TM-score 0.93) in purple. The structures were trimmed, keeping only the residues of the Foldseek alignments, and the top hit to a bacterium (AF-A0A2A7AF35-F1-v4) was omitted for clarity as it was near identical (TM-score 0.98). (**d**) Shows the predicted integrase of *Mycobacterium* phage Highbury (GenBank Accession: OR521086) in blue, and predicted recominbases/integrases from *Trichuris trichiura* (AFDB accession AF-A0A077ZEL6-F1-model_v4, sequence identity 21.4%, TM-score 0.92) in yellow, and *Enterococcus faecalis* (AFDB accession AF-A0A132P2T2-F1-model_v4, sequence identity 17.9%, TM-score 0.92) in red. The predicted recominbase from *Methanocaldococcus jannaschii* (AFDB accession AF- Q57813-F1-model_v4, sequence identity 22.6%, TM-score 0.78) was omitted for clarity. (**e**) Shows the predicted protein structures for a predicted 2OG-Fe(II) oxygenase on *Synechococcus* phage S-H9-1 (GenBank Accession: NC_070961) in blue and the human Prolyl hydroxylase EGLN3 (UniProt accession: Q9H6Z9) in yellow, with alignment sequence identity 18.6% and TM-score of 0.87. The structures were trimmed keeping only the residues aligned by Foldseek. (**f**) Shows the predicted protein structures for a predicted amidase on *Delftia* phage IME-DE1 (GenBank Accession: NC_028702) in blue and the human Peptidoglycan recognition protein 1 PGLYRP1 (UniProt accession: O75594) in yellow, with alignment sequence identity 16.0% and TM-score of 0.79. The structures were trimmed, keeping only the residues aligned by Foldseek.

structural categories such as tail and connector compared to hits against AFDB (Supplementary Fig. S19). For Phold DB, 345 836 (10.9%) had hits to *Bacteria* proteins, 157 877 (5.0%) to *Eukaryota* and 49 229 (1.6%) to *Methanocaldococcus jannaschii*, with 29 449 (0.9%) having hits to all three. For INPHARED 1419, the numbers of hits were 24 669 (16.7%), 12 737 (8.6%), and 3889 (2.6%), respectively, with 1913 (1.3%) shared. Of these 1913, the most common single annotation was 'integrase' (212). Many INPHARED 1419 integrases had extremely high (>0.8) TM-scores to proteins from different domains with sequence identities below 25%, suggesting integrases are extremely conserved structure and function across the tree of life [88]. For instance, the predicted integrase of *Mycobacterium* phage Highbury (GenBank Accession: OR521086) has TM-scores of 0.92, 0.92, and 0.78 with sequence identities of 21.4%, 17.9%, and 22.6% to predicted recominbases/integrases from *Trichuris trichiura*, *Methanocaldococcus jannaschii*, and *Enterococcus faecalis* respectively (Fig. 6D). Other core metabolic enzymes, such as thymidylate synthases, methyltransferases and glycosyltransferases, were commonly found in INPHARED 1419 phages with high structural similarity to proteins from other domains of life (Supplementary Table S17).

Finally, 54 406 Phold DB and 5 552 INPHARED 1419 proteins had *Homo sapiens* hits. Compared to hits shared between all three domains of life, hits to *Homo sapiens* were especially enriched towards phage non-structural PHROG categories of nucleic acid metabolism, other and moron, auxiliary metabolic gene and host takeover proteins (Supplementary Fig. S20). The most common PHROG groups from Phold DB 3.16M with *Homo Sapiens* hits generally had core nucleic acid metabolic functions, including to PHROG 1249 acetyltransferase (1423 proteins), PHROG 16 DNA helicase (1322 proteins), and PHROG 61 (2OG-Fe(II) oxygenase) (Supplementary Tables S18 and S19).

Analysis of the most hit *Homo Sapiens* protein for both Phold DB (2 459 hits) and INPHARED 1419 (303 hits), Prolyl hydroxylase EGLN3 (UniProt accession Q9H6Z9), revealed widespread distribution of structurally similar proteins throughout phages that infect marine bacteria such as *Synechococcus*, *Vibrio*,and *Cyanobacteria*. These proteins had amino acid sequence identities ranging from 8.4%–22.5% with EGLN3, with most annotated as '2OG-Fe(II) oxygenase' by Phold (Fig. 6E). Following structural MSA of these hits with FoldMason [89], we found that the two histidine residues comprising the iron cation binding sites [90] were highly conserved, while the next most conserved residue was glutamine, forming part of the conserved 2-His-1-carboxylate facial triad canonical for 2OG-Fe(II) oxygenases [91]. This suggests these proteins are correctly annotated as 2OG-Fe(II) oxygenases, which likely confer evolutionary advantages for these phages adapted to the marine environment. We also found 43 INPHARED 1419 (and 292 Phold DB) proteins with strong structural similarity to the human peptidoglycan recognition protein 1 PGLYRP1 (Uniprot accession O75594). These hits were present across a wide variety of phages annotated as 'amidase' or 'lysin' by Phold, with 14.5–26.1% amino acid similarity to PGLYRP1. This structural similarity suggests a similar mechanism of peptidoglycan recognition between these phages and PGLYRP1 in *Homo sapiens* (Fig. 6F).

Overall, such strong structural similarity between phage proteins encoding for core nucleic acid metabolism and other enzymatic functions, suggests a wealth of folds and functions that are present both in phages and across the tree of life [92].

Discussion

The availability of large-scale protein sequence and structure databases, combined with improvements in structural alignment algorithms, has allowed for the illumination of difficult-to-annotate protein dark matter [69, 93, 94]. Phage proteins are amongst the most difficult; sequence-based homology approaches leave over 65% of proteins on an average phage unannotated.

In this study, we present Phold, a framework that combines rapid structural information inference using the ProstT5 pLM with Foldseek, leveraging our database of over 1.36M phage protein structures to provide interpretable structure-based annotation with curated high-quality functional annotation labels. We show that Phold has wide applications to prokaryotic virus annotation, including identification of potential antimicrobial resistance proteins and anti-phage defence proteins. Our structural similarity analyses of Phold DB and annotated cultured phages show that many phage proteins, particularly those with nucleic acid metabolism and other enzymatic functions, have structural homologs across the tree of life, including in *Homo sapiens*. Given Phold's improved performance over sequence-based homology tools, we anticipate Phold will be extremely valuable for phage, archaeal virus, and viral metagenomic research communities. We also believe the functional annotations generated by Phold symbiotically provide a platform for guiding wet-lab studies to confirm or disprove these hypotheses, which can be used to continually improve functional labels for future phage annotation databases.

Acknowledgements

This work was supported with the assistance of resources and services from Phoenix HPC at the University of Adelaide and Pawsey Supercomputing Research Centre, which is supported by the Australian Government. We would like to thank Fabien Voisin and Sarah Beecroft for their assistance in operating ColabFold at scale at Phoenix and Pawsey, respectively, with extra acknowledgement to Sarah for containerizing ColabFold for use on Setonix's AMD GPUs. We would also like to thank Gemma Atkinson and Artyom Egorov for providing us with NetFlax protein structure predictions. R.A.E. was supported by awards from the Australian Research Council DP250103825 and FL250100019. M.S. acknowledges support by the National Research Foundation of Korea (grants 2020M3A9G7103933, RS-2021-NR061659, RS-2021-NR056571, and RS-2024-00396026), Samsung DS Research Fund, Creative-Pioneering Researchers Program, AI-Bio Research Grant through Seoul National University, and Novo Nordisk Foundation (NNF24SA0092560). M.M. acknowledges support from the National Research Foundation of Korea (grant RS-2023-00250470).

Author contributions: George Bouras (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Resources [lead], Software [lead], Validation [lead], Visualization [equal], Writing – original draft [lead], Writing – review & editing [lead]), Susanna Grigson (Conceptualization [supporting], Data curation [equal], Methodology [equal], Resources [supporting],

Software [supporting], Writing – review & editing [equal]), Milot Mirdita (Conceptualization [supporting], Methodology [supporting], Software [supporting], Supervision [supporting], Validation [supporting], Writing – review & editing [equal]), Michael Heinzinger (Conceptualization [supporting], Methodology [supporting], Software [supporting], Validation [supporting], Writing – review & editing [equal]), Bhavya Papudeshi (Methodology [supporting], Software [supporting], Validation [supporting], Writing – review & editing [equal]), Vijini Mallawaarachchi (Resources [supporting], Software [supporting], Validation [supporting], Writing – review & editing [equal]), Renee Green (Formal analysis [supporting], Visualization [equal], Writing – review & editing [equal]), Rachel Seongeun Kim (Methodology [supporting], Resources [supporting], Writing – review & editing [equal]), Victor Mihalia (Formal analysis [supporting], Methodology [supporting], Validation [supporting], Writing – review & editing [equal]), Alkis J. Psaltis (Funding acquisition [equal], Supervision [supporting], Writing – review & editing [equal]), Peter-John Wormald (Funding acquisition [equal], Supervision [supporting], Writing – review & editing [equal]), Sarah Vreugde (Funding acquisition [equal], Methodology [supporting], Supervision [equal], Writing – review & editing [equal]), Martin Steinegger (Conceptualization [supporting], Formal analysis [supporting], Methodology [supporting], Software [supporting], Supervision [equal]), and Robert A. Edwards (Conceptualization [supporting], Funding acquisition [equal], Methodology [equal], Resources [supporting], Supervision [equal], Writing – review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Data availability

Protein structure predictions for all Phold database proteins can be found at https://doi.org/10.5281/zenodo.16739199 (all Phold Search DB 1.36M structures) and https://doi.org/10.5281/zenodo.16741650 (all additional efam and enVhog structures without PHROG assignment in Phold DB 3.16M but not Phold Search DB 1.36M). Protein structure predictions, genomes, and GenBank-formatted annotation files for INPHARED 1419 viruses and the Cook, Crass, and Tara benchmarking datasets can be found at https://doi.org/10.5281/zenodo.17575988. Foldseek-formatted Phold databases are available at https://doi.org/10.5281/zenodo.16741548 and can be downloaded using 'phold install'.

Code availability

Phold is open-source software available at https://github.com/gbouras13/phold and https://doi.org/10.5281/zenodo.16750703. All other code required to recreate the results in this manuscript can be found at https://github.com/gbouras13/phold-analysis. A modified version of the ColabFold MSA search code using an additional database containing 129944764 non-redundant viral (predominantly phage) proteins is available at https://github.com/gbouras13/colabfoldv, while the database can be downloaded from https://doi.org/10.5281/zenodo.15045387.

References

1. Clokie MR, Millard AD, Letarov AV *et al.* Phages in nature. *Bacteriophage* 2011;1:31–45. https://doi.org/10.4161/bact.1.1.14942
2. Chikhi R, Lemane T, Loll-Krippleber R *et al.* Logan: planetary-scale genome assembly surveys life's diversity. biorxiv, https://doi.org/10.1101/2024.07.30.605881, 1 September 2025, preprint: not peer reviewed.
3. Cook R, Telatin A, Hsieh S-Y *et al.* Nanopore and Illumina sequencing reveal different viral populations from human gut samples. *Microbial Genomics* 2024;10:001236. https://doi.org/10.1099/mgen.0.001236
4. Yutin N, Benler S, Shmakov SA *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* 2021;12:1044. https://doi.org/10.1038/s41467-021-21350-w
5. Al-Shayeb B, Sachdeva R, Chen L-X *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* 2020;578:425–31. https://doi.org/10.1038/s41586-020-2007-4
6. Edwards RA, Vega AA, Norman HM *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* 2019;4:1727–36. https://doi.org/10.1038/s41564-019-0494-6
7. Camargo AP, Nayfach S, Chen I-MA *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res* 2023;51:D733–43. https://doi.org/10.1093/nar/gkac1037
8. Grigson SR, Giles SK, Edwards RA *et al.* Knowing and naming: phage annotation and nomenclature for phage therapy. *Clin Infect Dis* 2023;77:S352–9. https://doi.org/10.1093/cid/ciad539
9. Pirnay J-P, Djebara S, Steurs G *et al.* Personalized bacteriophage therapy outcomes for 100 consecutive cases: a multicentre, multinational, retrospective observational study. *Nat Microbiol* 2024;9:1434–53. https://doi.org/10.1038/s41564-024-01705-x
10. Cook R, Brown N, Redgwell T *et al.* INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. *PHAGE* 2021;2:214–23. https://doi.org/10.1089/phage.2021.0007
11. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8. https://doi.org/10.1038/nbt.3988
12. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60. https://doi.org/10.1038/nmeth.3176
13. Remmert M, Biegert A, Hauser A *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;9:173–5. https://doi.org/10.1038/nmeth.1818
14. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;39:W29–37. https://doi.org/10.1093/nar/gkr367
15. Larralde M, Zeller G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* 2023;39:btad214. https://doi.org/10.1093/bioinformatics/btad214
16. Bouras G, Nepal R, Houtak G *et al.* Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* 2023;39:btac776. https://doi.org/10.1093/bioinformatics/btac776
17. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;8:90. https://doi.org/10.1186/s40168-020-00867-0
18. Camargo AP, Roux S, Schulz F *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024;42:1303–12. https://doi.org/10.1038/s41587-023-01953-y

19. Figueroa JL, III, Dhungel E, Bellanger M *et al.* MetaCerberus: distributed highly parallelized HMM-based processing for robust functional annotation across the tree of life. *Bioinformatics* 2024;40:btae119. https://doi.org/10.1093/bioinformatics/btae119
20. Shaffer M, Borton MA, McGivern BB *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* 2020;48:8883–900. https://doi.org/10.1093/nar/gkaa621
21. The UniProt Consortium. UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–31. https://doi.org/10.1093/nar/gkac1052
22. Terzian P, Olo Ndela E, Galiez C *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* 2021;3:lqab067. https://doi.org/10.1093/nargab/lqab067
23. Trgovec-Greif L, Hellinger H-J, Mainguy J *et al.* VOGDB—database of virus orthologous groups. *Viruses* 2024;16:1191. https://doi.org/10.3390/v16081191
24. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* 2009;77:499–508. https://doi.org/10.1002/prot.22458
25. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2
26. van Kempen M, Kim SS, Tumescheit C *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024;42:243–6. https://doi.org/10.1038/s41587-023-01773-0
27. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;12:85–94. https://doi.org/10.1093/protein/12.2.85
28. Say H, Joris B, Giguere D *et al.* Annotating metagenomically assembled bacteriophage from a unique ecological system using protein structure prediction and structure homology search. bioRxiv, https://doi.org/10.1101/2023.04.19.537516, 21 April 2023, preprint: not peer reviewed.
29. Kim RS, Levy Karin E, Mirdita M *et al.* BFVD—a large repository of predicted viral protein structures. *Nucleic Acids Res* 2025;53:D340–7. https://doi.org/10.1093/nar/gkae1119
30. Mirdita M, Schütze K, Moriwaki Y *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 2022;19:679–82. https://doi.org/10.1038/s41592-022-01488-1
31. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;379:1123–30. https://doi.org/10.1126/science.ade2574
32. Odai R, Leemann M, Al-Murad T *et al.* The Viral AlphaFold Database of monomers and homodimers reveals conserved protein folds in viruses of bacteria, archaea, and eukaryotes. *Sci Adv* 2025;11:eadz8560. https://doi.org/10.1126/sciadv.adz8560
33. Nomburg J, Doherty EE, Price N *et al.* Birth of protein folds and functions in the virome. *Nature* 2024;633:710–7. https://doi.org/10.1038/s41586-024-07809-y
34. Litvin U, Lytras S, Jack A *et al.* Viro3D: a comprehensive database of virus protein structure predictions. *Mol Syst Biol* 2025;21:1599–617. https://doi.org/10.1038/s44320-025-00147-9
35. Heinzinger M, Littmann M, Sillitoe I *et al.* Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom Bioinform* 2022;4:lqac043. https://doi.org/10.1093/nargab/lqac043
36. Jha N, Kravitz J, West-Roberts J *et al.* Gaia: an AI-enabled genomic context–aware platform for protein sequence annotation. *Sci Adv* 2025;11:eadv5109. https://doi.org/10.1126/sciadv.adv5109
37. Grigson SR, Bouras G, Papudeshi B *et al.* Synteny-aware functional annotation of bacteriophage genomes with Phynteny. bioRxiv, https://doi.org/10.1101/2025.07.28.667340, 29 July 2025, preprint: not peer reviewed.
38. Cantu VA, Salamon P, Seguritan V *et al.* PhANNs, a fast and accurate tool and web server to classify phage structural proteins. *PLoS Comput Biol* 2020;16:e1007845. https://doi.org/10.1371/journal.pcbi.1007845
39. Flamholz ZN, Biller SJ, Kelly L. Large language models improve annotation of prokaryotic viral proteins. *Nat Microbiol* 2024;9:537–49. https://doi.org/10.1038/s41564-023-01584-8
40. Boulay A, Leprince A, Enault F *et al.* Empathi: embedding-based phage protein annotation tool by hierarchical assignment. *Nat Commun* 2025;16:9114. https://doi.org/10.1038/s41467-025-64177-5
41. Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;44:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381
42. Zhang Z, Wayment-Steele HK, Brixi G *et al.* Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc Natl Acad Sci* 2024;121:e2406285121. https://doi.org/10.1073/pnas.2406285121
43. Simon E, Zou J. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders. 2024.
44. Zayed AA, Lücking D, Mohssen M *et al.* efam: an expanded, metaproteome-supported HMM profile database of viral protein families. *Bioinformatics* 2021;37:4202–8. https://doi.org/10.1093/bioinformatics/btab451
45. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 2010;26:889–95. https://doi.org/10.1093/bioinformatics/btq066
46. Monzon V, Haft DH, Bateman A. Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinform Adv* 2022;2:vbab043. https://doi.org/10.1093/bioadv/vbab043
47. Pérez-Bucio R, Enault F, GC.: an extended view of the viral protein families on Earth through a vast collection of HMM profiles. *Peer Community J* 2025;5. https://doi.org/10.24072/pcjournal.627
48. Steinegger M, Meier M, Mirdita M *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;20:473. https://doi.org/10.1186/s12859-019-3019-7
49. Sahakyan H, Makarova KS, Koonin EV. Search for origins of anti-CRISPR proteins by structure comparison. *CRISPR J* 2023;6:222–31 https://doi.org/10.1089/crispr.2023.0011
50. Alcock BP, Huynh W, Chalil R *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 2023;51:D690–9. https://doi.org/10.1093/nar/gkac920
51. Varadi M, Bertoni D, Magana P *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;52:D368–75. https://doi.org/10.1093/nar/gkad1011
52. Tesson F, Hervé A, Mordret E *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* 2022;13:2561. https://doi.org/10.1038/s41467-022-30269-9
53. Roux S, Paul BG, Bagby SC *et al.* Ecology and molecular targets of hypermutation in the global microbiome. *Nat Commun* 2021;12:3076. https://doi.org/10.1038/s41467-021-23402-7
54. Ernits K, Saha CK, Brodiazhenko T *et al.* The structural basis of hyperpromiscuity in a core combinatorial network of type II toxin–antitoxin and related phage defense systems. *Proc Natl Acad Sci* 2023;120:e2305393120. https://doi.org/10.1073/pnas.2305393120
55. Liu B, Zheng D, Zhou S *et al.* VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* 2022;50:D912–7. https://doi.org/10.1093/nar/gkab1107
56. Sunagawa S, Acinas SG, Bork P *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Micro* 2020;18:428–45. https://doi.org/10.1038/s41579-020-0364-5
57. Hyatt D, Chen G-L, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119. https://doi.org/10.1186/1471-2105-11-119
58. Larralde M. Pyrodigal: python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *J Open Source Softw* 2022;7:4296. https://doi.org/10.21105/joss.04296

59. Mallawaarachchi V, Roach MJ, Decewicz P *et al.* Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics* 2023;39:btad586. https://doi.org/10.1093/bioinformatics/btad586
60. Li D, Liu C-M, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–6. https://doi.org/10.1093/bioinformatics/btv033
61. Nayfach S, Camargo AP, Schulz F *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;39:578–85. https://doi.org/10.1038/s41587-020-00774-7
62. Nurk S, Meleshko D, Korobeynikov A *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;27:824–34. https://doi.org/10.1101/gr.213959.116
63. Hu EJ, Shen Y, Wallis P *et al.* LoRA: Low-Rank Adaptation of Large Language Models. 2021. https://doi.org/10.48550/arXiv.2106.09685
64. Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat Commun* 2024;15:7407. https://doi.org/10.1038/s41467-024-51844-2
65. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57:702–10. https://doi.org/10.1002/prot.20264
66. Mariani V, Biasini M, Barbato A *et al.* lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 2013;29:2722–8. https://doi.org/10.1093/bioinformatics/btt473
67. Haghani M, Bhattacharya D, Murali TM. NEFFy: a versatile tool for computing the number of effective sequences. *Bioinformatics* 2025;btaf222. https://doi.org/10.1093/bioinformatics/btaf222
68. Varadi M, Anyango S, Deshpande M *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;50:D439–44. https://doi.org/10.1093/nar/gkab1061
69. Barrio-Hernandez I, Yeo J, Jänes J *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* 2023;622:637–45. https://doi.org/10.1038/s41586-023-06510-w
70. McKerral JC, Papudeshi B, Inglis LK *et al.* The promise and pitfalls of prophages. bioRxiv, https://doi.org/10.1101/2023.04.20.537752, 21 April 2023, preprint: not peer reviewed.
71. Heinzinger M, Weissenow K, Sanchez JG *et al.* Bilingual language model for protein sequence and structure. *NAR Genom Bioinform* 2024;6:lqae150. https://doi.org/10.1093/nargab/lqae150
72. Schwengers O, Jelonek L, Dieckmann MA *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* 2021;7:000685. https://doi.org/10.1099/mgen.0.000685
73. Cook R, Telatin A, Bouras G *et al.* Driving through stop signs: predicting stop codon reassignment improves functional annotation of bacteriophages. *ISME Commun* 2024;4:ycae079. https://doi.org/10.1093/ismeco/ycae079
74. Kallenborn F, Chacon A, Hundt C *et al.* GPU-accelerated homology search with MMseqs2. *Nat Methods* 2025;22:2024–7. https://doi.org/10.1038/s41592-025-02819-8
75. Sánchez-Osuna M, Cortés P, Llagostera M *et al.* Exploration into the origins and mobilization of di-hydrofolate reductase genes and the emergence of clinical resistance to trimethoprim. *Microb Genom* 2020;6:mgen000440. https://doi.org/10.1099/mgen.0.000440
76. Peters DL, McCutcheon JG, Stothard P *et al.* Novel Stenotrophomonas maltophilia temperate phage DLP4 is capable of lysogenic conversion. *BMC Genomics* 2019;20:300. https://doi.org/10.1186/s12864-019-5674-5
77. Tal N, Morehouse BR, Millman A *et al.* Cyclic CMP and cyclic UMP mediate bacterial immunity against phages. *Cell* 2021;184:5728–39. https://doi.org/10.1016/j.cell.2021.09.031
78. Doron S, Melamed S, Ofir G *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 2018;359:eaar4120. https://doi.org/10.1126/science.aar4120
79. Mahata T, Kanarek K, Goren MG *et al.* Gamma-Mobile-trio systems are mobile elements rich in bacterial defensive and offensive tools. *Nat Microbiol* 2024;9:3268–83. https://doi.org/10.1038/s41564-024-01840-5
80. Rousset F, Depardieu F, Miele S *et al.* Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe* 2022;30:740–53. https://doi.org/10.1016/j.chom.2022.02.018
81. Gao L, Altae-Tran H, Böhning F *et al.* Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* 2020;369:1077–84. https://doi.org/10.1126/science.aba0372
82. Patel PH, Maxwell KL. Prophages provide a rich source of antiphage defense systems. *Curr Opin Microbiol* 2023;73:102321. https://doi.org/10.1016/j.mib.2023.102321
83. Papudeshi B, Roach MJ, Mallawaarachchi V *et al.* Sphae: an automated toolkit for predicting phage therapy candidates from sequencing data. *Bioinform Adv* 2025;5:vbaf004. https://doi.org/10.1093/bioadv/vbaf004
84. Murtazalieva K, Karatzas E, Corona F *et al.* Elucidating the mechanisms of action and evolutionary history of phage anti-defence proteins. bioRxiv, https://doi.org/10.1101/2025.06.06.658234, 6 June 2025, preprint: not peer reviewed.
85. Ingles-Prieto A, Ibarra-Molero B, Delgado-Delgado A *et al.* Conservation of protein structure over four billion years. *Structure* 2013;21:1690–7. https://doi.org/10.1016/j.str.2013.06.020
86. Bernheim A, Cury J, Poirier EZ. The immune modules conserved across the tree of life: towards a definition of ancestral immunity. *PLoS Biol* 2024;22:e3002717. https://doi.org/10.1371/journal.pbio.3002717
87. Lescot M, Hingamp P, Kojima KK *et al.* Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *ISME J* 2016;10:1134–46. https://doi.org/10.1038/ismej.2015.192
88. Smyshlyaev G, Bateman A, Barabas O. Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. *Mol Syst Biol* 2021;17:e9880. https://doi.org/10.15252/msb.20209880
89. Gilchrist CLM, Mirdita M, Steinegger M. Multiple protein structure alignment at scale with FoldMason. bioRxiv, https://doi.org/10.1101/2024.08.01.606130, 1 August 2024, preprint: not peer reviewed
90. Das S, Shahnaz N, Keerthana C *et al.* Functional and comparative analysis of the FeII/2-oxoglutarate-dependent dioxygenases without using any substrate. *Biol Methods Protoc* 2025;10:bpae096. https://doi.org/10.1093/biomethods/bpae096
91. Martinez S, Hausinger RP. Catalytic mechanisms of Fe(II)- and 2-oxoglutarate-dependent oxygenases. *J Biol Chem* 2015;290:20702–11. https://doi.org/10.1074/jbc.R115.648691
92. Brüssow H. The not so universal tree of life or the place of viruses in the living world. *Philos Trans R Soc Lond B Biol Sci* 2009;364:2263–74. https://doi.org/10.1098/rstb.2009.0036
93. Yeo J, Han Y, Bordin N *et al.* Metagenomic-scale analysis of the predicted protein structure universe. bioRxiv, https://doi.org/10.1101/2025.04.23.650224, 26 April 2025, preprint: not peer reviewed.
94. Pavlopoulos GA, Baltoumas FA, Liu S *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* 2023;622:594–602. https://doi.org/10.1038/s41586-023-06583-7

Received: September 17, 2025. **Revised:** November 12, 2025. **Accepted:** November 13, 2025