Conditional Protein Denoising Diffusion Generates Programmable Endonucleases

Bingxin Zhou[1,2†], Lirong Zheng[1*†], Banghao Wu[1,3†], Kai Yi[4], Bozitao Zhong[1], Pietro Liò[5*], Liang Hong[1,2,6,7*]

[1]Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, 200240, China.
[2]Shanghai National Center for Applied Mathematics (SJTU center), Shanghai Jiao Tong University, Shanghai, 200240, China.
[3]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China.
[4]School of Mathematics and Statistics, University of New South Wales, , Sydney, 2052, NSW, Australia.
[5]Department of Computer Science and Technology, University of Cambridge, Cambridge, CB3 0FD, UK.
[6]Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai, 200240, China.
[7]Shanghai Artificial Intelligence Laboratory, Shanghai, 200232, China.

*Corresponding author(s). E-mail(s): lirong_zheng@sjtu.edu.cn; pl219@cam.ac.uk; hongl3liang@sjtu.edu.cn;
Contributing authors: bingxin.zhou@sjtu.edu.cn; wubh20@sjtu.edu.cn; kai.yi@unsw.edu.au; zbztzhz@gmail.com;
†These authors contributed equally to this work.

Abstract

Computation or deep learning-based functional protein generation methods address the urgent demand for novel biocatalysts, allowing for precise tailoring of functionalities to meet specific requirements. This emergence leads to the creation of highly efficient and specialized proteins with wide-ranging applications in scientific, technological, and biomedical domains. This study establishes a conditional protein diffusion model, namely CPDiffusion, to deliver diverse protein sequences with desired functions. While the model is free from extensive training data and the sampling process involves little guidance on the type of generated

amino acids, CPDiffusion effectively secures essential highly conserved residues that are crucial for protein functionalities. We employed CPDiffusion and generated **27** artificially designed Argonaute proteins, programmable endonucleases applied for easy-to-implement and high-throughput screenings in gene editing and molecular diagnostics, that mutated approximately $\mathbf{200-400}$ amino acids with **40%** sequence identities to those from nature. Experimental tests demonstrate the solubility of all **27** artificially-designed proteins (AP), with **24** of them displaying DNA cleavage activity. Remarkably, **74%** of active APs exhibited superior activity compared to the template protein, and the most effective one showcased a remarkable nearly nine-fold enhancement of enzymatic activity. Moreover, **37%** of APs exhibited enhanced thermostability. These findings emphasize CPDiffusion's remarkable capability to generate long-sequence proteins in a single step while retaining or enhancing intricate functionality. This approach facilitates the design of intricate enzymes featuring multi-domain molecular structures through *in silico* generation and throughput, all accomplished without the need for supervision from labeled data.

1 Introduction

Deep-learning-assisted functional protein design is an innovative and promising approach that harnesses the power of deep learning models to expedite and enhance the creation of novel proteins [1, 2]. By leveraging deep learning, researchers can generate novel proteins with customized functions tailored to specific criteria [3–5]. The integration of deep learning in generating novel proteins provides opportunities for optimizing protein stability, binding affinity, or enzymatic activity, enabling the creation of proteins with desired properties. Deep learning's capacity to generate and enhance a multitude of potential proteins enables researchers to extend the frontiers of protein engineering and discovery. This expansion in scope offers opportunities for the creation of proteins with unique and valuable functionalities. At present, several prominent deep learning models have been applied in designing novel functional proteins, such as ProteinGAN [6], ProGen [7], and RFdiffusion [8], highlighting their contributions to advancing this promising field of biological research.

As an emerging tool for generating diverse samples from complex distributions [9–13], denoising diffusion probabilistic models (DDPMs) stand out as natural and strong candidates for generating novel proteins, where the identification of dissimilar proteins with desired structures or functions is paramount. Both requirements can be satisfied by the design philosophy of DDPMs, which assembles trainable neural network layers to progressively denoise artificial perturbations of defined noise distribution and reveal the original data distribution [8, 14–18]. Through iterative denoising steps aligned with specific objectives, DDPMs unravel the implicit mapping rules connecting a protein's sequence and structure to its functions. Moreover, the denoising process can be conditioned on specific structural preferences or other characteristics of the protein of interest, guiding the output sampling distribution toward favorable directions. Consequently, a trained DDPM possesses the remarkable capability to generate diverse molecules conditioned a set of desired properties tailored for functional proteins.

In this study, we applied the DDPM methodology to generate novel programmable endonucleases called prokaryotic Argonaute (pAgo) proteins. pAgo protein is a class of endonucleases that play a crucial role in DNA interference in prokaryotes, and has gained significant attention in the field of biotechnology and bioengineering [19–24]. Their remarkable ability to target and cleave specific single-strand DNA/RNA sequences has led to important applications in diagnostics, which enables the design of molecular diagnostic assays for detecting and quantifying nucleic acid sequences associated with pathogens or cancer-related mutations [20, 22]. These assays offer improved early detection and precise treatment of diseases. Moreover, pAgo proteins exhibit high affinity to substrates and specific recognition of target sequences, making them valuable tools for nucleic acid labeling both *in vitro* and *in vivo* [25–27]. Additionally, the combined action of pAgo proteins and helicase holds potential for the development of gene therapies targeting disease-causing genes [28]. Thus, pAgo proteins open up new possibilities in therapeutics and genetic manipulation technologies.

Mesophilic pAgo proteins have been proposed to be a candidate for integration into isothermal nucleic acid detection and RNA editing techniques [24, 29, 30]. However, their widespread applicability is hindered by inherent limitations in exhibiting lower DNA/RNA cleavage activity compared with that of thermophilic pAgo proteins. Thus, strategies to enhance the enzymatic function of mesophilic pAgo proteins should be established. Considering the vast search space of the possible combination of amino acids for the protein, it is desired for the generation tool to sample diverse protein sequences that exhibit low sequence identities with respect to the template wild-type proteins (WT) as well as other sampling results. To this end, we design a function-oriented design scheme with conditional protein diffusion (CPDiffusion) for protein sequence generation. The proposed CPDiffusion learns implicit facilitation rules of robust candidates from wild-type proteins in nature, including both general-class proteins and a small set of other Ago proteins. To explore the reliability of CPDiffusion, we generate a set of mesophilic pAgo proteins based on *Kurthia massiliensis* Ago (KmAgo) protein's backbone [31]. Despite a considerably cost-efficient training dataset, our model managed to deliver 27 new mesophilic pAgo proteins that have $30 - 50\%$ different amino acids from the template KmAgo. The new sequences are also less than 40% similar to other WT Ago proteins in general. Moreover, the model learns to identify highly conserved regions with minimum guidance and tends to exhibit high variance in the remaining regions. The entire training and inference process involves minimum expert guidance and specification on the protein of interest.

We validate the enzymatic function and thermostability of AI-designed mesophilic pAgo proteins by biophysical and biochemical assays. Protein expression and purification experiments revealed successful expression and solubility of all the 27 artificial protein (AP) sequences. Significantly, 24 out of the 27 generated protein sequences displayed single-strand DNA (ssDNA) cleavage activity, where 20 of these sequences demonstrated a superior ssDNA cleavage activity to that of the reference KmAgo, with the highest activity being nearly nine times over the KmAgo's. Furthermore, 10 of these new APs exhibited enhanced thermostability. Thus, CPDiffusion generated novel mesophilic programmable endonucleases with a lengthy sequence (737 AAs) and multiple structural domains performing different functions, which could be

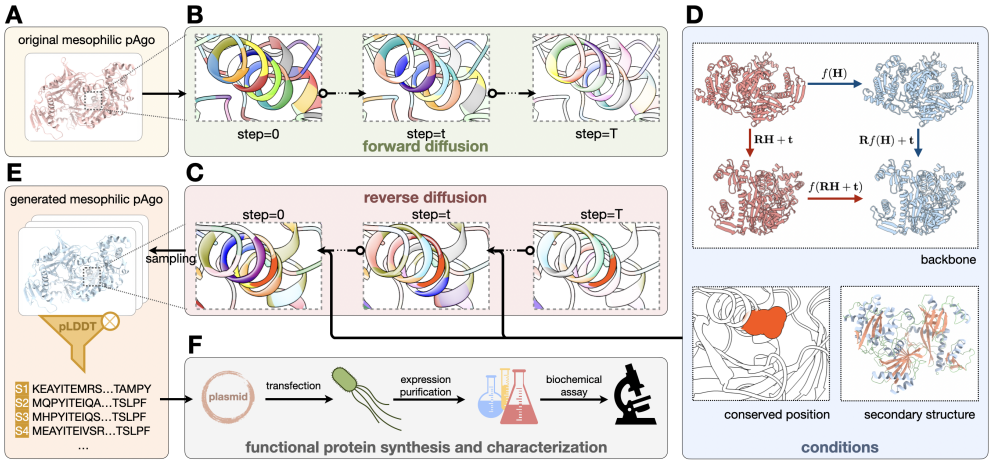

Fig. 1 Model Architecture of CPDiffusion. Arrows show the information flow among different components. (**A**) The original mesophilic prokaryotic Argonaute (pAgo) is processed to extract amino acid (AA)-level molecule biochemical and topological properties. See details in SI. (**B**) In an arbitrary zoom-in region, every AA type in the pAgo protein is corrupted in T steps following some transition probability matrix at the forward diffusion process to gradually reach uniform distribution. (**C**) The reverse diffusion process starts from randomly sampled across the 20 amino acids with a uniform distribution, following a progressive denoising process that (**D**) conditioned on protein backbone, secondary structure, and the fixed AA type at a conserved position. (**E**) A set of protein sequences will be sampled from the learned distribution, which is then filtered by their pLDDT calculated by AlphaFold2 [32]. (**F**) Wet experimental workflow of generated protein synthesis and characterization.

potential candidates for easy-to-implement and high-throughput screenings in gene editing and molecular diagnostics [24, 30]. The new approach for generating valid sequences of functional proteins showcased a distinctive feature characterized by low sequence identity, enhanced enzymatic function, thermostability, and high success rate of generalized functional proteins, accompanied by the simultaneous mutation of approximately 200 to 400 amino acids. This methodology marks a significant advancement in protein design and engineering for biomedical, biotechnological, and environmental applications.

2 Results

CPDiffusion trains a denoising diffusion model of 4 million parameters with the structure of 32 thousands natural proteins, including 694 WT Ago proteins. The model assigns categorical noise distributions to amino acids and learns their distribution via a conditional reverse diffusion. The conditions are key to recovering the distribution of protein sequences, where some desired choices include protein backbone, secondary structure, and transition matrix. Additionally, there are some optional conditions that can be applied to guide generating preferable protein sequences, such as the conserved positions (*e.g.*, the 713th AA (D) in KmAgo).

The trained CPDiffusion was applied to generate 100 mesophilic pAgo sequences guided by the KmAgo template, following an initial structure-based screening process

to reduce the experimental workload. Leveraging the structural conservation of Ago proteins, we adopted AlphaFold2 [32] to assess the quality of the new sequences by their structural prediction confidence with respect to the template KmAgo. Initially, the overall pLDDT (predicted Local Distance Difference Test) scores of the generated APs were compared against each other, where APs with their overall pLDDT scores lower than one standard deviation below the average performance among 100 samples were eliminated (Figure 2A, upper panel). Meanwhile, the residue-level discrepancies of APs and WT were measured by $\sigma(\Delta\text{pLDDT})$ (Figure 2A, middle panel) and $\text{count}(|\Delta\text{pLDDT}| > 10)$ (Figure 2A, lower panel), where $\Delta\text{pLDDT} = \text{pLDDT}_{\text{AP}_i} - \text{pLDDT}_{\text{WT}}$ describes per-residue pLDDT difference between APs and the WT. The former calculates the volatility of per-residue pLDDT difference between APs and the WT. While we desire APs with a smaller inconsistency (with respect to the WT), those with $\sigma(\Delta\text{pLDDT})$ exceeding one standard deviation above the average inconsistency were removed. The latter criterion disposes APs of more than 93 AA positions that have ΔpLDDT larger than 10. Similar to before, the threshold 93 was taken from one standard deviation above the average.

The above three initial screening criteria led to the exclusion of ~ 30 AP sequences. A more refined analysis is pursued subsequently by comparing the pLDDT scores of individual AAs within the APs with those in the WT. We finally selected 27 APs (the sequences are shown in Tables S2-S9), whose structures were determined by AlphaFold2, exhibit the highest local similarities to the WT KmAgo protein (Figure 2B and Figures S1-S27, the examples of excluded APs are shown in Figures S28-S33). Meanwhile, selected APs were subjected to structural alignment with the WT to confirm their structural similarity (Figure 2C and Figure S34-S35). Root Mean Square Deviation (RMSD) values were calculated to quantify this similarity, where all the 27 remaining proteins report considerably small RMSDs below 3 Å [33–35], falling within the resolution range typically observed in X-ray crystallography of pAgo protein (Figure 2D). Additionally, the TM scores also validate the foldability of APs similar to the WT (Figure S36). Furthermore, we examined the sequence similarity of the selected 27 APs to each other, the template KmAgo, other pAgo proteins, and other protein families (Figure 2E and Figure S37). Comparisons between AP and other APs, as well as between APs and the template KmAgo, indicated sequence similarity ranging from approximately 46% to 80% and 48% to 69%, respectively. In comparison to all natural pAgo proteins (establishment of pAgo database is discussed in Methods) and proteins from broader families, the sequence identity ranged within 35% and 45%. These findings suggest that our CPDiffusion discovers dissimilar sequences with inherent structural characteristics of template proteins.

As discussed above, the screening process returns 27 APs that exhibit reliable structures (the structure and domain architecture of KmAgo are shown in Figure 3A and B) (measured by pLDDT by AlphaFold2 [32]) and satisfying diversity at the sequence level. We next evaluate the expression, single-strand DNA (ssDNA) cleavage activity, and thermostability of these APs using wet-lab experiments. In order to measure the protein expression, a polyG linker was employed that fuses green fluorescence protein (GFP) and AP. Consequently, soluble expressed APs are observable to the unaided eye. Notably, under buffer conditions, all 27 APs exhibited distinct

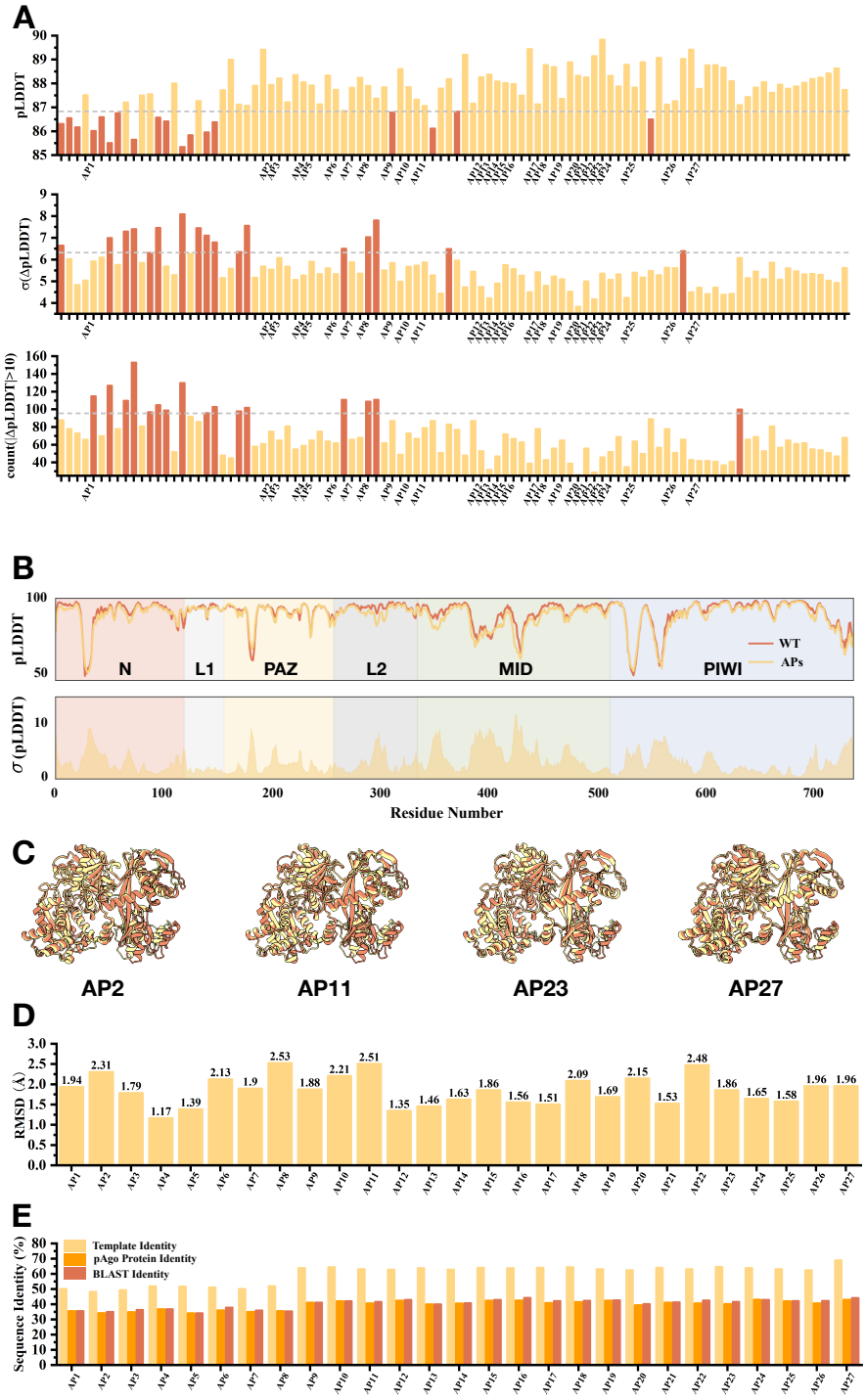

Fig. 2 (Caption on the next page.)

Fig. 2 Initial Screening of CPDiffusion-Generated Artificial Ago Proteins. (**A**) Upper panel: Average pLDDT of all APs. Middle panel: Residue-level pLDDT variation of all APs compared to the WT KmAgo. Bottom panel: Number of AAs in APs that have a disparate pLDDT (*i.e.*, the difference between pLDDTs of the AP and the WT is larger than 10) from KmAgo's. (**B**) Upper panel: The pLDDT comparison of the filtered 27 APs and the WT. Lower panel: Variation of AA-wise pLDDT between the 27 APs and KmAgo. (**C**) Four alignment examples of APs and the WT structures. Other alignments can be found in SI. (**D**) RMSD of the 27 filtered APs. (**E**) Sequence identity of the 27 APs compared to KmAgo, the family of pAgo proteins, and all proteins in the BLAST database.

green fluorescence signals, as depicted in Figure 3C, signifying successful expression and solubility. Additionally, we tested the foldability of APs. As shown in Figure S38, the circular dichroism signal of APs is similar to that of KmAgo, which indicates that APs are correctly folded in a buffer. Subsequently, we investigated the ssDNA cleavage activity of these APs. Figure 3D presents the cleavage process for pAgo proteins, where the presence of red fluorescence indicates effective cleavage of the target DNA (tDNA) and the absence of red fluorescence implies the uncleavage of tDNA (see details in methods). Remarkably, 24 out of the 27 APs demonstrated ssDNA cleavage activity at 37°C, while 21 proteins exhibited enhanced DNA cleavage activity than the WT, with the best-performing protein (AP23) exhibiting nearly 9 times the ssDNA cleavage activity of the WT's. Furthermore, we employed the fluorescence polarization assay on KmAgo and AP23 to study the binding affinity to gDNA and tDNA. The results, as shown in Figure S39, the dissociate constant Kd of AP23 is lower than that of KmAgo, demonstrating an increased binding affinity of AP23 to gDNA and tDNA. To evaluate the thermostability of the generated APs, we incubated them at 42°C for 2 minutes and 5 minutes, followed by assays for ssDNA cleavage activity, which were then normalized to their respective activities at 37°C. Figure 3F shows that 10 APs exhibited greater thermostability than the WT. Particularly noteworthy were some APs, such as AP5 and AP14 (Figure 3F), that shows concurrent enhancements in both activity and thermostability. These APs of enhanced biofunctions, especially for enzymatic activity and thermostability, showcase cpdd's promising ability towards function-oriented protein design, which is benefit from not only data-driven analysis and inference on understanding the implicit 'sequence-structure-function' relationship of protein data, but also biology-guided design of powerful deep neural networks for effectively learning, discovering, and exploring domain-specific knowledge.

To understand how the CPDiffusion-generated APs relate to all the natural Ago proteins in terms of their function and evolution relationship, we performed a comprehensive analysis based on their sequences and structures. First, we undertook an evolutionary analysis on the 27 APs to find their relationship with other natural Agos. A phylogenetic tree was built with multiple sequence alignment (MSA) of APs and the pAgo protein database (Figure 4A, Methods). The designed APs belong to the KmAgo lineage and has positioned within the long-A clade, which means APs possess similar sequences of KmAgo and evolutionary properties of long-A pAgos. Considering the conserved sites across all pAgo proteins, the distribution of conserved AAs from APs generated by CPDiffusion closely parallels those observed in naturally Ago proteins (Figure 4B, Methods). These conserved residue sites and distributions are confined within domains critical for Ago protein binding and cleavage of substrates (Figure 4C).

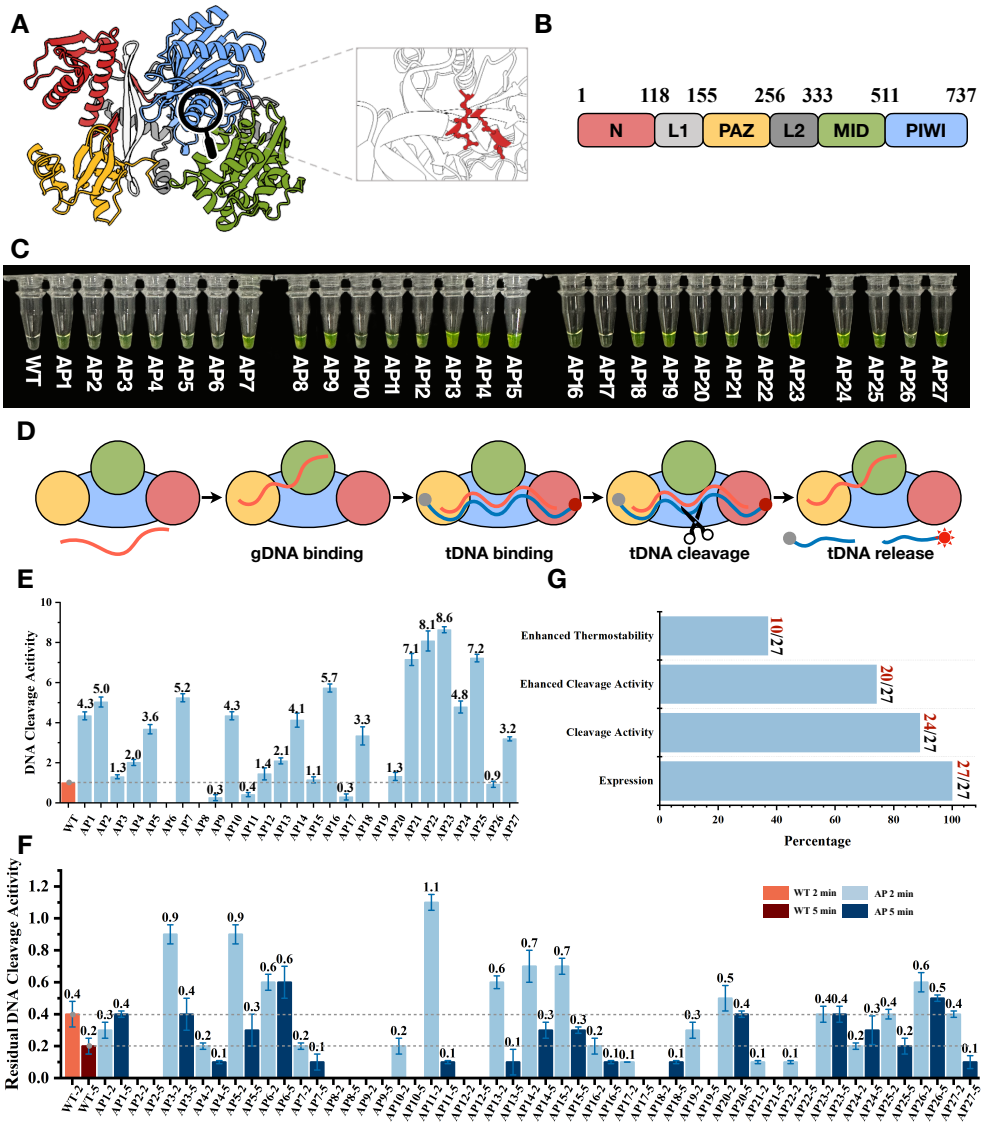

Fig. 3 Synthesis and Characterization of CPDiffusion-Generated Artificial Ago Proteins. (**A**) Left panel: Representation of the domain architectures of KmAgo. The N domain, Linker1, PAZ domain, Linker2, MID domain, and PIWI domain are colored in red, light gray, yellow, green, gray, and blue, respectively. Right panel: Catalytic sites of KmAgo. (**B**) Schematic diagram of the domain organization of KmAgo. (**C**) 27 APs exhibited soluble expression, which is visualized by linking GFP to APs. (**D**) Schematic diagram of DNA-catalytic cycle of KmAgo proteins. Cleavage activity of (**E**) KmAgo and APs at 37 °C and (**F**) those incubated at 42 °C for 2 minutes and 5 minutes before conducting ssDNA cleavage assay. Proteins are loaded with ssDNA guide and then incubated with ssDNA target in a 2 : 1 : 1 molar ratio (protein:guide:target). These activities are determined as the slope of the time dependence of fluorescence intensity [36] (see Figure **S46-S48**). Results from three independent experiments were quantified with the associated standard deviations represented by error bars. (**G**) Out of the 27 experimented APs, all of them can be expressed, with 24 exhibiting single-strand DNA cleavage activity, 17 surpassing the WT's cleavage activity, and 10 showing enhanced thermostability.

Moreover, the electrostatic surface properties of APs exhibit resemblances to those of the KmAgo template, including the negatively charged core for metal ion binding at the cleavage site, and the positively charged surface responsible for target nucleic acid binding (Figure 4D and Figure S40). These findings affirm CPDiffusion's ability to assimilate intrinsic sequential and structural characteristics, thereby ensuring the functional efficacy of the resultant proteins. Of particular interest is the observation that, despite the overall similarity in APs' structure to the WT's, there are discernible variations in the number of hydrogen bonds and salt bridges among the amino acids within proteins (Figure 4E). This discrepancy implies the existence of distinct protein-protein interactions within APs, which could potentially induce alterations in protein enzymatic activity and thermostability.

3 Discussions

In this study, we designed CPDiffusion for generating functional proteins with the three-dimensional backbone of KmAgo being the structural template. KmAgo protein exhibits a characteristic molecular architecture composed of two distinct lobes, a feature commonly observed in members of the long pAgo protein family (Figure 3A). The first lobe encompasses the N domain, Linker1, and PAZ domain, while the second lobe comprises the Linker2, MID domain, and PIWI domain (Figure 3A and B) [37]. Each domain plays a crucial role in the functional properties of KmAgo [38]. Similar to most long pAgo proteins, KmAgo primarily interacts with guide nucleic acids, with the 5'-end and 3'-end of the nucleic acid fitting into the basic pocket of the MID domain and the PAZ domain, respectively [39]. Once the guide nucleic acid is loaded onto KmAgo, the KmAgo-guide pair forms a complex that scans for complementary target strands by base-pairing with the seed region of the guide [34, 39]. This is followed by the propagation of the nucleic acid duplex [39]. Subsequently, the 3'-end of the guide is released from the PAZ domain, and the active site of the Glu-finger undergoes closure to facilitate correct base-pairing [39]. Conformational changes within the PIWI domain's catalytic tetrad enable the cleavage of the target strand [34]. Finally, the N domain aids in releasing the cleaved target, and the protein-guide complex reverts to its initial conformation in preparing the subsequent round of cleavage [40, 41]. The aforementioned sequential steps in the functional mechanism of KmAgo shed light on the intricate interplay between its various domains and their concerted actions during nucleic acid recognition and cleavage processes. This intricate interplay underscores the complexity of KmAgo as a multi-functional protein. Our results demonstrated that the CPDiffusion-generated APs exhibited low sequence identity (below 45%) compared with the naturally existing Ago proteins. Notably, biochemical analyses revealed that around 90% of artificial Ago proteins demonstrate ssDNA cleavage activity. Furthermore, CPDiffusion successfully generated a class of mesophilic Ago proteins with enhanced enzymatic activity and thermostability. In other words, our CPDiffusion does more than generate valid proteins with highly complex biofunctions, in the sense that it simultaneously evolves the template protein towards superior functions.

There has been considerable progress in designing new functional proteins using deep learning methods, which are validated experimentally. ProteinGAN [6] learns the

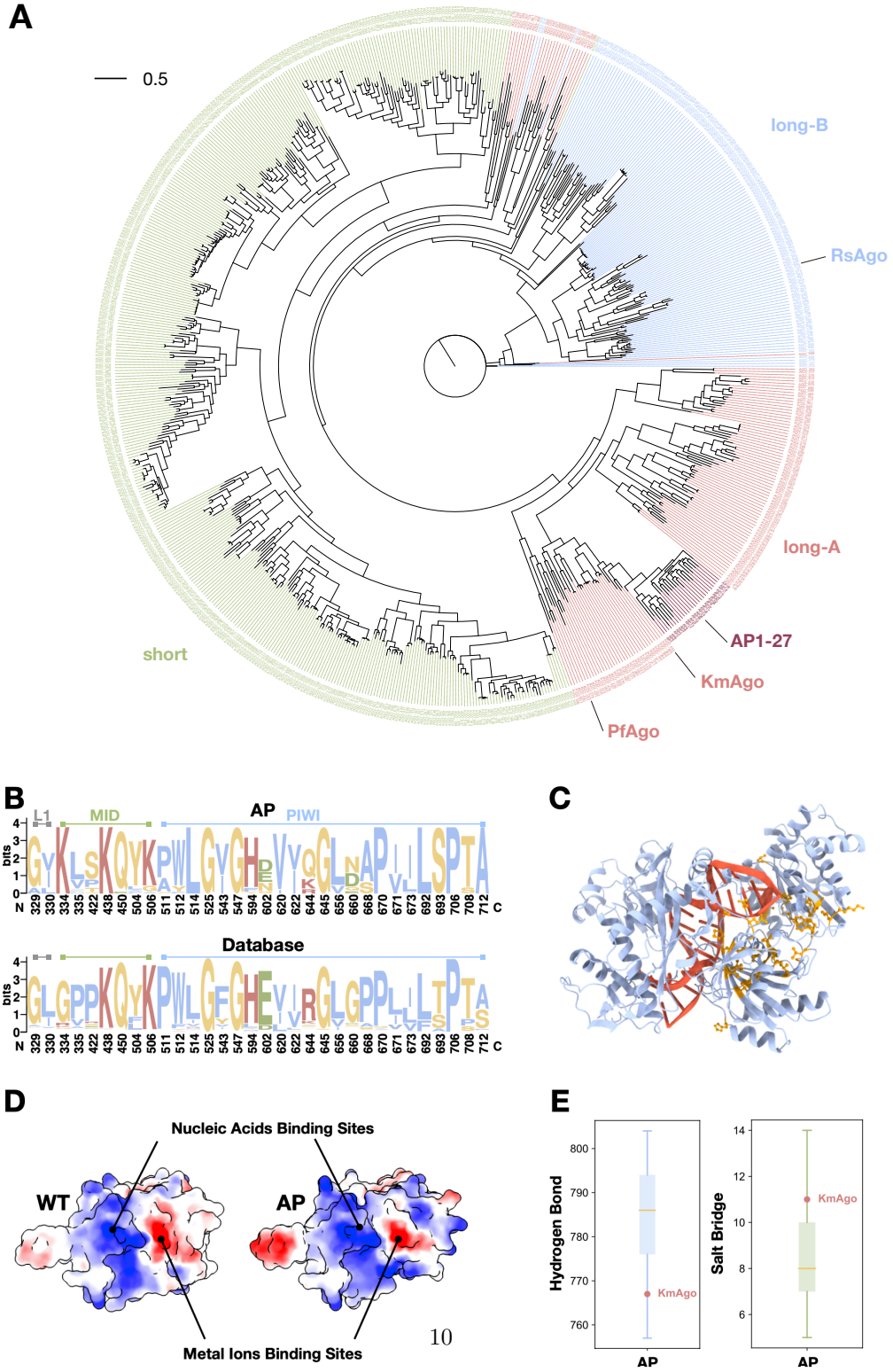

Fig. 4 Evolutionary Analysis of CPDiffusion Generated Artificial Ago Proteins. (**A**) Phylogenetic tree of generated sequences and natural sequences. (**B**) Conserved residues distribution among generated sequences (upper panel) and nature (lower panel). (**C**) Conserved residues are highlighted in orange in the protein structure. (**D**) Electrostatic surface of MID domain and PIWI domain in KmAgo and APs. The arrows indicate the nucleic acids binding sites and metal ions binding sites.(**E**) The number of hydrogen bond and salt bridge of KmAgo and CPDiffusion generated APs.

evolutionary relationships of protein sequences directly from the sophisticated multi-dimensional space of AA sequences. Out of 55 artificial sequences containing 321 AAs and one functional domain, 24% new proteins are soluble and display malate dehydrogenase catalytic activity. Madani et al. [7] pre-trained ProGen, a large language model with 280 million of protein sequences and billions of network parameters, for conditional protein design. Through fine-tuning the model on the particular protein family, it generates lysozyme ($\sim$ 120 amino acids) sequences of low sequence identities with thousands of protein samples observed activity. Lately, Watson et al. [8] delivered RFdiffusion base on structures, a function-oriented protein design method, that achieves outstanding performance on protein monomer design, protein binder design, symmetric oligomer design, enzyme active site scaffolding, and symmetric motif scaffolding. In comparison, CPDiffusion established a discrete denoising diffusion model for protein sequence generation, which implements sufficient inductive bias (such as structural learning and biology domain knowledge) in model design to fit a petite model at around 4 millions of parameters learned from only 30 thousands of natural proteins. The model learns sequential and structural information of proteins efficiently that directs designing proteins with excessively long protein sequences (737 AAs) and highly complex functionality. As shown in Figures S41-S42 and Figure 4B, the model efficiently learns the distribution of the conserved residue trained with the pAgo protein dataset. Remarkably, the generated functional artificial Ago proteins by CPDiffusion yields $\sim$ 90% of cleavage activity. It even discovers a number of artificially mesophilic Ago proteins in the sequence space that exhibited higher enzymatic activity and thermostability than the template KmAgo, indicating that the model designs functional proteins and evolves desired functions at the same time. These observations demonstrate the great significant potential of CPDiffusion in protein engineering, as enhancing the performance of mesophilic pAgo proteins opens possibilities for their application in isothermal molecular detection, labeling, gene editing, *etc.*

Utilizing our CPDiffusion, we can generate proteins with long sequences, multi-functional domains, enhanced enzymatic activity and thermostability, and low sequence identity. This approach offers a notable possibility to generate more complex proteins, such as crisper Case proteins, which could enable researchers to surpass the limitations imposed by patented protein sequences and expand the protein candidates library. Moreover, our CPDiffusion facilitates protein engineering endeavors. Directed evolution methods often involve laborious iterations, starting from single-site mutations and gradually combining multiple mutations to engineer a protein with the desired functionality. These approaches necessitate multiple rounds of experimentation, which can be time-consuming and costly. More importantly, the quality of the final product of the direct evolution is strongly depending on its starting wild type, as this method is only searching for mutants around the starting sequence with less than a few percent variations in sequence identity. The starting wild type is not necessarily the best starting point to search for the sequence with the globally best functionality as they might differ significantly in sequence landscape. Our model streamlines the process by directly generating proteins with high bio-performance in one go, which could provide many new start points for further directed evolution to rapidly search for the mutants of better properties. The efficiency and effectiveness of our approach offer

promising prospects for accelerating protein design and discovery, ultimately fostering advancements in protein engineering and design and corresponding applications.

4 Methods

4.1 Conditional Protein Denoising Diffusion

DDPMs [10] approximate a distribution by parameterizing the reversal of a discrete-time forward diffusion process. In the context of protein sequence generation, the forward diffusion operates on the distribution of AA types $\boldsymbol{X}_0^{\text{aa}}$ that involves the iterative noise addition for T steps until reaching $q(\boldsymbol{X}_T^{\text{aa}})$, a distribution indistinguishable from a reference independent transition distribution. The forward diffusion ($\boldsymbol{X}_0^{\text{aa}} \rightarrow \boldsymbol{X}_T^{\text{aa}}$) samples $\boldsymbol{X}_t^{\text{aa}}$ at every step t ($t = 1, \ldots, T$) using the forward diffusion kernel defined as $q(\boldsymbol{X}_t^{\text{aa}} \mid \boldsymbol{X}_{t-1}^{\text{aa}}) = \boldsymbol{X}_{t-1}^{\text{aa}} \boldsymbol{Q}_t$ until $\boldsymbol{X}_T^{\text{aa}} \sim q(\boldsymbol{X}_T^{\text{aa}})$, where $\boldsymbol{Q}_t$ denotes the transition probabilities at t. The reverse diffusion ($\boldsymbol{X}_T^{\text{aa}} \rightarrow \boldsymbol{X}_0^{\text{aa}}$) samples $\boldsymbol{X}_T^{\text{aa}} \sim p_\theta(\boldsymbol{X}_T^{\text{aa}}) \approx q(\boldsymbol{X}_T^{\text{aa}})$ to convert from the prior into samples of the learned data distribution, *i.e.*, $\boldsymbol{X}_0^{\text{aa}} \sim p_\theta(\boldsymbol{X}_0^{\text{aa}})$. During the progressive training process, the generative probability distribution $p_\theta(\boldsymbol{X}_{t-1}^{\text{aa}} \mid \boldsymbol{X}_t^{\text{aa}})$ is parameterized by a neural network $f(\theta)$ that embeds additional characterization of protein instances, such as their topology and physicochemical features. We represent proteins as graphs with nodes corresponding to AAs and implement equivariant graph neural networks (*e.g.*, , EGNN [42]) that perform rotation and translation equivariance on aggregating node attributes based on relative spatial relationships of corresponding neighboring AAs. Additional conditions are also possible to be inserted to guide the denoising process, such as the secondary structure and transition matrix. Notably, we defined a conserved AA at the 713th position as the known region and conditioned the reverse diffusion process of an unconditional DDPM for sequence inpainting [13]. The weights in the neural network $f(\theta)$ are optimized by minimizing the cross-entropy $\mathcal{L}_{\text{CE}}$ between $\hat{p}_\theta(\boldsymbol{X}^{\text{aa}})$, the network prediction that approximates $p_\theta(\boldsymbol{X}^{\text{aa}})$, and the observed AA types. Further details regarding the diffusion model can be found in SI.

4.2 Structure Prediction and Comparison

To predict the structure for the artificial Ago sequences generated by conditional protein denoising diffusion, we used AlphaFold2 in single-sequence mode with multiple sequence alignment (MSA) and with PDB templates. The highest-ranked predicted structure among the five models was used. The RMSD between WT and artificial Ago protein in our study are calculated using PyMOL (The PyMOL Molecular Graphics System, Version 2.0, Schrödinger, LLC.). Hydrogen bonds are defined as polar contacts in proteins, salt bridges are defined by locating oxygen atoms in acid amino acids, and nitrogen in base amino acid pairs has distances within 4 Å.

4.3 Prokaryotic Argonaute Protein Database

We constructed a dataset containing 694 pAgo proteins, selected from a comprehensive Ago dataset [43]. This curated set represents a broad diversity of pAgo types, including short, long-A, and long-B Ago proteins.

4.4 Sequence and Structure Analysis

The model generated 27 Ago proteins and 694 Ago proteins from the database were subjected to multiple sequence alignment using the MUSCLE v5 software with the super5 algorithm [44]. Conservation scores for each residue were computed by

$$R_{\text{seq}} = \log_2 N - \left(-\sum_{n=1}^{N} p_n \log_2 p_n \right),$$

where p_n is the observed frequency of AA n, and N=20 for all 20 kinds of AAs. Conservation score equals R_{seq} multiplied by the fraction of non-gapped positions in the column to apply the gap penalty. Residues in the Ago database scoring above 2.5 were selected. Residues in the design scoring below 0.2 were excluded to mitigate alignment discrepancies. The conserved residues were visualized using WebLogo [45]. The phylogenetic tree is computed via IQ-TREE v1.6 with 1500 ultrafast bootstraps and used BLOSUM62 as a substitution model [46]. FigTree v1.4.4 is used to visualize the phylogenetic tree (http://tree.bio.ed.ac.uk/software/figtree/). The electrostatic surface, or the Coulombic electrostatic potential (ESP) of the protein was subsequently assessed using the 'coulombic' command in Chimera visualization tool, which provided a graphical display of the distribution of electrostatic potential on the surface of the protein.

4.5 Protein Expression and Purification

A codon-optimized version of the KmAgo and its mutants gene was synthesized by Sangon Biotech (Shanghai, China). It was cloned into the pET15(b) plasmid (the construction of Ago protein is shown in Figure S43-S44) to construct pEX-Ago with an N terminal His-tag. The expression plasmid was transformed into *Escherichia coli* BL21(DE3) cells. A 30 ml seed culture was grown at 37 °C in LB medium with 50 μg/ml kanamycin and was subsequently transferred to 500m L of LB in a shaker flask containing 50 μg/ml kanamycin. The cultures were incubated at 37 °C until the OD600 reached 0.6-0.8, and protein expression was then induced by the addition of isopropyl-D-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM, followed by incubation for 16-20 h at 18 °C. Cells were harvested by centrifugation for 30 min at 4,000 rpm, and the cell pellets were collected for later purification. The cell pellets were resuspended in lysis buffer (25 mM Tris-HCl, 500 mM NaCl, 10 mM imidazole, pH 7.4) and then disrupted using a High-Pressure Homogenizer at 700-800 bar for 5 minutes (Gefran, Italy). The lysates were centrifuged for 30 min at 12,000 rpm at 4 °C, after which the supernatants were subjected to Ni-NTA affinity purification with elution buffer (25 mM Tris-HCl, 500 mM NaCl, 250 mM imidazole, pH 7.4). Further gel filtration purification using a Superdex 200 (GE Tech, USA) was carried out with an elution buffer. The fractions resulting from gel filtration were analyzed by SDS-PAGE (Figure S45), and fractions containing the protein were flash frozen at -80 °C in buffer (15 mM Tris–HCl pH 7.4, 200 mM NaCl, 10 glycerol).

4.6 Single-Strand DNA Cleavage Assay

For standard activity assays of KmAgo and its mutants, cleavage experiments were performed in a 2:1:1 molar ratio (protein:guide:target). For studying the effect of temperature on the activity of proteins, they were incubated at 42 °C for 2 min and 5 min before conducting the ssDNA cleavage assay. First, 5 µM protein was mixed with a synthetic 1 μM gDNA guide in the reaction buffer (15 mM Tris-HCl (pH 7.4), 200 mM NaCl, different concentrations of $MnCl_2$ or 5 mM EDTA). The solution was then pre-incubated at 37 °C for 20 min. After pre-incubation, 1 μM tDNA, which was labeled with the fluorescent group 6-FAM at the 5′-end and the quencher BHQ1 at the 3′-end, was added to the mixture. The cleavage experiments were performed at 37 °C. All experiments were performed in triplicate, and the fluorescence signals were traced by the quantitative real-time PCR QuantStudio 5 (Thermo Fisher Scientific, USA) with λ_{ex} = 495nm and λ_{em} = 520nm. The results were analyzed by QuantStudioTM Design & Analysis Software v1.5.1. The guide DNA and target DNA used for cleavage are listed in Table S1. Three independent experiments are conducted to determine the cleavage activity of Ago proteins.

4.7 Circular Dichroism spectroscopy

The CD measurements on the secondary structure of KmAgo and APs were performed in a Jasco J-1500 spectropolarimeter with a 1 mm pathlength cell. The concentrations of proteins were 0.1 mg/mL in 1× PBS buffer (pH = 7.4). For CD spectra, we used PBS buffer as the solvent instead of Tris-HCl buffer because the Tris-HCl buffer has a strong CD background, which will significantly affect the analysis of the CD signal of proteins [36].

4.8 Fluorescence Polarization Assay

To determine the apparent dissociation constant (Kd) for KmAgo and AP23 binding to either guide DNA or target DNA, a fluorescence polarization assay was conducted utilizing a multifunctional enzyme-linked immunosorbent assay plate reader (Spark, Tecan). A solution was prepared by combining 5 nM of 3'6-FAM labeled guide or target DNA with protein proteins in a range of concentrations from 0 to 1500 nM in a reaction buffer (15 mM Tris-HCl, pH 7.4, 200 mM NaCl, and varying concentrations of MnCl2). This mixture was incubated at 37°C for 20 minutes and then transferred to a light-protected 96-well ELISA plate. The degree of polarization was measured using the Spark Tecan plate reader, with an excitation wavelength of 485 nm and an emission wavelength of 525 nm. The resulting data were processed via non-linear regression using the one-site total binding model in GraphPad Software's Prism 8, to calculate the Kd value.

Supplementary Information. Figures S1-S47, Tables S1-S9.

Author Contributions. B.Zhou and L.Z. conceived and designed the study. B.Zhou and K.Y. designed and performed CPDiffusion and generation. L.Z. and B.W.

performed biochemical experiments. L.Z. and B.Zhong performed sequence and structure analysis. B.Zhou and L.Z. wrote the manuscript with feedback from L.H. and P.L. supervised the project.

Acknowledgments. The authors acknowledge the Center for High-Performance Computing at Shanghai Jiao Tong University for computing resources.

Funding. This work was supported by the National Natural Science Foundation of China (31630002), the Innovation Program of Shanghai Municipal Education Commission (2019-01-07-00-02-E00076), the Student Innovation Center at Shanghai Jiao Tong University, and Shanghai Artificial Intelligence Laboratory.

Additional Information. Correspondence and requests for materials should be addressed to hongl3liang@sjtu.edu.cn, pl219@cam.ac.uk, and lirong_zheng@sjtu.edu.cn.

Data availability. All the data are presented in SI.

Competing Interests. B.Zhou, L.Z., B.W., and L.H. are inventors on a provisional patent application related to the described work. The other authors declare no competing interests.

References

[1] Huang, P.-S., Boyken, S.E., Baker, D.: The coming of age of de novo protein design. Nature **537**(7620), 320–327 (2016)

[2] Pearce, R., Zhang, Y.: Deep learning techniques have significantly impacted protein structure prediction and protein design. Current opinion in structural biology **68**, 194–207 (2021)

[3] Lu, H., Diaz, D.J., Czarnecki, N.J., Zhu, C., Kim, W., Shroff, R., Acosta, D.J., Alexander, B.R., Cole, H.O., Zhang, Y., *et al.*: Machine learning-aided engineering of hydrolases for PET depolymerization. Nature **604**(7907), 662–667 (2022)

[4] Thean, D.G., Chu, H.Y., Fong, J.H., Chan, B.K., Zhou, P., Kwok, C., Chan, Y.M., Mak, S.Y., Choi, G.C., Ho, J.W., *et al.*: Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities. Nature Communications **13**(1), 1–14 (2022)

[5] Tan, Y., Zhou, B., Jiang, Y., Wang, Y.G., Hong, L.: Multi-level protein representation learning for blind mutational effect prediction. arXiv:2306.04899 (2023)

[6] Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., *et al.*: Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine Intelligence **3**(4), 324–333 (2021)

[7] Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos Jr, J.L., Xiong, C., Sun, Z.Z., Socher, R., et al.: Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 1–8 (2023)

[8] Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., et al.: De novo design of protein structure and function with rfdiffusion. Nature, 1–3 (2023)

[9] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265 (2015). PMLR

[10] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

[11] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831 (2021). PMLR

[12] Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)

[13] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11461–11471 (2022)

[14] Yi, K., Zhou, B., Shen, Y., Liò, P., Wang, Y.G.: Graph denoising diffusion for inverse protein folding. arXiv:2306.16819 (2023)

[15] Corso, G., Stärk, H., Jing, B., Barzilay, R., Jaakkola, T.S.: Diffdock: Diffusion steps, twists, and turns for molecular docking. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=kKF8_K-mBbS

[16] Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., Frossard, P.: Digress: Discrete denoising diffusion for graph generation. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=UaAD-Nu86WX

[17] Hoogeboom, E., Satorras, V.G., Vignac, C., Welling, M.: Equivariant diffusion for molecule generation in 3d. In: International Conference on Machine Learning, pp. 8867–8887 (2022). PMLR

[18] Gruver, N., Stanton, S., Frey, N.C., Rudner, T.G., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., Wilson, A.G.: Protein design with guided discrete diffusion. arXiv:2305.20009 (2023)

[19] Hegge, J.W., Swarts, D.C., Oost, J.: Prokaryotic argonaute proteins: novel genome-editing tools? Nature Reviews Microbiology **16**(1), 5–11 (2018)

[20] Song, J., Hegge, J.W., Mauk, M.G., Chen, J., Till, J.E., Bhagwat, N., Azink, L.T., Peng, J., Sen, M., Mays, J., *et al.*: Highly specific enrichment of rare nucleic acid fractions using thermus thermophilus argonaute with applications in cancer diagnostics. Nucleic acids research **48**(4), 19–19 (2020)

[21] Liu, Q., Guo, X., Xun, G., Li, Z., Chong, Y., Yang, L., Wang, H., Zhang, F., Luo, S., Cui, L., *et al.*: Argonaute integrated single-tube pcr system enables supersensitive detection of rare mutations. Nucleic acids research **49**(13), 75–75 (2021)

[22] Wang, F., Yang, J., He, R., Yu, X., Chen, S., Liu, Y., Wang, L., Li, A., Liu, L., Zhai, C., *et al.*: Pfago-based detection of sars-cov-2. Biosensors and Bioelectronics **177**, 112932 (2021)

[23] Xun, G., Liu, Q., Chong, Y., Guo, X., Li, Z., Li, Y., Fei, H., Li, K., Feng, Y.: Argonaute with stepwise endonuclease activity promotes specific and multiplex nucleic acid detection. Bioresources and Bioprocessing **8**, 1–12 (2021)

[24] Graver, B.A., Chakravarty, N., Solomon, K.V.: Prokaryotic argonautes for in vivo biotechnology and molecular diagnostics. Trends in Biotechnology (2023)

[25] Filius, M., Cui, T.J., Ananth, A.N., Docter, M.W., Hegge, J.W., Van Der Oost, J., Joo, C.: High-speed super-resolution imaging using protein-assisted dna-paint. Nano letters **20**(4), 2264–2270 (2020)

[26] Chang, L., Sheng, G., Zhang, Y., Shao, S., Wang, Y., Sun, Y.: Agofish: cost-effective in situ labelling of genomic loci based on dna-guided dttago protein. Nanoscale Horizons **4**(4), 918–923 (2019)

[27] Toudji-Zouaz, A., Bertrand, V., Barrière, A.: Imaging of native transcription and transcriptional dynamics in vivo using a tagged argonaute protein. Nucleic Acids Research **49**(15), 86–86 (2021)

[28] Vaiskunaite, R., Vainauskas, J., Morris, J.J., Potapov, V., Bitinaite, J.: Programmable cleavage of linear double-stranded dna by combined action of argonaute cbago from clostridium butyricum and nuclease deficient recbc helicase

from e. coli. Nucleic Acids Research **50**(8), 4616–4629 (2022)

[29] Li, X., Dong, H., Guo, X., Huang, F., Xu, X., Li, N., Yang, Y., Yao, T., Feng, Y., Liu, Q.: Mesophilic argonaute-based isothermal detection of sars-cov-2. Frontiers in microbiology **13**, 957977 (2022)

[30] Qin, Y., Li, Y., Hu, Y.: Emerging argonaute-based nucleic acid biosensors. Trends in Biotechnology **40**(8), 910–914 (2022)

[31] Kropocheva, E., Kuzmenko, A., Aravin, A.A., Esyunina, D., Kulbachinskiy, A.: A programmable pago nuclease with universal guide and target specificity from the mesophilic bacterium kurthia massiliensis. Nucleic acids research **49**(7), 4054–4065 (2021)

[32] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (2021)

[33] Hegge, J.W., Swarts, D.C., Chandradoss, S.D., Cui, T.J., Kneppers, J., Jinek, M., Joo, C., Van Der Oost, J.: Dna-guided dna cleavage at moderate temperatures by clostridium butyricum argonaute. Nucleic acids research **47**(11), 5809–5821 (2019)

[34] Sheng, G., Zhao, H., Wang, J., Rao, Y., Tian, W., Swarts, D.C., Oost, J., Patel, D.J., Wang, Y.: Structure-based cleavage mechanism of thermus thermophilus argonaute dna guide strand-mediated dna target cleavage. Proceedings of the National Academy of Sciences **111**(2), 652–657 (2014)

[35] Rivas, F.V., Tolia, N.H., Song, J.-J., Aragon, J.P., Liu, J., Hannon, G.J., Joshua-Tor, L.: Purified argonaute2 and an sirna form recombinant human risc. Nature structural & molecular biology **12**(4), 340–349 (2005)

[36] Zheng, L., Lu, H., Zan, B., Li, S., Liu, H., Liu, Z., Huang, J., Liu, Y., Jiang, F., Liu, Q., *et al.*: Loosely-packed dynamical structures with partially-melted surface being the key for thermophilic argonaute proteins achieving high dna-cleavage activity. Nucleic Acids Research **50**(13), 7529–7544 (2022)

[37] Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J., Van Der Oost, J.: The evolutionary journey of argonaute proteins. Nature structural & molecular biology **21**(9), 743–753 (2014)

[38] Lisitskaya, L., Aravin, A.A., Kulbachinskiy, A.: Dna interference and beyond: structure and functions of prokaryotic argonaute proteins. Nature communications **9**(1), 5165 (2018)

[39] Wang, Y., Sheng, G., Juranek, S., Tuschl, T., Patel, D.J.: Structure of the guide-strand-containing argonaute silencing complex. Nature **456**(7219), 209–213

(2008)

[40] Kwak, P.B., Tomari, Y.: The n domain of argonaute drives duplex unwinding during risc assembly. Nature structural & molecular biology **19**(2), 145–151 (2012)

[41] Miyoshi, T., Ito, K., Murakami, R., Uchiumi, T.: Structural basis for the recognition of guide rna and target dna heteroduplex by argonaute. Nature communications **7**(1), 11846 (2016)

[42] Satorras, V.G., Hoogeboom, E., Welling, M.: E(n) equivariant graph neural networks. In: International Conference on Machine Learning, pp. 9323–9332 (2021)

[43] Ryazansky, S., Kulbachinskiy, A., Aravin, A.A.: The expanded universe of prokaryotic argonaute proteins. MBio **9**(6), 10–1128 (2018)

[44] Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research **32**(5), 1792–1797 (2004)

[45] Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E.: Weblogo: a sequence logo generator. Genome research **14**(6), 1188–1190 (2004)

[46] Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R.: Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution **37**(5), 1530–1534 (2020)