bio2token: All-atom tokenization of any biomolecular structure with Mamba

Andrew Liu, Axel Elaldi, Nathan Russell & Olivia Viessmann

Flagship Pioneering

Cambridge, MA 02142, United States

{anliu, aelaldi, nrussell, oviessmann}@flagshippioneering.com

Abstract

Efficient encoding and representation of large 3D molecular structures with high fidelity is critical for biomolecular design applications. Despite this, many representation learning approaches restrict themselves to modeling smaller systems or use coarse-grained approximations of the systems, for example modeling proteins at the resolution of amino acid residues rather than at the level of individual atoms. To address this, we develop quantized auto-encoders that learn atom-level tokenizations of complete proteins, RNA and small molecule structures with reconstruction accuracies well below 1 Angstrom. We demonstrate that a simple Mamba state space model architecture is efficient compared to an SE(3)-invariant IPA architecture, reaches competitive accuracies and can scale to systems with almost 100,000 atoms. The learned structure tokens of bio2token may serve as the input for all-atom generative models in the future. Our implementation is available at https://github.com/flagshippioneering/bio2token.

1 Introduction

Background. Biomolecular structures can be represented as 3D point clouds, where each point corresponds to a chemical entity such as an atom, functional group, or molecular subunit. Generative modeling of these structures, especially for large biomolecules, often employs coarse-grained representations to manage complexity. Methods like denoising diffusion probabilistic models (DDPMs) and language models generate structures at varying levels of detail, from atoms to residues. While DDPMs have been applied to atomistic conformers (Hoogeboom et al., 2022) and protein-ligand design (Schneuing et al., 2022), scaling them to large proteins remains computationally challenging. Models like RFDiffusion-All-atom address this by diffusing only the protein backbone and reconstructing side chains separately (Krishna et al., 2024; Dauparas et al., 2022). Similarly, language models like ESM-3 (Hayes et al., 2024) rely on residue-level representations but still struggle with large proteins and complexes.

Achieving atomic-resolution modeling for large molecules requires reasoning over long-range interactions in sequence space, a challenge for traditional architectures like transformers and graph models. To address this, we leverage Mamba (Gu & Dao, 2023), a structured state-space model designed for long-context modeling, replacing transformer modules in structure tokenizers to enable efficient all-atom representations. Mamba has demonstrated scalability to tasks involving thousands to millions of tokens on standard GPU hardware.

3D structure tokenization for generative modeling. Turning 3D structures into discrete 1D sequences for generative language modeling, discrete diffusion or other downstream task has become a popular approach to biomolecular modeling. FoldSeek introduced the "3Di" structural interaction alphabet to convert three-dimensional protein backbone structures into one-dimensional sequences, facilitating faster structural alignment (van Kempen et al., 2022). Neural network-based quantized auto-encoders (QAEs) (Van Den Oord et al., 2017) have since been employed to learn 3D structure

tokenizers. ESM-3 utilizes a transformer-based QAE that encodes residue-level backbones and decodes to all-atom structures, with training limited to proteins with fewer than 512 residues and using a 600M parameters transformer model. FoldToken (Gao et al., 2024) and InstaDeep (Gaujac et al., 2024) also use QAEs with transformer and graph neural network architectures, respectively, focusing on residue-level tokenization but limited to backbone reconstruction. Alphafold-3 (AF-3) (Abramson et al., 2024) generates all-atom structures using a token-guided diffusion network. For small molecules, approaches include one-hot encoding of coordinate digit strings (Flam-Shepherd & Aspuru-Guzik, 2023; Zholus et al., 2024) and SE(3)-invariant QAEs like Geo2Seq (Li et al., 2024) and MolStructTok (Anonymous, 2024). Prior work predominantly relies on QAEs with various architectures and features, incorporating symmetries through structural features or invariant point attention. In contrast, our method uses neither engineered SE(3)-invariant features nor does it employ invariant network architectures.

In this work we present a simple, lightweight, and compute efficient Mamba-based structure tokenizer that converts 3D point clouds into 1D discrete tokens. We train small molecule-only, protein-only, and RNA-only vocabularies *mol2token, protein2token* and *rna2token*. We also train a unified tokenizer *bio2token* that encodes any of those biomolecules, ranging from tens to tens of thousands of atoms, that would be challenging for transformer-based methods to scale too.

1.1 Background: Transformers, State space models, and Mamba

Transformer. Transformers (Vaswani, 2017) use the *attention* mechanism to capture long-range dependencies in sequences. The attention mechanism has the update rule:

$$y = M(x)x, \tag{1}$$

where x is the input sequence, y is the latent representation, and $M(x) = \text{softmax}\left(Q(x)K(x)^T\right)$ is the attention matrix. This matrix multiplication formulation makes attention ideal for GPU processing. However, since $M(x)$ is generally dense and full-rank, transformers suffer from $O(N^2)$ compute and memory costs with respect to sequence length N.

Mamba. Recent alternatives such as deep structured state space models (SSM) (Gu et al., 2021; Gu & Dao, 2023; Dao & Gu, 2024) have gained traction in the field of sequence modeling thanks to their ability to overcome the quadratic bottleneck and scale to extremely long context lengths. The basic linear time-invariant (LTI) SSM is a linear recurrent neural network (RNN) with the update rule:

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Ch_t, \tag{2}$$

where x and y are the input and output sequences, respectively, h is the RNN state, and A, B, C are learnable parameters. The recent Mamba SSM generalizes LTI SSMs to have input-dependent parameters $B(x), C(x)$, allowing the model to selectively attend between token positions, much like a transformer. In fact, the SSM update can be written as an attention-like update $y = M(x)x$, where $M(x) = \textit{contraction}(A, B(x), C(x))$. Imposing scalar-times-diagonal structure on A makes $M(x)$ semi-separable (a form of low-rank), enabling efficient matrix multiplication via a parallel scan.

2 Methods

Our structure tokenizer model is a QAE, as shown in Fig. 1. We represent entire biomolecular systems as 3D atomic point clouds, encode atom positions into latent vectors, quantize said vectors into tokens, and finally decode tokens back into the 3D point cloud.

Tokenizing 3D point clouds via quantization. Quantization networks learn a discrete representation, or vocabulary, of the training data. Prior works such as ESM-3 use vector quantization (VQ) (Gray, 1984). VQ suffers from codebook collapse and requires auxiliary loss functions during training. Instead, we use Finite-Scalar Quantization (FSQ), which does not require regularization terms and produces a more efficient coverage of the

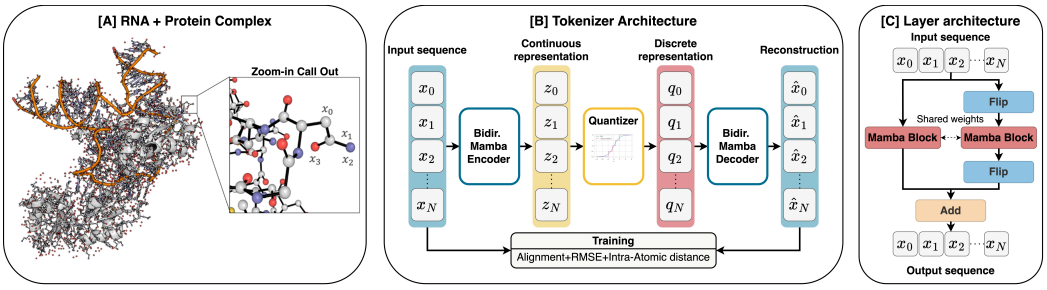

Figure 1: **[A]** Biomolecular system of many thousands of atoms **[B]** The tokenizer transforms point clouds into tokens and then back to point clouds. **[C]** Implementation details of the bidirectional Mamba layer. Following prior works, we use the flip operation to handle bidirectionality.

codebook (Mentzer et al., 2023) via a simple rounding scheme. FSQ projects the input into a hypercube of integer length L and dimensions D (where $D < 8$ usually), then rounds to the nearest integer set $\{0, 1, ..., L\}$. The final code/token is the product of all integer coordinates in the hypercube.

Loss function The ground truth and the decoded point clouds X and $\tilde{X}$ are aligned via Umeyama-Kabsch algorithm (Lawrence et al., 2019). The loss is then the sum of the SVD-aligned RMSE and the inter-atomic distance loss, which encourages isometry between the ground truth and reconstructed structures:

$$L(X, \tilde{X}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|x_i - \tilde{x}_i\|^2} + \sqrt{\frac{1}{n}\sum_{r}\frac{1}{d_r}\sum_{i\in\mathcal{R}_r}\sum_{\substack{j\in\mathcal{R}_r \\ j\neq i}} \left(\|x_i - x_j\| - \|\tilde{x}_i - \tilde{x}_j\|\right)^2}$$

Here, $\mathcal{R}_r$ is the set of atom indices in residue r, and d_r is the number of pairs in r. In the case of small molecules, this is calculated over the entire molecule.

3 Experimental Details

Datasets: An overview of all training and test data is provided in the Appendix table 2. We use the ∇^2DFT dataset for small molecules; CATH 4.2, CASP14 and CASP15 for proteins; and RNA3DB for RNA structures. We also test bio2token at inference time on multi-chain complexes and protein-RNA complexes. Note that neither such complexes were included in the training. See Appendix A.1 for more details about the datasets.

Architecture: Figure 1B and C gives an overview of each layer composition and full architecture. Each layer of our encoder and decoder is a bidirectional implementation of the original *Mamba block*[1]. We ran various hyperparameter studies on a protein2token training with the CATH 4.2 protein dataset. We tested the effects of varying encoder and decoder layers on the model performances in terms of RMSE and found that, given limited compute, 4 encoder layers and 6 decoder layers to work best as a trade-off between model size and batch size. We use a codebook size of 4096, which is in line with other published structure tokenizers. Additional details on the effect of the number of encoder layers, compressibility of tokens, and other architectural ablations are provided in Appendix A.3.

[1]The Mamba block contains two branches; the selective SSM branch with a linear projection, followed by a one-dimensional convolutional layer and a nonlinear activation; and the skip connection branch that is a linear projection followed by a non-linear activation. This is directly imported from the implementation of (Gu & Dao, 2023)

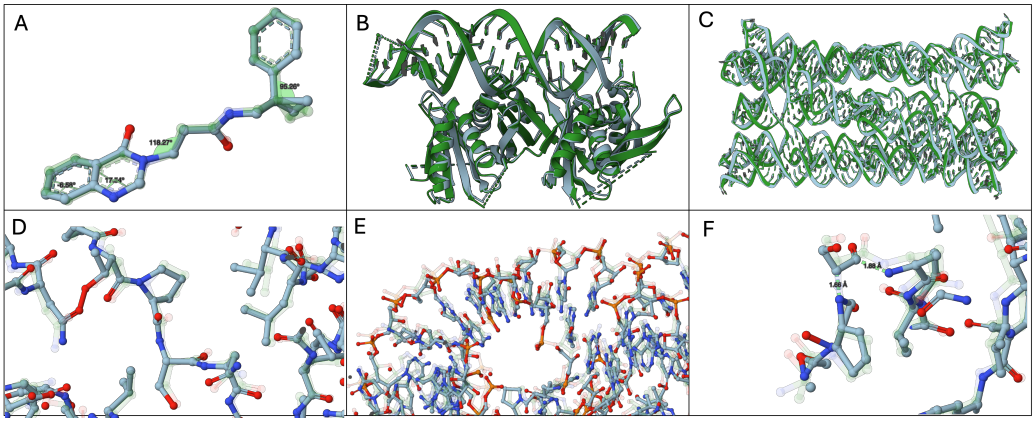

Figure 2: Ground truth molecules in green and reconstructions in blue. Ground truth molecules are transparent in the ball-and-stick panels. Visuals prepared with Mol* (Sehnal et al., 2021) (A), (B), and (C) are reconstructions of a small molecule by mol2token, RNA-protein complex by bio2token, and multi-chain RNA complex by bio2token, respectively. (D) neighborhood of residue on loop of 3WBM found near center of coordinate space (E) close up of RNA helix of 3WBM (F) Example of errors found near edge of coordinate space.

4 Results

Table 1 summarizes the results of bio2token on all test sets. A more detailed version with separate analysis on back-bones and side-chains as well as the numeric results for the domain-specific tokenizers mol2token, protein2token, and rna2token, and their out-of-domain performance are provided in the Appendix tables 6, 7 and 8. Fig. 3 visualizes all reconstruction RMSEs on all biomolecular test sets with in-domain, out-of-domain and all-domain (bio2token) tokenizers.

Best Model	**Test-set**	**RMSE ± std (95% CI) [Å]**	**Validity Test**
Mol2Token on small molecules	test-conformers test-structure test-scaffolds	**0.2±0.04 (0.01)** **0.2± 0.04 (0.01)** **0.2± 0.04 (0.01)**	41.7%
Bio2token on proteins	CATH4.2 test CASP14 CASP15	**0.56±0.06 (0.01)** **0.58±0.10 (0.02)** **0.59± 0.11 (0.02)**	TM_{prot}: 0.98±0.01 TM_{prot}: 0.99±0.01 TM_{prot}: 0.98±0.02
Bio2token on RNA	RNA3DB-test	**0.66± 0.21 (0.01)**	TM_{RNA}-score: 0.96 ± 0.12
ESM-3 Tokenizer on proteins	CASP14 CASP15	1.3 ± 0.2 1.7 ±0.4	–
InstaDeep on proteins	PDB sub-set	back-bone: 1.89	TM_{prot}: 0.94

Table 1: Summary of the best tokenizer models: Atom-wise RMSE between the ground truth structure point cloud and the reconstructed point cloud from the tokens. Validity tests are described in Appendix A.5.

Small molecules: mol2token reconstructs small molecule conformers of unseen molecules and unseen scaffold families with an average RMSE of 0.2Å versus 0.36Å for the combined model bio2token. Fig. 2A shows a valid reconstructed conformer. from the test set on top of the ground truth conformer. We found that 41.7% of all reconstructed molecules with mol2token passed all of our validity metrics.

Proteins: bio2token outperforms protein2token on CASP14 and CASP15 test hold-outs with RMSE values around 0.58Å and 0.59Å versus 0.61Å and 0.8Å. This is significantly lower than ESM-3's decoder reconstruction on CASP14 (1.3Å) and 15 (1.7Å) that infers all-atom structure from the residue-level only encodings. InstaDeep's back-bone tokenizer

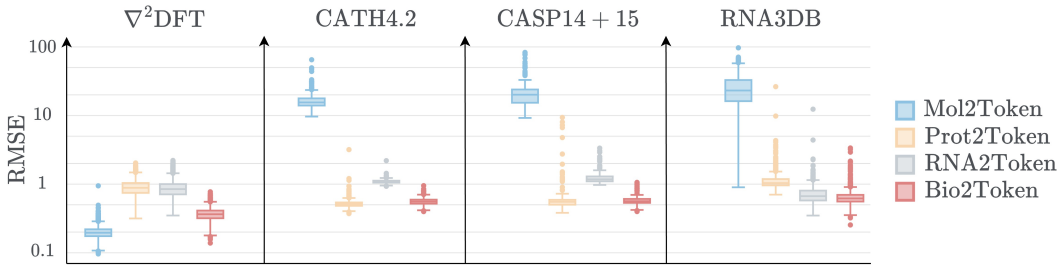

Figure 3: reconstruction results on all test data. Numeric values are provided in Appendix tables 6 - 8. The combined tokenizer Bio2Token achieves competitive reconstruction against domain-specific tokenizers for small molecules, proteins (CATH4.2, CASP14/15), and RNAs, achieving RMSEs of 0.25-0.35 Å, 0.56-0.59 Å, and 0.6Å, respectively.

compares with a back-bone RMSE of 1.89Å to bio2token's back-bone RMSEs of 0.52-0.55Å across the different protein test sets. Generally the TM_{prot} for Bio2Token are all above 0.99, indicating that structural homology in terms of tertiary structure is highly preserved.

RNAs: bio2token reconstructs the RNA3DB test dataset with the lowest RMSE average of 0.66Å on all atoms, compared to 0.73Å for rna2token. The largest RNA chain in the RNA3DB test data is 8toc.R with 4,269 nucleic acids and around 90,441 atoms. Rna2token achieves RMSE of 1.53Å on this structure, compared to Bio2token with 1.82Å.

Complexes: Here, we tested to what degree we can encode and reconstruct RNA-protein and multi-chain complexes with the QAE, despite having never trained on them. We achieve around $0.77 - 0.82$Å for protein-RNA and multi-chain RNA complexes, which range from 3,000-15,000 atoms.

Computational efficiency: Mamba versus Invariant Point Attention (IPA) We compare our Mamba QAE with IPA, which is the most popular choice for structural modeling due to its SE(3) invariance. Accuracies and run times are listed in Appendix A.3 table 5. We generally find that training protein2token with an IPA-decoder is roughly 3 times slower than the Mamba QAE under similar hyperparameter configurations.

Insights to what Bio2Token learns: For details concerning what Bio2Token learns, including its error distribution across different points and rotational variance, we refer the reader to Appendix A.7.

5 Discussion and Limitations

We explored Mamba's potential for encoding high-resolution biomolecular structures, demonstrating that a simple Mamba-based architecture enables scaling to large biomolecules without SE(3) invariance. Our tokenizer learns encodings across macromolecular classes at atomic resolution, achieving reconstruction accuracies of 0.5–0.6Å, from a 4096-token vocabulary. Moreover, Bio2Token scales much more favorably compared to IPA. The comparable small amount of data (127,000 macromolecules in total) used in our trainings signals that all-atom encoding might substantially enhance training efficiency, compared to more coarse-grained encoding that lack atomistic detail, and leverages more information from the structures. Atomistic detail is important for many biomolecular design applications – the precise positioning of individual atoms within a protein or RNA molecule can significantly impact its function and interactions with other molecules.

However, low RMSE alone does not ensure chemically valid reconstructions—minor coordinate deviations can result in steric clashes or incorrect bonding. As shown in Fig. 2F, our model sometimes misrepresents covalent connectivity. Future improvements could involve

larger datasets, physics-based post-processing such as those used in Abramson et al., or explicit structural constraints. Nonetheless, Mamba-based architectures offer a compelling alternative to transformers for atomic-resolution biomolecular modeling. Our quantized QAE formulation enables compatibility with language models.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Anonymous. Tokenizing 3d molecule structure with quantized spherical coordinates. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=UqrSyATn7F`. under review.

Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron LM Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622 (7983):637–645, 2023.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.*, 15:3130–3139, 2024. doi: 10.1039/D3SC04185A. URL `http://dx.doi.org/10.1039/D3SC04185A`.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*, 2023.

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z Li. Foldtoken: Learning protein language via vector quantization and beyond. *arXiv preprint arXiv:2403.09673*, 2024.

Benoit Gaujac, Jérémie Donà, Liviu Copoiu, Timothy Atkinson, Thomas Pierrot, and Thomas D Barrett. Learning the language of protein structure. *arXiv preprint arXiv:2405.15840*, 2024.

Sha Gong, Chengxin Zhang, and Yang Zhang. Rna-align: quick and accurate alignment of rna 3d structures based on size-independent tm-scorerna. *Bioinformatics*, 35(21):4459–4461, 2019.

Robert Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, et al. Nabla2dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. *arXiv preprint arXiv:2406.14347*, 2024.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384 (6693):eadl2528, 2024.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Jim Lawrence, Javier Bernal, and Christoph Witzgall. A purely algebraic justification of the kabsch-umeyama algorithm. *Journal of research of the National Institute of Standards and Technology*, 124:1, 2019.

Xiner Li, Limei Wang, Youzhi Luo, Carl Edwards, Shurui Gui, Yuchao Lin, Heng Ji, and Shuiwang Ji. Geometry informed tokenization of molecules for language model generation. *arXiv preprint arXiv:2408.10120*, 2024.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3:1–14, 2011.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, and Alexander S Rose. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1):W431–W437, 05 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab314. URL `https://doi.org/10.1093/nar/gkab314`.

Marcell Szikszai, Marcin Magnus, Siddhant Sanghi, Sachin Kadyan, Nazim Bouatta, and Elena Rivas. Rna3db: A structurally-dissimilar dataset split for training and benchmarking deep learning models for rna structure prediction. *Journal of Molecular Biology*, pp. 168552, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4): 702–710, 2004.

Artem Zholus, Maksim Kuznetsov, Roman Schutski, Rim Shayakhmetov, Daniil Polykovskiy, Sarath Chandar, and Alex Zhavoronkov. Bindgpt: A scalable framework for 3d molecular design via language modeling and reinforcement learning. *arXiv preprint arXiv:2406.03686*, 2024.

A Appendix

A.1 Datasets

Small molecules: Small molecules, typically organic molecules below a 500 Dalton weight, are not static. At standard temperatures and pressures they take on various 3D structural conformations, each having a specific conformational energy. We used the ∇^2DFT dataset (Khrabrov et al., 2024) of 1.9M small molecules with a total of 16M simulated structural conformations as a source of data. This dataset provides train and test splits for multiple levels of generalizability: a *test-conformer* split of unseen conformations of molecules, a *test-structure* split of unseen molecules and all their conformations, and a *test-scaffold* split of unseen scaffold classes of molecules and their conformations. The minimum number of heavy atoms in this dataset is 8 and the maximum is 27.

Proteins: We prototype and run various hyperparameter studies on the CATH 4.2 dataset of 18k protein structures from the PDB, with train-test splits on the CATH topology classifications as defined by Ingraham et al. (2019). This dataset comprises proteins of 40 to 500 amino acids in length, for a minimum and a maximum of 282 and 4,173 heavy atoms. We also tested on the CASP14 and CASP15 datasets, to compare to the values reported by ESM-3. CASP14 and 15 structures were published after CATH4.2 and are thus not contained in 4.2. CASP14 and 15 contain proteins up to 2,265 residues in length with the biggest structure having 18,042 heavy atoms. To train the large bio2token model we leverage the Alphafold database (AFDB). We use a random sub-set of 100k clusters from FoldSeek's sequence-structure clusters (Barrio-Hernandez et al., 2023), and collect one structure per cluster.

RNA: We train on RNA3DB, which splits the RNA structures in the PDB into sequence-based and structural homology classes (Szikszai et al., 2024). The structures span a range of 2 to 4,450 nucleic acids in lengths with 42 to 95,518 heavy atoms. For training efficiency, we limit the training dataset to structures with maximum $10,000$ sequence length, but run inference on all lengths of the test set.

Generalisation to complexes: We test bio2token at inference time on multi-chain complexes and protein-RNA complexes. Note that neither multi-chain nor mixed complexes were included in the training.

Dataset	Dataset size and splits	Structure size	Used in
∇^2DFT	**train**: 8.9M conformers (0.5M molecules) **test-conformer**: 1.5M conformers (1.5M molecules) **test-structure**: 1.2M conformers (176k molecules) **test-scaffold**: 1.1M conformers (177 molecules)	atoms min: 8 atoms max: 27	Mol2Token, Bio2Token
CATH4.2	**train**: 17k structures **test + val**: 1.6k structures	res/atoms min: 40/282 res/atoms max: 500/4.2k	Protein2Token, Bio2Token
CASP14	**test**: 88 structures	res/atoms min: 49/401 res/atoms max: 2.2k/18k	
CASP15	**test**: 155 structures	res/atoms min: 46/341 res/atoms max: 10k/7.9k	
RNA3DB	**train**: 10k structures **test**: 1.4k structures	res/atoms min: 2/42 res/atoms max: 4.5k/96k	RNA2Token, Bio2Token
AFDB sample	**train**: 100k structures	res/atoms min: 21/174 res/atoms max: 2.7k/22k	Bio2Token

Table 2: Summary of training and test datasets, including minimum and maximum number of residues and atoms.

A.2 Hyperparameters and training

We train four models, all with the same number of 4 encoder and 6 decoder layers, and a codebook size of 4096 for a total of 1.2M parameters (see section below for architecture study details). We use the Adam optimizer (Kingma, 2014), with polynomial learning rate

scheduler and a starting learning rate of $3e^{-4}$. Depending on the model, we use 1 or 8 NVIDIA A10 GPUs (24GB / 184GB GPU RAM). We train three biomolecule specific models, *mol2token*, *protein2token*, and *rna2token*, respectivily trained on the ∇^2DFT dataset, CATH4.2 dataset, and RNA3DB, and an harmonized *bio2token* model, trained on all three dataset and a subset of the AFDB dataset. Additionally, we use random rotation for data augmentation. Model specific parameters are:
mol2token: batch size=16, max seq length=64, 216k steps (44 hours), single GPU.
protein2token: batch size=16, max seq length=4160, 195k steps (68 hours), single GPU.
rna2token: effective batch size=32, max seq length=10000, 149k steps (38 hours),8 GPUs.
bio2token: effective batch size=32, max seq length= 10000, 257k steps (73 hours),8 GPUs.

A.3 Architecture studies

Effect of number of encoder blocks The encoder mixes the atom coordinates and the degree of mixing, or "spread" across atom positions is determined by the number of encoder Mamba blocks and hidden state size. To quantify the spread of local information we define the mixing radius as the number of positions that change their token id when the atom at position i is deleted. Here, we fix the hidden state size of 128 and train QAEs with increasing numbers of encoder blocks $n_{enc} = [2, 4, 5, 6]$ and find the mixing radius to be almost linear with a best fit for a second order polynomial, see Figure 4. This relationship is similar to what is expected from a convolution. For example 2 blocks result in a mixing of ± 2.7 positions to the left and right; and 6 blocks mix ± 5.3 positions.

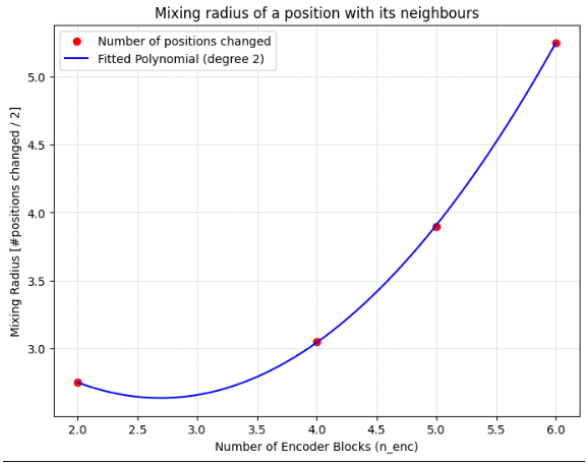

Figure 4: Average mixing radius of per-atom position information with increasing number of Mamba blocks in the encoder.

Codebook size We train protein2token on the CATH4.2 dataset, with a fixed model size. We vary codebook sizes by increasing quantization dimensions $D \in [4, 5, 6, 7, 8]$ with a fixed level of $L = 4$, for total codebook sizes of $[256, 1024, 4096, 16348, 65536]$. We find the accuracy versus codebook size relationship to approximately follow a power law, see Fig. 5. Ultimately, the choice of codebook size will be a trade-off between accuracy and downstream modeling. A tokenizer with increasing vocabulary will make downstream LLM generation harder. For the final training of bio2token we chose 4096 as our codebook size, which is in line with other published structure tokenizers, and allows for a fair comparison.

Effect of various design choices on RMSE We conduct an ablation study to evaluate the impact of additive architectural and training modifications on the performance of the Mamba QAE. All models are trained with identical quantization hyperparameters. We start with a baseline model consisting of 2 encoder and 4 decoder layers, and sequentially add data augmentation through random rotation, bi-directionality, deeper encoder and decoder, and finally the inter-atomic distance loss. The results are presented in Table A.3.

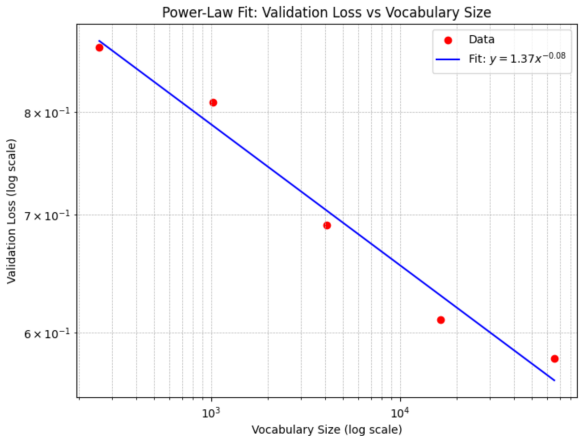

Figure 5: Protein2token (CATH dataset) reconstruction accuracy as a function of codebook size.

Modifying the original encoder/decoder layer to incorporate bi-directionality and increasing the number of layers resulted in a significant improvement, yielding a 22% reduction in reconstruction RMSE. Further enhancing the training strategy with random rotation augmentation and integrating an inter-atomic distance loss, we observe a total RMSE reduction of 28% compared to the baseline.

Model + sequential modification	RMSE (CI ±95%)	Improvement (↓)
Mamba small [2 encoder / 4 decoder layers]	0.72 ± 0.01	-
+ Data augmentation [Rotation]	0.70 ± 0.01	-1.91%
+ Bi-directionality	0.61 ± 0.01	-12.89%
+ Deeper [4 encoder / 6 decoder layers]	0.55 ± 0.01	-11.13%
+ Inter-atomic distance loss	0.52 ± 0.01	-4.53%

Table 3: Ablation study of final model and training choices. Ablation is run on protein2token training with the CATH 4.2 dataset.

Compressibility of tokens To test the compressibility of the token sequences we train the tokenizer with an additional 1D convolutional layer before and after the quantizer network (pooling after the encoder and up-sampling before the decoder). We compress with $k \in [1, 2, 4]$, to shorten the all-atom sequence of length N to N/k. RMSE increases by a factor of 1.7 and 2.6 for the compression factors of 2 and 4 respectively. This is similar to previously reported compressibilities for residue-level structure tokenizers (Gaujac et al., 2024)

Table 4 below shows the the relationship between test set reconstruction RMSD and compressibility factor with a codebook size of 4096. We also tested if increasing the SSM's hidden dimension could increase compressibility, which we found to not be the case.

Compression	**D_model hidden size**	**RMSE** [Å]	**factor of RMSE increase**
1	128	0.86	—
2	128	1.49	1.7
4	128	2.22	2.6
1	1280	0.84	—
2	1280	1.45	1.7
4	1280	2.15	2.6

Table 4: Effect of compression on RMSE. Increasing the hidden dimension does not help noticeably to recover accuracy.

A.4 Model efficiency comparisons

Computational efficiency and performance: Mamba versus IPA We train a protein2token tokenizer with a 2-layer transformer encoder and an IPA decoder with 4 recyclings. Due to GPU memory constraints, training is limited to protein structures of a maximum length of 2192 atoms, at a batch size of 1. We train an equivalent Mamba-based protein2token with 2 encoder Mamba-blocks and 4 decoder Mamba-blocks, with a batch size of 1 and the maximum batch size before GPU memory is exhausted, which is 32. We find that the IPA-based QAE requires 1 sec/step, compared to 0.3sec/step for an equivalent Mamba-based QAE. In terms of achieved validation accuracy IPA-based architecture is significantly worse than the Mamba-based QAE with an RMSE of 2.18 versus 0.81. Likely this is due to the "small" number of IPA-block recycles, often 8 (instead of 4) are cited in the literature. But this becomes prohibitive for sequences lengths of 2192. To compare at the full capacity of the GPU hardware, we find that training for 24 hours with the Mamba-based QAE with a maximum batch size of 32 has superior accuracy with 0.62Å.

Architecture	Time [sec/step]	Validation accuracy after 24h run time [Å]	Validation accuracy after 70k steps [Å]
Transformer encoder, IPA decoder, batch size = 1	1.0	2.18	2.18
Mamba, batch size = 1	0.3	0.81	0.91
Mamba, batch size = 32	0.7	0.62	0.65

Table 5: Effect of compression on accuracy RMSD. Increasing the hidden state size does not recover accuracy.

Codebook efficiency: learned versus spatial tesselation We explore how well the trained QAEs perform relative to idealized voxel partitions and learned voronoi tesselations. For a desired tesselation resolution a (the side length of a voxel), and a total cubic volume of side length A (the maximum spatial extent of biomolecular structures) results in a total number of voxels $N_v = (A/a)^3$. To calculate the average reconstruction accuracy of a point (atom) in a voxel, we calculate the average $rmsd_v$ to the voxel centre:

$$rmsd_v = \frac{8}{a^3} \int_0^{a/2} \int_0^{a/2} \int_0^{a/2} \sqrt{x^2 + y^2 + z^2} dx dy dz$$

With Monte-Carlo integration (not shown) this is approximately $0.48 \times a$. To tesselate a biomolecular structure of cubic volume with a side length A and a desired average accuracy $rmsd_v$, a total voxel count of

$$N_v = \left(\frac{0.48 \times A}{rmsd_v} \right)^3$$

Figure 6 plots the number of total Voronoi voxels needed to encode the 3D space of three exemplar cubes of side length $a = [10, 60, 80]$Å , representative for small molecules, proteins and RNA respectively . We center structures at zero, sample rotations and use k-means clustering to find 4096 cluster centers that are used as the centroids of Voronoi tesselations. Upon comparing these approaches we see that for the tested codebook size, the

QAE approach achieves lower $rmsd_v$, suggesting that it learns beyond the atom coordinate aaddress.

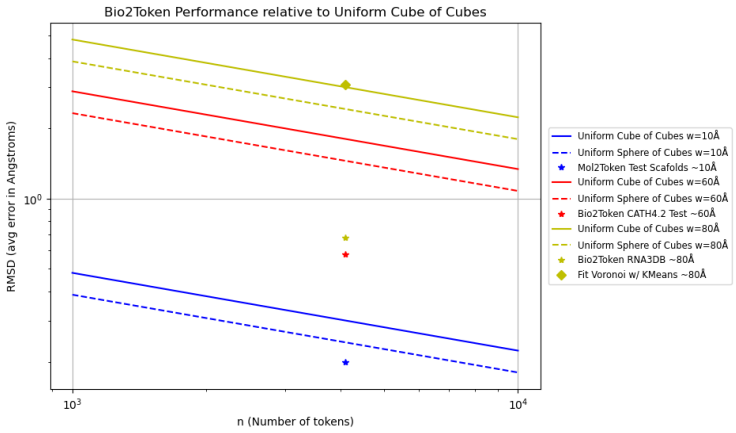

Figure 6: Comparing the reconstruction error between learned tokenizers, trained with 4096 codebook size, a naive tesselation of increasing number of voxels and a k-means Voronoi tesselation approach.

A.5 Validity tests

Small molecule validity: We convert the heavy atom point clouds into molecules by inferring covalent bonds using atom type and inter-atomic distances with OpenBabel (O'Boyle et al., 2011). We first evaluate whether the recovered molecular system is equivalent to the encoded structure and then evaluate bond lengths, angles, and torsion angles using the methods described by Buttenschoen et al. (2024). We use RDKit to compute the energy of the conformer and compare to the average energy of 25 RDKit generated conformers (Landrum, 2013). We compute these statistics for test set ground-truth conformers and the reconstruction to evaluate any change. A reconstruction is said to pass all tests if it passes the tests from PoseBusters as well as produces the same molecular graph as the input.

Proteins and RNA validity: We report the template modeling score (TM-Score) between ground truth and reconstructed point clouds. It captures local and global structural alignment and is designed to be size independent. The protein TM-score $\mathrm{TM}_{\mathrm{prot}}$ is calculated on the C_α of the amino acid back-bone (Zhang & Skolnick, 2004). The RNA TM-score $\mathrm{TM}_{\mathrm{RNA}}$ is calculated on the C3' of the nucleic acid back-bone (Gong et al., 2019). TM=0 means no structural similarity at all; TM=1.0 means structurally identical.

A.6 Tokenizer results

Model	Test-set	RMSE ± std (95% CI) [Å]	Validity Test
Bio2Token on small molecules	test-conformers test-structure test-scaffolds	**0.36±0.07 (0)** **0.37± 0.07 (0)** **0.36± 0.07 (0)**	< 1%
Bio2Token on proteins	CATH4.2 test	bb: 0.52± 0.07 (0.01) sc: 0.59± 0.06 (0.01) **all: 0.56±0.06 (0.01)**	TM_{prot}: 0.98±0.01
	CASP14	bb: 0.54±0.10 (0.02) sc: 0.62±0.09 (0.02) **all: 0.58±0.10 (0.02)**	TM_{prot}: 0.99±0.01
	CASP15	bb:0.55±0.12 (0.02) sc:0.63± 0.12 (0.02) **all: 0.59± 0.11 (0.02)**	TM_{prot}: 0.98±0.02
Bio2Token on RNA	RNA3DB-test	bb: 0.66± 0.21 (0.01) sc: 0.65 ± 0.22 (0.01) **all: 0.66± 0.21 (0.01)**	TM_{RNA}-score: 0.88 ± 0.12
ESM-3 tokenizer on proteins	CASP14	back-bone:0.61 ± 0.1 **all: 1.3 ± 0.2**	
	CASP15	back-bone: 1.0 ± 0.3 **all: 1.7 ±0.4**	
InstaDeep tokenizer on proteins	self-defined test set from the PDB	back-bone: 1.89 side-chains not modeled	TM_{prot}: 0.94

Table 6: Bio2token results: Atom-wise RMSE between the ground truth structure point cloud and the reconstructed point cloud from the tokens. "bb" and "sc" are the respective RMSEs over the back-bone and side-chain atoms in the case of proteins and RNAs. Bio2token is unable to preserve chemical validity of small molecules and mol2token should be used for these structures. For proteins and RNA we provide the TM-scores as a measure of tertiary structural similarity.

A.6.1 In-domain tokenizing

In-domain tokenizing	**Test-set**	**rmse ± std, (95% CI) [Å]**	**Validity Test**
mol2token on small molecules	test-conformers test-structure test-scaffolds	**0.20± 0.04(0.01)** **0.20± 0.04 (0.01)** **0.20± 0.04 (0.01)**	41.7% passed all chemical validity metrics
protein2token on proteins	CATH4.2 test	bb: 0.49±0.12 (0.01) sc:0.56±0.11 (0.01) **all: 0.53±0.12 (0.01)**	TM_{prot}: 0.99±0.01
	CASP14	bb: 0.57±0.21 (0.04) sc: 0.65±0.21 (0.04) **all: 0.61±0.21(0.04)**	TM_{prot}: 0.99±0.01
	CASP15	bb:0.76±1.21 (0.19) sc:0.85±1.25 (0.20) **all: 0.80±1.23 (0.19**)	TM_{prot}: 0.99±0.03
RNA2token on RNAs	RNA3DB-test	bb: 0.73±0.34 (0.02) sc: 0.72 ±0.40 (0.02) **all: 0.73±0.39 (0.02)**	TM_{RNA}-score: 0.86 ± 0.13
ESM-3 Tokenizer on proteins	CASP14	back-bone:0.61 ± 0.1 **all: 1.3 ± 0.2**	
	CASP15	back-bone: 1.3 ± 0.3 **all: 1.7 ±0.4**	
InstaDeep	self-defined test set from the PDB	back-bone: 1.89 side-chains not modeled	TM_{prot}: 0.94

Table 7: In-domain tokenizing: The reconstruction error is the atom-wise rmse between the ground truth structure point cloud and the reconstructed point cloud from the tokens. "bb" and "sc" are the respective rmses over the back-bone and side-chain atoms in the case of proteins and RNAs. Validity tests for small molecules are the chemical validity metrics as described in the main text and for proteins and RNA we provide the TM-scores as a measure of tertiary structural similarity

A.6.2 Out-of-domain tokenizing

Out-of-domain tokenizing	**Test-set**	**rmse ± std (95% CI) [Å]**	**Validity Test**
mol2token on proteins	CATH4.2 test	all: 16.40± 4.07 (0.24)	TM_{prot}: 0.13±0.04
	CASP14	all: 21.37± 10.44 (2.18)	TM_{prot}: 0.13±0.05
	CASP15	all: 23.23± 13.95 (2.20)	TM_{prot}: 0.13±0.06
mol2token on RNA	RNA3DB-test	all: 25.88±12.22 (0.65)	TM_{RNA}: 0.02 ±0.01
protein2token on RNAs	RNA3DB-test	all: 1.16±0.79 (0.04)	TM_{RNA}: 0.81 ±0.16
RNA2token on proteins	CATH4.2 test	all: 1.09±0.07 (0.01)	TM_{prot}: 0.96±0.03
	CASP14	all: 1.27±0.36 (0.08)	TM_{prot}: 0.96±0.04
	CASP15	all: 1.30±0.39 (0.06)	TM_{prot}: 0.96±0.04

Table 8: Applying tokenizers on out-of-domain molecules. Only all-atom rmses are shown here for simplicity. mol2token to proteins and RNAs: The rmse values show the insufficiency of learning larger biomolecular structures just from small molecules. protein2token on RNAs: The rmse if higher than the rna2token reconstruction error (reported in the main text), but is in close proximity. rna2token on proteins: the rmse is slightly worse than the protein2token errors reported in the main text on CATH4.2 and CASP14, but better on CASP15.

A.7 Insights into bio2token

A.7.1 RMSE per atom as a function of distance to centre

Figure 7 shows scatter plots for a sample of 10k points across all structure point clouds with their absolute distance to the centre and their RMSE. RMSE increases once the point's distance to centre increases past the common size range of the training structures. This can also be seen in Fig. 2F, where reconstructions deviate at the periphery of the coordinate space for a structure of about 16,000 atoms).

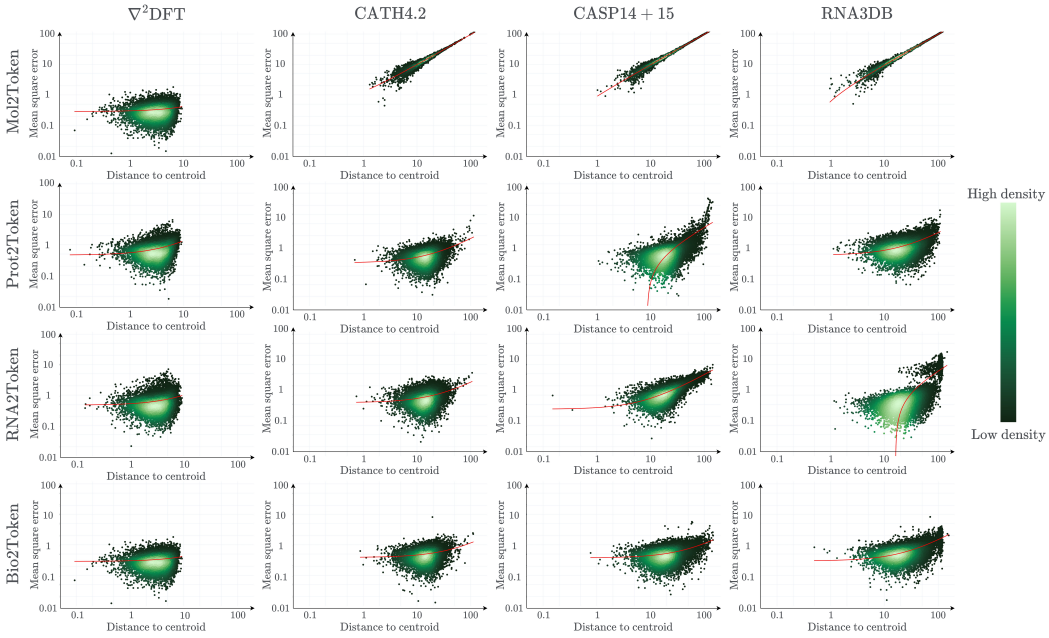

Figure 7: Reconstruction RMSE per point as a function of its distance to the centre. Each subplot is a random sample of 10.000 points across all point clouds of the respective dataset.

A.7.2 Rotational variance of tokens

Bio2Token does not exploit rotational invariance in it's architecture. The Bio2Token tokens are varying periodically with respect to rotations. To visualise the effect we show the individual amino acid `GLN` and its back-bone and side-chain atoms under a set of full 2π rotations around the z- and the x-axis. Fig. 8 shows how the atom token ids shift with respect to changes in orientation. Moreover, reconstruction errors are not biased towards any orientation, as seen in Fig. 9.

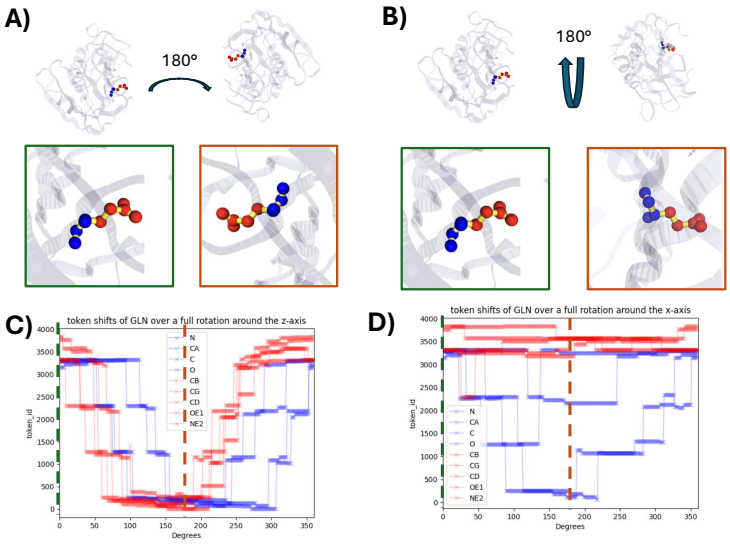

Figure 8: Token circularity with rotations. A and B visualise a π rotation of the protein around the z- and x-axis. The zoom into the `GLN` amino acid shows how the individual atoms are changing orientations with respect to the centre. The respective token ids of each atom on the highlighted `GLN`are plotted in C) and D) as a function of rotation angle. The green and red dotted lines correspond to the tokens at the positions in A) and B).

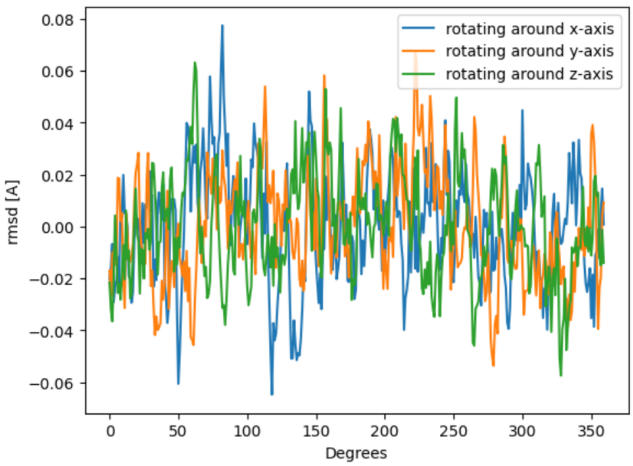

Figure 9: The reconstruction error of an exemplar protein under a full set of 2π rotations around all major axes. The reconstruction error shows no orientation bias.

A.8 Code Availability

Code and model weights are available for all trained tokenizers. Inference scripts are provided for pdb formated files at `https://anonymous.4open.science/r/bio2token-72F2`