Synnotate: Interpretable Genomic-Context-Based Annotation of the Prokaryotic and Viral Dark Proteome

Khoa Hoang*
Department of Bioengineering, Stanford University,
Stanford, CA, USA
E-mail: khoang99@stanford.edu

Swan Van
Department of Biochemistry and Molecular Biophysics, University of California San Diego
San Diego, CA, USA
E-mail: nvan@ucsd.edu

Janice Subroto
Computational and Systems Biology, University of California Los Angeles
Los Angeles, CA, USA
E-mail: jsubroto@g.ucla.edu

Nguyen Nguyen
Faculty of Biology and Biotechnology, VNUHCM University of Science
Ho Chi Minh City, Vietnam
E-mail: 23C64016@student.hcmus.edu.vn

Michael Snyder
Department of Genetics, Stanford University,
Stanford, CA, USA
E-mail: mpsnyder@stanford.edu

*Research reported in this publication was supported by the Human Virome Program, a program of the National Institutes of Health under award number U54AG089334. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Advances in genome sequencing have dramatically expanded the catalog of microbial and viral proteins, yet millions remain unannotated. Functional annotation relies predominantly on homology-based transfer, which performs poorly for proteins lacking informative sequence or structural similarity. Conserved genomic neighborhoods provide complementary evidence for functional inference by capturing evolutionarily conserved functional relationships among neighboring genes. Existing methods, however, either identify conserved neighborhoods without predicting protein function or rely on black-box machine learning models with limited interpretability. Here, we present Synnotate, an interpretable genome-context framework for microbial dark proteome annotation beyond sequence and structural homology. Synnotate represents genomic neighborhoods as standardized functional annotation tokens and infers the function of a target gene from its neighbors, along with calibrated confidence, exact neighbor-level attribution, and retrieved reference neighborhoods. Applied at scale, Synnotate assigns high-confidence functional predictions to 4.9% of previously unannotated genes in bacteriophage genomes and 2.0% in prokaryotic genomes, extending functional annotation and genome mining beyond the limits of sequence and structural similarity alone.

Keywords: Functional annotation; genome context; dark proteome; microbes; bacteriophages; machine learning; interpretability.

1. Introduction

The microbial and viral dark proteome represents one of the largest unresolved challenges in functional genomics, with an estimated 20–30% of bacterial proteins and nearly 65% of bacteriophage proteins remaining functionally uncharacterized.[1,2] This persistent annotation gap constrains our ability to interpret microbial and viral genomes at scale. Current annotation pipelines rely predominantly on sequence- and structure-based search and clustering for functional annotation,[3,17] with emerging protein language model (pLM)-based approaches improving the detection of remote homologs.[4] Because these strategies rely on previously characterized proteins in reference databases, they perform poorly for lineage-restricted, rapidly evolving, and highly divergent proteins.

Genomic context provides functional information complementary to sequence and structural similarity. As functionally related genes are frequently co-localized within conserved genomic neighborhoods, the genes surrounding an uncharacterized protein can provide evidence for its potential function even in the absence of detectable homologs. Tools such as MultiGeneBlast,[5] cblaster,[6] and antiSMASH[7] identify homologous gene clusters and can support post hoc functional inference from matched loci. However, they are designed primarily for cluster discovery rather than direct protein annotation, and their large-scale homology searches can be costly for genome-wide use. Deep learning methods such as Gaia[8] and Phynteny[9] have incorporated genomic context into scalable protein-function prediction. Because context is combined with pLM representations or other sequence-derived features, however, its independent contribution to individual predictions is difficult to disentangle.

Here, we present Synnotate, a scalable and interpretable framework for prokaryotic and viral protein annotation using genomic context. Synnotate takes as input a fixed-size window of annotated neighboring genes represented as standardized functional annotation tokens and

predicts the function of the central gene using a Transformer-based neural network (Figure 1). Synnotate uses genomic context as its primary predictive input and returns calibrated confidence, neighbor-level attribution, and retrieved reference neighborhoods for each prediction. This enables scalable, evidence-backed annotation of proteins lacking detectable sequence or structural homologs.

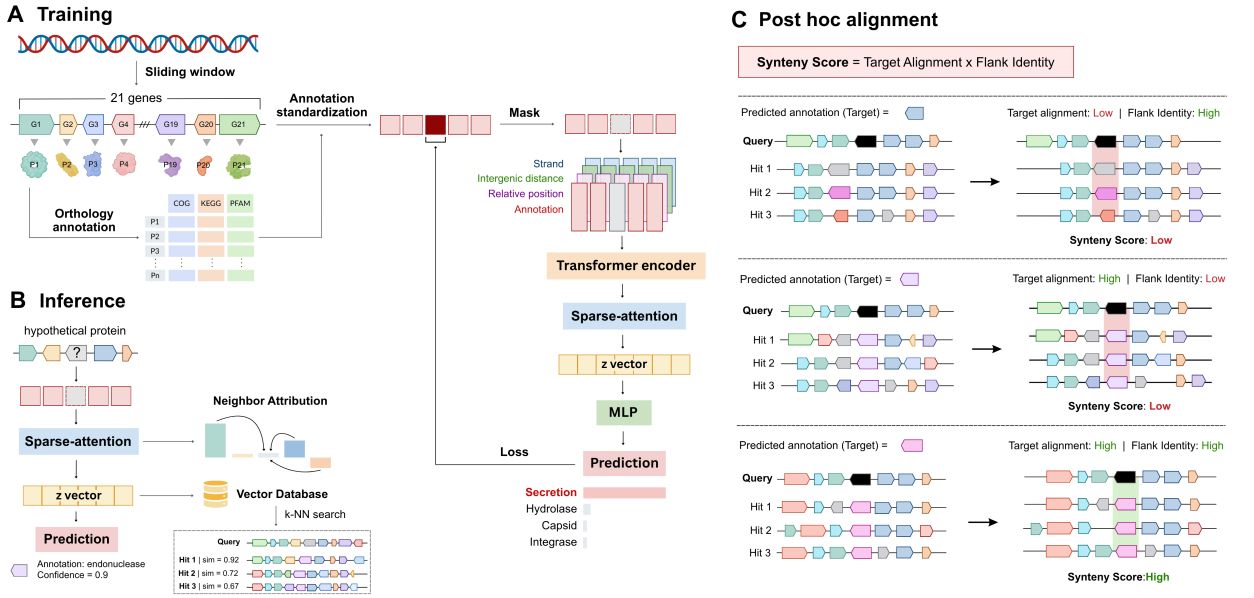


Figure 1. Overview of the Synnotate training, inference, and interpretation workflow. (A) Synnotate training schema. (B) Synnotate inference and interpretation schema. (C) Post hoc annotation alignment for synteny scoring.

2. Methods

2.1. *Training datasets*

2.1.1. *Prokaryotic dataset*

Bacterial and archaeal genomes were obtained from the Genome Taxonomy Database (GTDB)[10] release 232. GTDB species-representative genomes with NCBI RefSeq annotations were retained after quality filtering ($\geq$95% completeness, $\leq$5% contamination, and $\leq$500 contigs), excluding metagenome-assembled genomes, single-amplified genomes, and environmental assemblies. Protein-coding genes were extracted from RefSeq annotations, and GTDB taxonomy was used for phylogenetically informed dataset partitioning. The final dataset consisted of 33,496 genomes spanning 4,900 GTDB genera and approximately 143 million proteins.

2.1.2. *Bacteriophage dataset*

16,479 complete bacteriophage genomes were obtained from the INfrastructure for a PHAge REference Database (INPHARED)[11] (4 January 2026 release). Protein-coding genes were annotated using curated PHROG assignments[12] provided in the database, and phage taxonomy was obtained from the INPHARED metadata table.

2.2. *Standardized annotation vocabulary construction*

A standardized annotation vocabulary was generated by integrating orthology assignments, protein product annotations, and homology-based label propagation. Prokaryotic proteins were annotated with eggNOG-mapper (v2.1.13)[13] using DIAMOND[14] in sensitive mode. For each protein, the KO identifier[15] assigned to its eggNOG-mapper orthologous group was used when available. Otherwise, the group's first annotated Pfam domain was used.[16] Proteins lacking both assignments were retained without an initial orthology-derived group. Phage proteins were assigned to PHROG families.[12] Annotation tokens were defined by group identifiers and matched with the most frequent informative RefSeq product annotation. Genes lacking a group identifier but possessing an informative RefSeq product annotation were then grouped by their product annotation. Remaining unlabeled genes were annotated by homology-based label propagation. Labeled and unlabeled sequences were clustered together in a single MMseqs2 Linclust[17] run (30% amino-acid identity, 80% coverage). For an unlabeled gene, the cluster's majority label was transferred only when at least two labeled members were present and the majority label accounted for at least 80% of the labeled members in the cluster. Otherwise, the gene remained unlabeled and assigned a *DARK* token. Rare annotations (<50 occurrences) were collapsed into an *OTHER* token. Summary statistics of the training datasets and the proportions of *DARK* and *OTHER* tokens are provided in Table S1.

2.3. *Synnotate model architecture and training*

Synnotate was trained as a masked-token prediction model over genomic neighborhoods. For each target gene, a symmetric window of $\pm w$ neighboring genes was extracted from the same contig, yielding a fixed input of $2w{+}1$ positions ($w = 10$ by default). Windows extending beyond contig boundaries were padded with a reserved *CONTIG-EDGE* token. Each position was represented by the sum of 384-dimensional embeddings from four tracks: annotation, coding strand, intergenic distance, and position within the window (0 to $2w$). During training and inference, only the target gene's annotation token was replaced by the reserved *MASK* token; its strand, intergenic distance, and position features were retained. The input sequence was encoded by a two-layer Transformer (embedding dimension 384, four attention heads, feed-forward dimension 512, GELU activation, dropout 0.1). Encoded neighbor representations were aggregated by a target-conditioned sparse-attention pooling layer, in which a query derived from the masked target attended over neighboring genes (excluding the target itself) using sparsemax.[18] The pooled representation was passed to a linear classifier to predict the target functional token. The sparsemax-weighted linear classifier permits exact decomposition of each prediction into per-neighbor contributions (Fig. 1A, B).

Models were trained using the AdamW optimizer, cross-entropy loss, and a batch size of 2,048 for up to 40 epochs. Early stopping (patience = 5) was based on validation top-1 accuracy, with the best-performing checkpoint retained. To mitigate class imbalance, each annotation token contributed at most 3,000 training examples. Model performance was evaluated using a genus-disjoint five-fold cross-validation protocol over 4,900 prokaryotic and 1,397 phage genera. In each fold, one genus-disjoint partition was held out for testing, and the remaining genera were split into disjoint training and validation sets. Each annotated gene was

evaluated once in an outer test fold using a model trained and validated only on other genera. The deployment model used for genome-wide annotation was trained on the full dataset using an analogous genus-disjoint validation set (∼2%) for early stopping and confidence calibration.

2.4. *Vector database construction and retrieval*

To provide reference examples for each prediction, Synnotate indexes annotated genomic neighborhoods in the learned context-embedding space (*z-vector*, Fig. 1B). Specifically, each neighborhood in the training set was represented by the pooled embedding produced by the trained deployment model immediately before the final linear classifier. Embeddings were L2-normalized and indexed using FAISS[19] together with the corresponding genome metadata. For each query gene, its genomic neighborhood was embedded using the same model and used to query the reference vector database to retrieve the k nearest neighborhoods (default $k = 5$) from distinct genomes, excluding the genome of the query itself. Retrieved neighborhoods were used for interpretation and post hoc synteny scoring.

2.5. *Post-hoc annotation-token alignment and synteny score*

Retrieved neighborhoods were aligned to the query neighborhood using Needleman–Wunsch alignment[20] over their annotation token sequences (Fig. 1C). Identical non-special tokens scored +2, while mismatches and gaps scored −1. *MASK* and *CONTIG-EDGE* tokens contributed 0. To account for inverted syntenic blocks, each neighborhood was aligned in both forward and reverse orientations, retaining the higher-scoring alignment. After alignment, flank identity was calculated as the mean fraction of matching flanking tokens, and target alignment as the fraction of retrieved neighborhoods whose aligned central token matched the prediction. The overall synteny score was defined as

$$\text{Synteny score} = \text{Flank identity} \times \text{Target alignment}. \tag{1}$$

2.6. *Structural search and clustering*

Protein structural similarity searches were performed using Foldseek with ProstT5-derived 3Di representations.[21] Proteins were searched against AFDB-SwissProt for prokaryotes and BFVD[22] for bacteriophages (E-value $\leq 10^{-3}$, maximum 10 hits), retaining the lowest-E-value hit and its target annotation. Structural clusters were generated using Foldseek *cluster* (greedy set-cover, 80% coverage).

2.7. *Structure prediction and alignment*

Representative proteins from structural clusters and selected case-study proteins were modeled using localColabFold (ColabFold[23] v1.6.1) with the AlphaFold2-ptm model (three models, three recycles, no structural templates or Amber relaxation). Multiple sequence alignments were generated through the ColabFold public MMseqs2 API using the *mmseqs2_uniref_env* database. Models were ranked by predicted local distance difference test (pLDDT), and the top model was summarized by mean pLDDT and predicted TM-score (pTM). Pairwise structural

comparisons were performed with TM-align.[24] Structural similarity was evaluated using the TM-score normalized to the reference chain length, root mean square deviation (RMSD), aligned length, and sequence identity over the aligned region.

2.8. *Reference phylogenetic trees*

Reference bacterial phylogenies were obtained by pruning the GTDB bac120 reference tree (release 232) to the taxa of interest using DendroPy.[25] Reference phage phylogenies were inferred from whole-proteome similarity. The most complete genome was selected from each INPHARED genus. Mash[26] amino acid sketches (size 1,000) were used to compute all-versus-all distances, from which a neighbor-joining tree was constructed with DendroPy; each tip represents one phage genus.

2.9. *SeqHub annotation*

SeqHub[8] annotation was generated using Protein Context Search APIs, with the target protein and ±2 neighboring genes as input. Searches were run at all four *diversityFilter* settings. The target annotation was assigned by majority vote across retrieved target-protein annotations.

2.10. *cblaster annotation*

cblaster[6] was used to identify homologous gene clusters through its remote search API. For each query, a 21-gene neighborhood centered on the target protein was searched against the NCBI non-redundant database with `-hs=5`, and annotations of matched target proteins were retrieved through the NCBI Entrez API for comparison with Synnotate predictions.

2.11. *Phynteny benchmark*

Synnotate was benchmarked against three Phynteny variants, each hiding a target gene's annotation while keeping all other annotated genes in the genome visible: (1) the released model (`phynteny_transformer` v0.1.3) applied to the same held-out test folds; (2) a per-fold model retrained on the same genus-held-out training genomes using the published recipe for 100 epochs; and (3) a context-only model directly comparable to Synnotate, trained and evaluated as in (2) but with the `--zero_idx` option, which zeroes the masked target's embedding so that no target-sequence information reaches the model. Both retrained variants were calibrated on their fold's genus-disjoint validation fold. As a sequence-only control, a small per-fold multilayer perceptron was trained on mean `esm2_t33_650M_UR50D` embeddings,[35] the same representation Phynteny uses, and calibrated on the same validation folds. To compare in Phynteny's nine-category label space, each Synnotate annotation was mapped to its PHROG v4 category, annotation probabilities were summed within each category, and the highest-scoring category was taken as the prediction.

3. Results

3.1. *Synnotate enables confident and interpretable protein function prediction from genomic context*

To assess whether genomic context alone contains sufficient information for protein function prediction, we masked annotated genes and inferred their functional tokens solely from the surrounding genomic neighborhood. Performance was evaluated by genus-disjoint five-fold cross-validation over 4,900 prokaryotic and 1,397 bacteriophage genera and compared with two context-only baselines: (1) a *Logistic Regression* Baseline that omits gene order, strand, and intergenic distance, and (2) a *Neural Baseline* that replaces the Transformer encoder and sparse-attention pooling with mean pooling. Synnotate consistently outperformed both baselines across all levels of neighborhood annotation, achieving top-1 accuracies of 50.7% in prokaryotes and 83.2% in bacteriophages (Fig. 2A). Prediction accuracy increased with neighborhood annotation coverage and remained higher in bacteriophages within matched coverage bins, consistent with bacteriophages' compact, modular genome organization. We benchmarked Synnotate against Phynteny[9] on predicting the nine PHROG functional categories in the same held-out test sets (Fig. 2D, Section 2.11). The sequence-only ESM-2 model achieved consistently high top-1 accuracy at all levels of neighborhood annotation. The released and retrained Phynteny models performed comparably to each other and below ESM-2, suggesting that for Phynteny, genomic context adds noise when there is strong target sequence information. Because Synnotate uses only genomic context, we also evaluated Phynteny with the target's sequence embedding zeroed. This context-only variant performed substantially worse than Synnotate ($\sim$27 pp), likely due to differences in tokenization and Synnotate's train–test consistency. Balanced accuracy and macro-F1 are reported in Fig. S3.

We next examined whether Synnotate's confidence scores reflect prediction reliability. Higher softmax probabilities generally corresponded to greater accuracy, but the raw scores were imperfectly calibrated, so we applied isotonic regression on the validation folds to map them to observed accuracy (Fig. 2B). A few high-confidence errors persisted, reflecting a failure mode we term *context leakage* (Fig. S1): a gene at a shifted or inserted position receives the function typically found at a nearby conserved position. To filter these cases, we added a synteny score based on annotation-token alignment between the query and retrieved reference neighborhoods (Fig. 1C, Section 2.5). Together with calibrated confidence, this score defined trusted prediction regions that achieved the target accuracy on held-out test genes (Fig. 2C) and were applied across all genomes for dark-proteome annotation.

Finally, we examined the genomic features underlying Synnotate's predictions. Accuracy increased with neighborhood size up to $\sim \pm 5$ genes before plateauing (Fig. S2), suggesting that most functional information lies within compact neighborhoods, consistent with typical operon and phage module sizes.[27,28] Feature ablation showed that neighbor annotations contributed most of the predictive signal, while genomic organization (relative position, coding strand, and intergenic distance) provided smaller, complementary gains (Fig. 2E). Organization alone achieved 34.2% accuracy on bacteriophages, indicating that gene arrangement itself carries substantial functional information. Neighbor-level attribution showed that Synnotate relies mainly on immediately adjacent genes, with attribution decreasing monotonically with

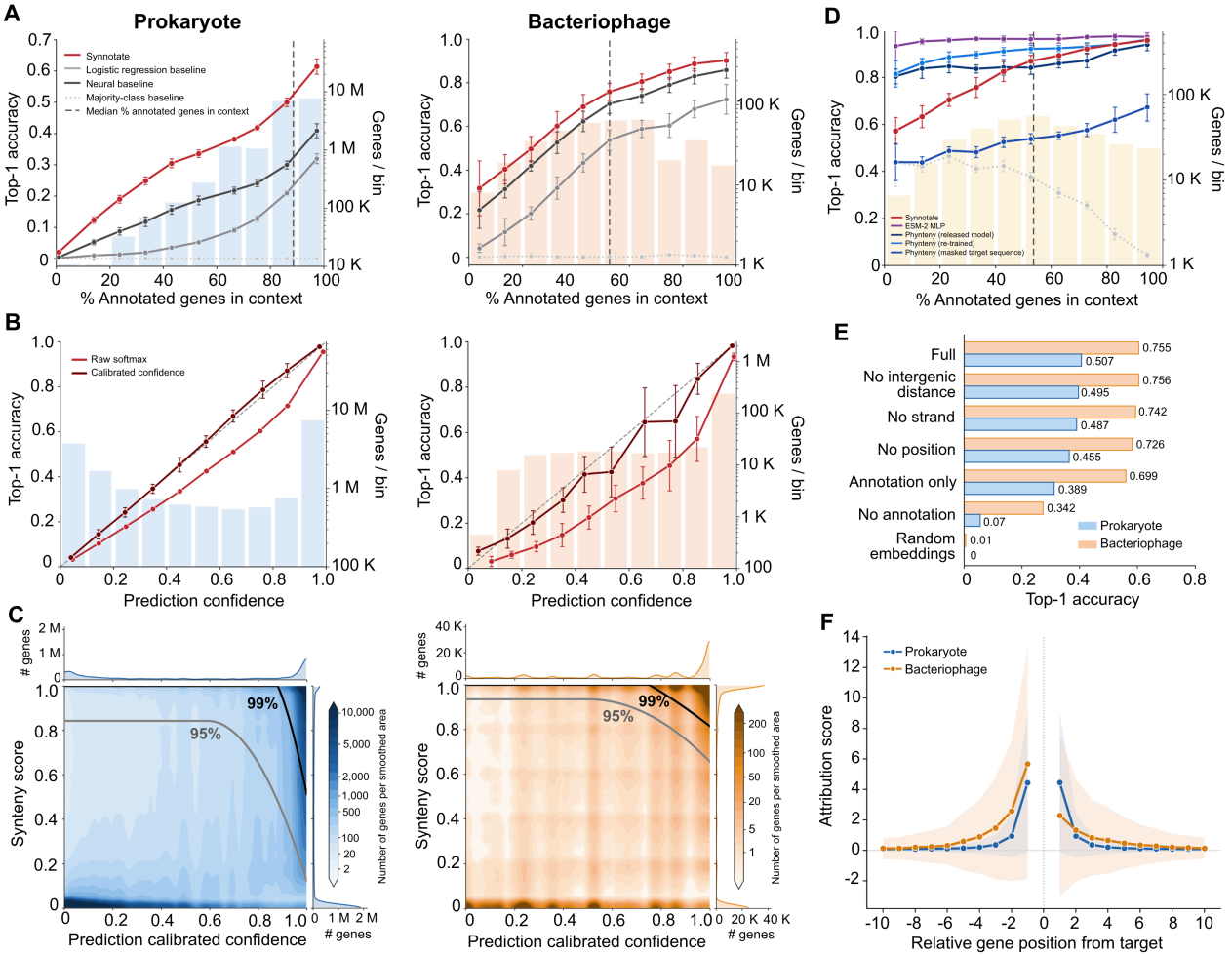


Figure 2. Evaluation of Synnotate on held-out annotated genes. (A) Top-1 accuracy on prokaryote and bacteriophage datasets, stratified by the percentage of annotated genes in the 21-gene neighborhood. (B) Observed top-1 accuracy versus raw softmax probability and isotonic-calibrated confidence. (C) Calibrated confidence versus synteny score for test genes pooled across five folds; grey and black lines mark thresholds for 95% and 99% expected accuracy. (D) Top-1 accuracy on bacteriophages of Synnotate, an ESM-2 embedding MLP, and three Phynteny variants (released, retrained, and retrained with the target sequence masked), stratified as in (A). (E) Feature ablation showing the effect of removing each feature on performance. (F) Neighbor attribution scores across five test folds. Error bars (A, B, D) and shading (F) denote ±1 SD across folds.

distance (Fig. 2F), and in bacteriophages was biased toward upstream neighbors, consistent with directional operonic organization.

3.2. *Confident annotation beyond sequence- and structure-based search: a Desulfovibrio piger case study*

We next tested whether Synnotate could annotate proteins left unresolved by sequence- and structure-based pipelines using *Desulfovibrio piger*. Of 2,430 protein-coding genes, 2,105 received functional annotations from RefSeq and eggNOG-mapper (Fig. 3A), while Foldseek

structural searches provided additional annotations for 45 proteins. Among the remaining 280 proteins lacking sequence- or structure-based annotations, Synnotate confidently resolved five unannotated proteins (Fig. 3B).

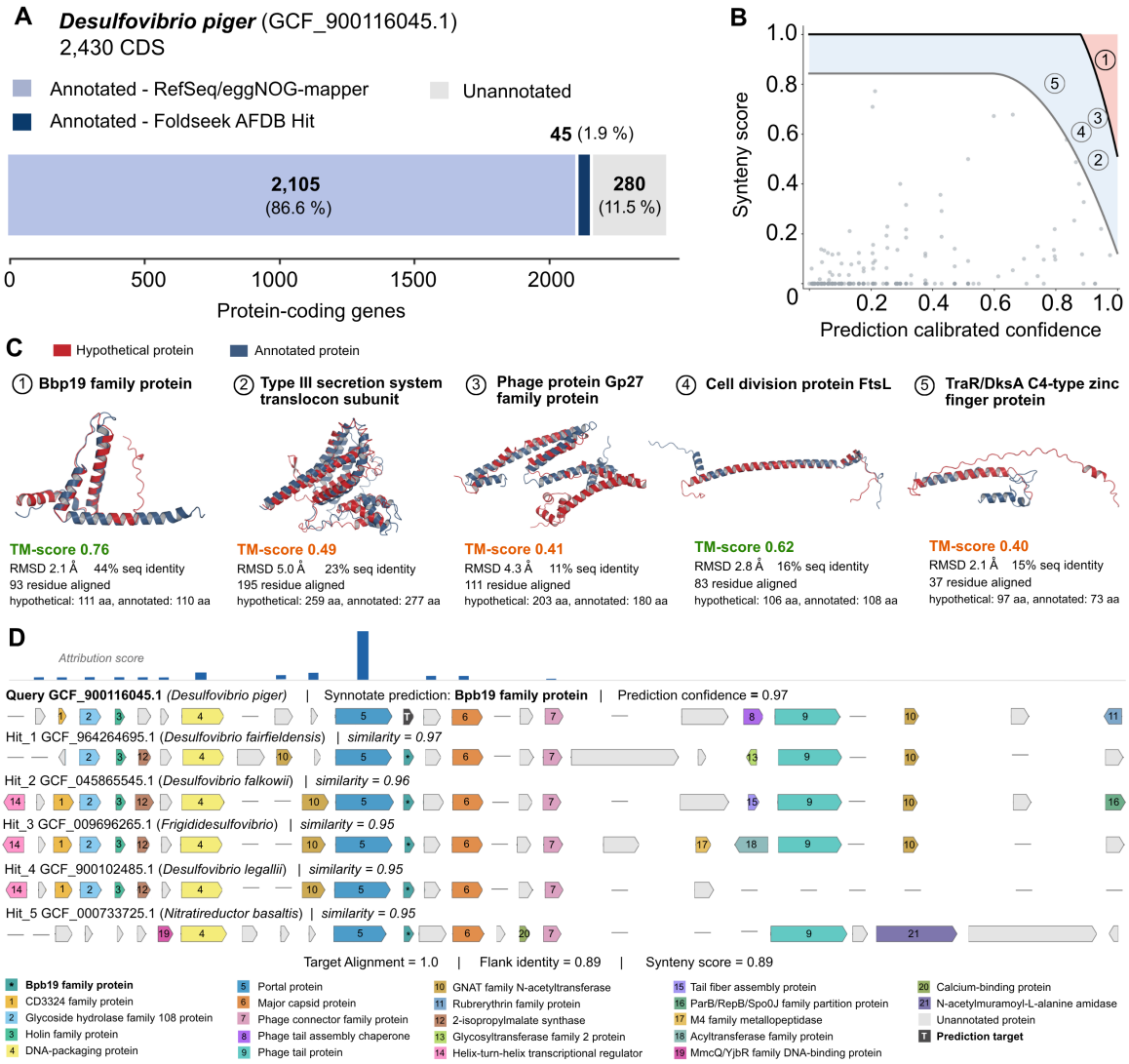


Figure 3. Case study of Synnotate annotation of previously unannotated genes in *Desulfovibrio piger*. (A) Breakdown of protein-coding genes by annotation tier: (i) genes annotated by RefSeq/eggNOG-mapper; (ii) genes unannotated in (i) but annotated via Foldseek against AFDB-SwissProt; (iii) genes remaining unannotated after (i) and (ii). (B) Synnotate prediction confidence and synteny scores for genes in tier (iii); five genes exceeding the 95% accuracy threshold are highlighted. (C) Structural alignments of the five Synnotate-predicted proteins to annotated proteins with the same predicted label from the nearest retrieved reference neighborhood. (D) Example interpretation output for the predicted Bbp19 gene, showing neighbor attribution scores, nearest-neighbor hits from the vector database, and post hoc annotation-token alignment.

The five predictions were (1) Bbp19 family protein, (2) type III secretion system translocon subunit SctB, (3) phage Gp27-family protein, (4) cell division protein FtsL, and (5)

TraR/DksA C4-type zinc-finger protein. For each prediction, the annotated protein from the nearest retrieved neighborhood carrying the predicted token was used as the structural reference. Two predictions (1, 4) showed strong structural agreement (TM-scores 0.76 and 0.62), whereas the remaining predictions showed conserved structural topologies and genomic neighborhoods (Fig. S4) despite low TM-scores (0.40–0.49) and sequence identity (5–23%), indicating that genomic context can reveal functional annotations for divergent proteins. Beyond prediction, Synnotate provides inspectable evidence through neighbor-level attribution and retrieval of conserved genomic neighborhoods. For example, annotation of the Bbp19 family protein was primarily supported by its immediately upstream portal protein and downstream major capsid protein, while retrieved neighborhoods independently confirmed conservation of the surrounding gene organization (Fig. 3D). We next compared Synnotate with existing context-aware annotation tools. SeqHub Protein Context Search[8] produced matching predictions for two of the five genes (SctB and FtsL), assigned a different function to the TraR/DksA C4-type zinc-finger protein (lipoprotein/spanin subunit), and returned uncharacterized predictions for the remaining two genes. cblaster[6] correctly recovered only SctB and failed to resolve the remaining targets (Sections 2.9 and 2.10). These results demonstrate that Synnotate can infer protein function from transferable signals encoded by neighboring genes within conserved genomic neighborhoods, enabling interpretable annotation of proteins unresolved by a variety of homology and deep learning methods.

3.3. *Synnotate recovers divergent prokaryotic dark proteins*

We then applied Synnotate to 17.6 million dark genes from 33,496 prokaryotic genomes. Using the trusted-region thresholds, Synnotate confidently annotated 108,991 genes (0.6%) at the 99% threshold and 359,060 genes (2.0%) at the 95% threshold, 17% of which returned a named hit via Foldseek search. Synnotate predictions were broadly distributed across the prokaryotic phylogeny, with a median of 8 genes annotated per genome (Fig. 4A, Fig. S5). Among the most frequently recovered functions was the inner-membrane protein YgfX, with 1,463 previously unannotated proteins expanding the known family by 30% across 475 genera, including 307 without previously annotated YgfX (Fig. 4B, Fig. 4E). Predicted YgfX proteins exhibited length distributions nearly identical to annotated YgfX proteins (Fig. 4C). A Foldseek structural similarity search against AFDB-SwissProt identified structural homologs for only 14.7% of the predicted proteins (Fig. 4D). However, when predicted and annotated proteins were structurally clustered together (Section 2.6), 97.7% of the predicted YgfX proteins clustered with previously annotated YgfX proteins, indicating that the apparent lack of structural homologs primarily reflected incomplete reference databases rather than novel protein folds. We next examined a subset of structurally unresolved YgfX proteins: a representative 96-amino-acid protein from *Aliivibrio* (confidence = 0.99, synteny support = 1.0) that lacked database matches and structural clustering, yet retained structural similarity to a context-retrieved YgfX reference (TM-score = 0.41) within a conserved genomic neighborhood (Fig. 4F–H). Together, these findings suggest that Synnotate can recover highly divergent members of conserved protein families that are missed by conventional homology-based annotation.

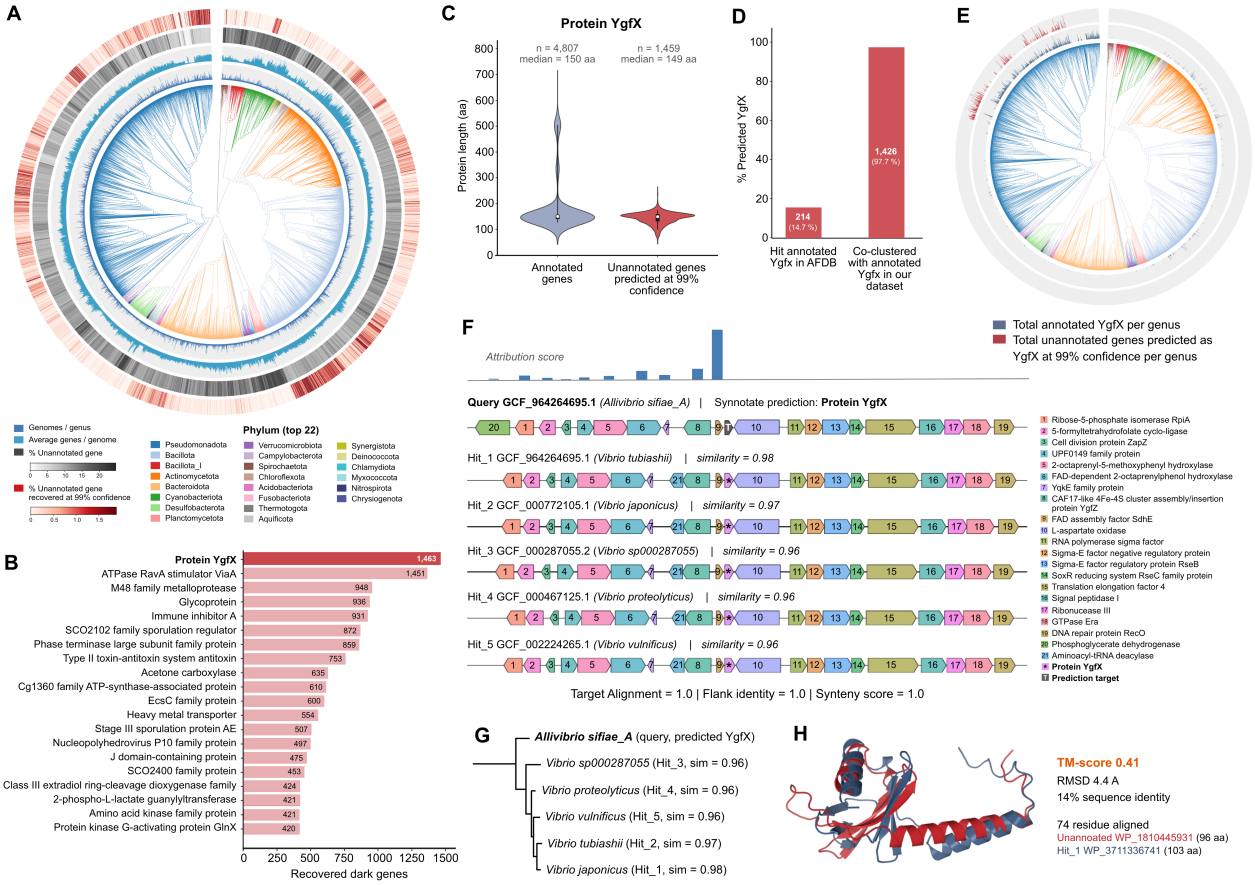


Figure 4. Synnotate annotation of the prokaryotic dark proteome. (A) Prokaryotic phylogeny showing dark-proteome genes and their Synnotate annotations, aggregated by genus. (B) Top 20 predicted annotations at 99% confidence. (C) Length distributions of annotated YgfX proteins and unannotated proteins predicted as YgfX. (D) Foldseek search of predicted YgfX proteins against AFDB and their clustering with annotated proteins from the prokaryotic dataset. (E) Distribution of annotated and predicted YgfX across the prokaryotic phylogeny. (F) Interpretation output (neighbor attributions, retrieved reference neighborhoods, and post hoc annotation-token alignment) for a predicted YgfX that neither co-clusters with nor matches annotated YgfX in AFDB. (G) Phylogeny of bacteria containing "dark" YgfX and its k nearest-neighbor hits from (F). (H) Structural alignment of a dark/predicted YgfX with the annotated YgfX from its nearest retrieved neighborhood.

3.4. *Synnotate expands annotation of the bacteriophage dark proteome*

Applying Synnotate to 873,701 unannotated genes from 16,479 bacteriophage genomes, we confidently annotated 3.7% at the 99% threshold and 4.9% of dark proteins at the 95% threshold, 22% of which returned a named hit via Foldseek search. Homing endonuclease-like proteins represented the most frequent high-confidence prediction (Fig. 5A), with 1,832 additional candidates expanding the annotated repertoire by 62% relative to 2,967 known members. Interestingly, these predicted proteins were substantially shorter than previously annotated homing endonucleases (median length 77 versus 233 amino acids) (Fig. 5B). Direct structural search and clustering provided little evidence for these candidates: Foldseek search against BFVD[22] identified informative structural homologs for only 0.7% of predicted proteins, and

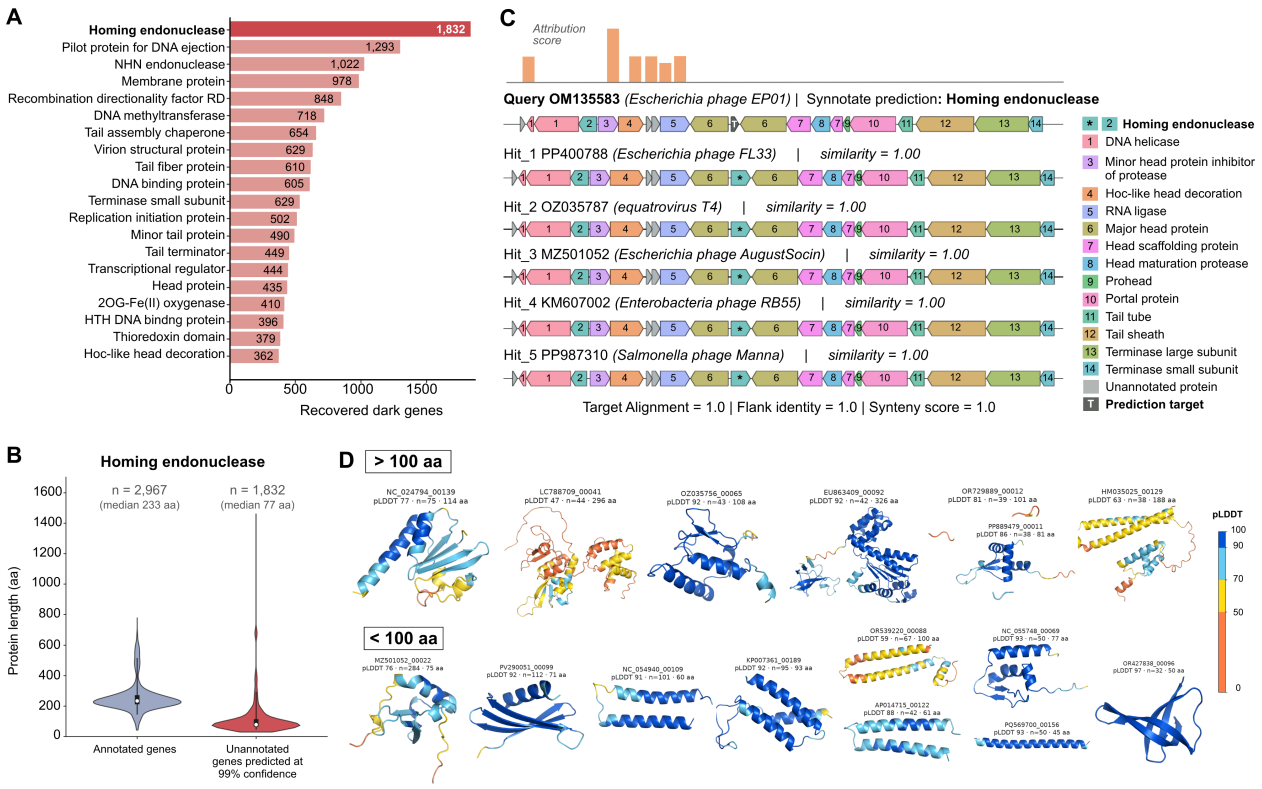


Figure 5. Synnotate annotation of the viral dark proteome. (A) Top 20 Synnotate-predicted annotations at 99% confidence. (B) Length distributions of annotated homing endonucleases and proteins predicted as homing endonucleases. (C) Neighbor attribution and synteny alignment between the query genomic context and retrieved examples. (D) Representative proteins from the top 16 clusters of predicted homing endonucleases.

only 0.6% clustered with previously annotated homing endonucleases (Section 2.6). Yet these proteins retained the canonical genomic position between capsid protein genes (Fig. 5C) and broadly overlapped the phylogenetic distribution of known homing endonucleases (Fig. S6B). When clustered independently, approximately twelve structural clusters accounted for half of all predicted homing endonucleases (Fig. S7, Fig. 5D). The predominance of short proteins was notable as GIY-YIG homing endonucleases contain compact catalytic domains[29,30] that can be separated from larger accessory regions, suggesting that some predictions may represent highly divergent, minimized forms, or highly degraded remnants of mobile elements.[31] Together, these findings suggest that conserved genomic position can persist even when the central proteins show little detectable sequence or structural similarity, allowing context-based models to prioritize otherwise overlooked viral proteins.

4. Discussion

Protein functional annotation has been largely driven by sequence and structural similarity, yet these relationships become progressively less informative for rapidly evolving and highly divergent proteins. Here, we show that genomic context provides functional information com-

plementary to sequence and structure. Synnotate translates this signal into a predictive and interpretable framework for large-scale protein function annotation beyond homology.

The proteins recovered by Synnotate illustrate three ways in which genomic context can extend annotation. First, some proteins appear to be known family members that are missing from current reference databases because sequence and structural diversity are incompletely sampled. In *Desulfovibrio piger*, the predicted Bbp19 and FtsL proteins lacked informative database hits but showed strong structural agreement with context-retrieved references (TM-scores 0.76 and 0.62). Likewise, 97.7% of predicted YgfX proteins clustered with annotated YgfX proteins despite limited direct database support. Second, some proteins may be more deeply diverged members of known families. This pattern includes the three lower-TM-score *D. piger* predictions—SctB, a Gp27-family protein, and a TraR/DksA C4-type zinc-finger protein—as well as a small subset of predicted YgfX proteins. These proteins retained genomic-context support and weak structural similarity to annotated references (TM-scores 0.41–0.49). Third, some predictions may represent poorly characterized or previously unrecognized evolutionary variants. The phage homing endonuclease-like proteins showed little detectable sequence or structural similarity to known examples but repeatedly occurred at the same conserved genomic position and within overlapping phage lineages. They may represent highly divergent variants, compact forms, decaying mobile-element remnants, or previously unrecognized branches of related proteins. Together, these examples show that genomic context can recover both overlooked members of known families and more deeply divergent proteins that fall outside the reach of current sequence- and structure-based annotation.

Several challenges remain. First, Synnotate depends on informative neighboring annotations and conserved gene order, so performance may decline in sparsely annotated or highly rearranged regions. Its functional vocabulary also inherits gaps and biases from existing annotation methods. In addition, confidence thresholds were determined based on annotated genes and applied to unannotated ones. Finally, although our datasets cover broad bacterial and viral diversity, they represent only a small fraction of known microbial genomes. Expanding training to larger and more diverse genome collections and metagenomes[32] may improve performance and extend annotation to less explored protein families.

More broadly, Synnotate can support genome mining beyond protein annotation. Its neighborhood embeddings and vector database make it possible to retrieve genes that occur in similar genomic contexts, even when their annotations or protein sequences differ. By combining context similarity with the diversity of proteins found at the central position, these representations could help identify mobile-element insertion sites, cassette-exchange hotspots, and regions of rapid gene turnover. In this way, genomic context could become not only an annotation signal but also a searchable representation of genome organization, revealing conserved loci and exchange hotspots that sequence-based comparisons fail to connect.

5. Code and Data Availability

Code and data: `https://github.com/khoa-yelo/Synnotate`. Supplementary materials: `https://github.com/khoa-yelo/Synnotate/blob/main/paper/supp.pdf`.

References

1. P. Escudeiro, C. S. Henry and R. P. M. Dias, Functional characterization of prokaryotic dark matter: the road so far and what lies ahead, *Curr. Res. Microb. Sci.* **3**, p. 100159 (2022), https://doi.org/10.1016/j.crmicr.2022.100159.
2. S. Nayfach *et al.*, Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome, *Nat. Microbiol.* **6**, 960 (July 2021), https://doi.org/10.1038/s41564-021-00928-6.
3. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek, *Nat. Biotechnol.* **42**, 243 (February 2024), https://doi.org/10.1038/s41587-023-01773-0.
4. Z. N. Flamholz, S. J. Biller and L. Kelly, Large language models improve annotation of prokaryotic viral proteins, *Nat. Microbiol.* **9**, 537 (February 2024), https://doi.org/10.1038/s41564-023-01584-8.
5. M. H. Medema, E. Takano and R. Breitling, Detecting sequence homology at the gene cluster level with MultiGeneBlast, *Mol. Biol. Evol.* **30**, 1218 (May 2013), https://doi.org/10.1093/molbev/mst025.
6. C. L. M. Gilchrist *et al.*, cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters, *Bioinform. Adv.* **1**, p. vbab016 (January 2021), https://doi.org/10.1093/bioadv/vbab016.
7. K. Blin *et al.*, antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation, *Nucleic Acids Res.* **53**, W32 (July 2025), https://doi.org/10.1093/nar/gkaf334.
8. N. Jha *et al.*, Gaia: an AI-enabled genomic context-aware platform for protein sequence annotation, *Sci. Adv.* **11**, p. eadv5109 (June 2025), https://doi.org/10.1126/sciadv.adv5109.
9. S. R. Grigson *et al.*, Synteny-aware functional annotation of bacteriophage genomes with Phynteny bioRxiv preprint (July, 2025), https://doi.org/10.1101/2025.07.28.667340.
10. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil and P. Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy, *Nucleic Acids Res.* **50**, D785 (September 2021), https://doi.org/10.1093/nar/gkab776.
11. R. Cook *et al.*, INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes, *Phage (New Rochelle)* **2**, 214 (December 2021), https://doi.org/10.1089/phage.2021.0007.
12. P. Terzian *et al.*, PHROG: families of prokaryotic virus proteins clustered using remote homology, *NAR Genom. Bioinform.* **3**, p. lqab067 (August 2021), https://doi.org/10.1093/nargab/lqab067.
13. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork and J. Huerta-Cepas, eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale, *Mol. Biol. Evol.* **38**, 5825 (October 2021), https://doi.org/10.1093/molbev/msab293.
14. B. Buchfink, K. Reuter and H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND, *Nat. Methods* **18**, 366 (April 2021), https://doi.org/10.1038/s41592-021-01101-x.
15. X. Mao, T. Cai, J. G. Olyarchuk and L. Wei, Automated genome annotation and pathway identification using the KEGG orthology (KO) as a controlled vocabulary, *Bioinformatics* **21**, 3787 (October 2005), https://doi.org/10.1093/bioinformatics/bti430.
16. E. L. Sonnhammer, S. R. Eddy and R. Durbin, Pfam: a comprehensive database of protein domain families based on seed alignments, *Proteins* **28**, 405 (July 1997), https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l.
17. M. Steinegger and J. Söding, MMseqs2 enables sensitive protein sequence searching for the

analysis of massive data sets, *Nat. Biotechnol.* **35**, 1026 (November 2017), https://doi.org/10.1038/nbt.3988.
18. A. F. T. Martins and R. F. Astudillo, From softmax to sparsemax: a sparse model of attention and multi-label classification arXiv:1602.02068 (February, 2016), https://doi.org/10.48550/arXiv.1602.02068.
19. M. Douze *et al.*, The Faiss library arXiv:2401.08281 (October, 2025), https://doi.org/10.48550/arXiv.2401.08281.
20. S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.* **48**, 443 (March 1970), https://doi.org/10.1016/0022-2836(70)90057-4.
21. M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Steinegger and B. Rost, ProstT5: bilingual language model for protein sequence and structure bioRxiv preprint (July, 2023), https://doi.org/10.1101/2023.07.23.550085.
22. R. S. Kim, E. L. Karin, M. Mirdita, R. Chikhi and M. Steinegger, BFVD—a large repository of predicted viral protein structures, *Nucleic Acids Res.* **53**, D340 (January 2025), https://doi.org/10.1093/nar/gkae1119.
23. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov and M. Steinegger, ColabFold: making protein folding accessible to all, *Nat. Methods* **19**, 679 (June 2022), https://doi.org/10.1038/s41592-022-01488-1.
24. Y. Zhang and J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Res.* **33**, 2302 (2005), https://doi.org/10.1093/nar/gki524.
25. J. Sukumaran and M. T. Holder, DendroPy: a Python library for phylogenetic computing, *Bioinformatics* **26**, 1569 (June 2010), https://doi.org/10.1093/bioinformatics/btq228.
26. B. D. Ondov *et al.*, Mash: fast genome and metagenome distance estimation using MinHash, *Genome Biol.* **17**, p. 132 (June 2016), https://doi.org/10.1186/s13059-016-0997-x.
27. T. Conway *et al.*, Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing, *mBio* **5**, e01442 (July 2014), https://doi.org/10.1128/mBio.01442-14.
28. E. S. Miller, E. Kutter, G. Mosig, F. Arisaka, T. Kunisawa and W. Rüger, Bacteriophage T4 genome, *Microbiol. Mol. Biol. Rev.* **67**, 86 (March 2003), https://doi.org/10.1128/MMBR.67.1.86-156.2003.
29. J. C. Kowalski *et al.*, Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings, *Nucleic Acids Res.* **27**, 2115 (May 1999), https://doi.org/10.1093/nar/27.10.2115.
30. K. W. Loedige *et al.*, Modular determinants of cleavage preference in GIY-YIG nucleases revealed by block-based DNA shuffling and directed evolution, *Nucleic Acids Res.* **54**, p. gkaf1430 (January 2026), https://doi.org/10.1093/nar/gkaf1430.
31. A. Barzel, U. Obolski, J. P. Gogarten, M. Kupiec and L. Hadany, Home and away—the evolutionary dynamics of homing endonucleases, *BMC Evol. Biol.* **11**, p. 324 (November 2011), https://doi.org/10.1186/1471-2148-11-324.
32. G. Ta *et al.*, MGnify genomes: a resource for biome-specific microbial genome catalogues, *J. Mol. Biol.* **435**, p. 168016 (July 2023), https://doi.org/10.1016/j.jmb.2023.168016.
33. K. Katoh, K. Misawa, K. Kuma and T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.* **30**, 3059 (July 2002), https://doi.org/10.1093/nar/gkf436.
34. Y. Shimoyama, pyMSAviz: MSA visualization python package for sequence analysis https://moshi4.github.io/pyMSAviz/, (2026), Accessed: Jul. 28, 2026.
35. Z. Lin, H. Akin, R. Rao, B. Hie *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**, 1123 (March 2023), https://doi.org/10.1126/science.ade2574.