ProTokens: A Machine-Learned Language for Compact and Informative Encoding of Protein 3D Structures

Xiaohan Lin[1*], Zhenyu Chen[1*], Yanheng Li[1], Xingyu Lu[2], Chuanliu Fan[3], Ziqiang Cao[3], Shihao Feng[4,†], Yi Qin Gao[1,4,†] and Jun Zhang[4,†]

1. Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.

2. Huawei Shanghai Research Institute, Huawei Technologies Co. Ltd., Shanghai 201206, China.

3. Institute of Artificial Intelligence, Soochow University, Suzhou 215006, China.

4. Changping Laboratory, Beijing 102200, China.

* These authors contributed equally to this work and should be considered as co-first authors.

† To whom correspondence should be addressed:

fengsh@cpl.ac.cn (S.F.), gaoyq@pku.edu.cn (Y.Q.G) or jzhang@cpl.ac.cn (J.Z.)

Abstract

Designing protein structures towards specific functions is of great values for science, industry and therapeutics. Although backbones can be designed with arbitrary variety in the coordinate space, the generated structures may not be stabilized by any combination of natural amino acids, resulting in the high failure risk of many design approaches. Aiming to sketch a compact space for designable protein structures, we present an unsupervised learning strategy by integrating the structure prediction and the inverse folding tasks, to encode protein structures into amino-acid-like discrete tokens and decode these tokens back to 3D coordinates. We show that tokenizing protein structures with proper perplexity can lead to compact and informative representations (ProTokens), which can reconstruct 3D coordinates with high fidelity and reduce the trans-rotational equivariance of protein structures. Therefore, protein structures can be efficiently compressed, stored, aligned and compared in the form of ProTokens. Besides, ProTokens enable protein structure design via various generative AI without the concern of symmetries, and even support joint design of the structure and sequence simultaneously. Additionally, as a modality transformer, ProTokens provide a domain-specific vocabulary, allowing large language models to perceive, process and explore the microscopic structures of biomolecules as effectively as learning a foreign language.

1. Introduction

Structure-oriented protein design is highly valuable in scientific research and medical development, provided that the function of proteins is largely determined by their 3D structures[1–4]. Successful design of protein structures may give rise to useful therapeutic or productive biomolecules like antibodies and enzymes[5–10]. Despite the fact that AlphaFold2[11] has achieved impressive performance in mapping amino acid sequence into folded protein structures, one of its inverse problems, that is, modeling the distribution of designable protein structures which can be accommodated or stabilized by a combination of natural amino acids, remains wide open[12,13].

Although many physics-based approaches have been proposed to generate novel protein structures, the success of the artificially designed proteins is still limited[14] due to poor understanding of the designable space of protein structures. Recently, data-driven generative methods are gaining increasing popularity and may lead to new solutions to this long-standing challenge[15–17]. However, protein 3D structures are defined by spatial coordinates of atoms which exhibit specific symmetry and are subjected to physics constraints (such as transrotational equivariance and polymer restraints), hence, significantly impeding the transaction of modern generative AI (e.g., generative adversarial[18], diffusion[19–21] and autoregressive[22–24] models etc.) for protein design. To combat the symmetry issue, some attempts transformed protein structures into a space consisting of trans-rotational operations and adapted diffusion models for the SO(3) group using specific diffusion kernels[5]. Other efforts were paid to introduce protein-specific priors in diffusion models in order to account for the polymer restraints[17,25]. Such methods enabled protein structure generation via diffusion-like processes but did not provide necessary compatibility with other advances of diffusion models such as speeded sampling[26,27] and conditional guidance[28–30], because most of these techniques only work on the regular and unconstrained vector space.

On the other hand, with the rise of large language models (LLMs) as a potential artificial general intelligence (AGI) candidate, many efforts are being paid to "unify the modality" of various signals (including images, videos and sounds etc.) to texts or tokens that LLMs are familiar with[31–33]. It is highly desired that a representation of proteins amenable to LLMs can be developed in order to give AGI an access to the microscopic biomolecular universe. However, it is unwise to treat 3D coordinates of proteins directly as input or output of an LLM due to the symmetry issue. Indeed, the Cartesian atom coordinates of proteins are known to be redundant for representing protein structures given the trans-rotational equivariance and the polymer nature of protein chains. Methods have been designed to compress and record protein structures using internal coordinates[34], and to encode local structural features of proteins such as backbone torsions in discretized forms and facilitate the comparison of protein structures[35].

In this research, we developed a novel strategy to represent protein structures in a compact way, and the redundancy of protein coordinates caused by trans-rotational equivariance and polymer restraints is largely reduced, leading to a regular latent vector space which can be decoded back and reproduce the 3D protein structures with high fidelity. Owing to the removal of group symmetry and polymer restraints, any modern generative methods can be straightforwardly applied on this latent space to generate and sample protein structures. Inspired by the success of latent generative models[36], by further quantizing this latent space into a limited number of discrete tokens (called "ProTokens"), we not only obtained a new solution to protein structure compression, but also created a "protein language" which encodes sufficient information about protein structures

and meanwhile remains amenable to LLMs. ProTokens are learned in a self-supervised manner, intuited by the profound connections between structure prediction and inverse folding. Based on ProTokens, protein structures can be compressed, aligned and compared efficiently. Moreover, the problem of protein structure prediction (or inverse folding) can be recast as a machine translation task from amino-acid sequences to ProToken sequences (or vice versa), which falls within the capacity of LLMs. Therefore, fine-tuning LLMs over ProToken vocabulary may enable AGI to demystify the structural semantics of proteins and explore the protein universe.

2. Methods

Overview

ProTokens are learned via an autoencoder-style neural network called "ProToken Distiller". The input and output of Distiller are both protein backbone structures, and ProTokens are extracted by the Distiller in an unsupervised manner. The ProToken Distiller consists of three essential parts: an encoder, a quantization module and a decoder, which are stacked together to encode (or compress) and decode (or reconstruct) a given protein backbone structure. Specifically, N, C, CA, O atom coordinates of each amino acid in a protein are given to the Encoder, producing a latent embedding (called "Structure Latent Vector", SLV) for each residue. The SLVs are then projected to a relatively low dimensional space and quantized as ProTokens. Finally, the Decoder receives the ProTokens and translates them into a 3D structure of the corresponding protein.

Mathematically, we denote the protein structure as $\mathbf{x}$, the latent embedding as $\mathbf{z}'$ and the corresponding ProToken as $\mathbf{z}$. The encoder serves to model the following posterior distribution,

$$\mathbf{z}' \sim p_\theta(\mathbf{z}'|\mathbf{x}) \approx p(\mathbf{z}'|\mathbf{x})$$

The continuous embedding $\mathbf{z}'$ is then discretized through the quantization module, either by means of a nearest-neighbor search with a "codebook" $\{\mathbf{e}_i\}_{i\in I}$[27] or finite scalar quantization[38]:

$$\mathbf{z} = \underset{\mathbf{e}\in\{\mathbf{e}_i\}_{i\in I}}{\operatorname{argmin}}\, d(\mathbf{z}', \mathbf{e})$$

where $d(\mathbf{z}', \mathbf{e})$ is a pre-defined distance metric, for example, Euclidean distance $\|\mathbf{z}' - \mathbf{e}\|$. Finally, the decoder produces the conditional probability (or perform reconstruction):

$$\mathbf{x} \sim p_\phi(\mathbf{x}|\mathbf{z}) \approx p(\mathbf{x}|\mathbf{z})$$

Specifically, a Boltzmann distribution is used to model $p(\mathbf{x}|\mathbf{z})$ on the basis of an SE(3)-invariant energy function in the following form,

$$E(\mathbf{x}; \mathbf{x}_\phi) = \mathcal{L}_{\mathrm{FAPE}}(\mathbf{x}; \mathbf{x}_\phi) + \mathcal{L}_{\mathrm{violation}}(\mathbf{x}_\phi)$$

$$p_\phi(\mathbf{x}|\mathbf{z}) \sim \exp\left(-\beta E(\mathbf{x}; \mathbf{x}_\phi)\right)$$

where $\mathbf{x}_\phi$ is the structure generated by the Decoder $\mathbf{f}_\phi^{\mathrm{Dec}}(\cdot)$, FAPE and violation losses are defined as in AlphaFold2.

Similar to other vector quantization learning, the training objective of ProToken Distiller is to maximize the log-likelihood of the conditional probability (or minimize the reconstruction loss). Due to the non-differentiability of quantization, the straight-through estimator [39] is implemented to allow the gradient being back-propagated to the encoder, and special care and tricks should be taken in order to stabilize the training of a VQ-VAE [37,38,40,41].

Protein Structure Encoder

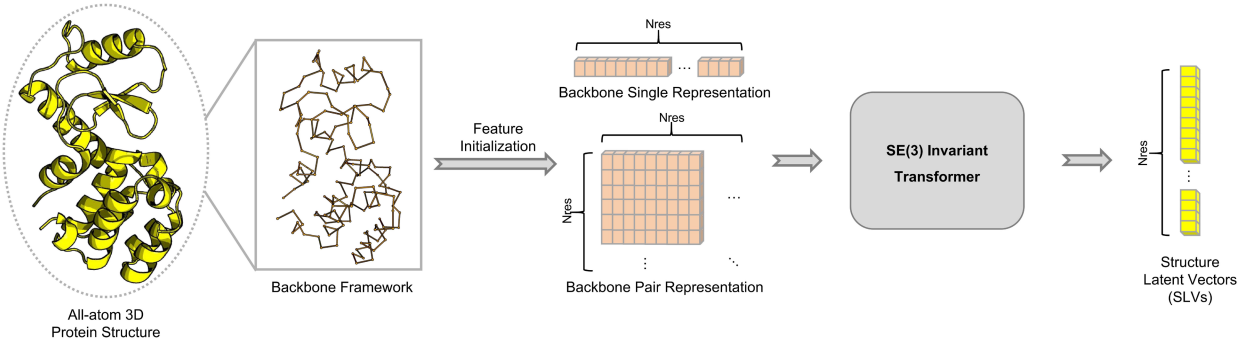

Fig. 1 Illustration of the Encoder model. The backbone of a protein is first transformed into single and pair features, and then input to the Encoder. The Encoder performs learning on both the single and pair features, and outputs the Structure Latent Vectors (SLVs) via an SE(3)-invariant Transformer.

The Enoder in the Distiller is composed of an SE(3)-invariant HyperFormer[42,43] and a structure-aware Transformer[11]. Specifically, the HyperFormer treats residues and inter-residue geometries as vertices and edges of a graph, respectively. Like EvoFormer in AlphaFold2, HyperFormer updates both the vertex and edge representations interdependently via hyper-attention mechanism[42,43]. Based on the refined vertex (or single) and edge (or pair) representations output by HyperFormer, invariant point attention (IPA) introduced in AlphaFold2 is adopted in the structure-aware Transformer, in order to efficiently learn the global geometric features (especially, long-range interactions) of a protein backbone. The Transformer finally outputs SLV for each residue in the protein. It contains sufficient information about the protein backbone and enables the Decoder to reconstruct the protein structure regardless of any trans-rotational transformation. It is noteworthy that no information about amino acid sequence is provided to the model, thus, the Encoder is completely agnostic about the side chain of the input protein.

Quantization Module & Protein Structure Decoder

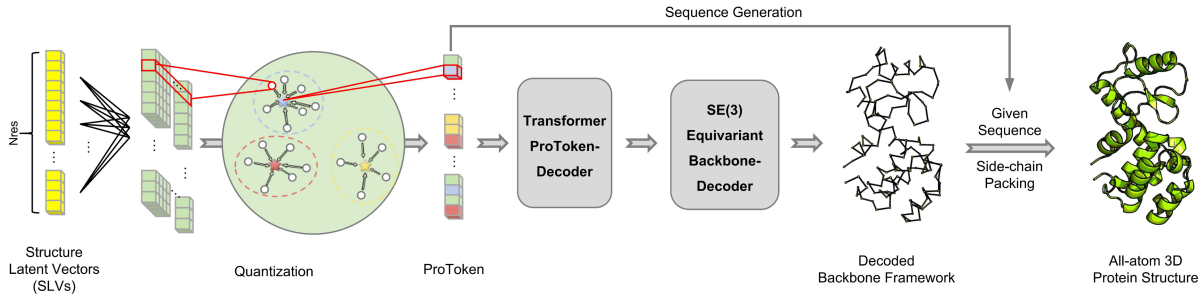

Fig. 2 Illustration of the Quantization Module and Decoder. The SLVs are first discretized into ProTokens through the Quantization Module. The Decoder transforms the ProTokens into the protein backbone 3D structure via an SE(3)-equivariant Transformer.

The SLVs produced by the Encoder are first discretized via a Quantization Module before being fed to the Decoder. As in VQ-VAE[37,40] and VQ-GAN[44,45], quantization is equivalent to clustering SLVs and degenerating SLVs within a cluster into one representation vector (also known as a "code"). Because the codebook (i.e., the set of cluster centers) is of finite size, quantization often plays the role of an "information bottleneck" in an auto-encoder network[36].

These quantized vectors are named as "ProTokens". In contrast to existing protein structure embedding methods[35] which are mainly based on manually extracted or local structural features, ProTokens are learned end-to-end from the spatial coordinates of the backbone, thus, naturally capturing both the local and global features about the protein structures.

The architecture of the Decoder is symmetric to that of Encoder. It is also composed of single and pair representation tracks, but equipped with an SE(3)-equivariant Transformer in order to directly produce protein 3D structures from ProTokens.

Assigned with a sequence of amino acids, either from a database (for protein structure compression) or inverse folding (for protein design), the decoded protein backbones can be further refined to all-atom structures by means of accurate side-chain packers[46,47].

3. Results & Applications

We trained the ProToken Distiller with varied codebook sizes over a curated protein structure dataset[48] containing well-folded protein chains without structural gaps. To demonstrate the efficacy of ProTokens for structure compression, we evaluated the quality of reconstructed protein structures decoded from ProTokens against the ground-truth structures in the validation set. Next, we investigated the structural features encoded by ProTokens, as well as the relation between ProTokens and amino acids, in order to elaborate on the feasibility of discretizing protein structures. Finally, we proposed ProTokens as a domain-specific language to describe

protein structure, and took advantage of powerful off-the-shell LLMs to solve domain-specific problems related to protein structures. By educating LLaMA, an open-sourced LLM, over the ProToken corpus, various challenging tasks, including structure design, inverse folding, as well as structure prediction, can be undertaken by the fine-tuned "Dr. LLaMA".

3.1 Efficient compression of protein 3D structures

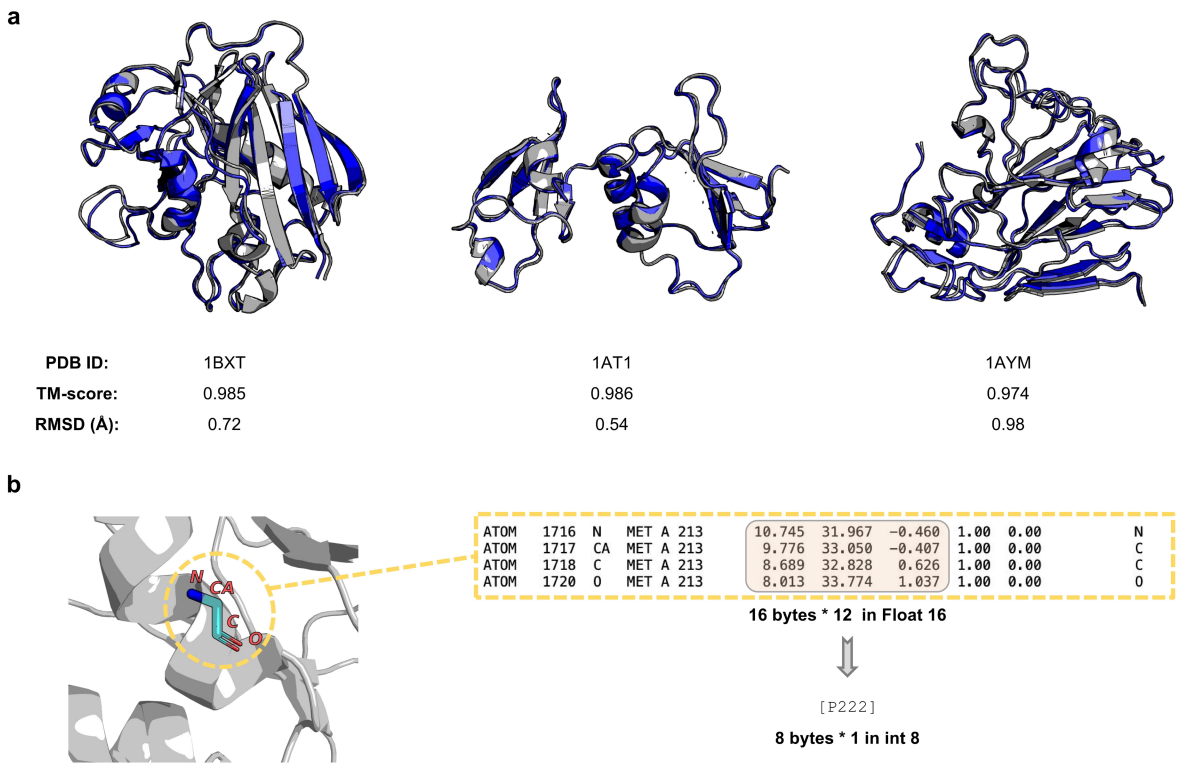

Fig. 3 Benchmark of ProTokens for structure compression. a. Alignment of reconstructed structures (blue) decoded from the compressed ProTokens against the ground-truth structures (gray) for protein 1BXT (PDB ID; left), 1AT1 (PDB ID; middle), and 1AYM (PDB ID; right), respectively. **b.** Comparison of storage format for protein backbone structures (PDB format v.s. ProToken format).

On the one hand, the rapid expansion of the protein structure database, either contributed by experiments[49] or augmented by computational tools[50], calls for efficient structure compression algorithms for the sake of tractable storage and mobility. On the other hand, although being widely used, the atomic positions are known to be redundant for recording protein structures, given the equivariance of Cartesian coordinates and the polymer nature of protein chains. Compression of redundant information into a more compact form is a central mission and trend for modern AI[51].

In typical file formats that record protein structures, such as mmCIF and PDB[52], for each residue 4×3=12 floats are needed to save the Cartesian coordinates (i.e., x, y, and z component, respectively) of the four backbone

heavy atoms (Fig. 3b). Thus, for a protein with N_{res} amino-acid residues, the backbone structure consumes up to $12 \times N_{\mathrm{res}} \times 16$ bytes (assuming the coordinates are stored in float16 format). In contrast, based on ProTokens, the backbone structure can be represented with Nres integers, taking merely $N_{\mathrm{res}} \times 16$ or $N_{\mathrm{res}} \times 8$ bytes (assuming the ProTokens are stored in int16 or int8 format, respectively), thus achieving a remarkable (more than 10-fold) compression ratio (Fig. 3b).

In spite of the promising compression efficiency, the fidelity of the proposed compression strategy, that is, the quality of the structures reconstructed from ProTokens, needs to be checked. The reconstruction quality is evaluated on a dataset containing more than 1k protein 3D structures. We compared the backbone structures of these proteins before and after compression. Given the ground-truth structures as reference, the average TM-score of the reconstructed structures exceeds 0.9, and the average RMSD is within 1.0 Å. Comparisons of several decoded structures and the corresponding ground-truth structures are illustrated in Fig. 3a, demonstrating that ProTokens can serve as an effective and high-fidelity compression method for protein 3D structures.

3.2 Comparing protein structures using ProTokens

Fig. 4 Benchmark of ProTokens for structure alignment and comparison. a. The structural dissimilarity matrices between proteins. Dissimilarity between the *i*-th and *j*-th proteins is defined as the Levenshtein distance between their ProTokens (lower triangle), or defined as 1-TMScore(*i,j*) (upper triangle), respectively. **b.** Alignment of two remote homologues from AlphaFold Database (AFDB) according to FoldSeek and ProTokens, respectively.

Comparison of protein structures is a common demand for various analyses involving protein structures, for instance, template searching, structure assessment, and homology detection, etc[35,53,54]. However, comparing two protein structures is not as straightforward as comparing two amino-acid sequences, because the spatial alignment is known to be time-consuming[54]. To this end, ProTokens remove the trans-rotational symmetries of protein structures, thus circumventing the need for spatial alignment. Moreover, like amino acids, ProTokens take the form of a sequence of discrete tokens (Fig. 4b). Thus, structures can be aligned efficiently with any method developed for sequence alignment[35,55]. As illustrated in Figure 4b, the structural discrepancy or

dissimilarity between two proteins can also be evaluated as the editing distance between their corresponding ProTokens.

To test whether ProTokens enable efficient alignment and comparison of protein structures, we propose Levenshtein distance between two ProToken sequences as a metric for structural discrepancy. Inspired by FoldSeek[35], a substitution cost matrix and an indel cost matrix are defined according to the similarity between ProTokens based on their secondary-structure preference and amino-acid preference, respectively (see Section 3.3 for more details). In Fig. 4a (lower triangle part), we show the computed ProToken editing distances between 54 proteins drawn from 3 FoldSeek[35] clusters. As a golden standard, the negative TM-Score (1 - TMScore) is computed for the same set of proteins (upper triangle part in Fig. 4a). It can be seen that ProTokens can cluster the protein folds well, forming three diagonal blocks in the distance matrix. Moreover, the dissimilarity score yielded by ProTokens also quantitatively agrees well with TM-Score, while ProToken editing distance can be computed orders-of-magnitude faster than TM-Scores. Such encouraging performance implies potential applications of ProTokens for (remote) homology detection and fast evaluation of protein structure similarities.

3.3 Physics interpretation of ProTokens

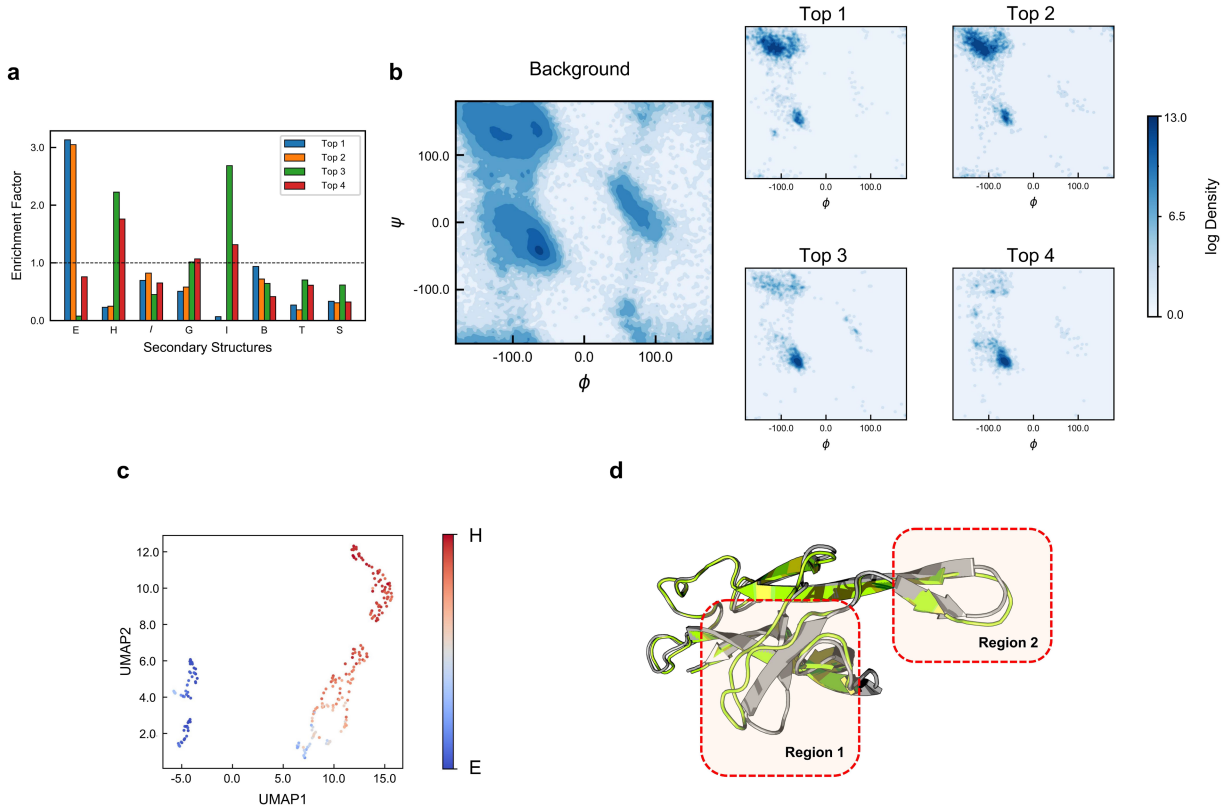

Fig. 5 Structural information encoded by of ProTokens. a. Secondary-structure preference for the top 4 mostly occupied ProTokens. Distribution of the secondary-structure elements in the validation set is chosen as the baseline for the enrichment factor. Symbols: 'E': β-strand, 'H': α-helix, 'G': helix-3, 'I': helix-5, 'B': β-bridge, 'T': turn, 'S': bend, '/': others. **b.** Ramachandra plots for the residues corresponding to the top 4 mostly occupied ProTokens. Ramachandra plot for the entire validation set is also shown

as reference. **c.** UMAP visualization of ProTokens. The similarity (or distance) metric is defined according to the distribution of backbone torsion angles. The relative enrichment of α-helix or β-strand residues is highlighted by red and blue, respectively. **d.** Mutating ProTokens causes changes in the decoded structures (green) to different degrees compared to the native structure (gray), where Region 1 contains five ProToken mutations.

It is appealing to investigate why ProTokens can perform well on structure compression and comparison. Since ProTokens are derived end-to-end from protein 3D coordinates, they are expected to contain structural information in different levels. To validate this assumption, we first explore the relationship between ProTokens and backbone torsional angles, which characterize critical local structural features of a protein chain. For each ProToken, we collect the torsional angles of the residues corresponding to it, and obtain the Ramachandra statistics. As shown in Fig. 5b (top 4 mostly used ProTokens), different ProTokens possess highly distinct patterns in the distribution of the backbone torsions. We also embed ProTokens into 2D space using uniform manifold approximation and projection[56] (UMAP) method according to the Wasserstein distance[57] computed based on the distribution of torsional angles (i.e., density of the Ramachandra plots as in Fig. 5b). The ProTokens show obvious cluster patterns in Fig. 5c, indicating that ProTokens are highly heterogeneous, and sensitive to local structural patterns of proteins.

Furthermore, we examine whether ProToken contains information about secondary structures. The secondary structure for each residue in the dataset is first annotated via DSSP[58]. We then plot the frequency histogram of the secondary structure elements occupied by those residues corresponding to a certain ProToken. The enrichment factor is proposed to quantify the preference of one ProToken over different secondary structure elements, which is defined as the ratio between the frequency of different secondary structures corresponding to a given ProToken, with respect to the background frequency drawn from the dataset. As shown in Fig. 5a, some ProTokens favor α-helices, while some exhibit a significant preference for β-strands, giving an explanation to the observed clusters in the Ramachandra plots (Fig. 5c)

In addition to local structural patterns, we also explore whether ProToken encodes non-local or long-range structural information, which may be key to the overall structural consistency and the reconstruction quality[11]. Specifically, given a sequence of ProTokens, we deliberately "mutate" one or more ProTokens into their nearest neighbor(s), defined as the ProToken sharing the most similar local structural preferences. The mutated ProToken sequence is then decoded into the 3D structure. As an example shown in Fig. 3f, the mutated ProTokens in Region 5 lead to non-negligible structural changes in a distal region of the protein (Region 2 in Fig. 5d) while keeping the local structures unchanged, reminiscent of the "allostery" phenomenon. This perturbation experiment demonstrates that the ProTokens are not only sensitive to local structural patterns, but also perceptive of the global structural interactions to some extent.

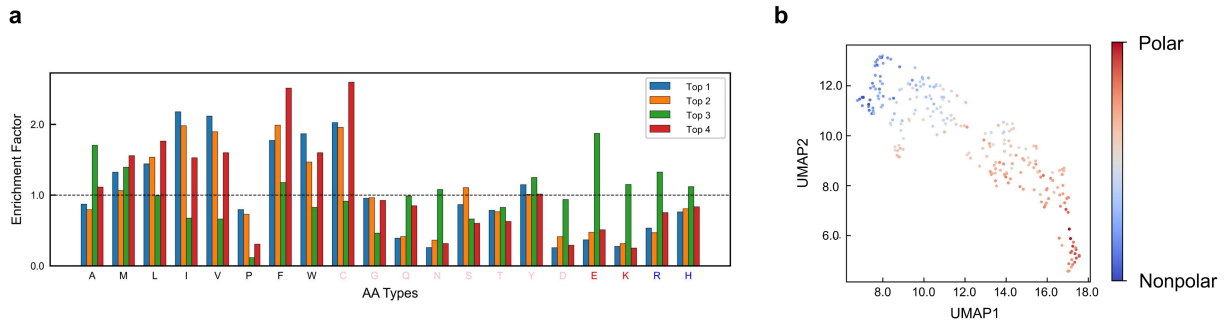

Fig. 6 Connections between ProTokens and amino acids. a. Amino-acid preference for the top 4 mostly occupied ProTokens. Distribution of the amino acids in the validation set is chosen as the baseline for the enrichment factor. Colors of residues: black for hydrophobic, pink for hydrophilic, red for acidic and blue for basic, respectively. **b.** UMAP visualization of ProTokens. The similarity (or distance) metric is defined according to the distribution of amino acid types. The relative enrichment of polar or nonpolar residues is indicated by red and blue, respectively.

Intuitively, it may seem unreasonable to discretize a continuous 3D structure into quantized tokens in terms of physics. However, we point out that such discretization is not only feasible but also solvable inspired by *nature*. Consider the following case: Suppose an ideal "inverse folder" serves as the Encoder, which can translate a well-folded backbone structure into the corresponding amino-acid sequence, and an ideal "structure predictor" serves as the Decoder, which can fold an amino-acid sequence to its native structure, then the vocabulary of natural amino acids becomes a natural solution to this discretization problem. Under this setting, ProTokens correspond exactly to amino acids, and they may arguably set the upper limit of the compression ratio for the foldable protein structures.

Due to the absence of the ideal inverse folder and the structure predictor, we have to trade the compression perplexity for the fidelity, that is, to expand the vocabulary size of ProTokens. However, we can still regard the Encoder, Decoder, and ProTokens as the "generalized" inverse folder, structure predictor, and amino-acid-like tokens, respectively. Following this reasoning, it is compelling to check whether ProTokens, though learned merely from backbone structures by a sequence-agnostic model, are related to amino acids. We calculate the enrichment factor for the ProToken so as to characterize its preference over amino acids. From Figure 6a it can be seen that amino acids corresponding to the ProToken are not uniformly distributed (note that the background distribution has been whitened). Some ProTokens (e..g, the rank-4 ProToken) favor hydrophobic amino acids including leucine and phenyl-alanine, etc., whereas some other ProTokens (e.g., the rank-3 ProToken) enrich more hydrophilic and charged amino acids including lysine, arginine, and glutamic acid, etc. Moreover, we embed ProTokens via UMAP according to their similarity in terms of the enrichment factor for amino acids (Fig. 6b), and find that ProTokens are heterogeneous in their relative preference over polar and non-polar amino acids. This observation corroborates that ProTokens bear close connections to amino acids, and that the ProToken Distiller can be interpreted as the unification of an inverse folding model and a structure prediction model.

3.4 ProTokens as a domain-specific language for LLMs

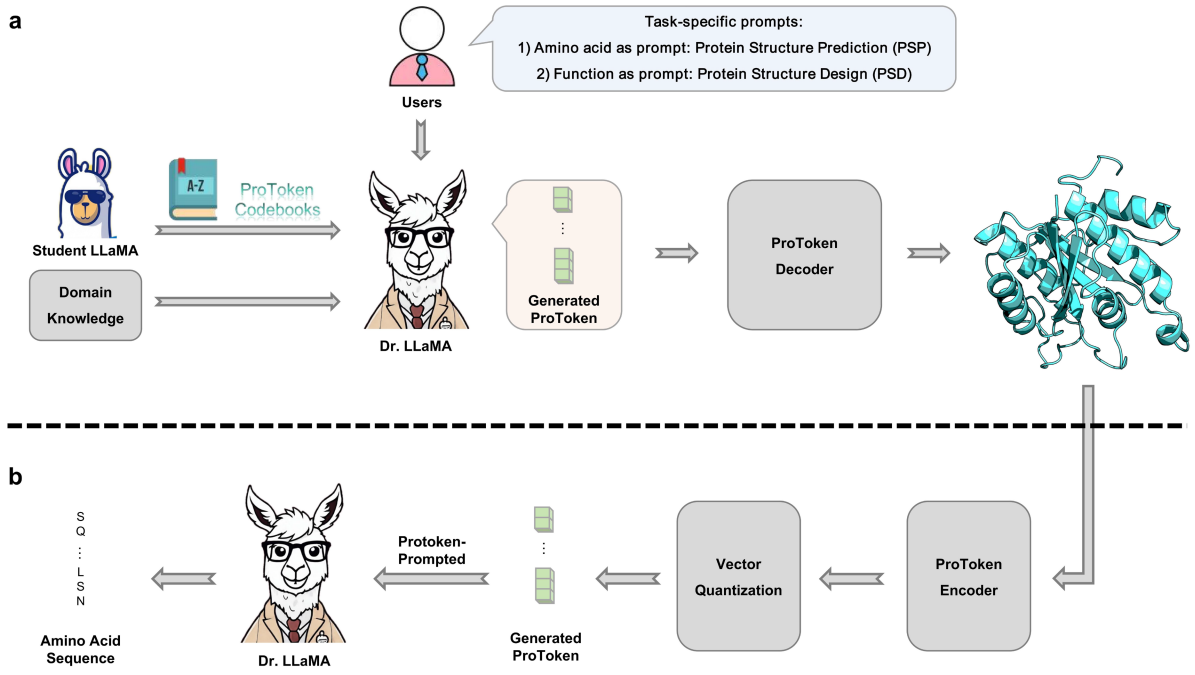

Fig. 7 Educating LLMs over ProTokens. a. Equipped with Decoder, Dr. LLaMA can produce protein 3D structures in the form of ProTokens, perform structure prediction according to the prompts specifying amino acid sequence, and perform structure design according to the prompts specifying desired functions. **b.** Equipped with Encoder, Dr. LLaMA can inversely fold a protein backbone to the corresponding amino acid sequence.

LLMs, such as GPT[59–61] and LLaMA[62,63], have shown impressive capabilities to process information based on text-formatted data, leading to a potential path to AGI. It is commonly believed that, transforming other modalities (including video, image, sound, etc.) to tokens, and teaching them as "foreign languages" to LLMs, may allow the recruitment of LLMs to solve domain-specific problems, or can extend LLM's ability to incorporate more diverse information. From this perspective, ProTokens naturally provide a vocabulary, presenting protein structures to LLMs. Thus, various protein-related tasks can be *unified* as sequence generation or machine-translation (or sequence-to-sequence) problems. As a very first attempt, we fine-tuned LLaMA with an expanded vocabulary including amino acid tokens and ProTokens, over instructions and tasks including protein structure prediction, design and inverse folding, etc. The fine-tuned model, named as “Dr. LLaMA”, is familiarized with the domain knowledge, hence, expected to help solve many tasks related to protein structures.

3.4.1 Machine translation between amino acids and ProTokens

On the basis of decodable ProTokens, the protein structure prediction problem, which refers to sampling the most probable structure(s) which a given amino-acid sequence folds into, can now be recast as a translation task from an amino-acid sequence to a ProToken sequence (Fig. 7a). Mathematically, given an amino-acid sequence

s, and its structure **x** (obtained from, e.g., experiments), we first transformed the structure **x** into ProTokens **z**, via the Encoder (Section 2.1).

On the other hand, the "inverse folding" task, that is, generating an amino-acid sequence **s** to stabilize a given protein backbone **x**, can also be formulated as a machine translation task from the ProToken **z** corresponding to **x** to **s** (Fig. 7b).

Following the convention in machine translation, we performed supervised fine-tuning (SFT) of LLaMA[62,63] with LoRA[64], following the instructions (Fig. 9) that specify the demand to translate **s** into **z** or vice versa. The training can be efficiently parallelized by means of teacher-forcing. Preliminary results demonstrate that Dr. LLaMA is able to fulfill both the structure prediction (Fig. 9a) and inverse folding (Fig. 9b) tasks well once fine-tuned properly..

3.4.2 De novo design of proteins with LLMs

Up to date, protein structure design is almost exclusively achieved via a continuous optimization (or generation) process over the protein coordinate space, e.g., with regard to a score function defined by physics heuristics[65], data mining[5,17,25], or qualified structure predictors[7,66]. In contrast, since protein structures have been discretized as ProTokens, generating a protein 3D structure now becomes no different from generating a text sequence.

Mathematically, it can be proved that maximizing the likelihood of a ProToken sequence is equivalent to maximizing a likelihood lower bound of the corresponding protein 3D structure,

$$p(\mathbf{x}) = \int p(\mathbf{z})\, \delta\big(\mathbf{f}_{\phi}^{\mathrm{Dec}}(\mathbf{z}), \mathbf{x}\big) \mathrm{d}\mathbf{z} \geq \int_{\mathbf{z}=\mathbf{f}_{\boldsymbol{\theta}}^{\mathrm{Enc}}(\mathbf{x})} p(\mathbf{z})\, \delta\big(\mathbf{f}_{\phi}^{\mathrm{Dec}}(\mathbf{z}), \mathbf{x}\big) \mathrm{d}\mathbf{z} = p\left(\mathbf{f}_{\boldsymbol{\theta}}^{\mathrm{Enc}}(\mathbf{x})\right)$$

where $\mathbf{f}_{\boldsymbol{\theta}}^{\mathrm{Enc}}(\cdot), \mathbf{f}_{\phi}^{\mathrm{Dec}}(\cdot)$ represent the encoder and the decoder, respectively. It is well-known that transformer-based autoregressive models[67] are particularly good at learning the distribution of sequences. Therefore, we fine-tuned a Transformer model to maximize the likelihood of ProTokens in the dataset, and new ProToken sequences can be easily sampled by a proper decoding strategy[68].

Moreover, ProTokens enable the joint generation of a "protein tuple", that is, a 3D structure (or the ProToken sequence) along with an amino-acid sequence (Fig. 8b). Mathematically, given the vocabulary of ProTokens $z \in \mathbb{Z}_n$, and the vocabulary of amino acid as $a \in \mathbb{Z}_m$, the Cartesian product of the two sets is defined as,

$$w = (z, a) \in \mathbb{Z}_n \otimes \mathbb{Z}_m = \mathbb{Z}_{nm}$$

The distribution of the protein tuple sequence $\mathbf{w}$, can also be conveniently learned by a Transformer-based model, which can efficiently generate a protein 3D structure and the associated sequence that are compatible with each other, thus providing a new and principled approach for protein designs.

A typical workflow for joint design of the protein structure and sequence is illustrated in Fig. 8b. Through experiments we show that, once a Transformer model (such as LLaMA) is trained properly and achieves low

perplexity, the ProTokens (or protein tuples) produced by the Transformer with low perplexity can reasonably fold into stable and well-defined 3D structures via the Decoder as shown in Fig. 8a, leading to a new approach to protein structure design.

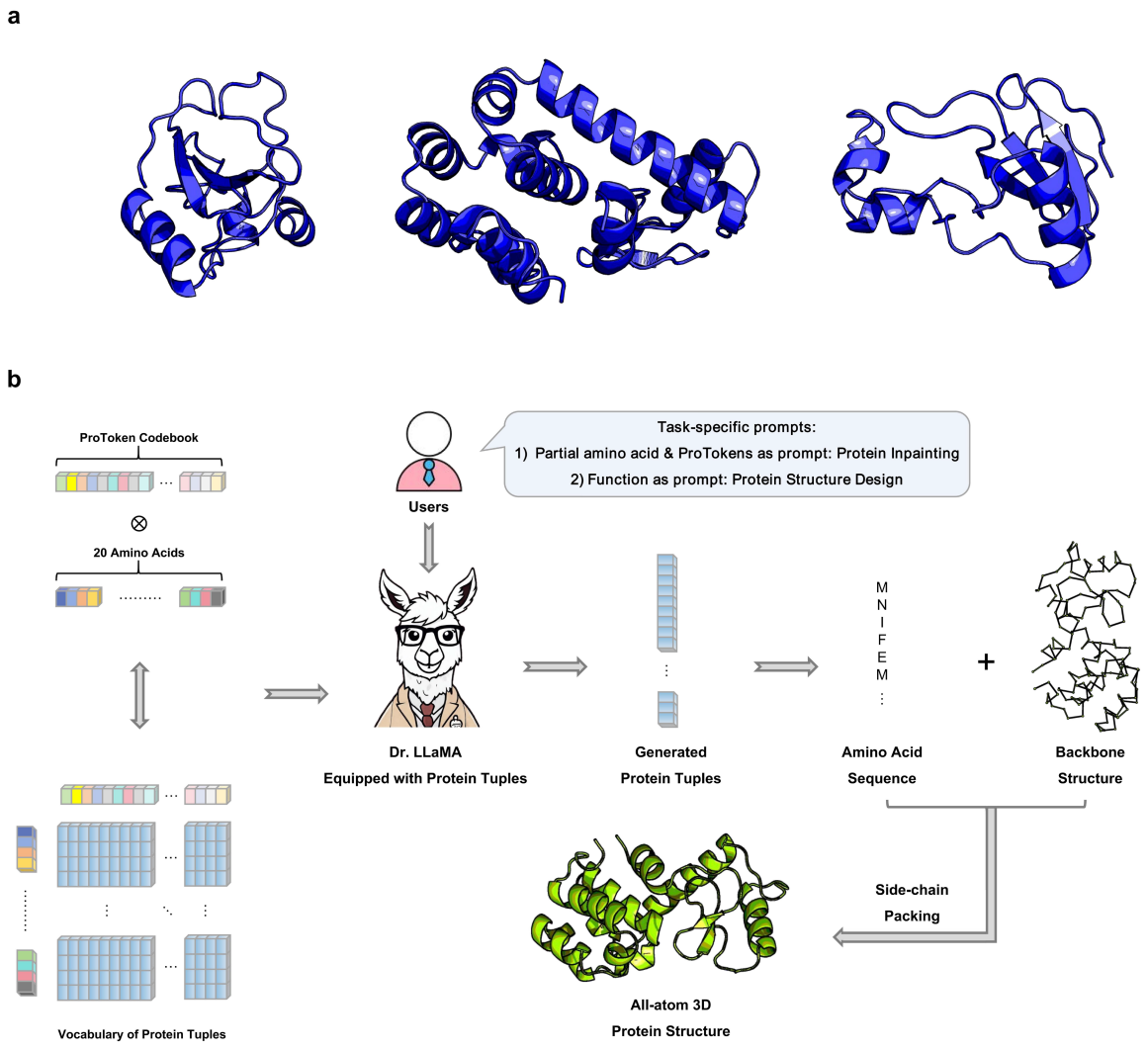

Fig. 8 Design protein structures via ProTokens. a. Showcases of designed structures from ProTokens generated by a Transformer. **b. Workflow** for joint design of protein tuples. The Cartesian product of the ProToken codebook and the set of 20 amino acids yields a new vocabulary, i.e., protein tuple. Based on the protein tuple, Dr. LLaMA can accept user-specified prompts, to jointly design the protein sequence and backbone structure,z which can be further fine-grained to the all-atom level through existing side-chain packers.

The distribution of the protein tuple sequence $\mathbf{w}$, can also be conveniently learned by a Transformer-based model. The workflow for joint design of the protein tuple is illustrated in Fig. 8b. Similar to ProToken Transformer, we also proved that a well-trained protein-tuple Transformer can efficiently generate a protein

3D structure and the associated sequence which are compatible with each other, thus, providing a new and principled approach for protein designs.

a

Users:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. Instruction: Please translate the following sequence into the corresponding ProTokens.

Input:

[A][S][A][G][V][A] ..... [G][L][A][F][K][S] [I][T][T][E]

Dr. LLaMA:

Output:

[P39][P7][P4][P198][P5][P97] ..... [P55][P223][P6][P195][P7][P12] [P195][P4][P96][P11]

b

Users:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. Instruction: Please translate the following ProTokens into the corresponding sequence.

Input:

[P39][P7][P4][P198][P5][P97] ..... [P55][P223][P6][P195][P7][P12] [P195][P4][P96][P11]

Dr. LLaMA:

Output:

[A][S][A][G][V][A] ..... [G][L][A][F][K][S] [I][T][T][E]

Fig. 9 Exemplary chat-boxes of Dr. LLaMA. Following the given instruction, Dr. LLaMA translates protein sequence to ProTokens or vice versa, to perform: **(a)** single sequence structure prediction, and **(b)** inverse folding tasks, respectively.

4. Concluding Remarks

Proteins serve as the most fundamental and important functional molecules for life. Understanding the relationship between the constituent of a given protein and its function remains central in various research scopes. It is commonly assumed that the function of a protein is largely determined by its 3D structure, and that the structure is mostly dictated by its amino acid sequence. Protein structure thus plays a central role linking its sequence and function, and two widely concerned problems naturally arise: 1) How to predict the functional protein structure given its sequence, and 2) how to generate proteins (including structures and sequences) to

fulfill a specific function. AlphaFold2 has achieved remarkable progress on the former challenge, however, the latter question remains wide open and is still calling interest from an active research community for protein design.

In principle, protein design requires joint generation of a protein structure and a sequence (i.e., a "protein tuple") which are mutually "compatible", that is, the sequence can stably fold into the designed structure. However, in common practice, the generation process is factorized: a protein backbone structure is first designed according to the desired functions, then an amino-acid sequence is assigned in order to stabilize the designed structure, completing the "protein tuple". Such factorization rests on a strong but often overlooked assumption of "designability", namely, there exists at least one sequence combo of (natural) amino acids that can fold dominantly if not exclusively to the designed protein structure. However, although structures can be generated with arbitrary variety in the coordinate space, these backbones may not be stabilized by any combination of natural amino acids, resulting in failure of the design. Unfortunately, the high failure risk of many factorized design approaches reflects the fact that our understanding of the "designable space" of protein structures is still quite limited. Therefore, proper modeling of the designable space of protein structures is urgently needed to improve the success of factorized protein design. Recently, data-driven generative methods are gaining increasing attention and provide new opportunities to this long-standing challenge. However, 3D protein structures are composed of atomic coordinates which exhibit specific symmetries and are subjected to physics priors, such as transrotational equivariance and polymer restraints, hence, significantly impeding the transaction of modern generative AI for protein structure design.

In this research, we developed an unsupervised learning strategy, which integrates the structure prediction and the inverse folding tasks, to encode protein structures into discrete ProTokens and decode them back to 3D coordinates. Although discretizing protein structures which lie in the continuous coordinate space into tokens may seem unreasonable at first sight, we demonstrated that this problem is indeed well-defined and solvable. Particularly, as one special solution to this problem, the ProToken encoder can be an ideal "inverse folder" which translates a backbone structure into a sequence of amino acids which stabilizes the structure; whereas the decoder can be an ideal "folder" which predicts the protein structure given the amino-acid sequence. In this ideal case, ProTokens are equivalent to the vocabulary of natural amino acids, and the designable protein space is compressed compactly. However, due to the difficulty in obtaining or consulting the "ideal" encoder and decoder, the compactness and informativeness of ProTokens has to be balanced. We showed that discretizing protein structures with proper perplexity can lead to informative and relatively compact ProTokens which can reconstruct 3D coordinates with high fidelity and reduce the trans-rotational equivariance as well as the polymer restraints of protein structures. Intriguingly, although ProTokens are learned end-to-end by a sequence-agnostic model which only inputs and outputs protein backbone structures, it is found that the extracted ProTokens still show close relations to the type or property of natural amino acids. Such a subtle linkage between ProTokens and amino acids suggests that the "designable space" of protein structures may be intrinsically discrete and even compact.

As appealing applications, protein structures can be efficiently compressed, stored, aligned and compared in the form of ProTokens. Besides, owing to the removal of group symmetry and polymer restraints, many modern generative methods, including VAE, flow and diffusion models, to name a few, can be straightforwardly applied on the ProToken latent space to generate protein structures. Autoregressive models can also be recruited to

directly sample ProToken sequences for the same purpose. Particularly, it is shown that ProTokens enables joint design of protein structure and sequence simultaneously which is hardly achievable by existing protein design approaches, circumventing the potential issues of factorized designing strategy.

Additionally, ProTokens transform the modality of protein structures into text-like tokens which are amenable for LLMs. Therefore, the problem of protein structure prediction (or inverse folding) can be recast as a machine translation task from amino-acid sequences to ProToken sequences (or vice versa), which falls within the reach of capable LLMs. We showed that once LLaMA is educated over the ProToken vocabulary, it can incorporate the domain-specific information (i.e., protein structures) effectively with the potentiality to demystify the structural semantics of proteins and help explore the protein universe.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2022ZD0115002). The authors thank Dr. Sirui Liu and Dr. Xing Che for useful discussion, and gratefully acknowledge the support from Huawei MindSpore team to this research.

5. References

1. Voet, D., Voet, J. G. & Pratt, C. W. *Fundamentals of biochemistry: life at the molecular level*. (John Wiley & Sons, 2016).
2. Laskowski, R. A., Watson, J. D. & Thornton, J. M. From protein structure to biochemical function? *Journal of Structural and Functional Genomics* **4**, 167–177 (2003).
3. Zhang, C. & Kim, S.-H. Overview of structural genomics: from structure to function. *Current opinion in chemical biology* **7**, 28–32 (2003).
4. Wild, D. L. & Saqi, M. A. Structural proteomics: inferring function from protein structure. *Current Proteomics* **1**, 59–65 (2004).
5. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
6. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
7. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
8. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
9. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. Preprint at http://arxiv.org/abs/2205.15019 (2022).
10. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. (2022).
11. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
12. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
13. Qing, R. *et al.* Protein Design: From the Aspect of Water Solubility and Stability. *Chem. Rev.* **122**, 14085–14179 (2022).

14. Lutz, I. D. *et al.* Top-down design of protein architectures with reinforcement learning. *Science* **380**, 266–273 (2023).
15. Strokach, A. & Kim, P. M. Deep generative modeling for protein design. *Current opinion in structural biology* **72**, 226–236 (2022).
16. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).
17. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* (2023) doi:10.1038/s41586-023-06728-8.
18. Goodfellow, I. *et al.* Generative adversarial nets. *Advances in neural information processing systems* **27**, (2014).
19. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020).
20. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32**, (2019).
21. Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).
22. Papamakarios, G., Pavlakou, T. & Murray, I. Masked autoregressive flow for density estimation. *Advances in neural information processing systems* **30**, (2017).
23. Chen, X., Mishra, N., Rohaninejad, M. & Abbeel, P. Pixelsnail: An improved autoregressive generative model. in *International Conference on Machine Learning* 864–872 (PMLR, 2018).
24. Salimans, T., Karpathy, A., Chen, X. & Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517* (2017).
25. Jing, B. *et al.* EIGENFOLD: GENERATIVE PROTEIN STRUCTURE PREDICTION WITH DIFFUSION MODELS. (2023).
26. Lu, C. *et al.* Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems* **35**, 5775–5787 (2022).
27. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
28. Sinha, A., Song, J., Meng, C. & Ermon, S. D2C: Diffusion-Decoding Models for Few-Shot Conditional Generation. in *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 12533–12548 (Curran Associates, Inc., 2021).
29. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021).
30. Ho, J. & Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
31. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
32. Banerjee, A. & Arora, V. wav2tok: Deep Sequence Tokenizer for Audio Retrieval. in *The Eleventh International Conference on Learning Representations* (2023).
33. Ryoo, M., Piergiovanni, A., Arnab, A., Dehghani, M. & Angelova, A. TokenLearner: Adaptive Space-Time Tokenization for Videos. in *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 12786–12797 (Curran Associates, Inc., 2021).

34. Kim, H., Mirdita, M. & Steinegger, M. Foldcomp: a library and format for compressing and indexing large protein structure sets.
35. Van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01773-0.
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10674–10685 (IEEE, 2022). doi:10.1109/CVPR52688.2022.01042.
37. Van Den Oord, A. & Vinyals, O. Neural discrete representation learning. *Advances in neural information processing systems* **30**, (2017).
38. Mentzer, F., Minnen, D., Agustsson, E. & Tschannen, M. Finite Scalar Quantization: VQ-VAE Made Simple. Preprint at http://arxiv.org/abs/2309.15505 (2023).
39. Liu, L., Dong, C., Liu, X., Yu, B. & Gao, J. Bridging Discrete and Backpropagation: Straight-Through and Beyond. Preprint at http://arxiv.org/abs/2304.08612 (2023).
40. Razavi, A., Van den Oord, A. & Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* **32**, (2019).
41. Chang, H., Zhang, H., Jiang, L., Liu, C. & Freeman, W. T. Maskgit: Masked generative image transformer. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11315–11325 (2022).
42. Zhang, J. *et al.* Unsupervisedly Prompting AlphaFold2 for Accurate Few-Shot Protein Structure Prediction. *J. Chem. Theory Comput.* acs.jctc.3c00528 (2023) doi:10.1021/acs.jctc.3c00528.
43. Zhang, J., Zhou, Y., Lei, Y.-K., Yang, Y. I. & Gao, Y. Q. Molecular CT: Unifying Geometry and Representation Learning for Molecules at Different Scales.
44. Esser, P., Rombach, R. & Ommer, B. Taming transformers for high-resolution image synthesis. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 12873–12883 (2021).
45. Yu, J. *et al.* Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627* (2021).
46. McPartlon, M. & Xu, J. An end-to-end deep learning method for protein side-chain packing and inverse folding. *Proceedings of the National Academy of Sciences* **120**, e2216438120 (2023).
47. Zhan, Y., Zhang, Z., Zhong, B., Misra, S. & Tang, J. DiffPack: A Torsional Diffusion Model for Autoregressive Protein Side-Chain Packing. *arXiv preprint arXiv:2306.01794* (2023).
48. Liu, S. *et al.* PSP: million-level protein sequence dataset for protein structure prediction. *arXiv preprint arXiv:2206.12240* (2022).
49. Burley, S. K. *et al.* Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols* 627–641 (2017).
50. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **50**, D439–D444 (2022).
51. Delétang, G. *et al.* Language modeling is compression. *arXiv preprint arXiv:2309.10668* (2023).
52. Westbrook, J. D. & Fitzgerald, P. M. The PDB format, mmCIF, and other data formats. *Methods Biochem Anal* **44**, 161–179 (2003).
53. Holm, L. Using Dali for protein structure comparison. *Structural Bioinformatics: Methods and Protocols* 29–42 (2020).
54. Zhang, Y. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302–2309 (2005).

55. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
56. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at http://arxiv.org/abs/1802.03426 (2020).
57. Peyré, G. & Cuturi, M. Computational Optimal Transport. Preprint at http://arxiv.org/abs/1803.00567 (2020).
58. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
59. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training.
60. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners.
61. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at http://arxiv.org/abs/2005.14165 (2020).
62. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at http://arxiv.org/abs/2302.13971 (2023).
63. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at http://arxiv.org/abs/2307.09288 (2023).
64. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at http://arxiv.org/abs/2106.09685 (2021).
65. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. in *Methods in Enzymology* vol. 383 66–93 (Elsevier, 2004).
66. Norn, C. *et al.* Protein sequence design by explicit energy landscape optimization. *BioRxiv* 2020.07. 23.218917 (2020).
67. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30**, (2017).
68. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).