JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau

Article

Accurate Machine Learning Prediction of Protein Circular Dichroism Spectra with Embedded Density Descriptors

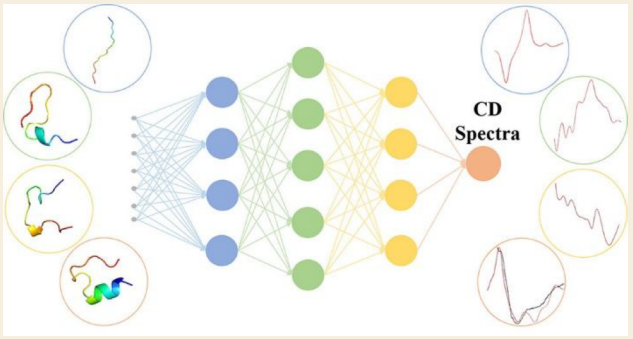
Luyuan Zhao,[#] Jinxiao Zhang,[#] Yaolong Zhang,[#] Sheng Ye, Guozhen Zhang, Xin Chen, Bin Jiang,* and Jun Jiang*

Cite This: *JACS Au* 2021, 1, 2377–2384

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A data-driven approach to simulate circular dichroism (CD) spectra is appealing for fast protein secondary structure determination, yet the challenge of predicting electric and magnetic transition dipole moments poses a substantial barrier for the goal. To address this problem, we designed a new machine learning (ML) protocol in which ordinary pure geometry-based descriptors are replaced with alternative embedded density descriptors and electric and magnetic transition dipole moments are successfully predicted with an accuracy comparable to first-principle calculation. The ML model is able to not only simulate protein CD spectra nearly 4 orders of magnitude faster than conventional first-principle simulation but also obtain CD spectra in good agreement with experiments. Finally, we predicted a series of CD spectra of the Trp-cage protein associated with continuous changes of protein configuration along its folding path, showing the potential of our ML model for supporting real-time CD spectroscopy study of protein dynamics.



KEYWORDS: *machine learning, electronic circular dichroism spectroscopy, transition dipole moment, protein dynamics, embedded density descriptors*

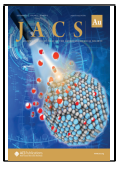
1. INTRODUCTION

Protein structure determination is crucial for understanding many biological functions.[1−3] Especially, tracking protein structural variation in real time is desirable for exploring underlying mechanisms.[4−6] Very recently, AlphaFold, the artificial intelligence (AI) program developed by DeepMind, has achieved accurate prediction of the folded structure of a protein based on its primary structure, i.e., amino acid sequence.[7] Despite the tremendous process in protein tertiary structure prediction, people still have little knowledge about the path through which a protein evolves from one configuration to another. Yet the structure information on protein dynamics is the key to understanding how they function and how their function can be modulated. Spectroscopy techniques may shed light into protein dynamics by directly probing protein structures and couplings along the dynamic process.[8−12] Among many spectroscopic techniques, electronic circular dichroism (CD) holds advantages of ease of operation and high sensitivity to subtle structural changes.[13,14] Aided by an advanced light source, it could be a powerful probe of real-time protein structure determination.

Spectroscopic measurement needs to be coupled with a rapid theoretical interpretation means, to realize real-time interpretation of structural variations. However, the huge computational cost required for accurately simulating protein spectra at the quantum chemistry (QC) level has long been a painful obstacle for such an ambitious goal. The rapidly developing machine learning (ML) techniques which have been successfully applied to physical and biological sciences to circumvent the challenge of solving complicated structure−property relationships offer an opportunity for addressing the longstanding problem. Recently, ML has been applied in various spectroscopic simulations,[15−17] as well as protein structure prediction.[7,18,19] Along this track, we have also developed ML tools for predicting protein infrared (IR) and ultraviolet (UV) spectra.[20−22]

Proteins with different secondary structure profiles have distinctive signatures in CD spectra, making them useful in studying protein dynamics such as folding and binding events.[13,23−25] The CD spectrum in the far UV is based on the energy of the electronic transitions that the peptide bond contributions dominate. And two key parameters of electric and magnetic transition dipole moments of peptide bonds are

Received: October 8, 2021
Published: November 25, 2021



https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377–2384

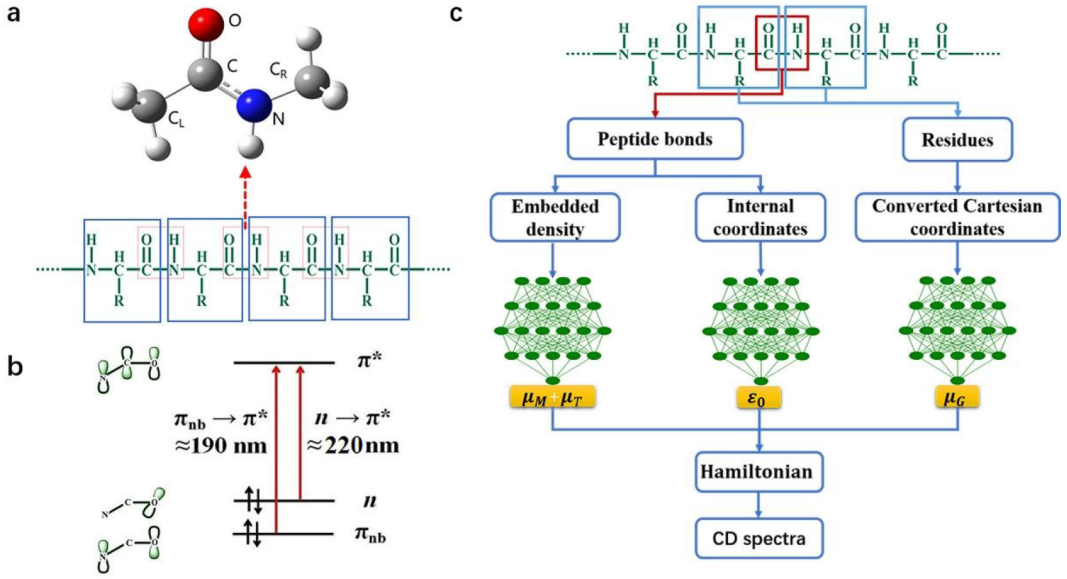


Figure 1. (a) NMA structure and protein structure. (b) Valence molecular orbitals and two electronic transitions of the peptide bond which are n → π^* or $\pi \rightarrow \pi^*$ transitions. (c) Machine learning protocol for predicting protein CD spectra.

necessary to simulate the CD spectra of different secondary structures.[26−28] However, accurate prediction of these two physical quantities by ML methods is difficult for two reasons: (1) they are vectors of multiple coordinate dependent components, which are covariant with system rotation or twisting, and (2) their vector directions are essentially determined by the corresponding types of electronic transitions, which are not well described by regular ML models and descriptors extracted directly from structural parameters. This difficulty can be overcome by the recently proposed tensorial embedded atom neural network (EANN) model.[29,30] Accordingly, we have developed a set of embedded density descriptors, to learn tensorial properties in a fully symmetry-adapted way. This descriptor considers each atom as impurities embedded in the electron gas generated by surrounding atoms and is constructed by the square of the linear combination of atomic orbitals from adjacent atoms. Consequently, the obtained descriptor is invariant with respect to the overall translation and rotation and permutation of identical atoms. Combining the virtual ML output obtained from the embedded density descriptor with atomic coordinate vectors by multiplication, we can get the symmetry-conservative tensors for describing electric dipole moments of the *N*-methylacetamide (NMA) molecule.[29]

In this study, we have constructed a ML protocol based on novel embedded density descriptors to predict the CD spectra of proteins with a comparable accuracy to density functional theory (DFT) calculations while significantly faster than the latter. The embedded density descriptors can learn both the electric and magnetic transition dipole moments of peptide bonds well, which integrate complex information on molecular chirality, atomic structure, electron and spin density, wave function transition, and so on. The simulated CD spectra by our ML model not only are in good agreement with experiment results but also help to distinguish various proteins with different secondary structure profiles. Moreover, we have successfully mapped the folding process of a protein with simulated CD spectra, showing the potential of our ML protocol in facilitating real-time observation of protein dynamics using CD spectra in the future.

2. THEORY AND COMPUTATION DETAIL

Proteins are composed of peptide bonds and amino acid residues (Figure 1a), and most CD responses in the far UV region come from two electronic excitations of the peptide bond: n → π^* transition around 220 nm and $\pi \rightarrow \pi^*$ transitions around 190 nm (Figure 1b). The model Hamiltonian of exciton can be constructed based on the Frenkel exciton model.[31−33] It is necessary to calculate the excitation energy ε_{ma} of the peptide bonds and the resonance coupling $J_{ma,nb}$ between excited states with the dipole approximation:[34,35]

$$\varepsilon_{ma} = \varepsilon_{0,ma} + \sum_{k} \frac{1}{4\pi\varepsilon\varepsilon_0}\left(\frac{\boldsymbol{\mu}_{T,ma}\cdot\boldsymbol{\mu}_{G,k}}{|\mathbf{r}_{mk}|^3} - 3\frac{(\boldsymbol{\mu}_{T,ma}\cdot\mathbf{r}_{mk})(\boldsymbol{\mu}_{G,k}\cdot\mathbf{r}_{mk})}{|\mathbf{r}_{mk}|^5}\right) \quad (1)$$

$$J_{ma,nb} = \sum_{m,n}^{m\neq n} \frac{1}{4\pi\varepsilon\varepsilon_0}\left(\frac{\boldsymbol{\mu}_{T,ma}\cdot\boldsymbol{\mu}_{T,nb}}{|\mathbf{r}_{mn}|^3} - 3\frac{(\boldsymbol{\mu}_{T,ma}\cdot\mathbf{r}_{mn})(\boldsymbol{\mu}_{T,nb}\cdot\mathbf{r}_{mn})}{|\mathbf{r}_{mn}|^5}\right) \quad (2)$$

where m (n) runs over peptide bonds, a and b denote the n → π^* or $\pi \rightarrow \pi^*$ transitions, respectively, $\varepsilon_{0,ma}$ denotes the excitation energy of the isolated peptide bond, $\boldsymbol{\mu}_{T,ma}$ and $\boldsymbol{\mu}_{G,k}$ denote the electric transition dipole moment of the peptide bond and the ground state dipole moment of the surrounding amino acid residues k, respectively, and $\mathbf{r}_{mn}$ denotes the distance vector between m and n. In addition, magnetic transition dipole moment $\boldsymbol{\mu}_M$ of peptide bonds is needed for calculating rotatory strength ($R = |\boldsymbol{\mu}_M|\cdot|\boldsymbol{\mu}_T|\cdot\cos\theta$).

The whole ML protocol for protein CD spectra is described as follows (Figure 1c). A total of 1000 different types of

https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377−2384

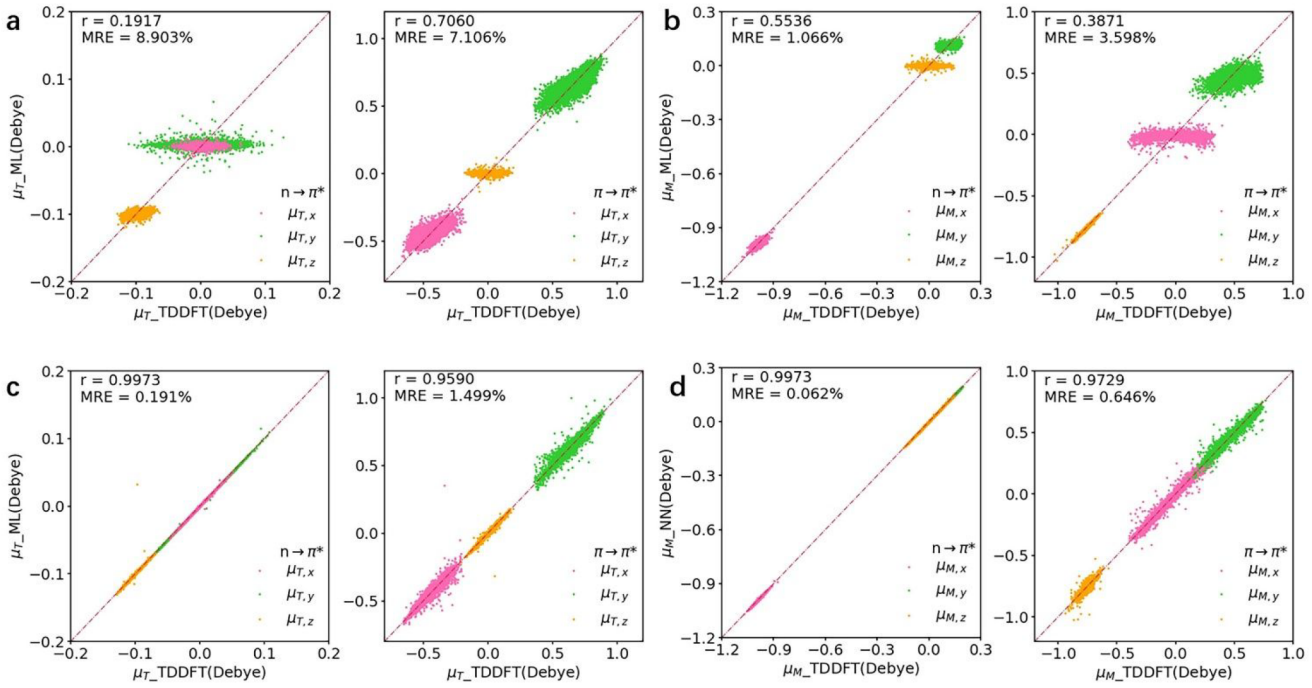


Figure 2. ML prediction of the electric and magnetic transition dipole moments of peptide bonds. (a) Correlation plots of the TDDFT and ML predicted electric transition dipole moments of the n → π^* and $\pi \rightarrow \pi^*$ transitions using CM with GBR. (b) Correlation plots of the TDDFT and ML predicted magnetic transition dipole moments of the n → π^* and $\pi \rightarrow \pi^*$ transitions using CM with GBR. (c) Same as (a) but using EANN. (d) Same as (b) but using EANN.

proteins are downloaded from the RCSB Protein Data Bank.[36] Proteins are first sliced into peptide bonds and amino acid residues. We then randomly selected 50 000 peptide bonds and 200 000 amino acid residues for data preparation. A total of 200 000 amino acid residues include 20 kinds of amino acids, each with 10 000 structures. We employ the time dependent DFT (TDDFT) method at the PBE0/cc-pVDZ level to calculate the excited state properties of peptide bonds (modeled by the *N*-methylacetamide molecule in Figure 1a) and the DFT method at B3LYP/6-311++G** level to calculate the ground state properties of amino acid residues. All the DFT and TDDFT simulations are performed in the Gaussian 16 package.[37]

Internal coordinates and converted Cartesian coordinates are chosen as the molecular descriptors for ε_0 of the peptide bond and $\boldsymbol{\mu}_G$ of the amino acid residue, respectively, while the embedded density descriptors[29] are used for better representing electric and magnetic transition dipole moments $\boldsymbol{\mu}_T/\boldsymbol{\mu}_M$ of the peptide bond, as discussed below. Starting with the DFT/TDDFT data sets, we run the ML data-training process to build the correlation between the descriptors and our prediction targets. For the ε_0 of peptide bonds and the $\boldsymbol{\mu}_G$ of amino acid residues, we use a neural network model with three hidden layers (32, 64, and 128 neurons, respectively). The rectified linear unit activation function is used for each hidden layer to resist the disappearance of the gradient and reduce the influence of noise.[38] L2 regularization is employed to solve the problem of overfitting.[39] In addition, we use the Adam optimizer[40] to avoid the local minima during the NN training. And we adjust the learning rate every 500 steps after setting the initial learning rate to 0.001. $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ of peptide bonds are predicted using embedded atomic neural networks (EANN) with the atom-wise embedded density descriptors.[29] Note that 36 descriptors are used as the input for representing the local environment of each atom of the NMA molecule, and each atomic neural network consists of 2 hidden layers (30 neurons in each layer).

With these parameters predicted by the ML model, ε_{ma} and $J_{ma,nb}$ can be calculated according to eq 1 and eq 2, yielding the effective Hamiltonian of exciton, and $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ can provide rotatory strength for the CD spectrum. The SPECTRON[41] program is used to diagonalize the Hamiltonian and finally output the CD spectrum of the selected protein (details in the Supporting Information).

3. RESULTS AND DISCUSSION

The accuracy and robustness of ML prediction are examined with the Pearson correlation coefficient (r) and the mean relative error (MRE) (details in the Supporting Information). Internal coordinates are chosen as the molecular descriptor to predict the excitation energies of peptide bonds. Each peptide bond has nine internal coordinates. For both n → π^* and $\pi \rightarrow \pi^*$ transitions, the ML predicted ε_0 are in good agreement with the TDDFT calculated one (Figure S1b). The results have high Pearson coefficients (0.9616 and 0.9512) and low MREs (0.363% and 0.252%). Meanwhile, the Cartesian coordinates are reoriented to the same reference coordinate system for the prediction of the ground state dipole moments of amino acid residues. The Cartesian coordinates can directly determine the structural features, and they have proved to be good descriptors for studying both the magnitude and direction of ground-state dipole moments in our previous work.[21] All correlation coefficients for $\boldsymbol{\mu}_G$ prediction are greater than 0.98, and most MRE values are below 5% (Figure S1c). The detailed results about ε_0 of peptide bonds and $\boldsymbol{\mu}_G$ of amino acid residues can be found in our previous work.[22] The results show that these simple molecular descriptors already warrant good accuracy and robustness of our ML models for ε_0 and $\boldsymbol{\mu}_G$.

https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377−2384

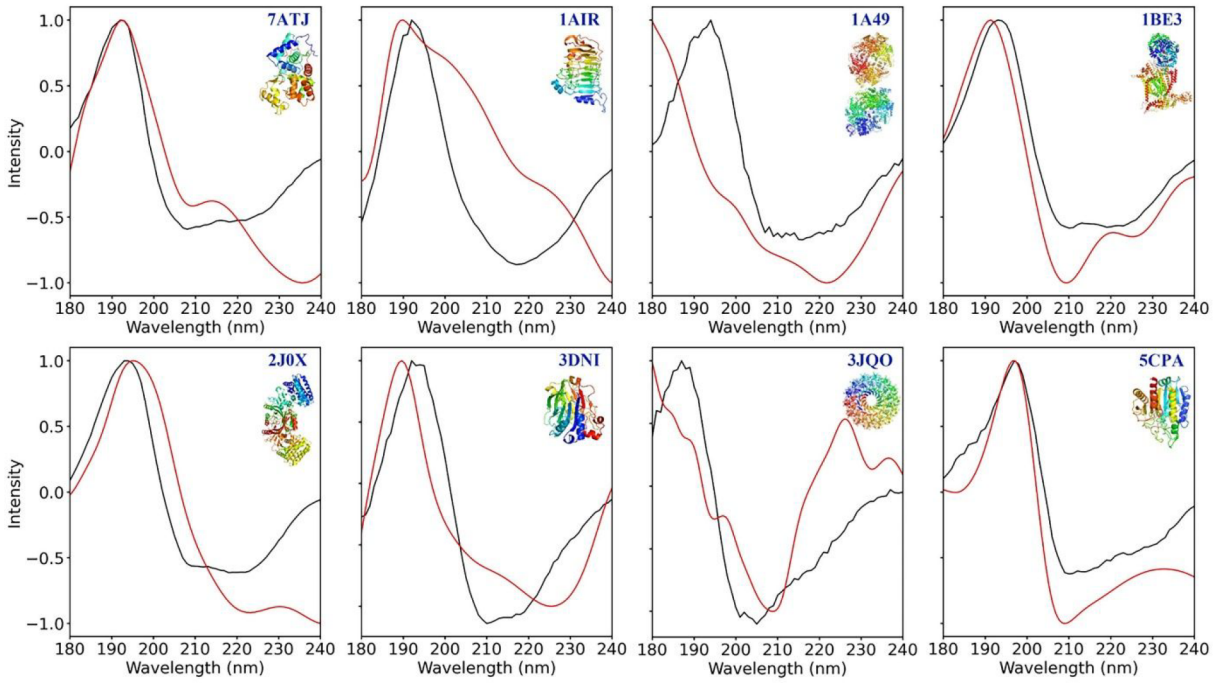


Figure 3. Experimental (black curves) and ML predicted (red curves) CD spectra of different types proteins. Intensity is scaled to have the same maximum intensity for each panel.

For the prediction of the more challenging electric and magnetic transition dipole moments of peptide bonds, we use the tensorial EANN model[29] based on the electron density-like atom-wise descriptors called embedded density descriptors, which are the square of a series of the linear combinations of Gaussian atomic orbitals from neighboring atoms.[30] As detailed in the SI, the EANN model can fit the direction of these transition dipole moments automatically and preserve their symmetry-covariant properties. To validate this, we compare the predictions of $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ with various molecular descriptors (details in the Supporting Information). Figure 2a,b shows the predicted $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ using the Coulomb matrix (CM) as the descriptor with a gradient boosting regression (GBR) algorithm. The results for both n $\rightarrow$ π^* and $\pi \rightarrow \pi^*$ transitions are unsatisfactory (as seen from poor Pearson coefficients), presumably because the $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ components in three directions are treated separately and not described as a whole for a given transition type (Figure S4 and Figure S5). In contrast, Figure 2c,d shows that the TDDFT calculated $\boldsymbol{\mu}_T/\boldsymbol{\mu}_M$ are perfectly predicted by EANN in all directions, with high Pearson coefficients ($r > 0.95$) and low MREs (<1.5%) for both n $\rightarrow$ π^* and $\pi \rightarrow \pi^*$ transitions.

To validate the advantage of the ML protocol, we compare DFT-based CD spectra and ML-based CD spectra of four different types of proteins (Figure S7). The essential matrix elements in the model Hamiltonian including ε_0, $\boldsymbol{\mu}_T$, $\boldsymbol{\mu}_M$, and $\boldsymbol{\mu}_G$ are calculated by DFT calculations and the ML model, respectively. Then CD spectra are generated by diagonalizing the model Hamiltonian using the SPECTRON program. As Table S1 shows, the simulated CD spectra from two computational protocols show reasonable agreement in terms of high Spearman rank coefficient (ρ). This coefficient is widely used for measuring the agreement between the spectra.[42−45] These results will be further improved in the future by increasing the prediction accuracy of the essential matrix elements in the model Hamiltonian, especially for the nondiagonal matrix elements (Figure S8c). Moreover, the speed for ML-based approach is mostly 4 orders of magnitude faster than the DFT-based one.

Next, we compare the spectra simulated by our protocol with the corresponding experimental results. The proteins used in Figure 3 are randomly selected from the RCSB Protein Data Bank[36] and are not included in our training data. The experimental spectra are all from the SMP180 and SP175 data sets of the Protein Circular Dichroism Data Bank (PCDDB).[46,47] The simulated spectra with different secondary structures are all in good agreement with the experiment (Table 1). More importantly, the simulated spectra of different secondary structures have different peak positions and line shapes, which lay the foundation for the interpretation of the experimental spectra. More results can be seen in the Supporting Information (Figure S10). The overall results show that our ML model is of high accuracy and good transferability.

Table 1. Comparison of the ML Simulated Protein CD Spectra with Experiments in Terms of Spearman Rank Correlation (ρ)

protein	PDB ID	secondary class	number of atoms	ρ
Peroxidase C1A	7ATJ	α	2944	0.74
Pectate lyase C	1AIR	β	2786	0.81
Pyruvate kinase	1A49	$\alpha + \beta$	34001	0.79
Cytochrome bc1 complex	1BE3	$\alpha + \beta$	16222	0.88
Aspartokinase III	2J0X	$\alpha + \beta$	6915	0.88
DNase I	3DNI	$\alpha + \beta$	2494	0.75
TraF protein	3JQO	$\alpha + \beta$	38842	0.83
Carboxypeptidase A	5CPA	$\alpha + \beta$	2753	0.97

To further demonstrate the applicability of our ML model, we run an MD simulation of four proteins with different

https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377−2384

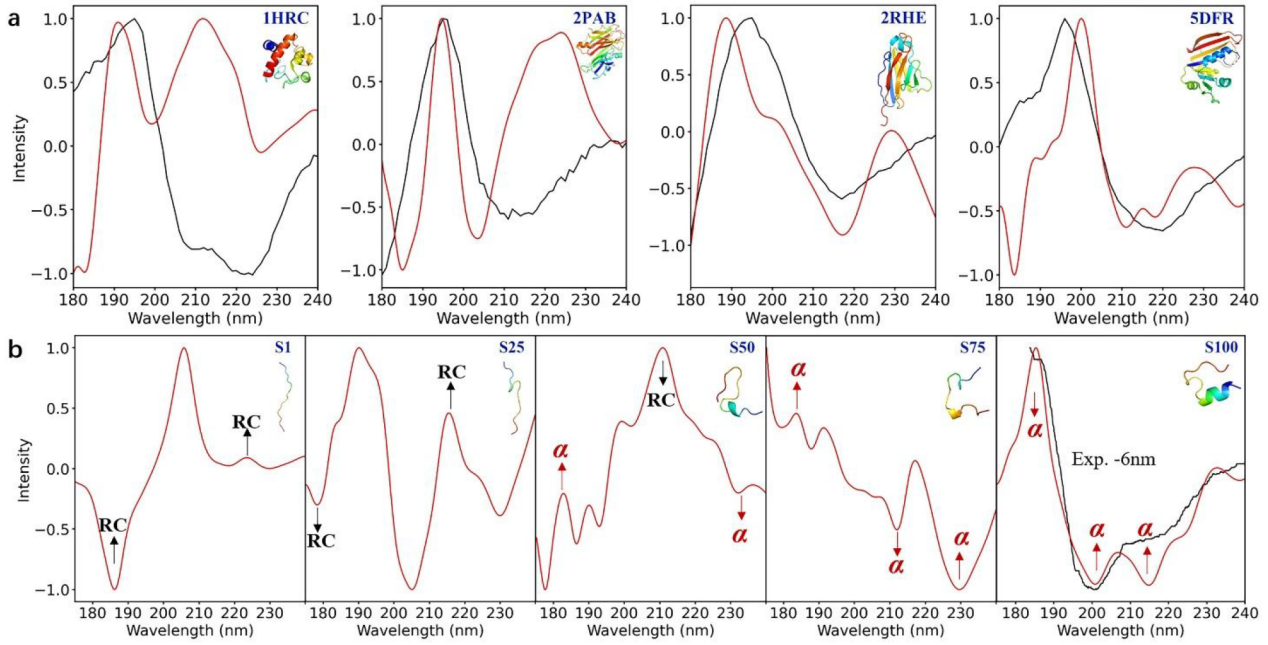


Figure 4. (a) Experimental (black curves) and ML predicted (red curves) CD spectra. The ML predictions are based on 1000 MD configurations. (b) The ML predicted CD spectra of the Trp-cage protein along its folding path (S1 → S100, S1: the original unfolded structure, S25: slightly folded along with the decrease of coil content, S50: folding faster and helical elements appear, S75: a cage formed with the rapid increase of α-helix, S100: the final stably folded structure). All spectra are averaged over 100 MD conformations for each state.

secondary structures (α-helix, β-sheet, and $\alpha + \beta$) to take the factor of environmental fluctuation into account and average the CD spectra of 1000 molecular dynamics (MD) conformations (details in the Supporting Information). The simulated CD spectra by our ML model are in good agreement with the experimental spectra in terms of overall line shape and the main peak position (Figure 4a). However, the ML-based spectra show a worse agreement than the spectra shown in Figure 3 when including the environment fluctuation by MD simulation. It is likely that a 2 ns trajectory is insufficient to cover configuration fluctuation of proteins in experimental measurement. In contrast, a single-frame structure from a PDB file represents the protein's average configuration in the crystalline condition, which is a better representative of its real configuration in spectra measurement. Therefore, CD spectra derived from a single-frame structure using our ML model are in better agreement with experimental results than those derived from an MD trajectory (Figure 3). In particular, the simulated CD spectra of different secondary structures show unique characteristics, which help to distinguish one from others. For comparison, the averaged IR spectra of these four proteins based on 1000 MD are obtained by the ML model proposed in our previous work.[20] The averaged IR spectra of two proteins with $\alpha + \beta$ structure (PDB ID: 2RHE, 5DFR) are very similar (Figure S11). It is thus clear that CD spectra are more sensitive to different secondary structures. These results indicate that our ML model can be applied to obtain CD spectra of different structures under environmental fluctuations with high accuracy and good transferability.

Protein folding is a key process for proteins to form unique three-dimensional structures and functions. Tracking structural changes in real time during the folding process can facilitate mechanistic understanding of proteins. Trp-cage (PDB ID: 1L2Y) is a mini protein that has been widely studied, and it is a convenient tool for studying folding dynamics.[48] Therefore, we use our ML model to monitor the folding process of the Trp-cage protein. All the CD spectra predicted by our ML model are based on 100 MD conformations retrieved from our previous study.[49] We have selected five representative states during the folding process (Figure 4b and Table S2). The initial state is in total coil structure (S1) and starts to fold as the content of coil decreases (S25). The structure folds faster and helical elements appear (S50). Then, a cage is formed with the appreciable increase of α-helix (S75). Finally it reaches the fully folded (S100) state. Clearly, the simulated CD spectra of the structure in different stages of the folding process are different. The random coil marked "RC" has a positive band at 212 nm and a negative one around 195 nm. The alpha helix marked "α" has negative bands at 222 and 208 nm and a positive one at 190 nm. During the folding process, the characteristics of "RC" decrease while those of "α" increase. Especially, the simulated CD spectrum of the completely folded state fits well with the experimental one[50] after the latter is blue-shifted as a whole by 6 nm (equivalent to ~0.2 eV). The above results of Trp-cage simulation show that our ML protocol can facilitate real-time CD spectroscopy study on protein folding.

4. SUMMARY

We have proposed a cost-effective machine learning protocol to simulate the electronic circular dichroism spectrum for proteins. This ML model benefits from the embedded density descriptors that give a robust and reliable prediction of the key tensorial parameters for the CD spectrum, including the electric and magnetic transition dipole moments of peptide bonds. Based on the parameters predicted by ML, we build the effective Hamiltonian and generate CD spectra of proteins. Our computational protocol not only significantly speeds up spectra simulation compared to the conventional first-principle calculation protocol but also obtains comparable results with

https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377−2384

experiments on a variety of different proteins, signifying the efficiency, accuracy, and transferability of our protocol. Further, the model is used for fast prediction of CD spectra of different conformations of Trp-cage along its folding pathway, demonstrating its power of efficiently mapping protein structures with their corresponding CD spectra for the study of protein dynamics. In summary, our ML protocol is a promising tool in the simulation of a protein CD spectrum and can be extended to other fields such as near-ultraviolet spectroscopy and two-dimensional spectroscopy.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.1c00449.

Computational details, ML prediction of the excitation energies of peptide bonds and the ground state dipole moments of 20 residues, ML prediction of electric and magnetic transition dipole moments using CM and ACSF, proteins of interest in this study, the comparison of DFT and ML simulated CD spectra, more experimental and ML predicted CD spectra of different types proteins, and ML predicted IR spectra of four different types proteins based on 1000 MD configurations (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jun Jiang − *Hefei National Laboratory for Physical Sciences at the Microscale, Collaborative Innovation Center of Chemistry for Energy Materials, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China;* orcid.org/0000-0002-6116-5605; Email: jiangj1@ustc.edu.cn

Bin Jiang − *Hefei National Laboratory for Physical Sciences at the Microscale, Collaborative Innovation Center of Chemistry for Energy Materials, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China;* orcid.org/0000-0003-2696-5436; Email: bjiangch@ustc.edu.cn

Authors

Luyuan Zhao − *Hefei National Laboratory for Physical Sciences at the Microscale, Collaborative Innovation Center of Chemistry for Energy Materials, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China*

Jinxiao Zhang − *Guangxi Key Laboratory of Electrochemical and Magneto-chemical Functional Materials, College of Chemistry and Bioengineering, Guilin University of Technology, Guilin 541006, P. R. China*

Yaolong Zhang − *Hefei National Laboratory for Physical Sciences at the Microscale, Collaborative Innovation Center of Chemistry for Energy Materials, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China*

Sheng Ye − *School of Artificial Intelligence, Anhui University, Hefei, Anhui 230601, P. R. China*

Guozhen Zhang − *Hefei National Laboratory for Physical Sciences at the Microscale, Collaborative Innovation Center of Chemistry for Energy Materials, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China;* orcid.org/0000-0003-0125-9666

Xin Chen − *Gusu Laboratory of Materials, Suzhou, Jiangsu 215123, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.1c00449

Author Contributions

#L.Z., J.Z., and Y.Z. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (Grants 2018YFA0208603), the CAS Project for Young Scientists in Basic Research (YSBR-005), the National Natural Science Foundation of China (Grants 22025304, 22033007, 22073089, 21790350, 21703221), and the Anhui Initiative in Quantum Information Technologies (AHY090200). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of University of Science and Technology of China.

■ REFERENCES

(1) Pan, X. J.; Thompson, M. C.; Zhang, Y.; Liu, L.; Fraser, J. S.; Kelly, M. J. S.; Kortemme, T. Expanding the space of protein geometries by computational design of de novo fold families. *Science* **2020**, *369* (6507), 1132.

(2) Chen, C. Y.; Chang, Y. C.; Lin, B. L.; Huang, C. H.; Tsai, M. D. Temperature-Resolved Cryo-EM Uncovers Structural Bases of Temperature-Dependent Enzyme Functions. *J. Am. Chem. Soc.* **2019**, *141* (51), 19983−19987.

(3) Mangubat-Medina, A. E.; Martin, S. C.; Hanaya, K.; Ball, Z. T. A Vinylogous Photocleavage Strategy Allows Direct Photocaging of Backbone Amide Structure. *J. Am. Chem. Soc.* **2018**, *140* (27), 8401−8404.

(4) Salvi, N.; Abyzov, A.; Blackledge, M. Analytical Description of NMR Relaxation Highlights Correlated Dynamics in Intrinsically Disordered Proteins. *Angew. Chem., Int. Ed.* **2017**, *56* (45), 14020−14024.

(5) Wu, S.; Wang, D.; Liu, J.; Feng, Y.; Weng, J.; Li, Y.; Gao, X.; Liu, J.; Wang, W. The Dynamic Multisite Interactions between Two Intrinsically Disordered Proteins. *Angew. Chem., Int. Ed.* **2017**, *56* (26), 7515−7519.

(6) Kennedy, D.; Norman, C. What Don't We Know? *Science* **2005**, *309*, 75−75.

(7) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577* (7792), 706−710.

(8) Wei, S.; Zou, X.; Tian, J.; Huang, H.; Guo, W.; Chen, Z. Control of Protein Conformation and Orientation on Graphene. *J. Am. Chem. Soc.* **2019**, *141* (51), 20335−20343.

(9) Orton, H. W.; Stanek, J.; Schubeis, T.; Foucaudeau, D.; Ollier, C.; Draney, A. W.; Le Marchand, T.; Cala-De Paepe, D.; Felli, I. C.; Pierattelli, R.; Hiller, S.; Bermel, W.; Pintacuda, G. Protein NMR Resonance Assignment without Spectral Analysis: 5D SOlid-State Automated Projection SpectroscopY (SO-APSY). *Angew. Chem., Int. Ed.* **2020**, *59* (6), 2380−2384.

(10) Pletka, C. C.; Nepravishta, R.; Iwahara, J. Detecting Counterion Dynamics in DNA-Protein Association. *Angew. Chem., Int. Ed.* **2020**, *59* (4), 1465−1468.

(11) Quiñones-Ruiz, T.; Rosario-Alomar, M. F.; Ruiz-Esteves, K.; Shanmugasundaram, M.; Grigoryants, V.; Scholes, C.; López-Garriga, J.; Lednev, I. K. Purple Fibrils: A New Type of Protein Chromophore. *J. Am. Chem. Soc.* **2017**, *139* (29), 9755−9758.
(12) Adamski, W.; Salvi, N.; Maurin, D.; Magnat, J.; Milles, S.; Jensen, M. R.; Abyzov, A.; Moreau, C. J.; Blackledge, M. A Unified Description of Intrinsically Disordered Protein Dynamics under Physiological Conditions Using NMR Spectroscopy. *J. Am. Chem. Soc.* **2019**, *141* (44), 17817−17829.
(13) Ianeselli, A.; Orioli, S.; Spagnolli, G.; Faccioli, P.; Cupellini, L.; Jurinovich, S.; Mennucci, B. Atomic Detail of Protein Folding Revealed by an Ab Initio Reappraisal of Circular Dichroism. *J. Am. Chem. Soc.* **2018**, *140* (10), 3674−3682.
(14) Rogers, D. M.; Jasim, S. B.; Dyer, N. T.; Auvray, F.; Refregiers, M.; Hirst, J. D. Electronic Circular Dichroism Spectroscopy of Proteins. *Chem.* **2019**, *5* (11), 2751−2774.
(15) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8* (10), 6924−6935.
(16) Dong, S. S.; Govoni, M.; Galli, G. Machine learning dielectric screening for the simulation of excited state properties of molecules and materials. *Chem. Sci.* **2021**, *12* (13), 4970−4980.
(17) Sommers, G. M.; Calegari Andrade, M. F.; Zhang, L.; Wang, H.; Car, R. Raman spectrum and polarizability of liquid water from deep neural networks. *Phys. Chem. Chem. Phys.* **2020**, *22* (19), 10592−10602.
(18) Heo, L.; Feig, M. High-accuracy protein structures by combining machine-learning with physics-based refinement. *Proteins: Struct., Funct., Genet.* **2020**, *88* (5), 637−642.
(19) Hanson, J.; Paliwal, K. K.; Litfin, T. D.; Yang, Y. Q.; Zhou, Y. Getting to Know Your Neighbor: Protein Structure Prediction Comes of Age with Contextual Machine Learning. *J. Comput. Biol.* **2020**, *27* (5), 796−814.
(20) Ye, S.; Zhong, K.; Zhang, J.; Hu, W.; Hirst, J. D.; Zhang, G.; Mukamel, S.; Jiang, J. A Machine Learning Protocol for Predicting Protein Infrared Spectra. *J. Am. Chem. Soc.* **2020**, *142* (45), 19071−19077.
(21) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (24), 11612.
(22) Zhang, J.; Ye, S.; Zhong, K.; Zhang, Y.; Chong, Y.; Zhao, L.; Zhou, H.; Guo, S.; Zhang, G.; Jiang, B.; Mukamel, S.; Jiang, J. A Machine-Learning Protocol for Ultraviolet Protein-Backbone Absorption Spectroscopy under Environmental Fluctuations. *J. Phys. Chem. B* **2021**, *125* (23), 6171−6178.
(23) del Villar-Guerra, R.; Trent, J. O.; Chaires, J. B. G-Quadruplex Secondary Structure Obtained from Circular Dichroism Spectroscopy. *Angew. Chem., Int. Ed.* **2018**, *57* (24), 7171−7175.
(24) Micsonai, A.; Wien, F.; Kernya, L.; Lee, Y.-H.; Goto, Y.; Réfrégiers, M.; Kardos, J. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (24), No. E3095.
(25) Oktaviani, N. A.; Matsugami, A.; Malay, A. D.; Hayashi, F.; Kaplan, D. L.; Numata, K. Conformation and dynamics of soluble repetitive domain elucidates the initial β-sheet formation of spider silk. *Nat. Commun.* **2018**, *9* (1), 2121.
(26) Sreerama, N.; Woody, R. W. Computation and Analysis of Protein Circular Dichroism Spectra. In *Methods in Enzymology*; Academic Press: 2004; Vol. *383*, pp 318−351.
(27) Brahms, S.; Brahms, J. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.* **1980**, *138* (2), 149−178.
(28) Rosenfeld, L. Quantenmechanische Theorie der natürlichen optischen Aktivität von Flüssigkeiten und Gasen. *Eur. Phys. J. A* **1929**, *52* (3), 161−174.
(29) Zhang, Y.; Ye, S.; Zhang, J.; Hu, C.; Jiang, J.; Jiang, B. Efficient and Accurate Simulations of Vibrational and Electronic Spectra with Symmetry-Preserving Neural Network Models for Tensorial Properties. *J. Phys. Chem. B* **2020**, *124* (33), 7284−7290.
(30) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. *J. Phys. Chem. Lett.* **2019**, *10* (17), 4962−4967.
(31) Abramavicius, D.; Palmieri, B.; Mukamel, S. Extracting single and two-exciton couplings in photosynthetic complexes by coherent two-dimensional electronic spectra. *Chem. Phys.* **2009**, *357* (1), 79−84.
(32) Abramavicius, D.; Jiang, J.; Bulheller, B. M.; Hirst, J. D.; Mukamel, S. Simulation Study of Chiral Two-Dimensional Ultraviolet Spectroscopy of the Protein Backbone. *J. Am. Chem. Soc.* **2010**, *132* (22), 7769−7775.
(33) Frenkel, J. On the Transformation of light into Heat in Solids. I. *Phys. Rev.* **1931**, *37* (1), 17−44.
(34) Kasha, M.; Rawls, H. R.; Ashraf El-Bayoumi, M. The exciton model in molecular spectroscopy. *Pure Appl. Chem.* **1965**, *11* (3−4), 371−392.
(35) Zhang, Y.; Luo, Y.; Zhang, Y.; Yu, Y.-J.; Kuang, Y.-M.; Zhang, L.; Meng, Q.-S.; Luo, Y.; Yang, J.-L.; Dong, Z.-C.; Hou, J. G. Visualizing coherent intermolecular dipole-dipole coupling in real space. *Nature* **2016**, *531* (7596), 623−627.
(36) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(37) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. *Gaussian 16*, Rev. A.03; Wallingford, CT, 2016.
(38) Maas, A. L.; Hannun, A. Y.; Ng, A. Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proceedings of the 30th International Conference on Machine Learning*; JMLR: 2013; p 3.
(39) Ng, A. Y. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*; Association for Computing Machinery: 2004; p 78.
(40) Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. Presented at the 3rd International Conference on Learning Representations, San Diego, CA, 2015.
(41) Abramavicius, D.; Palmieri, B.; Voronine, D. V.; Sanda, F.; Mukamel, S. Coherent multidimensional optical spectroscopy of excitons in molecular aggregates; quasiparticle versus supermolecule perspectives. *Chem. Rev.* **2009**, *109* (6), 2350−2408.
(42) Hirst, J. D.; Colella, K.; Gilbert, A. T. B. Electronic Circular Dichroism of Proteins from First-Principles Calculations. *J. Phys. Chem. B* **2003**, *107* (42), 11813−11819.
(43) Besley, N. A.; Hirst, J. D. Theoretical Studies toward Quantitative Protein Circular Dichroism Calculations. *J. Am. Chem. Soc.* **1999**, *121* (41), 9636−9644.
(44) Baumann, K.; Clerc, J. T. Computer-assisted IR spectra prediction — linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348* (1), 327−343.
(45) Henschel, H.; Andersson, A. T.; Jespers, W.; Mehdi Ghahremanpour, M.; van der Spoel, D. Theoretical Infrared Spectra: Quantitative Similarity Measures and Force Fields. *J. Chem. Theory Comput.* **2020**, *16* (5), 3307−3315.
(46) Whitmore, L.; Woollett, B.; Miles, A. J.; Klose, D. P.; Janes, R. W.; Wallace, B. A. PCDDB: the Protein Circular Dichroism Data Bank, a repository for circular dichroism spectral and metadata. *Nucleic Acids Res.* **2011**, *39*, D480−D486.
(47) Whitmore, L.; Miles, A. J.; Mavridis, L.; Janes, R. W.; Wallace, B. A. PCDDB: new developments at the Protein Circular Dichroism Data Bank. *Nucleic Acids Res.* **2017**, *45* (D1), D303−D307.
(48) Bellissent-Funel, M.-C.; Hassanali, A.; Havenith, M.; Henchman, R.; Pohl, P.; Sterpone, F.; van der Spoel, D.; Xu, Y.; Garcia, A. E. Water Determines the Structure and Dynamics of Proteins. *Chem. Rev.* **2016**, *116* (13), 7673−7697.

https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377−2384

(49) Jiang, J.; Lai, Z.; Wang, J.; Mukamel, S. Signatures of the Protein Folding Pathway in Two-Dimensional Ultraviolet Spectroscopy. *J. Phys. Chem. Lett.* **2014**, *5* (8), 1341−1346.
(50) Adams, C. M.; Kjeldsen, F.; Patriksson, A.; van der Spoel, D.; Gräslund, A.; Papadopoulos, E.; Zubarev, R. A. Probing solution- and gas-phase structures of Trp-cage cations by chiral substitution and spectroscopic techniques. *Int. J. Mass Spectrom.* **2006**, *253* (3), 263−273.

https://doi.org/10.1021/jacsau.1c00449
JACS Au 2021, 1, 2377−2384