nature communications

Article

https://doi.org/10.1038/s41467-026-69098-5

Uncovering Cas9 PAM diversity through metagenomic mining and machine learning

Received: 22 July 2025

Accepted: 23 January 2026

Published online: 08 February 2026

Check for updates

Tao Fang[1,5], Lea Bogensperger[2,5], Lilith Feer[2], Ahmed Allam[2], Valentyn Bezshapkin[3], Zsolt Balázs[2], Christian von Mering[4], Shinichi Sunagawa[3], Michael Krauthammer[2] ✉ & Gerald Schwank[1] ✉

Recognition of protospacer adjacent motifs (PAMs) is crucial for target site recognition by CRISPR–Cas systems. In genome editing applications, the requirement for specific PAM sequences at the target locus imposes substantial constraints, driving efforts to search for novel Cas9 orthologs with extended or alternative PAM compatibilities. Here, we present CRISPR-PAMdb, a comprehensive and publicly accessible database compiling Cas9 protein sequences from 3.8 million bacterial and archaeal genomes and PAM profiles from 7.4 million phage and plasmid sequences. Through spacer–protospacer alignment, we infer consensus PAM preferences for 8003 unique Cas9 clusters. To extend PAM discovery beyond traditional alignment-based approaches, we develop CICERO, a machine learning model predicting PAM preferences directly from Cas9 protein sequences. Built on the ESM2 protein language model and trained on the CRISPR–PAMdb database, CICERO achieves an average cosine similarity of 0.69 on test data and 0.75 on experimentally validated Cas9 orthologs. For Cas9 clusters where alignment-based predictions are infeasible, CICERO generates PAM profiles for an additional 50,308 Cas9 proteins, including 17,453 high-confidence predictions with CICERO confidence scores above 0.8. Together, CRISPR–PAMdb and CICERO enable large-scale exploration of PAM diversity across Cas9 proteins, accelerating design of next-generation CRISPR-Cas9 tools for precise genome engineering.

CRISPR–Cas systems, which provide adaptive immunity in prokaryotes, have been repurposed as powerful tools for modern genome editing. In CRISPR-Cas systems, such as type II (Cas9) and type V (Cas12), the recognition of protospacer adjacent motifs (PAMs)—short DNA sequences (usually 2–6 bases) adjacent to the target site—is critical. The PAM acts as a molecular signal, enabling Cas proteins to bind specific DNA sequences and distinguish between self and non-self DNA, thus preventing unintended cleavage of the host genome. Upon PAM recognition, Cas unwinds adjacent double-stranded DNA, allowing R-loop formation where the CRISPR RNA (crRNA) pairs with the target strand. If the crRNA spacer exhibits sufficient complementarity to the target sequence—particularly in the seed region—Cas-mediated DNA cleavage is initiated[1–4].

While critical for precision, PAM recognition imposes a major constraint on CRISPR-Cas systems by limiting the range of targetable genomic sites. This restriction presents a significant challenge for genome editing, particularly in therapeutic contexts where flexible targeting is crucial for treating diverse genetic disorders[5]. Expanding

[1]Institute of Pharmacology and Toxicology, University of Zurich, Zürich, Switzerland. [2]Department of Quantitative Biomedicine, University of Zurich, Zürich, Switzerland. [3]Institute of Microbiology, ETH Zurich, Zürich, Switzerland. [4]Department of Molecular Life Sciences, University of Zurich, Zürich, Switzerland. [5]These authors contributed equally: Tao Fang, Lea Bogensperger. ✉e-mail: michael.krauthammer@usz.ch; schwank@pharma.uzh.ch

PAM diversity, either through protein engineering or the discovery of natural Cas proteins with novel PAM specificities, is therefore a key priority for increasing the versatility and effectiveness of CRISPR-Cas technologies[6–8].

To uncover new CRISPR–Cas systems with novel PAM specificities, several studies have developed large-scale computational frameworks to infer PAM motifs from metagenomic data. Ciciani et al. established an automated pipeline that extracted and oriented CRISPR spacers, aligned them to viral and plasmid genomes, and identified conserved flanking motifs, generating one of the first large-scale bioinformatically inferred PAM catalogs for Cas9 orthologs[9]. Building on similar principles, Ruffolo et al.[10] and Nayfach et al.[11] applied related spacer–protospacer alignment strategies to vastly expanded metagenomic resources, reportedly inferring tens of thousands of PAM motifs across Cas8, Cas9, and Cas12 families. However, none of these studies provides publicly accessible datasets that link individual Cas proteins to their inferred PAM preferences. Consequently, despite their conceptual and technical advances, these resources cannot yet be used to identify or engineer Cas9 variants with defined PAM recognition profiles, or to train and benchmark machine learning models that relate Cas9 sequences to PAM specificity.

To address this gap, we introduce CRISPR-PAMdb, a publicly available, curated database of Cas9 protein sequences and their inferred PAMs, derived from mining over 3.8 million bacterial and archaeal genomes and more than 7.4 million phage and plasmid sequences. Inferred PAMs are associated with individual Cas9 proteins by aligning CRISPR spacers to protospacers within phage and plasmid sequences, followed by the extraction of consensus PAM sequences. Furthermore, we present CICERO (CRISPR Cas9 PAM predictor model), a machine learning model that is trained on the CRISPR-PAMdb database and capable of predicting PAM preferences directly from Cas9 protein sequences using an ESM2 protein language model backbone[12]. By providing these resources and the accompanying pipelines, we aim to facilitate broader exploration of the Cas9 family to accelerate rational design of next-generation genome-editing tools with alternative PAM preferences.

Results

Alignment-based pipeline for CRISPR-Cas9 system mining and PAM inference

We first developed a computational pipeline (Fig. 1a and "Methods" section) to systematically mine CRISPR-Cas systems from bacterial and archaeal (meta)genomes, along with their associated protospacers from phage and plasmid sequences. Due to computational constraints, we focused our analysis on CRISPR-Cas9 systems, although the pipeline is readily adaptable and could also be retrained on other CRISPR-Cas systems, such as CRISPR-Cas12. The only required modifications are the use of Hidden Markov Models specific to the target effector family and selection of consensus PAM motifs from either the upstream or downstream region of mapped protospacers, depending on the orientation of the specific CRISPR system (See Methods for details).

The pipeline first retrieved 3,747,151 bacterial and archaeal genomes from the mOTUs4 database, consisting of isolate genomes, single amplified genomes, and metagenome-assembled genomes[13], and an additional 25,371 archaeal genomes from the European Nucleotide Archive (ENA)[14]. From these, we identified CRISPR arrays, including repeat and spacer sequences, and flanking protein-coding genes. These genes were subsequently screened using curated Cas hidden Markov models[15], resulting in the identification of 21,265,037 Cas proteins and 430,685 Cas9 proteins. Based on the taxonomy information from mOTUs4, Cas proteins were identified in 1,658,140 bacterial genomes, of which 335,447 (20%) contain Cas9 proteins. In contrast, among the 23,352 archaeal genomes with Cas proteins, only 429 contain Cas9 proteins (Fig. 1b), consistent with previous observations of Cas9 rarity in archaea. This scarcity may reflect lower viral pressure in archaeal environments, potentially favoring alternative defense mechanisms. Alternatively, it might stem from higher fitness costs of maintaining Cas9 in energy-limited or extreme environments[16–20].

To identify the protospacer adjacent motifs (PAMs) associated with the Cas9 family, we grouped crRNAs (spacers) from 62,542 Cas9 protein clusters sharing ≥98% sequence identity. Among these, 31,190 clusters contained more than 10 spacers. For 21,137 of them, we successfully mapped protospacers within 7,413,108 phage and plasmid sequences derived from a comprehensive collection of mobile genetic elements, including plasmid sequences from IMG-PR[21], phage sequences from IMG-VR[22], and additional phage sequences from an expanded collection compiled from other studies (see "Methods" section for details; Supplementary Data 1). Importantly, the alignment of spacers to their corresponding protospacers allowed us to infer PAM preferences for 8003 Cas9 clusters using PAMpredict[9] (Fig. 1c); we term this resource CRISPR-PAMdb. Of note, the accuracy of spacer–protospacer-based PAM inference strongly depends on the availability of unique protospacers in phage sequences. We therefore evaluated the contribution of our collected mobile genetic elements database compared to the commonly used viral genome database, IMG-VR[22]. Interestingly, we found that over half of the mapped protospacers were unique to the mobile genetic elements dataset (Fig. 1d), highlighting the importance of expanding reference datasets for PAM discovery and suggesting limitations in the reliance on IMG-VR and IMG-PR as standard references.

To validate the accuracy of our mining pipeline, we first compared the inferred PAM preferences of three widely used Cas9 enzymes with their experimentally established PAM motives: SpCas9 with an NGG PAM[23,24], SaCas9 with an NNGR(A/G)R(A/G)T PAM[25,26], and CjCas9 with an NNNNR(A/G)Y(C/T)AC PAM[27–29]; Supplementary Figs. 1–3. For CjCas9, we identified an almost identical amino acid sequence in CRISPR-PAMdb (99.29% sequence similarity with 100% coverage of the canonical CjCas9 sequence), which showed an inferred PAM profile nearly identical to the reported CjCas9 PAM (NNNNAYAC; Supplementary Fig. 1). For SaCas9, we identified three highly similar amino acid sequences (~96% similarity with 100% coverage of the canonical SaCas9 sequence; Supplementary Fig. 2). One of these exhibited the reported NNGRRT profile of SaCas9, while the other two displayed different PAM profiles (NNGA and ATGAANT), likely reflecting natural diversification among closely related SaCas9 orthologs. For SpCas9, the most similar entry in CRISPR-PAMdb shared only 93% identity with full-length coverage, yet the inferred PAM profile retained identical to the reported NGG PAM profile of SpCas9 (Supplementary Fig. 3).

Next, we evaluated the performance of our pipeline on an additional set of 79 Cas9 proteins for which the PAM profiles were experimentally characterized by Gasiunas et al.[30]. We did this by identifying closely related Cas9 clusters to the external references via sequence similarity searches using blastp[31]. Only the top-matching cluster per reference that met our de-duplication thresholds (≥98% sequence identity and ≥90% coverage of the external reference sequence; see "Methods" for details) was then retained. Our mined Cas9 dataset included closely related Cas9 clusters for 65 of the 79 reference Cas9 proteins. After filtering, consensus PAM inferences were obtained for 26 out of the 65 Cas9 clusters (Supplementary Fig. 4b and Supplementary Data 2). We then compared the experimentally determined PAM profiles of the reference Cas9 variants with the bioinformatically inferred PAM profiles using augmented cosine similarity (this is referred to as accuracy, see Methods for details and Supplementary Information for a formal definition), where higher values indicate greater overall overlap in base preferences across all PAM positions. Confirming the high accuracy of our bioinformatic PAM identification approach, we obtained an average cosine similarity

https://doi.org/10.1038/s41467-026-69098-5

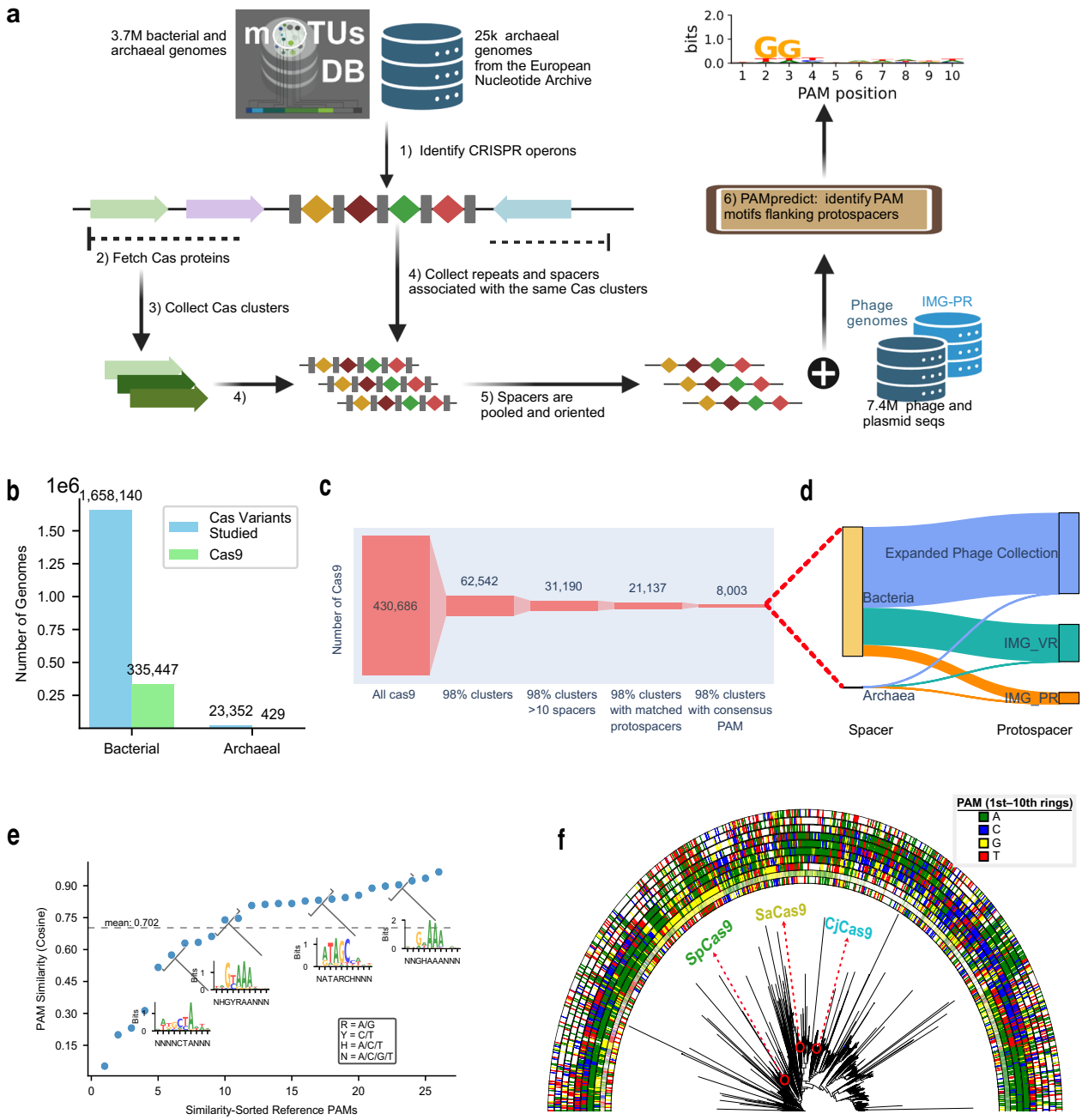

Fig. 1 | Pipeline overview and results of alignment-based PAM inference. **a** Schematic overview of our mining pipeline. (1) Identify CRISPR repeat-spacer arrays from bacterial and archaeal genomes; (2) Fetch Cas proteins; (3) Collect Cas clusters; (4) Collect repeats and spacers associated with the same Cas clusters; (5) Spacers are pooled and oriented consistently by clustering repeats; (6) Map pooled spacers to identify PAMs flanking protospacers in dereplicated phage and plasmid sequences using PAMpredict[9]. Created in BioRender. Fang, T. (2026) https://BioRender.com/58sntmw. **b** Count and sources of mined Cas proteins. **c** Count of Cas9 variants at each pipeline stage. **d** The Sankey plot illustrates the mapping between spacers in the bacterial and archaeal genomes and their mapped protospacers in phage and plasmid sequences for the final set of 8003 Cas9 variants with consensus PAM inferences. **e** Augmented cosine similarity between 26 reference PAM sequences with consensus PAM inferences derived from their most similar Cas9 protein clusters (≥98% sequence similarity and ≥90% coverage of the external reference sequence). Accompanying sequence logo plots illustrate the inferred PAM profiles across different levels of PAM similarity. The string shown below the x-axis in each plot represents the corresponding external reference PAM sequence. **f** Phylogenetic tree of Cas9 protein clusters sharing 98% sequence similarity. Annotations, from inner to outer rings, indicate the most likely nucleotide inferred at each of the 10 PAM positions. To facilitate interpretation, we also include and indicate the position of the three frequently employed Cas9 enzymes SpCas (NGG), SaCas9 (NNGRR(A/G)T), and CjCas9 (NNNNR(A/G)Y(C/T)AC) in the tree. Source data are provided as a Source Data file.

of 0.702, with 22 of the 26 Cas9 clusters (84.6%) showing cosine similarities above 0.5 (Fig. 1e).

We next constructed a phylogenetic tree of all 8003 identified Cas9 protein clusters with their consensus PAMs (Fig. 1f). Confirming the overall accuracy of our clustering and annotation strategy, the positions of well-characterized Cas9 enzymes, such as SpCas9, SaCas9, and CjCas9, aligned well with their established phylogenetic relationships[30,32]. Furthermore, we observed that closely related Cas9 sequences generally share similar PAM motifs, suggesting that PAM specificity has been largely conserved within evolutionary

lineages. At the same time, the tree underscores the remarkable diversification of PAM preferences across the entire Cas9 family, although global nucleotide usage analysis revealed a moderate but consistent enrichment of purines at most PAM positions, with purine-to-pyrimidine ratios at positions 1–10 of 1.3, 7.7, 4.7, 3.0, 1.4, 0.6, 2.1, 1.9, 1.6, and 1.0, respectively (Fig. 1f, see "Methods" for details).

Machine learning-based PAM prediction for Cas9 proteins using CICERO

For the majority of the 62,542 identified Cas9 protein clusters, we were unable to map a sufficient number of protospacers in phage and plasmid databases to infer PAMs. To overcome this limitation, we developed and trained CICERO (CRISPR Cas9 PAM predictor model), a machine learning model that is capable of predicting PAM profiles directly from CRISPR Cas9 sequences using a protein language model (pLM) backbone (ESM2)[12]. Our training pipeline, similar to Nayfach et al.[11], follows a two-phase scheme (Fig. 2a). In phase 1, CICERO is trained to predict PAMs directly for given Cas9 proteins as inputs. Thereby, the protein language model embeddings are extended by a multi-layer perceptron (MLP) network to predict PAM profiles, which reflect how likely each nucleotide is to appear at each position (see Methods for more technical details). In phase 2, which only takes place after phase 1 is completed, the setup is further refined to have another MLP layer on top of the trained model from phase 1 to predict corresponding confidence estimates of the predictions. Essentially, this confidence prediction model learns to estimate the accuracy of PAM predictions from phase 1, thereby providing insight into the reliability of each prediction.

To assess the impact of model scaling, we trained CICERO variants with ESM2 backbones ranging from 8 M to 3B parameters and validated them using five-fold cross-validation on the CRISPR-PAMdb dataset, which comprises 8003 Cas9 clusters with inferred PAM preferences (of which 7571 remained after applying a length filter, see "Methods" for details). Model performance varied only modestly across model/parameter sizes, with CICERO-650M achieving the highest accuracy, showing an average cosine similarity between predicted and inferred PAM profiles of 0.69 ± 0.03 across all held-out test splits (Fig. 2b and Supplementary Figs. 5 and 6). To further dissect the determinants of prediction accuracy, we analyzed multiple Cas9 and PAM-related factors. Accuracy increased for longer Cas9 sequences – likely reflecting the greater representation of long variants in the training data—and for shorter PAM profiles, which are inherently easier to predict (Supplementary Fig. 5c–d). Finally, stratification by sequence similarity revealed that performance improves with increasing similarity to the training set, but remains robust even for more divergent proteins (Supplementary Fig. 5e), indicating that CICERO-650M generalizes beyond closely related Cas9 orthologs.

To further validate the performance of CICERO-650M, we applied it on the well-characterized Cas9 nucleases SpCas9, SaCas9, and CjCas9 (Supplementary Figs. 1–3). For SpCas9, the model accurately predicted the canonical NGG PAM with high confidence (0.92). For CjCas9, the predicted NNNNYAAC motif (confidence of 0.87) closely matched the reported NNNNR(A/G)Y(C/T)AC motif, differing only at position 5. For SaCas9, CICERO-650M generated a PAM profile broadly consistent with the reported NNGR(A/G)R(A/G)T motif, correctly identifying R(A/G) at positions 4–5 and T at position 6. While the confidence of the SaCas9 PAM prediction was lower (0.63), this variation is consistent with the known diversity of PAM preferences among different SaCas9 family members in CRISPR-PAMdb. Finally, we also benchmarked CICERO-650M on a dataset of 79 Cas9 proteins with experimentally determined PAM profiles reported by Gasiunas et al.[30]. On this dataset, CICERO-650M achieved a median augmented cosine similarity of 0.75 (Fig. 2b and Supplementary Fig. 7).

Next, we tested whether the trained confidence network (i.e., confidence head) installed on top of the PAM prediction layer could reliably estimate prediction accuracy. Confidence scores indeed showed a strong correlation with actual prediction accuracy when evaluated on the set of experimentally validated Cas9 proteins Spearman correlation test: $r = 0.8$, $p = 8.8 \times 10^{-19}$; Fig. 2c and Supplementary Data 2. This prompted us to conduct a simple thresholding strategy, in which the predictions for the Cas9 proteins of the CRISPR-PAMdb test data were grouped into confidence bins (0.7–0.8, 0.8–0.9, >0.9). As expected, prediction accuracy increased markedly with higher confidence, reaching an accuracy of 0.86 for predictions with confidence >0.9 (Fig. 2d). Applying the binning strategy to the 79 experimentally validated Cas9 proteins showed similar results, with the median accuracy rising to 0.89 for predictions with confidence >0.8 (37/79 variants; Supplementary Data 2), and to 0.95 for those with confidence >0.9 (9/79 variants; Fig. 2e).

In a recent preprint, Nayfach et al. reported a related Cas9 PAM prediction model, Protein2PAM[11]. When evaluated on the same 79 experimentally validated Cas9 proteins, Protein2PAM achieved an accuracy of 0.81 (augmented cosine similarity), which—similar to CICERO—further increased when filtering for high-confidence predictions, reaching 0.89 for confidence >0.8 (40/79) and 0.92 for confidence >0.9 (17/79). Although these results indicate a slightly higher performance of Protein2PAM, the comparison remains limited by the small size of the benchmark dataset. Future benchmarking of both models on a larger set of Cas9 proteins with bioinformatically inferred PAM preferences would therefore be valuable. Such a head-to-head evaluation will become feasible once the Protein2PAM training data and model weights are publicly released.

We next applied CICERO-650M to the 54,539 Cas9 clusters for which we could not infer the PAM preference due to the lack of sufficient protospacers identified in phage and plasmid databases. To adhere to the same preprocessing routine as in training CICERO-650M, a length filter was applied, after which 50,308 sequences with a length between 200 and 1538 remained. These sequences were used as input for each trained fold of CICERO-650M, from which predictions and confidence scores were computed as the average across folds. This resulted in 32,016 predictions with confidence larger than 0.7, 17,453 predictions with confidence larger than 0.8, and 3119 predictions with confidence larger than 0.9 (Fig. 3a). Using this dataset, we also constructed a phylogenetic tree of Cas9 protein clusters with PAM predictions of confidence >0.7, providing a broad and systematic map of PAM diversity across 50,308 members of the Cas9 family (Fig. 3b).

Finally, we applied CRISPR-PAMdb and CICERO to illustrate how Cas9 orthologs with diverse PAM specificities can expand the target range of base editors. Conventional SpCas9 base editors, which couple catalytically impaired SpCas9 to deaminases for single-base conversion[33,34], are constrained by the strict NGG PAM requirement located ~10–20 bp upstream of the target base, leaving many pathogenic variants inaccessible. PAM-relaxed SpCas9 variants, such as SpG and SpRY[7], alleviate this limitation but at the cost of reduced specificity. To assess the extent to which Cas9 orthologs with similar specificities than SpCas9 could unlock additional target sites, we systematically analyzed pathogenic and likely pathogenic ClinVar variants, focusing on T·A → C·G and G·C → A·T transitions amenable to cytosine or adenine base editing (See "Methods" for details). Among 8897 T·A → C·G and 37,636 G·C → A·T transitions, only 47.0% and 48.2% were targetable with SpCas9, respectively, whereas theoretical maximum coverage was achieved when including Cas9 orthologs with PAM specificities ≥NGG, either mined from CRISPR-PAMdb or predicted by CICERO with confidence >0.8 (Supplementary Fig. 8 and Supplementary Data 3).

Discussion

In this study, we identified 62,542 unique Cas9 clusters from 3.8 million bacterial and archaeal genomes and screened 7.4 million phage and plasmid sequences to detect protospacer matches via

https://doi.org/10.1038/s41467-026-69098-5

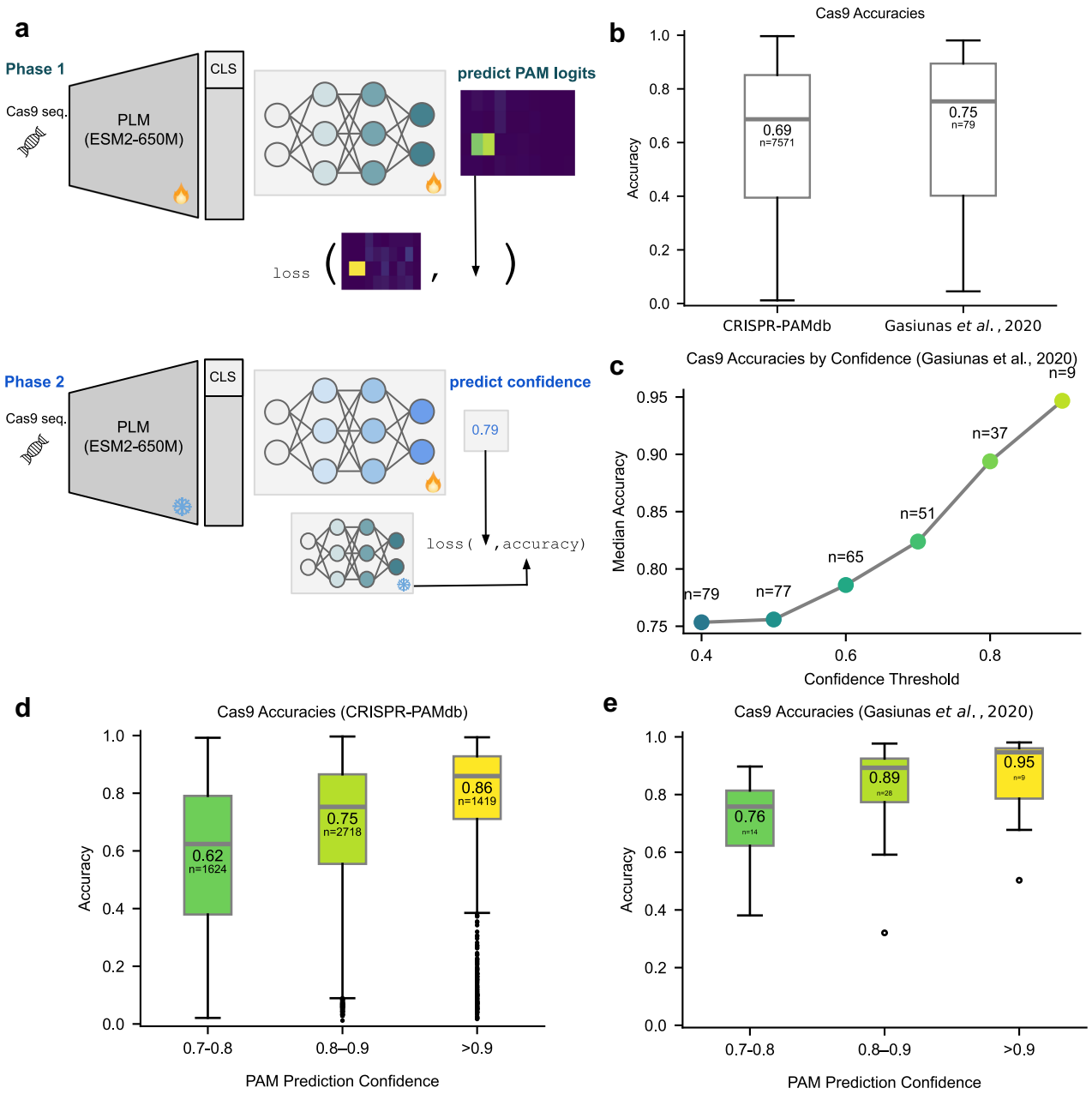

Fig. 2 | CICERO training pipeline and PAM prediction accuracies on CRISPR-PAMdb and external validation set (Gasiunas et al.[30]). a Two-phase training pipeline similar to Nayfach et al.[11] using a protein language model as shared backbone. In phase 1, the model is trained to predict PAM profiles given Cas9 input sequences. Phase 2 reuses the same backbone and adds a PAM-prediction MLP network as a confidence prediction module, which is trained to regress the PAM prediction accuracy, computed as augmented cosine similarity to the target PAM profile. **b** Median accuracies on CRISPR-PAMdb Cas9 sequences aggregated over all cross-validation folds and on the 79 external validation set. Box plots show the median (center line), interquartile range (IQR; bounds of the box correspond to the 25th and 75th percentiles), whiskers extending to 1.5× IQR from the quartiles, and individual points beyond the whiskers as outliers. n indicates the number of evaluated Cas9 samples. The same definition applies to (**d**, **e**). **c** Median accuracy for the 79 external validation set, filtered according to varying confidence thresholds. Retaining only predictions with a confidence of at least 0.8 yields a much higher accuracy of 0.89 (for $n = 37$ filtered samples) compared to the baseline accuracy of 0.75 based on all 79 samples. (**d**) Median accuracies on CRISPR-PAMdb Cas9 sequences aggregated over all cross-validation folds as in (**b**) with additional grouping by predicted confidence. The number of samples in each confidence bin is shown, out of a total of 7571 Cas9 samples. **e** Median accuracies on the external validation set as in (**b**) with additional grouping by predicted confidence. The number of samples in each confidence bin is shown, out of a total of 79 Cas9 samples. Source data are provided as a Source Data file.

spacer–protospacer alignment. This enabled the construction of CRISPR-PAMdb, a publicly available database containing 8003 Cas9 clusters with inferred PAM preferences. To further expand PAM coverage, we developed CICERO, a machine learning model capable of directly predicting PAM profiles from Cas9 protein sequences. CICERO extended PAM annotations to an additional 50,308 Cas9 clusters, providing a systematic and scalable framework for exploring PAM diversity across the Cas9 protein family.

The accuracy of alignment-based PAM inference is inherently dependent on the diversity and abundance of unique spacers and protospacers available in current genomic datasets. A key strength of our study is the use of what is, to our knowledge, the most

https://doi.org/10.1038/s41467-026-69098-5

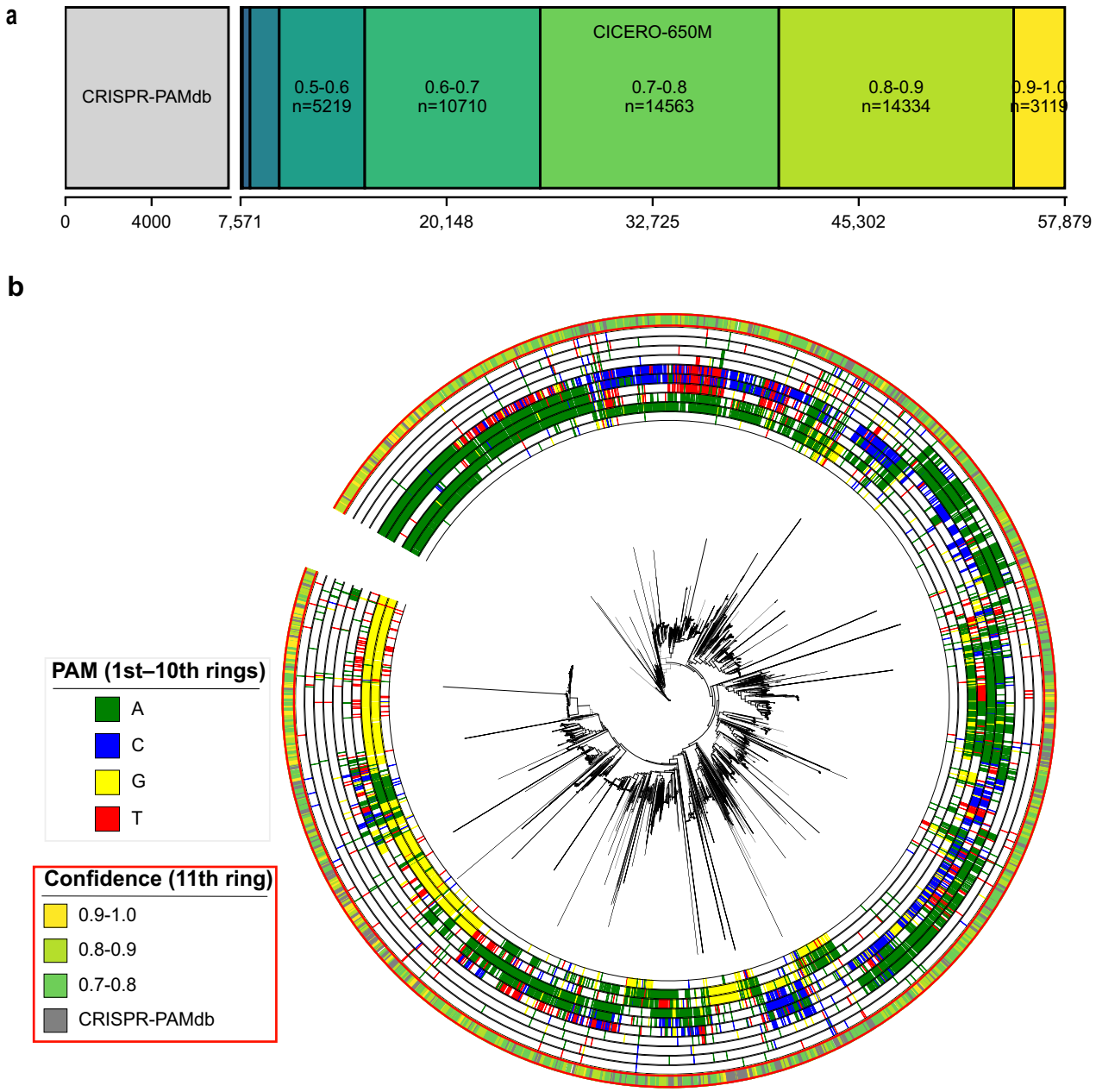

Fig. 3 | Expansion of Cas9 PAM diversity using CICERO predictions. a Expansion of Cas9 PAM predictions using CICERO-650M. The initial 8003 PAMs bioinformatically inferred from metagenomic datasets (gray) are extended to over 50,000 via CICERO-650M (colored). Each color represents a confidence bin of predicted PAMs. Notably, over 60% of the unlabeled Cas9 sequences received predictions with a confidence score above 0.7. **b** Phylogenetic tree of Cas9 protein clusters with consensus PAM inferred from metagenomic datasets or predicted with high-confidence (confidence score > 0.7) by CICERO. Annotations from the inner to outer rings represent the most likely nucleotide at each of the 10 PAM positions. The outermost ring indicates the source of the PAM prediction—inferred PAM (CRISPR-PAMdb) or CICERO—and, for CICERO-derived predictions, the associated confidence score. Source data are provided as a Source Data file.

comprehensive collection of bacterial, archaeal, phage, and plasmid genomes assembled to date. This breadth is essential, as spacer and protospacer availability is shaped by ongoing host–invader co-evolution[35,36]. New spacers are regularly acquired in response to emerging threats, while older ones may be lost to preserve array compactness and efficiency. Similarly, protospacers in phage and plasmid genomes are frequently mutated or deleted to evade CRISPR-mediated interference.

To improve the robustness of inference, we clustered Cas9 proteins into homologous groups (>98% homology), allowing the aggregation of multiple unique spacers and protospacers for consensus PAM inference. While this approach increases statistical power, it may obscure differences in PAM preference between highly similar Cas9 variants. Nevertheless, when tested on a set of Cas9 variants with experimentally validated PAM preferences, we observed strong similarities to our inferred PAM profiles, suggesting that the CRISPR-PAMdb database will be a valuable resource for selecting Cas9 orthologs for genome editing experiments.

Compared to previous efforts[9,10], our pipeline produced a higher number of Cas9 clusters with inferred PAM preferences, emphasizing the importance of leveraging comprehensive bacterial and archaeal genomic datasets, as well as comprehensive phage and plasmid

sequence collections, for alignment-based PAM inference. Utilizing the taxonomic annotations provided by mOTUs4, we determined that 1,658,140 of 3,709,852 bacterial genomes (45%) and 23,352 of 62,669 archaeal genomes (37%) harbor Cas proteins. At the species level, 31,513 of 64,205 bacterial species (approximately 49%) were found to harbor Cas proteins. These findings are consistent with prior reports indicating that approximately 50% of bacterial genomes contain CRISPR systems. However, the lower-than-expected prevalence of Cas proteins in archaeal genomes—relative to historical estimates nearing 90%—may reflect gaps in phylogenetic coverage within our genome database or limitations of current Hidden Markov Models (HMMs) in detecting highly divergent or previously uncharacterized CRISPR–Cas variants in archaea. Alternatively, it may indicate the presence of distinct, non-CRISPR defense strategies in archaeal lineages[37–40].

The potential of CICERO to predict PAMs from given Cas9 sequences considerably extends the coverage beyond traditional alignment-based bioinformatics tools. By leveraging a protein language model (ESM2) and incorporating a confidence scoring mechanism, CICERO enables accurate and scalable prediction of PAM preferences directly from Cas9 sequence data. High-confidence predictions (e.g., confidence > 0.9) corresponded to accuracies of up to 0.95 on benchmark datasets, offering a practical filter for prioritizing reliable predictions. Nonetheless, there is room for improvement. For example, future iterations of CICERO could benefit from expanded training datasets that include experimentally validated PAMs from more diverse Cas9 orthologs, particularly from underexplored phylogenetic clades. Moreover, CICERO is currently exclusively trained on Cas9 proteins, but the framework is inherently modular and training could be extended to other CRISPR effector families, such as Cas12 or Cas13.

Recently, Nayfach et al.[11] also reported a machine-learning model (Protein2PAM) for predicting Cas9 PAM preferences directly from protein sequence. Protein2PAM and CICERO share several core design choices: both leverage ESM2-based protein representations, predict PAMs as fixed-length 10 × 4 matrices, optimize distribution-level agreement between predicted and reference PAM profiles, and include confidence estimates to identify unreliable predictions. Despite these similarities, the two approaches differ in several respects. We systematically characterize our modeling approach through ablation studies across protein language-model backbones and across different input and output settings. In addition, CICERO employs a compact confidence head trained jointly with PAM prediction under direct supervision, without additional pretraining objectives or auxiliary sequence similarity-based embeddings. When benchmarking both models on a set of 79 Cas9 nucleases with experimentally validated PAMs, both achieved comparable performance, with Protein2PAM showing slightly higher median accuracy. However, interpretation of this comparison is limited by the small size of the benchmark, and a more comprehensive head-to-head assessment on larger, independent datasets would therefore be valuable in future work.

Taken together, CRISPR-PAMdb and CICERO provide a scalable and reproducible framework for large-scale Cas9–PAM annotation and prediction that integrates evolutionary inference with machine-learning–based generalization. By expanding the repertoire of Cas9 orthologs with defined PAM specificities, this resource facilitates the rational design of genome-editing tools beyond currently used Cas9 orthologs.

Methods

Genomic data acquisition and CRISPR-Cas system identification

3,747,151 isolated and metagenome-assembled bacterial and archaeal genomes were downloaded from the mOTUs4 database[13]. Among these, 37,298 were archaeal genomes. In addition, 25,371 archaeal genomes were retrieved from the European Nucleotide Archive (ENA) as of 2024-10-1[14]. To mine PAMs from mobile genetic elements, we established a reference dataset of phage sequences (N = 7,978,168; Supplementary Data 1) from previously published bulk-metagenomic, viral-particle enrichment and viral isolate studies such as IMG-VR[22] (v4.1) and added an in-house dataset of circular phages identified by mVIRs (v1.1.1, standard parameters, minimal length of circular element is 2000 bp)[41] as a circular element in the mOTUs-db assemblies[13] and classified as "virus" by geNomad[42] (v1.7.4, database version 1.7; end-to-end parameters: --disable-nn-classifcation, --sensitivity 7.5) (reference future downloadable file from mOTUs-db). The dataset was dereplicated using vClust[43] (v1.29, standard parameters for prefilter and align steps; --ani 1 --qcov 1 --tcov 1 --algorithm leiden for cluster step), which resulted in the total of 6,713,135 phage genomes. Additionally, we incorporated 699,973 plasmid genomes from IMG/PR[21] (2023-08-08_1 version), resulting in a total of 7,413,108 mobile genetic elements.

The CRISPR array containing contigs was extracted from these genomes using MinCED (v0.4.2, https://github.com/ctSkennerton/minced) and PILER-CR[44] (v1.06). Results from both tools were merged, overlapping contigs were removed, and only contigs longer than 5000 bp were retained. Protein-coding sequences within 20 kb upstream and downstream of CRISPR arrays in the contigs were predicted using Prodigal-gv[42] (v2.11.0) and Cas variants were then identified by hmmsearch[45] (v3.4) in these flanking regions using curated Cas Hidden Markov Models from CRISPRCasTyper[15] (v1.8.0). We evaluated the pipeline using a manually curated CRISPR-Cas loci benchmark dataset comprising 1106 genomes known to contain CRISPR-Cas9 systems[46]. Our pipeline successfully identified CRISPR-Cas9 systems in 1012 of these genomes and recovered all curated Cas proteins—including, but not limited to, Cas9—in the majority of cases (Supplementary Fig. 4a).

PAM Inference for CRISPR-Cas9 system

PAMs for the identified Cas9s were inferred by aligning associated spacers to protospacers in phage and plasmid genomes from our curated mobile genetic elements dataset (see above), discarding matches with more than four mismatches or gaps. Protospacers and their flanking regions (up to 10 nt on both sides since the orientation of protospacers are unknown) were aligned to derive consensus PAMs in the flanking region using PAMpredict[9] (v.1.0.2). To enhance PAMpredict's accuracy, which depends on the number and diversity of input spacers, spacers from nearly identical Cas9 proteins were grouped using MMseqs2 (v17.b804f, options: --min-seq-id 0.98 -c 1 --cov-mode 0)[47] before analysis, improving PAM inference sensitivity. Spacer sequences associated with nearly identical Cas9 proteins were also uniformly oriented prior to input into PAMpredict by aligning the associated repeats to an arbitrary but consistent reference orientation (without assuming spacer polarity). This was achieved through repeat clustering with cd-hit-est (v4.8.1, options: -c 0.8, -s 0.75, -r 1)[48]. Subsequently, based on the detected consensus PAM located either upstream or downstream of the protospacer, we determined the protospacer and spacer orientations using the necessary contextual information—for example, for Cas9 variants, the PAM is typically downstream of the protospacer, while for Cas12a, it is upstream[49,50]. Redundant spacers were removed using cd-hit-est (options: -c 0.95, -s 1, -r 0), and Cas9 clusters with fewer than 10 spacers after dereplication were excluded from PAM inference.

The inference output is a 10 × 4 information-content matrix representing PAM predictions across ten nucleotide positions and four bases (A, T, G, C). For biological interpretability, we derived consensus IUPAC[51] motifs from these 10 × 4 PAM profiles using information-content thresholds. Here, information content quantifies the positional nucleotide preference by measuring how much the observed base distribution at each site deviates from a uniform background, indicating how strongly a given base is favored. Nucleotides with an information content exceeding 0.5 bits were assigned to each PAM

position, generating consensus sequences that capture strong nucleotide preferences.

Phylogenetic trees of Cas9 protein clusters with 98% sequence similarity were aligned using MAFFT[52] (v7.526), constructed with FastTree[53] (v2.1.11), and visualized in iTOL[54] (v7). The IUPAC consensus motifs were used to calculate purine-to-pyrimidine ratios at PAM positions 1–10 by analyzing the presence of A/G versus C/T at each position, excluding those with strong preferences for both purines and pyrimidines.

Machine learning for PAM prediction

In our experiments, we train CRISPR Cas9 PAM predictor model (CICERO)—an ML model that uses Cas9 protein sequences as input and predicts the associated PAM as an output. Our models use ESM2 as the backbone[12]—a powerful protein language model that encodes information at the sequence level. After evaluating models with varying sizes (8M, 35M, 150M, 650M, and 3B parameters), we selected the 650M-parameter variant for our final approach, termed CICERO-650M. This model encodes each Cas9 protein sequence into a learned embedding that captures global contextual information across the entire protein, which is subsequently fine-tuned during training. In particular, we use the special [CLS] token (where a token refers to an amino acid), which is a meta token first introduced in the BERT architecture[55]. It is placed at the start of the input sequence and enables encoding the entire sequence into a single vector representation for downstream tasks, such as predicting the PAM motifs. This is then passed to an output layer for PAM prediction—a fully-connected two-layer MLP with a rectified linear unit (ReLU) activation, a hidden dimension of 1280 and a dropout rate of 0.2. The MLP outputs a 10×4 matrix of raw PAM logits, corresponding to a fixed 10-length PAM prediction over ten nucleotide positions and four possible bases (A, T, G, C). See Fig. 2a for a visualization of the employed pipeline.

The training objective (see Supplementary Information for a formal definition) combines a standard cross-entropy loss on the predicted PAM logits, combined with an augmented cosine similarity loss at the information level, similar to Nayfach et al.[11], to better align model predictions with biological PAM patterns. This augmented cosine similarity is also used as an evaluation metric for PAM predictions; it is considered a proxy for accuracy by quantifying how well the predicted PAM information content matches the target PAM profile inferred from metagenomic data by PAMpredict[9]. Intuitively, it captures both nucleotide agreement and the positional relevance of each base, encouraging the model to correctly identify high-information (i.e., biologically important) positions, while incorporating a fictitious "N" base to account for low-information regions where no specific base is strongly favored. All reported accuracy values throughout this paper refer to this metric. Representative examples illustrating a range of similarity scores and their interpretation are shown in Fig. 1e, Supplementary Fig. 6, and Supplementary Data 2, which list the similarity scores, predicted consensus IUPAC PAM strings, and corresponding experimentally validated PAMs. Of note, when the training objective and evaluation were adjusted to consider only positions corresponding to the consensus PAM string—excluding trailing "N" positions (see definition of the mask in the loss function as described in detail in Supplementary Information)—model performance improved, achieving an accuracy of 0.75 on test splits and 0.80 on the external dataset when excluding the predicted trailing "N" positions. However, for all analyses and reporting in this study, we focus on the model that evaluates all PAM positions without incorporating length-based bias, providing a more general and unbiased framework.

We additionally implement and train a confidence network to enhance interpretability. The idea of learned confidence estimates has been explored in several works[11,56,57]. Since deep neural networks often suffer from poor calibration by over- or underestimating predictive confidence, such approaches can improve calibration. In CICERO, this is realized through a simple two-layer MLP, which adds only minimal computational overhead while producing interpretable confidence estimates. This is realized in Phase 2 (Fig. 2a), where we freeze both the finetuned protein language model backbone and the PAM-prediction head learned in Phase 1. We then attach a new confidence prediction head on top of the frozen [CLS] embedding. Only this confidence head is trained in this second phase to predict the empirical accuracy of the PAM head on each Cas9 input sequence. It is supervised with real-valued accuracy scores (computed per sequence) and optimized with an L2 loss. Intuitively, the confidence head learns to identify features in the fine-tuned, frozen [CLS] embeddings that predict errors of the PAM prediction head, allowing it to highlight inputs where the PAM prediction is likely unreliable.

For data processing, we used 8003 Cas9 sequences for which a PAM was accurately inferred. As preprocessing, we apply length-based filtering by excluding sequences shorter than 200 amino acids, as such short fragments are unlikely to contain the PAM-interacting domain (PID), which is critical for PAM determination and sequences exceeding the 99th quantile in terms of length (i.e., sequences longer than 1538 amino acids). After filtering, we retain 7571 pairs of sequences and their corresponding PAMs. For evaluation, we performed five-fold cross-validation using a stratified split setup with train/validation/test splits in each fold. Training is conducted using a learning rate of 1e-4 for both the protein language model backbone and the MLP head. The checkpoint with the best validation loss (evaluated on the validation split) is retained during training, which runs for 15 epochs, and is shown to be sufficient for our experiments. The confidence network is trained similarly on each fold for 15 epochs, with a learning rate of 3e-4. The models are first evaluated on the test split of each fold and subsequently on an external data set of 79 Cas9 sequences with experimentally determined PAMs[30], which serves as additional ground truth.

Benchmarking the PAM mining pipeline and CICERO model performance

To validate the performance of our mining pipeline and CICERO model, we conducted sanity checks on three canonical enzymes (SpCas9, SaCas9, and CjCas9) to confirm expected behavior (Supplementary Figs. 1–3). For the mining pipeline, we assessed inferred PAMs of the most similar sequences from CRISPR-PAMdb relative to the canonical enzymes. For CICERO, we evaluated predicted PAMs of the most similar sequences from the 50,308 Cas9 clusters for which we could not infer PAM preferences using the mining pipeline. Sequence similarity to the canonical enzymes was calculated using blastp 2.14.1[31] (*e*-value threshold: 0.0001), retaining sequences with ≥90% identity and ≥90% coverage of the canonical reference. Among those passing these criteria, we selected the most similar sequence based on the highest identity.

We also benchmarked our mining pipeline using an external dataset of 79 Cas9 proteins with experimentally validated PAMs[30]. Sequence similarity between these external Cas9 sequences and the CRISPR-PAMdb Cas9 protein cluster representatives was assessed using blastp 2.14.1(*e*-value threshold: 0.0001). Only the clusters exhibiting the highest similarity and meeting the filtering criteria (≥98% sequence identity and ≥90% coverage of the external reference sequence) were retained for benchmarking. Ultimately, this yielded 26 closely related Cas9 clusters with inferred consensus PAMs for the external reference sequences (Fig. 1e). In contrast, to benchmark the CICERO model, we directly applied predictions to all 79 external Cas9 proteins (with no exclusions or de-duplication relative to the training data), see Supplementary Fig. 7.

Assessing the utility of the PAM mining pipeline and CICERO

To assess the utility and applicability of our CRISPR-PAMdb and CICERO resources, we systematically queried the ClinVar database

(2025-12-02 version) to extract disease-associated single-nucleotide variants (SNVs) classified as pathogenic or likely pathogenic, focusing on T:A-to-C:G and G:C-to-A:T transitions amenable to base editing. The ClinVar dataset was filtered to curate a high-confidence set of genetic variants. Only variants mapped to the GRCh38 genome assembly, with reviewed records, confirmed pathogenicity, and specific phenotype annotations were retained, while variants in cancer-related genes (BRCA1, BRCA2, MLH1, MSH2, PTEN) were excluded. The resulting dataset contained 8897 T:A-to-C:G and 37,636 G:C-to-A:T transitions amenable to base editing. Genomic context sequences, including 200 bp flanking regions on both sides, were generated for each variant using the GRCh38 reference assembly. A base-editing library was constructed by selecting transition variants and generating multiple target sites for each variant. The appropriate DNA strand was selected based on the target base to ensure compatibility with the base editor (A for adenine base editors, C for cytosine base editors). Target sites were created by shifting the mutation position across the protospacer (PAM-distal positions 1–10) and extracting the corresponding 20-bp protospacer and 10-bp PAM sequences. Each 20-bp protospacer and adjacent 10-bp PAM were then evaluated for editor targetability (requiring an appropriate PAM with the pathogenic base at positions 1–10) using Cas9 orthologs from CRISPR-PAMdb and CICERO-predicted editors (confidence > 0.8) (see Supplementary Fig. 8c and 8d for example). To minimize off-target potential, we retained only editors with specificity equal to or greater than that of SpCas9 with an NGG PAM. Specificity scores were computed by assigning values of 0, 1, 2, or 3 to each PAM position corresponding to sites allowing 4, 3, 2, or 1 possible nucleotides, respectively, and summing these scores across all 10 PAM positions. For example, "N" at a given position corresponds to a specificity score of 0, while "R" (A/G) corresponds to a score of 2.

Statistics and reproducibility

This study is entirely computational and does not involve biological experiments, clinical trials. No statistical method was used beyond the calculation of a Spearman rank correlation between predicted confidence scores and external validation accuracies. Sample sizes were defined by exhaustive data availability and predefined quality thresholds. No data were excluded except according to the following pre-established objective criteria: Cas9 sequences shorter than 200 or longer than 1538 amino acids were removed; clusters with fewer than 10 unique spacers after dereplication were excluded from alignment-based PAM inference; and spacer–protospacer alignments containing more than 4 mismatches or gaps were discarded. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. All analyses were performed using fully deterministic, publicly available code and data; results are therefore fully reproducible.

Reporting summary

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

Data availability

The data generated in this study have been deposited in Zenodo [https://doi.org/10.5281/zenodo.17855072][58]. These data include all mined Cas9 protein sequences, alignment-inferred PAM profiles (CRISPR-PAMdb), CICERO-predicted PAM profiles, benchmarking results, and processed ClinVar pathogenic variant analyses (Supplementary Data 2 and Supplementary Data 3). All files are openly accessible with no restrictions. The raw input datasets used in this study were obtained from public repositories. Bacterial and archaeal genomes were obtained from the mOTUs4 database [https://motus-db.org/]. Additional archaeal genomes were downloaded from the European Nucleotide Archive (ENA) (as of October 2024) [https://www.ebi.ac.uk/ena/browser/downloading-data]. Plasmid sequences were obtained from IMG/PR (version 2023-08-08_1) [https://genome.jgi.doe.gov/portal/IMG_PR/IMG_PR.home.html]. Phage sequences were obtained from the sources listed in Supplementary Data 1, all of which are included in the Zenodo deposition, except for one additional large mOTUs-db–derived phage dataset that will be released in the next public mOTUs-db update and is available from the authors upon request. Disease-associated single-nucleotide variants were obtained from ClinVar (version 2025-12-02) [https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/]. Source data are provided with this paper.

Code availability

CRISPR-PAMdb was developed in Python and CICERO was implemented in PyTorch. The complete source code, pipelines, and documentation are available under an open-source license at https://github.com/Schwank-Lab/CRISPR-PAMdb. The exact version of the code used for this publication has been archived in Zenodo [https://doi.org/10.5281/zenodo.17855426][59].

References

1. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
2. Gleditzsch, D. et al. PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. *RNA Biol.* **16**, 504–517 (2019).
3. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
4. Leenay, R. T. & Beisel, C. L. Deciphering, communicating, and engineering the CRISPR PAM. *J. Mol. Biol.* **429**, 177–191 (2017).
5. Collias, D. & Beisel, C. L. CRISPR technologies and the search for the PAM-free nuclease. *Nat. Commun.* **12**, 555 (2021).
6. Silverstein, R. A. et al. Custom CRISPR—Cas9 PAM variants via scalable engineering and machine learning. *Nature* https://doi.org/10.1038/s41586-025-09021-y (2025).
7. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).
8. Kleinstiver, B. P. et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
9. Ciciani, M. et al. Automated identification of sequence-tailored Cas9 proteins using massive metagenomic data. *Nat. Commun.* **13**, 6474 (2022).
10. Ruffolo, J. A. et al. Design of highly functional genome editors by modelling CRISPR–Cas sequences. *Nature* **645**, 518–525 (2025).
11. Nayfach, S. et al. Engineering of CRISPR-Cas PAM recognition using deep learning of vast evolutionary data. Preprint at https://doi.org/10.1101/2025.01.06.631536 (2025).
12. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
13. Dmitrijeva, M. et al. The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. *Nucleic Acids Res.* **53**, D797–D805 (2025).
14. O'Cathail, C. et al. The European Nucleotide Archive in 2024. *Nucleic Acids Res.* **53**, D49–D55 (2025).
15. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas. Loci. *CRISPR J.* **3**, 462–469 (2020).
16. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* **42**, 6091–6105 (2014).
17. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).

18. Rocha, E. P. C. & Bikard, D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biol.* **20**, e3001514 (2022).
19. Martínez-Alvarez, L. & Peng, X. Redefining paradigms in the archaeal virus-host arms race. Preprint at https://doi.org/10.1101/2025.04.20.649705 (2025).
20. Zaayman, M. & Wheatley, R. M. Fitness costs of CRISPR-Cas systems in bacteria. *Microbiology* **168**, 10.1099/mic.0.001209 (2022).
21. Camargo, A. P. et al. IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata. *Nucleic Acids Res.* **52**, D164–D173 (2024).
22. Camargo, A. P. 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.* **51**, D733–D743 (2023).
23. Karvelis, T. et al. Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biol.* **16**, 253 (2015).
24. Tang, L. et al. Efficient cleavage resolves PAM preferences of CRISPR-Cas in human cells. *Cell Regen.* **8**, 44–50 (2019).
25. Ran, F. A. et al. In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).
26. Tan, Y. et al. Rationally engineered *Staphylococcus aureus* Cas9 nucleases with high genome-wide specificity. *Proc. Natl. Acad. Sci. USA* **116**, 20969–20976 (2019).
27. Gao, S. et al. Genome editing with natural and engineered CjCas9 orthologs. *Mol. Ther. J. Am. Soc. Gene Ther.* **31**, 1177–1187 (2023).
28. Kim, E. et al. In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat. Commun.* **8**, 14500 (2017).
29. Yamada, M. et al. Crystal structure of the minimal Cas9 from Campylobacter jejuni reveals the molecular diversity in the CRISPR-Cas9 systems. *Mol. Cell* **65**, 1109–1121.e3 (2017).
30. Gasiunas, G. et al. A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat. Commun.* **11**, 5512 (2020).
31. Boratyn, G. M. et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29–W33 (2013).
32. Wang, K. et al. Structural insights into Type II-D Cas9 and its robust cleavage activity. *Nat. Commun.* **16**, 7396 (2025).
33. Averina, O. A., Kuznetsova, S. A., Permyakov, O. A. & Sergiev, P. V. Current knowledge of base editing and prime editing. *Mol. Biol.* **58**, 571–587 (2024).
34. Porto, E. M., Komor, A. C., Slaymaker, I. M. & Yeo, G. W. Base editing: advances and therapeutic opportunities. *Nat. Rev. Drug Discov.* **19**, 839–859 (2020).
35. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
36. Qi, C. et al. PAMPHLET: PAM prediction homologous-enhancement toolkit for precise PAM prediction in CRISPR-Cas systems. *J. Genet. Genom.* **52**, 258–268 (2025).
37. Hille, F. et al. The biology of CRISPR-Cas: backward and forward. *Cell* **172**, 1239–1259 (2018).
38. Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
39. Pourcel, C. et al. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz915 (2019).
40. Burstein, D. et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).
41. Zünd, M. et al. High throughput sequencing provides exact genomic locations of inducible prophages and accurate phage-to-host ratios in gut microbial strains. *Microbiome* **9**, 77 (2021).
42. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
43. Zielezinski, A. et al. Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nat. Methods* **22**, 1191–1194 (2025).
44. Edgar, R. C. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinform.* **8**, 18 (2007).
45. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinform.* **11**, 431 (2010).
46. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
47. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
48. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
49. Zetsche, B. et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).
50. Schubert, M. S. et al. Optimized design parameters for CRISPR Cas9 and Cas12a homology-directed repair. *Sci. Rep.* **11**, 19482 (2021).
51. Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984. *Nucleic Acids Res.* **13**, 3021–3030 (1985).
52. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
53. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
54. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
55. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long and Short Papers), 4171–4186 (Minneapolis, Minnesota. Association for Computational Linguistics, 2019).
56. DeVries, T. & Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks. Preprint at https://doi.org/10.48550/arXiv.1802.04865 (2018).
57. Guo, E., Draper, D. & Iorio, M. D. Annealing double-head: an architecture for online calibration of deep neural networks. Preprint at https://doi.org/10.48550/arXiv.2212.13621 (2023).
58. FANG, T. et al. Uncovering Cas9 PAM diversity through metagenomic mining and machine learning. Zenodo https://doi.org/10.5281/ZENODO.17855072 (2025).
59. TaoDFang, Feer, L. & Bogensperger, L. Schwank-Lab/CRISPR-PAMdb: V1.0.0. Zenodo https://doi.org/10.5281/ZENODO.17855426 (2025).

Acknowledgements

We thank the Schwank lab for their helpful discussions and feedback throughout the study and the Science IT team at the University of Zurich for the computational infrastructure used for data analysis. We also greatly appreciate the von Mering lab for their thoughtful input and for providing essential computational resources. This work was supported by core funding from ETH Zurich (S.S.), the University Research Priority Program (URPP) Human Reproduction Reloaded of the University of Zurich (L.B. and L.F.), the SNSF grant number 10003518 (L.B.), the Swiss National Science Foundation (SNSF) grant number 310030_185293 (G.S.), and the State Secretariat for Education, Research and Innovation–funded European Research Council Consolidator Grant (SERI-funded ERC-CoG) "GeneRepair" (G.S.).

Author contributions

T.F. performed data curation, formal analysis, methodology, project administration, writing—original draft, and writing—review & editing. L.B. performed formal analysis, methodology, project administration, writing—original draft, and writing—review & editing. L.F. developed the Snakemake pipeline, performed code review and documentation, and contributed to writing—review & editing. A.A. contributed methodology, supervision, and writing—review & editing. V.B. performed data curation and resource provision and contributed to writing—review & editing. Z.B., C.v.M., and S.S. provided supervision and contributed to writing—review & editing. M.K. and G.S. jointly supervised the work, acquired funding, provided resources, and contributed to writing—review & editing.

Competing interests

G.S. is a scientific advisor to Prime Medicine and a scientific cofounder of Nerai Bio. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-69098-5.

Correspondence and requests for materials should be addressed to Michael Krauthammer or Gerald Schwank.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.