UC Irvine

UC Irvine Previously Published Works

Title

Unraveling dynamic protein structures by two-dimensional infrared spectra with a pretrained machine learning model.

Permalink

https://escholarship.org/uc/item/9618v6hv

Journal

Proceedings of the National Academy of Sciences, 121(27)

Authors

Wu, Fan

Huang, Yan

Yang, Guokun

et al.

Publication Date

2024-07-02

DOI

10.1073/pnas.2409257121

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

PNAS RESEARCH ARTICLE | CHEMISTRY

Unraveling dynamic protein structures by two-dimensional infrared spectra with a pretrained machine learning model

Fan Wu[a,1], Yan Huang[a,1], Guokun Yang[a,1], Sheng Ye[b,2], Shaul Mukamel[c,2], and Jun Jiang[a,2]

Contributed by Shaul Mukamel; received May 9, 2024; accepted May 28, 2024; reviewed by Gregory D. Scholes and Martin T. Zanni

Dynamic protein structures are crucial for deciphering their diverse biological functions. Two-dimensional infrared (2DIR) spectroscopy stands as an ideal tool for tracing rapid conformational evolutions in proteins. However, linking spectral characteristics to dynamic structures poses a formidable challenge. Here, we present a pretrained machine learning model based on 2DIR spectra analysis. This model has learned signal features from approximately 204,300 spectra to establish a "spectrum-structure" correlation, thereby tracing the dynamic conformations of proteins. It excels in accurately predicting the dynamic content changes of various secondary structures and demonstrates universal transferability on real folding trajectories spanning timescales from microseconds to milliseconds. Beyond exceptional predictive performance, the model offers attention-based spectral explanations of dynamic conformational changes. Our 2DIR-based pretrained model is anticipated to provide unique insights into the dynamic structural information of proteins in their native environments.

ultrafast spectroscopy | protein dynamics | machine learning

Significance

Unraveling dynamic protein structures from two-dimensional infrared (2DIR) spectral signals presents a formidable task. By utilizing machine learning, we have correlated the intricate 2DIR spectroscopic features with the protein dynamic conformations, thereby tracing changes of secondary structure content along protein folding trajectories. This pretrained model facilitates universal transfer learning across diverse protein folding trajectories, offering valuable insights into the dynamic behavior of proteins and their biological function.

Protein structures are pivotal for elucidating their diverse biological functions. Significant experimental advancements have been made in the determination of protein structure (1–3). In recent years, AI has shown promising success in determining the lowest-energy state of proteins (4–18). Tools like AlphaFold2 (4, 5) and RoseTTAFold (6) can predict the three-dimensional structures of proteins from their amino acid sequences, while the integration of message passing neural network (MPNN) supplements the predictive capability of protein assemblies (8). The latest generative models can sample a broad variety of protein structures based on desired properties (13–16). These advancements have deepened our understanding of the lowest-energy static protein structures. Given that the dynamic characteristics of proteins ultimately shape their biological functions (19), integrating conformational dynamics information into machine learning (ML) training is therefore crucial for identifying dynamic protein structures that are relevant to biological processes (20–22).

Optical signals offer a unique window into protein dynamic responses. Two-dimensional infrared (2DIR) spectroscopy, based on femtosecond pulse sequences, has proven to be a powerful tool for determining protein structure and provides snapshots of protein folding events (23–30). However, unraveling dynamic protein structures from a series of 2DIR spectra present a formidable task, which typically requires days or weeks of manual analysis by a trained expert. Recent efforts in applying ML methods to extract structural information from spectroscopic signals (31–35) have paved the way for the potential of tracing protein dynamics. Therefore, it is imperative to develop data-driven ML protocols for automatically establishing correlations between protein 2DIR spectra and their dynamic conformations.

Here, we introduce a ML pretrained model utilizing the state-of-the-art transformer architecture (36), which effectively learns the signal features from 2DIR spectra and establishes a "spectrum-structure" correlation, thereby enabling the prediction of dynamic contents of various secondary structures in proteins. After being pretrained on approximately 204,300 simulated 2DIR spectra, the model shows exceptional transferability to real protein folding trajectories covering timescales from microseconds to milliseconds. Beyond its universal predictive power, this model also allows one to obtain signal interpretation for structure identification from the original spectra through attention maps. By tracing conformational changes in protein dynamics via 2DIR spectroscopy, this model can provide critical insights into the dynamic behavior of proteins and their biological function.

Results

Overall Schematic and ML Model Architecture. Our experimental workflow, illustrated in Fig. 1*A*, encompasses "construction of the 2DIR spectra dataset," "pretraining of the

Author affiliations: [a]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; [b]Anhui Provincial Engineering Research Center for Unmanned System and Intelligent Technology, School of Artificial Intelligence, Anhui University, Hefei 230601, Anhui, China; and [c]Department of Chemistry and of Physics & Astronomy, University of California, Irvine, CA 92697

Author contributions: F.W., S.Y., S.M., and J.J. designed research; F.W., Y.H., and G.Y. performed research; F.W., Y.H., and G.Y. contributed new reagents/analytic tools; F.W., Y.H., G.Y., S.Y., S.M., and J.J. analyzed data; and F.W., Y.H., G.Y., S.Y., S.M., and J.J. wrote the paper.

Reviewers: G.D.S., Princeton University; and M.T.Z., University of Wisconsin-Madison.

The authors declare no competing interest.

Copyright © 2024 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1] F.W., Y.H., and G.Y. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: yess@mail.ustc.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.2409257121/-/DCSupplemental.

Published June 25, 2024.

model," and "transfer learning to protein folding trajectories," all structured in a straightforward hierarchy. The 2DIR spectra pretraining dataset is generated by simulations, with detailed steps shown in Fig. 1*B*. We extracted 4,086 distinct homologous superfamily protein structures from the Orengo et al. developed CATH database, version 4.3 (37). The complete file index used in this study is available in *SI Appendix*. To capture dynamic information, 50 snapshots were taken for each trajectory at 1-ns intervals following molecular dynamics (MD) equilibrium, yielding a total of 204,300 protein conformations. The Hamiltonians for each protein conformation within the amide I spectral window were derived using semiempirical vibrational spectroscopic maps (38, 39), and the 2DIR signals were simulated employing the NISE code developed by Jansen et al. (40, 41). The contents of different secondary structures, including helix, strand, and coil were determined by utilizing the Stride program (42). The combined 2DIR spectra and secondary structure content formed the pretraining dataset.

The detailed architecture of our model is shown in Fig. 1*C*, based on the Vision Transformer (36). The 2DIR signals used for learning span the 1,575 to 1,725 cm^{-1} spectral window, where the horizontal and vertical axes correspond to coherence and detection frequency, respectively. Each spectrum is converted into a 224×224 matrix, segmented into 16×16 small patches, and subsequently flattened for model input. The position embeddings and an extra learnable regression [RGS] token to estimate the secondary structure contents were incorporated. The Transformer Encoder is composed of 12 alternating sets of multiheaded self-attention (MSA) layers and multilayer perceptron (MLP) layers, with pre-layernorm (LN) technique (43, 44) being utilized. The [RGS] token, as the output from the Transformer Encoder, is assumed to capture the essential spectral features of the 2DIR signal. It then proceeds through a MLP layer to predict the secondary structure contents of protein conformations.

The notable advantage of the pretrained model lies in its capability to show good predictive performance on entirely new, unseen datasets. In *Transfer Learning* at the bottom of Fig. 1*A*, the proficiency of our model in predicting the dynamic secondary structure contents during the reversible folding processes of Trp-cage and WW domain proteins serves as an illustrative example. The transferability to the folding trajectories of α3D and ubiquitin proteins will be demonstrated in the following. Additionally, the attention weights from the MSA layer can be extracted for visualization, showcasing which regions of the 2DIR spectra are most significant when predicting the contents of various secondary structures.

Pretraining Results on the CATH Dataset. The pretraining task is pivotal for determining the performance of the ensuing transfer learning process. During this phase, the model has to be trained on a vast and varied dataset to assimilate the intricate signal patterns and features of 2DIR spectra. Within our spectral dataset of 204,300 entries, 20% was allocated as a validation set, and 50 to

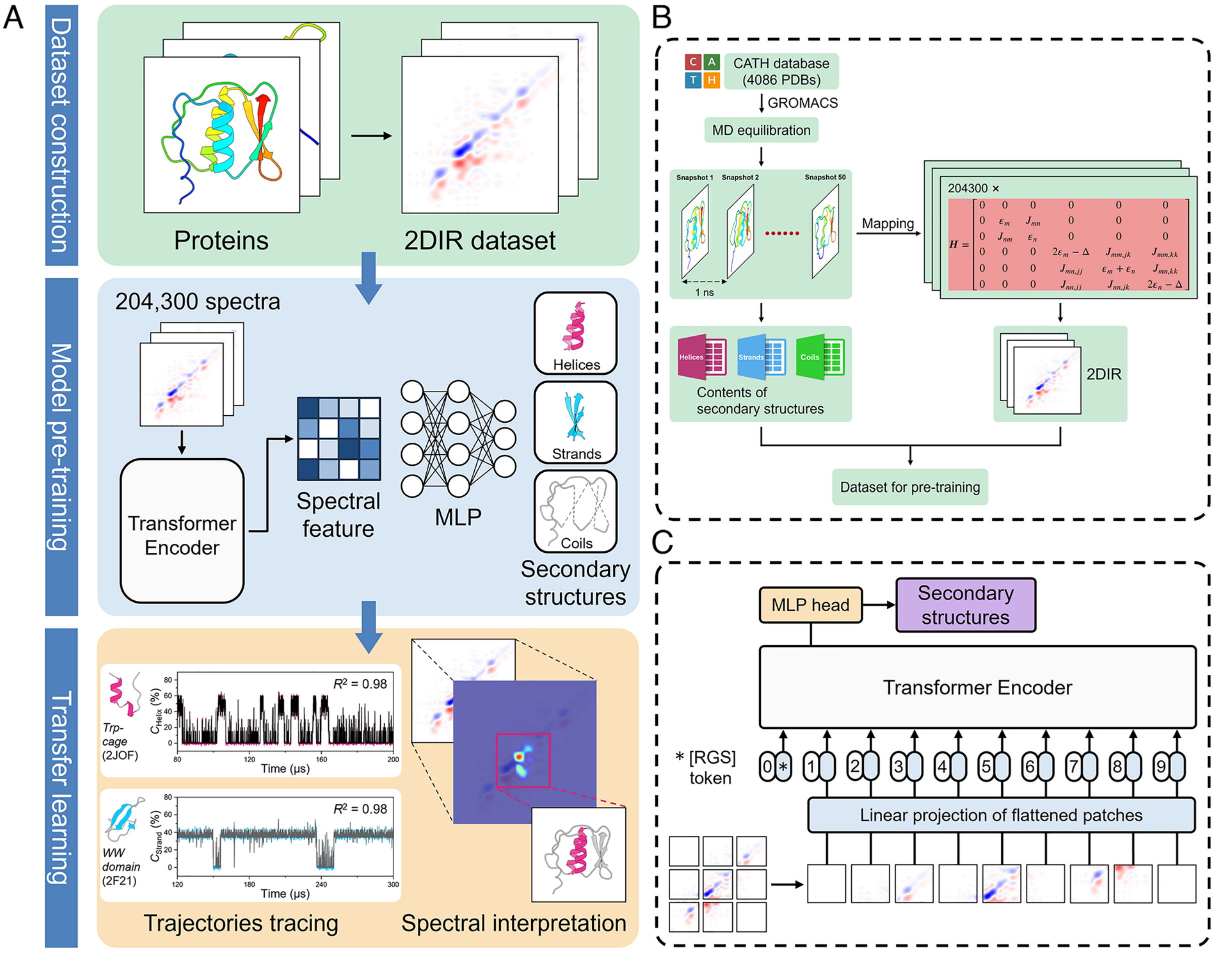

Fig. 1. Comprehensive workflow to predict dynamic protein secondary structures. (*A*) The overall experimental sequence flows from top to bottom, encompassing construction of the 2DIR spectra dataset, pretraining of the model, and transfer learning to protein folding trajectories. The C_{Helix} of Trp-cage and the C_{Strand} of WW domain respectively represent the contents of helices and strands within the overall secondary structure, expressed as a percentage. (*B*) For the pretraining dataset, protein entries underwent MD equilibration before each snapshot was taken to simulate 2DIR spectra and quantify secondary structure contents. (*C*) Detailed architecture of the ML model.

80% was employed as a training set to determine the minimum data amount required. Our findings suggest that utilizing at least 75% of the training set is necessary to achieve satisfactory prediction results, with a coefficient of determination (R^2) of 0.98 and a root mean squared error (RMSE) of 1.87 (*SI Appendix*, Fig. S1). The model has the capacity to concurrently predict the contents of three secondary structures for a single protein conformation. The error distribution of the validation set depicted in *SI Appendix*, Fig. S2 shows that the mean absolute error (MAE) for 96.54% of the predictions falls below 3.00. C_{Helix}, C_{Strand}, and C_{Coil} are represented as the content of the corresponding secondary structures, namely helix, strand, and coil. When examining the specific predictive contents for each secondary structure, as illustrated in the scatter plot of Fig. 2*A*, the R^2 value of helix, strand, and coil are 0.99, 0.98, and 0.98, respectively. Meanwhile, their RMSE are 1.51%,1.89%, and 1.71%, respectively, demonstrating the robust regression performance of the model.

To investigate the ability of the model to predict the dynamic content changes of secondary structures, we selected a total of 30 entries based on the protein molecular weight distribution in the CATH database. A 100-ns MD simulation was then performed, extending the initial 50-ns trajectory. The changes in secondary structure content over time were counted and the 2DIR spectra were calculated. As shown in *SI Appendix*, Fig. S3, the three major structural classes of proteins (mainly helix, mainly strand, and mixed helix/strand) in the database mostly contain about 20 to 200 amino acid residues. Thus, taking the mainly helix class as an example, for proteins with ≤200 amino acid residues, we began with a 20-residue protein and subsequently selected one every 30 residues, reaching a total of seven protein structures. For proteins

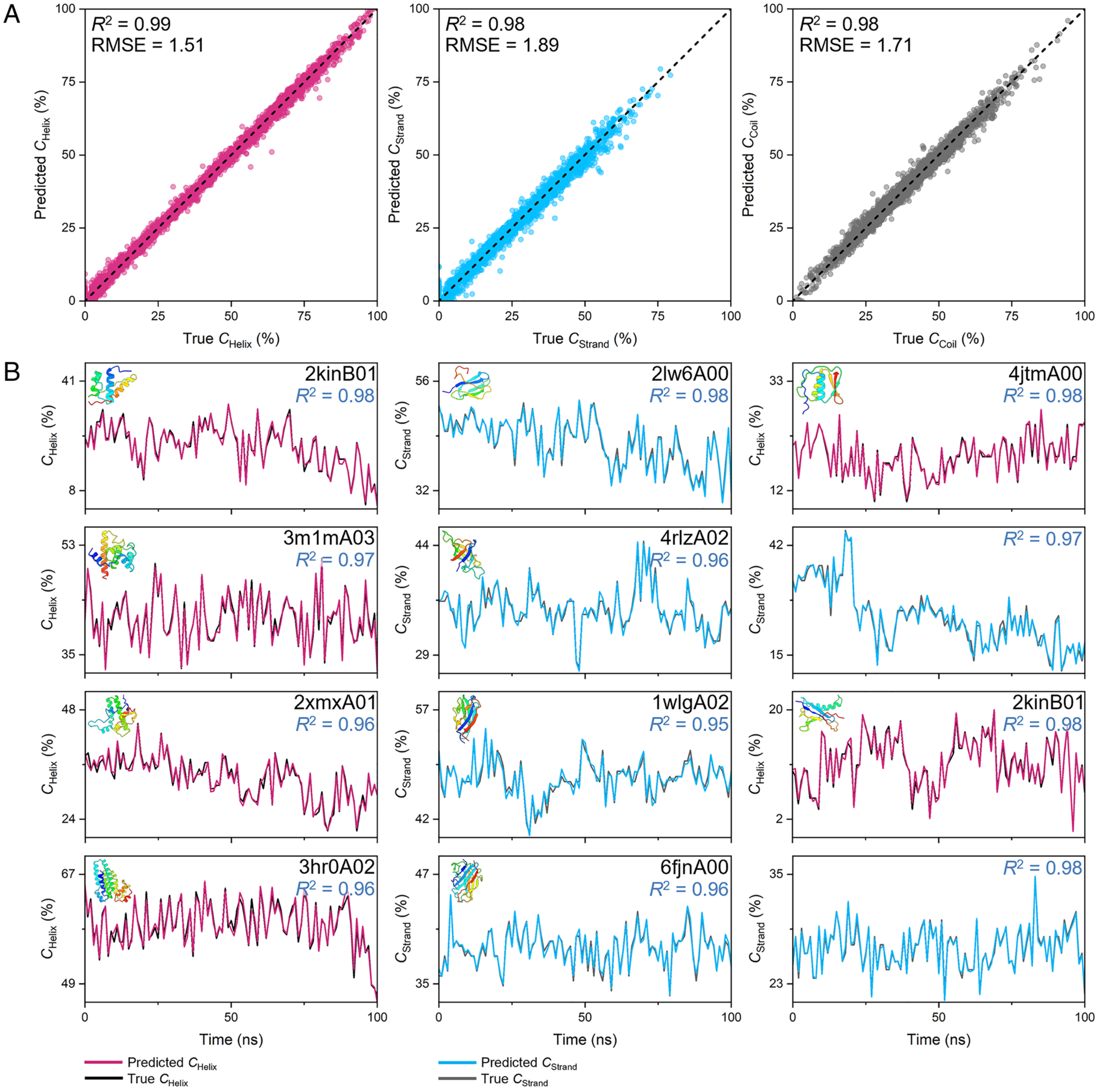

Fig. 2. Pretraining results on the CATH dataset. (*A*) Sequential scatter plots showing predictions for helix, strand, and coil contents, arranged from left to right. (*B*) Predictions for the dynamic secondary structure contents across three classes of protein trajectories: "mainly helix," "mainly strand," and mixed "helix/strand."

with more than 200 amino acid residues, we chose three proteins with about 250, 300, and 350 residues, to ensure a diverse and representative sampling from the original database. In Fig. 2*B*, for the three classes of proteins comprising approximately 80 to 170 amino acid residues, the pretrained model precisely predicts the dynamic content changes of the secondary structures, achieving an R^2 value of at least 0.95. From the complete predictions for the 30 trajectories depicted in *SI Appendix*, Fig. S4, there is an overall decreasing trend in the predictive performance of the model as the number of residues is incrementally raised from ~20 to 350, which is correlated with the overlap of the oscillator signals in the 2DIR spectra (45). Notably, even for the trajectory of a protein with up to 382 residues (CATH ID: 5o6hA00), the model still maintains good predictions with an R^2 value of at least 0.91. These findings demonstrate the effectiveness of our pretrained model in capturing the intricacies of 2DIR spectra, enabling it to provide accurate and reliable predictions on the secondary structure content of corresponding protein conformations.

Transfer Learning for Tracing Protein Folding Trajectories. In the previous section, the spectrum-structure dataset constructed through MD simulations is confined to the nanosecond scale due to computational constraints. However, protein folding processes typically occur over the microsecond to millisecond timescale (46, 47). The potential of our pretrained model to successfully transfer to datasets with broader timescales will thus be most valuable. As shown in Fig. 3, the reversible folding trajectories of four proteins with distinct secondary structure characteristics, simulated on the Anton supercomputer (48, 49), were utilized to evaluate the transferability of our model. Specifically, Trp-cage (PDB ID: 2JOF) and α3D (PDB ID: 2A3D) contain only helices, WW domain (PDB ID: 2F21) consists solely of strands (50), and ubiquitin (PDB ID: 1UBQ) features a mixture of both (51). For each complete folding trajectory, approximately 10,000 conformations were harvested at equal time intervals. The first 40% of this dataset was allocated for fine-tuning to update the pretrained weights, and this portion was randomly divided into training and validation sets in a 9:1 ratio. The remaining 60% served as the test set to verify transfer performance. The correlation curves between the predictive performance and the amount of fine-tuning data in the training set for different proteins are detailed in *SI Appendix*, Fig. S5 and Table S1.

For the Trp-cage folding trajectory in Fig. 3*A*, the prediction results of the test set start at 80 μs (see *SI Appendix*, Fig. S6 for the complete folding trajectory). Here, high or low variations in the helix content, correspond to the Trp-cage protein in a folded or unfolded state, respectively. Alternating folding and unfolding events can be clearly observed within the predicted 120-μs trajectory. The pretrained model, which was not fine-tuned using any 2DIR spectra, had an R^2 value of only 0.92 on the validation set. The predictive accuracy improves with the increasing amount of feeding data. Notably, when the amount of data reaches ~1800, an optimal performance is achieved, with R^2 value equals to 0.98 (*SI Appendix*, Fig. S5*A*). We highlighted five characteristic conformations marked by significant helix content variations and zoomed in to display the prediction details at surrounding times. These results reveal that our model can precisely trace the dynamic secondary structure content changes of Trp-cage along its folding process. The helices of the shown conformations are colored in red to enhance the visual representation, whereas the tertiary structures were not

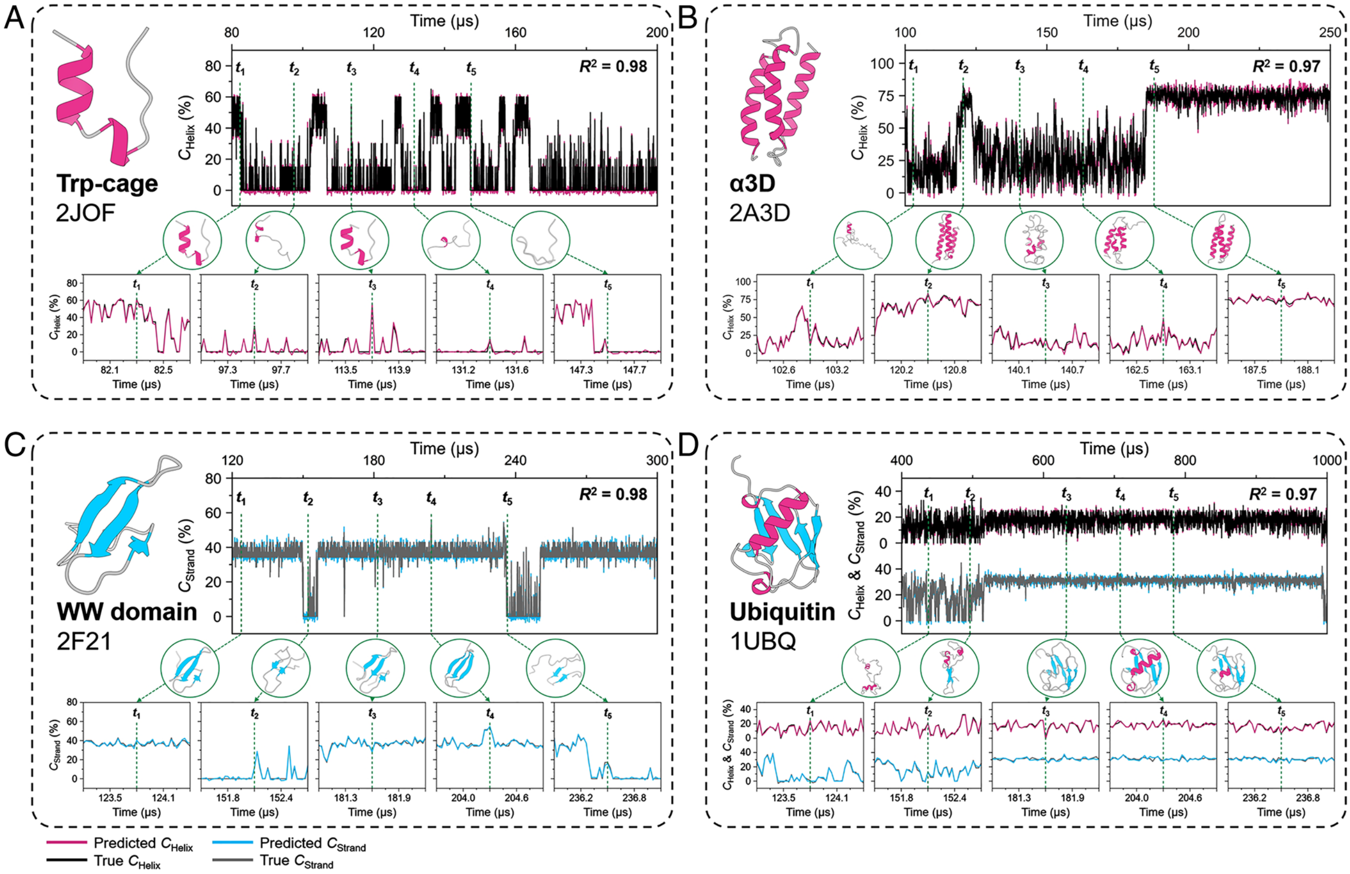

Fig. 3. Transfer learning results of protein folding trajectories simulated on the Anton supercomputer. Panels (*A*), (*B*), (*C*), and (*D*) correspond to the prediction results of Trp-cage, α3D, WW domain, and ubiquitin, respectively. In each panel, we highlight five conformations with significant changes in secondary structure contents, labeled t_1–t_5, and further display prediction details at surrounding times.

predicted. The α3D protein, similar to Trp-cage in containing only helices, possesses about three times the number of amino acid residues, leading to more complex 2DIR signals. At the same time, folding and unfolding events are less frequent within the prediction time span (Fig. 3*B*), increasing the prediction challenge. However, even without fine-tuning, the model provides an R^2 value of 0.84. Increasing the fine-tuning data to ~3,000 led to satisfactory predictions, with R^2 value reaching 0.97 (*SI Appendix*, Fig. S5*B*). The WW domain primarily consists of strand-type secondary structures. The 2DIR spectral features associated with strands were effectively captured during the previous pretraining step, facilitating quite a smooth transfer learning process. The R^2 values before and after fine-tuning were 0.90 and 0.98, respectively, with a fine-tuning data usage of ~2,400 (Fig. 3*C* and *SI Appendix*, Fig. S5*C*). Ubiquitin, widely found in eukaryotic cells, has 76 amino acid residues and features both helix and strand secondary structures. Fig. 3*D* shows that our fine-tuned model can accurately predict the helices and strands content changes for each conformation throughout the folding process. Upon using nearly all of the spectra from the training set, the R^2 value increased from 0.85 to 0.97 (*SI Appendix*, Fig. S5*D*). It is noteworthy that the CHARMM22* force field slightly underestimates the stability of helices in the folded state of ubiquitin (52), leading to a greater variance in helix content compared to the strand. After the model was pretrained on a vast and varied spectral dataset to optimize its weights, only a minimal amount of additional data was required for fine-tuning to achieve exceptional predictive accuracy on entirely new datasets. Please note that for these four trajectories, training the same model from scratch resulted in the R^2 values of only 0.71, 0.67, 0.67, and 0.63, respectively (*SI Appendix*, Table S1). We thus believe that this pretrained model can facilitate universal transfer learning for predicting the dynamic secondary structure content changes of protein trajectories. The code and weights files of our model are freely accessible on our GitHub repository (53) (https://github.com/SaintCloud-0013/2DIR-ML), and future downloads and verifications are highly appreciated.

Attention-Based Interpretation of Spectral Signals. Beyond its universal transfer learning capability, our model offers a significant advantage by providing visual interpretations on the original spectra for its prediction results. By analyzing the attention maps corresponding to different secondary structures, we gain insights into the characteristic regions of 2DIR spectra that the model focuses on when making predictions. After integrating the attention weights across all the MSA layers (54), Fig. 4 provides the digital interpretation of the model for ubiquitin conformations during its folding process. Sequentially displayed from top to bottom are the corresponding 2DIR spectra, helix attention maps, and strand attention maps. It is evident that the transition signals at different vibrational energy levels are effectively captured (in the 2DIR spectra, blue and red denote transitions of the vibrational quantum numbers $v = 0 \rightarrow 1$ and $v = 1 \rightarrow 2$, respectively). Both diagonal and off-diagonal (cross) peak regions contribute to the structure identification (33). There is a clear distinction in the attention distribution for the helix and strand. The model focuses primarily on the spectral region around 1,650 cm^{-1} for the helix, while for the strand, the attention is mainly on the region around 1,630 cm^{-1}, with a minor contribution around 1,680 cm^{-1}. This observation aligns with generally accepted understanding (55), suggesting that the model effectively captures the relevant features from the original 2DIR spectra during its learning process. In addition, we find that as the helix content increases, the corresponding attention undergoes a redshift, a similar redshift is noted with an increase in strand

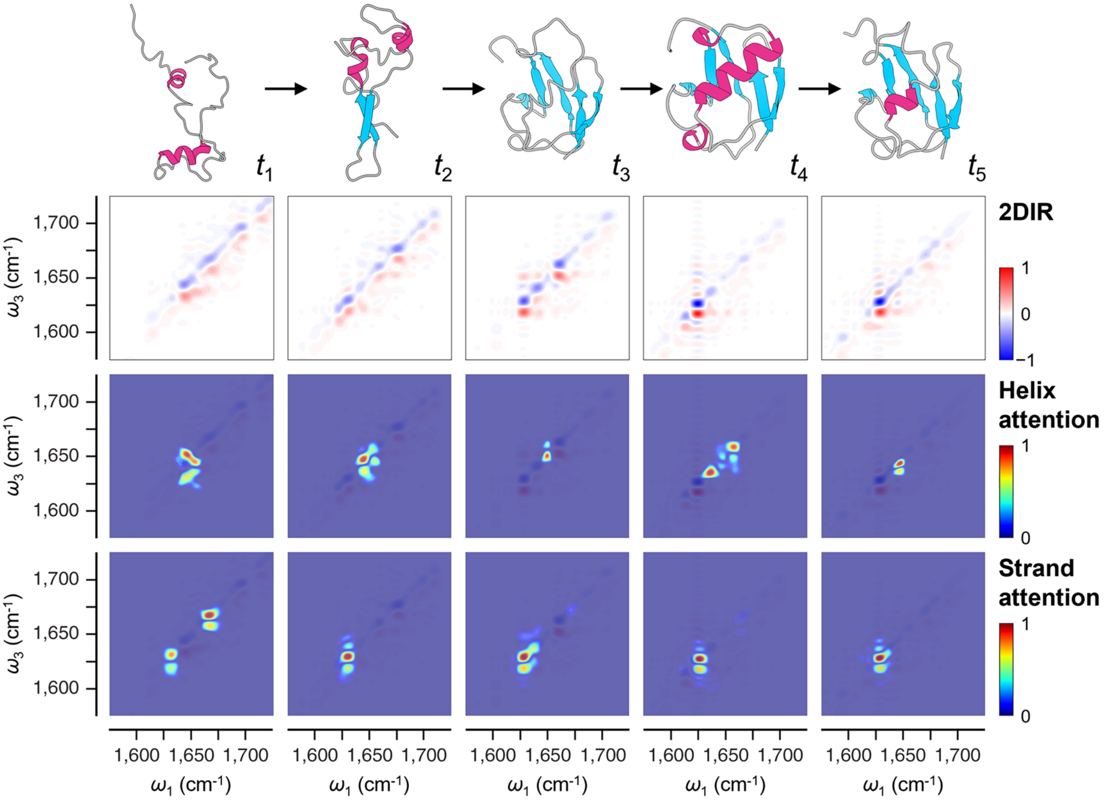

Fig. 4. Attention maps of different secondary structures for ubiquitin conformations. The figure displays, from top to bottom, the conformations along the ubiquitin folding trajectory, followed by the corresponding 2DIR spectra, helix attention maps, and strand attention maps.

content, consistent with previous work (56–59). This attention-based spectral interpretation enhances the transparency of the ML black box, fostering greater trust in the pretrained model and making it more reliable and effective for practical applications.

Discussion

In summary, we have developed a pretrained ML model based on the state-of-the-art transformer architecture, which captures features of 2DIR spectra and establishes a spectrum-structure correlation. This facilitates a universal prediction of dynamic secondary structure content changes along protein trajectories. During the pretraining phase, about 204,300 2DIR spectra were used to optimize the model weights. The pretrained model demonstrates high fidelity in predicting helix, strand, and coil contents of protein conformations in the validation set, with coefficient of determination R^2 values of 0.99, 0.98, and 0.98, respectively. Remarkably, in the transfer learning process, even without fine-tuning, the model still achieved fair R^2 values over 0.84 in predicting dynamic content changes for real folding trajectories of four typical proteins, namely Trp-cage, α3D, WW domain, and ubiquitin. Following a time-efficient and data-minimal fine-tuning step, R^2 values reached satisfactory 0.98, 0.97, 0.98, and 0.97, respectively, underscoring its universal transferability and high performance. Moreover, the model can provide an attention-based interpretation of spectroscopic signals by elucidating the dynamics of different secondary structures along protein trajectories. By using 2DIR spectroscopy to trace conformational structures during protein dynamics, this model can provide useful insights into the dynamic behavior of proteins in their biological functions. With the significant boost in computing power, the pretraining protocol detailed in this article should provide a powerful methodology for the protein dynamics community, and trigger the development and applications of ML models in related fields.

Materials and Methods

As described in the previous sections, 2DIR spectra can serve as a tool for tracing the dynamic conformations of proteins. However, extracting oscillator signals from these spectra and converting them into quantitative structural information presents a major challenge. Our aim is to establish a spectrum–structure correlation through ML, facilitating the prediction of protein secondary structure contents. Additionally, by leveraging pretraining techniques in combination with a vast and varied dataset, we strive to achieve universal transferability across various protein folding trajectories.

Dataset Construction. The 2DIR spectra used in both the pretraining and transfer learning datasets were generated through simulations derived from protein PDB files. The computational protocol of these spectra is detailed below. The initial structures for constructing the pretraining dataset were sourced from the CATH database v4.3 (37), developed by Orengo et al. This database is organized according to the protein secondary structure categories. From it, we extracted 4,086 distinct homologous superfamily protein structures using three major categories: "Mainly Alpha," "Mainly Beta," and "Alpha Beta." The complete index of files can be found in *SI Appendix*. To capture the dynamic conformations, MD simulations were conducted for each protein using the Gromacs (60) software, with detailed settings provided below. For each trajectory, 50 snapshots were harvested at 1-ns intervals following the NVT and NPT equilibria. The entire pretraining dataset comprises a total of 4,086 × 50 = 243,000 spectra. The transfer learning dataset features protein reversible folding trajectories, simulated on the Anton supercomputer across timescales from microseconds to milliseconds. This collection encompasses structures such as the Trp-cage, α3D, WW domain, and ubiquitin. For each trajectory, we sampled approximately 10,000 conformations at equal time intervals to compute the 2DIR spectra. The secondary structure assignments for protein conformations were determined by calculating hydrogen bond energies (61) anrd mainchain dihedral angles from the atomic coordinates of the snapshots, using the Stride (42) program.

2DIR spectra simulations. We employed the Frenkel exciton Hamiltonian within the amide I spectral window:

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

Here, $b_i^{\dagger}$ and $\boldsymbol{b}_i$ represent the Bosonic creation and annihilation operators for individual peptide unit, respectively. ω_i is the vibrational frequency of the local mode, J_{ij} denotes the coupling between two local modes, and Δ_i refers to the anharmonicity. The frequency of an oscillator is calculated using the Skinner map (62), where the environment of the C and N atoms is considered:

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

ω_{map}, $P_{i,map}$, and $\mathbf{E}_{i,map}$ are, respectively, the vibrational frequency, electric potential, and electric field, as predefined in the map. P_i and $\mathbf{E}_i$ were computed as follows:

$$P = \sum_{j} \frac{q_j}{|\mathbf{r}_j|},$$

$$E_x = \sum_{j} \frac{q_j}{|\mathbf{r}_j|^3} (\mathbf{r}_j \cdot \hat{\mathbf{x}}).$$

E_x is the electric field in the x-direction, E_y and E_z can be calculated in a similar manner. Additionally, frequency shifts for each neighboring amide groups are incorporated based on Ramachandran angles (63).

The coupling between the neighboring and nonneighboring oscillators is determined through the transition charge coupling (TCC) (64) and glycine dipeptide (GLDP) methods (65), respectively:

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

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

The dipole moment of each oscillator was obtained from the relative positions between C, O, and N atoms (66):

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

The 2DIR spectrum is the imaginary part of the sum of the rephasing (photon echo, $\mathbf{k}_{\text{I}} = -\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$) and nonrephasing signals ($\mathbf{k}_{\text{II}} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$), where $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ are the wave vectors of the incoming infrared fields with time delays of t_1, t_2, and t_3.

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

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

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

where ω_1 and ω_3 represent the frequencies of t_1 and t_3 after Fourier transformation, respectively. GB, SE, and EA represent the contributions from different Liouville space pathways, known as ground-state bleach, stimulated emission and excited-state absorption, respectively.

MD simulations. The protein PDB entries within the CATH database, were subjected to MD sampling using Gromacs 2018. In the aqueous environment, the all-atom OPLS-AA/L force field was utilized in combination with TIP3P water molecules, and Na^+ or Cl^- ions were used to balance the charge of the system. To avoid the influence of periodic images, the protein molecule was centrally positioned in a cubic box, ensuring at least 1.0 nm distance from the edges. A 50,000-step energy minimization was applied to eliminate steric clashes or inappropriate geometries. This was followed by two-phases NVT and NPT equilibration steps, each lasting 100 ps. Production dynamics was then performed for a period of 50 ns with a 2-fs timestep. The system was maintained at 373 K and 1 atmosphere using v-rescale Berendsen thermostat (67) and Parrinello–Rahman barostat (68), respectively. Fifty snapshots were collected every 1 ns along each production trajectory.

Model Architecture. In the 2DIR spectra, information on frequencies and couplings of vibrational modes is stored in a contour plot with excitation frequency (ω_1) and detection frequency (ω_3) serving as the coordinate axes. A self-attention-based transformer is adopted to capture the features of these two-dimensional matrices. The model primarily consists of three components: segmented embeddings of 2DIR matrices, a backbone network of Transformer Encoder, and an MLP head for regression tasks. The input 2DIR signals covering the 1,575 to 1,725 cm^{-1} spectral window are resized to 224×224 matrices and then segmented into 16×16 small patches. Each patch is linearly embedded into the network, incorporating position embeddings and an extra learnable regression [RGS] token for numerical regression of the secondary structure contents. The Transformer Encoder comprises 12 alternating sets of MSA and MLP layers, with LN technique being utilized. The [RGS] token, processed by the Transformer Encoder, captures the essential spectral features of 2DIR signal. It then proceeds through an MLP layer to predict the secondary structure contents of protein conformations.

Pretraining and fine-tuning configurations. During the pretraining phase, we utilized the AdamW (69) optimizer, configured with a learning rate of 10^{-4} and a weight decay of 10^{-2}. The pretraining dataset was randomly divided into training and validation sets in an 8:2 ratio. The model was trained for 100 epochs with a batch size of 256. A learning rate warmup strategy was implemented for the initial 5% of the total epochs, followed by a linear decay to zero for the remaining.

For the fine-tuning step, pretrained model weights are loaded to optimize for predicting specific protein folding trajectory. The optimizer and learning rate strategy are consistent with the pretraining step, and the weights across all layers are updated. The 2DIR spectra calculated from the first 40% snapshots of the complete folding trajectory were used to construct the fine-tuning dataset, which was randomly split into training and validation sets in a ratio of 9:1. The remaining 60% is served as a test set to evaluate the performance of the fine-tuned model. Given the relatively small size of the dataset used for the fine-tuning, the model was trained with a batch size of 32 during 40 epochs.

Data, Materials, and Software Availability. Protein files with distinct secondary structure characteristics are available from the CATH database (37) (https://www.cathdb.info/browse/tree). Protein folding trajectories are available from D. E. Shaw Research (50, 51) (https://www.deshawresearch.com/resources.html). Code is available at our GitHub repository (53) (https://github.com/SaintCloud-0013/2DIR-ML). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. We thank David E. Shaw and his colleagues for providing data and assistance. This work was financially supported by Hefei Comprehensive National Science Center, the Innovation Program for Quantum Science and Technology (2021ZD0303303), the Chinese Academy of Sciences (CAS) Project for Young Scientists in Basic Research (YSBR-005), and the National Natural Science Foundation of China (22025304, 22033007, and 22303091). We thank the Hefei Advanced Computing Center for providing numerical computations; and the University of Science and Technology of China (USTC) Supercomputing Center for providing computational resources for this project. S.Y. acknowledges the support of the National Natural Science Foundation of China (22203001, 62236002, and 61921004), the University Synergy Innovation Program of Anhui Province (GXXT-2022-062), and the Natural Science Foundation of Anhui Province (2208085Y04). S.M. acknowledges the support of the National Science Foundation (CHE-2246379). S.M. is a senior fellow at the Hagler Institute for Advanced Study at Texas A&M University.

1. G. Brändén, R. Neutze, Advances and challenges in time-resolved macromolecular crystallography. *Science* **373**, eaba0954 (2021).
2. S. Ahlawat, K. R. Mote, N. A. Lakomek, V. Agarwal, Solid-state NMR: Methods for biological solids. *Chem. Rev.* **122**, 9643–9737 (2022).
3. Y. Cheng, Single-particle cryo-EM—How did it get here and where will it go. *Science* **361**, 876–880 (2018).
4. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. K. Tunyasuvunakool *et al.*, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
6. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
7. I. R. Humphreys *et al.*, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).
8. J. Dauparas *et al.*, Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
9. S. Mosalaganti *et al.*, AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **376**, eabm9506 (2022).
10. A. Fryszkowska *et al.*, A chemoenzymatic strategy for site-selective functionalization of native peptides and proteins. *Science* **376**, 1321–1327 (2022).
11. I. D. Lutz *et al.*, Top-down design of protein architectures with reinforcement learning. *Science* **380**, 266–273 (2023).
12. M. L. Hekkelman, I. de Vries, R. P. Joosten, A. Perrakis, AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).
13. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, in press. https://doi.org/10.1038/s41586-024-07487-w.
14. R. Krishna *et al.*, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
15. J. L. Watson *et al.*, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
16. J. B. Ingraham *et al.*, Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
17. L. Lu *et al.*, De novo design of drug-binding proteins with predictable binding energy and specificity. *Science* **384**, 106–112 (2024).
18. M. Baek *et al.*, Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**, 117–121 (2024).
19. S. Bhatia, J. B. Udgaonkar, Heterogeneity in protein folding and unfolding reactions. *Chem. Rev.* **122**, 8911–8935 (2022).
20. H. K. Wayment-Steele *et al.*, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
21. W. Lu *et al.*, DynamicBind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat. Commun.* **15**, 1071 (2024).
22. E. Rennella, D. D. Sahtoe, D. Baker, L. E. Kay, Exploiting conformational dynamics to modulate the function of designed proteins. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2303149120 (2023).
23. 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).
24. A. Ghosh, J. S. Ostrander, M. T. Zanni, Watching proteins wiggle: Mapping structures with two-dimensional infrared spectroscopy. *Chem. Rev.* **117**, 10726–10759 (2017).
25. J. P. Kraack, P. Hamm, Surface-sensitive and surface-specific ultrafast two-dimensional vibrational spectroscopy. *Chem. Rev.* **117**, 10623–10664 (2017).
26. C. R. Baiz *et al.*, Vibrational spectroscopic map, vibrational spectroscopy, and intermolecular interaction. *Chem. Rev.* **120**, 7152–7218 (2020).
27. N. T. Hunt, Using 2D-IR spectroscopy to measure the structure, dynamics, and intermolecular interactions of proteins in H_2O. *Acc. Chem. Res.* **57**, 685–692 (2024).
28. 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).
29. R. Hu *et al.*, Ultrafast two-dimensional infrared spectroscopy resolved a structured lysine 159 on the cytoplasmic surface of the microbial photoreceptor bacteriorhodopsin. *J. Am. Chem. Soc.* **144**, 22083–22092 (2022).
30. M. J. Ryan, L. Gao, F. I. Valiyaveetil, A. A. Kananenka, M. T. Zanni, Water inside the selectivity filter of a K^+ ion channel: Structural heterogeneity, picosecond dynamics, and hydrogen bonding *J. Am. Chem. Soc.* **146**, 1543–1553 (2024).
31. V. Barone *et al.*, Computational molecular spectroscopy. *Nat. Rev. Methods Primers* **1**, 38 (2021).

32. 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).
33. H. Ren *et al.*, Machine learning recognition of protein secondary structures based on two-dimensional spectroscopic descriptors. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2202713119 (2022).
34. P. Klukowski, R. Riek, P. Güntert, Time-optimized protein NMR assignment with an integrative deep learning approach using AlphaFold and chemical shift prediction. *Sci. Adv.* **9**, eadi9323 (2023).
35. T. Yang *et al.*, Catalytic structure design by AI generating with spectroscopic descriptors. *J. Am. Chem. Soc.* **145**, 26817–26823 (2023).
36. A. Dosovitskiy *et al.*, An image is worth 16×16 words: Transformers for image recognition at scale. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2010.11929 (Accessed 22 October 2020).
37. I. Sillitoe *et al.*, CATH: Increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
38. H. Kim, M. Cho, Infrared probes for studying the structure and dynamics of biomolecules. *Chem. Rev.* **113**, 5817–5847 (2013).
39. B. Blasiak, C. H. Londergan, L. J. Webb, M. Cho, Vibrational probes: From small molecule solvatochromism theory and experiments to applications in complex systems. *Acc. Chem. Res.* **50**, 968–976 (2017).
40. T. Jansen, J. Knoester, Nonadiabatic effects in the two-dimensional infrared spectra of peptides: Application to alanine dipeptide. *J. Phys. Chem. B* **110**, 22910–22916 (2006).
41. T. L. Jansen, J. Knoester, Waiting time dynamics in two-dimensional infrared spectroscopy. *Acc. Chem. Res.* **42**, 1405–1411 (2009).
42. D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).
43. R. Xiong *et al.*, On layer normalization in the transformer architecture. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2002.04745 (Accessed 12 February 2020).
44. L. Liu, X. Liu, J. Gao, W. Chen, J. Han, Understanding the difficulty of training transformers. arXiv [preprint] (2020). https://doi.org/10.48550/arXiv.2004.08249 (Accessed 17 April 2020).
45. W. Zhuang, T. Hayashi, S. Mukamel, Coherent multidimensional vibrational spectroscopy of biomolecules: Concepts, simulations, and challenges. *Angew. Chem. Int. Ed.* **48**, 3750–3781 (2009).
46. H. S. Chung, K. McHale, J. M. Louis, W. A. Eaton, Single-molecule fluorescence experiments determine protein folding transition path times. *Science* **335**, 981–984 (2012).
47. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
48. D. E. Shaw *et al.*, Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
49. D. E. Shaw *et al.*, "Millisecond-scale molecular dynamics simulations on Anton" in *SC'09: International Conference for High Performance Computing, Networking, Storage and Analysis* (ACM, 2009), pp. 1–11.
50. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
51. S. Piana, K. Lindorff-Larsen, D. E. Shaw, Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5915–5920 (2013).
52. S. Piana, K. Lindorff-Larsen, D. E. Shaw, How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–49 (2011).
53. F. Wu *et al.*, Unraveling dynamic protein structures by two-dimensional infrared spectra with a pretrained machine learning model. Github. Available at https://github.com/SaintCloud-0013/2DIR-ML. Deposited 2 May 2024.
54. S. Abnar, W. Zuidema, Quantifying attention flow in transformers. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2005.00928 (Accessed 2 May 2020).
55. M. D. Fayer, *Ultrafast Infrared Vibrational Spectroscopy* (Taylor & Francis, Boca Raton, FL, 2013), p. 475.
56. Z. Ganim *et al.*, Amide I two-dimensional infrared spectroscopy of proteins. *Acc. Chem. Res.* **41**, 432–441 (2008).
57. Z. Lai, N. K. Preketes, S. Mukamel, J. Wang, Monitoring the folding of Trp-cage peptide by two-dimensional infrared (2DIR) spectroscopy. *J. Phys. Chem. B* **117**, 4661–4669 (2013).
58. H. S. Chung, M. Khalil, A. W. Smith, Z. Ganim, A. Tokmakoff, Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 612–617 (2005).
59. H. S. Chung, Z. Ganim, K. C. Jones, A. Tokmakoff, Transient 2D IR spectroscopy of ubiquitin unfolding dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14237–14242 (2007).
60. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
61. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
62. 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).
63. T. la Cour Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst, J. Knoester, Modeling the amide I bands of small peptides. *J. Chem. Phys.* **125**, 44312 (2006).
64. P. Hamm, S. Woutersen, Coupling of the amide I modes of the glycine dipeptide. *Bull. Chem. Soc. Japan* **75**, 985–988 (2002).
65. 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).
66. 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).
67. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
68. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
69. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1711.05101 (Accessed 14 November 2017).