ProTokens: Probabilistic Vocabulary for Compact and Informative Encodings of All-Atom Protein Structures

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

Affiliations:

1. Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.
2. Changping Laboratory, Beijing 102200, China.
3. Institute of Artificial Intelligence, Soochow University, Suzhou 215006, 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 develop *probabilistic tokenization theory* for metastable protein structures. We present an unsupervised learning strategy, which conjugates inverse folding with structure prediction, to encode protein structures into amino-acid-like tokens and decode them back to atom coordinates. We show that tokenizing protein structures *variationally* can lead to compact and informative representations (ProTokens). Compared to amino acids — the Anfinsen's tokens — ProTokens are easier to detokenize and more descriptive of finer conformational ensembles. Therefore, protein structures can be efficiently compressed, stored, aligned and compared in the form of ProTokens. By unifying the discrete and continuous representations of protein structures, ProTokens also enable all-atom protein structure design via various generative models without the concern of symmetry or modality mismatch. We demonstrate that generative pretraining over ProToken vocabulary allows scalable foundation models to perceive, process and explore the microscopic structures of biomolecules effectively.

I. 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 3-dimensional (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 artificial intelligence (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. In contrast to the 3D structure, the amino acid tokens of a protein, also known as the first order or 1D structure, is formally discrete and SE(3)-invariant, thus being a compelling candidate as input to the LLM. Many methods have been developed for protein understanding and designing tasks which educate LLMs over amino acid sequences[34,35]. This paradigm is plausible given Anfinsen's hypothesis[36] that the protein 3D structure is fully determined by its amino acid sequence in most cases.

Consequently, the modality difference of protein structure representations (1D v.s. 3D) causes significant divergence in the research paradigms of proteins, particularly in the realm of protein design. In this research, we provide a novel and unified perspective for protein 1D and 3D structures based on protein physics, with an important conclusion that although protein 3D structures are continuous, the function-related conformational ensembles (or metastable states) are countable at a proper observation timescale. Particularly, the conventionally defined 1D structure of proteins (i.e., amino acid sequence) is equivalent

to a probabilistic tokenization of protein 3D structure ensemble given a large observation timescale when all the (un-)folded structures collapse to a single metastable state. Although being compact, amino acid tokens bear some shortcomings in terms of informativeness, including the difficulty of being tokenized to 3D structures and the mode collapse issue that alternative conformations corresponding to different functions become degenerate.

In this research, starting from the *Anfinsen's tokens* (i.e., the set of amino acids), we attempt to expand the probabilistic token vocabulary (ProTokens) for protein 3D structures, trying to strike better balance between the compactness and informativeness of these discretized tokens. ProTokens are learned via an unsupervised *ProToken Distiller*, intuited by the profound connections between structure prediction and inverse folding. ProTokens can be faithfully decoded back and reconstruct the metastable protein structures with high quality while the trans-rotational equivariance and polymer restraints over protein coordinates are reduced, thus allowing protein structures to be compressed, aligned and compared efficiently. Moreover, ProTokens are amenable for both LLM-based and diffusion-based foundation models due to the unified modality of 1D and 3D protein structures.

In summary, **our main contributions are five-fold**:

1. We propose and justify the probabilistic tokenization of all-atom protein structures, by factoring the continuous distribution of spatial coordinates into discrete parts representing function-relevant metastable states and continuous parts accounting for the conformational fluctuations within the metastable state.
2. We develop a data-driven and physics-informed approach to extracting amino-acid-like ProTokens in an unsupervised and variational way. The ProTokens are compact and informative representations for all-atom protein structure ensembles.
3. We unify the 1D and 3D modality of protein structures by merging amino acids as subset of ProToken vocabulary, and equip ProTokens with *Janus* representations, making them ready for both discrete language-based foundation models and continuous diffusion-based foundation models.
4. We summarize caveats and pitfalls of applying a transformed representation like ProTokens in tasks related to protein structures and develop mathematical guidelines to ameliorate these issues.
5. We present a simple and scalable pretraining objective based on ProTokens and demonstrate the pretraining yields zero-shot generalization (or emerging) capability for various tasks related to protein structures.

II. Theory & Methods

A. Transform Representation of Protein 3D Structures

Protein 3D structures are usually presented as the spatial coordinates of the atoms which belong to the SE(3)-symmetry group. Such symmetry makes it unwise to treat 3D coordinates of proteins directly as input or output of any symmetry-free models including the state-of-the-art foundation models[37–39]. Furthermore, protein is a special kind of biopolymer, and its structure is subjected to physics restraints such as peptide bonding interactions which are usually not relevant to conformational changes or its functions. Consequently, the spatial coordinates of protein structure are a redundant representation and not amenable to modern AI.

1. Can continuous 3D structures be reasonably discretized?

Transforming Cartesian coordinates of a protein 3D structure $\mathbf{x} \in \mathbb{R}^{3\times N_{\text{res}}}$ into a SE(3)-invariant representation $\mathbf{z} \in \mathbb{R}^{d}$ is non-trivial considering that this transform is inevitably invertible due to the alteration of the symmetry. Intuitively, such a transform may be learned through an auto-encoding-like training with symmetry-specific encoder and decoder architectures[40]. However, with finite amount of data, there exist infinite transforms $f\colon \mathbf{x} \to \mathbf{z}$, that can fit the data asymptotically perfectly. Proper regularization is needed to make this transform reasonable in terms of mathematics and physics. In this work, we are focusing on protein structures that attribute to the protein's functions. Such structures should come from metastable states[41] (or metastable conformations) in terms of protein physics, because they should exhibit sufficiently long lifetime in order to play the role. The metastable state is a consequence of the separation of timescales of protein dynamics[42,43]. The conformational change within a metastable state is defined as intrastate relaxation or fluctuation. The timescale of relaxation is termed as relaxation time τ_{rlx}, which should be much smaller than the lifetime τ_{life} of the metastable state.

We note that, although the structure of a protein can be continuously changed in the Cartesian space, the *set of its metastable states* are countable, hence, can be well-defined discretely according to the landscape theory[44,45]. Specifically, the definition of metastable states depends on the observation timescale τ_{obs}: A metastable can be only defined if $\tau_{\text{rlx}} < \tau_{\text{obs}} < \tau_{\text{life}}$. According to the landscape theory[46,47], the smaller τ_{obs} is, the larger amount of metastable states can be defined[48,49] (Fig. 1a). In a limit case, when τ_{obs} is comparable to the (un-)folding timescale (τ_{fold}) of a protein, most proteins are known to exhibit a two-state kinetics between the folded and unfolded states, hence, the only folded metastable state can be well defined by the amino-acid sequence of the protein.

2. Probabilistic tokenization of protein 3D structures

Thanks to the metastability, a continuous distribution of function-related protein structures can now be reasonably factored into discrete and continuous parts,

$$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} = \int p(\boldsymbol{\epsilon}|\mathbf{z}(\mathbf{x}))p(\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\epsilon} \tag{1}$$

where the discrete part $\mathbf{z}$ stands for the metastable state, $p(\mathbf{z})$ is a discrete distribution specifying the number of metastable states which $\mathbf{x}\sim p(\mathbf{x})$ consists of, and the continuous part $p(\boldsymbol{\epsilon}|\mathbf{z}(\mathbf{x}))$ covers the intrinsic structure fluctuations within this metastable state. The first equation holds because $\mathbf{z}$ is a deterministic function of $\mathbf{x}$.

Equation 1 is the foundation of our *probabilistic tokenization theory* for protein structures. It justifies that a discrete prior (denoting the metastable states) for continuous protein structures is reasonable. It also implies that, if the metastable states are few (given a large τ_{obs}), $\boldsymbol{\epsilon}$ should account for large and complicated variations within each state. On the contrary, if the metastable states are defined at a small timescale, $\boldsymbol{\epsilon}$ is only responsible to very subtle structural variations within each state, but the number of $\mathbf{z}$ could quickly explode. Put it in other words, there is a trade-off between the *compactness* of $\mathbf{z}$ (i.e., the number of defined metastable states) and the *informativeness* of $\mathbf{z}$ (i.e., the residual intra-state variation that has to be explained by $\boldsymbol{\epsilon}$), as illustrated in Fig. 1a.

Inspired by the lattice model for proteins[50], it is plausible that the number of metastable states grow with the length of proteins, so we can assign a finite number of discrete states to each residue (namely, tokens), and the combination of these residue-wise tokens, define the overall state of the protein. From this perspective, amino acid(s) is indeed one kind of such token, i.e., *Anfinsen's token*(s). Anfinsen's tokens are *probabilistic* in nature because they do not correspond to a single snapshot of protein 3D structure, but to all the folded structures with $\tau_{\text{obs}} \approx \tau_{\text{fold}}$ according to the Anfinsen's hypothesis. However, Anfinsen's tokens are extremely compact (with a small vocabulary size of only 20), thus, leaving the intrastate variations large and hard to estimate. That explains why conformational prediction based on amino acids is a tremendously challenging task. Due to the absence of an effective *detokenization* algorithm (although significant advance has been made since AlphaFold[11,51,52]) which can trustworthily map back amino acids to the folded conformations of the protein, they are often not regarded as a 3D representation for protein structures.

Compared to Anfinsen's tokens, tokenizing metastable states at finer timescale $\tau_{\text{obs}} < \tau_{\text{fold}}$ has several compelling advantages: i) more detailed changes in structures corresponding to functional switch can be described including alternative conformations; ii) tokens can strike good balance between being compact and being informative; iii) a more efficient detokenizer can be obtained to backmap the tokens. Specifically, given a protein 3D structure $\mathbf{x}$, we tokenize the metastable structure ensemble $\{\mathbf{x}\}$ associated with $\mathbf{x}$ into probabilistic, amino acid-like tokens, which can be detokenized back to conformations from the corresponding metastable state.

3. ProTokens: Factorized tokenization of all-atom protein 3D structures with unimodality and Janus representations

Due to the fact that the relaxation of side chains is often much faster than the backbones, metastable conformations of the backbone and sidechain can be tokenized separately[53]. The token for an all-atom protein 3D structure ensemble, termed as ProToken, is thus defined by the Cartesian product of the backbone token and the sidechain token.

A *wildcard backbone token* is introduced: When it is combined with any sidechain token in the form of amino acid, it becomes a valid ProToken equivalent to Anfinsen's token. Particularly, this special subset of

ProTokens degenerately encodes all the folded conformations of the protein. Therefore, ProToken unifies the modality of 1D and 3D representations of protein structures. In addition to the unified modality, we also equip each ProToken with *Janus representations*, that is, two mutually mappable representations: One is a discrete token index, which is amenable for language-based foundation models; The other is a continuous symmetry-free token embedding, which is ready for diffusion-based foundation models.

We devise an unsupervised deep learning approach called *ProToken Distiller* to extract ProTokens automatically from experimental and physics-simulated data. Inspired by the profound connections between ProTokens and amino acids, the *ProToken Distiller* is indeed a joint optimizer for the classical inverse folding and the structure prediction problems (Fig. 1b), except that the conventional predefined amino acid vocabulary is replaced by a learnable set of ProTokens so that the gradient flow can be back-propagated end to end.

B. Pitfalls of a Learned Protein Structural Representation

After transforming the spatial coordinates of metastable structures into SE(3)-invariant (continuous or discrete) representations like ProTokens, one may apply them for downstream tasks related to protein structures. However, we reveal that high risk exists if the transformed representations are not optimized or implemented with caveats, and summarize several potential pitfalls of learning and applying a transformed representation. Awareness of these issues are reflected in the specifically designed model components and training objectives which will be elaborated in the following sections.

1. The space of the transformed representation should be compact.

Taking ProTokens of a N_{res}-long protein as example, the real-valued token embeddings take the shape of (N_{res}, d). If $d \gg 3\bar{N}_{\mathrm{atom}}$ (where $\bar{N}_{\mathrm{atom}}$ stands for the average number of atoms per residue), diffusion in the transformed embedding space will be less efficient (both in terms of data and training) than spatial diffusion approaches like RFDiffusion and AlphaFold3[5,51], provided that the likelihood of the model decays exponentially with the dimensionality of the representation.

Besides, the number of all possible combinations of ProTokens grows combinatorially with respect to the vocabulary size K. Since the number of foldable protein structures (as well as functionally relevant metastable states) is unlikely to exceed $20^{N_{\mathrm{res}}}$, the reasonable number of tokens used during training should lie in the range between tens to a few hundreds.

To extract compact ProTokens, we design a *Compressing* module (see Section II-C and SI Section I-C. for more details) in the *ProToken Distiller* to properly compress the ProToken embeddings lengthwise or depth-wise. Besides, we derive a variational information bottleneck (VIB)[54] to quantify the necessity of ProTokens , and the vocabulary size K is optimized with respect to the VIB loss by *variational clustering* during training (Eq. 11; see Section II-D and SI Section I-C for more details).

2. The transformed representation should be robust against intrinsic structure fluctuations.

In terms of protein physics, fluctuations are intrinsic to metastable protein structures, which may cause subtle structure changes but do not alter the function. One particular concern of a learned structural representation is that the *robustness* of the yielded embeddings and tokens against subtle structure

perturbation may not be guaranteed (Fig. 1c), while the susceptibility to intrinsic fluctuations will be harmful to downstream tasks such as function predictions.

To alleviate this issue, we introduce adversarial examples by adding physics-informed fluctuations to any input structure, and adversarially train the model to behave robustly to the *adversarial attacks*. Besides, we also train a *Deduplicator* module to relax fluctuated structures within a metastable ensemble towards a single stable representative structure, and collapse the embeddings of the fluctuated structures (see Section II-C and SI Section I-B for more details).

3. Existence v.s. uniqueness: The degeneracy of transformed representation should be addressed.

Noteworthy, the *uniqueness* of a learned ProToken corresponding to a certain structure $\mathbf{x}$ cannot be guaranteed through data-driven training. *Duplicate* or degenerate ProTokens may exist that can be decoded to (almost) the same structure (Fig. 1c). Degeneracy is particularly poisonous when ProTokens are used for maximum likelihood estimation (MLE) of protein structures by generative models such as auto-regressive and diffusion models. To be more specific, after transforming the representation of protein structure $\mathbf{x}$ to ProToken $\mathbf{z}$, which can be detokenized back via a decoder $\mathbf{x} = g_\phi(\mathbf{z})$, the (log-)likelihood of the structure $\mathbf{x}$ should be computed in the following way as in latent generative models[55,56],

$$\log p(\mathbf{x}) = \log \int p(\mathbf{z}) p_\phi(\mathbf{x}|\mathbf{z})\, \mathrm{d}\mathbf{z} = \log \sum_{\mathbf{z}_i \in \mathcal{Z}(\mathbf{x};\phi)} P(\mathbf{z}_i) \tag{2}$$

where $p_\phi(\mathbf{x}|\mathbf{z}) = \delta\big(g_\phi(\mathbf{z}), \mathbf{x}\big)$ is defined as "*duplicate distribution*", which consists of all ProTokens $\{\mathbf{z}\}$ that can be decoded back to $\mathbf{x}$ through the decoder. Since $\mathbf{z}$ is discrete, integration of $p_\phi(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ over $\mathbf{z}$ is equivalent to the summation of $P(\mathbf{z}_i)$ over a finite and countable *duplicate set* $\mathcal{Z}(\mathbf{x};\phi)$.

Since we barely have information about $\mathcal{Z}(\mathbf{x};\phi)$ *a priori*, the exact likelihood in Eq. 2 is hard to compute in practice. To address this issue, we derive computationally efficient lower bounds Eq. S4 & Eq. S7 for the likelihood and develop a *Duplicator* module (see Section II-C for more details) to explore and expand the duplicate set to reduce the bias and variance for likelihood estimation.

C. Models

Tokenizing metastable states associated with a given protein structure can be cast as a *conditional generative learning problem* as in invertible coarse graining[57]. We design a deep neural network system, *ProToken Distiller*, to achieve this goal (Fig. 2a).

1. Probabilistic conditional decoder

Specifically, we aim at constructing a metastable conformational distribution $p_\phi(\mathbf{x})$ and sampling from it according to a structure $\mathbf{x}$, which can be achieved through a probabilistic conditional *Decoder* g_ϕ (Fig. 2b) using the reparameterization trick as in VAE[58] or GAN[18],

$$\mathbf{x} \sim p_\phi(\mathbf{x}); \qquad p_\phi(\mathbf{x})\mathrm{d}\mathbf{x} = g_\phi\big(\boldsymbol{\epsilon}, \mathbf{z}(\mathbf{x})\big)\mathrm{d}\boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\epsilon}$ is random noise from a known prior like Gaussian distribution, and $\mathbf{z}(\mathbf{x})$ is the embedding of a tokenized $\mathbf{x}$ provided as the conditional information to the *Decoder*. According to the probabilistic tokenization theory (Eq. 1), the ProToken $\mathbf{z}$ specifies the identity of the metastable state, whereas ϵ accounts for the conformational fluctuations within the state.

The Decoder is a composite function consisting of a *Token Duplicator* module and a *Detokenizer* module: The *Token Duplicator* is responsible to expand and sample from the duplicate set in Eq. 2, whereas the *Detokenizer* module is a SE(3)-equivariant generative model which samples protein structures from a metastable ensemble corresponding to a given ProToken string. More details about the *Token Duplicator* module and *Detokenizer* module can be found in SI Section I-A.

The conditional embedding $\mathbf{z}(\mathbf{x}) = h_\theta \circ f_\theta(\mathbf{x})$ is obtained via a composite of *Encoder* f_θ and *Tokenizer* h_θ, which transforms the all-atom protein 3D structure $\mathbf{x}$ into SE3-invariant embeddings of discrete tokens.

2. SE(3)-invariant encoder

The *Encoder* f_θ comprises an SE(3)-invariant *Structure Encoder* module, and a *Deduplicator* module (Fig. 2c). As explained, the *Deduplicator* module is introduced to improve the *robustness* of the yielded embeddings against intrinsic structure fluctuations (see SI Section I-A for more details about the *Duplicator*). Given the separation of timescales of backbone and sidechain motions, the sidechain and backbone structures are encoded separately,

$$f_\theta(\mathbf{x}) = f_\theta(\mathbf{x}_{\mathrm{BB}}, \mathbf{x}_{\mathrm{SC}}) \approx f_{\theta_1}(\mathbf{x}_{\mathrm{BB}}) \otimes f_{\theta_2}(\mathbf{x}_{\mathrm{SC}}) \tag{4}$$

where $\mathbf{x}_{\mathrm{BB}}, \mathbf{x}_{\mathrm{SC}}$ denote the backbone and sidechain structures, respectively; and $\otimes$ denotes Cartesian product (i.e., concatenation of tensors). The resulting $f_\theta(\mathbf{x})$ is a continuous SE(3)-invariant embedding for the protein structure. More details about the backbone and sidechain structure encoders can be found in SI Section I-B.

Considering that metastable structure ensembles can be reasonably represented by discrete tokens, we prepend a Tokenizer h_θ to the Encoder in order to *variationally cluster* (or discretize) $f_\theta(\mathbf{x})$ into quantized ProTokens.

3. Variational tokenizer

The Tokenizer $h_\theta = r_\theta \circ s_\theta$ discretizes the structural embeddings $f_\theta(\mathbf{x})$ into ProTokens $\mathbf{z}(\mathbf{x})$ (Fig. 2d), consisting of a composition of a *Clustering* module s_θ and a *Compressing* module r_θ (see more details about the Compressing module in SI Section I-C). The *Clustering* module aggregates the continuous embedding learned by the Encoder into *K* clusters (i.e., "codes" or "tokens"), and each cluster is assigned with a *d*-dimensional vector as the cluster center (or token embedding). Since the backbone and sidechain embeddings are obtained through two independent tracks, the Clustering module also operates separately for backbone and sidechain embeddings,

$$\mathbf{z}_{\mathrm{BB}} = s_{\theta_1}(\mathbf{v}_{\mathrm{BB}});\ \mathbf{z}_{\mathrm{SC}} = s_{\theta_2}(\mathbf{v}_{\mathrm{SC}}) \tag{5}$$

$\mathbf{z}_{\mathrm{BB}}, \mathbf{z}_{\mathrm{SC}}$ denote the backbone and sidechain tokens, respectively. The ProToken for all-atom structure is then assembled by Cartesian product $\mathbf{z} = \mathbf{z}_{\mathrm{BB}} \otimes \mathbf{z}_{\mathrm{SC}}$. Each ProToken has *dual representations* which are

mutually mappable: one corresponds to the discrete cluster index, the other is the embedding of the cluster center which lies in a continuous vector space.

To make the clustering procedure end-to-end differentiable, we approximate the gradient flow with straight-through estimators[59,60] for backbone tokenization, which will be elaborated in the next section (Section II-D). Details about the sidechain tokenization can be found in SI Section I-C.

Noteworthy, there is a fundamental difference between the Clustering module and common VQ models such as VQ-VAE[61] or VQ-GAN[62]: The clustering is performed *variationally*, that is, the number of alive codes should be as small as possible in order to tighten the variational information bottleneck (see more details in Section II-D), which contrasts sharply to state-of-the-art VQ training where a high usage of codes is usually preferred[63,64].

D. Optimization of ProToken Distiller

Overall, the ProToken Distiller (g_ϕ, and $f_\theta \circ h_\theta$) is optimized towards the following coupled objectives:

i) minimizing the divergence between $p_\phi(\mathbf{x})$ and $p_D(\mathbf{x})$;

ii) maximizing the mutual information between ProTokens $\mathbf{z}$ and $g_\phi(\boldsymbol{\epsilon}; \mathbf{z}(\mathbf{x}))$, while minimizing the mutual information between ProTokens $\mathbf{z}$ and the input structure $\mathbf{x}$;

iii) minimize the divergence between $f_\theta(\mathbf{x})$ and the encoding of the adversarial example $f_\theta(\mathbf{x}')$.

Intuitively, the first objective ensures the *sufficiency* of ProTokens as a transformed representation of metastable protein structures. The second objective guarantees the *necessity* and *non-redundancy* of the transformed representation. The last objective improves the robustness of ProTokens against intrinsic structural fluctuations. Technically, these objectives can be achieved by minimizing an InfoGAN loss[65] regularized by variational information bottleneck[54] and adversarial training[66]. As a result, the final loss function L_{PD} for the ProToken Distiller to be minimized is a linear combination of the sufficiency loss L_{suf}, necessity loss L_{nec}, and robustness loss L_{rob} with a reasonable set of hyperparameters,

$$L_{\mathrm{PD}}(\theta, \phi) = L_{\mathrm{suf}}(\theta, \phi) + L_{\mathrm{nec}}(\theta, \phi) + L_{\mathrm{rob}}(\theta) \tag{6}$$

1. Data preparation

In order to train the probabilistic Decoder, given each structure sample $\mathbf{x}_D$ from the training set, data augmentation is performed by means of metastable perturbation sampling (see SI Section II-A for more details), yielding a structure ensemble $\{\mathbf{x}; \mathbf{x}_D\}$ representing conformers from the same metastable state associated with $\mathbf{x}_D$. Samples from $\{\mathbf{x}; \mathbf{x}_D\}$ are used to compute the generative loss defined in L_{suf}.

Furthermore, we construct adversarial examples $\mathbf{x}' \in \{\mathbf{x}; \mathbf{x}_D\}$ against $\mathbf{x}_D$ by setting a similarity cutoff TM-score$(\mathbf{x}', \mathbf{x}_D) > 0.9$. These examples are provided to the model for the calculation of L_{rob}.

We note that both metastable conformers and adversarial examples can be prepared offline prior to the training, thus, incurring no extra overhead for training. Particularly, after the training proceeds, the adversarial examples can also be constructed online where the Decoder itself can serve as a perturbative sampler of an input $\mathbf{x}_D$. We will show that using these decoded structures as adversarial examples is indeed equivalent to the mutual information loss in L_{nec}.

2. Sufficiency

For the first objective, we adopt a loss function inspired by conditional GAN[67], which guides the *Decoder* to generate structures from the metastable states associate to an input structure $\mathbf{x}_D$,

$$L_{\text{GAN}}(\theta,\phi) = -\mathbb{E}_{\mathbf{x}_D\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|x_D),\boldsymbol{\epsilon}} D\left(g_\phi(\boldsymbol{\epsilon},\mathbf{z})||\{\mathbf{x};\mathbf{x}_D\}\right)\right] \tag{7}$$

$$p_\theta(\mathbf{z}|\mathbf{x}_D) = h_\theta \circ f_\theta(\mathbf{x}_D)$$

where $D(\cdot)$ is the critic measuring the divergence between two distributions; $\mathbf{z}$ is the token embedding (i.e., the identity of the metastable state) for $\mathbf{x}_D$ yielded by the Encoder and Tokenizer, and $p_\theta(\mathbf{z}|\mathbf{x}_D)$ depends on the code search algorithm during variational clustering (we adopt the nearest-code search[61] by default). $\boldsymbol{\epsilon}$ represents random noises which are used to model the intrinsic structural fluctuations within the metastable state.

We repurpose the structure module of AF2 with the dropout trick[68] for $g_\phi(\boldsymbol{\epsilon},\mathbf{z})$, where $\boldsymbol{\epsilon}$ denotes the random dropout mask, allowing the SE(3)-equivariant structure module of AF2 to generate different structures conditioned on the same token $\mathbf{z}$. Noteworthy, other unconditional generative models for protein backbone structures[5] can also be adopted and conditionally fine-tuned.

To optimize this GAN objective, we implement maximum-mean discrepancy (MMD) as the critic $D(\cdot)$[69], a non-parametric integral divergence metric between two distributions, which is known to stabilize and simplify the training of GANs[69]. The similarity kernel required by MMD is defined via frame aligned point error (FAPE)[11], a Fréchet-like distance metric[70] for protein 3D structures.

In order to differentiate through the Clustering module h_θ in Tokenizer, we approximate the gradient flow of backbone tokens via the straight-through (ST) estimator[59,60], and implement a commitment loss to reduce the errors of ST estimator,

$$L_{\text{ST}}(\mathbf{z}_e,\mathbf{z}_q) = (1-\beta_{\text{ST}})\left\|\mathbf{z}_e - \text{stop\_grad}(\mathbf{z}_q)\right\|_2^2 + \beta_{\text{ST}}\left\|\mathbf{z}_q - \text{stop\_grad}(\mathbf{z}_e)\right\|_2^2 \tag{8}$$

where $\mathbf{z}_e,\mathbf{z}_q$ represent vectors before and after quantization respectively. The final sufficiency loss for ProToken Distiller takes the following form,

$$L_{\text{suf}} = L_{\text{GAN}} + \lambda_1 L_{\text{ST}} \tag{9}$$

3. Necessity

We perform *variational clustering* in the Tokenizer according to a VIB objective. Specifically, the training objective of *ProToken Distiller* can be recast in terms of VIB theory[54],

$$L_{\text{VIB}} = \mathbb{E}_{\mathbf{x}}\left[-\log p_\phi(\mathbf{x}|\mathbf{z}) + \text{KL}[p_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]\right] \tag{10}$$

In VIB, $p_\phi(\mathbf{x}|\mathbf{z})$ is known as the prediction or reconstruction model (i.e., the Decoder), whereas $p_\theta(\mathbf{z}|\mathbf{x})$ is the inference model (i.e., the Encoder). By appending a Tokenizer to the Encoder, the prior $p(\mathbf{z})$ and posterior $p_\theta(\mathbf{z}|\mathbf{x})$ become discrete, the VIB in Eq. 10 simplifies to,

$$L_{\text{VIB}} = \mathbb{E}_{\mathbf{x}}\left[-\log p_\phi(\mathbf{x}|\mathbf{z}) + \log K\right] \tag{11}$$

where K is the vocabulary size. Therefore, to minimize L_{VIB} is equivalent to gradually pruning the backbone token vocabulary and minimizing K, which is achieved by re-clustering the embeddings into a smaller number of clusters during training[71].

Furthermore, to prevent the ignorance of conditional information (i.e., the ProTokens) by the generative *Decoder*, we also include a mutual information regularizer similar to InfoGAN[65], except that the auxiliary posterior estimator in InfoGAN is replaced by the conditional Encoder. This adaptation leads to a self-consistency term in the loss function,

$$L_{\mathrm{MI}}(\theta,\phi) = -\mathbb{E}_{\mathbf{x}_D\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x}_D),\tilde{\mathbf{x}}\sim G(\boldsymbol{\epsilon},c)} \log\ p_\theta(\mathbf{z}|\tilde{\mathbf{x}}) - H\big(p_\theta(\mathbf{z}|\tilde{\mathbf{x}})\big)\right] \tag{12}$$

where the entropy term $H\big(p_\theta(\mathbf{z}|\tilde{\mathbf{x}})\big)$ is similar to the entropy regularization introduced in VQ training[72]. The final necessity loss for ProToken Distiller takes the following form,

$$L_{\mathrm{NEC}} = \lambda_2 L_{\mathrm{MI}}(\theta,\phi) + \log K \tag{13}$$

4. Robustness

To help ProToken Distiller be better immune to adversarial attacks (i.e., structural fluctuations within a metastable state), we reuse the perturbative sampling data and present them as adversarial samples $\mathbf{x}' \in \{\mathbf{x}; \mathbf{x}_D\}$ to the Encoder for each $\mathbf{x}_D$, and apply an adversarial training loss for the Encoder,

$$L_{\mathrm{ROB}}(\theta) = -\lambda_3 \mathbb{E}_{\mathbf{x}_D\sim\mathcal{D},\mathbf{x}'\sim\{\mathbf{x};\mathbf{x}_D\}}\left[\mathbb{E}_{\mathbf{z}_D\sim p_\theta(\mathbf{z}|\mathbf{x}_D)} \log p_\theta(\mathbf{z}_D|\mathbf{x}')\right] \tag{14}$$

3. Experiments & Results

We train the ProToken Distiller using a curated protein structure dataset[73]consisting of cleaned single chains collected from the Protein Data Bank[74] with a truncation date of Oct. 13th. 2021 (prior to CASP14). We filter these structures and keep chains without structural gaps, remove structures shorter than 30 residues and those from NMR experiments (due to the possibility of lacking metastability), and hold out over 500 structures for validation and use the others (~552 K) as the training data. During training, each single chain is cropped up to 256 residues.

We initialize the codebook of learnable backbone tokens with a size of $K = 768$. During training, the codebook size is reduced to $K = 513$ through variational clustering including a default wildcard token. The number of sidechain tokens is fixed to 20 without further optimization, thus, resulting in a total number of $513 \times 20 = 10260$ ProTokens. The dimensions of the backbone and sidechain token embeddings after compression is 32 and 8, respectively, leading to a final dimension of 40 for each ProToken embedding. More detailed training settings can be found in SI Section II-A.

Because the sidechain parts in ProTokens are set as default and not deliberately optimized, to keep notations simple and consistent, in the following sections regarding the benchmark of optimized ProTokens, unless specified otherwise ProTokens are used interchangeably with backbone tokens, and protein structures refer to the backbone structures.

A. Variationally optimized ProTokens are informative and generalizable representation of protein structures

1. ProTokens preserve both global and local patterns of metastable protein 3D structures

To demonstrate the efficacy of ProTokens for 3D structure representation, especially for backbone conformations, we evaluate the quality of the reconstructed backbone protein structures decoded from the backbone ProTokens against the ground-truth structures in the test set. Provided that the Decoder is a generative model which may yield different structures given the same ProToken, during benchmark we turn off any randomness so that the model runs deterministically.

The test set contains 87 and 44 single chain targets from CASP14 and CASP15 which are publicly accessible. These test data are used to reasonably assess the quality of backbone ProTokens. According to Fig. 3a, the average TM-score of reconstruction (rTM-score) exceeds 0.96 and the median rTM-score exceeds 0.97 for both test sets. Note that such subtle deviations fall in the range comparable to the intrinsic fluctuations of a metastable protein structure and are considered as adversarial perturbations during the training of ProToken Distiller. In conclusion, this experiment demonstrates that ProTokens can sufficiently preserve the overall patterns of the protein backbone structure. Beises, we also provide examples of protein structures (Fig. 3b) which can be reasonably detokenized from ProTokens, but not from the amino acid sequences, confirming that ProTokens are more informative than Anfinsen's tokens.

Next, we wonder whether the efficacy of ProTokens is sensitive to the length or shape of a protein, so we zoom in over the results of the test set in scatters (Fig. 3c). In general, the reconstruction quality is relatively consistent with respect to the length and shape (in terms of the radius of gyration) of the protein. However,

when the protein exhibits a relatively extreme shape, for instance, very extended or in a nearly-linear form (Fig. 3c), the fidelity (defined as the reconstruction quality) of ProTokens drops significantly regardless of whether the protein is short or long. Considering that extremely extended and linear protein structures mostly contain a limited number of stabilizing interactions (Fig. 3d), such geometry usually lacks sufficient stability in physiological conditions. This phenomenon implies that the efficacy of ProTokens may be susceptible to the stability, but not to the length or shape, of a protein structure.

We further examine whether ProTokens discriminate against certain local structural patterns, particularly, those local elements that are less abundant in the training data. It can be concluded from Fig. 3e that ProTokens can reasonably reconstruct the local environment (in terms of residue local distance difference test, lDDT[75]) of residues taking various second-order structures, showing no preference or discrimination over specific patterns.

Given the reliable reconstruction quality of ProTokens, we released a compressed PDB dataset (PT-PDB), containing ~550k single chains with the backbones being compressed and stored in ProToken indices, leading to a compression ratio of 24. We also released the embeddings for the ProTokens, based on which the compressed backbone structures can be decoded back to Cartesian coordinates of atoms.

2. ProTokens generalize to the dark protein universe, disconnected domain assemblies and multimers

The test set in the above experiment consists exclusively of structures from experiments, which may be subjected to human biases. Because ProTokens are trained solely on experimental structures, it is possible that such biases are inherited. Based on an AI-predicted protein structure dataset (AFDB)[76], previous research reveals that there is a large volume of unexplored dark space of the protein universe[77]. Therefore, we extract high-confidence structures from the AFDB dark cluster dataset and interrogate ProTokens' capability of reconstructing these out-of-distribution (OOD) folds. Fig. 3a shows that these OOD single-chain structures can be represented by ProTokens as well, and the reconstruction quality shows no significant difference from that of the experimental targets statistically. This experiment demonstrates that ProTokens are generalizable for tertiary structural patterns.

Next, we wonder whether ProTokens are able to capture quaternary structural patterns, despite the fact that no such examples were presented during training. Considering that multi-domain proteins are often regarded as transitioning from tertiary to quaternary structures, we first collect several samples containing multi-domain chains from CASP14 and CASP15 test sets. These single chains are manually chopped into annotated domains according to the official definition, yielding a set of disconnected multi-domain assemblies which resemble protein complexes. Noteworthy, the training set of ProTokens merely contains single-chain structures without discontinuity (i.e., without structural gaps), so these multi-domain assemblies are also OOD examples. Nevertheless, we allow ProTokens to treat them by manually setting a residue index gap between domains during inference. We surprisingly find that ProTokens can reconstruct the assembled structures reasonably well (Fig. 4a) when the inter-domain contacts are not sparse.

Based on this encouraging result, we further test ProTokens over real-world protein complexes, including homomers and heteromers. The overall performance is satisfying, both the single-chain and the complex structure can be well reconstructed (Fig. 4b). This result confirms ProTokens' capability of characterizing both the tertiary and quaternary structure patterns, although slight performance degradation is observed when the protein complex involves more chains (Fig. 4b).

In addition to the overall shape, the accuracy of the complex interface is of particular interest when assessing a multimer structure. Motivated by this, we compute the DockQ score[78] for the tested dimers, including antigen-antibody (Ag-Ab) complexes[79]. Figure 4c shows that the interface is relatively accurate[78], though not perfect, with an average and median DockQ exceeding 0.49. Noteworthy, compared to multi-domain assemblies or heterodimers, we find that the reconstructed Ag-Ab complex interface shows larger variance, although each single chain is reconstructed well (Fig. 4d). This may be due to the fact that the interaction patterns between Ag-Ab are often sparse and largely determined by the side chains, hence, are less comparable to the tertiary structural patterns that ProTokens are trained on. Besides, the flexibility of the Ag-Ab interface may also impact the fidelity of ProTokens.

Nevertheless, these experiments on multi-chain structures validate the generalization of ProTokens, which are trained solely on tertiary structures, over quaternary interactions. Our observation also implies that the quaternary structure patterns between protein chains may share deep connections with tertiary patterns within a chain, both arising from fundamental physics interactions between amino acid residues.

3. ProTokens are descriptive of finer functional protein conformations

Researchers are particularly concerned with alternative conformations that may lead to functional switching of a protein. Despite many efforts being paid[68,80–82], decoding alternative protein conformations according to the amino acid sequence remains an unresolved challenge. According to the probabilistic tokenization theory, such difficulty is largely due to the compactness of Anfinsen's tokens (Fig. 1a), leading to large and multimodal intra-state variation (i.e., $p(\boldsymbol{\epsilon}|\mathbf{z})$ in Eq. 1) which is hard to model. Compared to the Anfinsen's tokens, ProTokens are designed to be more informative by compromising its compactness. In principle, different metastable conformational ensembles of one protein, which are degenerate in terms of Anfinsen's tokens, can now be distinguished by ProTokens.

We thus conduct experiments to test whether ProTokens are able to characterize alternative conformations of a protein. We first prepare a test set consisting of proteins with a pair of alternative conformations from 60 PDBFlex[83] clusters with the local RMSD larger than 2.0 Å. Note that these conformers are all resolved by experiments, showing sufficient stability. We then calculate the reconstruction quality for each pair of the alternative conformers and find that ProTokens can not only characterize well the structure of each conformer, but also authentically preserve the relative difference between a pair of conformers (Fig. 4e). This experiment verifies the capability of ProToken as informative representations for finer-scale conformations of a protein.

Many proteins are known to undergo conformational changes or adaptations during binding to ligand(s), we thus wonder whether ProTokens can characterize the binding-altered backbone conformations of a protein. Similar to the previous experiment, we test the reconstruction quality of a set of proteins with different *apo*- and *holo*-form conformations (defined as conformers mutually different with a minimum backbone RMSD larger than 2.0 Å with and without ligand binding). According to the results shown in Fig. 4f, ProTokens are able to faithfully describe both the *apo* and *holo* conformers of a ligand-binding protein and preserve their relative differences. This important feature permits the use of ProTokens for the design of ligand-binding proteins involving conformational changes.

B. Probabilistic tokenization yields robust and interpretable encodings of metastable protein backbone structures

1. Similarity in ProTokens parallels similarity in 3D structures

It is compelling to analyze the reasons behind ProTokens' ability in representing protein 3D structures. A reasonable hypothesis is that the structure similarity may be reflected by ProTokens, and we start our analysis from the ProToken embeddings. Provided that each residue in a protein 3D structure is assigned with a ProToken, we wonder whether the ProToken embedding of the residue characterizes its local structure environment, inspired by the classical word embedding theory[84].

On the one hand, we compute the cosine similarity between each pair of ProToken embeddings (Fig. 5a), which reflects the token similarity learned and determined by the optimized model. On the other hand, we compute the BLOSUM (short for blocks substitution matrix)[85] of ProTokens using residue pairs with similar local environments by convention. These residue pairs are drawn from a dataset specifically for developing structure alignment methods like FoldSeek[77]. A higher BLOSUM coefficient between two tokens indicates a higher chance of them being shared by two residues with similar local structural environments. Intriguingly, we find that the BLOSUM of ProTokens correlates significantly well with the similarity matrix of ProToken embeddings (Fig. 5a). This result reveals an important feature of ProToken: if two residues adopt ProTokens with high similarity, they are likely to be positioned in a similar 3D structural environment.

Since the residue-wise similarity in ProTokens has been proved to characterize local structural similarity, it is likely that the sequence-wise distance measured by ProTokens may also quantify the overall structural difference. We thus collect a set of clustered structures AFDB[76,77], calculate the TM-score as well as the ProToken Similarity Score (PT-score) between each pair of the structures. It can be concluded from Fig. 5b that a higher PT-score between proteins indicates a higher structural similarity measured by TM-score. Specifically, PT-score is not only able to distinguish similar structure pairs from unsimilar ones, but also display a quantitative correlation with TM-score so that it can reasonably sort and rank structures according to the similarity with respect to a reference structure. Therefore, the PT-score can be a promising alternative option of quantifying the structural difference between proteins that is required by structure search and clustering algorithms.

2. Fidelity of ProTokens reflects the (meta)stability of 3D structures

As observed in the previous section, the fidelity of ProTokens seem to correlate with the stability of a protein structure (Fig. 3d), possibly due to the probabilistic tokenization process, that is, the metastable structure patterns are prioritized by ProTokens over the transient and dynamic snapshots of a protein 3D structure. To further investigate this phenomenon, we zoom in our analysis over protein residues, considering that even within a single protein, the stability of different residues may be heterogeneity. For instance, it is common that some substructures are intrinsically more dynamic or less stable than the other regions, which may result in less confident or even unresolved parts in experimentally determined structures.

Therefore, we disassemble the test samples into residues, and categorize them according to the B-factors reported by the crystallization experiments. In general, a larger B factor indicates larger intrinsic dynamics (or less stability) of the residue in its 3D structure[86]. Not surprisingly, we find that the fidelity of a ProToken

corresponding to a certain residue correlates with the B-factor of that residue (Fig. 5c). In other words, if the local structure is less stable, it is also more likely to be smeared by ProTokens. We further corroborate this conclusion based on another test set consisting of AF2-predicted structures. Previous research revealed that the predicted confidence (predicted lDDT) of AF2 can reflect to a certain degree the structural flexibility of the residue[87]. In line with previous findings, we observe that on average, when the plDDT of a residue lowers, the fidelity of ProTokens for that residue gets worse (Fig. 5d).

Additionally, as many loop or intrinsic disordered regions of the protein usually cannot be resolved by crystallization due to lack of metastability, it is likely that ProTokens are not able to represent them either. To test this hypothesis, we manually fix and fill the unresolved regions of the experimentally determined structures using PDB-Fixer[88] and AF2, then compute the reconstruction quality of ProTokens for these regions. Consistent with the previous conclusion, we observe a significant drop in the fidelity of ProTokens for these disordered regions (Fig. 5e). However, we notice that, unlike the manually fixed random regions, the experimentally resolved loop conformation of an antigen-bound antibody can be well reconstructed from ProTokens (Fig. 5e). This observation implies that the antigen-binding loop of an antibody is a metastable conformer of a free-form antibody, in line with the thermodynamics theory of protein binding. Unlike intrinsically disordered loops, the difficulty in experimentally resolving these metastable loop conformers alone may result from the existence of other competing metastable conformations.

C. Generative Pre-training of ProTokens with Foundation Models (DiT) Demonstrates Zero-Shot Generalization for Various Tasks

ProTokens (which refer to the combination of the backbone and sidechain tokens hereafter, and we will distinguish the backbone structure with the all-atom structure explicitly) unify the 1D and 3D modality of protein structures, allowing AI models to perceive an all-atom protein structure base on a concise unimodal input. Moreover, ProTokens can be evoked in either of the Janus representations: the continuous token embedding or the discrete token index. The former is amenable to diffusion-based models whereas the latter is suitable for language-based models. It is thus convenient to train a foundation model based on ProTokens. As in state-of-the-art foundation models, the training is usually initialized by a pretraining stage, which relies on a useful objective that is unsupervised and connected to downstream tasks.

It is verified in previous sections that ProToken is a sufficient representation of all-atom metastable and functional structures, so we suppose that the generative learning of ProTokens can be a reasonable pre-training objective for foundation models. Therefore, we train a diffusion-based foundation model, diffusion transformer (DiT), which has been proved to obey the scaling law[39], based on our pre-training objective, and name this pretrained model as ProToken-DiT (PT-DiT).

Training of PT-DiT is quite straightforward thanks to the merit that the ProToken embeddings lie in a regular Euclidean vector space without constraint, in line with a DiT for images or videos[38,39]. So, a minimum of modifications to the DiT model is required, and we only reduce the size of DiT in order to fit the limited computational budget.

PT-DiT is trained on an augmented dataset which expands the training samples of ProTokens by high-quality AI-predicted structures[73]. Besides, as explained in Section II-B, we emphasize that *duplicate*

augmentation is vital to a stable and converged training of PT-DiT as a latent diffusion model. More detailed training settings of DiT can be found in SI Section II-B.

Noteworthy, the PT-DiT is trained with the single pre-training objective till convergence. No further finetuning is conducted for the following experiments.

1. PT-DiT enables de novo design and controlled evolution of all-atom protein structures

According to Eq. 2, the pre-training objective of PT-DiT is indeed equivalent to the *de novo* generation of all-atom protein structures via a latent diffusion model. Therefore, we first perform experiments to test whether PT-DiT is able to generate reasonable all-atom protein structures. Unlike most existing structure-based approaches for protein design, PT-DiT does not factorize the generation process into a backbone generation step followed by a sidechain generation (or inverse folding) step. In contrast, PT-DiT samples a set of all-atom ProTokens via a single reverse diffusion trajectory, which can be further detokenized into all-atom structures through the Decoder conveniently. This difference marks PT-DiT as a novel approach that accomplishes the *joint design* of protein backbone and side chains.

In Fig. 6a, we show that PT-DiT can yield folded-like all-atom protein structures. These generated structures display diverse overall geometries or shapes, and show different preferences for varied secondary structure elements (Fig. 6a). This observation concludes that PT-DiT is able to efficiently explore the universe of foldable protein structures, and no evidence of mode collapse is observed due to the expressive capacity of DiT.

To test the validity of the generated structures, we design experiments to check two widely concerned properties regarding the designed proteins: i) whether the generated backbone is compatible with the sidechain; ii) whether the generated structure is stable enough to be folded. Previous research has revealed that a sufficiently stable (or hyper-stable) protein structure can be reasonably predicted by single sequence predictors[81] such as ESM-Fold[78] and AlphaFold2. Therefore, we perform single sequence structure prediction based on the generated sidechain tokens (that is, the amino acid sequence), and compare the predicted output with the generated structures. We find that the generated sidechain tokens can be folded into meaningful structures rather than random coils by ESM-Fold, indicating the stability of the generated amino acid sequences. Moreover, we observe that the structure consistency between predicted structures and the generated backbones is encouragingly high (Fig. 6a), concluding that the de novo designed backbone structures of PT-DiT are likely to be stable and foldable by the jointly designed sidechains.

Pretraining of PT-DiT yields a compact prior space for ProToken embeddings. As in other latent generative models, this prior vector space is a higher-level abstraction of the input space (i.e., protein structures), and allows us to manipulate the protein structures in this abstract space (Fig. 6b). Mathematically, the forward probabilistic flow ordinary differential equation (PF-ODE)[21] transforms a ProToken $\mathbf{z}$ representing an all-atom proteins structure into a Gaussian-distributed vector $\boldsymbol{\zeta}'$. Therefore, we can extrapolate from $\boldsymbol{\zeta}$ to a perturbed $\boldsymbol{\zeta}'$, and initialize a backward PF-ODE with $\boldsymbol{\zeta}'$ and obtain a perturbed ProToken $\mathbf{z}'$. This paradigm of latent extrapolation may enable controlled evolution of protein structures possibly towards specific external objectives, which can be useful in various scenarios.

As a proof of concept, we conduct the latent extrapolation experiment given all-atom protein structures, and provide examples shown in Fig. 6b. By randomly perturbing experimentally resolved structures in the

latent space, we can generate new all-atom structures with quite different folds. It should be noted that, the ability of controlled evolution of protein structures arises from the abstraction of ProToken embedding space, which is absent for structure generative models based on atomic coordinates like RFDiffusion.

2. PT-DiT shows emerging capability for inverse folding

Inverse folding is an important task in structure-based protein design when the sidechain and backbone structures are designed in a factorized way, and many efforts have been paid to develop powerful inverse folding models. Noteworthy, all of these models underwent supervised training towards the specific inverse folding objective, regardless of whether they make use of other pretrained models[6,89,90].

Due to the unified modality of 1D and 3D structures in ProTokens, the inverse folding task can be translated as in-filling or generating part of the ProTokens (i.e., sidechain tokens) given the remaining part of ProTokens (i.e., backbone tokens). Intriguingly, in the framework of PT-DiT, this definition coincides with the well-known inpainting task in image generation (Fig. 6c). Therefore, we can directly transact useful tools and techniques developed for image inpainting to the inverse folding task. Particularly, it is known that image inpainting can be achieved in a zero-shot manner based on a well-trained diffusion model[91]. We thus implement a simple and useful zero-shot inpainting technique which allows PT-DiT to conduct inverse folding tasks without any fine-tuning. Specifically, we first encode a given backbone structure into backbone token embeddings, then guide the PT-DiT to generate the remaining embeddings (i.e., the sidechain tokens), and obtain the sidechain tokens corresponding to the amino acid sequence (Fig. 6c).

In Fig. 6d, we showcase that given a backbone structure, even without fine tuning, PT-DiT can generate side chains with relatively high sequence recovery with respect to a known all-atoms structure. However, the recovery is not an ideal measure for inverse folding considering that many different amino acid sequences may fold into nearly the same backbone structures. Since PT-DiT is a generative model, we can sample different sidechain tokens given the same backbone structure, and yield diverse inversely folded sequences. We find that many of the generated sequences, despite being less identical to the reference sequence, are predicted to fold into the designated backbone structure according to a structure predictor (Fig. 6d). From these results we can conclude that PT-DiT shows emerging capability for probabilistic inverse folding task, that is, being able to generate conservative as well as diverse amino acid sequences compatible with a designated (foldable) backbone structure without any fine-tuning.

3. PT-DiT enables flexible protein engineering through contextual design

In addition to de novo design, many researchers are interested in contextual design, for instance, design structures conditioned on part of known structures or sequences. Similar to inverse folding, in the framework of PT-DiT many contextual design tasks are all equivalent to inpainting, hence, may be accomplished by a pretrained PT-DiT without fine tuning. We consider several contextual design scenarios that may find wide applications, and interrogate PT-DiT's capability of dealing with these challenging problems.

In the first scenario, PT-DiT is invoked to design a scaffold that can accommodate a specified binding pocket of a ligand (e.g., a small molecule), given the all-atom coordinates of the binding pocket as the context. This scenario is particularly relevant to designing ligand-binding proteins, which remains difficult to both sequence-based and structure-based protein design methods.

In Fig. 7a, we showcase designed proteins for two small molecule ligands. In each case, the pocket is obtained by cropping the structure of a ligand-binding protein and only keeping the binding pocket in a continuous form (that is, no gap between residue indices), mimicking a pre-designed binding site. ProTokens (both backbone and sidechain tokens) for this binding pocket are first obtained via the Encoder, and the length of the flanking tokens to the pocket is randomly sampled. The remaining task is to infill the ProToken embeddings for the added flanking blanks. After inpainting is done, we first detokenize the infilled ProTokens into 3D structures via the Decoder (Fig. 7a). It can be seen that the shape of the binding pocket is preserved in the decoded structure, with a designed flanking sequence as the scaffold. We also validate the generated all-atom structure through a structure predictor, and confirm that the predicted structure aligns well with the generated structure (Fig. 7a), further demonstrating the potentiality of PT-DiT as a zero-shot designer for ligand binding proteins.

In the second scenario, PT-DiT is implemented to design the framework regions (FWRs) as well as complementary determining regions 1 and 2 (CDR1 and CDR2), given the CDR3 of an antibody heavy chain as the context. This setting is reminiscent of the CDR grafting process, where functional CDR loops (particularly, CDR3) need to be transplanted into a human-like antibody sequence.

In a simplified experiment, we crop the CDR3 of an antibody heavy chain with known structure (PDB ID: 5JXE). Unlike binding pocket design, merely specifying a CDR3 loop as context cannot ensure the generated structures belonging to the family of antibodies which exhibit specific structural constraints. Therefore, we need to precondition the contextual ProTokens towards antibody-like structures. Specifically, we first select a human germline structure as template, then replace its CDR3 region with the to-be-grafted loop by superimposition. The backbone of this artificial grafted structure is encoded into backbone tokens, among which the CDR3 loop is cropped and set as context along with CDR3 sidechain tokens. To ensure the humanness of the grafted antibody, the sidechain tokens of CDR1 and CDR2 can be sampled according to human germlines and also added to the context, but we do not include them in our experiment. The lengths of the flanking FWRs and CDRs can be sampled according to the distribution of the human germlines[92] , while we set them the same as in 5JXE for simplicity. Through inpainting sampling, we can obtain ProTokens that can be decoded to all-atom structures containing the target CDR3 loop as well as the amino acid sequence for the entire chain.

According to our experiments, the preconditioned CDR3 ProTokens can successfully bias the generated structures towards antibody-like folds through inpainting (Fig. 7b). More importantly, according to the structures decoded from the generated ProTokens, the CDR3 conformation is largely preserved and deviates only slightly with the reference. Therefore, the functions of the grafted antibodies are also likely to be preserved. Besides, among the grafted structures, some share significant germline similarity and sequence identity with the human-sourced antibody (Fig. 7b), indicating that PT-DiT can yield human-like antibodies that accommodate a designated CDR3 loop conformation.

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. AlphaFold 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 backbone structure and side-chains which are mutually compatible, that is, the sequence can well 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. 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 a quite different approach, protein can also be designed by directly generating the amino acid tokens. In contrast to the 3D structure, the amino acid sequence, also known as the 1D structure of a protein, is formally discrete and SE(3)-invariant. Therefore, the amino acid sequence becomes a compelling candidate as input to language models. Many methods have been developed for protein understanding and designing tasks by educating the LLMs over amino acid sequences. Unfortunately, the modality difference of protein structure representations (1D v.s. 3D) causes significant divergence in the research paradigms of protein design. Although both paradigms have witnessed a lot of advances, little has been investigated about the connections between the continuous and discrete representations of protein structures, and whether they can be interpreted by a unified theory, hence, eliminating the practical differences in 1D- and 3D-based protein design.

In this study, we proposed probabilistic tokenization theory and provided a unified perspective for the 1D- and 3D-based protein structure representations. The probabilistic tokenization theory underpins the feasibility of transforming continuous metastable protein structures into discrete tokens. Particularly, amino acids are special tokens corresponding to long observation timescale. Furthermore, we developed an unsupervised and variational optimization strategy, which integrates the structure prediction and the inverse

folding tasks, to encode protein structures into an expanded vocabulary in addition to amino acids, called *ProTokens*, and detokenize them back to 3D coordinates. With the implementation of several important regularizers based on information theory, the optimization process ensures the sufficiency, necessity as well as robustness of ProTokens as representations of metastable protein structures.

With theoretical support, we showed that discretizing protein structures with proper perplexity can lead to informative and relatively compact ProTokens. Using limited vocabulary of tokens, ProTokens can reconstruct 3D coordinates with high fidelity and reduce the trans-rotational equivariance as well as the polymer restraints of protein structures. We also find that ProTokens trained on single-chain dataset can generalize to multi-chain complex structures, indicating the transferability of basic physics interaction patterns that govern the tertiary and quaternary protein structures. More intriguingly, although ProTokens are able to describe alternative conformations (particularly function-related conformations) of proteins, including ligand binding sites and antigen-bound antibody loops, we find ProTokens cannot reproduce well the experimentally unresolved or ultra-dynamic regions of a protein such as with high B-factors. These findings demonstrate that by means of probabilistic tokenization, ProTokens are good at describing metastable conformations that exhibit a relatively long lifetime, rather than overfit to transient structure snapshots that lack stability. Therefore, it would be an appealing direction to further investigate how the detokenized fidelity of ProToken correlates with the foldability of a protein backbone structure in the future.

Last but not least, we listed possible pitfalls for the usage of a transformed protein structure representation such as ProTokens for downstream tasks. The robustness of the transformed tokens is vital to understanding tasks like function prediction because subtle fluctuations of protein structures mostly do not change its function. On the other hand, the degeneracy or duplicate tokens could be very harmful to the latent generative model based on ProTokens due to a poor MLE bound. We also introduced guidelines to ameliorate these issues. As a compelling application, generative pretraining of ProTokens based on DiT with remedies to these pitfalls demonstrates emerging capability for various protein structure-related tasks, including de novo and contextual structure design, inverse folding and controlled evolution. Particularly, ProTokens allow all-atom design of protein backbone and side chains simultaneously without factorization, adding a powerful option to the AI-enabled protein design toolkit. It is appealing to devise task-specific instructions to further fine-tune the pretrained model, or align the pretrained model with physics simulators or experimental measurement, which we leave for future research. Besides diffusion models, it is also possible to integrate ProTokens as a domain-specific vocabulary to LLMs, allowing LLMs to demystify the structural semantics of proteins and help explore the protein universe.

Acknowledgments

This work was supported by the 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.

Additional Information

Conflict of Interest

Changping Laboratory is in the process of applying for provisional patent (Beijing Changping Laboratory) covering the discretization of biopolymer (including protein) 3D structures for the purpose of compressing, searching and generating 3D structures, that lists X.L., Z.C., J.Z. and Y.Q.G. as inventors. The other authors declare no conflict of interest.

Code & Data Availability

Training set of ProToken is a cleaned version of the PSP Dataset which can be accessed via {http://ftp.cbi.pku.edu.cn/psp/true_structure_dataset/pdb/}.

Backbone ProToken embeddings for the PDB dataset is available on {http://ftp.cbi.pku.edu.cn/psp/true_structure_dataset/protoken/}.

A precompiled version of the ProToken model with tutorials is available online via {https://colab.research.google.com/drive/15bBbfa7WigruoME089cSfE242K1MvRGz#scrollTo=vxQ7xhQQCBEc}, including the Encoder and Decoder as well as the Token vocabulary. Issues of released ProTokens can be reported in the channel {https://discord.gg/NXeG6NcaHc}.

Codes and checkpoints of more comprehensive model components and downstream foundation models will be released upon publication. Earlier accession for academic and collaborative purposes is possible by contacting the correspondence authors.

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? *J. Struct. Funct. Genomics* **4**, 167–177 (2003).
3. Zhang, C. & Kim, S.-H. Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7**, 28–32 (2003).
4. Wild, D. L. & Saqi, M. A. Structural proteomics: inferring function from protein structure. *Curr. 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. *Curr. Opin. Struct. Biol.* **72**, 226–236 (2022).
16. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* 15820–15831 (Curran Associates Inc., Red Hook, NY, USA, 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. in *Advances in Neural Information Processing Systems* vol. 27 (Curran Associates, Inc., 2014).
19. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. in *Advances in Neural Information Processing Systems* vol. 33 6840–6851 (Curran Associates, Inc., 2020).
20. Song, Y. & Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
21. Song, Y. *et al.* Score-Based Generative Modeling through Stochastic Differential Equations. Preprint at https://doi.org/10.48550/arXiv.2011.13456 (2021).
22. Papamakarios, G., Pavlakou, T. & Murray, I. Masked Autoregressive Flow for Density Estimation. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
23. Chen, X., Mishra, N., Rohaninejad, M. & Abbeel, P. PixelSNAIL: An Improved Autoregressive Generative Model. Preprint at https://doi.org/10.48550/arXiv.1712.09763 (2017).
24. Salimans, T., Karpathy, A., Chen, X. & Kingma, D. P. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. Preprint at https://doi.org/10.48550/arXiv.1701.05517 (2017).
25. Jing, B. *et al.* EigenFold: Generative Protein Structure Prediction with Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2304.02198 (2023).
26. Lu, C. *et al.* DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *Adv. Neural Inf. Process. Syst.* **35**, 5775–5787 (2022).
27. Song, J., Meng, C. & Ermon, S. Denoising Diffusion Implicit Models. Preprint at https://doi.org/10.48550/arXiv.2010.02502 (2022).
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. in *Advances in Neural Information Processing Systems* vol. 34 8780–8794 (Curran Associates, Inc., 2021).
30. Ho, J. & Salimans, T. Classifier-Free Diffusion Guidance. Preprint at https://doi.org/10.48550/arXiv.2207.12598 (2022).
31. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/arXiv.2010.11929 (2021).
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. Fang, Y. *et al.* Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2306.08018 (2024).
35. Lv, L. *et al.* ProLLaMA: A Protein Large Language Model for Multi-Task Protein Language Processing. Preprint at https://doi.org/10.48550/arXiv.2402.16445 (2024).
36. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223–230 (1973).
37. OpenAI *et al.* GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
38. Liu, Y. *et al.* Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. Preprint at https://doi.org/10.48550/arXiv.2402.17177 (2024).
39. Peebles, W. & Xie, S. Scalable Diffusion Models with Transformers. Preprint at https://doi.org/10.48550/arXiv.2212.09748 (2023).
40. Zhang, Z. *et al.* Protein Representation Learning by Geometric Structure Pretraining. Preprint at https://doi.org/10.48550/arXiv.2203.06125 (2023).
41. Olivieri, E. & Vares, M. E. *Large Deviations and Metastability*. (Cambridge University Press, Cambridge, 2005). doi:10.1017/CBO9780511543272.
42. Hänggi, P., Talkner, P. & Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* **62**, 251–341 (1990).
43. Klippenstein, S. J., Pande, V. S. & Truhlar, D. G. Chemical Kinetics and Mechanisms of Complex Systems: A Perspective on Recent Theoretical Advances. *J. Am. Chem. Soc.* **136**, 528–546 (2014).
44. Ghosh, D. K. & Ranjan, A. The metastable states of proteins. *Protein Sci.* **29**, 1559–1568 (2020).
45. Eaton, W. A. *et al.* Fast Kinetics and Mechanisms in Protein Folding1. *Annu. Rev. Biophys.* **29**, 327–359 (2000).
46. Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. (Cambridge University Press, Cambridge, 2004). doi:10.1017/CBO9780511721724.
47. Konovalov, K. A., Unarta, I. C., Cao, S., Goonetilleke, E. C. & Huang, X. Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. *JACS Au* **1**, 1330–1341 (2021).
48. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
49. Zhang, J. *et al.* Deep Representation Learning for Complex Free-Energy Landscapes. *J. Phys. Chem. Lett.* (2019) doi:10.1021/acs.jpclett.9b02012.
50. Taketomi, H., Ueda, Y. & Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7**, 445–459 (1975).
51. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
52. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
53. Monticelli, L. *et al.* The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **4**, 819–834 (2008).
54. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep Variational Information Bottleneck. Preprint at https://doi.org/10.48550/arXiv.1612.00410 (2019).
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis With Latent Diffusion Models. in 10684–10695 (2022).
56. Yu, J. *et al.* Vector-quantized Image Modeling with Improved VQGAN. Preprint at https://doi.org/10.48550/arXiv.2110.04627 (2022).

57. Zhang, J., Lin, X., E, W. & Gao, Y. Q. Machine-Learned Invertible Coarse Graining for Multiscale Molecular Modeling. Preprint at https://doi.org/10.48550/arXiv.2305.01243 (2023).
58. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2022).
59. Bengio, Y., Léonard, N. & Courville, A. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. Preprint at https://doi.org/10.48550/arXiv.1308.3432 (2013).
60. Liu, L., Dong, C., Liu, X., Yu, B. & Gao, J. Bridging Discrete and Backpropagation: Straight-Through and Beyond. *Adv. Neural Inf. Process. Syst.* **36**, 12291–12311 (2023).
61. Oord, A. van den, Vinyals, O. & Kavukcuoglu, K. Neural Discrete Representation Learning. Preprint at https://doi.org/10.48550/arXiv.1711.00937 (2018).
62. Esser, P., Rombach, R. & Ommer, B. Taming Transformers for High-Resolution Image Synthesis. in 12873–12883 (2021).
63. Sun, P. *et al.* Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. Preprint at https://doi.org/10.48550/arXiv.2406.06525 (2024).
64. Tian, K., Jiang, Y., Yuan, Z., Peng, B. & Wang, L. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. Preprint at https://doi.org/10.48550/arXiv.2404.02905 (2024).
65. Chen, X. *et al.* InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. Preprint at https://doi.org/10.48550/arXiv.1606.03657 (2016).
66. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. Preprint at https://doi.org/10.48550/arXiv.1412.6572 (2015).
67. Mirza, M. & Osindero, S. Conditional Generative Adversarial Nets. Preprint at https://doi.org/10.48550/arXiv.1411.1784 (2014).
68. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
69. Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y. & Póczos, B. MMD GAN: Towards Deeper Understanding of Moment Matching Network. Preprint at https://doi.org/10.48550/arXiv.1705.08584 (2017).
70. Jing, B., Berger, B. & Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. Preprint at https://doi.org/10.48550/arXiv.2402.04845 (2024).
71. Łańcucki, A. *et al.* Robust Training of Vector Quantized Bottleneck Models. Preprint at https://doi.org/10.48550/arXiv.2005.08520 (2020).
72. Chang, H., Zhang, H., Jiang, L., Liu, C. & Freeman, W. T. MaskGIT: Masked Generative Image Transformer. Preprint at https://doi.org/10.48550/arXiv.2202.04200 (2022).
73. Liu, S. *et al.* PSP: Million-level Protein Sequence Dataset for Protein Structure Prediction. Preprint at https://doi.org/10.48550/arXiv.2206.12240 (2022).
74. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
75. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
76. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
77. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
78. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE* **11**, e0161879 (2016).
79. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).
80. del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).

81. Zhang, J. *et al.* Unsupervisedly Prompting AlphaFold2 for Accurate Few-Shot Protein Structure Prediction. *J. Chem. Theory Comput.* (2023) doi:10.1021/acs.jctc.3c00528.
82. Kalakoti, Y. & Wallner, B. AFsample2: Predicting multiple conformations and ensembles with AlphaFold2. 2024.05.28.596195 Preprint at https://doi.org/10.1101/2024.05.28.596195 (2024).
83. Hrabe, T. *et al.* PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res.* **44**, D423-428 (2016).
84. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. Preprint at https://doi.org/10.48550/arXiv.1310.4546 (2013).
85. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**, 10915–10919 (1992).
86. Sun, Z., Liu, Q., Qu, G., Feng, Y. & Reetz, M. T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* **119**, 1626–1665 (2019).
87. Ma, P., Li, D.-W. & Brüschweiler, R. Predicting protein flexibility with AlphaFold. *Proteins Struct. Funct. Bioinforma.* **91**, 847–855 (2023).
88. openmm/pdbfixer. OpenMM (2024).
89. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. 2022.04.10.487779 Preprint at https://doi.org/10.1101/2022.04.10.487779 (2022).
90. Ren, M., Yu, C., Bu, D. & Zhang, H. Accurate and robust protein sequence design with CarbonDesign. *Nat. Mach. Intell.* **6**, 536–547 (2024).
91. Lugmayr, A. *et al.* RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. in 11461–11471 (2022).
92. Lefranc, M.-P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2015).

Figures & Legends

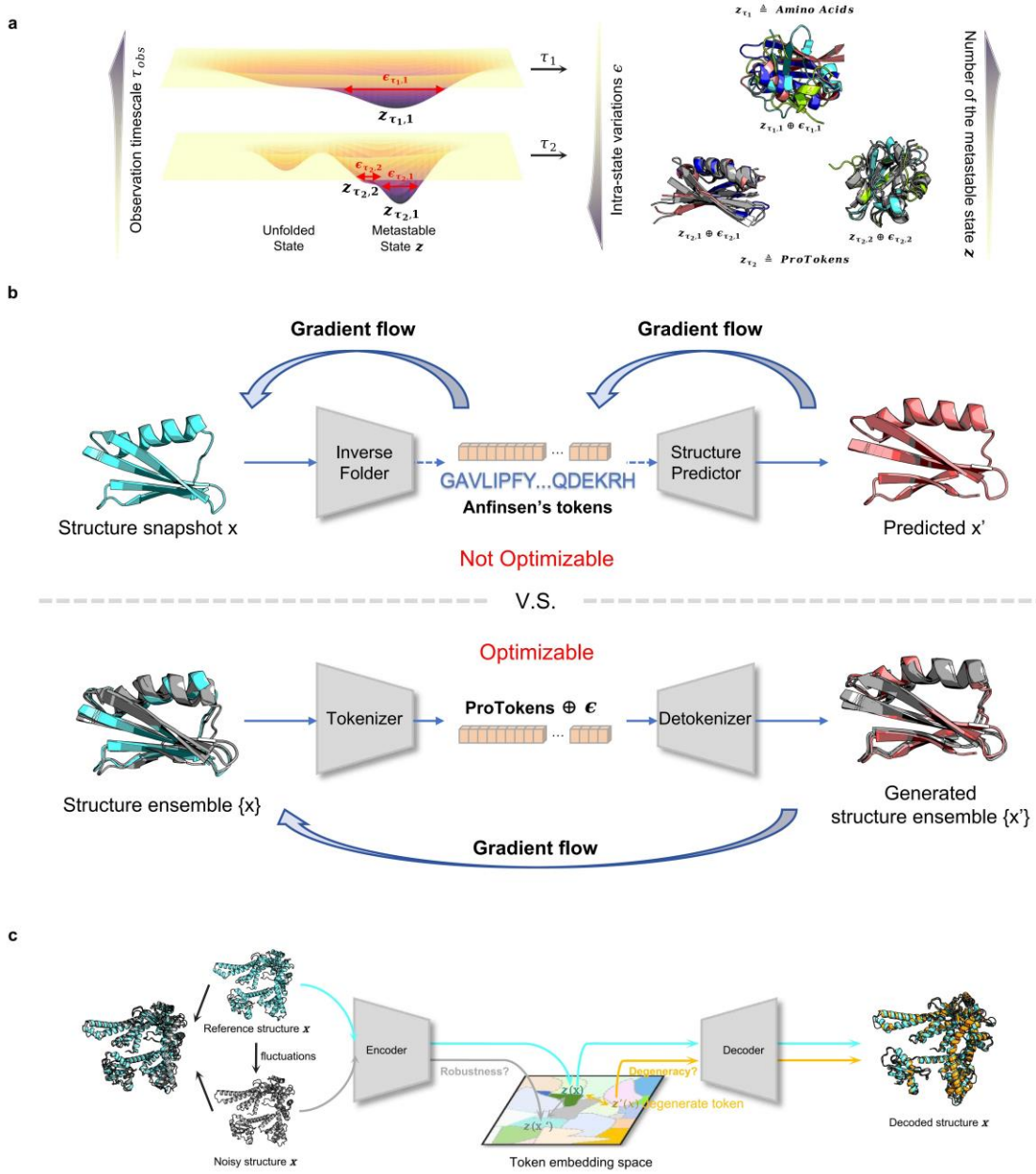

Figure 1. Probabilistic tokenization of protein 3D structures. **(a)** Metastable states defined at different timescales can lead to different tokens of protein structures, including amino acids. **(b)** ProTokens can be learned by connecting inverse folding and structure prediction models end-to-end, and replacing amino acids by optimizable vocabulary. **(c)** Machine-learned tokens may suffer from non-robustness and degeneracy.

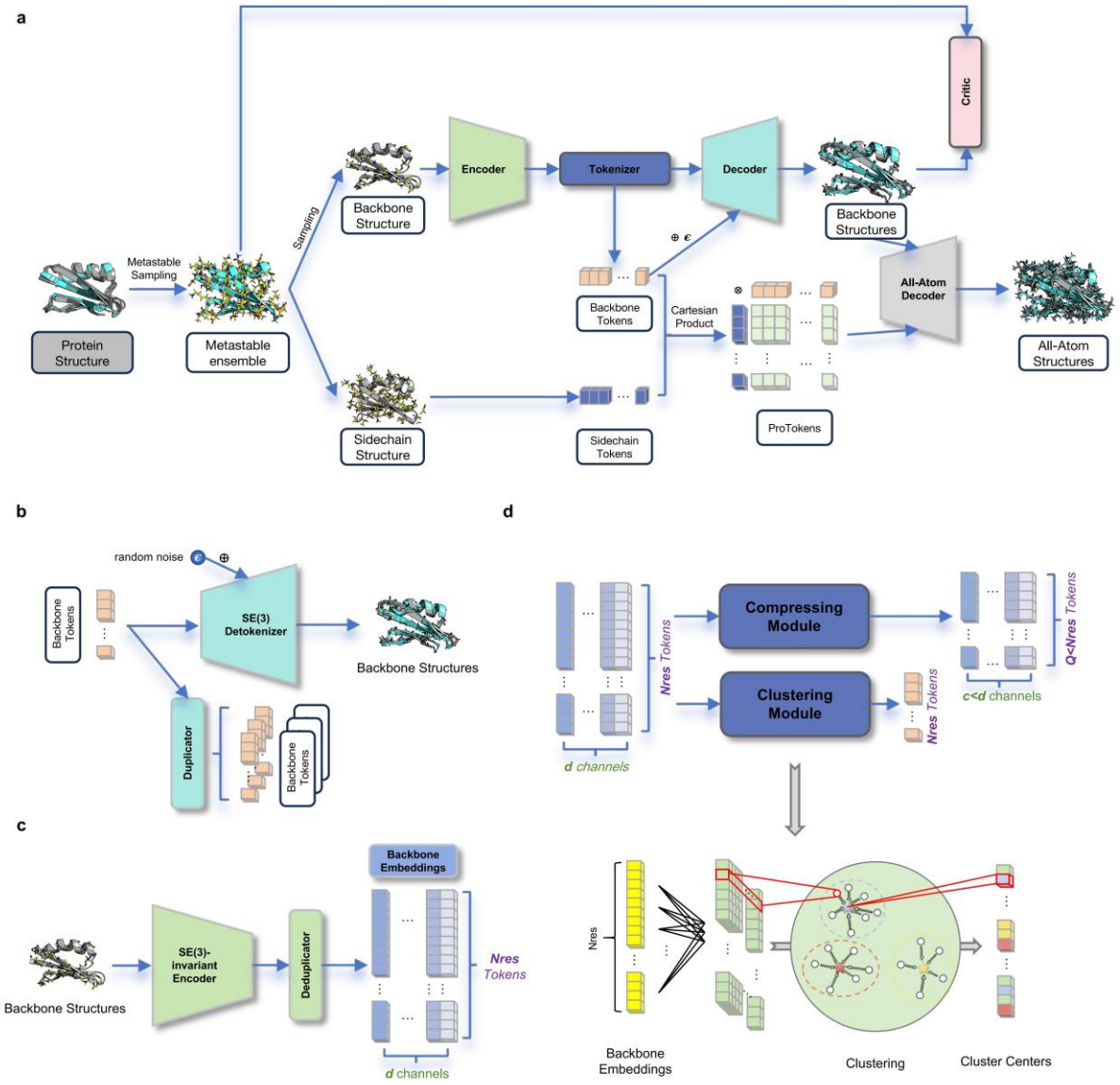

Figure 2. Architecture of ProToken Distiller. **(a)** Overview of the ProToken Distiller. The all-atom structure(s) is tokenized through the backbone track and sidechain track separately, leading to backbone tokens and side chain tokens, respectively. ProToken is the Cartesian product of the two. ProToken Distiller consists of a generative Decoder **(b)**, an Encoder **(c)**, and a Tokenzier **(d)**.

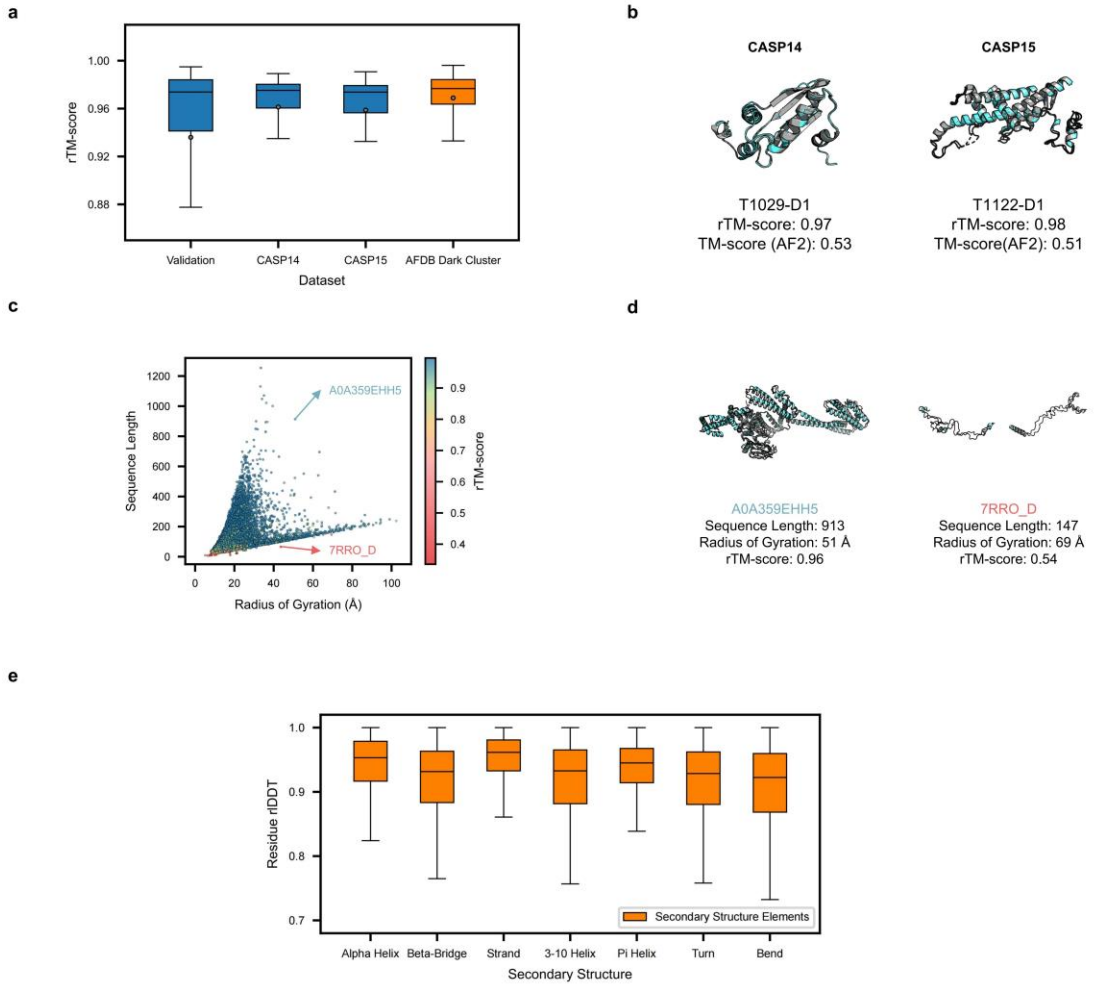

Figure 3. ProTokens are informative representations for tertiary structures. **(a)** Statistics of the TM-score between the decoded structure from ProTokens and the reference structure, rTM-score, over the validation and test sets. **(b)** Examples of ProToken-decoded structures (cyan) superimposed with the experimental reference (gray), which are more consistent than Anfinsen's token-decoded structures measured in TM-score. **(c)** ProTokens are relatively robust to the length and shape of proteins, but susceptible to extremely extended structures. **(d)** Comparison of examples when ProTokens succeed (left) or fail (right) in describing the tertiary structures. **(e)** ProTokens are locally accurate by characterizing various secondary structure elements well.

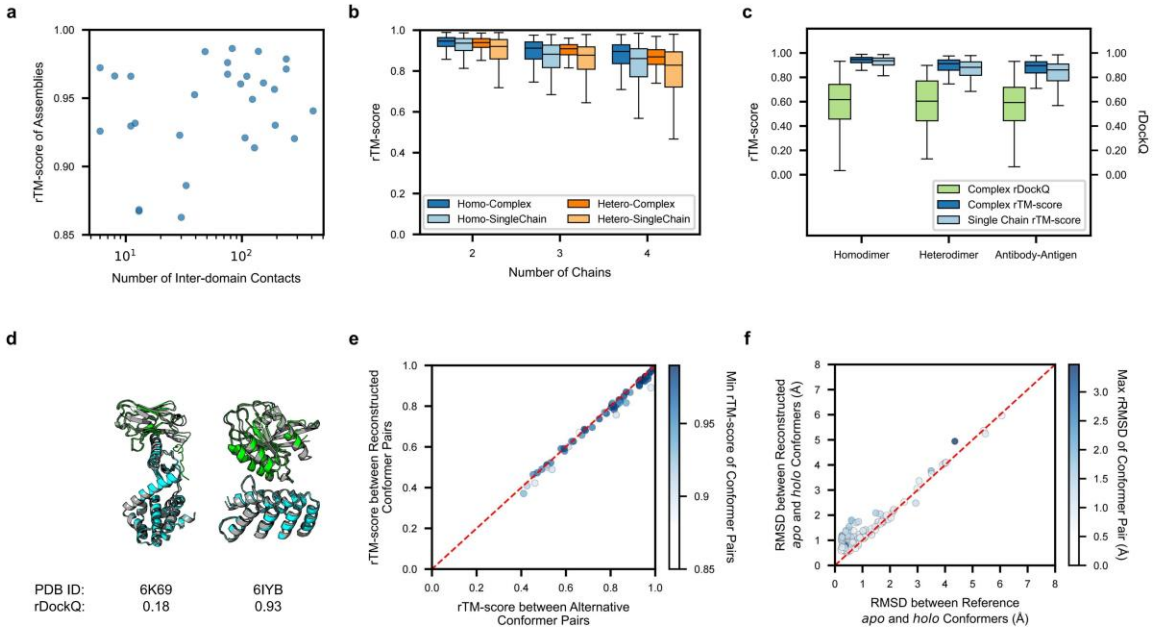

Figure 4. ProTokens generalize well to quaternary structural patterns. **(a)** ProTokens can reconstruct in high quality the disconnected domain assemblies with sufficient inter-domain interactions. **(b)** ProTokens characterize both the tertiary and quaternary structure patterns in protein complexes. **(c)** The interface between homo- and hetero-dimers can be reasonably reproduced by ProToken. **(d)** Compared to heterodimer (PDB ID: 6IYB), degraded fidelity is observed for the antigen-antibody interface (PDB ID: 6K69) with sparser quaternary interactions. **(e)** Assessing the fidelity of ProTokens for representing alternative metastable conformations. **(f)** Assessing ProTokens' capability of distinguishing *apo* and *holo* conformers for ligand-binding proteins.

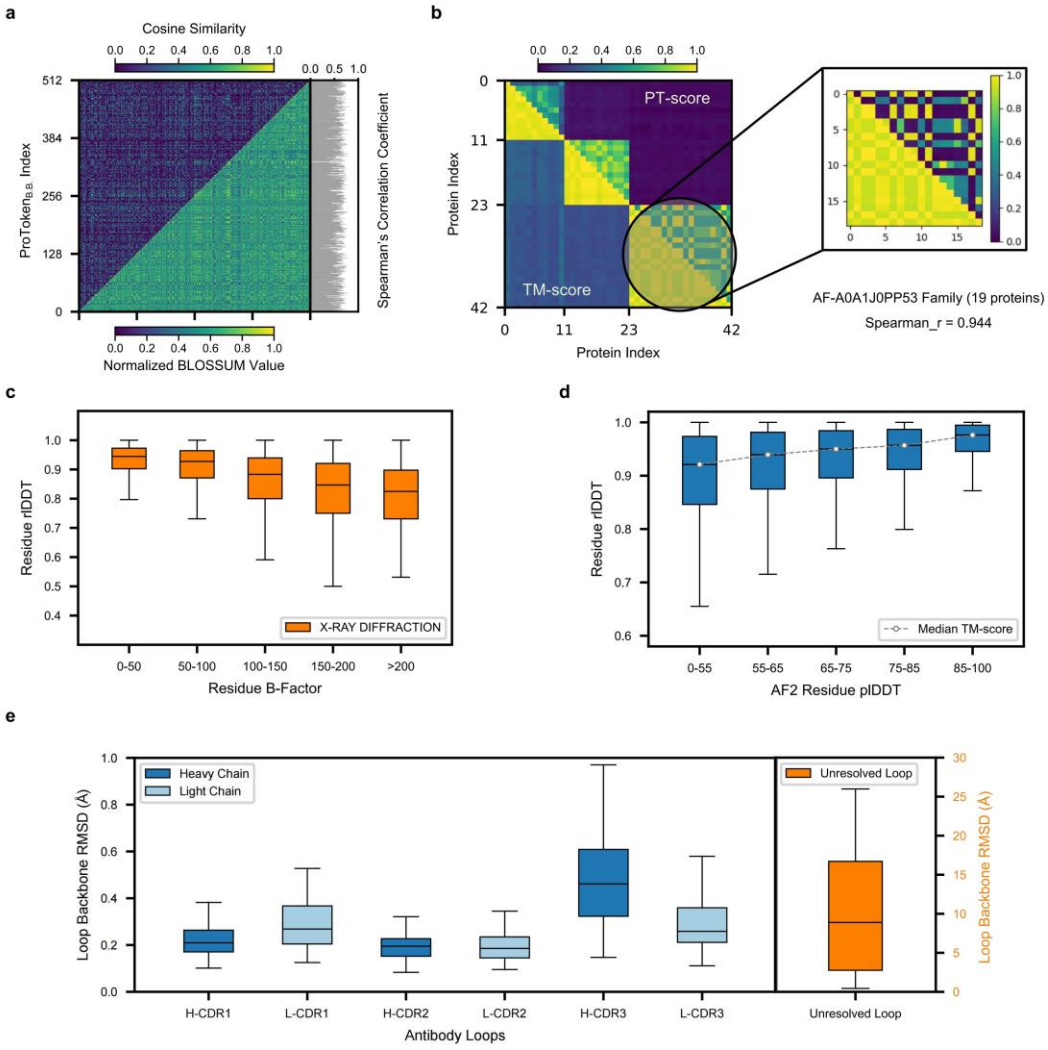

Figure 5. ProTokens reflect similarity and metastability of protein structures. **(a)** The cosine similarity matrix between ProToken embeddings correlates significantly with the BLOSUM. **(b)** PT-score (defined as the difference in ProTokens) correlates quantitatively with TM-score, and can be used for structure comparison and clustering. **(c)** The fidelity of ProTokens correlates negatively with the flexibility of local structures as measured by experimental B factors. **(d)** The fidelity of ProTokens correlates positively with the stability of local structures as indicated by AlphaFold2 plDDT values. **(e)** Antigen-binding conformations of antibody loops are represented well by ProTokens (left panel), indicating metastability, in contrast to experimentally unresolved loops (right panel; note the scale of y-axis is different from the left panel).

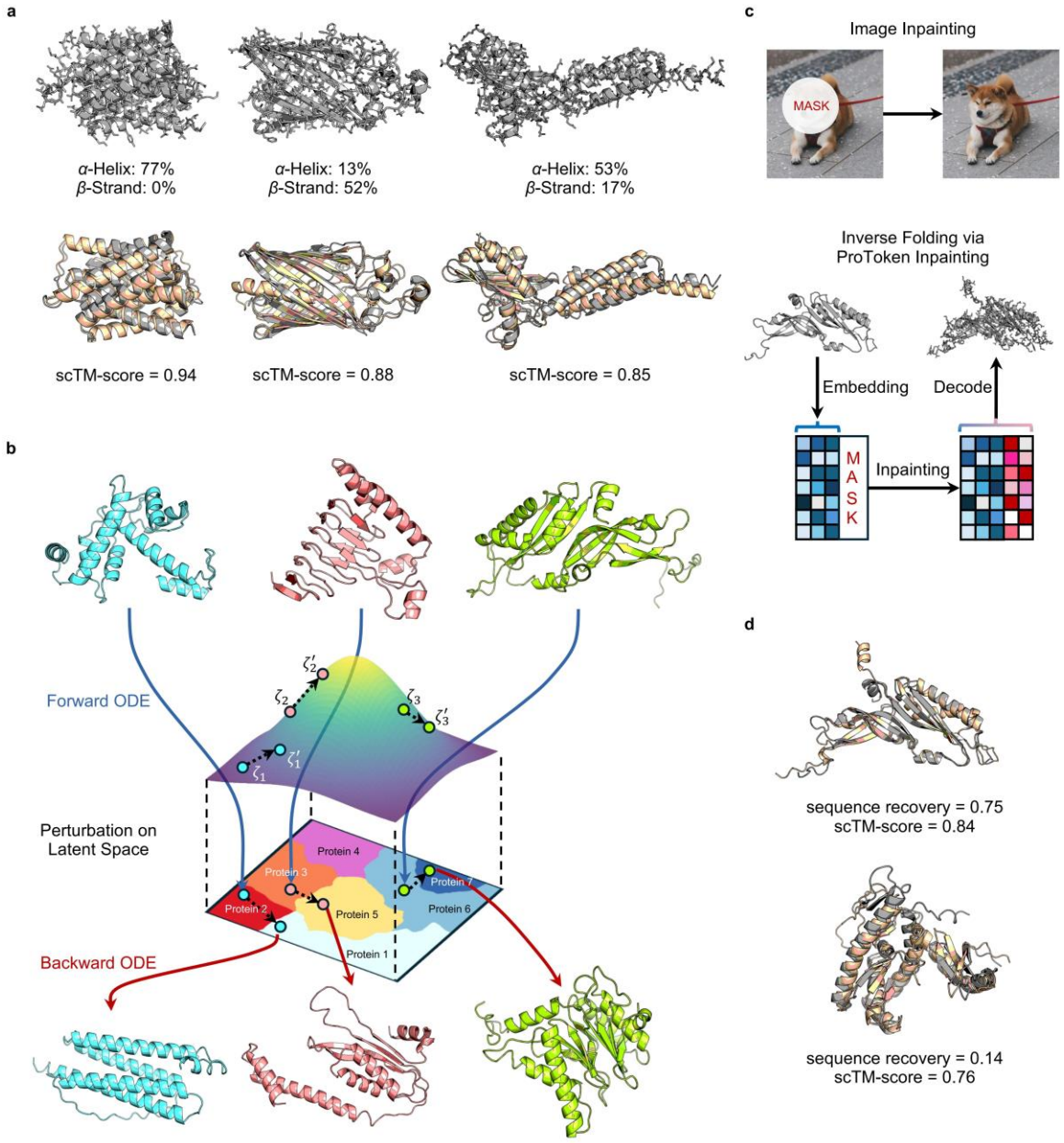

Figure 6. PT-DiT enables de novo design, controlled evolution and inverse folding of protein structures. **(a)** Upper panel: Three cases of de novo designed all atom structures (shown in gray) with different compositions of the second order structure elements. Lower panel: The superimposition of the predicted structure (yellow) with respect to the designed structure (gray). The self-consistency TM-score is annotated. **(b)** Illustration of controlled evolution over the PT-DiT latent space. Three all-atom proteins (colored in cyan, red and green, respectively) are evolved into new folds by perturbative extrapolation of their embeddings over the PT-DiT latent space. **(c)** Parallel in image inpainting (upper panel) and ProToken embedding, which can be used for zero-shot inverse folding (lower panel). **(d)** Examples of zero-shot inpainting via PT-DiT, both high-recovery and high-diversity sequences can be obtained.

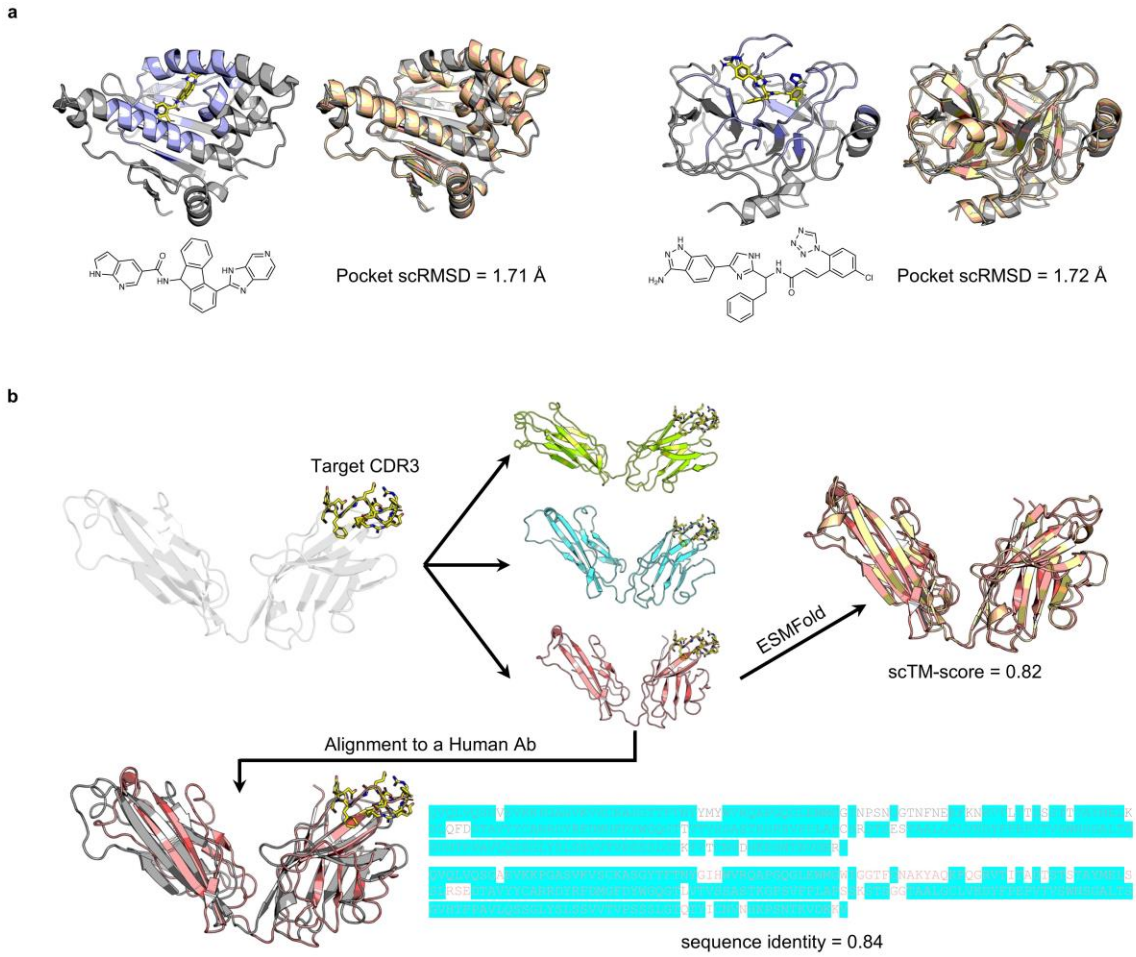

Figure 7. PT-DiT enables zero-shot contextual design of all atom protein structures. **(a)** Design of scaffold (gray) for ligand (yellow) binding proteins based on pocket structures (blue). The predicted structure (yellow) given the designed sequence is superimposed with the designed folds (gray), showing preservation of the desired pocket structure. Ligands are shown at bottom. **(b)** Grafting of a target CDR3 (colored region with side chains shown explicitly) into antibody-like structures (colored in green, cyan and red, respectively), among which an antibody-like structure (red) with human-like sequence is shown (sequence alignment is shown with identical amino acid colored in cyan).

Supporting Information

I. Model Details

A. Decoder

The Decoder is a composite function consisting of a *Token Duplicator* module and a *Detokenizer* module.

1. Detokenizer

The Detokenizer module is a SE(3)-equivariant generative model which samples protein structures from a metastable ensemble corresponding to a given ProToken string. More specifically, the all-atom structures are generated in a factorized way by a backbone detokenizer g_{ϕ_1} and a sidechain detokenizer g_{ϕ_2}, respectively,

$$\mathbf{x}\sim p_{\phi}(\mathbf{x}_{\mathrm{BB}},\mathbf{x}_{\mathrm{SC}}) = p_{\phi_1}(\mathbf{x}_{\mathrm{BB}})p_{\phi_2}(\mathbf{x}_{\mathrm{SC}}|\mathbf{x}_{\mathrm{BB}}) \tag{S1}$$

$$p_{\phi_1}(\mathbf{x}_{\mathrm{BB}})\mathrm{d}\mathbf{x}_{\mathrm{BB}} = g_{\phi_1}(\boldsymbol{\epsilon};\mathbf{z}_{\mathrm{BB}}))\mathrm{d}\boldsymbol{\epsilon} \tag{S2}$$

$$p_{\phi_2}(\mathbf{x}_{\mathrm{SC}}|\mathbf{x}_{\mathrm{BB}})\mathrm{d}\mathbf{x}_{\mathrm{SC}} = g_{\phi_2}(\boldsymbol{\epsilon};\mathbf{z}_{\mathrm{SC}},\mathbf{x}_{\mathrm{BB}})\mathrm{d}\boldsymbol{\epsilon} \tag{S3}$$

where the backbone structure is first generated according to the $p_{\phi}(\mathbf{x}_{\mathrm{BB}})$ conditioned on the backbone tokens $\mathbf{z}_{\mathrm{BB}}$, followed by the generation of sidechains from $p(\mathbf{x}_{\mathrm{SC}}|\mathbf{x}_{\mathrm{BB}})$ conditioned on the backbone structure $\mathbf{x}_{\mathrm{BB}}$ as well as the sidechain tokens $\mathbf{z}_{\mathrm{SC}}$.

Noteworthy, any unconditional backbone generative model (such as RFDiffusion[1]) can be adopted as an initializer for the backbone detokenizer further optimized through conditional fine-tuning. Similarly, any generative sidechain packer can be directly plugged-in (or fine-tuned) as the sidechain detokenizer. We implemented DLPacker[2] in this work and did not fine-tune it.

2. Token Duplicator

The duplicate set defined in Eq. 2 depends both on the volume of the ProToken space and the over-capacity of Decoder. The duplicate set associated with a less compact ProToken space, or a less invertible Decoder, intuitively tends to be larger, which is harmful to protein structure generation tasks as will be shown later.

Although the exact likelihood (Eq. 2) is hard to compute in practice, fortunately, we can develop proper lower bounds that can be conveniently used for maximum likelihood estimation of $\mathbf{x}$ through the transformed representation $\mathbf{z}$.

The first lower bound ($\mathcal{L}_1$) can be derived via the *truncation trick*. Given a (truncated) subset $\widetilde{\mathbb{Z}}(\mathbf{x}) \subseteq \mathbb{Z}(\mathbf{x};\phi)$,

$$\log p(\mathbf{x}) = \log \sum_{\mathbf{z}\in\mathbb{Z}(\mathbf{x};\phi)} p(\mathbf{z}_i) \geq \log \sum_{\mathbf{z}\in\widetilde{\mathbb{Z}}(\mathbf{x})} p(\mathbf{z}_i) \coloneqq \mathcal{L}_1\left(\widetilde{\mathbb{Z}}(\mathbf{x})\right) \tag{S4}$$

$\mathcal{L}_1$ exhibits an important property that,

$$\text{If } \widetilde{\mathbb{Z}}_2(\mathbf{x}) \supset \widetilde{\mathbb{Z}}_1(\mathbf{x}), \text{then } \mathcal{L}_1\left(\widetilde{\mathbb{Z}}_2(\mathbf{x})\right) \geq \mathcal{L}_1\left(\widetilde{\mathbb{Z}}_1(\mathbf{x})\right) \tag{S5}$$

The equality holds if elements in the difference set, $\widetilde{\mathbb{Z}}_2(\mathbf{x}) - \widetilde{\mathbb{Z}}_1(\mathbf{x})$, have zero probability density in total. $\mathcal{L}_1$ is still inconvenient to compute in practice because the summation of probability is prior to logrithm. Thus, we further develop another lower bound ($\mathcal{L}_2$) based on $\mathcal{L}_1$ according to Jensen's inequality,

$$\mathcal{L}_1\left(\widetilde{\mathbb{Z}}(\mathbf{x})\right) = \log\frac{\sum_{\mathbf{z}\in\widetilde{\mathbb{Z}}(\mathbf{x})} p(\mathbf{z}_i)}{\left|\widetilde{\mathbb{Z}}(\mathbf{x})\right|} + \log\left|\widetilde{\mathbb{Z}}(\mathbf{x})\right| \geq \frac{\sum_{\mathbf{z}\in\widetilde{\mathbb{Z}}(\mathbf{x})} \log p(\mathbf{z}_i)}{\left|\widetilde{\mathbb{Z}}(\mathbf{x})\right|} + \log\left|\widetilde{\mathbb{Z}}(\mathbf{x})\right| \tag{S6}$$

$$\log p(\mathbf{x}) \geq \mathcal{L}_1\left(\widetilde{\mathbb{Z}}(\mathbf{x})\right) \geq \mathbb{E}_{z\sim\tilde{p}(\mathbf{z}|\mathbf{x},\phi)} \log p(\mathbf{z}) + \log\left|\widetilde{\mathbb{Z}}(\mathbf{x})\right| \coloneqq \mathcal{L}_2\left(\widetilde{\mathbb{Z}}(\mathbf{x})\right) \tag{S7}$$

Different from $\mathcal{L}_1$, Eq. S7 allows us to perform minibatch optimization by unbiasedly sampling from a $\widetilde{\mathbb{Z}}(\mathbf{x})$ during training. Similarly, it also follows straightforwardly from Eq. S7 that the larger $\left|\widetilde{\mathbb{Z}}(\mathbf{x})\right|$ is, the tighter the lower bound $\mathcal{L}_2$ is, indicating the importance of expanding the subset $\widetilde{\mathbb{Z}}(\mathbf{x})$.

The remaining issue is how to construct and sample from the duplicate (sub-)set. A naïve approach is to approximate the duplicate set by a singleton set, i.e., $\widetilde{\mathbb{Z}}(\mathbf{x}) = \{f_\theta(\mathbf{x})\}$, which leads to a suboptimal truncation for $\mathcal{L}_1$. Or one can determistically set $\mathbf{z} = f_\theta(\mathbf{x})$ and estimate the expectation term in $\mathcal{L}_2$, which leads to a highly biased and high-variance one-sample Monte Carlo estimator.

To circumvent these drawbacks, a *Token Duplicator* module, $q_\phi(\mathbf{z}|\mathbf{x}, g_\phi)$, is prepended to the Detokenizer, which is responsible to expand and sample from the duplicate subset $\widetilde{\mathbb{Z}}(\mathbf{x})$. In general, any conditional generative model which yields diverse ProTokens which can be decoded back to $\mathbf{x}$ via the Decoder can be used as $q_\phi(\mathbf{z}|\mathbf{x}, g_\phi)$.

In this work, given a structure $\mathbf{x}$, we initialize the duplicate set with $\{\mathbf{z} = f_\theta(\mathbf{x})\}$, and implement a sampling-based, Monte-Carlo-style *Token Duplicator* to expanding the duplicate set. We randomly mutate the residue-wise backbone tokens in $\mathbf{z}$ according to the similarity matrix of token embeddings, yielding a perturbed $\mathbf{z}'$. We accept the proposed mutation $\mathbf{z}'$ with a probability proportional to the reconstruction quality of $g_\phi(\mathbf{z}')$, based on which the next residue-wise mutation is performed. If $g_\phi(\mathbf{z}')$ is sufficiently close to $\mathbf{x}$ (defined as TM-score($\mathbf{x}$, $g_\phi(\mathbf{z}')$) > 0.9), we add it to the duplicate set.

During generative training of protein structures $\mathbf{x}$ via ProTokens $\mathbf{z}$, we first resort to the *Token Duplicator* and construct a sufficiently large duplicate subset $\widetilde{\mathbb{Z}}(\mathbf{x})$. The $\mathcal{L}_1$ or $\mathcal{L}_2$ is then optimized based on $\widetilde{\mathbb{Z}}(\mathbf{x})$ in order to perform maximum likelihood estimation of $\mathbf{x}$. This approach is similar to data (or label) augmentation which plays key role in many state-of-the-art generative AI models.

B. Encoder

1. Structure Encoder

Considering the separation of timescales, we encode the sidechain and backbone structures separately as in Eq. 4. As for the *backbone structure encoder* $f_{\theta_1}(\mathbf{x}_{\mathrm{BB}})$, an SE(3)-invariant Transformer based on invariant point attention[3], is designed to transform the Cartesian coordinates of the N, CA, C, O atoms of each residue to a SE(3)-invariant vector. The model is composed of an SE(3)-invariant HyperFormer[4] and a structure-

aware Transformer. 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. Based on the refined vertex (or single) and edge (or pair) representations output by HyperFormer, invariant point attention 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.

On the other hand, since the equilibration of sidechain conformations is much faster than the backbone, all relaxed conformations of a sidechain (e.g., according to Boltzmann distribution) in the context of a given backbone can be considered as a single metastable state and reasonably embedded by a single vector which can distinguish the chemical identity of the sidechain fragment. Therefore, we transact the amino acid embedding learned by AF2 as the s*idechain structure encoder* $f_{\theta_2}(\mathbf{x}_{\mathrm{SC}})$ without further optimization.

The embedding of an all-atom structure is obtained via the Cartesian product operation, that is, by concatenating the backbone embedding and sidechain embedding.

2. Deduplicator

Note that our choice of sidechain encoding is naturally invariant to perturbations of sidechain conformations. However, unlike the *sidechain structure encoder*, one particular concern of a learned *backbone structure encoder* is that the *robustness* of the yielded codes against subtle structure perturbation is not guaranteed. To alleviate this issue, we append a *Deduplicator* u_θ to the *backbone structure encoder* (Fig. 2c), which is trained to surjectively map structures which belong to the same metastable states (or merely differ mutually up to negligible perturbations) to almost the same embedding,

$$u_{\theta_1} \circ f_{\theta_1}(\mathbf{x}_{\mathrm{BB}}) \approx u_{\theta_1} \circ f_{\theta_1}(\mathbf{x}_{\mathrm{BB}} + \Delta\mathbf{x}_{\mathrm{BB}}) \tag{S8}$$

where $\Delta\mathbf{x}_{\mathrm{BB}}$ denotes the structural fluctuation within a metastable state. In terms of physics, the *Deduplicator* behaves like a "relaxation simulator" which relaxes fluctuated structures within a metastable ensemble towards a single stable representative structure, and consequently, degenerates the embeddings for fluctuated structures.

C. Tokenizer

The Tokenizer is a composition of a *Clustering* module s_θ (which has been elaborated at length in the main text), and a *Compressing* module r_θ.

1. Compressing Module

The ProToken of all-atom structure is obtained via $\mathbf{z} = \mathbf{z}_{\mathrm{BB}} \otimes \mathbf{z}_{\mathrm{SC}}$, that is, the Cartesian product of the backbone token set and the sidechain token set. The continous ProToken embedding is equivalent to concatenating the the backbone embedding and sidechain embedding together, which has the shape of $(N_{\mathrm{res}}, d = d_{\mathrm{BB}} + d_{\mathrm{SC}})$ for a N_{res}-long protein.

Note that without processing, d_{BB} and d_{SC} are usually large such that $d \gg 3\bar{N}_{\mathrm{atom}}$ where $\bar{N}_{\mathrm{atom}}$ stands for the average number of atoms per residue. For instance, in our model, $d_{\mathrm{BB}} = 32$ during training and $d_{\mathrm{SC}} = 256$ in consistency with AF2. Despite of being SE(3)-invariant, the dimensionality of such a representation is still much too higher than the intrinsic degrees of freedoms of a protein (which is upper bounded by the number of Cartesian coordinates of its structure).

Therefore, we include a *Compressing* module $r_\theta(\mathbf{v})$: $\mathbb{R}^{N_{\mathrm{res}}*d} \rightarrow \mathbb{R}^{Q*c}$ to concentrating the information of ProToken by lowering its dimensionality. The compression can be performed lengthwise and (or) depth-wise. As shown in Fig. 2d, for the lengthwise compression, we transformed a ProToken string of shape (N_{res}, d) into (Q, d), with Q being a predefined number independent of and usaually smaller than the average N_{res}. In this research, we performed the depth-wise compression (Fig. 2d), which reduces the shape of (N_{res}, d) into $(N_{\mathrm{res}}, c \ll d)$, and we achieved this goal by means of dimensionality reduction methods[5].

II. Optimization Details

A. ProToken Distiller

1. Training settings

ProToken Distiller is trained with a batch size of 288. The learning rate is set to 5e-4 with a cosine decay down to 2e-5 after 80,000 steps. The training is executed on 48 NVIDIA A100 GPUs.

The training is split into two stages. For the first 100,000 steps, we implemented the robustness loss L_{ROB} in Eq. 14, with a prepared set of adversarial examples but turned off the mutual information loss L_{MI} in Eq. 12. For the remaining 100,000 steps, we switched off the robustness loss, instead, MI loss was turned on.

2. Metastable Perturbation Sampling (MPS)

To sample more metastable conformations of a 3D structure $\mathbf{x}_0$ associated with the same function, we performed metastable perturbation sampling according to the given 3D structure and yield $\{\mathbf{x}\}$. Specifically, we recommend several options that can serve for MPS: 1) resorting to a MD simulation engine and running temporal proximal sampling[6]; 2) resorting to AI-based sampler which can yield perturbed conformations like AF-Cluster[7]; 3) self-distillation of a pre-trained probabilistic ProToken distiller.

In this research, we implemented the first two strategies to obtain perturbed metastable conformations corresponding to a given reference 3D structure.

B. ProToken Diffusion Transformer (PT-DiT)

To prepare the training data of PT-DiT, we first performed Encoder inference for the training set of protein structures, yielding a basic set of ProTokens containing ~550k ProToken samples.

In order to counteract the biased estimation of likelihood of the latent diffusion model, we augmented the basic ProToken set with the duplicate set obtained by the *Token Duplicator*. Through experiments, we found empirically that augmentation of latent duplicates is vital for the success of training PT-DiT.

The generation quality of PT-DiT can be susceptible to errors that cause “token switch”, so we introduced anisotropic diffusion kernel[8] for the variance-preserving diffusion process, in order to better align with the objective of the latent diffusion model

We trained PT-DiT with a batch size of 256, learning rate of 2e-4, on 8 NVIDIA A100 GPUs.