PNAS RESEARCH ARTICLE | CHEMISTRY

AI protocol for retrieving protein dynamic structures from two-dimensional infrared spectra

Sheng Ye[a,1], Lvshuai Zhu[a,1], Zhicheng Zhao[a,1], Fan Wu[b], Zhipeng Li[a], BinBin Wang[a], Kai Zhong[c,2], Changyin Sun[a,2], Shaul Mukamel[d,2], and Jun Jiang[b,2]

Affiliations are included on p. 7.

Contributed by Shaul Mukamel; received November 25, 2024; accepted January 16, 2025; reviewed by Martin T. Zanni and Jin Wang

Understanding the dynamic evolution of protein structures is crucial for uncovering their biological functions. Yet, real-time prediction of these dynamic structures remains a significant challenge. Two-dimensional infrared (2DIR) spectroscopy is a powerful tool for analyzing protein dynamics. However, translating its complex, low-dimensional signals into detailed three-dimensional structures is a daunting task. In this study, we introduce a machine learning-based approach that accurately predicts dynamic three-dimensional protein structures from 2DIR descriptors. Our method establishes a robust "spectrum-structure" relationship, enabling the recovery of three-dimensional structures across a wide variety of proteins. It demonstrates broad applicability in predicting dynamic structures along different protein folding trajectories, spanning timescales from microseconds to milliseconds. This approach also shows promise in identifying the structures of previously uncharacterized proteins based solely on their spectral descriptors. The integration of AI with 2DIR spectroscopy offers insights and represents a significant advancement in the real-time analysis of dynamic protein structures.

protein dynamics | machine learning | spectrum-structure relationship

Understanding protein function requires insights into the dynamic evolution of their atomic structures (1), which has led to significant efforts in developing tools for structure determination (2–7). Advances in AI have revolutionized the prediction of a protein's fully folded three-dimensional structure from its primary amino acid sequence, with models like AlphaFold and RoseTTAFold significantly enhancing our understanding of static protein structures (8–14). However, the intermediate states that control proteins' dynamic behavior are less explored (15). This hinders our comprehension of critical processes such as transmembrane transport, ligand binding, conformational changes, and protein folding. Real-time monitoring of protein structures and dynamics is therefore essential for a complete understanding of their function.

Spectroscopic techniques have long been used to monitor protein dynamics, as they offer detailed temporal and spatial insights into structural changes (16–20). Among these, two-dimensional infrared (2DIR) spectroscopy stands out for its high spectral resolution and ability to capture nanometer-scale conformational changes on picosecond to nanosecond timescales (21–25). Compared to one-dimensional infrared spectroscopy, which produces congested spectra for larger molecules, 2DIR utilizes an additional frequency dimension to achieve higher resolution. This added clarity allows for the detection of interactions between vibrational modes, making 2DIR especially effective for analyzing complex systems and accounting for environmental effects (23, 25). Furthermore, 2DIR data can be generated through both theoretical calculations and experimental measurements, bridging the gap between theory and experiment.

Despite its advantages, interpreting complex 2DIR signals and correlating them with specific protein structural characteristics remains a challenging task (23, 25). While cross peaks in 2DIR spectra provide valuable information about atomic distances between atoms such as carbon, these data alone do not provide a complete picture of the overall protein structure. A similar challenge exists with NMR, which provides proton–proton distances but cannot uniquely resolve full structures from these data alone (26, 27). NMR requires hundreds of experimentally derived constraints, including chemical shifts, coupling constants, and NOE distances, to construct accurate three-dimensional models by revealing atomic environments and spatial relationships (27–29). In contrast, 2DIR spectroscopy captures structural information across the entire spectral matrix, where each pixel reflects interactions between vibrational modes, thus providing a detailed map

Significance

Dynamic protein structures are essential for understanding their diverse biological functions. Two-dimensional infrared (2DIR) spectroscopy is a powerful technique for monitoring protein dynamics, yet converting spectroscopic signals into detailed 3D structures remains a challenge. Here, we introduce a machine learning–based method that predicts 3D protein structure dynamics from 2DIR descriptors. By establishing a precise "spectrum–structure" relationship, our approach effectively predicts structures across a range of proteins, capturing folding trajectories and identifying structures of previously uncharacterized proteins. The integration of AI with 2DIR spectroscopy marks a significant advancement in real-time analysis of dynamic protein structures.

Author contributions: S.Y., L.Z., and Z.Z. designed research; S.Y., L.Z., F.W., Z.L., B.W., and K.Z. performed research; S.Y., L.Z., and Z.Z. analyzed data; S.Y., L.Z., Z.Z., F.W., Z.L., B.W., and K.Z. discussed the results and commented on the manuscript; C.S., S.M., and J.J. directed the project discussed the results and commented on the manuscript; and S.Y., C.S., S.M., and J.J. wrote the paper.

Reviewers: M.T.Z., University of Wisconsin-Madison; and J.W., Stony Brook University.

The authors declare no competing interest.

[1] S.Y. L.Z and Z.Z. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: k.zhong@rug.nl, cysun@ahu.edu.cn, smukamel@uci.edu, or jiangj1@ustc.edu.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2424078122/-/DCSupplemental.

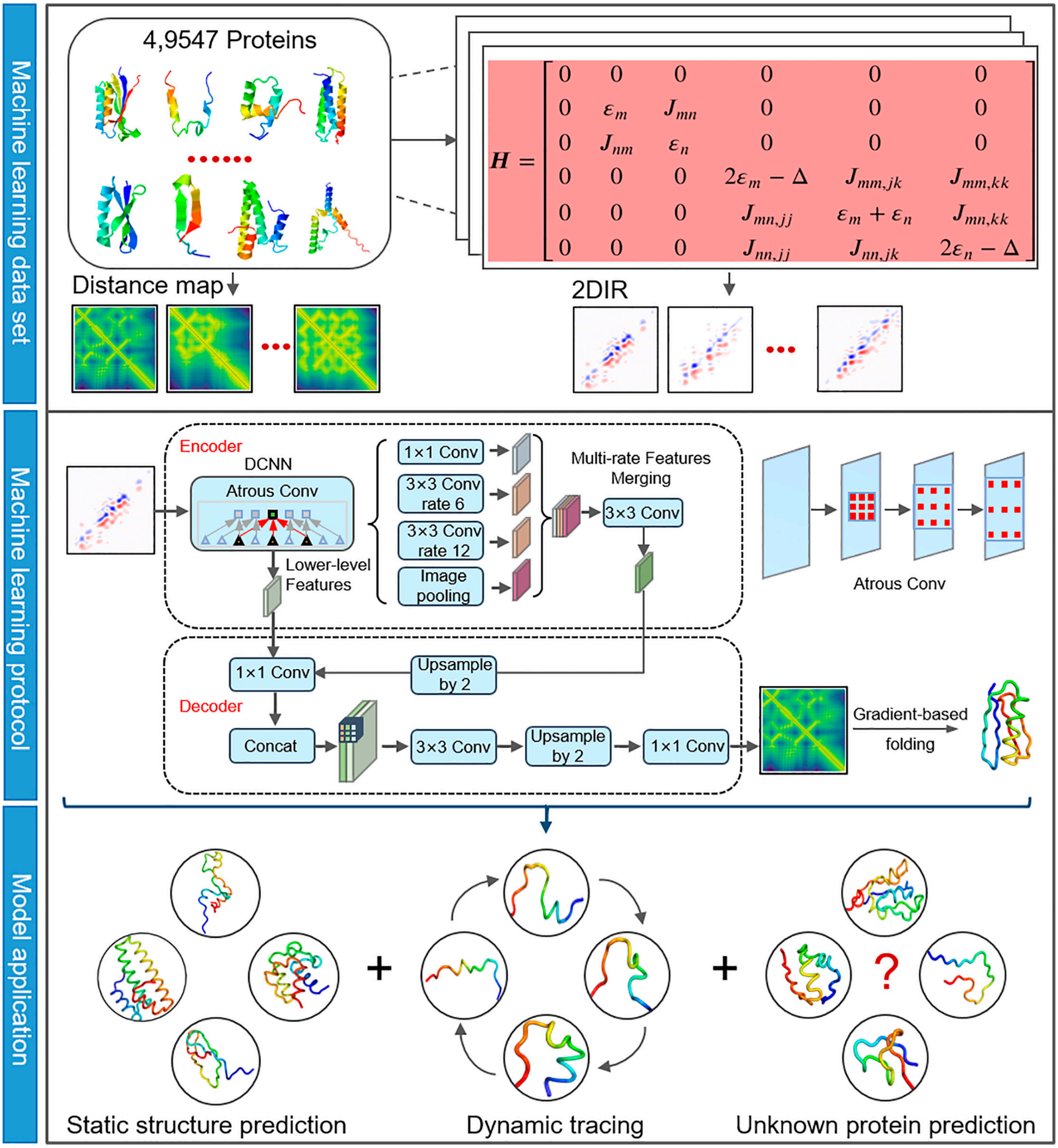
Published February 14, 2025.

of vibrational couplings (23). However, due to the complexity of the data, extracting clear structural insights from 2D IR spectra remains difficult. AI plays a critical role in addressing these limitations by analyzing incomplete datasets, uncovering hidden correlations, and serving as a powerful tool for modeling complex structure–pr operty relationships (30).

In this study, we present a machine learning (ML)–based protocol that leverages 2DIR spectral descriptors to accurately predict three-dimensional protein backbone structures. This approach not only predicts the real-time structural evolution of proteins on microsecond to millisecond timescales but also holds great potential for predicting the structures of previously uncharacterized proteins. Integrating AI with 2DIR spectroscopy opens possibilities for real-time detection of dynamic protein structures and the characterization of unknown proteins.

Results and Discussion

ML Protocol Workflow. The workflow of our ML-based protocol for predicting dynamic protein structures using 2DIR spectroscopy is divided into three components: “ML Dataset,” “ML Protocol,” and “Model Application.” (Fig. 1). We first collected 49,547 protein structures, each containing up to 100 residues (*SI Appendix*, Fig. S1), from the RCSB Protein Data Bank and SWISS-PROT library (31, 32). Due to the rarity of experimental 2DIR spectral data, we employed theoretical simulations to create a foundational ML database. The 2DIR signals were generated using the Frenkel exciton Hamiltonian, created for each protein conformation within the amide I spectral window, based on vibrational spectroscopic maps that align well with experimental findings (19, 33, 34). We subsequently calculated protein alpha carbon ($C\alpha$) distance maps, where each matrix element corresponds to the distance between



Machine learning data set
4,9547 Proteins
$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \varepsilon_m & J_{mn} & 0 & 0 & 0 \\ 0 & J_{nm} & \varepsilon_n & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\varepsilon_m - \Delta & J_{mm,jk} & J_{mm,kk} \\ 0 & 0 & 0 & J_{mn,jj} & \varepsilon_m + \varepsilon_n & J_{mn,kk} \\ 0 & 0 & 0 & J_{nn,jj} & J_{nn,jk} & 2\varepsilon_n - \Delta \end{bmatrix}$
Distance map
2DIR
Machine learning protocol
Encoder
DCNN
Atrous Conv
Lower-level Features
1×1 Conv
3×3 Conv rate 6
3×3 Conv rate 12
Image pooling
Multi-rate Features Merging
3×3 Conv
Atrous Conv
Decoder
1×1 Conv
Upsample by 2
Concat
3×3 Conv
Upsample by 2
1×1 Conv
Gradient-based folding
Model application
Static structure prediction
Dynamic tracing
Unknown protein prediction
?

Fig. 1. ML protocol for predicting protein structures with 2DIR descriptors. The experimental sequence flows from *Top* to *Bottom*, encompassing ML Dataset, ML Protocol, and Model Application.

the $C\alpha$ atoms of amino acids in the protein structure, which serve as initial predictions for the ML model (9, 10).

The ML architecture was then designed using the DeepLabV3 (35, 36) model and includes three key components: feature extraction, spatial dimension restoration through upsampling convolutional layers, and a final regression output. The DeepLabV3 model first extracts features from 2DIR images within the 1,575 to 1,725 cm^{-1} spectral window, capturing both coherent and detection frequencies. The 2DIR signals are converted into 3 × 224 × 224 RGB images, from which high-level features (2,048 × 28× 28) are extracted. The process uses atrous convolutions and feature fusion to enhance multiscale information capture and improve segmentation accuracy (36). In our approach, effective constraints are primarily derived from the diagonal and off-diagonal (cross-peak) regions near the diagonal of the 2DIR spectrum (*SI Appendix*, Fig. S2). Diagonal peaks represent fundamental vibrational frequencies and local environments, while cross-peaks provide information about specific vibrational couplings (23). These regions offer valuable insights into molecular structure, with approximately 2,050 pixels on average serving as meaningful constraints in our study (*SI Appendix*, Fig. S2). Subsequent layers progressively upsample and reduce dimensions to ultimately produce the structural predictions. To ensure comprehensive feature utilization, lower-level features from intermediate layers are concatenated with features before the final upsampling layer (37). To handle proteins of varying sizes, padding and a Maskloss function are employed, focusing on the nonpadded sections of our data, thus ensuring robustness in training and prediction (*SI Appendix*, Fig. S3). A gradient-based folding algorithm (10, 38) is then used to generate the three-dimensional protein backbone structures.

Finally, in the Model Application section, the trained ML model is used to predict both static protein structures and dynamic changes during protein folding. Additionally, this AI-based protocol, using 2DIR spectral descriptors, shows potential in predicting the structures of previously uncharacterized proteins. By integrating AI with 2DIR spectroscopy, this method offers a promising approach for analyzing protein dynamics and characterizing unknown protein structures.

Protein Static Structure Prediction. The ML model's ability to predict protein $C\alpha$ distance maps was evaluated using Mean Absolute Error (MAE) and precision metrics such as Top-L/5, Top-L/2, and Top-L for long-range predictions (38, 39) (Fig. 2*A*). The accuracy of the predicted three-dimensional protein backbone structures was assessed using RMSD. The training and testing loss curves indicate effective parameter optimization, with rapid initial convergence followed by stabilization, demonstrating the model's strong generalization capabilities. The close alignment of the training and testing losses, achieved by implementing batch normalization (BN) in each layer to stabilize data distribution and accelerate training (40), suggests minimal overfitting. The robustness of the model was further supported by a model-saving strategy based on validation performance (*SI Appendix*, Fig. S3).

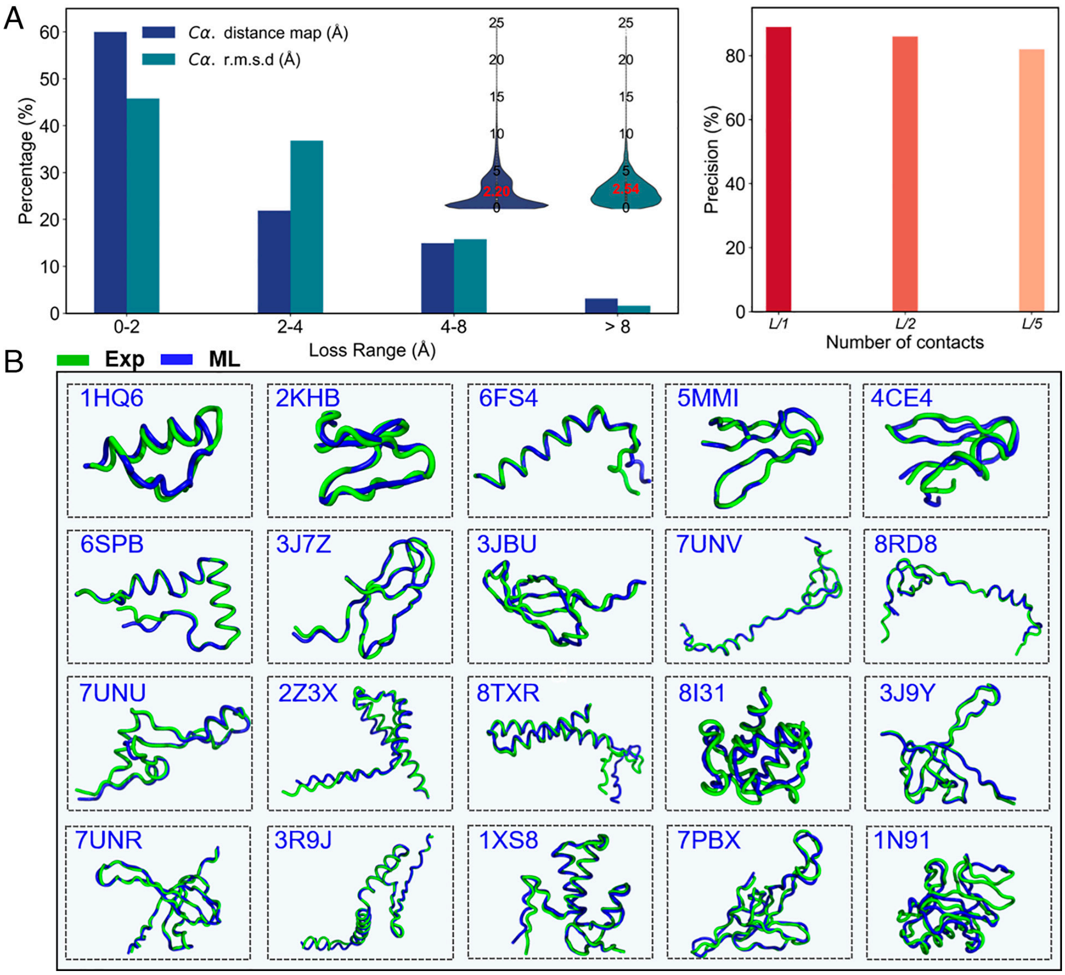


Fig. 2. ML Prediction of Protein Static Structures. (*A*) Evaluation metrics for ML predictions including $C\alpha$ distance map and precision metrics (Top-L/5, Top-L/2, and Top-L precision), alongside the RMSD of the predicted and experimental protein 3D backbone structures. (*B*) Comparison of ML predicted with experimentally determined structures for various proteins, ranging from 10 to 100 residues.

Comprehensive cross-validation results showed low MAE values (average $C\alpha$ distance map: 2.20 Å) and high long-range precision (accuracy >0.8) across 4,954 diverse protein test sets, confirming the model's reliability in predicting protein distance maps, as illustrated in Fig. 2*A*. We also assessed the model's performance across different protein sizes. As shown in *SI Appendix*, Fig. S4, very small proteins (0 to 20 amino acids) lack stable structural features, leading to higher prediction errors due to increased conformational variability. For proteins in the 20 to 40 amino acid range, more recognizable motifs emerge, allowing the model to perform better. However, as protein size exceeds 40 amino acids, increasing complexity and long-range interactions lead to a rise in errors. Overall, protein structures with loosely connected or unstable regions exhibit higher prediction deviations due to greater conformational variability (*SI Appendix*, Fig. S5). Future work will focus on improving prediction accuracy by incorporating biophysical constraints to enhance the model's accuracy. Despite these trends, the model achieved RMSD values comparable to experimental protein structures (average $C\alpha$ RMSD: 2.54 Å), as illustrated in Fig. 2 and *SI Appendix*, Table S1, underscoring the effectiveness of 2DIR spectral descriptors in predicting accurate three-dimensional protein structures.

To evaluate the model's performance on more complex proteins, we tested it on proteins ranging from 100 to 150 amino acids, derived from 10,000 structures in the RCSB database (*SI Appendix*, Fig. S6). Transfer learning was employed to enhance the model's transferability (Details in *SI Appendix*). Despite the limitations of the dataset and the increased complexity associated with larger proteins, the model performed well after fine-tuning with 70% of the dataset, while the remaining 30% was used for testing (*SI Appendix*, Fig. S6*B*). In 70% of the pretraining cases, the model achieved an average $C\alpha$ RMSD of approximately 3.33 Å (*SI Appendix*, Fig. S6*C*). Expanding the training dataset could further reduce error (*SI Appendix*, Fig. S6*B*), highlighting the model's strong transferability to larger, more structurally complex proteins. We also assessed the model's robustness at varying temperatures by investigating the molecular dynamics trajectories of the BsUDG–p56 complex (PDB: 3z0q) (*SI Appendix*, Fig. S7), which inhibits uracil removal and protects the phage genome from host DNA repair (41), at two temperatures (320 K and 348 K). These trajectories, sourced from a publicly available dataset (42), were used to study how temperature influences the complex's structural properties. A total of 2,500 conformations were

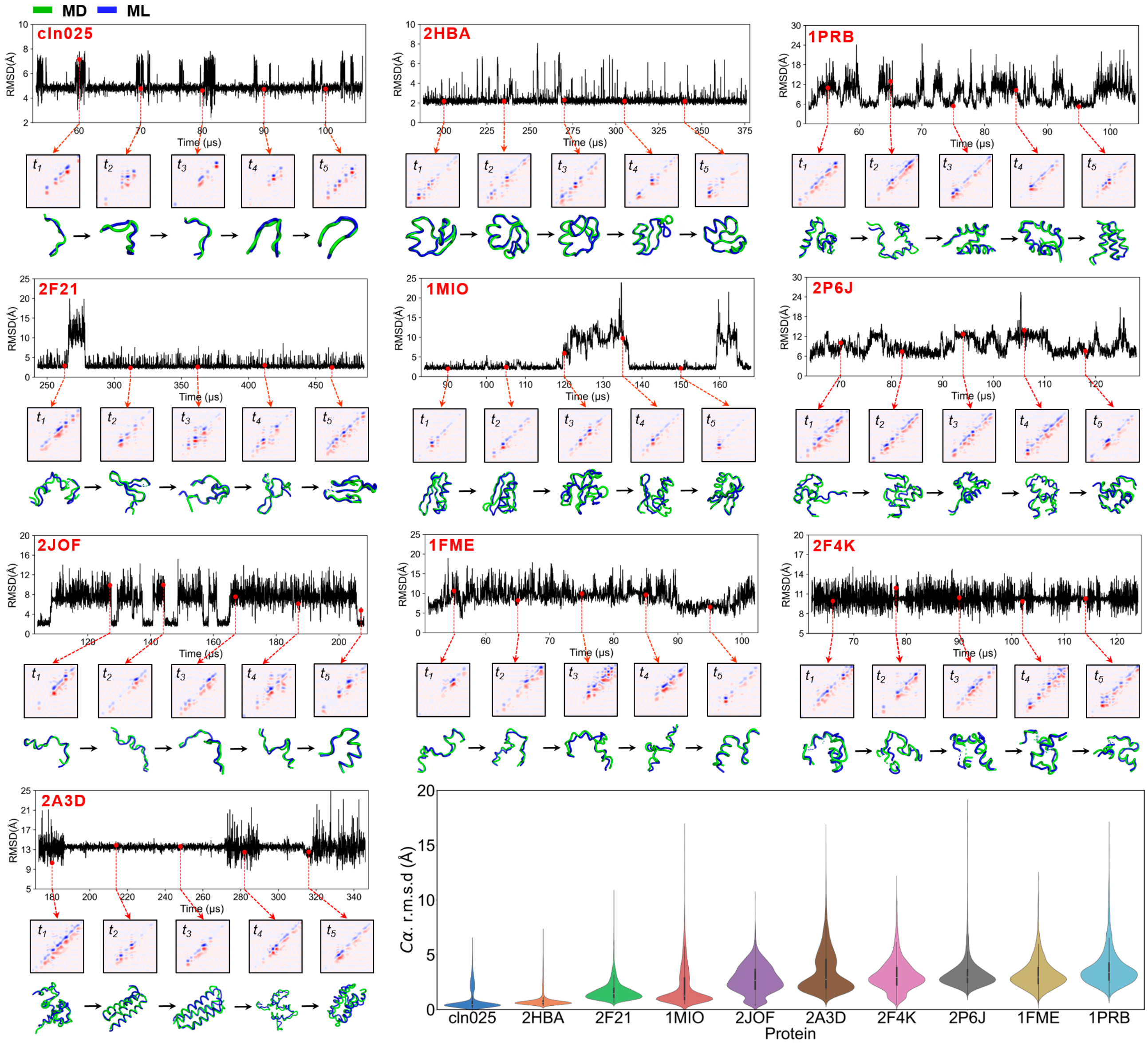


Fig. 3. ML prediction of protein dynamic structures in a reversible folding process.

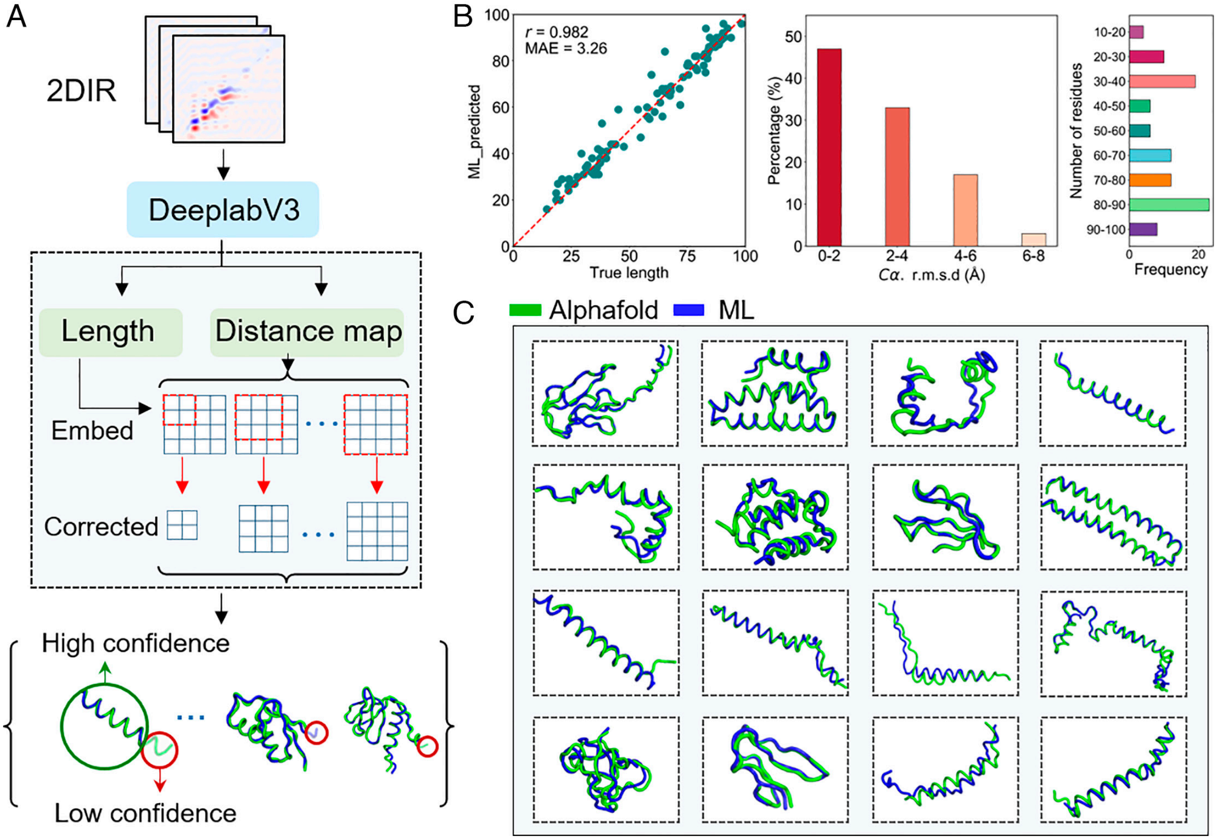


Fig. 4. ML prediction of unknown protein with 2DIR descriptor. (*A*) ML framework predicts unknown proteins. (*B* and *C*) ML prediction of 100 unknown proteins length and comparison of ML predicted with AlphaFold determined structures for various proteins.

generated at 1 ns intervals. We then used 50% of the dataset to fine-tune the model and update the pretrained weights via transfer learning, while the remaining 50% was used as a test set to validate the model's transferability. The results—average *Cα* RMSD of 2.63 Å at 320 K and 2.85 Å at 348 K—demonstrate that the model exhibits good transferability in predicting protein structures across different temperatures (*SI Appendix*, Fig. S7).

Then, to further evaluate whether the ML model accurately captures the underlying protein spectrum-structure relationship, we conducted an additional test. We first generated random 2DIR spectra that did not correspond to any real protein and input them into the trained model to produce protein structures. Using the Frenkel exciton Hamiltonian, we then calculated the 2DIR spectra for these generated structures. The calculated spectra were compared to the originally created spectra using the Structural Similarity Index Measure (SSIM), a widely used metric for image similarity (43) (*SI Appendix*, Fig. S8). The high SSIM values (>0.85) confirm that the model effectively captures and replicates the protein spectrum-structure relationship in a closed-loop manner.

Additionally, we created a 2DIR spectrum by combining spectra from purely α-helical and β-sheet proteins. The structure generated by the model from this combined spectrum exhibited features of both α-helices and β-sheets. Notably, an increase in β-sheet content correlated with a higher proportion of β spectra in the combination (*SI Appendix*, Fig. S9). This result demonstrates the model's ability to construct accurate and nuanced protein spectrum–structure relationships, reinforcing its potential for precise protein structure prediction and analysis.

Protein Dynamic Structure Prediction. 2DIR spectroscopy is renowned for its high spectral resolution and its ability to track conformational changes on the nanometer scale over picosecond to nanosecond timescales (23). To assess the ability of our ML model to predict dynamic protein structures, we selected ten well-characterized fast-folding proteins and extensively studied both theoretically and experimentally (44) (Fig. 3). These proteins, ranging from 10 to 80 amino acid residues, contained no disulfide bonds or prosthetic groups, and represented the three major structural classes: α-helical, β-sheet, and mixed α/β (44). The reversible folding trajectories of these proteins, which unfold over microseconds to milliseconds, were simulated using the Anton supercomputer, developed by the David E. Shaw research group (44–46).

Then approximately 10,000 conformations were sampled for each protein at consistent time intervals throughout the entire folding trajectory to calculate the corresponding 2DIR spectra. To improve the model's ability to predict dynamic protein structures, we employed transfer learning techniques. Predictive accuracy increased as more data were incorporated, with optimal performance achieved when 50% of the dataset was used (*SI Appendix*, Fig. S10). The first half of the dataset was used for fine-tuning and updating the pretrained weights, while the remaining half served as a test set to evaluate the model's transferability, as detailed in *SI Appendix*. Test set predictions began from the remaining 50% of the MD conformations, where alternating folding and unfolding events were clearly observed.

As shown in Fig. 3 and *SI Appendix*, Figs. S11–S13, the ML model accurately predicted the dynamic structures of the ten proteins throughout their folding processes, closely matching the reference MD structures over microsecond to millisecond timescales, with an average RMSD of 2.51 Å (*SI Appendix*, Table S2). For detailed visualization, we highlighted five characteristic states (t_1, t_2, t_3, t_4, and t_5) at equal intervals, focusing on significant changes during the reversible folding process. These states are further magnified in Fig. 3 to showcase the prediction details at corresponding times. The model's predictions align closely with the MD-simulated structures, effectively capturing the dynamic evolution of protein configurations. This level of accuracy demonstrates the model's ability to decode the complex, high-dimensional geometric data embedded in 2DIR spectral descriptors, further validating our approach for probing dynamic protein structures.

Application to Uncharacterized Proteins. The diversity and complexity of proteins are immense, with currently known proteins representing only a small fraction of the possible variations (47, 48). Discovering new proteins is crucial for advancing biomedical research. The ML model we developed, which uses 2DIR spectral descriptors as inputs, has been refined to detect previously uncharacterized proteins. The size of the distance map matrix generated by the model corresponds to the number of residues in a protein, effectively representing the protein's length. As shown in Fig. 4*A*, when the input is the 2DIR spectrum of an unknown protein, the model predicts both the elements of the distance matrix and the protein's length. It then combines this information to create the final distance map matrix. The three-dimensional structure of the protein is subsequently generated using a gradient-based folding algorithm.

To further validate the ML model's capability to detect unknown proteins, we first tested the ML model's prediction of protein length on a base dataset of 49,547 proteins. A high Pearson correlation coefficient of 0.957 and a MAE of 4.46 achieved in test set after cross-validation indicate that the ML model can accurately predict protein lengths using 2DIR descriptors (*SI Appendix*, Fig. S14*A*). Next, we randomly harvested 100 unknown proteins from the AlphaFold

Protein Structure Database (48), each containing up to 100 residues (*SI Appendix*, Fig. S14*B*), which were predicted by the AlphaFold program based on their initial sequences and have not yet been experimentally confirmed (Sequences information in *SI Appendix*, Table S3). By inputting their 2DIR spectra into the ML model, it was able to predict protein lengths with a Pearson correlation coefficient (r) of 0.982, an MAE of 3.26, and an average $C\alpha$ distance map deviation of 2.3 Å (Fig. 4*B* and *SI Appendix*, Fig. S14*B*). Finally, we compared the three-dimensional structures predicted by the ML model to those generated by AlphaFold, resulting in a RMSD of only 2.38 Å (Fig. 4*C* and *SI Appendix*, Fig. S15). These results indicate that the 2DIR descriptor can accurately detect unknown proteins length and structure.

Conclusions. In summary, our ML-based protocol accurately predicts three-dimensional protein structures using 2DIR spectroscopy descriptors. By establishing a precise "spectrum–structure" relationship in a closed-loop manner, this method not only predicts dynamic structures with high accuracy but also demonstrates broad applicability across various protein folding trajectories, spanning timescales from microseconds to milliseconds. Additionally, our protocol shows great potential for determining the structures of previously uncharacterized proteins based on their spectral data. Currently, the model focuses on $C\alpha$ distances to balance protein topology with computational efficiency. Expanding to all-atom positions could provide more detailed structural features, such as side-chain orientations and interactions. Future improvements will incorporate atomic-level data, including backbone torsion angles, side-chain dihedrals, and physics-based constraints, enabling more accurate predictions and a deeper understanding of protein structure and function. Furthermore, the accuracy of coupling and frequency models plays a crucial role in interpreting 2DIR spectra. Variations in coupling strengths can shift peak positions, intensities, and shapes, leading to biases and error accumulation, particularly in highly coupled modes. To mitigate these effects, feature scaling of both original and adjusted datasets is useful, especially for linear changes, while experimental validation is essential for nonlinear variations. Future work will focus on analyzing coupling variations, developing optimization algorithms, and validating models with experimental data. The integration of AI with 2DIR spectroscopy opens promising avenues for the real-time analysis of dynamic protein structures.

Methods

2DIR Spectra Simulations. We use the general framework of Frenkel excitons Hamiltonian to describe the amide I mode (25), specifically,

$$\mathbf{H} = \sum_i^N \ \omega_i \boldsymbol{b}_i^\dagger \boldsymbol{b}_i + \sum_{i,j}^N \ J_{ij} \boldsymbol{b}_i^\dagger \boldsymbol{b}_j - \sum_i^N \tfrac{\Delta_i}{2} \boldsymbol{b}_i^\dagger \boldsymbol{b}_i^\dagger \boldsymbol{b}_i \boldsymbol{b}_i,$$

where ω_i represents the fundamental vibrational frequency of the local amide I mode i, J_{ij} is the vibrational coupling between two modes i and j, and Δ_i is the anharmonicity of mode i. For the amide I vibrational frequency (ω_i) in whole Hamiltonian is calculated with the Skinner map (49):

$$\omega = \omega_{map} + \ \sum_i \ P_{i,map} P_i + (\mathbf{E}_{i,map} \cdot \mathbf{E}_i),$$

Here, the ω_{map}, $P_{i,map}$, and $\mathbf{E}_{i,map}$ represent the vibrational frequency, electric potential, and electric field, respectively, as predefined in the map. *P*_*i* and E_*i* were computed as:

$$P = \ \sum_j \ \tfrac{q_j}{|\mathbf{r}_j|};$$

$$E_i = \ \sum_j \ \tfrac{q_j}{|\mathbf{r}_j|^3} (\mathbf{r}_j \cdot \hat{\mathbf{i}}).$$

E_i represents the electric field in the directions of x, y, and z. Then the vibrational coupling between each local amide I mode was calculated through the transition charge coupling (50, 51):

$$J = \frac{1}{4\pi\varepsilon} \sum\nolimits_{a,b} \left(\frac{dq_a dq_b}{|\mathbf{r}_{ab}|} - \frac{3q_a q_b (\mathbf{v}_a \cdot \mathbf{r}_{ab})(\mathbf{v}_b \cdot \mathbf{r}_{ab})}{|\mathbf{r}_{ab}|^5} - \frac{dq_a q_b (\mathbf{v}_b \cdot \mathbf{r}_{ab}) - q_a dq_b (\mathbf{v}_a \cdot \mathbf{r}_{ab}) - q_a q_b (\mathbf{v}_a \cdot \mathbf{v}_b)}{|\mathbf{r}_{ab}|^3} \right),$$

$$\begin{aligned} J = (1-u)(1-t) \cdot map\left(\left\lfloor \frac{\phi}{30} \right\rfloor \left\lfloor \frac{\psi}{30} \right\rfloor \right) + (1-u)\, t \cdot map\left(\left\lfloor \frac{\phi}{30} \right\rfloor \left\lfloor \frac{\psi}{30} \right\rfloor \right) \\ + u(1-t) \cdot map\left(\left\lfloor \frac{\phi}{30} \right\rfloor \left\lfloor \frac{\psi}{30} \right\rfloor \right) + u \cdot t \cdot map\left(\left\lfloor \frac{\phi}{30} \right\rfloor \left\lfloor \frac{\psi}{30} \right\rfloor \right) \end{aligned}.$$

Additionally, the dipole moment of each oscillator, derived from the relative positions of the C, O, and N atoms, is given by (52):

$$\mu = 2.73(\mathbf{s} - ((\mathbf{CO} \cdot \mathbf{s}) + \tfrac{\sqrt{|\mathbf{s}|^2 - (\mathbf{CO} \cdot \mathbf{s})^2}}{\tan 10})\mathbf{CO}).$$

Finally, the total 2DIR signal is generated by three laser pulses with wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_{3.}$ The absorptive signals are obtained by adding the rephrasing ($\mathbf{k}_{\text{I}} = \ -\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$) and nonrephasing ($\mathbf{k}_{\text{II}} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$) spectra, These signals are emitted in different directions, corresponding to the rephasing and nonrephasing pathways. The 2DIR signal can be decomposed into three primary contributions: ground-state bleach (GB), stimulated emission (SE), and excited-state absorption (EA), with each contribution having both rephasing and nonrephasing components.

$$\begin{aligned} S^{(\text{I})}(\omega_3, t_2, \omega_1) = \int_0^\infty \int_0^\infty dt_3 dt_1 (S_{\text{GB}}^{(\text{I})}(t_3, t_2, t_1) + S_{\text{SE}}^{(\text{I})}(t_3, t_2, t_1) \\ + S_{\text{EA}}^{(\text{I})}(t_3, t_2, t_1)) exp(i(\omega_3 t_3 - \omega_1 t_1)) \end{aligned},$$

$$\begin{aligned} S^{(\text{II})}(\omega_3, t_2, \omega_1) = \int_0^\infty \int_0^\infty dt_3 dt_1 (S_{\text{GB}}^{(\text{II})}(t_3, t_2, t_1) + S_{\text{SE}}^{(\text{II})}(t_3, t_2, t_1) \\ + S_{\text{EA}}^{(\text{II})}(t_3, t_2, t_1)) exp(i(\omega_3 t_3 + \omega_1 t_1)) \end{aligned}.$$

The total response function is the sum of the rephrasing and nonrephasing signals:

$$I_{2D}(\omega_3, t_2, \omega_1) = S^{(I)}(\omega_3, t_2, \omega_1) + S^{(II)}(\omega_3, t_2, \omega_1),$$

where ω_1 and ω_3 correspond to the frequencies of t_1 and t_3 after double Fourier transformation, respectively.

The Hamiltonian was generated using the AIM program (53), and 2DIR simulations were performed with the NISE_2017 spectral simulation package (53). To expedite the calculations, couplings smaller than 0.01 cm^{-1} were neglected, and the anharmonicity was set to 16 cm^{-1} (54). The 2DIR spectra were calculated within the spectral window of 1,550 to 1,750 cm^{-1}, using coherence times ranging from 0 to 2.56 ps with 20 fs increments, with a waiting time (t_2) of 0 ps. Smoothing was applied using an exponentially decaying function with an effective lifetime of 1.8 ps. Since the spectrum was calculated for a single snapshot, only a single averaging step was performed to further accelerate the process.

ML Protocol. The whole ML protocol was designed based on the DeepLabV3 model (35, 36) to predict the three-dimensional structure of proteins from 2DIR spectral images. The architecture comprises three key components: a feature extraction module, a spatial resolution recovery module, and a regression output module.

In protein 2DIR spectroscopy, the frequency and coupling information of vibrational modes are stored in a contour map, where the excitation frequency (ω_1) and detection frequency (ω_3) serve as the coordinate axes (25). First, we utilize a pretrained DeepLabV3 model with a ResNet-50 backbone (55) to extract features from 2DIR images within the 1,575 to 1,725 cm^{-1} spectral window. The input 2DIR signals are converted into 3 × 224 × 224 RGB images. The DeepLabV3 network then extracts high-level features of 2,048 × 28 × 28, leveraging atrous convolutions to enhance multiscale information capture and improve segmentation accuracy (56). These features contain rich spatial information, aiding subsequent structural prediction.

Next, the model recovers spatial dimensions through a series of upsampling convolutional layers. The upsampling process involves increasing the feature map from 2,048 channels to 512 channels while doubling the spatial resolution. ReLU activation functions and BN layers enhance nonlinear expressiveness and stability (57). Lower-level features from intermediate DeepLabV3 layers are concatenated with features before the final upsampling layer. Further upsampling reduces the feature map to 128 channels, then 32 channels, and finally outputs a single-channel structural prediction. An adaptive average pooling layer adjusts the output to the target size. To handle proteins of different sizes, we introduce padding and a Maskloss function to focus on nonpadding data parts (58). This approach ensures the model's robustness to variable-length sequences and avoids interference from padding during training. The protein distance map obtained through these steps, combined with a gradient-based folding algorithm (10), ultimately generates the three-dimensional backbone structure of the protein.

Data, Materials, and Software Availability. The machine learning code and simulation dataset is available at https://github.com/ZhuLvs/2DIR (59). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was financially supported by the National Natural Science Foundation of China (Grants 22203001, 22033007, 21633007, 22025304, 22273093, 62236002, and 61921004), Innovation Program for Quantum Science and Technology grant 2021ZD0303303, the Chinese Academy of Sciences Project for Young Scientists in Basic Research (YSBR-005), K. C. Wong Education (GJTD-2020-15), the University Synergy Innovation Program of Anhui Province (GXXT-2022-062), the Natural Science Foundation of Anhui Province (Grants 2208085Y04), and the open research fund of Key Laboratory of Precision and Intelligent Chemistry. S.M. gratefully acknowledges the support of NSF grant CHE-2246379.

Author affiliations: [a]Engineering Research Center of Autonomous Unmanned System Technology, Ministry of Education, Anhui Provincial Engineering Research Center for Unmanned System and Intelligent Technology, School of AI, Anhui University, Hefei 230601, China; [b]State Key Laboratory of Precision and Intelligent Chemistry, Hefei National Research Center for Physical Sciences at the Microscale, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei 230026, Anhui, China; [c]Zernike Institute for Advanced Materials, Department of Nanoscience and Materials Science, University of Groningen, Groningen 9747AG, Netherlands; and [d]Department of Chemistry and Department of Physics & Astronomy, University of California, Irvine, CA 92697

1. D. Whitford, *Proteins: Structure and Function* (John Wiley & Sons, 2013).
2. H. N. Chapman *et al.*, Femtosecond X-ray protein nanocrystallography. *Nature* **470**, 73–77 (2011).
3. G. Brändén, R. Neutze, Advances and challenges in time-resolved macromolecular crystallography. *Science* **373**, eaba0954 (2021).
4. Y. Cheng, Single-particle cryo-EM—How did it get here and where will it go. *Science* **361**, 876–880 (2018).
5. Y. Cheng, Single-particle cryo-EM at crystallographic resolution. *Cell* **161**, 450–457 (2015).
6. A. Mittermaier, L. E. Kay, New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224–228 (2006).
7. I. Pupeza *et al.*, Field-resolved infrared spectroscopy of biological systems. *Nature* **577**, 52–59 (2020).
8. C. Ewen, Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures. *Nature* **634**, 525–526 (2024).
9. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
11. J. Dauparas *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
12. S. Mosalaganti *et al.*, AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **376**, eabm9506 (2022).
13. X. Fang *et al.*, A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat. Mach. Intell.* **5**, 1087–1096 (2023).
14. J. Xu, M. McPartlon, J. Li, Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat. Mach. Intell.* **3**, 601–609 (2021).
15. T. J. Lane, Protein structure prediction has reached the single-structure frontier. *Nat. Methods* **20**, 170–173 (2023).
16. I. Pupeza *et al.*, Field-resolved infrared spectroscopy of biological systems. *Nature* **577**, 52–59 (2020).
17. V. A. Lorenz-Fonfria, Infrared difference spectroscopy of proteins: From bands to bonds. *Chem. Rev.* **120**, 3466–3576 (2020).
18. A. Barth, Infrared spectroscopy of proteins. *Biochim. Biophys. Acta Bioenerg.* **1767**, 1073–1101 (2007).
19. C. R. Baiz *et al.*, Vibrational spectroscopic map, vibrational spectroscopy, and intermolecular interaction. *Chem. Rev.* **120**, 7152–7218 (2020).
20. S. Mukamel *et al.*, Coherent multidimensional optical probes for electron correlations and exciton dynamics: From NMR to X-rays. *Acc. Chem. Res.* **42**, 553–562 (2009).
21. J. P. Kraack, P. Hamm, Surface-sensitive and surface-specific ultrafast two-dimensional vibrational spectroscopy. *Chem. Rev.* **117**, 10623–10664 (2017).
22. H. T. Kratochvil *et al.*, Instantaneous ion configurations in the K$^+$ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **353**, 1040–1044 (2016).
23. A. Ghosh, J. S. Ostrander, M. T. Zanni, Watching proteins wiggle: Mapping structures with two-dimensional infrared spectroscopy. *Chem. Rev.* **117**, 10726–10759 (2017).
24. N. T. Hunt, Using 2D-IR spectroscopy to measure the structure, dynamics, and intermolecular interactions of proteins in H2O. *Acc. Chem. Res.* **57**, 685–692 (2024).
25. P. Hamm, M. Zanni, *Concepts and Methods of 2D Infrared Spectroscopy* (Cambridge University Press, 2011).
26. V. K. Shukla, G. T. Heller, D. F. Hansen, Biomolecular NMR spectroscopy in the era of artificial intelligence. *Structure* **31**, 1360–1374 (2023).
27. G. Karunanithy, V. K. Shukla, D. F. Hansen, Solution-state methyl NMR spectroscopy of large non-deuterated proteins enabled by deep neural networks. *Nat. Commun.* **15**, 5073 (2024).
28. P. Klukowski, R. Riek, P. Güntert, Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat. Commun.* **13**, 6151 (2022).
29. A. L. Ptaszek, J. Li, R. Konrat, G. Platzer, T. Head-Gordon, UCBShift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures. *J. Am. Chem. Soc.* **146**, 31733–31745 (2024).
30. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
31. P. W. Rose *et al.*, The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271–D281 (2016).
32. B. Boeckmann *et al.*, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
33. H. Kim, M. Cho, Infrared probes for studying the structure and dynamics of biomolecules. *Chem. Rev.* **113**, 5817–5847 (2013).
34. T. L. C. Jansen, J. Knoester, Waiting time dynamics in two-dimensional infrared spectroscopy. *Acc. Chem. Res.* **42**, 1405–1411 (2009).
35. L.-C. Chen, Rethinking atrous convolution for semantic image segmentation. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1706.05587 (Accessed 17 June 2017).
36. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation" in *Proceedings of the European conference on computer vision (ECCV)*, (2018), pp. 801–818.
37. J. Yang, C. Wu, B. Du, L. Zhang, Enhanced multiscale feature fusion network for HSI classification. *IEEE Trans. Geosci. Remote Sens.* **59**, 10328–10347 (2021).
38. A. W. Senior *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
39. Z. Guo, J. Liu, J. Skolnick, J. Cheng, Prediction of inter-chain distance maps of protein complexes with 2D attention-based deep neural networks. *Nat. Commun.* **13**, 6963 (2022).
40. S. Santurkar, D. Tsipras, A. Ilyas, A. Madry, How does batch normalization help optimization? Advances in neural information processing systems. arXiv[Preprint] (2018). https://doi.org/10.48550/arXiv.1805.11604 (Accessed 29 May 2018).

41. J. I. Baños-Sanz *et al.*, Crystal structure and functional insights into uracil-DNA glycosylase inhibition by phage ϕ29 DNA mimic protein p56. *Nucleic Acids Res.* **41**, 6761–6773 (2013).
42. A. Mirarchi, T. Giorgino, G. De Fabritiis, mdCATH: A large-scale MD dataset for data-driven computational biophysics. *Sci. Data* **11**, 1299 (2024).
43. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–613 (2004).
44. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
45. D. E. Shaw *et al.*, Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
46. D. E. Shaw *et al.*, "Millisecond-scale molecular dynamics simulations on Anton" in *Proceedings of the conference on High Performance Computing Networking, Storage and Analysis*, (2009), pp. 1-11.
47. C. L. Worth, S. Gong, T. L. Blundell, Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **10**, 709–720 (2009).
48. M. Varadi *et al.*, AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
49. L. Wang, C. T. Middleton, M. T. Zanni, J. L. Skinner, Development and validation of transferable amide I vibrational frequency maps for peptides. *J. Phys. Chem. B* **115**, 3713–3724 (2011).
50. P. Hamm, S. Woutersen, Coupling of the amide I modes of the glycine dipeptide. *Bull. Chem. Soc. Jpn.* **75**, 985–988 (2002).
51. R. D. Gorbunov, D. S. Kosov, G. Stock, Ab initio-based exciton model of amide I vibrations in peptides: Definition, conformational dependence, and transferability. *J. Chem. Phys.* **122**, 224904 (2005).
52. H. Torii, M. Tasumi, Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di-and tripeptides and considerations on the conformation of the extended helix. *J. Raman Spectrosc.* **29**, 81–86 (1998).
53. K. E. van Adrichem, T. L. C. Jansen, AIM: A mapping program for infrared spectroscopy of proteins. *J. Chem. Theory Comput.* **18**, 3089–3098 (2022).
54. P. Hamm, M. Lim, R. M. Hochstrasser, Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *J. Phys. Chem. B* **102**, 6123–6134 (1998).
55. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 770–778.
56. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–856 (2017).
57. S. Ioffe, Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv [Preprint] (2015). https://doi.org/10.48550/arXiv.1502.03167 (Accessed 11 February 2015).
58. O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015" in *18th International Conference*, (Springer, Munich, Germany), pp. 234–241. Proceedings, part III 8.
59. S. Ye *et al.*, AI protocol for retrieving protein dynamic structures from two-dimensional infrared spectra. GitHub. https://github.com/ZhuLvs/2DIR. Deposited 28 August 2024.

Supplementary Information for

AI Protocol for Retrieving Protein Dynamic Structures from Two-Dimensional Infrared Spectra

Sheng Ye[1]†, Lvshuai Zhu[1]†, Zhicheng Zhao[1]†, Fan Wu[2], Zhipeng Li[1], BinBin Wang[1], Kai Zhong[3*], Changyin Sun[1]*, Shaul Mukamel[4]* and Jun Jiang[2]*

[1]Engineering Research Center of Autonomous Unmanned System Technology, Ministry of Education, Anhui Provincial Engineering Research Center for Unmanned System and Intelligent Technology, School of Artificial Intelligence, Anhui University, Hefei, Anhui 230601, China.

[2]State Key Laboratory of Precision and Intelligent Chemistry, Hefei National Research Center for Physical Sciences at the Microscale, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei 230026, Anhui, China.

[3]Zernike Institute for Advanced Materials, Department of Nanoscience and Materials Science, University of Groningen, Groningen 9747AG, Netherlands.

[4]Department of Chemistry and Department of Physics & Astronomy, University of California, Irvine, California 92697, United States.

†These authors contributed equally to this work.

Table of Contents

Training and testing set and cross-validation

The accuracy and robustness of the machine learning predictions were evaluated using a cross-validation technique. The dataset was randomly divided into ten equal-sized subsets. In each iteration, one subset was used as the test set, while the remaining nine subsets were used for training. This process was repeated for all subsets to ensure reliable performance assessment.

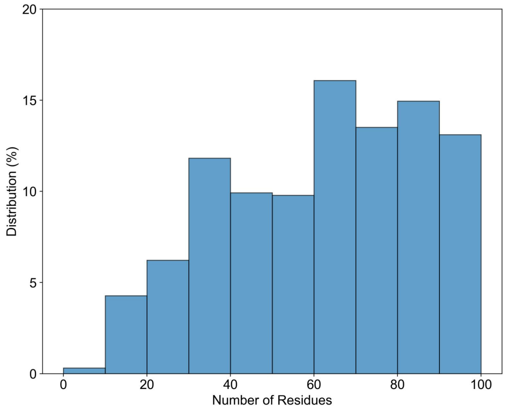
Transfer Learning details

During the pre-training phase, we used the Adam optimizer with an initial learning rate of 0.001 and employed a linear learning rate decay strategy[1]. The pre-training dataset was randomly split into training and validation sets in a 9:1 ratio. The model was trained for 100 epochs with a batch size of 32, maintaining the linear decay strategy throughout.

In the transfer learning step, we loaded the pre-trained model weights to optimize predictions for specific protein dynamic folding trajectories. The optimizer and learning rate strategy remained consistent with those used in the pre-training phase, and all layers' weights were updated. For instance, in the case of protein dynamic structure prediction, the fine-tuning dataset consisted of 2DIR spectra calculated from approximately 10,000 conformations sampled at equal time intervals from complete folding trajectories[2]. This dataset was randomly split into training and validation sets with a 5:5 ratios and the fine-tuning configuration mirrored that of the pre-training phase.

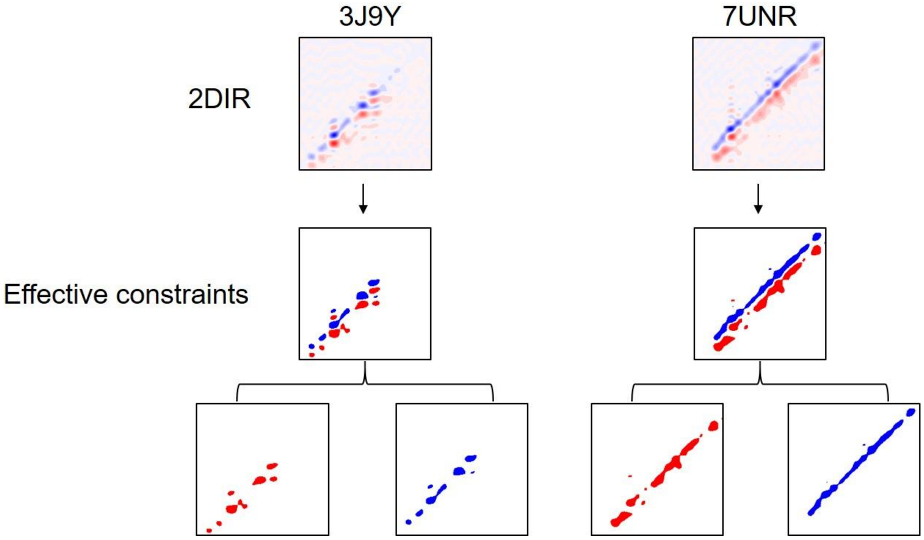
Supplementary information for Figures

Distribution of protein length



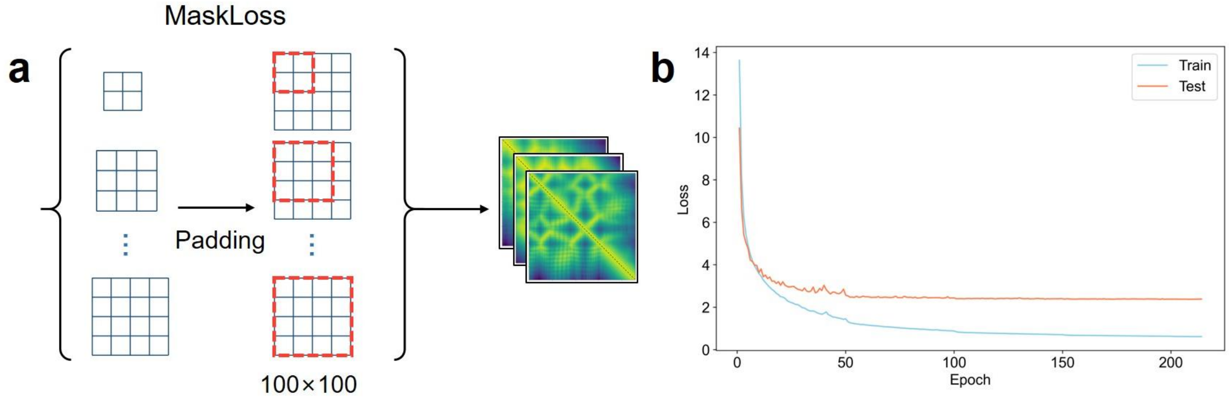
Supplementary Fig. 1 | Distribution of protein residue numbers in the basic dataset.

Effective constraints in 2DIR spectra



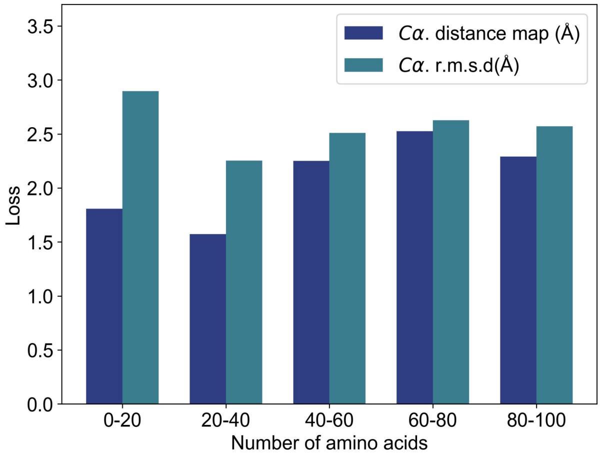
Supplementary Fig. 2 | Effective constraints in 2DIR spectroscopy for machine learning prediction of protein structures (example proteins: 3J9Y and 7UNR).

Mask loss and loss curves

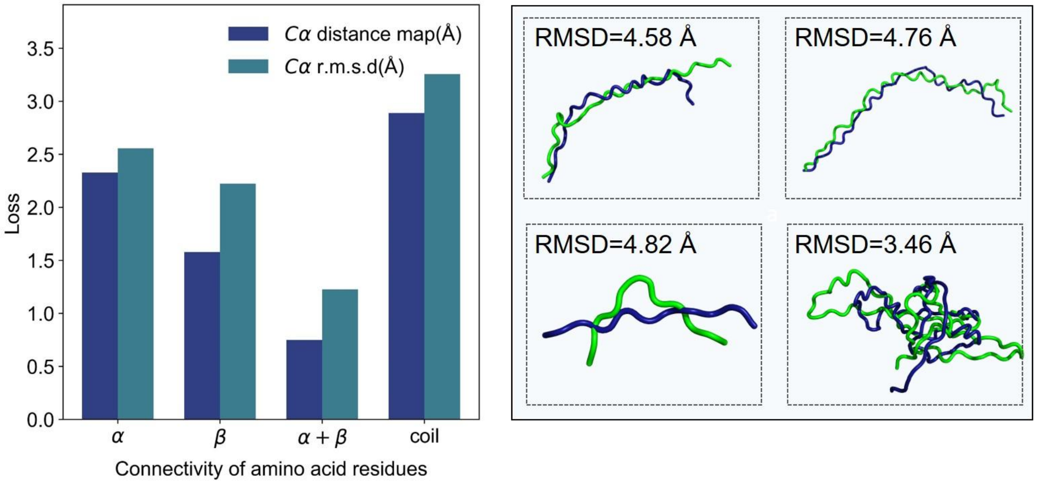


Supplementary Fig. 3 | (**a**) Standardizing proteins of varying lengths to a uniform size through padding and using MaskLoss to focus on learning from the non-padded regions. (**b**) The loss curves in whole training process.

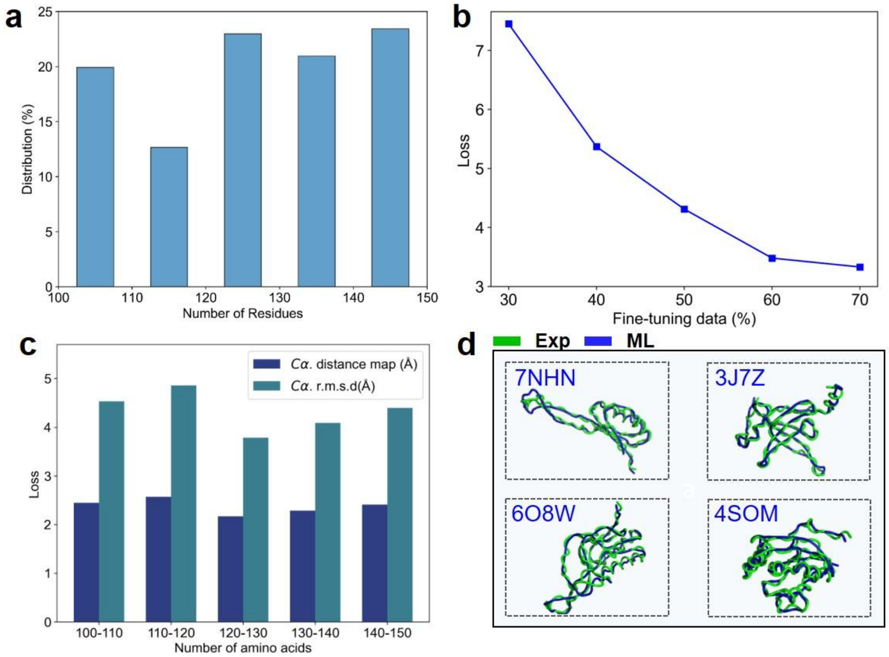
ML model evaluation



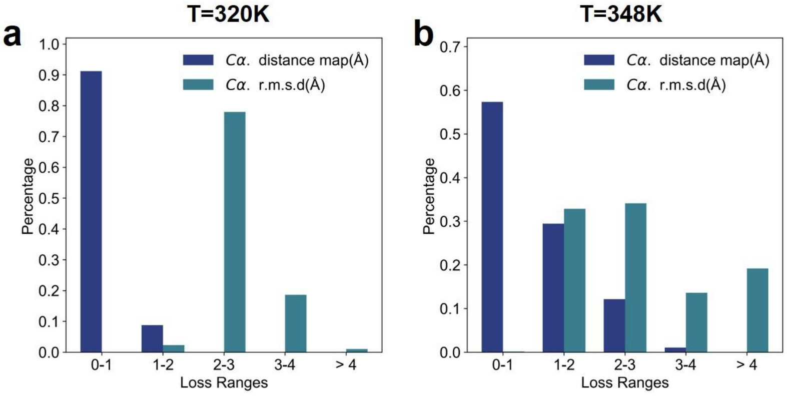
Supplementary Fig. 4 | Machine learning prediction of protein structure across varying protein sizes.



Supplementary Fig. 5 | Machine learning structure prediction of different protein types.

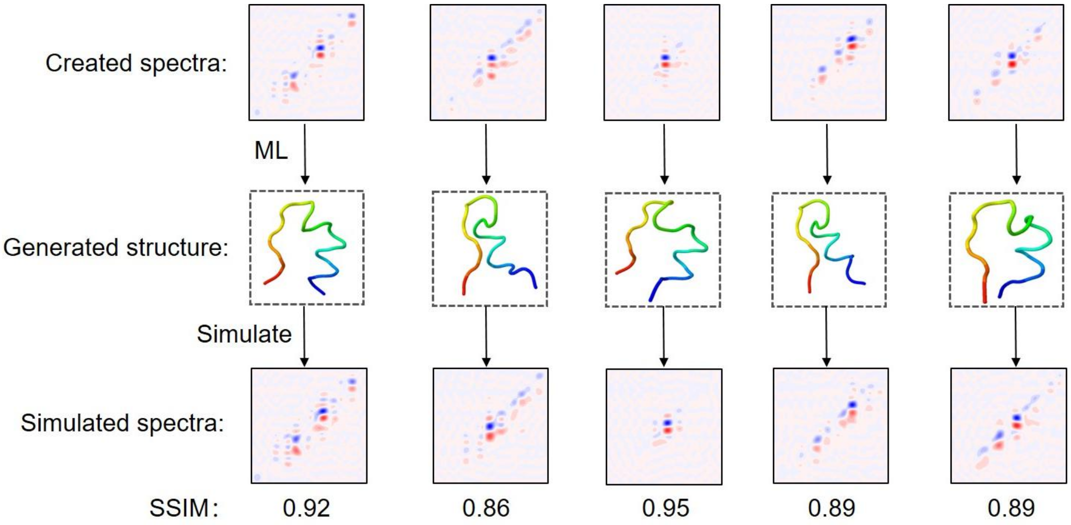


Supplementary Fig. 6 | (a) Distribution of protein lengths between 100 and 150 residues. **(b)** Proportion of fine-tuning datasets used for predicting protein structures of 100-150 residue length in transfer learning. **(c)** Evaluation metrics for machine learning predictions, including the *Cα* distance map and the RMSD between predicted and experimentally determined 3D protein backbone structures. **(d)** Comparison of machine learning-predicted structures with experimentally determined structures for proteins ranging from 100 to 150 residues.



Supplementary Fig. 7 | Machine learning prediction of BsUDG-p56 complex (PDB:3z0q) protein structure at different temperatures.

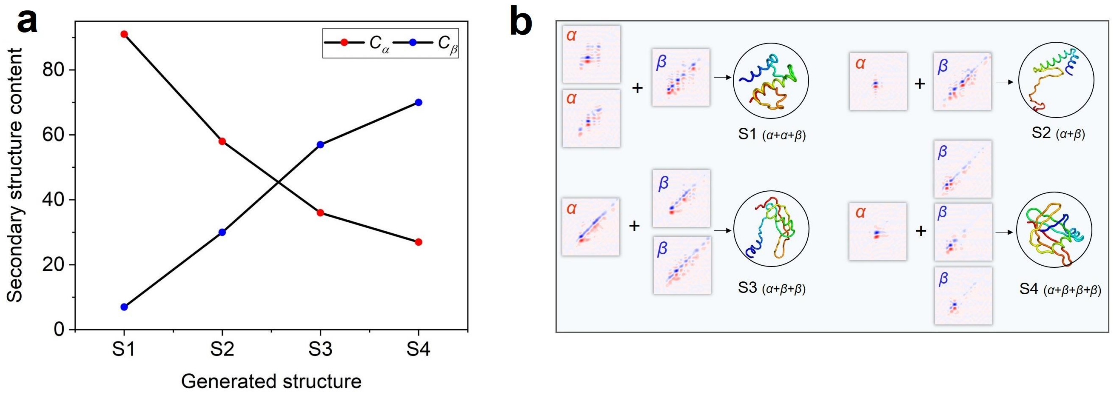
ML model verification



Supplementary Fig. 8 | Verified the correctness of the ML model by comparing the randomly created spectra and the calculated spectra.

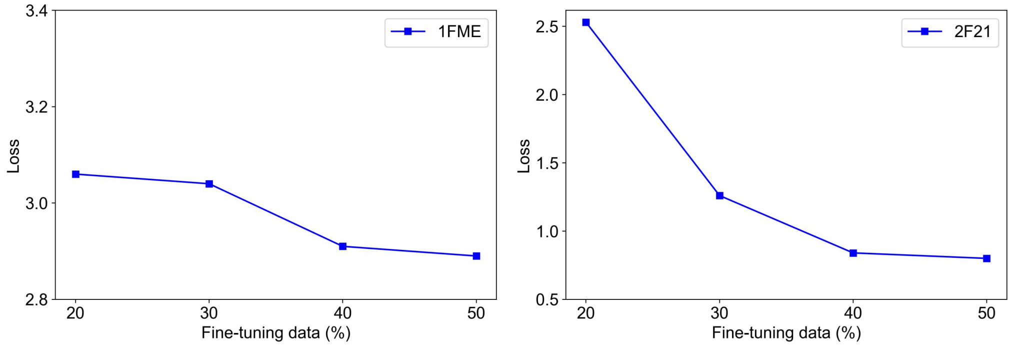
The formula for calculating the Structural Similarity Index (SSIM) is[3]:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2+\mu_y^2 + c_1)\quad(\sigma_x^2 + \sigma_y^2 + c_2)}$$

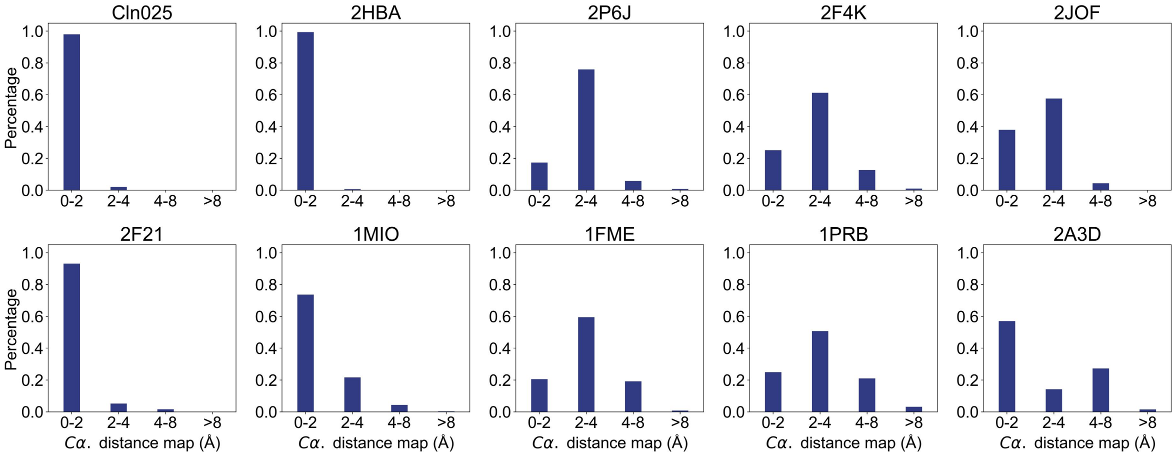


Supplementary Fig. 9 | (**a**) Validated the ML model by correlating spectral features with secondary structural elements. (**b**) Created new 2DIR spectrum by combining different proportions spectra from purely *α*-helical and *β*-sheet proteins.

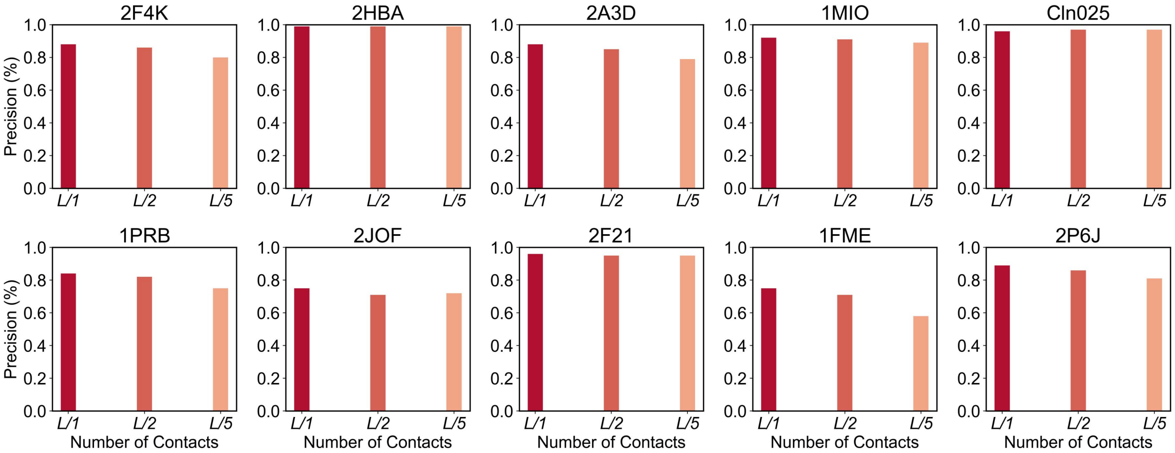
ML prediction of protein dynamic structures



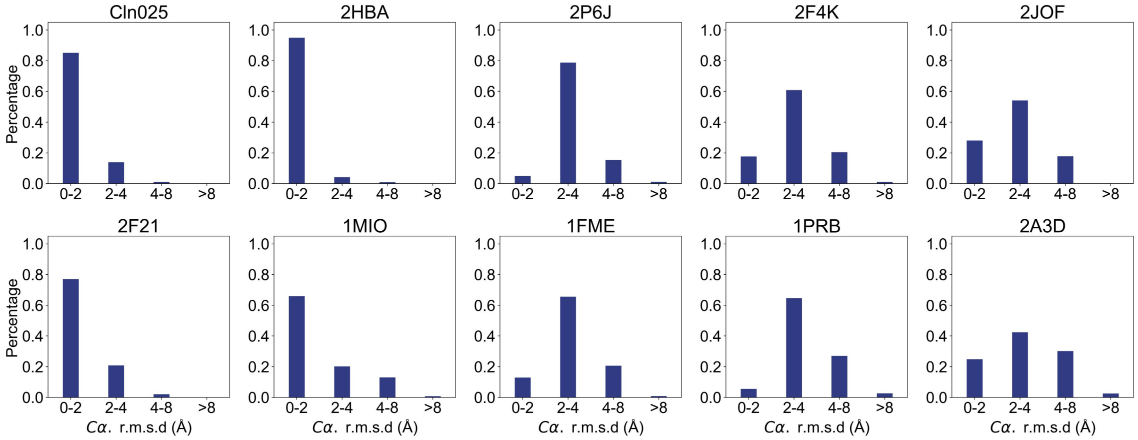
Supplementary Fig. 10 | Proportion of fine-tuning datasets utilized for predicting protein (1FME and 2F21) dynamic structures in transfer learning.



Supplementary Fig. 11 | ML predictions of *Cα* distance map for ten proteins dynamic structures in reversible folding process.

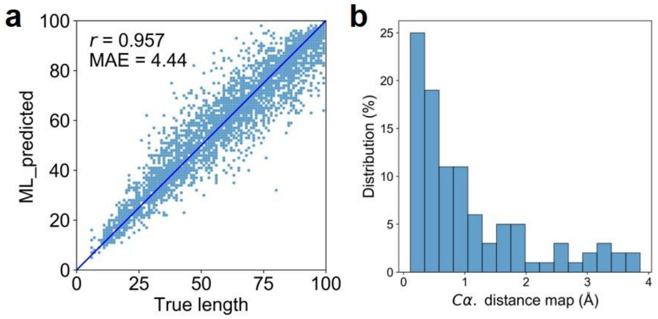


Supplementary Fig. 12 | ML predictions for ten proteins dynamic structures in reversible folding process with alongside precision metrics (Top-L/5, Top-L/2, and Top-L precision).

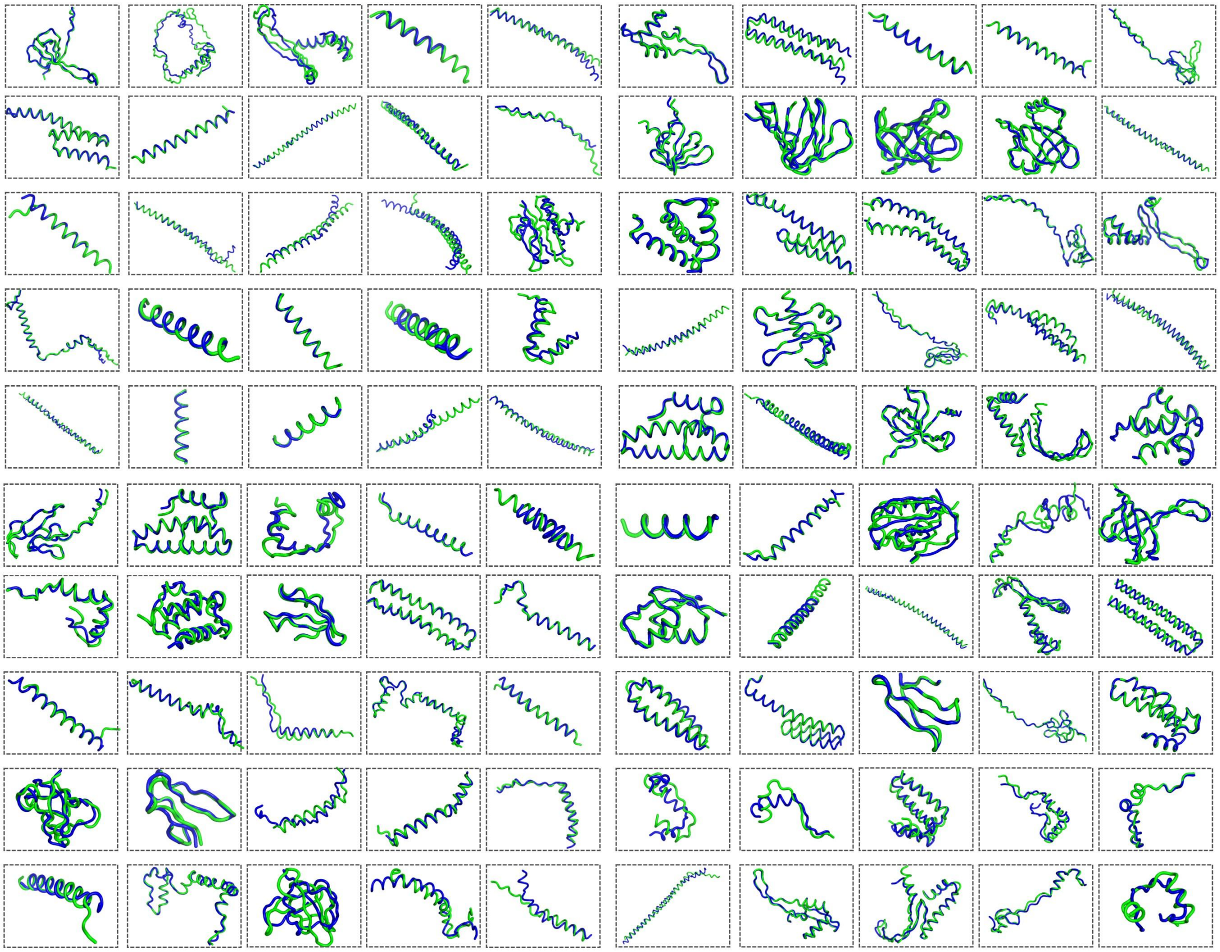


Supplementary Fig. 13 | ML predictions for ten proteins dynamic structures in reversible folding process with alongside root mean square deviation (*Cα.* r.m.s.d) of predicted versus actual 3D structures of *Cα* atoms.

Unknown protein prediction



Supplementary Fig. 14 | (**a**) ML prediction of protein length in a basic dataset. (**b**) ML predictions *Cα* distance map of 100 unknown proteins.



Supplementary Fig. 15 | Comparison of ML predicted with AlphaFold determined structures for various proteins.

Supplementary information for Tables

Supplementary Table 1. Comparison of ML predicted with experimentally determined structures for various proteins, ranging from 10 to 100 residues.

PDB ID	Number of residues	RMSD (Å)
1HQ6	26	2.06
2KHB	29	0.90
6FS4	32	1.40
5MMI	37	0.14
4CE4	41	1.17
6SPB	49	1.96
3J7Z	55	1.15
3JBU	55	2.59
7UNV	59	1.84
8RDB	60	3.02
7UNU	63	3.13
2Z3X	70	3.35
8TXR	76	3.22
8I31	78	2.68
3J9Y	84	0.59
7UNR	88	1.03
3R9J	88	1.24
1XS8	92	1.26
7RBX	96	1.64
1N91	98	3.44

Supplementary Table 2. ML prediction of protein dynamic structures in reversible folding process.

Protein name	PDB ID	MAE (Å)	RMSD(Å)
Cln025	Cln025	0.43	0.87
NTL9	2HBA	0.44	0.81
WW-domain	2F21	0.85	1.68
NuG2	1MIO	1.36	2.08
Trp-cag	2JOF	2.19	2.78
A3D	2A3D	2.55	3.44
UVF	2P6J	2.76	3.27
2F4K	2F4K	2.75	3.14
BBA	1FME	3.12	3.26
PRB	1PRB	3.27	3.73

Supplementary Table 3. Sequence information for 100 unknown protein from AlphaFold Protein Structure Database[4].

Index	Length	Sequence
1	84	MATKKAGGSTKNGRDSNPKMLGVKMYGGQAVTAGNIIVRQRGTEFHAGTNVGMGRDHTLFATADGVIKFEVKGQFGRRYVSVEA
2	89	MALTNADRAEIVAKFARAENDTGSPEVQVALLTAQINDLQGHFKEHKHDHHSRRGLIRMVNQRRKLLDYLKGKDATRYSDLIAALGLRR
3	36	ARNVLAALMDIIEATGATQVFYNHLYDPVSLVRDHR
4	34	MEVNILAFIATTLFILVPTAFLLIIYVKTASQND
5	27	MVVFLVGVLFLSIFVLLLLLAAISGIL
6	44	MKRTYQPSRLVRKRRHGFRARMATVGGRRVIGNRRAKGRKRLSA
7	79	MSEIADKVKKIVVEHLGVEESKVTPEASFIDDLGADSLDTVELVMAFEEAFNVEIPEDAAEKIATVKDAIDYIEKQKAA
8	41	MKIRNSLKSAKVRDKDCRVVRRRGRVYVINKKNPRMKARQG
9	82	MNPLIGAASVLAAGLAVGLAAIGPGMGQGTAAGYAVEGIARQPEAEGKIRGALLLSFAFMESLTIYGLVVALALLFANPFAS
10	38	MLTLKIFVYTVVTFFVSLFIFGFLSNDPGRNPGQKDLD
11	31	MALSDSQIFIALFTALITGILAVRLGIALYK
12	44	MESPAFFYTIFLWCLLLSITGYSIYVGFGPPSKTLRDPFEEHED
13	42	MTTKKSSYTYPIFTVRWLAVHALAVPTVFFLGSITAMQFIQR
14	81	MSGATGERPFSDILTSIRYWVIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTEDRQEAPLITDRFNALEQVKQLSE
15	34	MEVNILGLIATALFIIIPTSFLLILYVKTASQQP
16	67	MDTGTVKWFNDSKGFGFITPDAGGDDLFAHFSEVQGDGFKTLAENQKVSYETKRGPKGMQAANISPL
17	37	MKVRASVRKICDNCRLIRRKRKIMVICSNPKHKQRQG
18	31	MKKMSIKFKKLQTIRKKIVLVLKQNANFMNI
19	41	MKNFTTYLSTAPVVALIWFTFTAGLLIEINRFFPDPLVFSF
20	38	MVKPNPNKQSVELNRTSLYWGLLLIFVLAVLFSSYIFN
21	31	MAITESQIFIALFLSLITGILAVRLGIELYK
22	83	MSGATGERPFSDILTSIRYWVIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTEDRQEAPLITDRFNALEQIKELSQVD
23	67	METGTVKWFNDAKGFGFITPDGGGEDLFAHFSEIRIEGFKTLQENQKVTYEVKTGPKGKQAANIKPA
24	43	MALIIAKLPEAYAPFDPIVDVLPVIPVLFLALAFVWQASVSFR
25	32	LVYTFLLVGTLGIIFFAIFFREPPKVPSKGKK
26	77	MARVCQVTGKGPMSGNNVSHANNRTKRRFLPNLQNRRFFVESENRWVRLRVSNAGLRLIDKNGIDAVLADLRARGEI
27	75	MDVNAAKMIGAGLAAIGMGLAALGVGNVFAQFLAGALRNPGAADSQQGRLFIGFAAAELLGLLAFVTMIILVFVA
28	30	MKKTSLKMLATSLFTIFSRLQWVFQKRHAA
29	38	MIDEYVKKILDEKVTMYKQTSIHIQSLKRKALYFNSIK
30	84	MATKKAGGSTKNGRDSNPKMLGVKMYGGQAVTAGNIIVRQRGTEFHAGANVGMGRDHTLFATADGVIKFEVKGQFGRRYVSVEV
31	78	MSNKGQLLQDPFLNALRKEHVPVSIYLVNGIKLQGNIESFDQYVVLLRNTVTQMVYKHAISTVVPARPVNFHPDAESS

32	94	MLNVSEYFDGKVKSIGFDSVTIGRASVGVMAEGEYTFGTGQPEEMTVVSGALKVLLPGESEWKWYEAGSVFNVPGHSEFHLQVAEPTSYLCRYL
33	69	MAKIKGQVKWFNESKGFGFITPADGSKDVFVHFSAIQGNGFKTLAEGQAVEFEIQDGQKGPAAVNVTAI
34	94	MTTFNAEVRKEQGKGASRRLRVANKFPAIIYGGNEAPVAIELDHDVVMNLQAKPEFYTDVLTIVVDGKEIKVKAQAVQRHPFKPKLHHIDFVRA
35	68	MAPVKNIIMQRCFLESLLFLQIVFFYSKIISLSFIKWLNNINLKRDNIRYYALKNKRKSNKTKRLSKY
36	66	MQKYCELVRQKYAEIGSGDLGYVPDAIGCALNALNDIAANSALNSSVREQAAYAAANLLVSDYVDE
37	92	MANTAQARKRARQAAKANSHNSALRSKFRTAIKAVRKAIDAGDQAKAAEVFKSSVKTMDTIADKKIVHKNKAARHKSRLAAAIKGLQASAAQ
38	89	MTPESVMQIGQEAMRIALMLAAPLLLAALVSGLIISLLQAATQVNEQTLSFIPKILAVAATAVIAGPWMLNLVLDYMRNLFTNLPYIIG
39	85	MAHKKAGGSTRNGRDSNAKRLGVKRFGGESVLAGSIIVRQRGTKFHAGTNVGCGRDHTLFATADGKVQFEVKGPNNRKYISIVAE
40	90	MNKSQLIDKIAADADISKAAAGRVLDAFMGSVSDALKGGDEVALVGFGTFSVRERAARTGRNPQTGKEITIPAGKVPGFRAGKALKDSVN
41	59	MFVWNFFCLQIICCFLFNLVIQKIFRFYNCKSEKLILYFQFNKSISNLHRNQYRFFINI
42	82	MVTIRLARHGAKKRPFYQVVVTDSRNARNGRFIERVGFFNPIASGQAEGLRLDMDRIEHWVGQGATLSDRVNALIKEAKKAA
43	86	MAHKKAGGSSRNGRDSESKRLGVKVYGGQAINAGGIIVRQRGTRMHPGENVGIGKDHTLFALTDGHVQFTTKGAAKKHTVVVVPAA
44	92	MANSPSAKKRAKQAEKRRSHNASLRSMVRTYIKNVVKAIDAKDAEKAQAAYVLAVPVIDRMADKGIIHKNKAARHKSRLNGHVKALNVAAAA
45	74	KKKEEVERVQKEKADELNTVHEKIKQVVAKKDETFSNLKQQYEAACKRADHLEGLLEQQRHLMLKKQANSNKID
46	89	MALSVEEKAQIVTDYQQAVGDTGSPEVQVALLTANINKLQGHFKANGKDHHSRRGLIRMVNQRRKLLDYLKGKDVNRYSTLIGRLGLRR
47	68	MNLEDRVTDLESRLAFQDDTIQALNDVLVEQQRIVERLQLQMAALLKRQEEMAGQFESFEEEAPPPHY
48	90	MNDSVKTSLKRTLVGKVVSNKMDKTVTVLVEHRVKHPIYGKYVVRSKKYHAHDDANTYNEGDLVEIQETRPISKTKAWVVSKLLEAARVI
49	90	MNKSELIDAIAASADIPKAAAGRALDAVIESVTGALKAGDSVVLVGFGTFSVTDRPARVGRNPQTGKTLQIAAAKKPGFKAGKALKEAVN
50	78	MSNIEQQVKKIVAEQLGVSEAEVKNESSFQDDLGADSLDTVELVMALEEAFGCEIPDEEAEKITTVQLAIDYINAHNG
51	87	MSENKNVRTLQGKVVSDKMDKTVTVLVERKVKHPLYGKIIRLSTKIHAHDENNQYGIGDVVVIAESRPLSKTKSWVVKELVEKARTV
52	96	MALSLTDVEKIAKLSRLSLTEEEKGKTLSELNDIFAMVEKMQSVNTDGVEPMAHPHEAALRLREDTVTETDHAAEYQAVAPEVRNRLYIVPQVIEE
53	91	MNKSELIQAIADEAELSKRAASEFVDAFVSVVTQAMKDGKDVTLVGFGSFHTAQSAERKGRNPKTGEPLTIAARKTPKFRAGKALKEAVNR
54	27	MVFFNILFWKGPFFKRSFSRSFFGKVL
55	55	MLTDSAQFVQNQYAKQLKAKKEVSKKMLVLDSIKNTITTILDSSDKFIKLLYWKC
56	90	MANHSSAKKAARQTVKRTLISKKRSSAIKTFIKKVVHEISLGNKENANLALSVAQSKIMQGVKKNIIKLNTASRKISRLSKQIKSLNESK
57	35	MTSKELTGLNLRVSFATANIKIALISSLITLRLSL
58	74	MEEWEQGGSDICSKFGAVIERVRDGMQSKIRAFSAINTRMADHSKALDERERSLQQEKELLVKETGRVVETKSR

59	60	EDKRRRNTAASARFRAKKKEREHAMESRCKNLESKVGDLERECEALRRENGWLKGLVVGV
60	29	ICVDSVAELDKTGAENVESNEIVEGLDIS
61	29	MCYSSHYYYYYYYYYYYYYYYYYYSMSQK
62	65	MQISERKNALILHIDAFQELNLVLLIRAVLNGLIQFLSFYHLYALNNNLNKKILSFKKICFSRKI
63	59	MSCIFNGIQKSHYETDKNLTYCFMLKFYGLIVLIGLSWYLNIKFDYSLKRDDSLIVKAD
64	61	MLSRIFFLILCFKIHNAKINEVKKLEAKIARSIFENFLKEWTSFFKLSLKLFFHKIGILKF
65	83	MLTIRLALGGSKKRPFYHLNVTDSRNPRDGSHKEQVGFFNPIARGQEIRLSVNQERVAYWLSVGAQPSERVAQLLKDAAKAAA
66	73	MATQTVEGSSRSGPRRTTVGNLLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN
67	26	MNEYILRAAYIFYIYIYIIYYIYIYI
68	23	MYLYMLFFFFFFFFFFFLFLKFF
69	29	MHLALIKRYFMSLIFIMLSYIMIEIVKDK
70	33	MKIHLSPDEVNLLVYRYLVENGFVHTSFSFFNA
71	66	MTSEVLIYLILEIFFSASIYILNVHKIIELNITFFVIITLIYIYIYIYIELIFFFFFVSIIKANIR
72	20	MEYRFIFIFVHYYFIIFFYL
73	20	MVLLFYIFIFYISLLSYLFN
74	53	MMMIMMTIMMMIMITIMMMIIMWITTITTTMMITILIIIIKTHRIIIHIKTKE
75	68	MEVSLNNVNNDIKDVKEHITNFKEYVEKRIKDINNIMDMNRKEIDEKIEHICMNQKKLMGDFYPYKKN
76	16	FFFFFFFFFFFFVIVI
77	31	MPISISVIFSFFLFFFFFSFLLYFIQLLPYI
78	90	MKTNKRITIIGKVQGVFFRKSTKAKAEELDISGWVRNERDGSVFAEIEGNRHAVKAMEAWLSQGPPKALVENLLIEAGEEQGYTGFEIKE
79	56	MAHSNIWFSHPRKYGKGSRQCRVCANQGGIIRKYGLDICRQCFREKADAIGFVKNR
80	87	MATTERNLRKERIGKVVSDKMDKSITVAVERRVKHPIYGKFVAKTTKFMVHDENNECGSGDLVKISETRPLSKNKRWRLVEIIEKAK
81	58	MATITIKQIGSPIRRPESQRKILIGLGLNKMHKVVTRQDTPEVRGAIAKIPHLVTVID
82	40	MLFMVRRIQVFLVHFFKGRFFGQLFFDKFFLFRTFDVDNE
83	82	MEAATETHNDLVNLLNEKNIAITGHKAEAAQSAESVNKIIEENCQLKAEVERLQAEIRELRIKLWDATEEERKMNAGRLRDV
84	90	MNKTELIAKVAETSELTKKDATKAVDAVLDAISDALKEGDKVQLIGFGNFEVRERAARKGRNPQTGEEIEIASSKIPAFKPGKQLKDSIK
85	73	MNTLALAIGIIFGLAALGGAIGNSLVISRTIEGVARQPEARGSLMGLMLLGVGLVEAVPIIAVAVGFILYSQM
86	62	MTIAFQLAVFALIATSSILLISVPVVFASPDGWSSNKNVVFSGTSLWIGLVFLVGILNSLIY
87	86	MANIKSQIKRIKTNEKARQRNQSVKSSVKTAIRKFREAAESGDKAKAVELQQAAARALDKAASKGVIHANQAANKKSAMAKRVNQL
88	37	MKVQPSVKKICDKCKVIRRHGRIMVICENLRHKQRQG
89	85	MAHKKGASSSRNGRDSNPQYLGVKRYGGQLVNAGEILVRQRGTKFHPGLNVGRGGDDTLFALAAGTVEFGAKRGRKTVNIVPAEA
90	89	MALSTDEKKSILTEYGLHESDTGSPEAQVALLTKRIIGLTEHLKVHKHDHHSRRGLLLLVGRRRRLLNYVMKVDIERYRSLIQRLGLRR
91	32	LAALMDIIGATGATQVVYNHLYDPVSLVRDHR

92	24	MLTNIMYVTVYVTDQDRALEFYTE
93	87	MKTFDELFAELTDRARNRPEGSGTVEALDAGVHAQGKKVLEEAGEVWIAAEHESDDRLAEEISQLLYRVQVLMLGRGLTTEDVYRYL
94	47	MSKGKRTFQPNNRRRARTHGFRLRMRTRAGRAILSARRRKGREKLSA
95	44	MKRTYQPSVTRRKRTHGFRVRMKTRGGRAVINARRAKGRKRLAV
96	76	MLNYGRKTNTFTNCDSNVRDARFQLTGIYLRLLLIDCKLRNLNDFEYNLAKTSIKLENTKNCLKGNKKRYKNWITL
97	78	MAKVCQLTGKRPMSGNNVSHAQNKTRRRFLPNLQSRRFWVESENRWVRLRLSTNALRTIDKKGIDAVLAEMRANGQKV
98	96	MTKSELIAALMRKHPHLQLKDINLIVNTVFGAISKSLADNNRVELRGFGAFSIKERDPRVGRNPKTGEQVQVSKKFIPFFKTGKELHARINKARES
99	56	MAVQQNKPTRSKRGMRRSHDALTTAALSVDKVSGETHLRHHITADGYYRGRKVITK
100	27	AILYFLEKGAQPTVTVHDILRKAEFFK

References

1. You, K., Long, M., Wang, J. & Jordan, M. I. How does learning rate decay help modern neural networks? *arXiv preprint arXiv:1908.01878* (2019).
2. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517-520 (2011).
3. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600-612 (2004).
4. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439-D444 (2022).