Acta Pharmaceutica Sinica B 2026;16(4):2282−2298

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

TOOLS

SIM: Discovery of novel RNA-targeting argonautes by self-iterative learning from scarce data

Shuze Peng[f,†], Feiming Huang[a,b,c,d,†], Nuolan Li[f], Karl Luigi Loza Vidaurre[a,b,c], Luer Chen[f], Yu Yang[f], Jiaying Hu[f], Yanyu Kou[f], Wei He[a,b,c], Shiwei Wang[a,b,c], Lei Shi[a,b,c], Kehao Tao[a,b,c], Bo Sun[g], Xiaoxuan Song[g], Hao Yang[f], Hainan Zhang[h,i], Lin Yang[i], Zixin Deng[f], Yanqiang Han[a,c,d,e,*], Yan Feng[f,*], Qian Liu[f,*], Jinjin Li[a,c,d,e,*]

[a]*National Key Laboratory of Advanced Micro and Nano Manufacture Technology, Shanghai Jiao Tong University, Shanghai 200240, China*
[b]*School of Integrated Circuits (School of Information Science and Electronic Engineering), Shanghai Jiao Tong University, Shanghai 200240, China*
[c]*Inner Mongolia Research Institute, Shanghai Jiao Tong University, Hohhot 010010, China*
[d]*Shanghai Jincheng Technology Co., Ltd., Shanghai 201109, China*
[e]*State Key Laboratory of Biocatalysis and Enzyme Engineering, Hubei Key Laboratory of Industrial Biotechnology, School of Life Sciences, Hubei University, Wuhan 430062, China*
[f]*State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China*
[g]*School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China*
[h]*HuidaGene Therapeutics Co., Ltd., Shanghai 200131, China*
[i]*Institute of Neuroscience, State Key Laboratory of Neuroscience, Key Laboratory of Primate Neurobiology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China*

Received 19 July 2025; received in revised form 24 October 2025; accepted 21 November 2025

*Corresponding authors.
E-mail addresses: hanyanqiang@sjtu.edu.cn (Yanqiang Han), yfeng2009@sjtu.edu.cn (Yan Feng), liuqian1018@sjtu.edu.cn (Qian Liu), lijinjin@sjtu.edu.cn (Jinjin Li).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2026.01.009

KEY WORDS

Prokaryotic argonaute;
Self-iterative learning;
RNA-targeting nuclease;
Data scarcity;
Enzyme discovery;
RNA biotechnology;
RNA modification sensing

Abstract The limited repertoire of experimentally validated RNA-targeting nucleases has constrained both mechanistic studies and the efficient discovery of novel enzymes for RNA biotechnology. This challenge is particularly pronounced for prokaryotic Argonaute (Ago) proteins, where the scarcity of confirmed RNA-targeting members and a lack of clarity regarding RNA specificity determinants hinder systematic exploration. Although machine learning offers a potential solution, its application is often impeded by the scarcity of labeled training data in this field. To address these limitations, we developed the self-iterative hierarchical ensemble model (SIM), which integrates hierarchical ensemble learning with a self-training strategy. This approach bypasses the dependency on large-scale experimental datasets, allowing SIM to iteratively expand its predictive capability from minimal initial labeled data. When applied to prokaryotic Agos, SIM identified six high-confidence RNA-targeting candidates, five of which were experimentally validated (83% success rate). Notably, SIM identified three uncharacterized Agos harboring a novel N-terminal domain, defining a previously unrecognized subclass. Biochemical and *in vivo* validations of *Haloferax profundi* Ago (*Hp*Ago) confirmed its RNA cleavage activity and a distinctive RNA modification-sensing capability. We leveraged this latter finding to develop a rapid, cost-effective method for quantifying modified RNAs. Our study not only expands the repertoire of RNA-targeting tools but also establishes SIM as a generalizable framework for protein function prediction under data-scarce conditions. This work has broad implications for both RNA biotechnology and the application of machine learning in data-limited fields.

1. Introduction

RNA manipulation tools have strengthened the understanding of RNA biology and provide potential methods for clinical diagnosis and therapies, offering increased safety and adaptability compared to current approaches like gene editing and nucleic acid detection[1,2]. The utilization of these RNA manipulation tools is prevalent in various RNA biotechnologies, encompassing knockdown, imaging, base editing, modification editing, and diagnostics, amongst others[3−7]. CRISPR/Cas discoveries significantly contributed towards the formulation of RNA-directed RNA targeting tools for medical diagnoses and therapy[8]. Nonetheless, enhancing the proficiency and optimizing the performance of these RNA-targeting instruments remain viable avenues for future advancement. For example, Cas13's off-target effect causes cytotoxicity and Csm complex's requirement of large-subunits poses challenges for delivery[9]. The pursuit of novel programmable nucleases with RNA targeting abilities could stimulate dynamic advancements in RNA manipulation technologies.

More recently, prokaryotic Argonautes' (pAgos) programmable capability for targeting arbitrary sequences without a recognition motif, and reduced size for efficient delivery, has sparked growing interest[10−17]. As known, eukaryotic Argonaute (eAgo) is a crucial protein component of RNA interference (RNAi) that binds to small RNA guides and facilitates target RNA binding/cleavage reactions *in vivo*[18]. pAgos have diverse catalytic properties, however, most reported pAgos are DNA-targeting Agos, acting on single strand DNA target (ssDNA)[19,20] (Supporting Information Table S1). Given RNA's availability *in vivo* as the single-stranded form, RNA-targeting Agos functioning at moderate temperatures are particularly valuable in RNA manipulation applications[21−23]. It's conceivable that Agos may uncover an RNA targeting capacity akin to Cas enzymes, offering more versatility and compact tools for cutting-edge applications across various RNA manipulations. To the best of our knowledge, it's currently unclear how these pAgos proteins select their target nucleic acids (DNA or RNA) from biochemical-identified RNA-targeting pAgos. For instance, a clade of pAgos including *Pseudooceanicola lipolyticus* Ago (*Pli*Ago) binds both ssDNA and ssRNA targets but specifically cleaves RNA[24]. Based on the structure of *Pli*Ago complex, mutation on residue K565 as a specific spot in the *Pli*Ago clade did not change the target selection. In short, limited information on the RNA-targeting mechanism has hindered the development of RNA-targeting Agos as well as its application.

Despite the significant variability in their amino acid sequences, Agos often adopt similar three-dimensional structures and perform analogous functions in different organisms[25,26]. pAgos are primarily composed of four distinctive domains: the N-terminal, PAZ (Piwi-Argonaute-Zwille), MID (Middle), and PIWI (P-element induced wimpy testis) domains[27,28]. The N-terminal domain is proposed to facilitate the dissociation of cleaved or fragmented target chains from the complex[23,25,29−31]. The PAZ domain is responsible for anchoring the 3'-end of the guide strand. The PIWI domain is composed of a DEDX catalytic tetrad that functions as the catalytic center of Ago proteins. Due to this structural and functional convergence, identifying functionally similar proteins through traditional sequence-based alignment or phylogenetic analyses remains challenging[10,32]. Currently, the identification of pAgos not only challenges computational predictions but also complicates the functional classification, especially for challenges on the complexity of RNA-targeting function. Therefore, it is essential to develop more direct and efficient screening methods that prioritize structural features and functional roles over sequence-level information to identify RNA-targeting Agos.

Machine learning (ML) has emerged as a powerful tool for elucidating the sequence-structure-function relationships of proteins[33]. Previously, researchers explored novel functional proteins by employing unsupervised clustering methods[34,35]. H. Altae-Tran et al.[34] utilized a protein sequence-based clustering method, FLSHclust, to efficiently identify 188 novel CRISPR-Cas systems. Huang et al.[35] discussed a protein structured-based clustering approach that grouped large-scale protein structures to extract new

functional deaminase. Given that clustering methods do not depend on functional labels for proteins, determining the specific functions of each cluster is challenging. Additionally, the presence of functional heterogeneity within clusters further complicates the precise identification of proteins with desired functions[36,37]. In contrast, supervised learning relies on labeled data to establish clear relationships between inputs and outputs, enabling accurate predictions. This dependence means that the quality and quantity of the labeled data are critical for the model's performance. However, the extreme scarcity of experimentally confirmed RNA-targeting Agos hinders the training of robust supervised ML models, as they require diverse and extensive datasets to generalize effectively. Therefore, developing methodologies that can effectively model and generalize from small datasets is crucial for advancing research in fields where experimental data is scarce, such as the study of RNA-targeting Agos.

In this study, we developed a SIM to address the challenges of identifying RNA-targeting Agos, particularly in data-scarce environments. The SIM was initially trained on only 31 experimentally labeled Agos and ultimately achieved high predictive accuracy, with an AUROC of 0.86 in four-fold cross-validation for identifying RNA-targeting Agos. Its efficiency derives from a self-training algorithm that expands the training set by assigning pseudo-labels to unlabeled data. To further enhance experimental efficiency and capture functional diversity, SIM was integrated into a multiscale screening pipeline. Using this pipeline, we identified five RNA-targeting Agos, including a novel class characterized by previously unreported N-terminal structures. One homolog, *Hp*Ago, was experimentally validated for its ability to recognize and cleave mismatched and modified RNA sequences with high specificity. *Hp*Ago demonstrated RNA-related versatility, including its potential for gene regulation and the detection of RNA modifications such as 8-oxo-G with positional accuracy.

2. Materials and methods

2.1. Sequence and structure-based cluster and search

To establish a performance baseline and justify the need for a machine learning-based approach, we systematically evaluated the efficacy of traditional homology-based methods for predicting the functional subtypes of the 31 labeled Agos. The evaluation was conducted using both qualitative clustering and quantitative search-based analyses. For sequence clustering, we used MMseqs2 to group the 31 Agos sequences[38]. For structure clustering, we first predicted the 3D structure of each Ago using ColabFold and then clustered them using Foldseek[39,40]. The resulting clusters from both methods were analyzed for functional coherence by examining the distribution of the four defined functional labels (RNA/DNA targeting at room/non-room temperature) within each cluster.

An all-*vs*-all homology search was performed on the dataset of 31 labeled Agos. For sequence-based analysis, BLASTP was used with an *E*-value cutoff of 0.001. For structure-based analysis, Foldseek was used with an *E*-value cutoff of 0.001. The 31 Agos were divided into four query groups based on their function: RNA-targeting ($n = 15$), DNA-targeting ($n = 16$), room-temperature ($n = 17$), and non-room-temperature ($n = 14$). For each protein in a query group, a search was conducted against the entire database of 31 Agos. The performance of each search was measured using precision and recall. Precision was defined as the fraction of retrieved homologs that shared the same functional label as the query protein. Recall represented the fraction of all proteins belonging to a specific functional class that were successfully retrieved by a query from that same class. The final reported metrics are the average precision and recall calculated across all queries within a functional category.

2.2. Data collection and preprocessing

The training dataset for SIM includes both labeled and unlabeled Agos, derived from literature mining that catalogs details including Ago names, operating temperatures, and nucleic acid-targeting types. The experimentally labeled dataset consists of 31 Agos, with 15 Agos that target RNA categorized as positive samples and 16 Agos that target DNA classified as negative samples. For operating temperature classification, 17 Agos that operate at room temperature are labeled as positive samples, and 14 Agos that operate at non-room temperature are classified as negative samples. A total of 937 unlabeled Agos were identified in the study by Ryazansky et al.[41]. These Agos are differentiated into 367 long Agos and 570 short Agos based on their protein sequence lengths. Long pAgos possess a modular structure that closely resembles their eukaryotic counterparts. Their sequence contains four key domains: N-terminal domain, PAZ domain, MID domain and PIWI domain. Short pAgos, in contrast, have a more streamlined sequence and are characterized by the absence of the N-terminal and PAZ domains. The short forms are utilized for self-iterative training, whereas the long forms are employed to predict operating temperatures and nucleic acid-targeting types.

To validate the generalizability of SIM on a different protein family also characterized by scarce and imbalanced data, we compiled a dataset of Class 2 CRISPR proteins. We retrieved experimentally annotated Class 2 CRISPR proteins from the CasPEDIA database and filtered them using the "Nuclease Activity" criterion[42]. This resulted in a labeled dataset of 29 proteins, consisting of 7 RNA-targeting CRISPRs and 22 DNA-targeting CRISPRs. Consistent with the Ago dataset, RNA-targeting proteins were assigned positive labels, and DNA-targeting proteins were assigned negative labels. Additionally, we downloaded 1341 unlabeled Class 2 CRISPR proteins from UniProt and retrieved their corresponding predicted structures from the AlphaFold Database.

To comprehensively characterize Agos, we employed a multifaceted analytical approach. Using protein sequence analysis, we extracted amphiphilic pseudo amino acids[43] as discerning features, while simultaneously delving into protein structural nuances to derive hierarchical atomic composition features. The hierarchical atomic composition descriptor is a sophisticated method for characterizing proteins at the atomic level, highlighting how the surrounding atomic environment of a target atom affects its chemical properties and overall protein function. This method uses cumulative integration, beginning with the calculation of atom-type frequencies within an initial shell around the target atom and progressively incorporating subsequent shells. Each additional layer contributes to the overall atomic composition, providing a detailed representation that includes both immediate and extended neighborhoods of the target atom. Upon receiving a protein sequence of length N, we applied hierarchical atomic composition for its characterization. Subsequently, we computed the average vector across all amino acids for the entire protein, resulting in a 40-dimensional representation of the Ago protein. The amphiphilic pseudo-amino acid composition (AmPAC), developed by Chou et al.[43],

quantifies amphiphilicity in proteins, which involves the coexistence of hydrophobic and hydrophilic characteristics in amino acid residues. This realization involves incorporating sequence order information and physicochemical properties of amino acids into a pseudo-amino acid composition. AmPAC extends beyond basic composition analysis by incorporating the orientation of amino acids relative to the protein's surface and interior, thereby capturing the nuanced behavior of proteins in aqueous environments. This is crucial for understanding their structure and function. AmPAC generates a 16-dimensional feature representing Ago. These features were constructed using iFeatureOmega with default parameters[44]. Finally, 40-dimensional feature from hierarchical atomic composition and 16-dimensional feature from AmPAC were then concatenated to form a single 56-dimensional feature for each protein. This combined feature, capturing both sequence and structural information, served as the complete input for training models.

2.3. Few-shot learning methods

To benchmark SIM's performance, we compared it against two few-shot learning methods designed to handle data scarcity. The first method involved fine-tuning the pre-trained ESM2-640M protein language model[45]. To adapt this model for our classification task, all layers except the final one were frozen, preventing their parameters from being updated during training. The protein embeddings generated by the final layer were then averaged across the sequence length. This averaged representation was fed into a multi-layer perceptron (MLP) that projected the 1280-dimensional vector down to 640 dimensions and subsequently to a single output dimension for classification. The model was trained using a BCEWithLogitsLoss function with 0.0001 learning rate.

The second method was a mean teacher model, a semi-supervised approach that leverages unlabeled data to improve performance. This framework consists of two models: a "student" and a "teacher". The student model is trained conventionally on the small labeled dataset using a standard classification loss. It is also trained on the unlabeled dataset using a consistency loss, which penalizes differences between the student model's predictions and the teacher model's predictions on the same unlabeled data. The teacher model is not trained *via* backpropagation; instead, its weights are an exponential moving average (EMA) of the student model's weights from previous training steps. This EMA update makes the teacher a more stable and accurate ensemble of the student model over time, providing reliable pseudo-labels for the unlabeled data and guiding the student model to learn more generalizable features.

2.4. Construction and evaluation of SIM

SIM is a hierarchical ensemble model using a self-training strategy. It has two main components: data-level ensemble and model-level ensemble. In the data-level ensemble, individual models are first trained using a four-fold cross-validation approach and then combined. The formula can be expressed as Eq. (1):

$$M_{\text{data}-\text{level}} = \left(\frac{1}{k}\right)\sum_{i=1}^{k} M_i \tag{1}$$

where i represents the current fold and k represents the number of folds. Here, we compared initial accuracies and chose XGBoost and AdaBoost as the baseline models. XGBoost and AdaBoost are implemented through Python's scikit-learn package with default parameters[46]. The model-level ensemble is executed by combining the models' post-data-level ensemble to harness their unique strengths. We applied Bayesian optimization to determine the optimal weight ratios for maximizing the AUROC for tasks involving nucleic acid-targeting type and operating temperature classification. The formula can be expressed as Eqs. (2) and (3):

$$\left(w_{\text{xg}}, w_{\text{ada}}\right) = \arg\max_{w_{\text{xg}}, w_{\text{ada}}} \mathbb{E}_{p(\theta|D)}\left[\text{AUROC}\left(w_{\text{xg}} M_{\text{xgboost}} + w_{\text{ada}} M_{\text{adaboost}}\right)\right] \tag{2}$$

$$M_{\text{model}-\text{level}} = w_{\text{xg}} M_{\text{xgboost}} + w_{\text{ada}} M_{\text{adaboost}} \tag{3}$$

where w_{xg} and w_{ada} are the optimal weights for the XGBoost and AdaBoost models, respectively. Initially, a data-level ensemble was conducted, which involved cross-validating the XGBoost and AdaBoost models. Subsequently, the best-performing, cross-validated XGBoost and AdaBoost models were combined in a model-level ensemble to finalize SIM. In the data-level ensemble, we considered the diversity of cross-validation datasets, hence we averaged the outputs of the models from each fold. When performing a model-level ensemble, we initially selected the optimal cross-validation-ensembled XGBoost and AdaBoost from the iterative process, assigning them an initial weight of 0.5 each. We then applied Bayesian optimization to determine the optimal weight ratios for maximizing the AUROC for predicting the nucleic acid targeting type and operating temperature. Ultimately, the final prediction probability, used to assess the likelihood of Agos targeting RNA and operating at room temperature, is calculated as the weighted sum of the model output probabilities.

In the self-training strategy, pseudo-labels are uniquely added to each training set generated by four-fold cross-validation, enhancing dataset diversity. To incorporate pseudo-labeled samples into the training set, we select the highest probability samples as positive and the lowest as negative. The new training set T_{new} can be represented as Eq. (4):

$$T_{\text{new}} = T + \left\{x_i \mid P(x_i) \geq \theta_{\text{high}}\right\} \cup \left\{x_i \mid P(x_i) \leq \theta_{\text{low}}\right\} \tag{4}$$

where $P(x_i)$ is the probability of x_i being positive, and θ_{high} and θ_{low} are probability thresholds used to select the highest and lowest probability samples. To avoid the repetitive use of training data in predictions, we categorized unlabeled Agos into short and long Agos, utilizing short Agos for self-training and long Agos for prediction. In each iteration, SIM predicts the classification probabilities of unlabeled short Agos. These Agos are then ranked by their probability values in descending order, selecting the highest probability samples as positive and the lowest as negative until achieving balance in the training set. Once balanced, we incrementally expanded the training set by adding one positive and one negative sample per iteration. Through this method, in each iteration, we add different high-confidence pseudo-labeled samples to the training set for the next round of training in the four-fold cross-validation. This iterative process continued until the model's predictions on the unlabeled pool stabilized, defined as the point where the highest prediction probability dropped below 0.5 or the lowest rose above 0.5.

Further screening for high-quality Agos that target RNA and operate at room temperature is conducted using criteria such as pLDDT, structural similarity, and solubility probability. pLDDT is a metric used by AlphaFold2 to assess the confidence in protein structure predictions. Structural similarity is determined by

selecting the maximum value of structural similarity between the predicted Agos and known RNA-targeting Agos. Solubility probability is calculated by integrating results from two solubility prediction software tools: SoluProt[47] and NetSolP[48]. The final solubility probability for each protein is the average of the values predicted by these tools.

2.5. Protein structure prediction and alignment

We employed ColabFold, a variant of AlphaFold2 integrated with MMseqs2, for protein structure prediction[39]. This method combines the advanced protein structure prediction capabilities of AlphaFold2 with the MMseqs2 sequence search tool, enhancing template-based modeling by rapidly and sensitively detecting homologous sequences. We used Foldseek for aligning protein structures, clustering, and conducting similarity searches[40]. The Foldseek provides TM-score to evaluate the structural similarity between two protein structures. The TM-score is robust against the size of the proteins, ensuring that it gives a reliable indication of structural resemblance irrespective of the protein length. This makes it a preferred metric in structural biology for assessing the accuracy of protein structure predictions and comparing structural models.

2.6. Protein expression and purification

The gene encoding *Hp*Ago (WP_058572808.1), *Nt*Ago (WP_006090832.1), *Ne*Ago (WP_097378604.1), *Tom*Ago (WP_076626087.1), *Ls*Ago (WP_058998822.1) and *Plu*Ago (WP_100217698.1) were synthesized *via* codon optimization and subsequently introduced into the pET28a expression vector with an N-terminal tag for expression in *E. coli* BL21(DE3). The 2A mutants referring to the two D in the conservative catalytic residual DEDX that they are mutated into A, including *Hp*Ago D678A D748A, *Nt*Ago D671A D741A, *Ne*Ago D663A D738A, *Tom*Ago D566A D629A, *Ls*Ago D851A D939A, were generated using site-directed mutagenesis. The mutants N-del, and OB-fold domain were synthesized and introduced into the pET28a expression vector with an N-terminal tag for expression in *E. coli* BL21(DE3). Protein expression was achieved by initially transferring *E. coli* BL21(DE3) cells containing the expression plasmid into 10 mL of LB medium supplemented with 50 μg/mL kanamycin, followed by overnight incubation at 37 °C. All the proteins were first heterologous expressed in *E. coli*, and subsequently purified using nickel column affinity, molecular exclusion, and heparin column affinity chromatography, respectively. The following day, the cells were transferred to 1 L of LB medium supplemented with 50 μg/mL kanamycin and incubated at 37 °C until the OD_{600} reached 0.6−0.8. Subsequently, the culture temperature was reduced to 16 °C, and the cells were induced with isopropyl β-D-thiogalactopyranoside (IPTG) at a final concentration of 0.5 mmol/L and then aerated for 14 h at 16 °C. The cells were harvested *via* centrifugation and stored at −80 °C for further purification.

Following centrifugation, the cellular pellet was resuspended in a lysis buffer (20 mmol/L Tris−HCl, 500 mmol/L NaCl, 10 mmol/L imidazole, 2% glycerol, 0.05% Triton X-100, pH 7.5) containing PMSF at a final concentration of 1 mmol/L. The resuspended pellets were crushed for 5 min at 750 bar using a High-Pressure Homogenizer (Gefran, Italy). The resulting lysate solution was centrifuged at 15,000 rpm for 30 min at 4 °C, and the supernatant was collected and added to a Ni-NTA column (Cytiva) that had been pre-balanced using the lysis buffer. The nickel column was then washed using 10 column volumes of wash buffer (20 mmol/L Tris−HCl, 500 mmol/L NaCl, 20 mmol/L imidazole, 2% glycerol, pH 7.5). Finally, proteins were eluted using 10 mL of elution buffer (20 mmol/L Tris−HCl pH 7.5, 500 mmol/L NaCl, 250 mmol/L imidazole, and 5 mmol/L-mercaptoethanol). Subsequently, the imidazole was eliminated using a PD-10 desalting chromatography column (Sephadex G-25M, Cytiva) and then eluted with desalting buffer (20 mmol/L Tris−HCl pH 7.5, 500 mmol/L NaCl, 10% (*v*/*v*) glycerol, and 2 mmol/L DTT). The desalted protein was further purified using 25 mL of Superdex 200 Increase 10/300 GL (Cytiva). The column was then washed and subsequently eluted with buffer A (20 mmol/L Tris−HCl and 50 mmol/L NaCl, pH 7.5). The NaCl in the eluted protein solution was diluted to 50 mmol/L with buffer A, and further purification was performed using a Heparin HiTrap column (Cytiva). The protein was eluted using a 0.05−2 mol/L NaCl gradient. The final *Hp*Ago protein was examined using 15% SDS-PAGE, and its concentration was determined by BCA kit (Yeasen) analysis.

2.7. Single-stranded nucleic acid cleavage assays

The single-strand nucleic acid cleavage activity of *Hp*Ago, *Nt*Ago, *Ne*Ago, *Tom*Ago, *Ls*Ago and *Plu*Ago and their 2A mutants were tested using 45 nt target nucleic acid chains labeled with 5′-FAM and 21 nt 5′phosphate-(5′P-)guide DNA that were synthesized by Sangon Biotech (Shanghai). All experiments were performed at a molar ratio of 4:4:1 (2000 nmol/L Ago: 2000 nmol/L guide: 500 nmol/L target) with 2 mmol/L $MnCl_2$ at 25 °C for 1 h.

The cleavage specificity evaluation of *Hp*Ago was performed using target nucleic acid chains labeled with 5′-FAM and guide nucleic acid chains with 5′phosphate-(5′P-) or 5′hydroxyl-(5′OH-) groups that were synthesized by Sangon Biotech (Shanghai). The cleavage assays were conducted in a reaction mixture containing purified *Hp*Ago, guide nucleic acid chains, and target nucleic acid chains at a molar ratio of 3:5:2 (1200 nmol/L *Hp*Ago: 2000 nmol/L guide: 800 nmol/L target) with 2 mmol/L $MnCl_2$ at 37 °C unless otherwise indicated. After mixing in a reaction buffer (20 mmol/L Tris−HCl, 250 mmol/L NaCl), the sample was incubated at 37 °C for 1 h. Supporting Information Table S1 lists the guide and target oligonucleotide sequences used in these experiments.

To investigate the temperature dependence of *Hp*Ago cleavage of ssRNA at a molar ratio of 4:4:1 (2000 nmol/L *Hp*Ago: 2000 nmol/L guide: 500 nmol/L target) with 1 mmol/L $MnCl_2$, the samples were subjected to a range of temperatures from 10 to 50 °C for 1 h using a polymerase chain reaction thermocycler (T100, Bio-Rad). Additionally, the effects of divalent cations ($MnCl_2$, $MgCl_2$, $CaCl_2$, $CoCl_2$, $CuCl_2$, and $ZnCl_2$) on the cleavage activity of *Hp*Ago were analyzed by incubating different samples at 37 °C for 1 h at a molar ratio of 4:4:1 (2000 nmol/L *Hp*Ago: 2000 nmol/L guide: 500 nmol/L target) with 1 mmol/L divalent cations.

The optimal length of 5′P-guide DNA/5′P-guide RNA adopted by *Hp*Ago was determined in the presence of 10−30 nt guide DNA/guide RNA targeted with 45 nt ssRNA at a molar ratio of 3:5:2 (1200 nmol/L *Hp*Ago: 2000 nmol/L guide: 800 nmol/L target) with 2 mmol/L $MnCl_2$ at 37 °C for 1 h. The guide-target mismatch tolerance was assessed by introducing a single and double mismatch at positions 1−16 of guide DNA/guide RNA targeted with 45 nt ssRNA at a molar ratio of 3:5:2 (1200 nmol/L *Hp*Ago: 2000 nmol/L guide: 800 nmol/L target) with 2 mmol/L $MnCl_2$ at 37 °C for 1 h.

Analysis of the effects of RNA modifications was performed using synthesized RNA targets containing modified nucleotides (ACCURATE BIOTECHNOLOGY CO.LTD) and a set of 16 nt guides (Sangon Biotech) for indicated time intervals at 37 °C for 1 h at a molar ratio of 3:5:2 (1200 nmol/L *Hp*Ago: 2000 nmol/L guide: 800 nmol/L target) with 2 mmol/L $MnCl_2$.

In all cases, the reaction was quenched by adding loading buffer (95% formamide, 0.5 mmol/L EDTA, 0.025% bromophenol blue, and 0.025% xylene cyanol FF), and the cleavage products were resolved by 16% polyacrylamide gel electrophoresis with urea and then stained with GelRed dye (Invitrogen). The cleavage products were visualized using a Tanon-3500 Gel Imaging System and analyzed using ImageJ and GraphPad Prism 8.0.

2.8. Procedures of the HpAgo-based 8-oxo-G detection assay

*Hp*Ago and 5′P- guide DNA for 22 nt Target were incubated at a molar ratio of 3:5 (600 nmol/L *Hp*Ago: 1000 nmol/L guide) with 1 mmol/L $MnCl_2$ at 25 °C for 15 min. After mixing in a reaction buffer (20 mmol/L Tris-HCl, 50 mmol/L NaCl), the samples containing a mixture of 22 nt 8-oxo-G modified Target and 22 nt unmodified Target in various specific ratios, ranging from 100% 8-oxo-G modified to 0% modified RNA (total concentration: 100 nmol/L) were incubated at 37 °C for 1 h.

The typical ligation reaction mixture, included 50 nmol/L Probe L, 50 nmol/L Probe R, 1x SplintR reaction buffer (NEB). This mixture, along with the cleaved RNA sample, was first denatured at 85 °C for 1 min and then slowly cooled down to room temperature. Then, SplintR Liagse (NEB) was added and incubated for 20 min at 30 °C. PrimerSTAR (Takara) was used in the PCR stage with 1 μmol/L Primer F/R. 1 × LAMP Fluorescent Dye (NEB) was added to observe the amplification product in real time.

2.9. Electrophoretic mobility shift assay (EMSA) of OB-fold domain

For analysis of ssDNA, and ssRNA binding by the electrophoretic mobility shift assay (EMSA), 500 nmol/L nucleic acid substrates (45 nt target DNA labeled with 3'-FAM, or 45 nt target RNA labeled with 3'-FAM) were mixed with increasing concentrations of OB-fold domain (0, 100, 500. 1000, 1500, 2000, 3000, 4000, 5000 nmol/L) in binding buffer (20 mmol/L Tris-HCl (pH 7.5), 50 mmol/L NaCl, 20 μmol/L EDTA) and incubated for 15 min at room temperature. The samples were mixed with 5 × loading buffer (5 × TBE, 50% glycerol, 0.005% Bromophenol Blue), loaded onto a 1.5 mm thick non-denaturing polyacrylamide gel (6%) in 0.5 × TBE, run at constant voltage (100 V) for 120 min in an ice-cold bath, and visualized by phosphor imaging (Tanon 3500BR).

2.10. Single-molecule FRET assay and data analysis

For smFRET experiments, HPLC-purified DNA and RNA oligonucleotides with or without labels were used to construct the templates and were purchased from ACCURATE BIOTECHNOLOGY, China, respectively. The sequences of the oligonucleotides are listed in Supporting Information Fig. S2.

SmFRET experiments were performed using TIRF microscopy. Utilizing an Andor EMCCD camera, fluorescence images were obtained with an integration time of 100 ms (10-Hz frame rate). Quartz slides and coverslips (Fisher Scientific, USA) for the construction of the flow chamber were first coated with polyethylene glycol (PEG). Next, the coverslip was treated with amino silane and coated with a mixture of 99% mPEG (m-PEG-5000, Laysan Bio, Inc.) and 1% biotin-PEG (biotin-PEG-5000, Laysan Bio, Inc.). Next, the chamber was incubated with streptavidin (10 μg/mL) for 5 min. The Cy3-labeled ssDNA and ssRNA targets at a concentration of 20 pmol/L were introduced into the chamber and immobilized on the quartz surface. After immobilization, free substrates were washed out of the chamber using 200 μL of the reaction buffer (20 mmol/L Tris-HCl (pH 7.5), 250 mmol/L NaCl, 5 mmol/L $MnCl_2$, 0.8% D-glucose, 1 mg/mL glucose oxidase, 0.4 mg/mL catalase, and 4 mmol/L Trolox). 1 μmol/L Cy5-labeled 5'P-guide DNA was incubated with 1 μmol/L *Hp*Ago and N-del mutants were incubated in the reaction buffer (20 mmol/L Tris-HCl (pH 7.5), 250 mmol/L NaCl, 5 mmol/L $MnCl_2$) for 20 min to form the binary complex. Then, 1 nmol/L binary complex of *Hp*Ago or N-del mutant diluted by Deaerating buffer (0.8% D-glucose, 1 mg/mL glucose oxidase, 0.4 mg/mL catalase, and 4 mmol/L Trolox) was injected into the chamber before data acquisition. All images were recorded with an exposure time of 100 ms for 1000 frames. Each frame was further processed to extract single-molecule fluorescence intensities. Only fluorescence spots in the acceptor channel were used for analysis to avoid missing or inactivating the acceptors. The FRET efficiency of a single molecule was approximated as Eq. (5):

$$E=\frac{\mathrm{IA}}{\mathrm{ID}+\mathrm{IA}} \tag{5}$$

where ID and IA are the donor and acceptor's background and leakage-corrected emission intensities, respectively. Each experiment was performed at least three times to ensure reproducibility. Single-molecule trajectories were collected using SPARTAN 3.7.0.

2.11. Analysis of cell growth

The real-time course of cellular growth within an 18−24 h period was evaluated using an automatic biological growth-monitoring reactor (Biosan). *E. coli* BL21(DE3) cells with the expression plasmid pET28a-*Hp*Ago or an empty vector were used. The strains were cultivated overnight in LB with 50 mg/mL kanamycin, and subsequently inoculated into 30 mL of fresh LB medium that had been supplemented with 50 mg/mL kanamycin at a ratio of 1%. Upon reaching an OD_{600} of 0.4, IPTG was added at a final concentration of 0.5 mmol/L was performed. Growth curves were plotted using data from three independent biological replicates.

2.12. Plasmid loss assays

To conduct experiments examining plasmid loss, *E. coli* BL21 (DE3) strains containing or lacking the expression plasmid pET28a-*Hp*Ago were co-transformed with the pBluescript II SK (+) plasmid. Following this, several colonies of transformed cell cultures from LB plates were placed into liquid LB medium supplemented with 50 mg/mL kanamycin or 50 mg/mL kanamycin and 50 mg/mL ampicillin and then cultured at 37 °C until an OD_{600} of 0.5 was attained. After the addition of 0.5 mmol/L IPTG, the cell culture was passaged at 12 h intervals under identical conditions. After five passages, the percentage of plasmid loss was determined by plating the cultures on a solid medium with either kanamycin or both kanamycin and ampicillin.

2.13. Western blot analysis

The cells were cultured and lysed in RIPA buffer. Protein concentration was determined using the BCA protein assay. Protein lysates were mixed with the loading buffer (Beyotime, China) and heated at 100 °C for 10 min. Samples (12 μg) were loaded onto SDS-PAGE gels and transferred to nitrocellulose membranes. Next, the membranes were blocked with 5% non-fat milk for 1 h at room temperature and then incubated with primary antibodies (1:1000) overnight at 4 °C. The membranes were washed with TBS containing 0.1% Tween 20 three times for 10 min, incubated with secondary antibodies (1:3000) for 2 h at room temperature, and then washed with TBS containing 0.1% Tween 20 three times. Finally, the protein bands were visualized using NcmECL Ultra Stabilized Per-oxidation Reagent (NCM, China) and an Amersham Imager 600 (GE, USA).

2.14. Cell culture, transfection and flow cytometry analysis

HEK293T cell lines were purchased from the Stem Cell Bank, Chinese Academy of Sciences, and cultured with DMEM (Gibco) supplemented with 10% fetal bovine serum (Gibco), 1% penicillin−streptomycin (Thermo Fisher Scientific) and 0.1 mmol/L nonessential amino acids (Gibco) in an incubator at 37 °C with 5% CO_2. Cells were cultured in T25 flasks (Thermo Fisher Scientific), and consolidation was performed in 24-well plates. The day before transfection, 32×10^5 cells were plated per well in 0.5 mL of complete growth medium. After 12 h, 1 μg Ago plasmids (*Hp*Ago, N-del, *Mbp*Ago), 0.5 μg reporter plasmids and 20 ng DNA guides were transfected into cells with 3.75 μg Polyethylenimine (PEI) (DNA/PEI ratio of 1:2.5) per well. Forty-eight hours after transfection, EGFP expression and mCherry fluorescence were analyzed by BD FACS Aria III, BD LSRFortessa X-20 or Beckman CytoFLEX S. Flow cytometry results were analyzed with FlowJo v.10.5.3.

2.15. RNA extraction and quantitative RT-PCR

Total RNA was extracted using the TRIzol reagent according to the manufacturer's protocol (Invitrogen). The quality and yield of the extracted RNA were analyzed using agarose gel electrophoresis and spectrophotometry, respectively. cDNAs were synthesized from the total RNA (1 mg) using PrimeScript II Reverse Transcriptase (Takara). Then, qRT-PCR was performed using TB Green Prime Ex Taq II and the standard protocols of the Takara 3-step instrument (Roche LightCycler 480). Relative changes in gene expression were calculated using the $2^{-\Delta\Delta C}{}_T$ methods and analyzed by GraphPad Prism 10 software. β-Actin-RNA expression was used as an internal control. The primer pair sequences are listed in Supporting Information Table S2.

2.16. Statistical analysis

All statistical analyses were performed using GraphPad Prism 8.0. Unless otherwise stated, all quantitative data are presented as the mean ± standard deviation (SD) from at least three independent biological replicates, as specified in the figure legends. The statistical significance of differences between two groups was determined using an independent samples two-tailed *t*-test. Linear regression analysis was used to evaluate the correlation between 8-oxo-G content and C_t values. The specific *P*-values are indicated in the figures, with a *P*-value <0.05 considered statistically significant.

3. Results

3.1. SIM exhibits high accuracy and generalization capability

Accurate functional classification of Agos based on sequence or structure alone is a significant challenge (Fig. 1a). High sequence divergence among Agos prevents sequence-based clustering from yielding functionally coherent groups (Supporting Information Fig. S1a). Conversely, the high structural homology within the family confounds structure-based clustering, leading to the grouping of functionally distinct proteins (Fig. S1b). This limitation was quantitatively confirmed by evaluating standard homology search tools. Both sequence- and structure-based approaches performed poorly, achieving average precision and recall scores of approximately 0.5 for predicting nucleic acid targets and optimal operating temperatures—a performance level equivalent to random chance (Fig. S1c−S1f). Moreover, traditional supervised machine learning models are prone to overfitting on such limited datasets, resulting in poor generalization to new candidates. To overcome these challenges, we developed SIM, a framework designed for robust and accurate classification from a minimal set of labeled samples.

SIM's architecture utilizes self-training, a semi-supervised learning technique, to improve generalization from limited labeled data. The model iteratively expands its training set by assigning pseudo-labels to proteins from a pool of 570 unlabeled short Agos (Fig. 1b and c). During this process, the model's performance on nucleic acid target prediction reached maximum AUROC scores of 0.86 (XGBoost) and 0.70 (AdaBoost) (Fig. 1d). For operating temperature prediction, both components achieved an AUROC of 0.98, at iterations 6 and 43, respectively (Fig. 1e). Other metrics, including AUPRC, recall, and precision, also showed significant improvement throughout the self-training iterations (Supporting Information Fig. S2). To optimize the composition of the unlabeled dataset, we investigated mixtures of short and long Agos. The model's classification performance was not significantly affected when up to one-third of the unlabeled pool consisted of long Agos (Supporting Information Fig. S3). However, a 50% mixture, while increasing recall for RNA-targeting, depleted the long Ago pool. Since long Agos are of high interest for downstream gene editing applications and were reserved for the inference set, we opted to use only short Agos for the unlabeled self-training data. The final SIM model was constructed by ensembling the best-performing XGBoost and AdaBoost models from the self-training process using Bayesian optimization. The resulting ensemble weights and detailed performance metrics are provided in Supporting Information Tables S3 and S4.

SIM demonstrated both superior prediction accuracy and broad generalizability, proving highly effective in data-scarce scenarios. The model outperformed traditional machine learning algorithms, such as Decision Tree (DT), AdaBoost, and XGBoost, and established few-shot learning approaches, such as fine-tuning ESM2 and a consistency regularization method called Mean Teacher (MT). This superiority was most pronounced in the RNA-targeting classification task (Fig. 1f and g), where SIM achieved a maximum AUROC of 0.859, along with significant improvements in recall and precision (Supporting Information Fig. S4). To assess its generalizability, we applied SIM to a dataset of class 2 CRISPR

proteins, a different protein family. Using 29 RNA or DNA targeting labeled proteins and 1341 unlabeled proteins for self-training, the strategy again substantially improved classification performance (Supporting Information Fig. S5a and S5b). On this dataset, SIM achieved an average AUROC of 0.904, surpassing the other models and confirming its robustness for predicting protein function from limited data (Fig. S5c).

3.2. Expanding RNA-targeting Agos using SIM

We advanced the identification of RNA-targeting Agos by integrating the SIM into a robust screening pipeline (Fig. 2a). Long pAgos exhibit high structural similarity with eukaryotic Agos (eAgos) and possess specific cleavage functions, highlighting their potential for gene editing and molecular diagnostics as alternative gene manipulation tools[49]. Consequently, we focused on 367 unlabeled long pAgos for prediction. We use short Agos for self-training and long Agos for inferring to ensures that the proteins we ultimately selected for experimental validation were never seen by the model during any training or iterative improvement phase, avoiding any circular reasoning. The pipeline began with applying the SIM model to predict Agos capable of RNA targeting and functionality at room temperature, resulting in 107 candidate Agos. These candidates were subsequently subjected to a stringent filtering process based on three critical criteria: the presence of the complete DEDX (X = N, D, K, or H) motif to ensure endonuclease activity, a predicted local distance difference test (pLDDT) score exceeding 0.7 to guarantee structural reliability, and a solubility probability greater than 0.4 to enhance functional viability. This multi-scale screening narrowed the selection to 39 Agos. We then calculated the structural similarity between these 39 filtered Agos and previously characterized RNA-targeting Agos. To promote diversity within our validation set, we established specific similarity thresholds: we selected two Agos (*Hp*Ago and *Natronorubrum tibetense* Ago (*Nt*Ago)) with maximum TM-scores below 0.8, two Agos (*Pseudomonas luteola* Ago (*Plu*Ago) and *Leptolyngbya* sp. *NIES-2104* Ago (*Ls*Ago)) with TM-scores ranging from 0.8 to 0.9, and two Agos (*Tateyamaria omphalii* Ago (*Tom*Ago) and *Natrinema ejinorense* Ago (*Ne*Ago)) with TM-scores exceeding 0.9. We set 0.8 and 0.9 as cut-offs due to the inherent high structural homology within the Ago protein family, for which the maximum TM-scores between 39 filtered Agos and previously characterized RNA-targeting Agos all fell within a very narrow range of 0.759 to 0.986. This stratified sampling approach resulted in a final set of six Agos, each representing distinct structural characteristics, which were subsequently chosen for experimental validation.

These candidates, along with designing their corresponding variants catalytically inactive with the DEDX motif (Supporting Information Fig. S6), were further biochemically evaluated. Except for *Plu*Ago expressed as insoluble, the other five Agos showed high purity and expected sizes after purification (Supporting Information Fig. S7). We next tested the nuclease activity using fluorescent-labeled nucleic acids as target in cleavage assays. The experiments showed that all purified Agos were RNA-targeting Agos under room temperature, with *Hp*Ago demonstrating the highest efficiency among the five new Agos (Fig. 2a and Supporting Information Fig. S8). Additionally, to elucidate the optimal conditions for *Hp*Ago's catalysis, we evaluated various factors affect cleavage, including working temperature, preferred metal ions, and metal ion concentration (Supporting Information Fig. S9). The results showed that *Hp*Ago had the best cleavage activity at 37 °C. Especially, *Hp*Ago preferred Mn^{2+} and Co^{2+} ions over Mg^{2+} ions.

Structural analysis revealed that three of five identified RNA-targeting Agos—*Hp*Ago, *Nt*Ago, and *Ne*Ago—possess a unique N-terminal structure characterized by a lengthy loop connecting two distinct regions (Fig. 2b). Comparative structural analysis between *Hp*Ago and the previously reported RNA-targeting *Pli*Ago (PDB: 7R8F) revealed that this distinct N-terminal structure is absent in *Pli*Ago (Supporting Information Fig. S10). Subsequently, we searched the PDB using the N-terminal structure of *Hp*Ago and discovered a high degree of structural similarity with the OB domain (PDB: 2KEN), achieving a TM-score of 0.86. To elucidate the functional implications of this unique N-terminal structure, we performed a structural similarity analysis between all OB-fold domains within the SCOP database and the three N-terminal structures from *Hp*Ago, *Ne*Ago, and *Nt*Ago. Structures exhibiting a TM-score greater than 0.6 were selected for heatmap visualization (Fig. 2c). Our analysis revealed that the N-terminal structure bears significant resemblance to domains within the "Nucleic Acid-Binding Proteins" superfamily. Furthermore, cluster analysis indicated a strong association with domains classified under the "Single-strand DNA-binding Domains" family. Based on these findings, we hypothesize that the distinctive N-terminal region may facilitate DNA binding and contribute to related biological functions.

To conclude, SIM can efficiently and accurately validate the RNA-targeting function under room temperature of Agos. Experimental assays confirmed that five new Agos successfully cleaved RNA at room temperature, achieving an accuracy rate of up to 83.3% (5/6, excluding one *Plu*Ago).

3.3. The OB-fold domain at N-terminal regulates RNase activity of HpAgo

For *Hp*Ago with N-terminal OB-fold domain is the first time reported, we purified and obtained such N-terminal structure of *Hp*Ago (Fig. 3a and Supporting Information Fig. S11). The OB-fold domain was highly expressed at a level of 8 mg/L and then was performed with EMSA experiment to verify its nucleic acid binding ability (Fig. 3b). The curve-like band was observed as the OB-fold domain concentration increased, indicating a binary complex formed with ssDNA. In previous research, the OB-fold domain generally functions *via* multimerization rather than as a monomer during its binding process[50]. Hence, we proposed that multiple conformations of the OB-fold domain may coordinate binding with various target concentrations to yield migrate binary profiles. In contrast, no binary complex was observed when using ssRNA, demonstrating OB-fold domain binds to ssDNA over ssRNA.

To elucidate the role of the OB-fold domain on *Hp*Ago's cleavage activity, a N-del mutant lacking N-terminal structure (N-del) was constructed (Fig. 3a and Fig. S11). First, cleavage activity on ssRNA target was analyzed (Fig. 3c). The N-del showed significantly higher ssRNA cleavage activity compared to the wild type when using 5′-P guide DNA (Fig. 3c and Supporting Information Fig. S12a), but at the same cleavage activity when using 5′-P guide RNA (Fig. 3c and Fig. S12b). This result raised the possibility that ssDNA binding capability of OB-fold domain could affect the RNase activity of *Hp*Ago *via* guide binding process. Considering that ssDNA can also be provided as target in cells, we then set the cleavage assays by adding both ssDNA and ssRNA at equal molar amounts as targets. We observed the N-del mutant exhibited similar ssRNA cleavage activity compared with the assay using solely ssRNA substrate, and no ssDNA cleavage

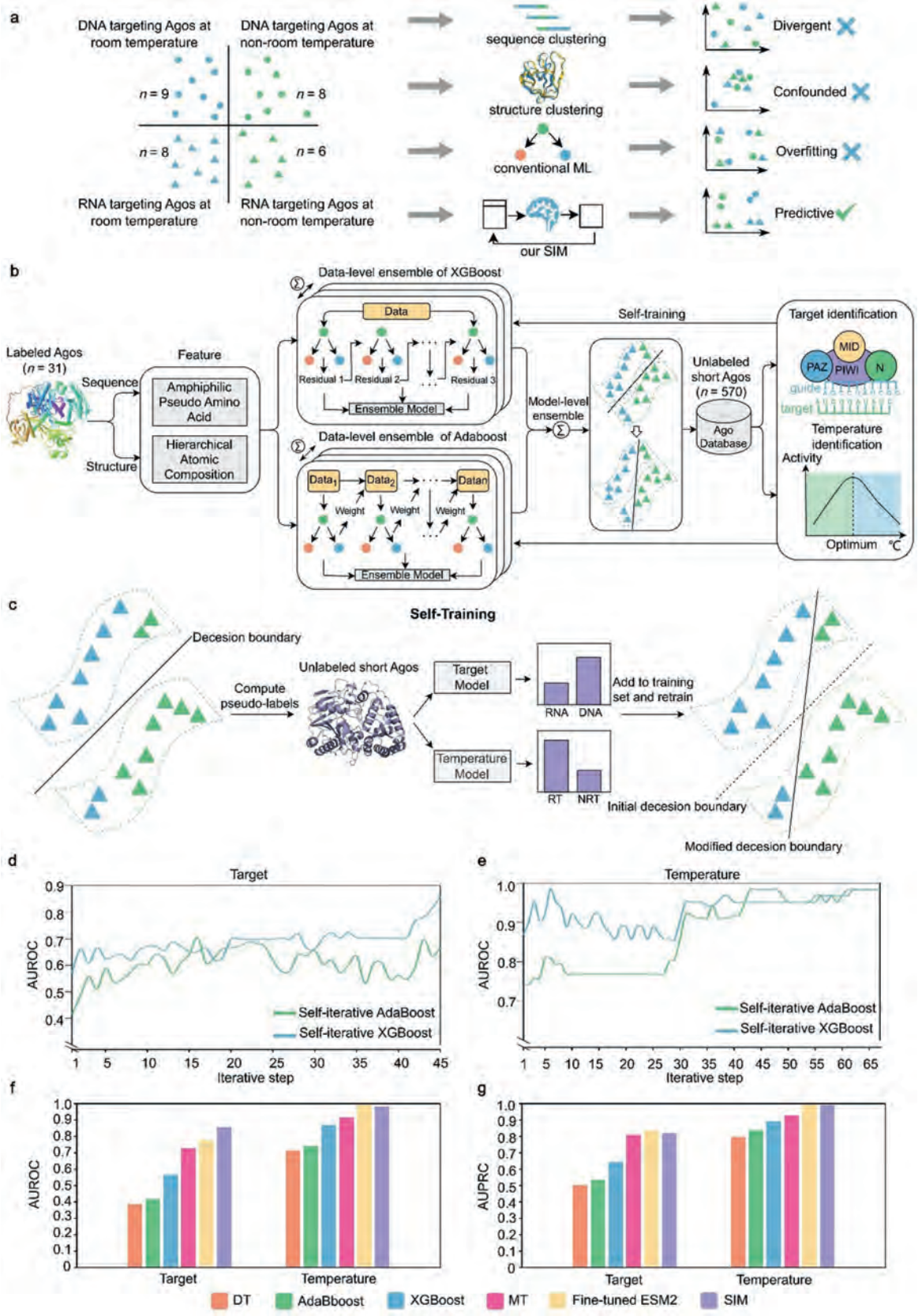

Figure 1 Design and test of SIM. (a) Schematic comparing the functional classification of Agos using different approaches: sequence clustering, structure clustering, conventional ML, and our SIM. (b) The model architecture of SIM. The input protein is encoded at both the structure and sequence levels before being input into the baseline model. Using hierarchical ensemble and self-training strategies, SIM is developed and can be utilized for downstream classification. (c) Schematic of the self-training strategy. The SIM model assigns pseudo-labels to the unlabeled Agos, which are then

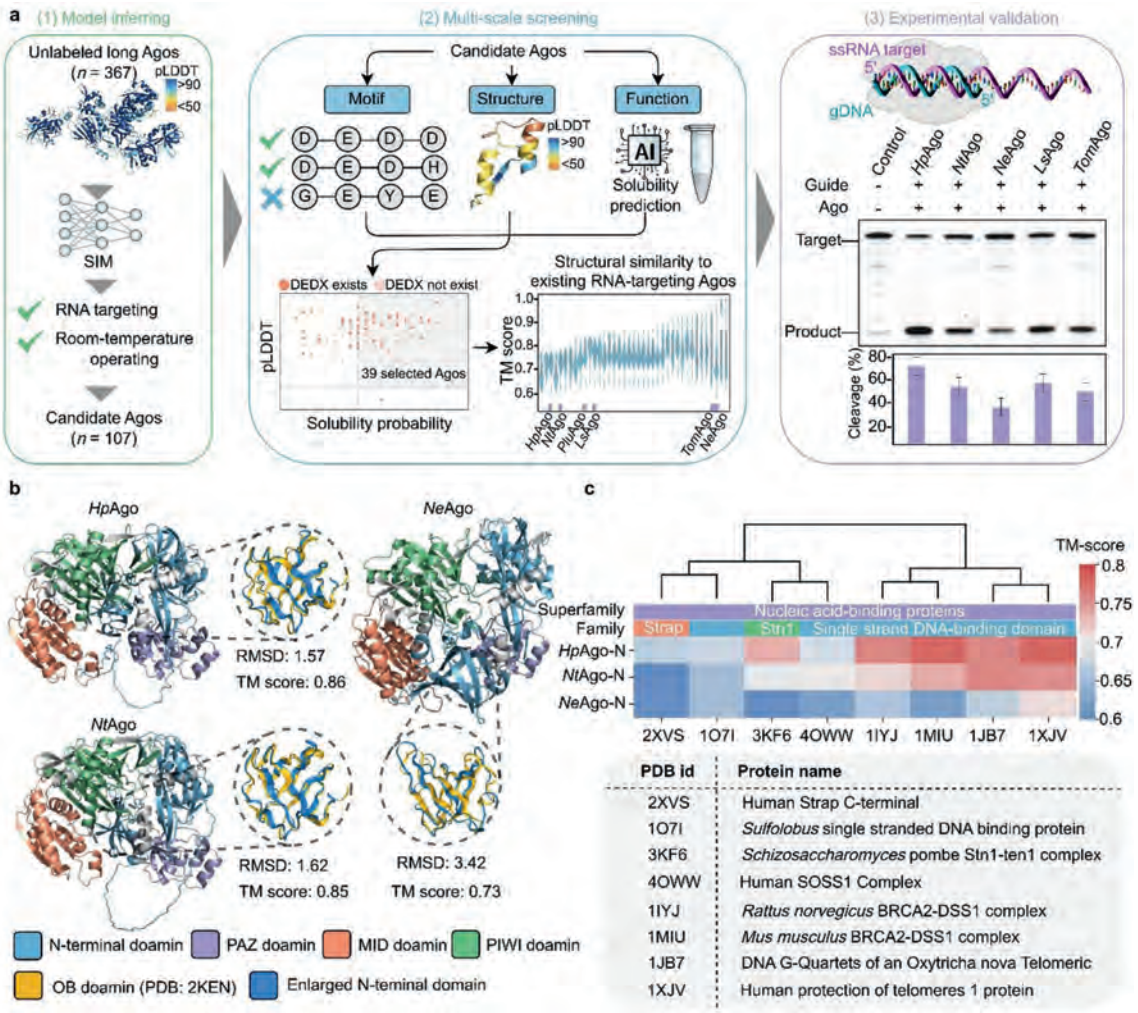

Figure 2 Identification and structural analysis of the new Agos. (a) Schematic diagram of the pipeline for efficiently mining RNA-targeting Agos. The SIM was first employed to select candidate Agos with the desired function. These candidates were then subjected to a multi-scale screening pipeline to identify six structurally and functionally viable Agos. Then, experimental validation of these six Agos confirmed that five possessed RNase activity under room temperature. (b) The structural schematic of *Hp*Ago, *Nt*Ago and *Ne*Ago highlights the unique N-terminal domain, which is enlarged to show five highly coiled antiparallel β-sheets capped by an α-helix, presenting a binding cleft on the opposite side. (c) Analysis of structural similarities between the N-terminal domains of *Hp*Ago, *Nt*Ago, *Ne*Ago, and various OB-fold domains.

was observed as expected, implying OB-fold domain may not coordinate in the target binding process.

To verify the impact of the OB-fold domain, we mixed the OB-fold domain and the N-del mutant at an equal molar concentration for a cleavage assay and observed decreased cleavage activity compared to the N-del mutant alone but higher than the *Hp*Ago WT using guide DNA (Supporting Information Fig. S13a), while *Hp*Ago WT, N-del mutant and mixed OB-fold domain with N-del mutant showed the same cleavage activity using guide RNA (Fig. S13b). These results indicate that the difference in activity between *Hp*Ago and the N-del mutant arises from the OB-fold domain's guide DNA binding affinity, thereby reducing the amount of guide available for *Hp*Ago loading and consequently diminishing its RNA cleavage efficiency.

In addition, to elucidate OB-fold domain participating in guide binding rather than target recognition, we applied a single-molecule fluorescence resonance energy transfer (smFRET) approach that can detect interactions from transient to long-lived in real-time. We planned to observe the binding process on the target nucleic acid by comparing profiles of *Hp*Ago wild type and the N-del mutant to explain the function of OB-fold domain. The fluorophores Cy3 were conjugated to the target as donors, and Cy5 to the guide-Ago complexes as acceptors, respectively (Fig. 3d). Thus, the FRET phenomenon occurred when the guide-Ago complexes bound to the target (Fig. 3e). We fit the survival probability of dwell times of the zero FRET state of ssDNA and ssRNA with a single exponential decay for *Hp*Ago wild type (Supporting Information Fig. S14) and found that *Hp*Ago wild

incorporated into the initial training set for retraining. The initial model's decision boundaries have not been correctly defined. The retrained model can adjust its decision boundaries to better distinguish between different data. (d, e) Performance progression of the self-iterative AdaBoost and XGBoost for predicting nucleic acid target and operating temperature. The AUROC for both classifiers improve significantly across training iterations. (f, g) Comparison of AUROC and AUPRC across different models in the prediction of nucleic acid targeting type and operating temperature.

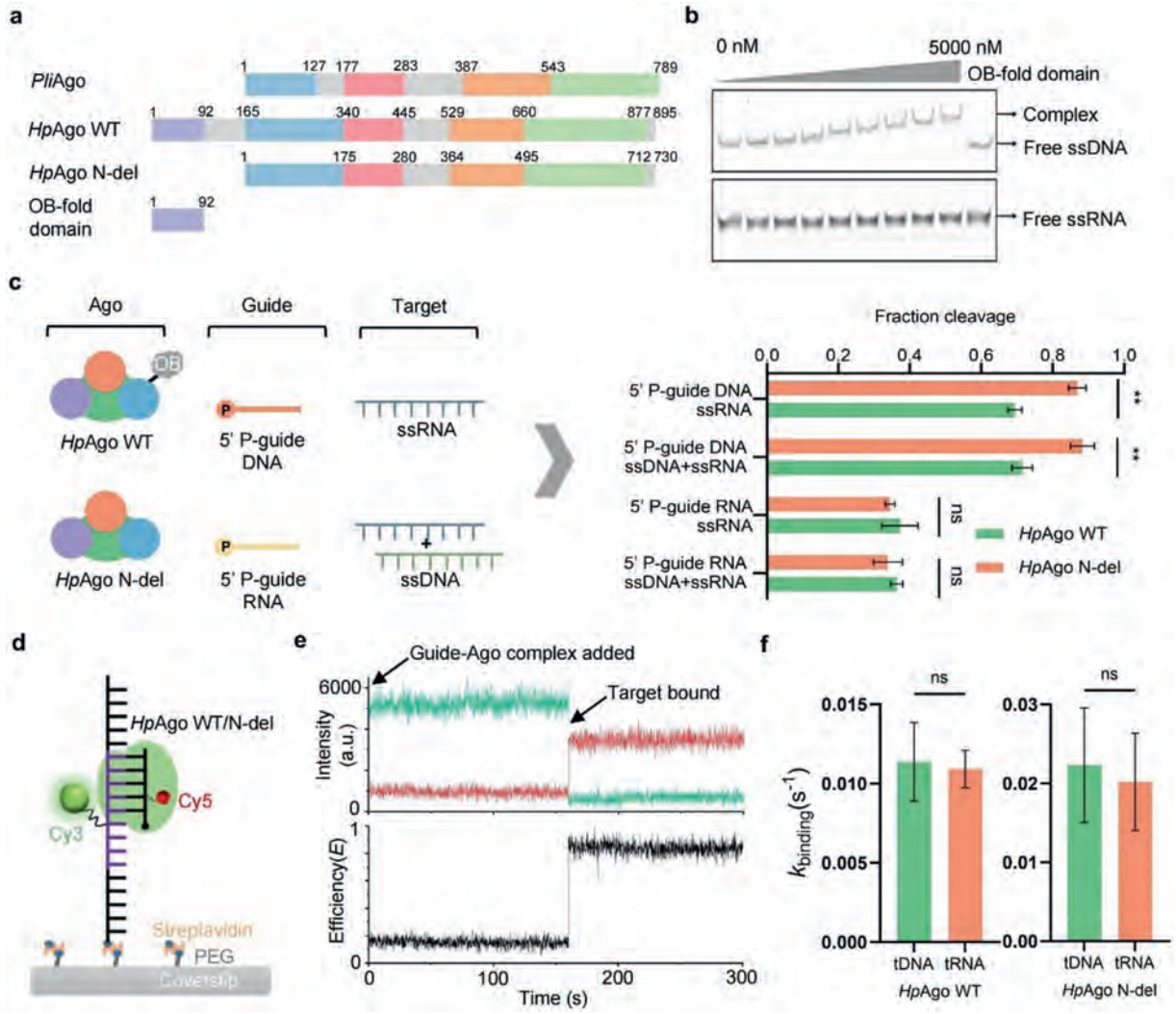

Figure 3 Functional characterization of *Hp*Ago's N-terminal structure. (a) Construction of N-del mutant and OB-fold domain based on domain architecture analysis. (b) Binding affinity of OB-fold domain to ssDNA or ssRNA substrate *via* EMSA experiments. (c) Cleavage assays comparing the activity of N-del and *Hp*Ago on ssRNA substrates alone or mixed with ssDNA, when using DNA and RNA as guides. (d) Schematic diagram of a smFRET experiment. (e) Representative smFRET time trajectories of stably bound Ago-DNA in the presence of 1 nmol/L Ago-DNA in solution. (f) Comparison of binding affinities of ssDNA and ssRNA substrates by *Hp*Ago WT and N-del mutant. The experiment was independently repeated three times, error bars indicate mean ± SD ($n = 3$, biological replicates). $P > 0.05$ is conducted by independent samples *t*-test.

type has a similar preference for binding to both DNA and RNA targets (Fig. 3f). We then analyzed the N-del mutant's preference in the same way, observing the same FRET phenomenon as that of the wild type, further demonstrating that the OB-fold domain has no significant effect on substrate binding.

In general, *Hp*Ago contains a unique N-terminal OB-fold domain with ssDNA binding function, which participates in 5'-P guide DNA binding and subsequently regulates the RNA cleavage activity of *Hp*Ago.

3.4. HpAgo's specificity to RNA modifications apply for detection

Guide-target base-pairing is crucial for the cleavage efficiency and specificity of Agos, which provides the fundamental information for Agos' applications. To assess *Hp*Ago's potential to mediate specific cleavage in various RNA-related technologies, we evaluated the effects of mismatches and target modifications on RNA targets. First, we found *Hp*Ago was able to use a 10−30 nt guide to cleave ssRNA (Supporting Information Fig. S15a−S15c). We then introduced a single mismatched nucleotide at various positions of the guide and found mismatches in the guide seed regions inhibited cleavage (Fig. S15d). When dinucleotide mismatches were introduced, *Hp*Ago exhibited heightened sensitivity to mismatches, particularly in m4m5 and m14m15 (Fig. 4a and b and Fig. S15e).

To further investigate *Hp*Ago's ability to sense RNA modifications, we used RNA targets containing various modifications. For each modified target, we designed a series of guide DNAs to position the modified RNA at different positions relative to the 10-11 cleavage site. *Hp*Ago showed significant specificity for RNA modifications 2'-OMe-G and 8-oxo-G (Fig. 4c and d),

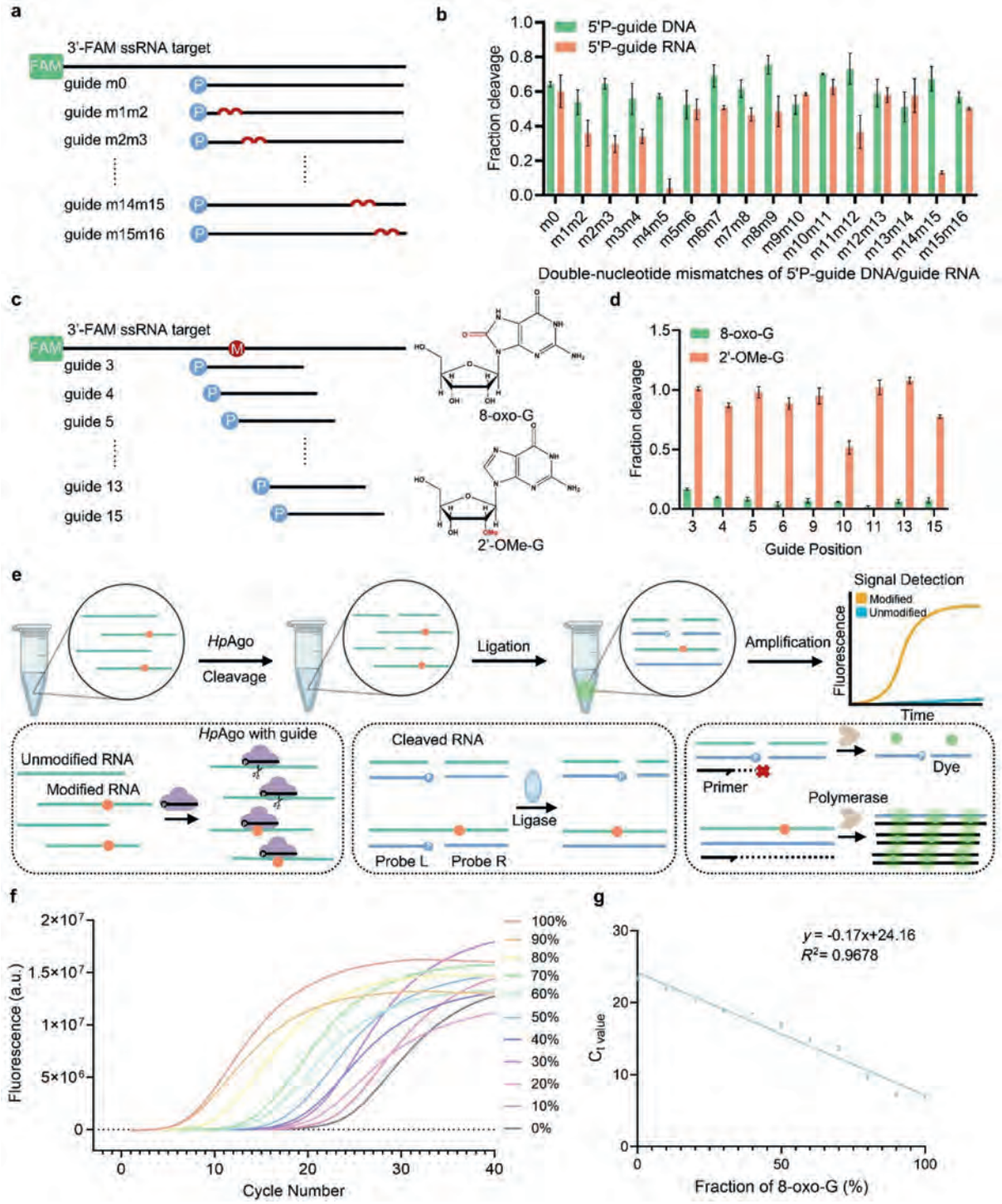

Figure 4 The impact of guide on cleavage activity tested with RNA targets carrying mismatches or modification. (a, b) The impacts of double-base mismatches between the guide and target on *Hp*Ago cleavage activity. Means and standard deviations from three independent experiments are shown. (c, d) Modified nucleotide residues with detected effects on the target RNA cleavage. The target RNA sequence is shown on the top (M, indicated as modified nucleotide), the guides are designed with various paired positions to modified target. (e) Schematic diagram for detecting RNA targets modified with 8-oxo-G. (f) Real-time fluorescence amplification curves showing the detection of various ratios of 8-oxo-G by mixing 22 nt 8-oxo-G modified target with unmodified Target. (g) The linear correlation of the threshold cycle (Ct) value with the samples of various ratios of 8-oxo-G target and quantified by PCR assays.

while the other two modifications only weakly affected RNA cleavage (Supporting Information Fig. S16). Specifically, the presence of 8-oxo-G strongly affected the cleavage activity of *Hp*Ago, especially the guide designed as the modified RNA positioned from the 5th to the 13th site. Although 2'-OMe-G is better tolerated by *Hp*Ago, it can still be sensed, causing a decreased cleavage activity to 50% when the modified RNA is positioned at the 10th position. Compared to the results using mismatched guides, the cleavage effect of the 8-oxo-G modified target is more similar to that of a double-nucleotide mismatched guide. This implies that the 8-oxo-G sensing mechanism may be related to the continuous disruption of Watson-Crick pairing, thus preventing Hoogsteen pairing from compensating for cleavage activity.

For detecting 8-oxo-G RNA modifications, O8G-seq is currently the only commercial method using antibody enrichment followed by high-throughput sequencing[51], which causes low specificity, complex operation, and high costs. Considering the distinguishing ability of *Hp*Ago for 8-oxo-G modified RNA, we developed *Hp*Ago-based detection for 8-oxo-G modification at single-base resolution. The principle of the method mainly relies on *Hp*Ago's cleavage on non-modified RNA to retain the modified RNA, then allow the ligase-assisted PCR to accumulate the modified RNA, finally raise the fluorescent signal as readout (Fig. 4e). Due *Hp*Ago's high substrate specificity between non-modified and 8-oxo-G modified RNA target, the non-modified RNA was cut and unable to accomplish ligation with paired two probes in the following ligase-assisted PCR process, while the modified RNA was amplified efficiently and monitored easily by fluorescence. To assess the feasibility of the method, we prepared various mixed samples containing 22 nt RNA with various ratios of 8-oxo-G and unmodified form (ranging from 0% to 100% modified RNA, Fig. 4f). The amplification curves were monitored in real-time, and the C_t values were measured as the quantitative value. There is a linear correlation between the 8-oxo-G contents and the C_t values (Fig. 4g), suggesting that the *Hp*Ago would offer a relative quantitative evaluation for 8-oxo-G modification RNA. Notably, the whole procedure is quite convenient and easily operated and can be finished within 2 h, which is obviously faster and more economical than the available O^8G-seq.

3.5. HpAgo represses targets expression in cells

The eAgo system is known for its RNA targeting and involvement in the RNAi pathway, while pAgos functions to help host cells defend against invading DNA[11,26]. Previous studies have indicated that the expression of *Km*Ago and *Cb*Ago does not significantly affect cell growth but decreases plasmid content and/or plasmid transformation efficiency in host cells[22,52]. Therefore, we investigated whether the RNA-targeting *Hp*Ago might play a cellular role similar to RNAi-related eAgo or other reported pAgos. We first evaluated the expression of *Hp*Ago in *E. coli*'s growth and its effect on exogenous plasmid invasion. Our results showed that expressing *Hp*Ago exhibited no toxicity to the *E. coli* host (Supporting Information Fig. S17). Additionally, *Hp*Ago expression did not increase the proportion of plasmid-free cells after five passages (Fig. 5a and b), suggesting that *Hp*Ago does not participate in DNA-related host defense mechanisms in the *E. coli* system.

We then tested the RNA-targeting function of *Hp*Ago in cells. We constructed *Hp*Ago with an N-terminal nuclear localization signal (NLS) and optimized it with human codon usage. Using a HEK293T cell line, we performed transient co-transfection with Enhanced Green Fluorescent Protein (EGFP) as the target and measured the relative target expression level *via* flow cytometry (Fig. 5c). We first co-transfected the reporter plasmid, *Hp*Ago expression plasmid, and matched guide DNA into cells, using non-specific guide DNA and the absence of guide DNA as controls. The *Hp*Ago coupled with targeted guide DNA showed a significant decrease in fluorescence intensity compared to controls with non-specific guide DNA or no guide DNA (Fig. 5d). Unexpectedly, the fluorescence intensity of cells transfected with *Hp*Ago also decreased to that of control with the expression vector, which we speculated maybe the unknown mechanism of guide generation in the cell.

We also compared the RNA-targeting capabilities of the N-del mutant of *Hp*Ago and the well-characterized mesophilic RNA-targeting Ago (*Mbp*Ago, WP_008504757.1, *Mucilaginibacter paludism*)[53] in cells (Supporting Information Fig. S18) by assessing their ability to silence EGFP expression. Compared to the N-del mutant, the *Hp*Ago wild type exhibited significantly stronger repressing of EGFP expression, suggesting that the unique OB-fold domain of *Hp*Ago may play a crucial role in RNA silencing. In contrast, *Mbp*Ago showed no discernible knockdown effect. Surprisingly, *Hp*Ago showed a similar repressing efficiency compared with shRNA. We also validated the flow cytometry results using RT-qPCR (Supporting Information Fig. S19). The results indicated that both *Hp*Ago and the N-del mutant reduce EGFP expression by decreasing the mRNA levels. Western blot analysis was performed to confirm that *Hp*Ago, N-del, and *Mbp*Ago were properly expressed in cells at comparable levels, thereby excluding the possibility that differences in target EGFP repression resulted from abnormal protein expression (Supporting Information Fig. S20). These results demonstrate that *Hp*Ago holds greater potential for RNA editing applications in mesophilic RNA-targeting Agos family.

4. Discussion

Our study presents two primary contributions to the field: first, the development and validation of SIM, a novel machine learning framework designed to accurately predict protein function from extremely limited data; and second, the application of this framework to discover and characterize a new class of RNA-targeting Agos, exemplified by *Hp*Ago. The synergy between these two contributions is central to our work, as the significant biological discovery serves as a powerful, real-world validation of our computational method's efficacy.

The primary methodological advance of our work is the SIM pipeline. Progress in functional genomics is often hampered by the scarcity of experimentally validated proteins, creating a "low-data" barrier for traditional supervised machine learning. Our findings demonstrate that SIM effectively overcomes this challenge. By employing a self-training strategy on a minimal set of only 31 labeled Agos, SIM achieved a high predictive accuracy and a remarkable 83.3% success rate in experimental validation. Importantly, SIM's performance surpassed that of established few-shot learning approaches, highlighting its superior design for data-scarce scenarios. Furthermore, its successful application to the CRISPR-Cas dataset demonstrates that its utility is not confined to Agos, establishing SIM as a generalizable framework for

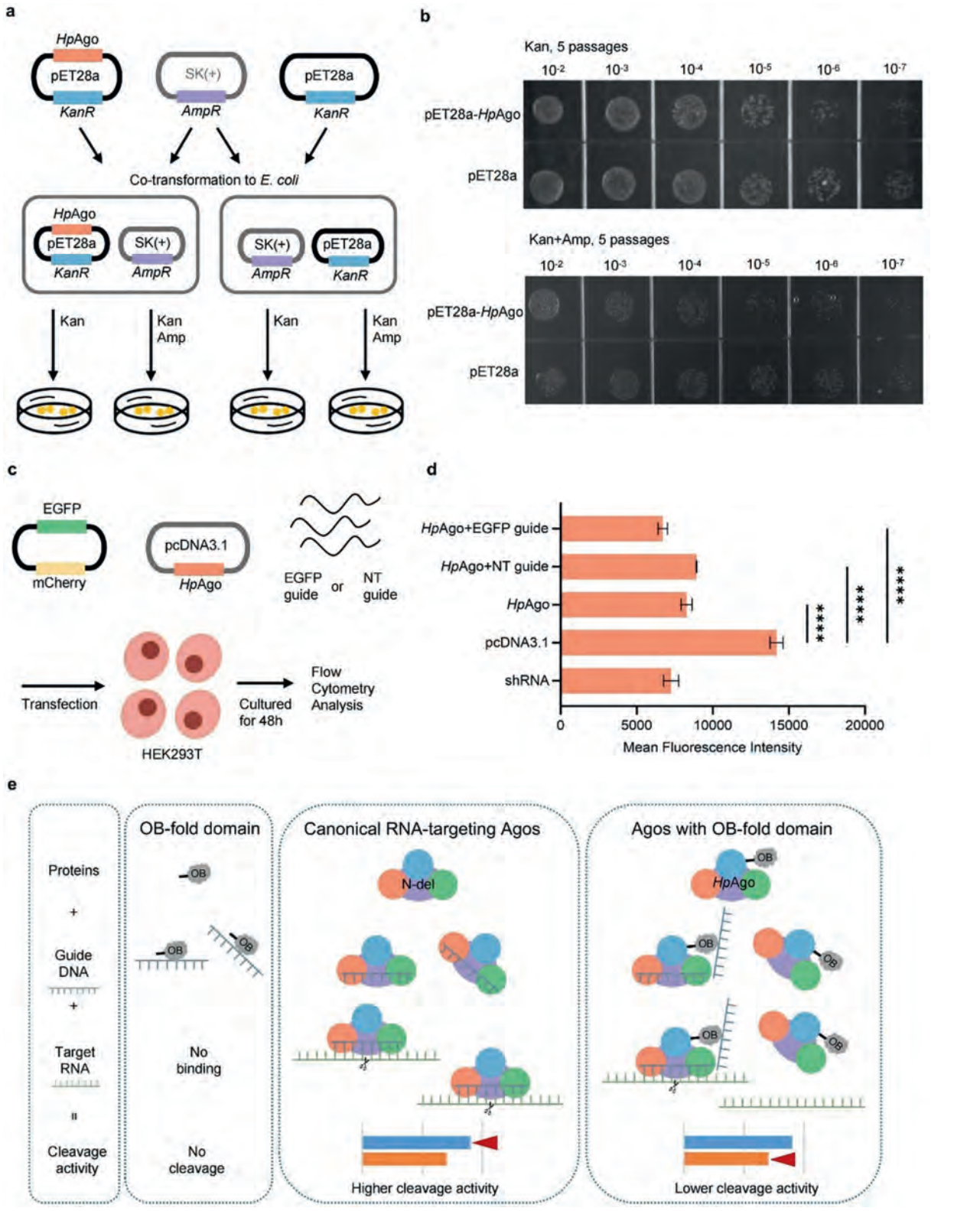

Figure 5 The role of *Hp*Ago in target knockdown. (a) Flowchart of plasmid maintenance experiment. (b) *E. coli* BL21(DE3) was co-transformed with pBluescript SK (+) (AmpR) and *Hp*Ago expression plasmid (KanR) or empty vector pET28a (KanR). The cells were cultivated in Luria-Bertani (LB) medium supplemented with kanamycin and Isopropyl β-D-Thiogalactoside (IPTG) with or without ampicillin and incubated at 37 °C for 5 passages. After the final passage, serial dilutions of bacterial cultures were plated on LB agar dishes with antibiotics to compare the number of cells containing both plasmids. (c) A schematic diagram for *in vivo* function of *Hp*Ago. (d) Flow cytometry analysis measured EGFP fluorescent expression level. The experiment was independently repeated three times, error bars indicate mean ± SD ($n = 3$, biological replicates). ****$P < 0.0001$ is conducted by independent samples *t*-test. (e) Model illustrates that OB-fold domain regulates Agos' cleavage activity *via* binding guide.

accelerating functional discovery in other protein families where labeled data are limited.

Among the predicted Agos, a class characterized by previously unreported N-terminal OB-fold domains has been confirmed to possess RNA-targeting functions through both biochemical assays and *in vivo* tests. In vitro assays and single-molecule FRET (smFRET) experiments revealed that mutants lacking the OB-fold domain showed no change in target binding but exhibited increased RNA cleavage activity. Moreover, this enhanced reactivity was reduced upon the addition of higher amounts of guide DNA, suggesting that the OB-fold domain may regulate *Hp*Ago activity by binding guide single-stranded DNA (ssDNA) (Fig. 5e). In a previous study, *Ng*Ago's N-terminal domain was characterized as the repA domain and was involved in DNA nicking cleavage. The fusion of the extra N-terminal domain with the canonical pAgo architecture has raised interest in naturally evolved domain recombination to confer novel functions and has prompted the exploration of uncharacterized proteins possessing such specific extra domains.

*Hp*Ago demonstrates precise recognition of RNA modifications, offering significant potential for applications in both *in vitro* nucleic acid detection and *in vivo* gene editing and RNA modification. Its sensitivity to RNA modifications, particularly 8-oxo-G, enables accurate identification of RNA oxidative damage in nucleic acids, allowing single-base resolution detection without sequence limitations. Leveraging these catalytic properties, we have developed, for the first time, an *Hp*Ago-based method for detecting modified RNA without the need for antibody enrichment or deep sequencing. This approach greatly reduces detection costs and improves accuracy, holding tremendous clinical application potential. Furthermore, it can specifically detect 8-oxo-G modified targets in samples and determine their positional information. This method could become a milestone in the development of next-generation RNA modification detection technologies, contributing to fundamental research and potential diagnostic applications related to RNA modifications.

Furthermore, compared to other reported RNA-targeting Agos (*e.g.*, *Mbp*Ago), *Hp*Ago demonstrated a highly effective ability to inhibit target EGFP expression in cells, whereas *Mbp*Ago had no effect. This suggests that *Hp*Ago may possess a unique function in *in vivo* gene regulation and holds potential as an alternative gene editing tool. Although target expression was reduced when the *Hp*Ago plasmid was transfected without a guide, this indicates that plasmid co-transfection may generate a guide through unknown inherent pathways (such as recruiting other proteins) within the complex cellular environment, thereby reducing target EGFP expression. Additionally, *Hp*Ago can accurately detect mismatches and two types of modifications on target RNA, which could be valuable for developing RNA-related diagnostic and therapeutic applications. We believe that these new discoveries and editing efforts will contribute to the development of promising RNA editing technologies, offering advantages such as smaller size, simple guide design, specificity for target mismatches and modification types, and independence from PAM sequences.

5. Conclusions

In this study, we advanced the understanding and application of RNA-targeting Agos by developing a self-iterative hierarchical ensemble Model and a multiscale screening pipeline capable of efficiently identifying RNA-targeting Agos in data-limited environments. This approach led to the successful identification of five RNA-targeting Agos, including *Hp*Ago, which demonstrated unique properties and broad potential for both basic research and practical applications. Notably, we characterized a novel N-terminal OB-fold domain in *Hp*Ago, revealing its regulatory role in RNA cleavage activity and its interaction with guide DNA. Our findings highlight *Hp*Ago's exceptional sensitivity to mismatches and RNA modifications, particularly its ability to detect 8-oxo-G modifications with high specificity and positional accuracy. This capability enabled the development of a novel *Hp*Ago-based RNA modification detection method, offering new avenues for studying RNA modifications and advancing diagnostic technologies. Additionally, *Hp*Ago's ability to inhibit target gene expression and its potential as a gene editing tool further underscore its versatility and promise as an alternative to CRISPR-based systems. These results not only broaden the scope of RNA-targeting Agos but also pave the way for innovative RNA-related diagnostic and therapeutic applications. The compact size, specificity, and flexibility of *Hp*Ago, combined with its independence from PAM sequences, position it as a powerful tool for precise RNA manipulation and editing, contributing to future advancements in RNA biology and biotechnology.

Acknowledgments

Research reported in this publication was supported by the National Key R&D Program of China (No. 2025YFA0922400), the Shanghai Zhangjiang National Independent Innovation Demonstration Zone Special Development Fund Major Project (ZJ2025-ZD-023), the Shanghai Science and Technology Project of China (Nos. 24010700100 and 23JC1402300), the National Natural Science Foundation of China (Nos. 32301040 and 32270051), the Science and Technology Cooperation Program of Shanghai Jiao Tong University in Inner Mongolia Autonomous Region—Action Plan of Shanghai Jiao Tong University for "Revitalizing Inner Mongolia through Science and Technology" (No. 2025XYJG0001-01-07), the Shanghai Jiao Tong University AI for Engineering Initiative, the Shanghai Pilot Program for Basic Research-Shanghai Jiao Tong University (21TQ1400204, China) and the Natural Science Foundation of Beijing Municipality (L234050, China).

Author contributions

Shuze Peng: Validation, Investigation, Writing - Original Draft, Visualization. Feiming Huang: Methodology, Software, Formal analysis, Writing - Original Draft. Nuolan Li: Data Curation, Validation. Karl Luigi Loza Vidaurre: Resources, Visualization. Luer Chen: Validation, Investigation. Yu Yang: Investigation, Visualization. Jiaying Hu: Investigation, Formal analysis. Yanyu Kou: Validation, Formal analysis. Wei He: Software, Formal analysis. Shiwei Wang: Resources, Software. Lei Shi: Formal analysis. Kehao Tao: Data Curation. Bo Sun: Validation. Xiaoxuan Song: Validation. Hao Yang: Investigation. Hainan Zhang: Resources. Lin Yang: Resources. Zixin Deng: Resources. Yanqiang Han: Writing - Review & Editing, Supervision. Yan Feng: Conceptualization, Supervision. Qian Liu: Conceptualization, Resources, Supervision, Funding acquisition. Jinjin Li: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

All data that support the findings of this study are available in the main text and the supporting information. The source code for SIM is available at https://github.com/hfm1026/SIM.

Appendix A. Supporting information

Supporting Information to this article can be found online at https://doi.org/10.1016/j.apsb.2026.01.009.

References

1. Li W, Jiang X, Wang W, Hou L, Cai R, Li Y, et al. Discovering CRISPR-Cas system with self-processing pre-crRNA capability by foundation models. *Nat Commun* 2024;**15**:10024.
2. Liu Q, Guo X, Xun G, Li Z, Chong Y, Yang L, et al. Argonaute integrated single-tube PCR system enables supersensitive detection of rare mutations. *Nucleic Acids Res* 2021;**49**:e75.
3. Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, et al. RNA editing with CRISPR-Cas13. *Science* 2017;**358**: 1019−27.
4. Toudji-Zouaz A, Bertrand V, Barrière A. Imaging of native transcription and transcriptional dynamics *in vivo* using a tagged Argonaute protein. *Nucleic Acids Res* 2021;**49**:e86.
5. Li Y, Tang X, Wang N, Zhao Z, Man S, Zhu L, et al. Argonaute-DNAzyme tandem biosensing for highly sensitive and simultaneous dual-gene detection of methicillin-resistant *Staphylococcus aureus*. *Biosens Bioelectron* 2024;**244**:115758.
6. Wilson C, Chen PJ, Miao Z, Liu DR. Programmable m6A modification of cellular RNAs with a Cas13-directed methyltransferase. *Nat Biotechnol* 2020;**38**:1431−40.
7. Villiger L, Joung J, Koblan L, Weissman J, Abudayyeh OO, Gootenberg JS. CRISPR technologies for genome, epigenome and transcriptome editing. *Nat Rev Mol Cell Biol* 2024;**25**:464−87.
8. Shi P, Wu X. Programmable RNA targeting with CRISPR-Cas13. *RNA Biol* 2024;**21**:1−9.
9. Liu G, Lin Q, Jin S, Gao C. The CRISPR-Cas toolbox and gene editing technologies. *Mol Cell* 2022;**82**:333−47.
10. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin EV, et al. The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 2014;**21**:743−53.
11. Meister G. Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet* 2013;**14**:447−59.
12. Qin Y, Li Y, Hu Y. Emerging Argonaute-based nucleic acid biosensors. *Trends Biotechnol* 2022;**40**:910−4.
13. Song J, Hegge JW, Mauk MG, Chen J, Till JE, Bhagwat N, et al. Highly specific enrichment of rare nucleic acid fractions using *Thermus thermophilus* argonaute with applications in cancer diagnostics. *Nucleic Acids Res* 2020;**48**:e19.
14. Kropocheva EV, Lisitskaya LA, Agapov AA, Musabirov AA, Kulbachinskiy AV, Esyunina DM. Prokaryotic Argonaute proteins as a tool for biotechnology. *Mol Biol* 2022;**56**:854−73.
15. He R, Wang L, Wang F, Li W, Liu Y, Li A, et al. *Pyrococcus furiosus* Argonaute-mediated nucleic acid detection. *Chem Commun Camb Engl* 2019;**55**:13219−22.
16. Shin S, Jung Y, Uhm H, Song M, Son S, Goo J, et al. Quantification of purified endogenous miRNAs with high sensitivity and specificity. *Nat Commun* 2020;**11**:6033.
17. Beskrovnaia M, Agapov A, Makasheva K, Zharkov DO, Esyunina D, Kulbachinskiy A. Sensing of DNA modifications by pAgo proteins *in vitro*. *Biochimie* 2024;**220**:39−47.
18. Wilson RC, Doudna JA. Molecular mechanisms of RNA interference. *Annu Rev Biophys* 2013;**42**:217−39.
19. Swarts DC, Hegge JW, Hinojo I, Shiimori M, Ellis MA, Dumrongkulraksa J, et al. Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Res* 2015;**43**:5120−9.
20. Swarts DC, Jore MM, Westra ER, Zhu Y, Janssen JH, Snijders AP, et al. DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* 2014;**507**:258−61.
21. Graver BA, Chakravarty N, Solomon KV. Prokaryotic Argonautes for *in vivo* biotechnology and molecular diagnostics. *Trends Biotechnol* 2024;**42**:61−73.
22. Liu Y, Li W, Jiang X, Wang Y, Zhang Z, Liu Q, et al. A programmable omnipotent Argonaute nuclease from mesophilic bacteria *Kurthia massiliensis*. *Nucleic Acids Res* 2021;**49**:1597−608.
23. Lee KZ, Mechikoff MA, Kikla A, Liu A, Pandolfi P, Fitzgerald K, et al. NgAgo possesses guided DNA nicking activity. *Nucleic Acids Res* 2021;**49**:9926−37.
24. Lisitskaya L, Shin Y, Agapov A, Olina A, Kropocheva E, Ryazansky S, et al. Programmable RNA targeting by bacterial Argonaute nucleases with unconventional guide binding and cleavage specificity. *Nat Commun* 2022;**13**:4624.
25. Lisitskaya L, Aravin AA, Kulbachinskiy A. DNA interference and beyond: structure and functions of prokaryotic Argonaute proteins. *Nat Commun* 2018;**9**:5165.
26. Wu J, Yang J, Cho WC, Zheng Y. Argonaute proteins: structural features, functions and emerging roles. *J Adv Res* 2020;**24**:317−24.
27. Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin EV, et al. The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 2014;**21**:743−53.
28. Sheng G, Zhao H, Wang J, Rao Y, Tian W, Swarts DC, et al. Structure-based cleavage mechanism of *Thermus thermophilus* Argonaute DNA guide strand-mediated DNA target cleavage. *Proc Natl Acad Sci U S A* 2014;**111**:652−7.
29. Hauptmann J, Dueck A, Harlander S, Pfaff J, Merkl R, Meister G. Turning catalytically inactive human Argonaute proteins into active slicer enzymes. *Nat Struct Mol Biol* 2013;**20**:814−7.
30. Kwak PB, Tomari Y. The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nat Struct Mol Biol* 2012;**19**:145−51.
31. Wang Y, Juranek S, Li H, Sheng G, Tuschl T, Patel DJ. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* 2008;**456**:921−6.
32. Ryazansky S, Kulbachinskiy A, Aravin AA. The expanded universe of prokaryotic Argonaute proteins. *mBio* 2018;**9**:e01935-18.
33. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583−9.
34. Altae-Tran H, Kannan S, Suberski AJ, Mears KS, Demircioglu FE, Moeller L, et al. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* 2023;**382**.
35. Huang J, Lin Q, Fei H, He Z, Xu H, Li Y, et al. Discovery of deaminase functions by structure-based protein clustering. *Cell* 2023; **186**:3182−3195.e14.
36. Ghayoumi B, Fasoli A, Ko CY, Bers DM, Sato D. Post-translational modifications promote functional heterogeneity in cardiac myocytes. *Biophys J* 2021;**120**:151a.
37. Bärthel S, Falcomatà C, Rad R, Theis FJ, Saur D. Single-cell profiling to explore pancreatic cancer heterogeneity, plasticity and response to therapy. *Nat Cancer* 2023;**4**:454−67.
38. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017; **35**:1026−8.
39. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods* 2022;**19**:679−82.

40. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024;**42**:243−6.
41. Ryazansky S, Kulbachinskiy A, Aravin AA. The expanded universe of prokaryotic Argonaute proteins. *mBio* 2018;**9**:e01935-18.
42. Adler BA, Trinidad MI, Bellieny-Rabelo D, Zhang E, Karp HM, Skopintsev P, et al. CasPEDIA Database: a functional classification system for class 2 CRISPR-Cas enzymes. *Nucleic Acids Res* 2024;**52**: D590−6.
43. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinforma Oxf Engl* 2005;**21**: 10−9.
44. Chen Z, Liu X, Zhao P, Li C, Wang Y, Li F, et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Res* 2022;**50**:W434−47.
45. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123−30.
46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825−30.
47. Hon J, Marusiak M, Martinek T, Kunka A, Zendulka J, Bednar D, et al. SoluProt: prediction of soluble protein expression in *Escherichia coli*. *Bioinforma Oxf Engl* 2021;**37**:23−8.
48. Thumuluri V, Martiny H-M, Almagro Armenteros JJ, Salomon J, Nielsen H, Johansen AR. NetSolP: predicting protein solubility in *Escherichia coli* using language models. *Bioinforma Oxf Engl* 2022; **38**:941−6.
49. Hegge JW, Swarts DC, van der Oost J. Prokaryotic Argonaute proteins: novel genome-editing tools? *Nat Rev Microbiol* 2018;**16**:5−11.
50. Lohman TM, Overman LB, Datta S. Salt-dependent changes in the DNA binding co-operativity of *Escherichia coli* single strand binding protein. *J Mol Biol* 1986;**187**:603−15.
51. Seok H, Lee H, Lee S, Ahn SH, Lee H-S, Kim G-WD, et al. Position-specific oxidation of miR-1 encodes cardiac hypertrophy. *Nature* 2020;**584**:279−85.
52. Kuzmenko A, Yudin D, Ryazansky S, Kulbachinskiy A, Aravin AA. Programmable DNA cleavage by Ago nucleases from mesophilic bacteria *Clostridium butyricum* and *Limnothrix rosea*. *Nucleic Acids Res* 2019;**47**:5822−36.
53. Li W, Liu Y, He R, Wang L, Wang Y, Zeng W, et al. A programmable pAgo nuclease with RNA target preference from the psychrotolerant bacterium *Mucilaginibacter paludis*. *Nucleic Acids Res* 2022;**50**:5226−38.