Supplementary Information for

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

Luyuan Zhao,[1,#] Jinxiao Zhang,[2,#] Yaolong Zhang,[1,#] Sheng Ye,[3] Guozhen Zhang,[1] Xin Chen, [4] Bin Jiang,[*,1] Jun Jiang[*,1]

[1]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

[2]Guangxi Key Laboratory of Electrochemical and Magneto-chemical Functional Materials, College of Chemistry and Bioengineering, Guilin University of Technology, Guilin 541006, P. R. China

[3]School of Artificial Intelligence, Anhui University, Hefei, Anhui 230601, P. R. China

[4]Gusu Laboratory of Materials, Suzhou, Jiangsu 215123, P. R. China

[#] L. Z., J.Z. and Y. Z. contributed equally to this work.

*Corresponding author: jiangj1@ustc.edu.cn, bjiangch@ustc.edu.cn

Table of contents

1 Computational details

1.1 The Frenkel excition model

Proteins are composed of peptide bonds and amino acid residues (Figure 1a). The absorption of partially delocalized peptide bonds of the backbone contributes to the response of the protein electronic circular dichroism (CD) spectrum in the 190–250 nm region. The electronic excitation of the peptide bonds is easily affected by fluctuations in the surrounding environment (amino acid residues) and can be described effectively by the Frenkel exciton model [1-3].

$$\hat{H} = \sum_{ma} \varepsilon_{ma} \hat{B}_{ma}^{\dagger} \hat{B}_{ma} + \sum_{ma,nb}^{m \neq n} J_{ma,nb} \hat{B}_{ma}^{\dagger} \hat{B}_{nb} \qquad (1)$$

where a and b represent two excited state transitions in peptide bonds, which are n→π* transition around 220 nm and π→π* transitions around 190 nm (Figure 1b). And ma denotes the a electronic transition of the m-th peptide bond. $\hat{B}_{ma}^{\dagger}$ is the creation Pauli operator which promotes the peptide bond m from the ground state into the excited state a, and the $\hat{B}_{ma}$ is the corresponding annihilation operator. ε_{ma} is the electronic excitation energy, and $J_{ma,nb}$ is the electronic coupling between excited states. ε_{ma} can be written as:

$$\varepsilon_{ma} = \varepsilon_{0,ma} + \sum_{k} \frac{1}{4\pi\varepsilon\varepsilon_0} \iint d\boldsymbol{r}_m d\boldsymbol{r}_k \left(\frac{[\rho_{T,ma}(\boldsymbol{r}_m) - \rho_{G,m}(\boldsymbol{r}_m)] \cdot \rho_{G,k}(\boldsymbol{r}_k)}{|\boldsymbol{r}_m - \boldsymbol{r}_k|} \right) \qquad (2)$$

$\varepsilon_{0,ma}$ is the excitation energy of the isolated peptide bond, and the second term is the electrostatic interaction between the peptide bond and the surrounding environment. $\rho_{T,ma}$ and $\rho_{G,m}$ denote the charge density of the *a th* excited state and ground state of the peptide bond *m*, respectively. $\rho_{G,k}$ represents the ground state charge density of the amino acid residue *k*, and k runs over all amino acid residues. r is the spatial coordinate. In addition, $J_{ma,nb}$ in Eq. (1) can be written as:

$$J_{ma,nb} = \frac{1}{4\pi\varepsilon\varepsilon_0} \iint d\boldsymbol{r}_m d\boldsymbol{r}_n \frac{\rho_{T,ma}(\boldsymbol{r}_m)\rho_{T,nb}(\boldsymbol{r}_n)}{|\boldsymbol{r}_m - \boldsymbol{r}_n|} \qquad (3)$$

In order to simplify the calculation of the two-electron integral in Eq. (2) and Eq. (3), the dipole approximation can be applied to deal with the electronic interaction [4-5]. By the dipole approximation, the excitation energy ε_{ma} of the peptide bonds can be described as:

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

the second part in Eq. (4) is the interaction between $\boldsymbol{\mu}_{T,ma}$ and $\boldsymbol{\mu}_{G,k}$, which are the transition dipole moment of the peptide bond and the ground state dipole moment of the surrounding amino acid residues *k*,

respectively. Also based on the dipole approximation, the resonance coupling $J_{ma,nb}$ between excited states can be described as:

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

According to Eq. (4) and Eq. (5), excitation energy ε_0 and electric transition dipole moment $\boldsymbol{\mu}_T$ of all peptide bonds, and the ground state $\boldsymbol{\mu}_G$ of all amino acid residues need to be calculated to construct the Hamiltonian of a protein by dipole approximation. In addition, the simulated CD spectrum requires the magnetic transition dipole moment $\boldsymbol{\mu}_M$ of peptide bond to provide the information about rotor strength.

1.2 The calculation of rotor strength.

Rotor strength is defined as the imaginary part of the scalar product of the electric and magnetic transition dipole moments of an electronic transition [6].

$$R = |\boldsymbol{\mu}_M| \cdot |\boldsymbol{\mu}_T| \cdot \cos\theta$$

Therefore the transition electric and magnetic dipole moments of peptide bonds should be provided to SPECTRON [7] to calculate the rotatory strength of the CD spectrum.

1.3 The machine learning protocol

1.3.1 Data collection

We downloaded 1000 different types of protein PDB files from the RCSB Protein Data Bank (http://www.rcsb.org) [8]. These PDB files contain 8 kinds of common proteins: fibrous protein, globular protein, keratin, collagen, chaperone, myoglobin, hemoglobin and denaturation. This ensures the diversity of data. We split the PDB files into peptide bonds and amino acid residues, and then we extracted 50 peptide bonds and 200 amino acid residues (20 kinds of amino acid residues, 10 for each residue) from each protein randomly. A total of 50,000 peptide bonds and 200,000 amino acid residues (20 kinds of amino acid residues, 10,000 for each residue) are used for training. Density functional theory (DFT) and time-dependent density functional theory (TDDFT) calculations are used to collect the data about peptide bonds and residues.

DFT calculations at B3LYP/6-311++G** level are used to obtain the ground state dipole moment of residues. B3LYP/6-311++G** is employed, because it is a reliable method when doing ground state calculations [9-11]. For peptide bonds, we use the TDDFT method at the PBE0/cc-pVDZ level to calculate the excitation energy, electric and magnetic transition dipole moment of the peptide bonds. The functional PBE0 has been recommended for TDDFT calculation of excitation energies of different molecules [12-13] and it has been employed in previous TDDFT study of NMA [14]. And the calculation using PBE0 is cost-effective, which is friendly to our calculation for peptide bonds with 50,000 sets of data. The lowest 10 excitation states are calculated. We have carried out the phase correction [15] with Multiwfn code [16] to solve

the related mutation problem of structure and properties caused by the random phase of the wave function. NOSYMM keyword is used to avoid structural reorientation and polarizable continuum models (PCM) with water ad solvent is required for all the calculations. All the DFT/TDDFT simulation are performed in Gaussian 16 package [17].

1.3.2 ML Model Training

For the prediction of ε_0 of peptide bond and $\boldsymbol{\mu}_G$ of amino acid residue, we select 80% of the data for training set and 20% for test set randomly. In order to avoid that the different range of raw input values may undermine the robustness NN results, all the input data are normalized to the dimensionless data in range 0 to 1:$x_n = \frac{x_i - x_{min}}{x_{max} - x_{min}}$, where x_i are the input data, x_{min} and x_{max} are minimum and maximum values of all the input data, and x_n are the normalized data.

The neural network (NN) architecture based on Tensorflow framework [18] has one input layer, three hidden layers (32, 64, and 128 neurons, respectively) and one output layer. The rectified linear unit (ReLU) [19] is used as the activation function to resist the disappearance of the gradient and reduce the influence of noise to a certain extent. L2 regularization is employed to combat the overfitting that occurs during the training process [20]. L2 regularization can makes the NN tend to learn a smaller weight value w. And a smaller weight value indicates that the complexity of the network is lower, and the fitting of the data is just suitable (also known as Occam's razor). We also use the Adam optimizer [21], which is most commonly used in machine learning and usually works well in comprehensive situations. The learning rate is set to 0.001.

For the prediction of $\boldsymbol{\mu}_T/\boldsymbol{\mu}_M$ of peptide bond, the EANN model with our newly-developed embedded density descriptors are employed, as discussed below.

The Pearson coefficient (r) and the mean relative error (MRE) are used to evaluate the robustness of our ML model. The pearson coefficient is calculated by a module of Python called Numpy. And the mean relative error is defined as follows:

$$\mathrm{MRE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right|$$

where A_i is the actual value and P_i is the predicted value of molecule i, respectively. n is the molecular number.

1.3.3 ML Prediction of CD spectra

Based on the trained ML model, the corresponding ε_0, $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ of peptide bond and $\boldsymbol{\mu}_G$ of amino acid residue can be obtained by inputting the geometric information contained in the PDB file of a new protein. With these predicted parameters, ε_{ma} and $J_{ma,nb}$ can be calculated using Eq. (4) and Eq. (5). Then they can construct the exciton Hamiltonian. $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ can provide rotor strength for CD spectrum. The

SPECTRON [7] program is used to diagonalize the Hamiltonian and finally output the CD spectrum of the selected protein.

1.4 Descriptors of electric and magnetic transition dipole moment

The prediction for $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$ is challenging because they are vectors involving multiple coordinate dependent components, which are covariant with system rotation. For the prediction of $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_M$, we select some machine learning algorithm and molecular descriptors, which are gradient boosting regression (GBR) [22] with coulomb matrix (CM) [23] and atom-centered symmetry functions (ACSF) [24], and the embedded atom neural network (EANN) [25] with embedded density descriptors [26].

The coulomb matrix (CM) is used to describe a central atom and its surrounding environment. For a central atom k, its CM can be expressed as:

$$M_{ij}(k) = \begin{cases} \frac{1}{2} Z_i^{2.4} \cdot f_{ik}^2 & i = j \\ \frac{Z_i Z_j}{||R_i - R_i||} & i \neq j \end{cases}$$

where i, j, and k are the labels of the atom. Z is the nuclear charge, and R is the Cartesian coordinates in Euclidean spaces. f_{ij} is a long-range correction function:

$$f_{ij} = \begin{cases} 1 & ||R_i - R_i|| \leq r - \Delta r \\ \frac{1}{2}\left(1 + \cos\left(\pi \frac{||R_i - R_i|| - r + \Delta r}{\Delta r}\right)\right) & r - \Delta r < ||R_i - R_i|| \leq r - \Delta r \\ 0 & ||R_i - R_i|| > r \end{cases}$$

The atom-centered symmetry functions (ACSF) uses a series of radial functions and angle functions to represent the environment around the central atom. For atom i, there are three radial symmetric functions:

$$G_i^1 = \sum_j f_c(R_{ij})$$

$$G_i^2 = \sum_j e^{-\eta(R_{ij} - R_s)^2} \cdot f_c(R_{ij})$$

$$G_i^3 = \sum_j \cos(k\, R_{ij}) \cdot f_c(R_{ij})$$

where R_{ij} is the distance between atom i and j, and $f_c(R_{ij})$ is the cutoff function function:

$$f_c(R_{ij}) = \begin{cases} 0.5 \cdot \left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right] & R_{ij} \leq R_c \\ 0 & R_{ij} > R_c \end{cases}$$

where R_c is the cutoff radius. It can be seen that G_i^1 is the simple sum of all the cutoff functions; G_i^2 is the sum of the product of the Gaussian function and the cutoff function. η and R control the peak width and shift of the Gaussian function, respectively. G_i^3 is the sum of the product of a damped cosine function and a radial function. The ACSF also includes two angle functions:

$$G_i^4 = 2^{1-\xi} \sum_{j,k \neq i}^{all} (1 + \lambda cos\theta_{ijk})^{\xi} \quad \cdot e^{-\eta\left(R_{ij}^2 + R_{ik}^2 + R_{jk}^2\right)^2} \cdot f_c(R_{ij}) \cdot f_c(R_{ik}) \cdot f_c(R_{jk})$$

$$G_i^5 = 2^{1-\xi} \sum_{j,k \neq i}^{all} (1 + \lambda cos\theta_{ijk})^{\xi} \quad \cdot e^{-\eta\left(R_{ij}^2 + R_{ik}^2\right)^2} \cdot f_c(R_{ij}) \cdot f_c(R_{ik})$$

G_i^4 and G_i^5 include the same angle part, but the radial part included is different. ξ is a parameter used to modify the distribution of angles centered of atom i. All the parameters described above are decided from best performing parameters reported by Marquetand et al [27].

The embedded atom neural network (EANN) reported in our previous works is a great approach to describe the transition dipole moments in an efficient way. Starting from the commonly used Gaussian-type orbitals (GTOs) and combined with empirical embedded atom method, the new density-like descriptors can be described as:

$$\boldsymbol{\rho}_{L,\alpha,r_s}^{i} = \sum_{l_x,l_y,l_z}^{l_x + l_y + l_z = L} \frac{L!}{l_x! l_y! l_z!} \left(\sum_{j=1}^{n_{atom}} c_j \varphi_{l_x l_y, l_z}^{\alpha, r_s}(r_{ij})\right)^2$$

where r_{ij} represents the Cartesian coordinates of the atom i relative to atom j, and the $\varphi_{l_x l_y, l_z}^{\alpha, r_s}(r_{ij})$ is the GTOs centered at each atom, c_j is an element-dependent weight which can be optimized in the training process. The prefactor consisted of the factorials l_x, l_y, l_z and L can transform the description of $\boldsymbol{\rho}_{L,\alpha,r_s}^{i}$ to an angular basis as Takahashi et al. realized. n_{atom} means the number of neighboring atoms close to the center atom with a sphere with a cutoff radius.

In order to verify the symmetry invariance of the descriptor, we define $f(r_{ij}) = \exp\left(-\alpha|r_{ij} - r_s|^2\right)$. According to multinomial theorem, $\boldsymbol{\rho}_{L,\alpha,r_s}^{i}$ can be derived as:

$$\boldsymbol{\rho}_{L,\alpha,r_s}^{i} = \sum_{j,k \neq i} c_j f(r_{ij}) c_k f(r_{ik}) r_{ij}^L r_{ik}^L (\cos\theta_{ijk})^L$$

Apparently, this density-like descriptors is invariant with respect to the overall translation and rotation, and permutation of identical atoms. And we also can see these density-like descriptors are constructed

relying on Cartesian coordinates only.

Electric and magnetic transition dipole moments subject to the three dimensional rotation group SO(3) meanwhile the Cartesian coordinate vector is naturally compatible with SO(3) symmetry when the molecule is rotated. Therefore, EANN constructs the desired tensor by multiplying virtual NN outputs with atomic coordinate vectors, while keeping structural descriptors symmetry invariant. Therefore, electric or magnetic transition dipole moment $\boldsymbol{\mu}_T$ is defined as:

$$\boldsymbol{\mu}_T^j = \sum_{i=1}^{N} q_i^j \boldsymbol{r}_i \quad (j = 1,2)$$

where N is the atom number and q_i^j is the two different outputs of the same atomic NN using the density-like descriptors. The cross product of $\boldsymbol{\mu}_T^1$ and $\boldsymbol{\mu}_T^2$ will be perpendicular to the plane defined by them:

$$\boldsymbol{\mu}_T^3 = \sum_{i=1}^{N} q_i^3 (\boldsymbol{\mu}_T^1 \times \boldsymbol{\mu}_T^2)$$

where q_i^3 is the third output of the same atomic NN. Then the sum of these three vectors $\boldsymbol{\mu}_T^{NN} = \boldsymbol{\mu}_T^1 + \boldsymbol{\mu}_T^2 + \boldsymbol{\mu}_T^3$ can describe the transition electric and magnetic dipole moment successfully because it is not restricted in the molecular plane with the correct rotational covariance.

1.5 The Spearman rank correlation coefficients

Spearman rank correlation coefficients quantitatively determine the monotonic relationship between two variables which was widely used measure for the agreement between the predicted and experimental spectra [28-31]. And it is defined as:

$$\rho = 1 - \frac{6\sum_i d_i^2}{n \cdot (n^2 - 1)}$$

where n is the number of elements in each vector, d_i is the difference between the ranks of x_i (absorption intensities of experiment/DFT simulated spectra) and y_i (absorption intensities of predicted spectra) in their respective data set.

1.6 Molecular dynamic simulations

Molecular dynamics (MD) simulations are performed for proteins in Figure 4a (PDB ID: 1HRC, 2PAB, 2RHE, 5DFR) using the GROMACS code [32]. Periodic boundary conditions with OPLS-AA force field and TIP3P water are employed to proteins. Particle-mesh Edwald is used to deal with long-range electrostatic interactions and short-range coulomb interactions are truncated at 1.2 nm. 50000 cycles energy minimization are carried out for each protein. After NVT equilibration with an integration timestep

of 2 fs ran for 0.5 ns at 300K, the 2 ns MD simulations are performed at 300K and 1 atm in NPT ensemble. In this process, 1000 configurations were extracted at 2 fs intervals to calculate the CD spectrum. The MD conformations of Figure 4b (S1, S25, S50, S75, S100) are retrieved from our previous reported work about Trp-cage [33]. The sequence of the Trp-cage is "NLYIQWLKDGG PSSGRPPPS", and the extended conformation is built as an initial structure. Based on this initial structure, 50 trajectories with different initial conditions were simulated. These trajectories covered 10 μs of the protein folding simulations. 100 state points along the folding pathway were selected by clustering method. And they are labeled as S1, S2, ..., S100.

2 Supplementary Figures

2.1 ML prediction of the excitation energies of peptide bonds and the ground state dipole moments of twenty residues

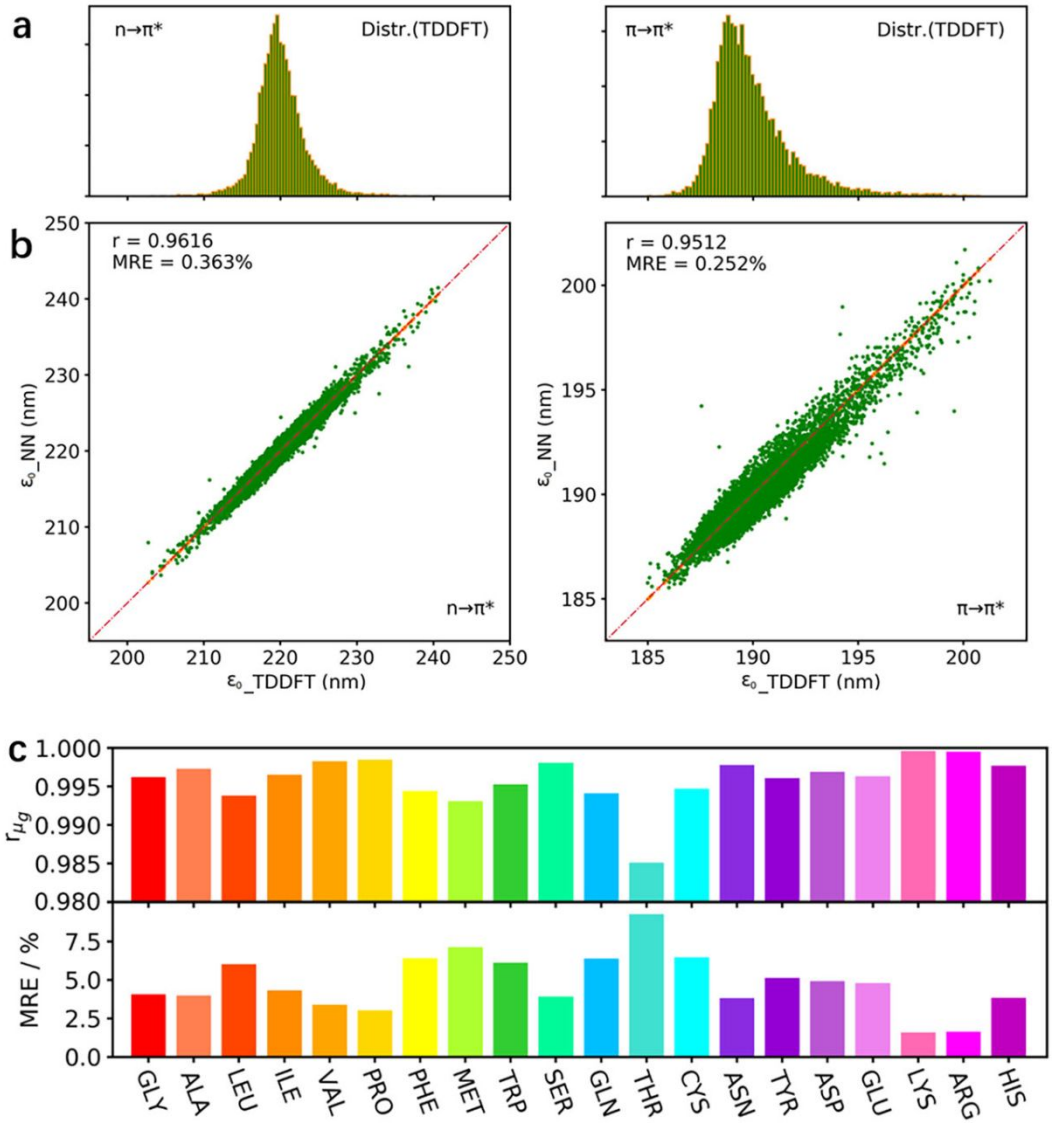

Figure S1. (a) Data distribution of the TDDFT calculated excitation energies of peptide bonds. (b)

Correlation plots of the TDDFT and ML predicted excitation energies of peptide bonds. (c) Pearson correlation coefficients (r) and the mean relative error (MRE) of twenty residues.

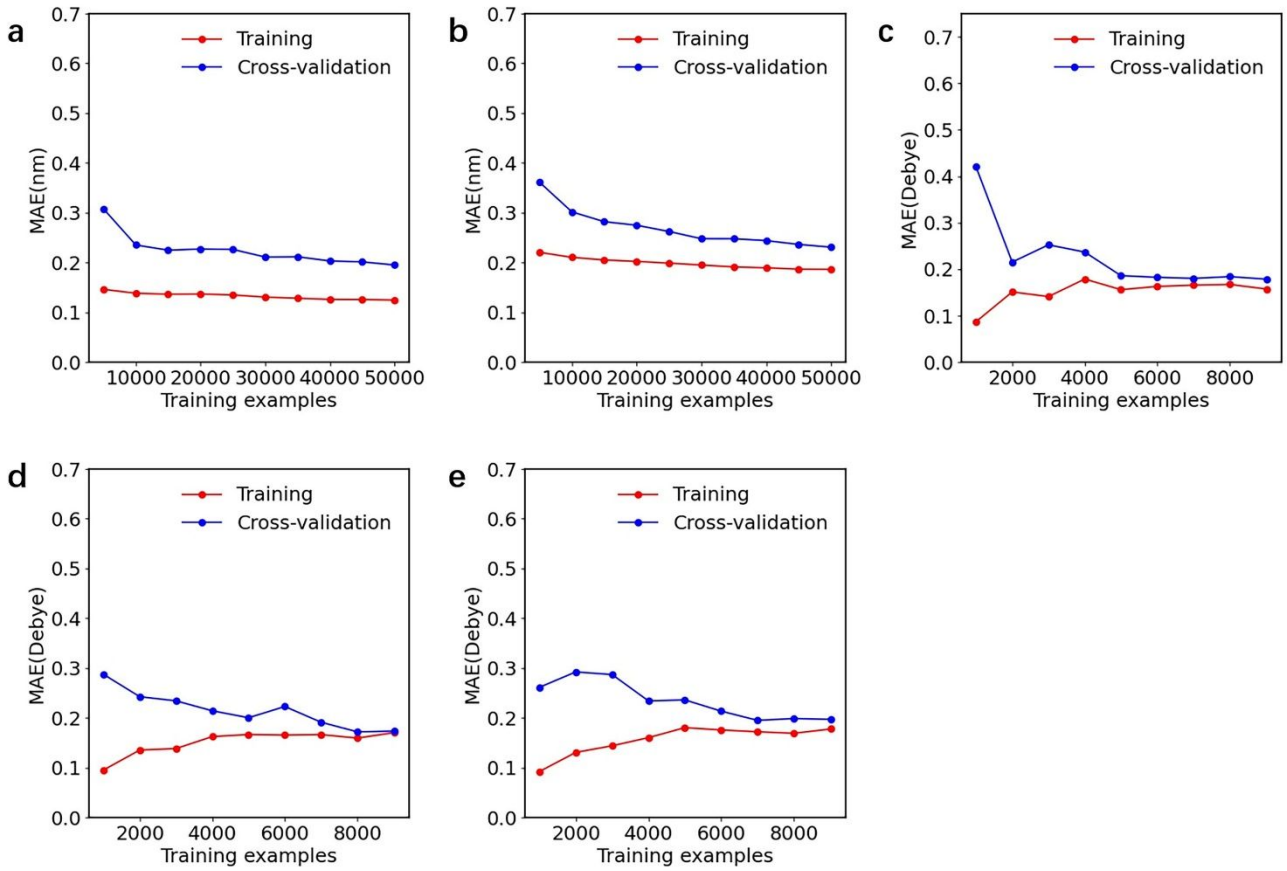

Figure S2. (a) The learning curve for predicting the excitation energies of NMA of the n → π* transition. (b) The learning curve for predicting the excitation energies of NMA of the π → π* transition. (c-e) The learning curves for predicting the ground state dipole moments of CYS in the x, y, z direction.

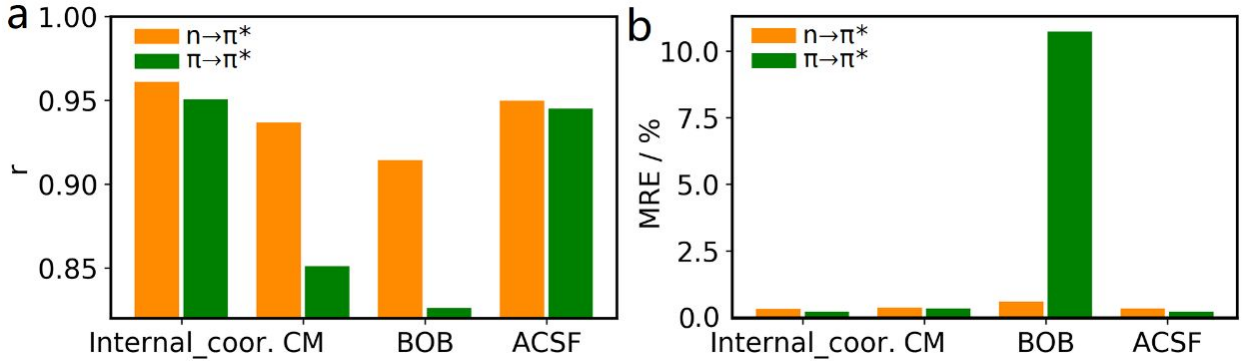

Figure S3. Comparison of results of ML prediction for peptide bond excitation energies with internal coordinates, coulomb matrix (CM), BOB (Bag of bonds) and atom-centred symmetry functions (ACSF) as molecular descriptors.

2.2 ML prediction of electric and magnetic transition dipole moments using CM and ACSF

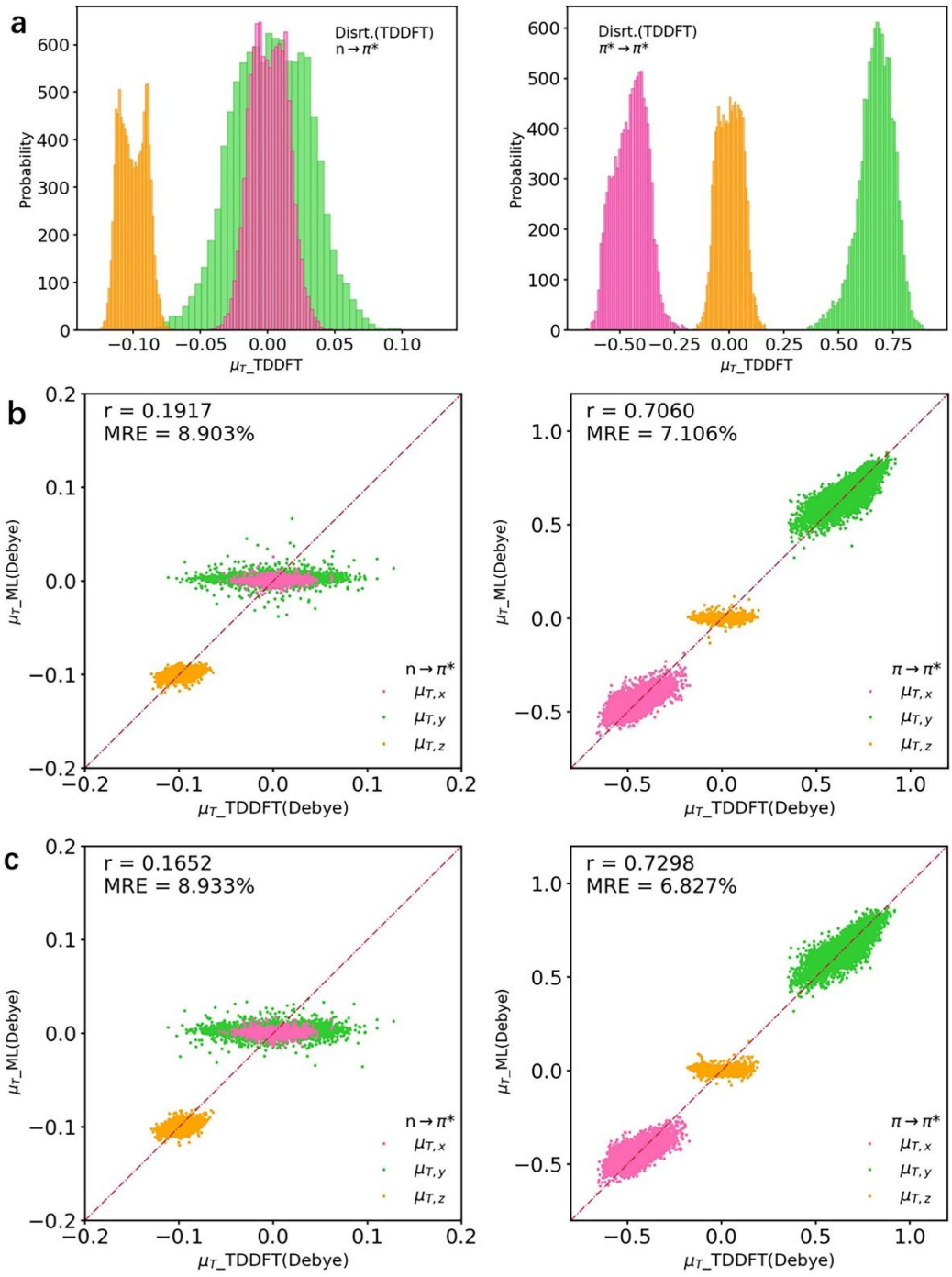

Figure S4. (a) Data distribution of the TDDFT calculated electric transition dipole moments of peptide bonds. (b) Correlation plots of the TDDFT and ML predicted electric transition dipole moments of the n→π* and π → π* transition using CM with GBR. (c) Same as (b) but using ACSF.

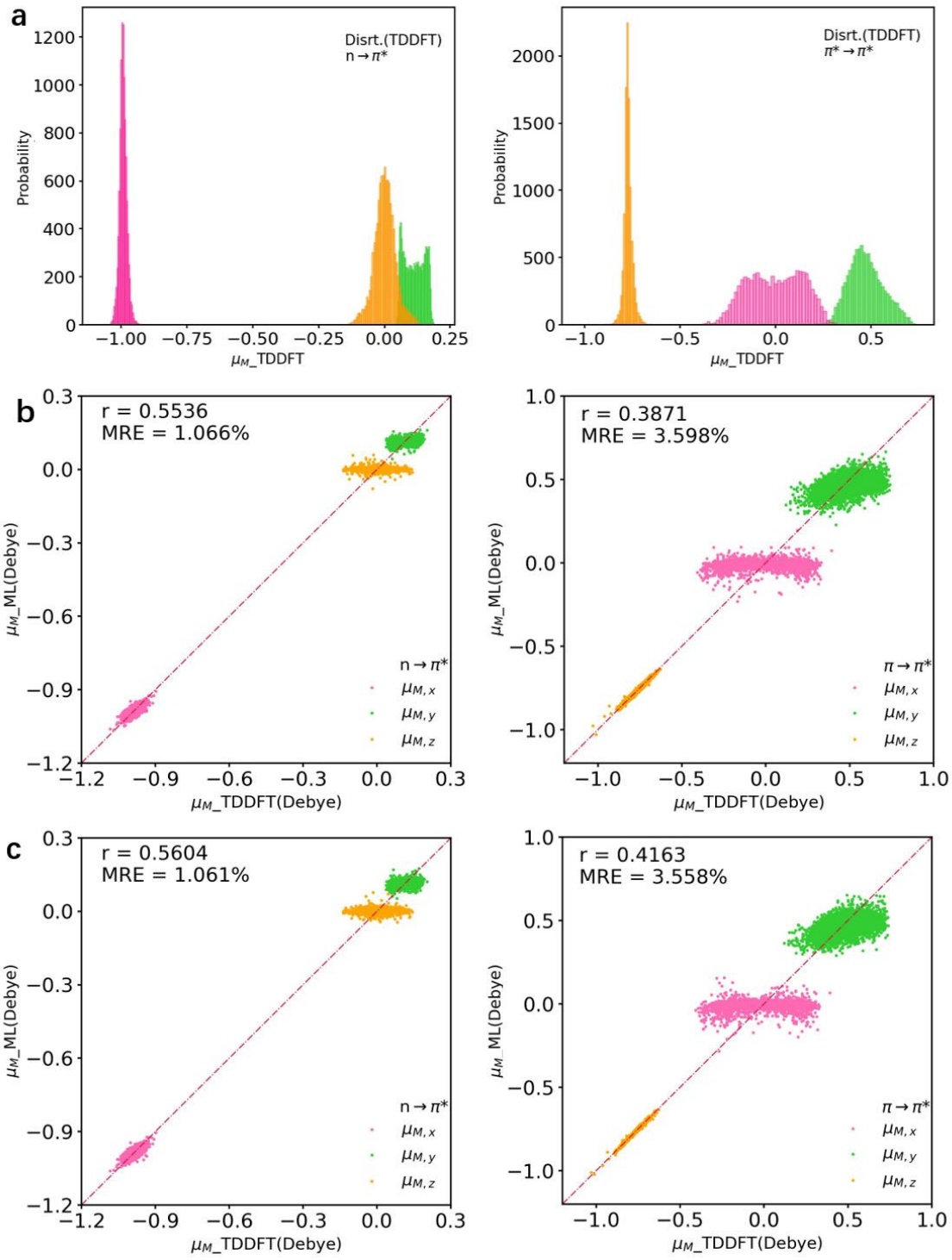

Figure S5. (a) Data distribution of the TDDFT calculated magnetic transition dipole moments of peptide bonds. (b) Correlation plots of the TDDFT and ML predicted magnetic transition dipole moments of the n→π* and π → π* transition using CM with GBR. (c) Same as (b) but using ACSF.

2.3 Proteins of interest in this study

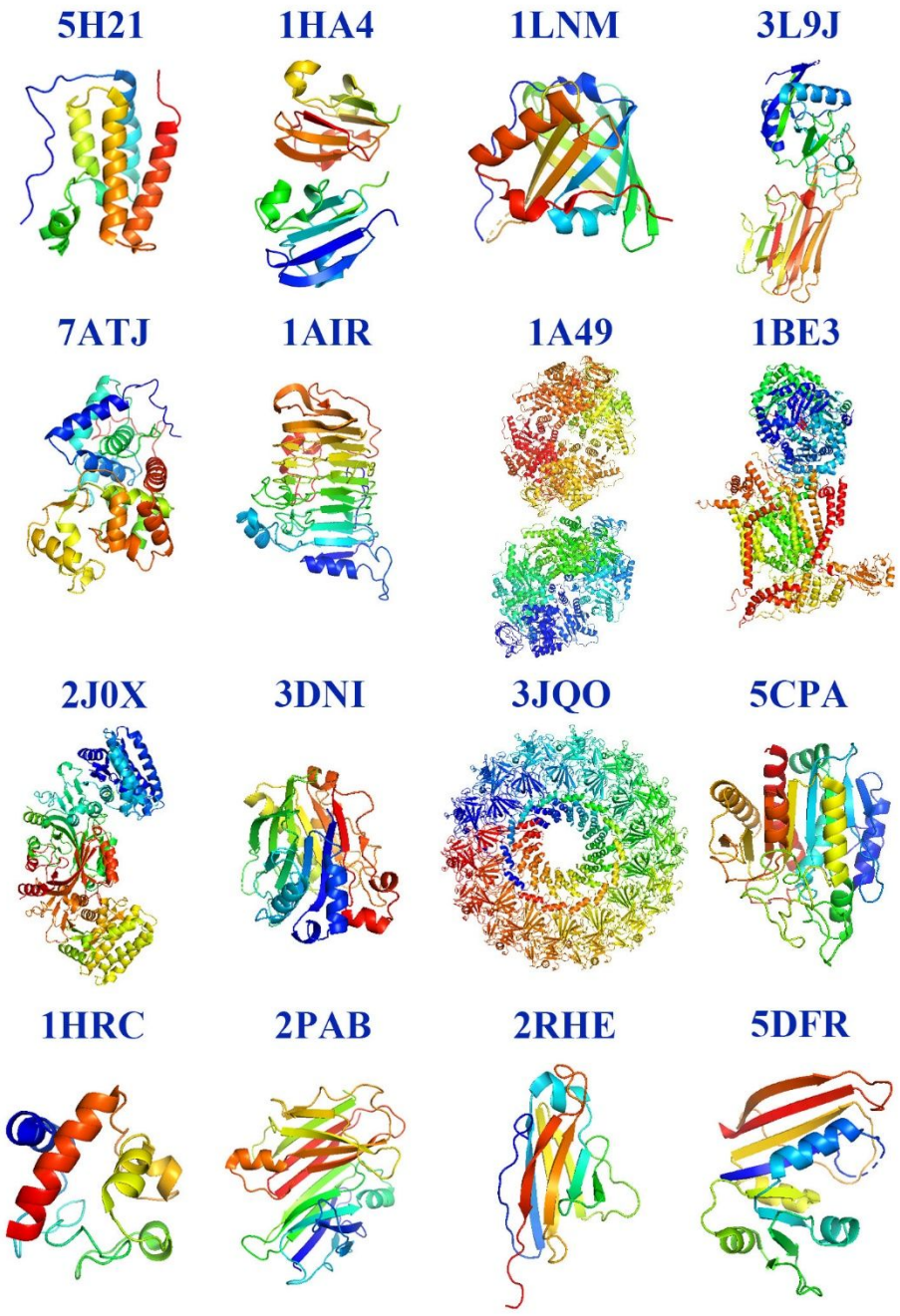

Figure S6. Protein structures (labeled by their respective PDB code) in the main text.

2.4 The comparison of DFT and ML simulated CD spectra

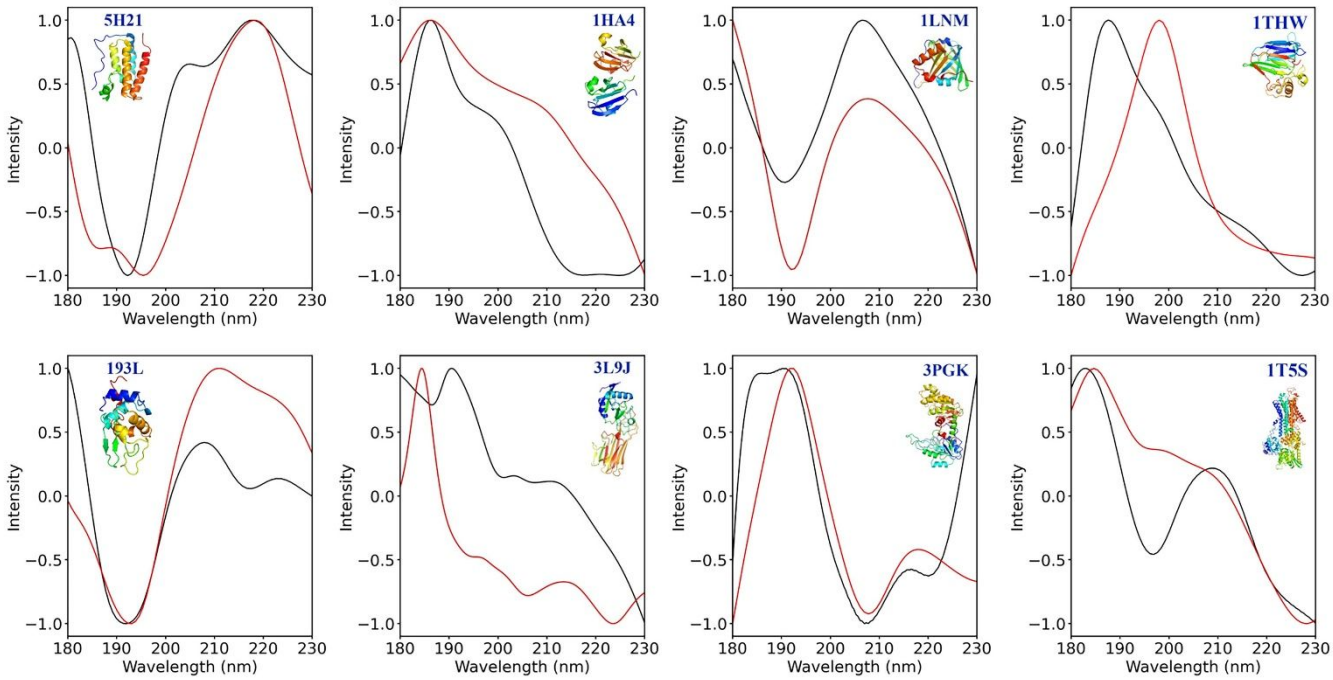

Figure S7. DFT (black curves) and ML (red curves) simulated CD spectra of different types of proteins (i.e., different fractions of α-helix and β-sheet) based on Frenkel exciton model. The ε_0, $\boldsymbol{\mu}_T$, $\boldsymbol{\mu}_M$, and $\boldsymbol{\mu}_G$ are first calculated with the DFT/ML methods. These parameters then construct an exciton Hamiltonian which is further diagonalized to get protein CD spectra. The electric and magnetic transition dipole moments of the peptide bonds constructing the CD in this figure are predicted by EANN.

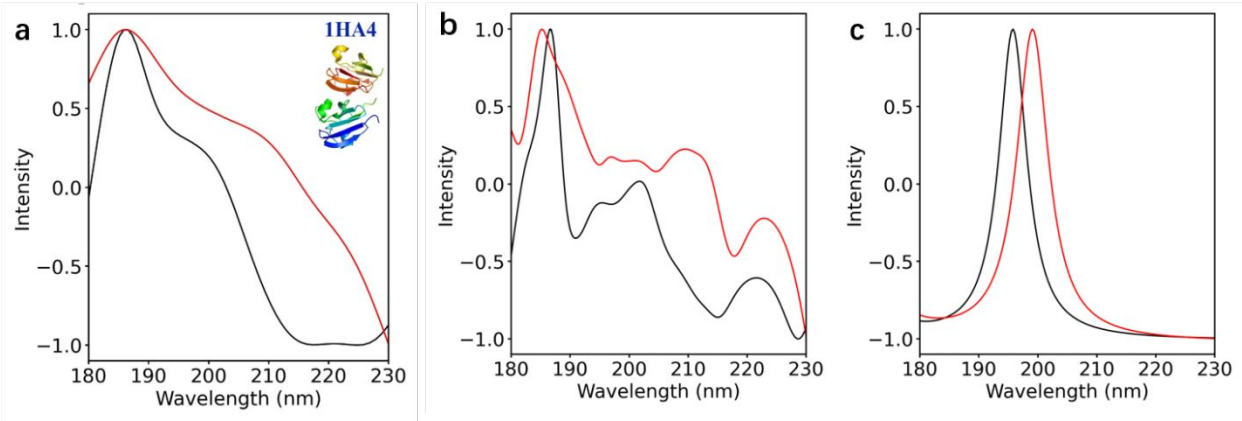

Figure S8. DFT (black curves) and ML (red curves) simulated CD spectra of the protein (PDB ID: 1HA4) based on the Frenkel exciton model. The broadening factor is set to (a) 1000 and (b) 750. (c) All the non-diagonal terms are set to zero and the simulated spectra are based on diagonal terms from DFT calculation and ML model, respectively. And the broadening factor is set to 750.

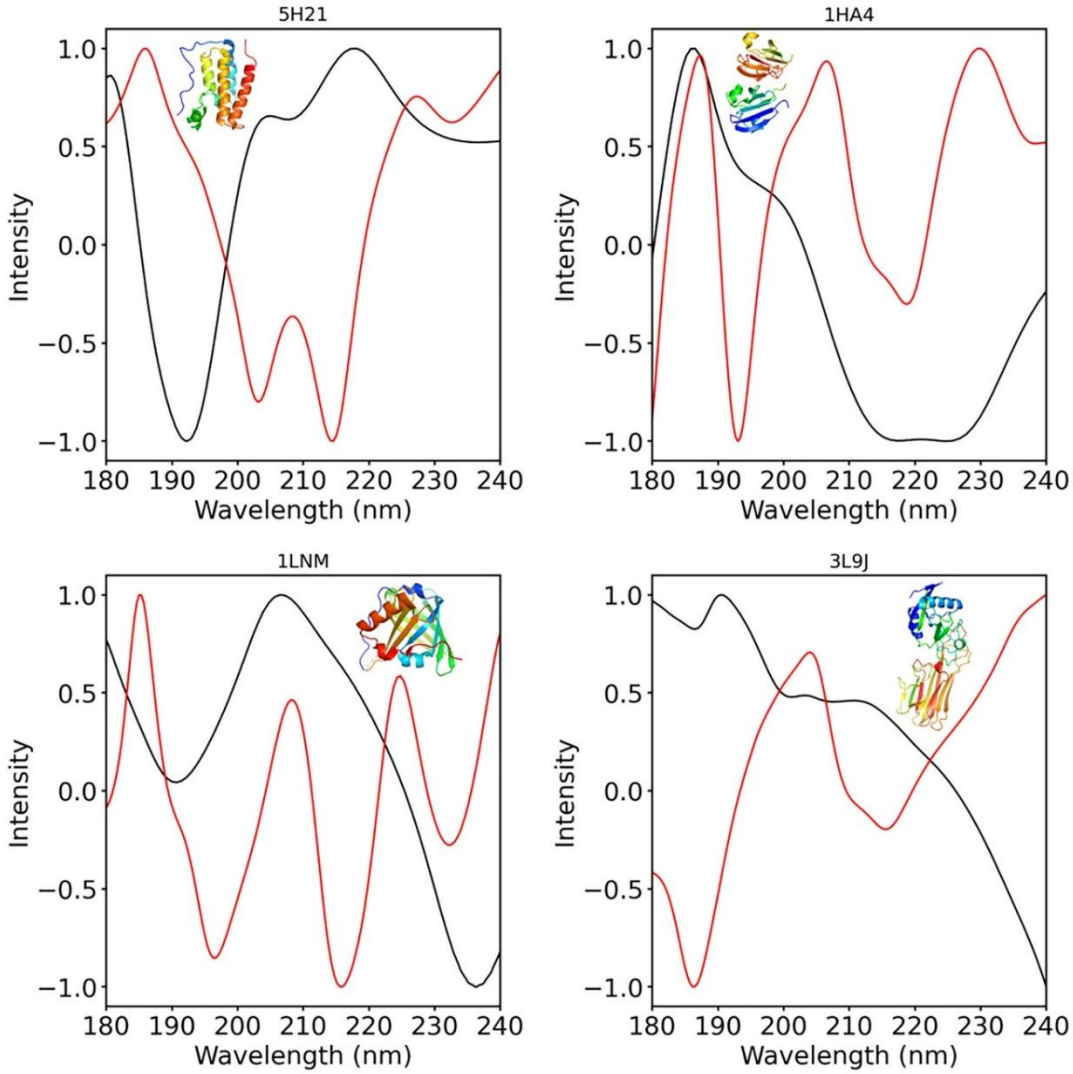

Figure S9. DFT (black curves) and ML (red curves) simulated CD spectra of different types proteins (i.e., different fractions of α-helix and β-sheet) based on Frenkel exciton model. The method is same as Fig. S6 but the electric and magnetic transition dipole moments of the peptide bonds constructing the CD in this figure are predicted by GBR with CM.

2.5 More experimental and ML predicted CD spectra of different types proteins

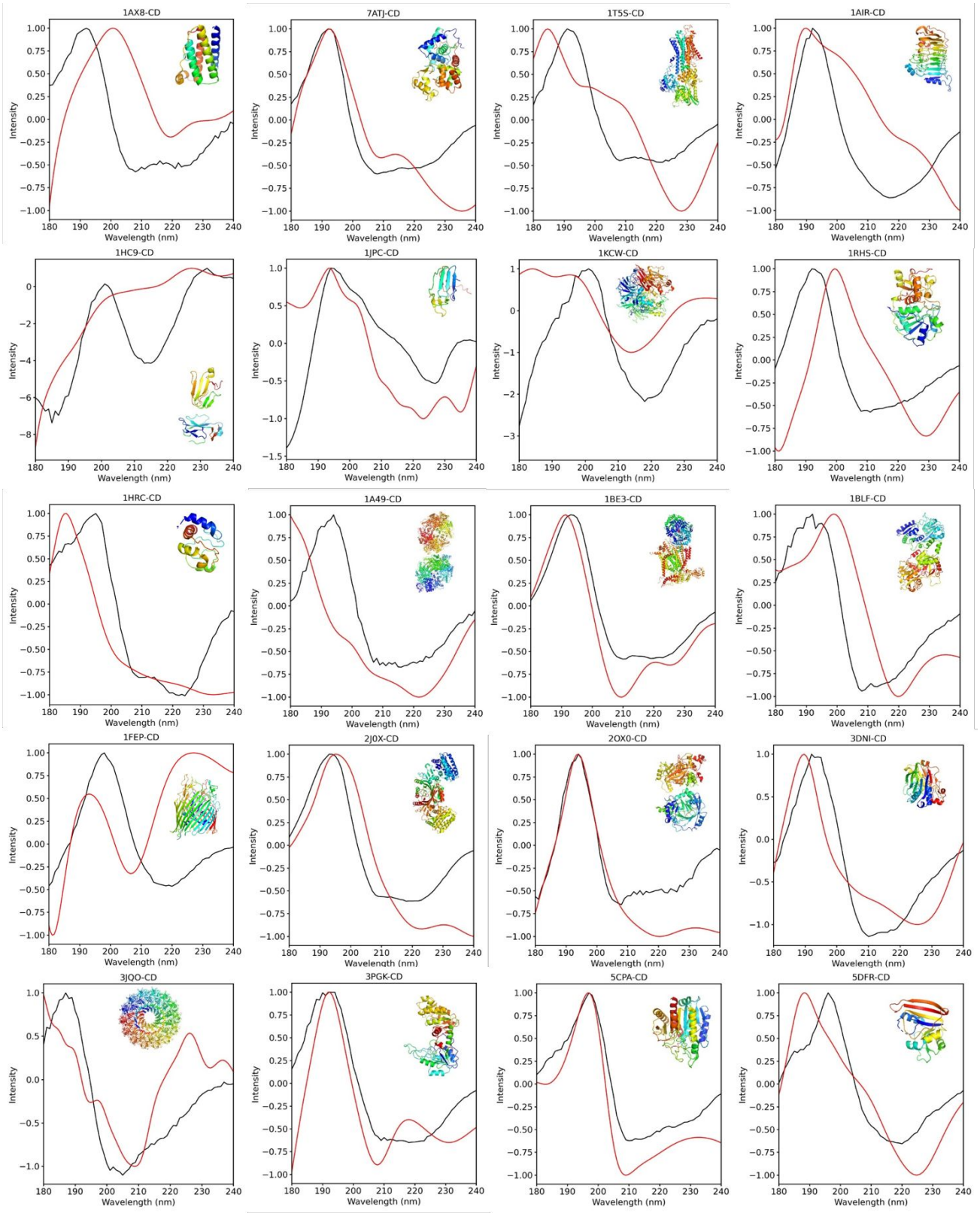

Figure S10. Experimental (black curves) and ML predicted (red curves) CD spectra of different types proteins (i.e., different fractions of α-helix and β-sheet). Intensity is scaled to have the same maximum intensity for each panel.

2.6 ML predicted IR spectra of four different types proteins based on 1000 MD configurations

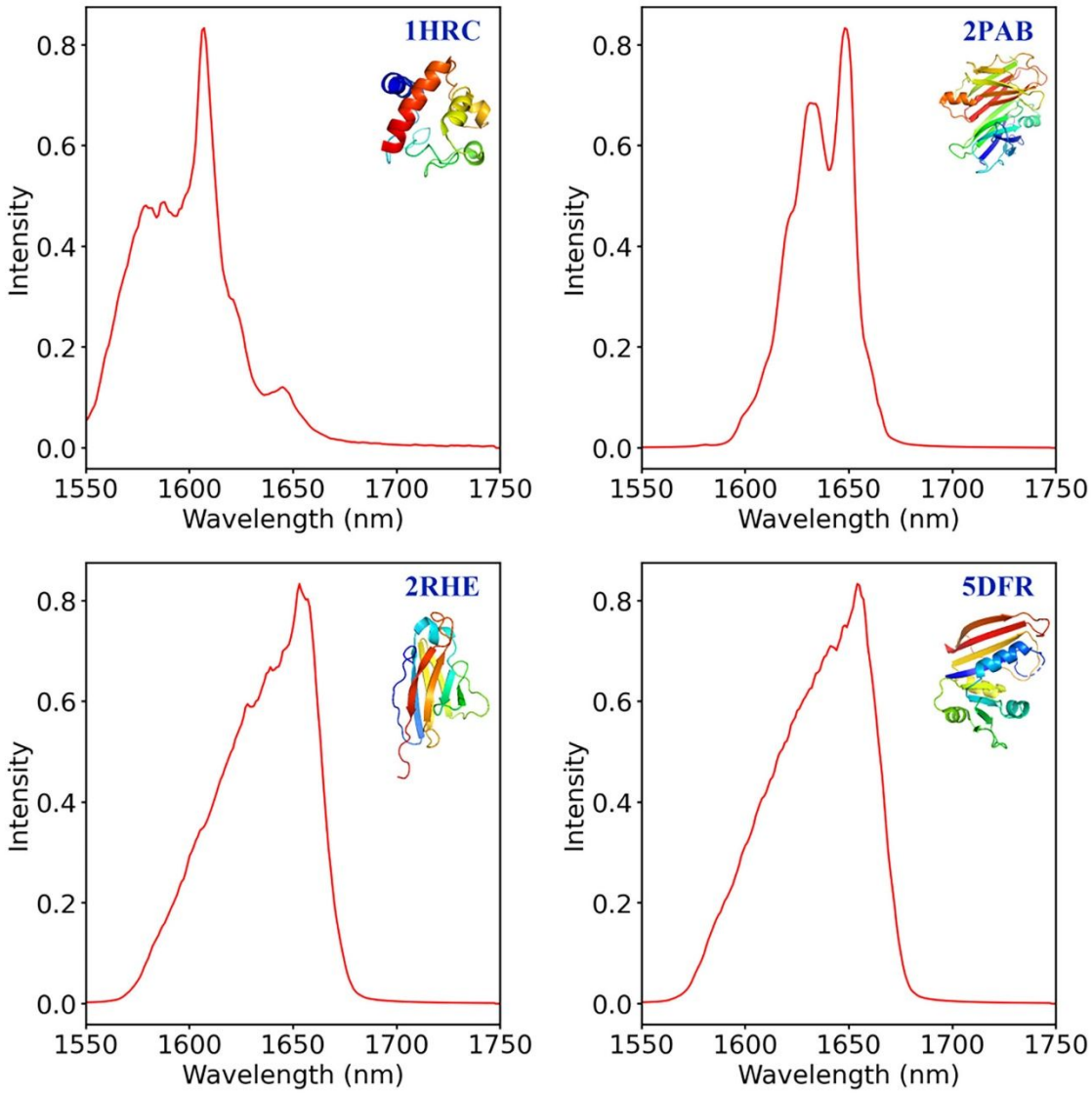

Figure S11. ML predicted IR spectra of four different types proteins (i.e., different fractions of α-helix and β-sheet). The ML predictions are based on 1000 MD configurations.

3 Supplementary Tables

Table S1. Comparison of the CD spectra simulated by DFT and the ML model based on Frenkel exciton model in terms of spearman rank correlation (ρ) and time ratio. This time is the total time for calculating the parameters required to construct the Hamiltonian.

Protein	PDB ID	Secondary class	Number of atoms	ρ	Time ratio (DFT/ML)
BrD4	5H21	α	1216	0.94	4214
Gamma S-Crystallin	1HA4	β	1558	0.97	3099
DigA16	1LNM	α+β	1340	0.86	2728.5
Thaumatin	1THW	α+β	1667	0.70	3589.5
Lysozyme	193L	α+β	1156	0.73	2177.5
TNFalpha	3L9J	α+β	2453	0.89	4782.5
Phosphoglycerate kinase	3PGK	α+β	3191	0.75	6878.5
ATPase	1T5S	α+β	7746	0.76	11002.7

Table S2. The averaged secondary structure contents of mini Trp-cage along its folding process. Each state is based on 100 MD conformations.

Content / State	Coil/%	β-Turn/%	α-helix/%	3_{10}-helices/%
S1	99.1	0.9	0.0	0.0
S25	73.0	25.0	0.0	0.0
S50	49.7	38.5	8.2	3.7
S75	48.9	29.9	19.7	1.6
S100	37.6	16.6	37.1	8.7

Supplementary References

1.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.

2.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.

3.Frenkel, J., On the Transformation of light into Heat in Solids. I. *Phys. Rev.* **1931,** *37* (1), 17-44.

4.Kasha, M.; Rawls, H. R.; Ashraf El-Bayoumi, M., The exciton model in molecular spectroscopy. *Pure Appl. Chem.* **1965,** *11* (3-4), 371-392.

5.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.

6.Rosenfeld, L., Quantenmechanische Theorie der natürlichen optischen Aktivität von Flüssigkeiten und Gasen. *Zeitschrift für Physik* **1929,** *52* (3), 161-174.

7.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.

8.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.

9.Mei-Rong, Y.; Yu, S.; Yong-Jin, X., Vibrational spectroscopic, NMR parameters and electronic properties of three 3-phenylthiophene derivatives via density functional theory. *SpringerPlus* **2014,** *3* (1), 701.

10.Xiao-Hong, L.; Xiang-Ru, L.; Xian-Zhou, Z., Calculation of vibrational spectroscopic and NMR parameters of 2-Dicyanovinyl-5-(4-N,N-dimethylaminophenyl) thiophene by ab initio HF and density functional methods. *Comput. Theor. Chem.* **2011,** *969* (1), 27-34.

11.Karthick, T.; Balachandran, V.; Perumal, S.; Nataraj, A., Vibrational (FT-IR and FT-Raman) spectra and quantum chemical studies on the molecular orbital calculations, chemical reactivity and thermodynamic parameters of 2-chloro-5-(trifluoromethyl) aniline. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2013,** *107*, 72-81.

12.Laurent, A. D.; Jacquemin, D., TD-DFT benchmarks: A review. *Int. J. Quantum Chem* **2013,** *113* (17), 2019-2039.

13.Leang, S. S.; Zahariev, F.; Gordon, M. S., Benchmarking the performance of time-dependent density functional methods. *The Journal of Chemical Physics* **2012,** *136* (10), 104101.

14.De Silva, N.; Willow, S. Y.; Gordon, M. S., Solvent Induced Shifts in the UV Spectrum of Amides. *The Journal of Physical Chemistry A* **2013,** *117* (46), 11847-11855.

15.Westermayr, J.; Gastegger, M.; Menger, M. F. S. J.; Mai, S.; González, L.; Marquetand, P., Machine learning enables long time scale molecular photodynamics simulations. *Chem. Sci.* **2019,** *10* (35), 8100-8107.

16.Lu, T.; Chen, F., Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **2012,** *33* (5), 580-592.

17.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. *Gaussian 16 Rev. A.03*, Wallingford, CT: 2016.

18.Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X., TensorFlow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, USENIX Association: Savannah, GA, USA, 2016; pp 265–283.
19.Maas, A. L.; Hannun, A. Y.; Ng, A. Y. In *Rectifier Nonlinearities Improve Neural Network Acoustic Models*, Proceedings of the 30th International Conference on Machine Learning: 2013; p 3.
20.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: Banff, Alberta, Canada, 2004; p 78.
21.Kingma, D.; Ba, J., Adam: A Method for Stochastic Optimization. *arXiv: Learning* **2014**.
22.Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.; Sheridan, R.; Feuston, B., Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comp. Sci.* **2003,** *43 6*, 1947-58.
23.Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012,** *108* (5), 058301.
24.Behler, J., Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011,** *134* (7), 074106.
25.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.
26.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.
27.Gessulat, S.; Schmidt, T.; Zolg, D.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **2019,** *16*.
28.Hirst, J. D.; Colella, K.; Gilbert, A. T. B., Electronic Circular Dichroism of Proteins from First-Principles Calculations. *The Journal of Physical Chemistry B* **2003,** *107* (42), 11813-11819.
29.Besley, N. A.; Hirst, J. D., Theoretical Studies toward Quantitative Protein Circular Dichroism Calculations. *J. Am. Chem. Soc.* **1999,** *121* (41), 9636-9644.
30.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.
31.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.
32.Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C., GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005,** *26* (16), 1701-1718.
33.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.