Briefings in Bioinformatics, 2025, **26(3)**, bbaf202
https://doi.org/10.1093/bib/bbaf202
Problem Solving Protocol

Predicting the structures of cyclic peptides containing unnatural amino acids by HighFold2

Cheng Zhu[1,‡], Sen Cao[2,‡], Tianfeng Shang[3], Jingjing Guo[2], An Su[1], Chengxi Li[4], Hongliang Duan[2,*]

[1]College of Pharmaceutical Sciences, Zhejiang University of Technology, Chaowang Road, Gongshu District, Hangzhou 310014, China
[2]Faculty of Applied Sciences, Macao Polytechnic University, R. de Luís Gonzaga Gomes, Macao 999078, China
[3]Artificial Intelligence Department, Shenzhen Highslab Therapeutics. Inc, Guangke 1st Road, Pingshan District, Shenzhen 518000, China
[4]College of Chemical and Biological Engineering, Zhejiang University, Yuhangtang Road, Xihu District, Hangzhou 310027, China

[*]Corresponding author. Faculty of Applied Sciences, Macao Polytechnic University, R. de Luís Gonzaga Gomes, Macao 999078, China. E-mail: hduan@mpu.edu.mo
[‡]Cheng Zhu and Sen Cao have contributed equally to this work.

Abstract

Cyclic peptides containing unnatural amino acids possess many excellent properties and have become promising candidates in drug discovery. Therefore, accurately predicting the 3D structures of cyclic peptides containing unnatural residues will significantly advance the development of cyclic peptide-based therapeutics. Although deep learning-based structural prediction models have made tremendous progress, these models still cannot predict the structures of cyclic peptides containing unnatural amino acids. To address this gap, we introduce a novel model, HighFold2, built upon the AlphaFold-Multimer framework. HighFold2 first extends the pre-defined rigid groups and their initial atomic coordinates from natural amino acids to unnatural amino acids, thus enabling structural prediction for these residues. Then, it incorporates an additional neural network to characterize the atom-level features of peptides, allowing for multi-scale modeling of peptide molecules while enabling the distinction between various unnatural amino acids. Besides, HighFold2 constructs a relative position encoding matrix for cyclic peptides based on different cyclization constraints. Except for training using spatial structures with unnatural amino acids, HighFold2 also parameterizes the unnatural amino acids to relax the predicted structure by energy minimization for clash elimination. Extensive empirical experiments demonstrate that HighFold2 can accurately predict the 3D structures of cyclic peptide monomers containing unnatural amino acids and their complexes with proteins, with the median RMSD for Cα reaching 1.891 Å. All these results indicate the effectiveness of HighFold2, representing a significant advancement in cyclic peptide-based drug discovery.

Keywords: cyclic peptides; unnatural amino acids; structure prediction; multi-scale modeling

Introduction

Peptides are compounds composed of amino acids linked by peptide bonds [1], with molecular weights between small molecules and macromolecules. Renowned for their ability to selectively modulate diverse protein–protein interactions, peptides are pivotal in drug development [2, 3]. Peptide-based therapeutics offer unique advantages: compared to small-molecule drugs, they exhibit greater specificity and reduced toxicity; relative to protein-based drugs, they possess diminished immunogenicity and lower production costs [4]. These attributes position peptides as compelling candidates for treating a broad spectrum of diseases [5]. Nonetheless, linear peptides are susceptible to being hydrolyzed by proteases [6]. Cyclization, coupled with the incorporation of unnatural amino acids, can markedly enhance their stability, binding affinity, specificity, and membrane permeability, thus enabling them to target intracellular proteins effectively [7, 8]. Given these distinctive benefits of cyclic peptides with unnatural amino acids, accurate prediction of their monomeric and complex spatial structures has become increasingly critical.

The rapid evolution of artificial intelligence algorithms has revolutionized structural prediction models. Tools such as AlphaFold2 [9], AlphaFold-Multimer [10], RoseTTAFold [11], and ESMFold [12] have achieved remarkable accuracy in predicting the structures of protein monomers and protein–protein complexes. Building upon these milestones, subsequent advancements have extended predictive capabilities to encompass protein-nucleic acid [13], protein–peptide [14], protein-small molecule [15], and protein-cyclic peptide complexes [16]. Recently, models like RoseTTAFold All-Atom [17] and AlphaFold3 [18] have further broadened the horizon of deep learning-based spatial structure prediction to the joint structure of complexes involving proteins, nucleic acids, small molecules, ions, and modified residues. However, despite these advancements, current methodologies remain limited. They either predict structures for cyclic peptides containing natural amino acids or linear peptides with unnatural amino acids, none specifically address the structural prediction of cyclic peptides incorporating unnatural amino acids and their associated complexes.

In this study, we leveraged AlphaFold-Multimer to achieve accurate spatial structure predictions for cyclic peptides with unnatural amino acids and their complexes. Considering peptides have certain small molecule properties, we integrated a

Received: February 3, 2025. **Revised:** March 25, 2025. **Accepted:** April 9, 2025

neural network that characterizes peptide atomic properties into AlphaFold-Multimer, enabling multi-scale modeling of peptide molecules and distinguishing different unnatural residues. Furthermore, we extended the rigid groups of natural amino acids to encompass unnatural amino acids. These enhancements allowed us to fine-tune the combined model using 3D structural data containing unnatural amino acids. While this model could be directly trained on cyclic peptides with unnatural amino acids, the scarcity of such structures in the Protein Data Bank (PDB) [19, 20] necessitated an alternative approach. So, we use the zero-shot learning method, first fine-tuning the model using linear peptides with unnatural amino acids and then modifying the relative position encoding matrix to enable accurate cyclic peptides and their complexes predictions.

The prediction results demonstrate that HighFold2 can accurately predict the spatial structures of cyclic peptides containing unnatural amino acids and their complexes. Compared to the native structure, the median RMSD (root-mean-square deviation) between the Cα atoms in the predicted and native structures is 1.891 Å, while the median RMSD for all atoms in the structure reaches 2.872 Å, and the median RMSD between atoms in the unnatural amino acids is 2.579 Å. Subsequent ablation experiments further demonstrated the efficacy of the individual components of the model in predicting cyclic peptide structures with unnatural amino acids. Prediction results also showed that HighFold2 performed well on the test set of linear peptides containing unnatural amino acids. Additionally, we parameterized the unnatural amino acids and performed energy minimization to relax the predicted structures, thereby eliminating potential spatial clashes.

In summary, we have developed a method termed HighFold2 capable of accurately predicting the structures of cyclic peptides containing unnatural amino acids and their complexes. This method involves training a deep learning model based on AlphaFold-Multimer using linear peptide structures with unnatural amino acids, then modifying the model's relative position encoding matrix, enabling it to predict the cyclic peptide structures successfully. Then, relaxation is performed to refine the spatial structure further. We believe that this method will serve as a powerful tool for the development of cyclic peptide-based therapeutics.

Methods and materials

Datasets

In this study, we utilized two datasets: linear peptides with unnatural amino acids and cyclic peptides with unnatural amino acids. The linear peptide dataset with unnatural amino acids was used to train the model, while the cyclic peptide dataset with unnatural amino acids was used exclusively for testing the model.

The linear peptide data containing unnatural amino acids were sourced from the ModPep dataset [21], which includes 501 samples of peptides with unnatural amino acid residues. Most of these are complexes formed from linear peptides with unnatural residues and proteins. Using the provided PDB IDs, we downloaded the corresponding crystal structures from the PDB database, removed duplicate entries, and ultimately obtained 419 unique 3D structures. We removed solvents, hydrogen atoms, and heteroatoms from the crystal structures to facilitate subsequent model training and testing and retained only each structure's first state.

To minimize redundancy in the crystal structures and reduce computational demands during training, we applied the following rules to remove certain chains from the spatial structure: chains with fewer than 50 amino acid residues were classified as peptide chains, while others were defined as protein chains. If a structure contained at least one protein chain, we aligned the sequences of all peptide chains. For peptide chains with identical or partially overlapping sequences separated by a distance greater than 5 Å, only the longest peptide chain was retained, and the others were discarded. Additionally, if at least two protein chains remained in this structure, we retained only the protein chain that contacts most with the peptide chain.

Then, we analyzed the types of unnatural amino acids in the crystal structures. Any structure containing the unnatural amino acid that was only found once was excluded. Samples containing nucleic acids or acetyl groups were also removed. After this screening, 382 samples remained, representing 23 distinct unnatural amino acid residues. The statistics of this dataset can be seen in Table S1.

Besides, cyclic peptides with unnatural amino acids were sourced from the cPEPmatch webserver [22]. To effectively evaluate the model's accuracy on cyclic peptide data, we filtered the dataset to retain samples that only contained the same 23 types of unnatural amino acid residues as in the linear peptide dataset. We excluded samples with other unnatural amino acids and limited the cyclization forms to head-to-tail cyclization or disulfide bridge formation. Additionally, we also searched the PDB database to identify some structures that met these criteria. After downloading the corresponding crystal structures from the PDB, we removed solvent molecules, hydrogen atoms, and heteroatoms, retaining only the first state of each structure. We also excluded samples containing nucleic acids and duplicate entries. In the end, we obtained a total of 34 samples of cyclic peptides containing unnatural amino acids.

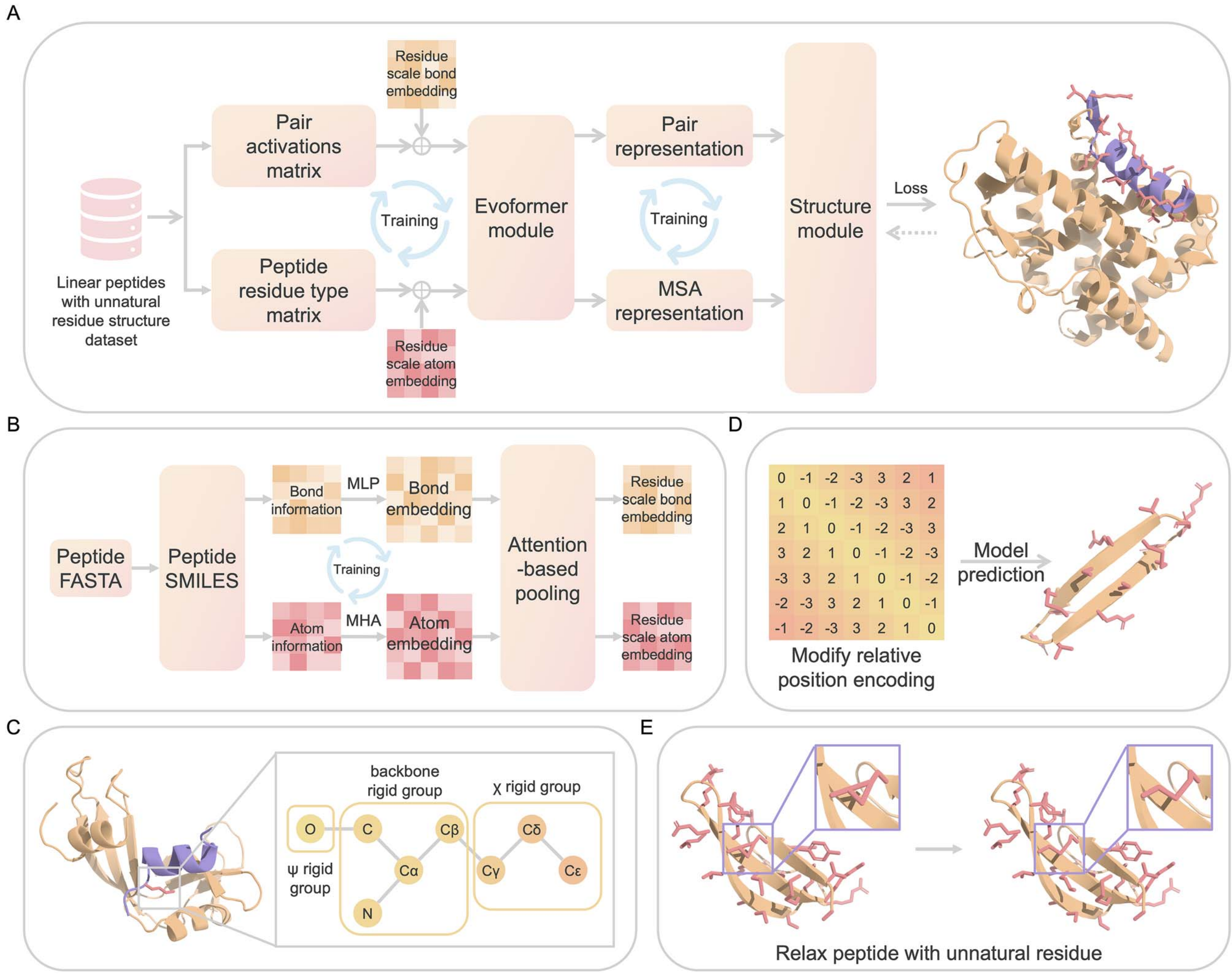
The architecture

HighFold2 is based on AlphaFold-Multimer and fine-tuned with its parameters. To enable the prediction of the 3D structures of peptide monomers containing unnatural amino acids and their complexes, we introduced two major modifications to the AlphaFold-Multimer architecture: the integration of a neural network module to extract atomic-scale features of peptide molecules and the extension of AlphaFold-Multimer's rigid groups to include unnatural amino acids.

Atomic-scale feature extraction

AlphaFold-Multimer originally extracts features of peptides at the residue level, omitting atomic-scale information. To address this limitation and distinguish various unnatural residues, we incorporated an additional neural network module to capture atomic-scale representations of peptides. These representations were then combined with the residue-level features from AlphaFold-Multimer to achieve multi-scale modeling, as shown in Fig. 1B. The atomic-scale features were derived from atom element and bond information, which was extracted by converting chains with fewer than 50 amino acid residues in crystal structures from FASTA [23] to SMILES [24] using RDKit.

Atom element features were encoded using one-hot encoding. To reduce the sparsity of the encoding matrix, we separately encoded common atoms (B, C, F, I, N, O, P, S, Br, Cl), while rare atoms were assigned a shared encoding. The resulting one-hot matrix was processed using a multi-head attention mechanism



0	-1	-2	-3	3	2	1
1	0	-1	-2	-3	3	2
2	1	0	-1	-2	-3	3
3	2	1	0	-1	-2	-3
-3	3	2	1	0	-1	-2
-2	-3	3	2	1	0	-1
-1	-2	-3	3	2	1	0

Figure 1. Overview of HighFold2. (A) The training process of structure prediction model. (B) The architecture of the atomic-scale feature extraction module for peptides. MLP and MHA represent the multi-layer perceptron and multi-head attention, respectively. (C) The definition of rigid groups for unnatural amino acids. (D) The method for modifying the relative position encoding in the prediction model to enable the prediction of cyclic peptide monomers and complexes containing unnatural amino acids. (E) The relaxation of 3D structures containing unnatural amino acids to eliminate spatial clashes.

(MHA) [25] defined as follows:

$$MultiHead(x) = Concat\left(Head_1, Head_2, \ldots, Head_i\right) W,$$

$$Head_i = Attention\left(xW_i^Q, xW_i^K, xW_i^V\right),$$

$$Attention\left(Q, K, V\right) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) V,$$

where, x represents the matrix of encoded atom element features with eight heads in total, Q, K, and V denote query, key, and value matrices, and d_K is the dimension of the key matrix. All Ws represent learnable parameters. Then, the multi-head attention outputs were passed through a multi-layer perceptron (MLP) [26] to transform their dimensions to 21 (corresponding to the 20 natural amino acids and one unknown amino acid). The i-th layer of MLP is defined as:

$$MLP_i\left(x\right) = ReLU\left(W_i x + b_i\right),$$

where x denotes the input atom element feature embeddings of the i-th MLP layer. W_i and b_i are the learnable weights and biases.

To integrate atomic-scale features into AlphaFold-Multimer, we applied an attention-based pooling mechanism [27], mapping atomic-level features to residue-level representations:

$$F = \frac{1 - M + \epsilon}{M - \epsilon},$$

$$A_{score} = Softmax\left(x_{atom} + F\right),$$

$$Pooling_{atom} = x_{atom}^T A_{score},$$

where M is the mask matrix ($N \times N_r$, where N is the number of atoms, and N_r is the number of residues), indicating the residue to which each atom belongs, ϵ is a small positive value to prevent division by zero, and x_{atom} represents atomic element embeddings after the MLP layers. The resulting $Pooling_{atom}$ ($N_r \times 21$) was added to the AlphaFold-Multimer's one-hot matrix of peptide residue type and then input into the Evoformer module to generate the single representation. To keep it concise, we omit the dimensional extension and reduction operations of variables in the formula.

Bond information between atoms was also encoded using the one-hot approach, covering five bond types (single, double, triple, aromatic, and ionic bond). The resulting matrix ($N \times N \times 5$) was transformed via a linear layer to reduce its dimensionality to N × N × 1. Then, using a pooling mechanism like that described above, the bond information was mapped to residue-level representations and added to AlphaFold-Multimer's pair activations matrix of peptide to generate the pair representation. These modifications enabled the model to effectively capture and integrate atomic-scale details into the residue-level framework of AlphaFold-Multimer, facilitating the prediction of 3D structures for peptides containing unnatural amino acids and their complexes.

Extend residue rigid groups to unnatural amino acids

The structure module in AlphaFold-Multimer leverages the single representation and pair representation derived from the Evoformer module to predict the torsion angles of each amino acid residue. Then, these predicted torsion angles, along with predefined rigid groups and atomic coordinates, facilitate the construction of all atomic coordinates in the spatial structure, simplifying the prediction complexity. We implemented a similar approach by defining rigid groups and initializing atomic coordinates for unnatural amino acids to extend the model to these residues, as shown in Fig. 1C.

Each atom within an amino acid residue can be categorized into one of five rigid groups based on its dependence on specific torsion angles. These groups include the backbone rigid group, which contains the backbone atoms Cα, Cβ, C, and N, while the ω and ϕ rigid groups, associated with the hydrogen atoms on Cα and the amino group, respectively, are excluded since our model does not predict hydrogen atoms. The ψ rigid group encompasses the oxygen atom in the carboxyl group, and the χ rigid group includes all side chain atoms. Specific details of the rigid group assignments for unnatural amino acids are documented in Table S2.

After defining the rigid groups, their initial coordinates are determined based on the crystallographic structures of unnatural amino acids. For the backbone rigid group, the coordinate system is centered at Cα, with C positioned along the positive x-axis, N in the xy-plane, and Cβ calculated relative to these references. The ψ rigid group is initialized with C as the origin, Cα on the negative x-axis, and the nitrogen of the subsequent residue within the xy-plane, allowing the position of the oxygen atom to be established. The side chain χ rigid group which are organized into four subgroups dependent on different torsion. Despite some amino acids having more than four torsion angles, smaller angles with negligible effects are disregarded. Within each subgroup, the third atom is set as the origin, the second atom is aligned to the negative x-axis, and the first atom lies in the xy-plane. The relative position of the fourth atom is rotated into the xy-plane using a rotation matrix R_x, as follows:

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta & -sin\theta \\ 0 & sin\theta & cos\theta \end{bmatrix},$$

$$Coordinate_{init} = R_x \bullet Coordinate_{relative},$$

where θ represents the rotation angle between the relative position of the fourth atom and the xy-plane when it rotates around the x-axis.

Modify relative position encoding for cyclization

To represent cyclic position information, we construct a cyclic position matrix by calculating the shortest distance for any two residues, as shown in Fig. 1D. For cyclic peptides with the head-to-tail constraint, the cyclic position can be formulated as:

$$p_{ij} = \left(N_r - \left|N_r - 2d_{ij}\right|\right) \times S\left(i - j\right)/2,$$

$$(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & if\ x = 0 \\ -1, & if\ x < 0 \end{cases},$$

where p_{ij} denotes the relative position between residue i and residue j in the cyclic position matrix, d_{ij} denotes the relative distance between residue i and residue j, N_r denotes the number of residues (including the unnatural residues), |•| denotes the symbol of absolute value.

For the scenario of cyclization with disulfide bridge constraints, the cyclic position matrix becomes more complicated due to the existence of disulfide bridges. As illustrated in HighFold [16], the relative position between any two residues is calculated as the shortest distance between them in the residual topological graph based on the classic Floyd-Warshall algorithm. The signal assignment is conducted according to the directionality of amide bonds in the practical position offset. For brevity, we do not elaborate further here, and please refer to the HighFold model for details.

Training

HighFold2 is trained based on the dataset of linear peptides, dividing it into training, validation, and test sets at a ratio of 7.5:1.5:1. Each unnatural amino acid was ensured to appear at least once in both the validation and test sets. Features required for training and validation were generated using ColabFold [28], while template features were excluded to prevent data leakage. Due to the limitation of computational resources, we cropped the number of amino acid residues in each training sample to 220 while the validation set remained uncropped. To mitigate overfitting, early stopping was implemented based on the loss of the validation set. We fine-tuned all five sets of parameters from AlphaFold-Multimer, allowing us to generate five predicted structures during testing, like the procedure of AlphaFold-Multimer. Table S3 shows the hardware used in our training.

Relaxation

Parameterization of unnatural amino acid residues

To be compatible with Amber force fields [29], force field parameters of unnatural amino acid residues were generated by following its official tutorial. In brief, each unnatural amino acid residue was capped with N-methyl and acetyl groups in the form of ACE-XXX-NME in GaussView. Structure optimization was then performed with Gaussian 16 by using density functional theory (DFT) [30] at the B3LYP/6-31G(d) level, followed by a calculation at the level of HF/6-31G(d) on the optimized structure to reproduce the electrostatic potential. Subsequently, the atomic partial charges of the unnatural amino acid residue were generated using the restricted electrostatic potential (RESP) [31] method with Antechamber, during this process the charge of the capping groups was fixed. The penalty function of RESP is formulated as follows:

$$X_{penalty} = a \sum_{n=1}^{nuclei} \left(\sqrt{q_n^2 + b^2} - b \right),$$

where a defines the asymptotic limit of the penalty, q_n represents the partial charge of the n-th atom, and b is the width parameter.

The charge sets of the ACE and NME groups are identical to the Amber ff14SB force fields [32]. Afterward, the capping groups were removed to generate the force field parameters of the unnatural amino acid residue. Finally, the .prepin and .frcmod files were generated based on the General Amber Force Field (GAFF) [33].

Energy minimization

The PDB file is loaded in AmberTools [34], and the structural topology and coordinate files were created by combining the Amber ff14SB, GAFF, and the generated unnatural amino acid residue parameters. For cyclic peptides, the 'bond' command is used to establish connections between the pair of atoms; for complex systems, ParmEd [35] is employed to specify the chain names for each peptide chain. The generated topology and coordinate files are then loaded into OpenMM [36] for an iterative restrained energy minimization procedure. Harmonic restraints with a spring constant of 10 kcal/mol·Å^2 are applied to heavy atoms to maintain proximity to the input structure. Violations are resolved by identifying residues with clashes, removing their restraints, and re-minimizing iteratively until all issues are addressed. The process is performed with OpenMM's default tolerance of 2.39 kcal/mol and no step limit, ensuring structural integrity and proper hydrogen placement.

Metrics

RMSD calculation

We evaluated the model's predictive performance by calculating the root-mean-square deviation (RMSD) between the predicted and actual structures. Specifically, the pocket-aligned peptide $\mathrm{RMSD}_{\text{all-atom}}$, pocket-aligned peptide $\mathrm{RMSD}_{C\alpha}$, and pocket-aligned peptide $\mathrm{RMSD}_{\mathrm{unAA}}$ were used to assess the structural accuracy of complexes containing unnatural amino acids, while $\mathrm{RMSD}_{\text{all-atom}}$, $\mathrm{RMSD}_{C\alpha}$, and $\mathrm{RMSD}_{\mathrm{unAA}}$ were employed for evaluating the structures of peptide monomers with unnatural amino acids.

The pocket-aligned peptide $\mathrm{RMSD}_{\text{all-atom}}$ and pocket-aligned peptide $\mathrm{RMSD}_{\mathrm{unAA}}$ are defined as follows: for each crystal structure, chains with fewer than 25 residues are identified as peptide chains, and the remaining chains are defined as protein chains. Backbone atoms (N, C, Cα) of the protein chain within 10 Å of the peptide chain are extracted, similarly to the way in RoseTTAFold All-Atom [17]. The kabsch algorithm is then used to align the crystal and predicted protein structures based on these backbone atoms. The same rotation and translation matrices derived from this alignment are applied to superimpose the peptide chains. The pocket-aligned peptide $\mathrm{RMSD}_{\text{all-atom}}$ quantifies the RMSD of all heavy atoms in the peptide chain, whereas $\mathrm{RMSD}_{\mathrm{unAA}}$ is calculated exclusively for the heavy atoms of unnatural amino acid residues. The pocket-aligned peptide $\mathrm{RMSD}_{C\alpha}$ follows a similar procedure, and the primary difference is that only the backbone atoms of the protein chains within 10 Å of the Cα atoms in the peptide chain are selected for alignment. Additionally, the RMSD calculation is restricted to the Cα atoms in the peptide chains.

$\mathrm{RMSD}_{\text{all-atom}}$ is determined by aligning all heavy atoms in the crystal and predicted structures using the kabsch algorithm and subsequently calculating the RMSD across all heavy atoms. When the calculation is limited to unnatural amino acid residues, the metric is denoted as $\mathrm{RMSD}_{\mathrm{unAA}}$. $\mathrm{RMSD}_{C\alpha}$ is derived by aligning the Cα atoms from the crystal and predicted structures and computing the RMSD between these atoms exclusively.

Structure assessments

The structures were independently evaluated using the online validation server MolProbity [37], which is a structural analysis tool that provides information on macromolecular accuracy by assessing the quality of the structure based on atomic contact analysis, geometry, and backbone torsion angles. This tool is commonly used for validating the overall structural integrity and stereochemistry of macromolecules.

Result and discussion

Building upon the AlphaFold-Multimer framework, we developed a model capable of accurately predicting the 3D structures of peptide monomers and complexes containing unnatural amino acids. During the construction of multiple sequence alignments (MSA), all unnatural amino acids were initially treated as the unknown residue (X). To overcome this limitation, we integrated atom-level embedding information specific to peptides into the AlphaFold-Multimer model, as shown in Fig. 1B. This enhancement allows the model to distinguish among various unnatural amino acid residues while enabling multi-scale modeling of peptide molecules. Since AlphaFold-Multimer models each atom by predicting seven torsion angles for every amino acid along with their predefined rigid groups and atomic coordinates, we extended this methodology to unnatural amino acids. Specifically, we defined corresponding rigid groups and initial atomic coordinates for these residues, as shown in Fig. 1C, enabling the model to predict their atomic coordinates in an end-to-end manner. Next, we fine-tuned the original five AlphaFold-Multimer parameter sets using linear peptides containing unnatural amino acids, as illustrated in Fig. 1A. Furthermore, by modifying the relative position encoding matrix, we expanded this model's capabilities to predict the structures of cyclic peptides with unnatural residues, as shown in Fig. 1D. To enhance the model further, we parameterized the unnatural amino acids, allowing for the relaxation of spatial structures containing these unnatural residues, as illustrated in Fig. 1E. To the best of our knowledge, this work represents the first effort to fine-tune AlphaFold-Multimer specifically for the structure prediction of cyclic peptides with unnatural amino acids, more details about the workflow can be seen in Methods and Materials section.

Accurately predict the structure of cyclic peptides with unnatural amino acids

HighFold2 can accurately predict the 3D structures of cyclic peptide monomers containing unnatural amino acids and their complexes with proteins. Since we trained five distinct models, five predicted structures can be generated during testing. The model automatically ranks these predicted structures according to predefined criteria, and we only select the rank one structure for comparison with the native structure. As shown in Fig. 2A, the model achieves a median $\mathrm{RMSD}_{C\alpha}$ (for complexes, $\mathrm{RMSD}_{C\alpha}$ refers to pocket-aligned peptide $\mathrm{RMSD}_{C\alpha}$) of 1.891 Å on the test set of cyclic peptides with unnatural amino acids. Most samples have $\mathrm{RMSD}_{C\alpha}$ values near 2 Å, with 20 out of 34 test samples having $\mathrm{RMSD}_{C\alpha}$ less than 2 Å. This suggests the model's predictions of cyclic peptide backbones containing unnatural amino acids are highly accurate. The median $\mathrm{RMSD}_{\text{all-atom}}$ (for complexes, $\mathrm{RMSD}_{\text{all-atom}}$ refers to pocket-aligned peptide $\mathrm{RMSD}_{\text{all-atom}}$) is 2.872 Å, with most samples clustering around 3 Å. Six samples had $\mathrm{RMSD}_{\text{all-atom}}$ values under 2 Å, indicating the model's reasonable ability to predict side-chain conformations. Given that side-chain positions are not fixed in real scenarios but can undergo torsional adjustments, the model's performance for $\mathrm{RMSD}_{\text{all-atom}}$ is considered acceptable. To evaluate the model's precision in predicting structures of unnatural amino acids, we computed $\mathrm{RMSD}_{\mathrm{unAA}}$ (for complexes, $\mathrm{RMSD}_{\mathrm{unAA}}$ refers to pocket-aligned peptide $\mathrm{RMSD}_{\mathrm{unAA}}$), which yielded a median value of 2.579 Å.

This indicates that the model accurately predicts most unnatural amino acid residues. Figure 2B shows all RMSD distributions in this test set, and the detailed predictions for each unnatural cyclic peptide sample can be found in Table S4. Additionally, as shown in Table S5 and Table S6, the predictive performance of HighFold2 does not decline regardless of increases in sequence length or the cyclization residues. We also compared HighFold2 with Chai-1 [38] and Protenix [39], both AlphaFold3-like structure prediction models, despite neither explicitly accounting for cyclic peptides. Unsurprisingly, HighFold2 achieved the lowest prediction error, as shown in Table S7.

Like AlphaFold-Multimer, HighFold2 provides a prediction confidence score for each amino acid residue. To investigate the relationship between these confidence scores and $\mathrm{RMSD}_{\mathrm{C}\alpha}$, $\mathrm{RMSD}_{\text{all-atom}}$, and $\mathrm{RMSD}_{\mathrm{unAA}}$, we obtained the predicted local-distance difference test for peptides (peptide pLDDT) and the predicted local-distance difference test for unnatural amino acid residues (unAA pLDDT). For complexes, peptide pLDDT is the average pLDDT score of the peptide ligand, while for peptide monomers, it is the average pLDDT score of the entire structure. As shown in Fig. 2C and Fig. 2D, peptide pLDDT is linearly correlated with both $\mathrm{RMSD}_{\mathrm{C}\alpha}$ and $\mathrm{RMSD}_{\text{all-atom}}$. As peptide pLDDT increases, the RMSD values between predicted and true structures decrease. In our test set of cyclic peptides with unnatural amino acids, the coefficient of determination (R^2) between peptide pLDDT and $\mathrm{RMSD}_{\mathrm{C}\alpha}$ is 0.456, while the R^2 between peptide pLDDT and $\mathrm{RMSD}_{\text{all-atom}}$ is 0.537. This indicates that peptide pLDDT can serve as a useful metric for roughly assessing the accuracy of predictions for cyclic peptides containing unnatural amino acids, both in monomeric form and as peptide ligands. Figure 2E demonstrates that unAA pLDDT is also linearly correlated with $\mathrm{RMSD}_{\mathrm{unAA}}$, with an R^2 value of 0.377. Although the correlation is the weakest among the three scenarios, this still indicates that unAA pLDDT can serve as a valuable indicator for evaluating the prediction accuracy of unnatural amino acids. The details of peptide pLDDT and unAA pLDDT can be seen in Table S8.

We also visualized the model's predicted structures for cyclic peptide monomers containing unnatural amino acids and complexes with this kind of cyclic peptide ligands. Figure 2F shows the predicted structure of mutant human alpha-defensin 1 (PDB ID 3LO6), where the $\mathrm{RMSD}_{\mathrm{C}\alpha}$ is as low as 0.643 Å and the $\mathrm{RMSD}_{\text{all-atom}}$ is 1.254 Å. The predicted position of the unnatural amino acid alpha-aminobutyric acid closely matches that in the crystal structure, achieving an $\mathrm{RMSD}_{\mathrm{unAA}}$ of 0.426 Å. Notably, this structure is cyclized through three pairs of disulfide bonds, which the model also successfully predicted. Figure 2G presents the predicted structure of a complex between HIV-1 integrase and a cyclic peptide containing an unnatural amino acid (PDB ID 3WNG). The model not only accurately predicted the binding site between the cyclic peptide ligand and the protein but also predicted the cyclic peptide's spatial structure. The pocket-aligned peptide $\mathrm{RMSD}_{\mathrm{C}\alpha}$ is 1.166 Å, and the pocket-aligned peptide $\mathrm{RMSD}_{\text{all-atom}}$ is 1.701 Å. This cyclic peptide contains a D-proline residue, and the model's prediction for this unnatural amino acid is also close to the crystal structure's position. Furthermore, the model accurately predicted the head-to-tail cyclization feature of this peptide.

Ablation experiments

We conducted a series of ablation studies to elucidate the contributions of individual modules in HighFold2 to the accuracy of predicting cyclic peptide structures with unnatural residues. We systematically removed the ensemble, modifications of the relative position encoding matrix (MRPE), and atomic-scale feature extraction (AFE) from the original model. Then, we evaluated its performance on the test set of cyclic peptides. The results, shown in Fig. 3A, indicate that the full model achieved the lowest average $\mathrm{RMSD}_{\mathrm{C}\alpha}$ (2.152 Å), average $\mathrm{RMSD}_{\text{all-atom}}$ (3.436 Å), and average $\mathrm{RMSD}_{\mathrm{unAA}}$ (2.965 Å), underscoring the indispensability of these three components. Detailed ablation results for each test sample are provided in Tables S9, S10, and S11.

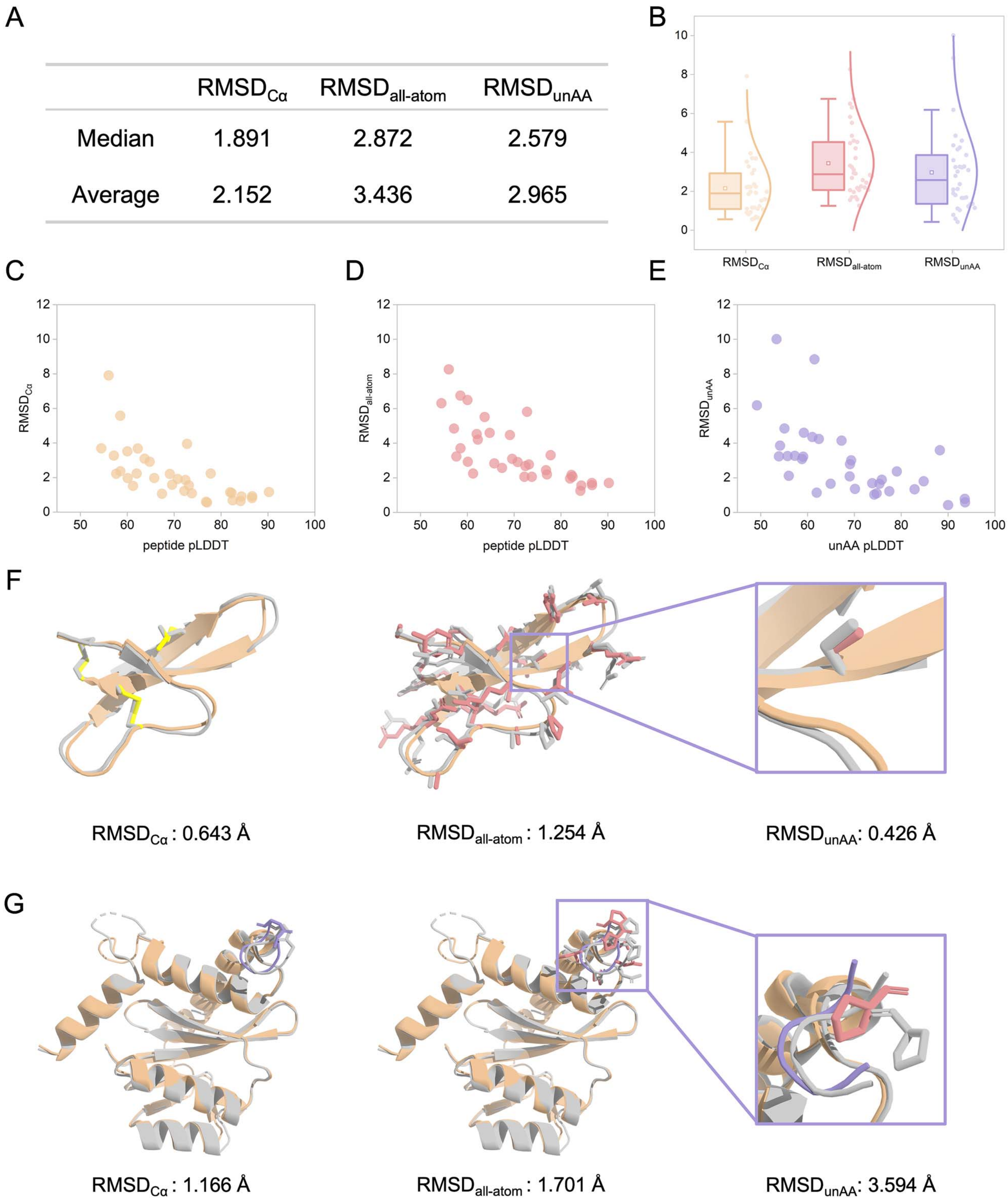
During testing on cyclic peptides, the model generates five predicted structures for each sample and ranks them using AlphaFold-Multimer's original criteria. When predictions relied solely on the first parameter set, prediction accuracy diminished, with the average $\mathrm{RMSD}_{\mathrm{C}\alpha}$ increasing by 0.266 Å, the average $\mathrm{RMSD}_{\text{all-atom}}$ by 0.164 Å, and the average $\mathrm{RMSD}_{\mathrm{unAA}}$ by 0.258 Å, as illustrated in Fig. 3B. This highlights the effectiveness of the ensemble approach. Reliance on a single parameter set tends to significant structural errors for certain samples, such as the β-sheet structure in 2M1P, which was mispredicted, as visualized in Fig. 3C. Table S12 also shows the prediction result for other fine-tuning parameter sets. Additionally, as shown in Table S13, fine-tuning does not impact HighFold2's predictive performance for cyclic peptides composed solely of natural amino acids.

Building on the model without the ensemble module, we further removed modifications to the relative position encoding matrix. By modifying the relative position encoding matrix, we ensured the precise representation of the relative positions of cyclized residues, allowing accurate modeling of their connectivity. Without this adjustment, prediction performance deteriorated markedly. As shown in Fig. 3D, many cyclic peptide samples were erroneously predicted as linear peptides, leading to significant increases in all RMSD. The largest increase for $\mathrm{RMSD}_{\mathrm{unAA}}$, 8.594 Å, occurred for 1T9E, as shown in Fig. 3E, where the model failed to predict its head-to-tail cyclization.

Then, we evaluated the removal of the atomic-scale feature extraction module, which integrates atomic-level peptide features with AlphaFold-Multimer's residue-level representation to enable multi-scale peptide modeling. Compared to predictions made using only the first fine-tuned parameter set, all RMSD metrics showed a slight increase, as illustrated in Fig. 3F, indicating that the atomic-scale feature extraction module contributes to improving the structural predictions of cyclic peptides. Figure 3G visualizes the comparison between the predicted structure from this ablation model and the native structure (PDB ID 2MSQ). The overall differences are substantial, with significant deviations in the predicted positions of the disulfide bonds.

Accurately predict the structure of linear peptides with unnatural amino acids

We also evaluated the model's predicted performance on the test set of linear peptides containing unnatural amino acids. As shown in Fig. 4A, HighFold2 achieved a median $\mathrm{RMSD}_{\mathrm{C}\alpha}$ of 0.994 Å, a median $\mathrm{RMSD}_{\text{all-atom}}$ of 1.906 Å, and a median $\mathrm{RMSD}_{\mathrm{unAA}}$ of 1.971 Å on the linear peptide test set. These results confirm the model's strong ability to predict the 3D structures of linear peptide monomers and their complexes involving unnatural amino acids. Detailed prediction results for each test sample are available in Table S14. Compared to its performance on cyclic peptides, the model performed better on linear peptides. This improved performance likely arises from the training data being based on linear peptide structures. We hypothesize that as more cyclic peptide structures are resolved in the PDB database, incorporating such data into the training process will further enhance the



	$RMSD_{C\alpha}$	$RMSD_{all\text{-}atom}$	$RMSD_{unAA}$
Median	1.891	2.872	2.579
Average	2.152	3.436	2.965

Figure 2. The performance of predicting the structure of cyclic peptide monomers and complexes containing unnatural amino acids. (A) The overall accuracy of the model in the independent cyclic peptide test set. (B) The distributions of all RMSD for cyclic peptide samples. (C) The correlation between peptide pLDDT and $RMSD_{C\alpha}$ in the independent cyclic peptide test set. (D) The correlation between peptide pLDDT and $RMSD_{all\text{-}atom}$ in the independent cyclic peptide test set. (E) The correlation between unAA pLDDT and $RMSD_{unAA}$ in the independent cyclic peptide test set. (F) An example of the prediction for the cyclic peptide monomer containing the unnatural amino acid. (G) An example of the prediction for the cyclic peptide complex containing the unnatural amino acid.

model's predictive accuracy for cyclic peptides with unnatural amino acids.

In the linear peptide test set, the peptide pLDDT also exhibited a linear relationship with both $RMSD_{C\alpha}$ and $RMSD_{all\text{-}atom}$. As shown in Fig. 4C and Fig. 4D, the R^2 between peptide pLDDT and $RMSD_{C\alpha}$ is 0.372, while the R^2 between peptide pLDDT and $RMSD_{all\text{-}atom}$ is 0.397. Notably, the linear relationship between peptide pLDDT and $RMSD_{all\text{-}atom}$ is stronger than that with $RMSD_{C\alpha}$, A

Metrics	Full model	w/o ensemble	w/o MRPE	w/o AFE
Average $RMSD_{C\alpha}$	**2.152**	2.418	3.413	2.478
Average $RMSD_{all\text{-}atom}$	**3.436**	3.600	4.417	3.653
Average $RMSD_{unAA}$	**2.965**	3.223	3.912	3.370

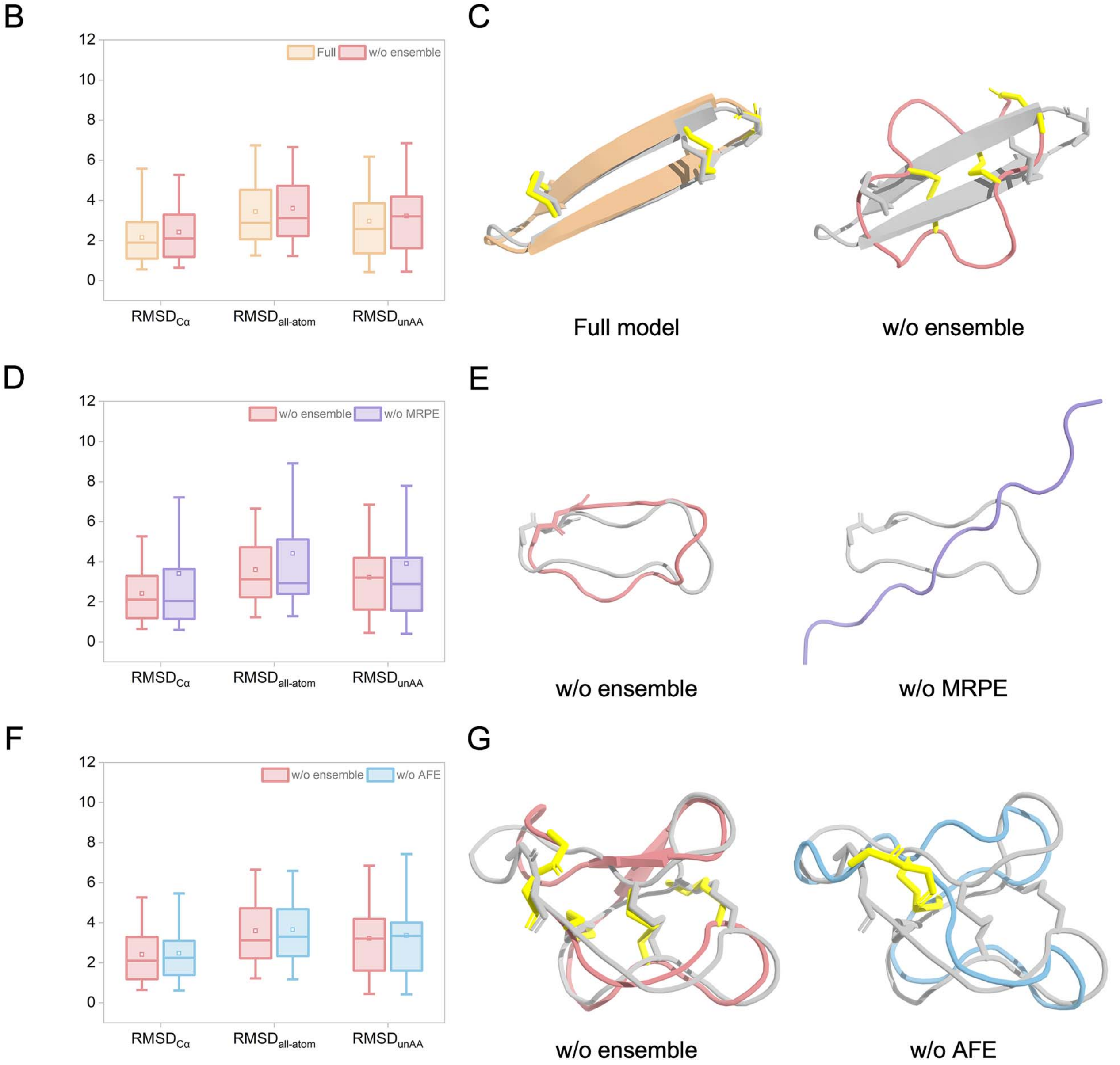
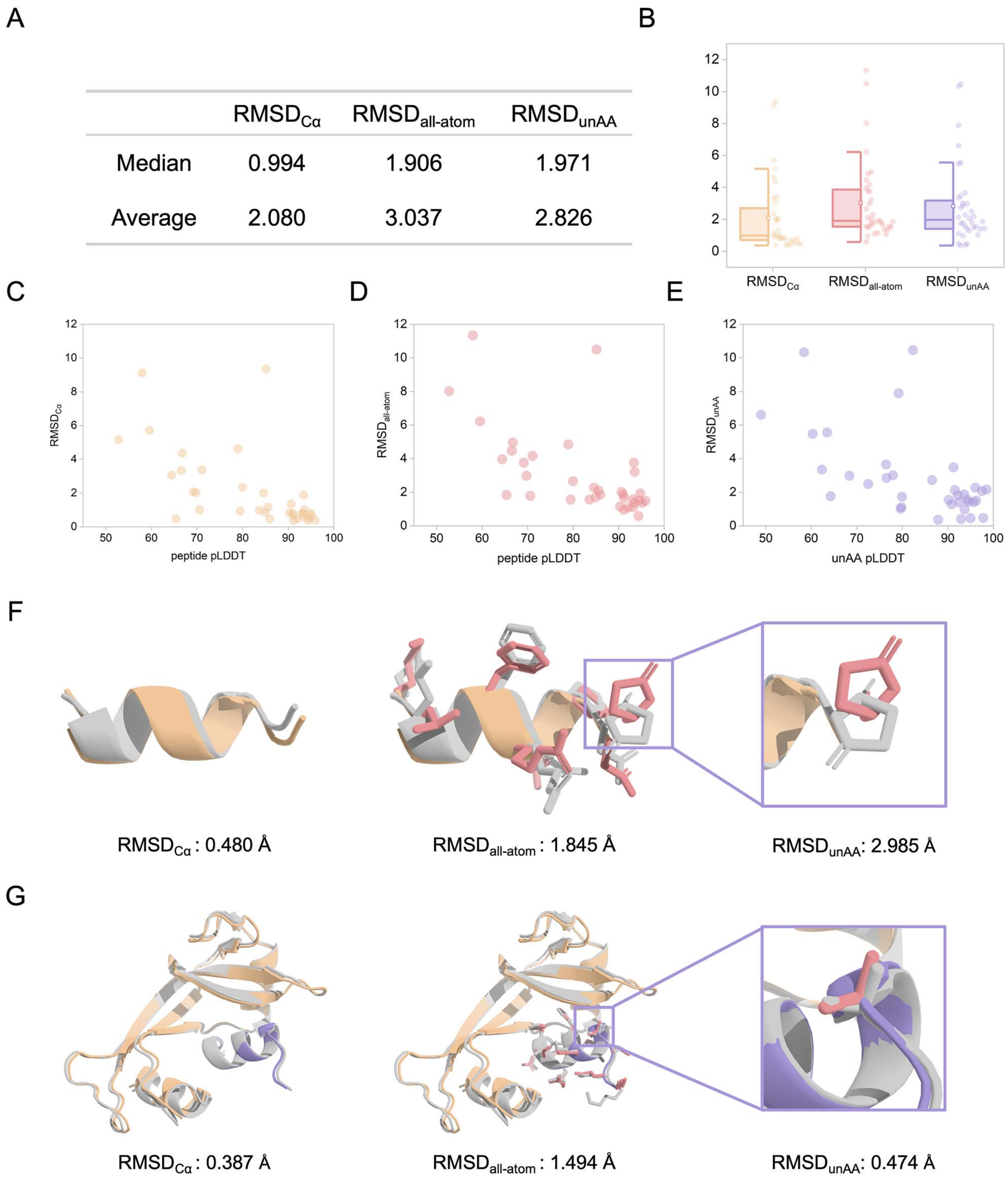


Figure 3. The analysis of the ablation experiments. (A) Predictive performance comparison between the full model and its variants on the independent cyclic peptide test set. w/o stands for without. MRPE and AFE represent the modified relative position encoding and atomic-scale feature extraction, respectively. The best score shows bold. (B) The comparison of distributions between the full model and the model without the ensemble. (C) A comparison of the prediction for the cyclic peptide containing the unnatural amino acid between the full model and the model without the ensemble. (D) The comparison of distributions between the model without the ensemble and the model without the modified relative position encoding. (E) A comparison of the prediction for the cyclic peptide containing the unnatural amino acid between the model without the ensemble and the model without the modified relative position encoding. (F) The comparison of distributions between the model without the ensemble and the model without the atomic-scale feature extraction. (G) A comparison of the prediction for the cyclic peptide containing the unnatural amino acid between the model without the ensemble and the model without the atomic-scale feature extraction.



	$RMSD_{C\alpha}$	$RMSD_{all\text{-}atom}$	$RMSD_{unAA}$
Median	0.994	1.906	1.971
Average	2.080	3.037	2.826

Figure 4. The performance of predicting the structure of linear peptide monomers and complexes containing unnatural amino acids. (A) The overall accuracy of the model in the linear peptide test set. (B) The distributions of all RMSD for linear peptide samples. (C) The correlation between peptide pLDDT and $RMSD_{C\alpha}$ in the linear peptide test set. (D) The correlation between peptide pLDDT and $RMSD_{all\text{-}atom}$ in the linear peptide test set. (E) The correlation between unAA pLDDT and $RMSD_{unAA}$ in the linear peptide test set. (F) An example of the prediction for the linear peptide monomer containing the unnatural amino acid. (G) An example of the prediction for the linear peptide complex containing the unnatural amino acid.

consistent with observations in the cyclic peptide test set. This suggests that peptide pLDDT reflects not only the accuracy of the predicted backbone but also the side chain accuracy. The R^2 between unAA pLDDT and $RMSD_{unAA}$ is 0.351, also showing the weakest correlation like in the cyclic peptide dataset. Table S15 shows the detailed peptide pLDDT and unAA pLDDT.

Figure 4F and Fig. 4G visualize the predicted structures of a linear peptide monomer and a peptide-protein complex, respectively. Figure 4F shows the predicted structure of the NK1 agonist Phyllomedusin, which includes a pyroglutamic acid residue (PDB ID 2NOR). The predicted structure achieved the $RMSD_{C\alpha}$ of 0.480 Å, indicating high accuracy in backbone prediction. However, the

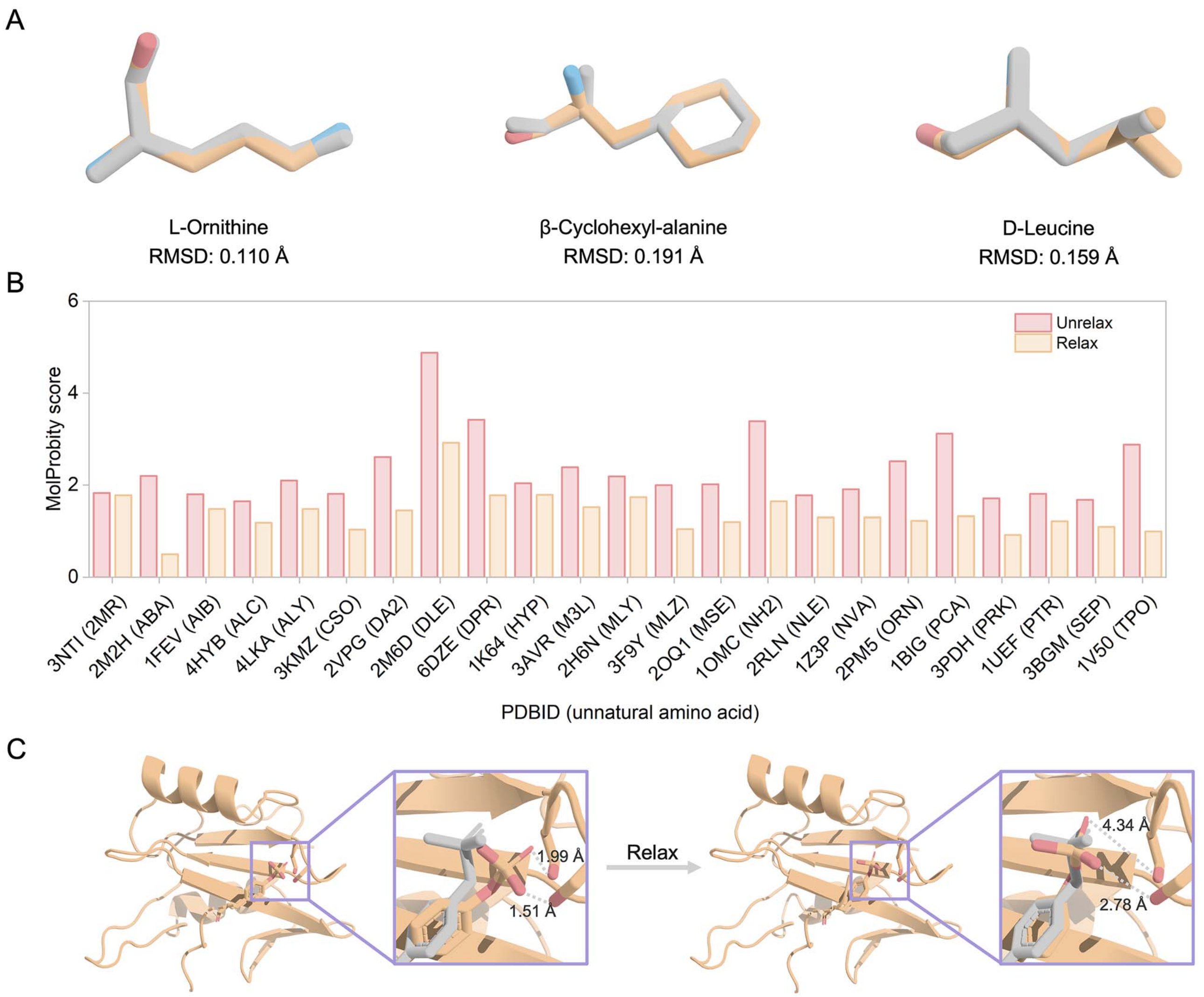


Figure 5. The relaxation for the predicted structure. (A) Structural alignment between the unnatural amino acid after QM (B3LYP/6-31G*) optimization and crystal structure. (B) Quality improvement in structures containing unnatural amino acids after relaxation. MolProbity integrates multiple evaluation criteria, with lower scores indicating better structure quality. (C) The comparison between the structures before and after relaxation for PDB ID 2CI9 containing O-phosphotyrosine.

$RMSD_{all\text{-}atom}$ is 1.845 Å, reflecting some inaccuracies in the side chain predictions. Figure 4G illustrates the predicted structure of a peptide-protein complex containing an alpha-aminoisobutyric acid residue (PDB ID 1FEV). The predicted structure had the $RMSD_{C\alpha}$ of 0.387 Å and the $RMSD_{unAA}$ of 0.474 Å, indicating not only accurate backbone prediction but also accurate predictions for the unnatural residue.

Relaxation of structure with unnatural residues

In this work, based on the efficient relaxation steps from AlphaFold-Multimer, we introduce an additional set of force field parameters for unnatural amino acids, which are obtained through ab initio calculations and compatible with the Amber force field. These parameters aim to improve the accuracy of the local geometry of the predicted models. The improvements include optimizing protein folding accuracy, reducing clashes, and enhancing the stereochemical correctness of protein residues.

Force field parameters for unnatural amino acids are essential for the relaxation process. When custom parameters are not available, the reliability of relaxation is limited. While general force fields like GAFF or CGenFF [40] can assign parameters to various compounds, their lack of specific optimization for unnatural amino acids often results in inaccurate structural predictions. Figure 5A shows that after 6-31G* structural optimization, the unnatural amino acids are highly consistent with the crystal structures, with an average RMSD of 0.153 Å, significantly improving the correlation between the unnatural amino acid conformations and the real structures.

Subsequently, we generated topology and coordinate files for the target structure in tleap. By modifying the OpenMM input files, we overcame the limitation of tleap's inability to output multichain systems, enabling the automated relaxation process for more complex systems. As shown in Fig. 5B, after relaxation, the quality of the structures for all unnatural residues improved significantly compared to the results from the structural prediction models. Table S16 shows that there were minimal or no spatial clashes, unfavorable rotamers, or abnormal Ramachandran values, with most residues in favorable regions in the independent cyclic peptide test set. In Fig. 5C, due to the lack of consideration for reasonable atomic contacts, some non-bonded atomic pairs in the predicted structure were found to be too close. These errors were corrected through an energy minimization procedure without significantly increasing computational costs. Even though the prediction model provided a high-quality overall structure, some side-chain deviations and unreasonable geometry were still observed, suggesting further molecular dynamics simulations or quantum mechanics/molecular mechanics (QM/MM) calculations to improve residue stereochemistry.

To further evaluate the optimization effect, we selected the covalent bonds between the natural amino acid portion and the modified group in unnatural amino acids as an evaluation criterion. Since bond lengths in the predicted models are based on constants, they cannot be adjusted according to environmental changes. Table S17 shows that after relaxation, the bond lengths of the unnatural amino acids are closer to those in the crystal structures [41]. Finally, we analyzed the impact of the relaxation process on all RMSD values. As shown in Table S18 and Table S19, the $\text{RMSD}_{C\alpha}$, $\text{RMSD}_{\text{all-atom}}$, and $\text{RMSD}_{\text{unAA}}$ values remain little changed for the cyclic and linear peptide test sets. Although side-chain adjustments introduced some local changes, they did not negatively affect the overall accuracy of the structural predictions. Moreover, the incorporation of relaxation did not impose a significant computational burden. The prediction and relaxation times of HighFold2 for the cyclic and linear peptide test sets are presented in Table S20.

We corrected unreasonable regions in the predicted structures through relaxation, significantly improving their geometric quality and accuracy. This process reduced the need for manual intervention and enhanced the efficiency and precision of handling systems containing unnatural amino acids.

Conclusion

In this work, we developed a method called HighFold2, based on AlphaFold-Multimer, capable of accurately predicting the 3D structures of cyclic peptide monomers containing unnatural amino acids and their complexes. To adapt AlphaFold-Multimer for the prediction of structures involving unnatural amino acids, we added a neural network module to characterize the atom-level properties of peptides. This not only enabled the distinction between different unnatural amino acids but also facilitated the multi-scale modeling of peptide molecules. We also extended the predefined rigid groups and initial atomic coordinate information for natural amino acids in AlphaFold-Multimer to encompass unnatural amino acids, successfully enabling the prediction of their structures. Given the scarcity of cyclic peptide structural data, we employed a zero-shot learning strategy, initially training the model using linear peptide data containing unnatural amino acids. Then, by modifying the relative position encoding matrix in the model, we achieved accurate predictions of cyclic peptides containing unnatural amino acids. Furthermore, by combining Gaussian, AmberTools, and OpenMM, we developed an efficient relaxation workflow, significantly reducing spatial clashes and improving geometric accuracy. Further testing demonstrated that HighFold2 showed outstanding prediction performance on both independent cyclic peptide and linear peptide test sets, and various ablation experiments validated the effectiveness of its modifications.

While HighFold2 has made notable progress, it also has certain limitations. First, since our cyclic peptide predictions rely on modifications to the relative position encoding matrix, this approach merely ensures spatial proximity among cyclization residues without offering precise control over the cyclization strategy. Consequently, the current version of HighFold2 supports only simple cyclization modes, such as head-to-tail cyclization and disulfide bond formation. We hypothesize that incorporating this relative position encoding modification strategy into AlphaFold3 could enable the prediction of more diverse cyclization patterns. Additionally, HighFold2 is currently incapable of predicting peptides containing unnatural backbones, as it is based on α-amino acid. This restriction arises from the fact that AlphaFold-Multimer predicts only three backbone torsion angles, making it inadequate for handling structures with additional torsional degrees of freedom. Introducing more backbone torsion angles and retraining the model could potentially extend its applicability to a broader range of unnatural backbone configurations.

Key Points

- We proposed a novel model, HighFold2, based on AlphaFold-Multimer, for accurately predicting the 3D structures of cyclic peptide monomers containing unnatural amino acids and their complexes.
- By incorporating additional atom-level information for peptide molecules in AlphaFold-Multimer and expanding the predefined rigid groups and initial atomic coordinates, we enabled HighFold2 to predict the 3D structures of unnatural amino acids.
- Due to the scarcity of cyclic peptide structures, we first fine-tuned AlphaFold-Multimer using linear peptide structures with unnatural amino acids. Then, by modifying the relative position encoding matrix, we enabled HighFold2 to be applicable for the prediction of cyclic peptide structures containing unnatural residues.
- We also parameterized the unnatural amino acids, allowing HighFold2 to perform energy minimization to eliminate spatial clashes in the predicted structures.

Acknowledgements

This work was supported by the Natural Science Foundation of Zhejiang Province (LD22H300004) and the Science and Technology Development Fund of Macau (no. 0177/2023/RIA3).

Author contributions

Cheng Zhu: Conceptualization; Data curation; Methodology; Software; Writing—original draft; Writing—review and editing. Sen Cao: Methodology; Writing— original draft; Writing—review and editing. Tianfeng Shang: Supervision; Writing—review and editing. Hongliang Duan: Supervision; Writing—review and editing.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

None declared.

Data availability

All data this work uses is available at https://github.com/hongliangduan/HighFold2 or the Supporting Information.

Code availability

The code of this work is available at https://github.com/hongliangduan/HighFold2.

References

1. Rodnina MV, Wintermeyer W. Peptide bond formation on the ribosome: Structure and mechanism. *Curr Opin Struct Biol* 2003;**13**:334–40. https://doi.org/10.1016/S0959-440X(03)00065-4
2. Lu H, Zhou Q, He J. *et al.* Recent advances in the development of protein–protein interactions modulators: Mechanisms and clinical trials. *Sig Transduct Target Ther* 2020;**5**:1–23. https://doi.org/10.1038/s41392-020-00315-3
3. Mannes M, Martin C, Menet C. *et al.* Wandering beyond small molecules: Peptides as allosteric protein modulators. *Trends Pharmacol Sci* 2022;**43**:406–23. https://doi.org/10.1016/j.tips.2021.10.011
4. Wang L, Wang N, Zhang W. *et al.* Therapeutic peptides: Current applications and future directions. *Sig Transduct Target Ther* 2022;**7**:1–27. https://doi.org/10.1038/s41392-022-00904-4
5. Fetse J, Kandel S, Mamani U-F. *et al.* Recent advances in the development of therapeutic peptides. *Trends Pharmacol Sci* 2023;**44**:425–41. https://doi.org/10.1016/j.tips.2023.04.003
6. Xu S, Tan P, Tang Q. *et al.* Enhancing the stability of antimicrobial peptides: From design strategies to applications. *Chem Eng J* 2023;**475**:145923. https://doi.org/10.1016/j.cej.2023.145923
7. Dougherty PG, Sahni A, Pei D. Understanding cell penetration of cyclic peptides. *Chem Rev* 2019;**119**:10241–87. https://doi.org/10.1021/acs.chemrev.9b00008
8. Ji X, Nielsen AL, Heinis C. Cyclic peptides for drug development. *Angew Chem Int Ed* 2024;**63**:e202308251. https://doi.org/10.1002/anie.202308251
9. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
10. Evans R, O'Neill M, Pritzel A. *et al. Protein Complex Prediction with AlphaFold-Multimer*. BioRxiv 2021. https://doi.org/10.1101/2021.10.04.463034.
11. Baek M, DiMaio F, Anishchenko I. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6. https://doi.org/10.1126/science.abj8754
12. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
13. Baek M, McHugh R, Anishchenko I. *et al.* Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat Methods* 2024;**21**:117–21. https://doi.org/10.1038/s41592-023-02086-5
14. Tsaban T, Varga JK, Avraham O. *et al.* Harnessing protein folding neural networks for peptide–protein docking. *Nat Commun* 2022;**13**:176. https://doi.org/10.1038/s41467-021-27838-9
15. Bryant P, Kelkar A, Guljas A. *et al.* Structure prediction of protein-ligand complexes from sequence information with Umol. *Nat Commun* 2024;**15**:4536. https://doi.org/10.1038/s41467-024-48837-6
16. Zhang C, Zhang C, Shang T. *et al.* HighFold: Accurately predicting structures of cyclic peptides and complexes with head-to-tail and disulfide bridge constraints. *Brief Bioinform* 2024;**25**:bbae215. https://doi.org/10.1093/bib/bbae215
17. Krishna R, Wang J, Ahern W. *et al.* Generalized biomolecular modeling and design with RoseTTAFold all-atom. *Science* 2024;**384**:eadl2528. https://doi.org/10.1126/science.adl2528
18. Abramson J, Adler J, Dunger J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500. https://doi.org/10.1038/s41586-024-07487-w
19. Berman HM, Westbrook J, Feng Z. *et al.* The protein data Bank. *Nucleic Acids Res* 2000;**28**:235–42. https://doi.org/10.1093/nar/28.1.235
20. Burley SK, Bhikadiya C, Bi C. *et al.* RCSB protein data Bank (RCSB.org): Delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* 2023;**51**:D488–508. https://doi.org/10.1093/nar/gkac1077
21. Singh S, Singh H, Tuknait A. *et al.* PEPstrMOD: Structure prediction of peptides containing natural, non-natural and modified residues. *Biol Direct* 2015;**10**:73. https://doi.org/10.1186/s13062-015-0103-4
22. Santini BL, Wendel S, Halbwedl N. *et al.* cPEPmatch webserver: A comprehensive tool and database to aid rational design of cyclic peptides for drug discovery. *Comput Struct Biotechnol J* 2024;**23**:3155–62. https://doi.org/10.1016/j.csbj.2024.08.008
23. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science* 1985;**227**:1435–41. https://doi.org/10.1126/science.2983426
24. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6. https://doi.org/10.1021/ci00057a005
25. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. New York, United States: Curran Associates Inc., 2017;6000–10. https://dl.acm.org/doi/10.5555/3295222.3295349.
26. Murtagh F. Multilayer perceptrons for classification and regression. *Neurocomputing* 1991;**2**:183–97. https://doi.org/10.1016/0925-2312(91)90023-5
27. Krapp LF, Abriata LA, Cortés Rodriguez F. *et al.* PeSTo: Parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun* 2023;**14**:2175. https://doi.org/10.1038/s41467-023-37701-8
28. Mirdita M, Schütze K, Moriwaki Y. *et al.* ColabFold: Making protein folding accessible to all. *Nat Methods* 2022;**19**:679–82. https://doi.org/10.1038/s41592-022-01488-1
29. Ponder JW, Case DA. Force fields for protein simulations. *Adv Protein Chem* 2003;**66**:27–85. https://doi.org/10.1016/S0065-3233(03)66002-X
30. Kohn W, Becke AD, Parr RG. Density functional theory of electronic structure. *J Phys Chem* 1996;**100**:12974–80. https://doi.org/10.1021/jp960669l
31. Schauperl M, Nerenberg PS, Jang H. *et al.* Non-bonded force field model with advanced restrained electrostatic potential charges (RESP2). *Commun Chem* 2020;**3**:1–11. https://doi.org/10.1038/s42004-020-0291-4
32. Maier JA, Martinez C, Kasavajhala K. *et al.* ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* 2015;**11**:3696–713. https://doi.org/10.1021/acs.jctc.5b00255
33. Wang J, Wolf RM, Caldwell JW. *et al.* Development and testing of a general amber force field. *J Comput Chem* 2004;**25**:1157–74. https://doi.org/10.1002/jcc.20035

34. Case DA, Aktulga HM, Belfon K. *et al.* AmberTools. *J Chem Inf Model* 2023;**63**:6183–91. https://doi.org/10.1021/acs.jcim.3c01153
35. Shirts MR, Klein C, Swails JM. *et al.* Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *J Comput Aided Mol Des* 2017;**31**:147–61. https://doi.org/10.1007/s10822-016-9977-1
36. Eastman P, Galvelis R, Peláez RP. *et al.* OpenMM 8: Molecular dynamics simulation with machine learning potentials. *J Phys Chem B* 2024;**128**:109–16. https://doi.org/10.1021/acs.jpcb.3c06662
37. Davis IW, Leaver-Fay A, Chen VB. *et al.* MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 2007;**35**:W375–83. https://doi.org/10.1093/nar/gkm216
38. Discovery C, Boitreaud J, Dent J. *et al.* Chai-1: Decoding the molecular interactions of life. BioRxiv 2024. https://doi.org/10.1101/2024.10.10.615955.
39. Team BAA, Chen X, Zhang Y. *et al.* Protenix - advancing structure prediction through a comprehensive AlphaFold3 reproduction. BioRxiv 2025. https://doi.org/10.1101/2025.01.08.631967.
40. Vanommeslaeghe K, Hatcher E, Acharya C. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* 2010;**31**:671–90. https://doi.org/10.1002/jcc.21367
41. Qiu Y, Smith DGA, Boothroyd S. *et al.* Development and benchmarking of open force field v1.0.0—The parsley small-molecule force field. *J Chem Theory Comput* 2021;**17**:6262–80. https://doi.org/10.1021/acs.jctc.1c00571

Briefings in Bioinformatics, 2025, **26(3)**, bbaf202
https://doi.org/10.1093/bib/bbaf202
Problem Solving Protocol