THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

A Machine-Learning Protocol for Ultraviolet Protein-Backbone Absorption Spectroscopy under Environmental Fluctuations

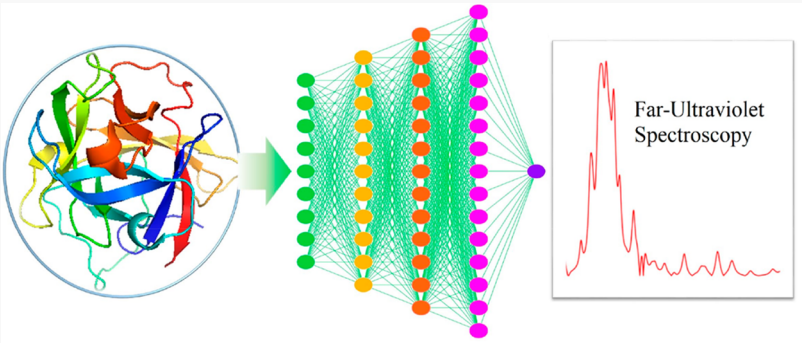
Published as part of The Journal of Physical Chemistry virtual special issue "125 Years of The Journal of Physical Chemistry".

Jinxiao Zhang,[‖] Sheng Ye,[‖] Kai Zhong,[‖] Yaolong Zhang, Yuanyuan Chong, Luyuan Zhao, Huiting Zhou, Sibei Guo, Guozhen Zhang, Bin Jiang, Shaul Mukamel, and Jun Jiang*

Cite This: *J. Phys. Chem. B* 2021, 125, 6171−6178

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Ultraviolet (UV) absorption spectra are commonly used for characterizing the global structure of proteins. However, the theoretical interpretation of UV spectra is hindered by the large number of required expensive ab initio calculations of excited states spanning a huge conformation space. We present a machine-learning (ML) protocol for far-UV (FUV) spectra of proteins, which can predict FUV spectra of proteins with comparable accuracy to density functional theory (DFT) calculations but with 3−4 orders of magnitude reduced computational cost. It further shows excellent predictive power and transferability that can be used to probe structural mutations and protein folding pathways.

1. INTRODUCTION

Protein structure determination is vital for understanding their function.[1] Spectroscopy is a primary tool for accomplishing this goal.[2,3] Ultraviolet electronic absorption has the ability to monitor the photoresponse of proteins with high sensitivity.[4−7] The far-ultraviolet absorption (FUV, 180−240 nm) region represents the n → π^* and $\pi \rightarrow \pi^*$ electronic transitions in the peptide skeleton, is the basis for several powerful spectroscopy techniques, including linear dichroism (LD) for measuring orientation information,[8] circular dichroism (CD) for identifying the optical isomerism and secondary structures,[9] and two-dimensional ultraviolet absorption (2DUV) for tracking energy transfer.[10] However, the practical utilization of FUV spectroscopy has been limited by the expensive required theoretical effort.[5] A major challenge has been the mapping between FUV signals and protein structures under environmental fluctuations which requires electronic structure calculations of thousands of molecular dynamics (MD) conformations in order to properly sample fluctuations of the proteins and the surrounding solvent.

Accurate FUV spectra of molecules with hundreds of atoms are accessible with the quantum mechanical (QM) method. For example, the sTDA-xTB method is a fully QM method for calculating UV−vis and CD spectra averaged along structures from a short MD simulation for biomolecules with up to thousands of atoms.[11] However, repeated QM computations for thousands of MD structures or more and environmental fluctuations are still time demanding. In addition, for larger proteins with tens of thousands of atoms, it is challenging for QM calculation to simulate UV−vis and CD spectra. A Frenkel exciton Hamiltonian that assumes local excitations in each structural unit with pairwise coupling of excitations can be employed. It has been widely employed in previous theoretical studies of energy transfer within peptide backbones in the FUV

Received: April 12, 2021
Revised: May 14, 2021
Published: June 4, 2021

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

region. However, computing the electronic couplings involves expensive calculations of two-electron integrals.[12] Approximate semiempirical methods based on empirical parametrized Hamiltonians derived from atomic coordinates are less expensive.[13] Maps are created by fitting polynomial functions of the electric field or the electrostatic potential at reference points obtained from either high-level theoretical simulations or experiment.[14] Some limitations of traditional spectroscopy map methods include the fitting errors when applied to structures that are too distant from the structures in the training set, unreliable local electrostatics caused by the unoptimized force-field charge, and the limitation of employing simple models to represent highly complex spectra.[15]

Over the past decade, machine learning (ML) has shown enormous potential in chemical science, including synthesis prediction, drug design, materials discovery, and spectra simulations.[16−21] In particular, deep learning, which is a subset of ML methods built on artificial neural networks (ANNs) with feature learning, has been extensively used for solving complex nonlinear problems in chemistry, such as the prediction of spectra, dipole moments, and protein structure.[22−24] ML holds clear advantages by carrying out a very high-dimensional regression on many different features to map molecular properties so as to construct structure−property relationships. ML-driven generalization of spectroscopic mapping procedures should provide an efficient high-throughput framework for the FUV spectroscopy simulations of proteins.

Here we develop a ML protocol for FUV spectra of proteins in a fluctuating environment. A set of structural descriptors is identified during the ML training process, which helps establish structure−property relationships. Our protocol gives a reasonably good prediction of FUV spectra for various proteins with comparable accuracy to much more expensive density-functional-theory (DFT)-based simulations and to experiment. The computational cost is 3−4 orders of magnitude lower than DFT. Structure changes caused by protein mutations and folding pathways under environmental fluctuations described by MD ensembles are predicted.

2. THEORETICAL METHODS

2.1. Theory of the Calculation of FUV Spectra. A protein is made of a polypeptide backbone and multiple amino acid residues (Figure 1a). Its FUV photoresponse is mostly derived from electronic excitations of the peptide backbone coupled with environmental fluctuations. Our divide-and-conquer strategy uses the Frenkel exciton model Hamiltonian for the electronic excitations.[12,25]

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

The indices $m(n)$ run over the peptide bonds and $a(b)$ represents n → π^* or $\pi \rightarrow \pi^*$ transitions. $\hat{B}_{ma}^{\dagger}$ and $\hat{B}_{ma}$ are creation and annihilation operators of excitations of the mth peptide bond, respectively. ε_{ma} is the excitation energy of peptide m, $J_{ma,nb}$ is the resonant coupling between the ath excited state of peptide m and the bth excited state of peptide n. Here we used the dipole approximation to compute the electrostatic interaction between two subjects;[26,27] more details can be found in the Supporting Information.

2.2. Simulation of FUV Spectra. The computational bottleneck in simulations of FUV spectra is getting the

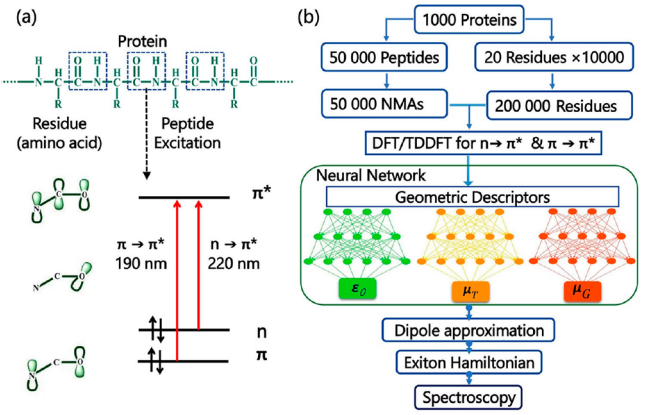


Figure 1. (a) Protein structure and the two electronic transitions of peptide (n → π^* and $\pi \rightarrow \pi^*$) which contribute to FUV adsorption. (b) Schematic of the machine-learning protocol for FUV protein spectroscopy.

parameters $\varepsilon_{0,ma}$, the electronic transition dipole moments of the peptide bond (m, a) $\boldsymbol{\mu}_{T,ma}$, and the ground state dipole moment of a residue l $\boldsymbol{\mu}_{G,l}$ in eqs 4 and 5 in the Supporting Information, which requires expensive DFT/time-dependent-DFT (TDDFT) computations for various MD conformations. We predict these three parameters from protein geometric descriptors based on a large data set of DFT/TDDFT calculations. The protocol involves five steps (Figure 1b): (1) The data sets of 50 000 peptides and 200 000 residues (10 000 residue structures for each type of amino acid, 20 in total) are randomly harvested from 1000 different types of proteins taken from the RCSB Protein Data Bank,[28] which ensure the data set diversity (Table S2). Each peptide bond in this data set is modeled as a *N*-methylacetamide (NMA) molecule. The residues are directly extracted from PDB files and link the dangling bonds with hydrogen atoms. Here, residues (H_2N−CHR−COH) refer to the remaining parts of each amino acid after their formation of protein-backbone peptides and removing of waters. TDDFT simulations for the data set based on n → π^* or $\pi \rightarrow \pi^*$ transitions at the PBE0/cc-pVDZ level and phase correlation[29] were performed for each NMA molecule to obtain its excitation energy (ε_0) and transition dipole moments ($\boldsymbol{\mu}_T$). PBE0/cc-pVDZ has been successfully used to calculate the excitation properties of NMA molecules in previous work.[30,31] The n → π^* transition is derived from the transition from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO), for which the TDDFT excitation energies are in agreement with experiment.[32] The $\pi \rightarrow \pi^*$ transition exhibits a multireference feature involving higher excited states, which is challenging even for a highly accurate post Hartee−Fock method such as EOM-CCSD.[30,33] Moreover, the calculated transition energy error of TDDFT is within acceptable limits (0.3 eV).[34,35] Given the limited computational resource and relative high accuracy of TDDFT, PBE0/cc-pVDZ makes a reasonable choice for the simulations of NMA molecules. (2) We further looked at individual residues in this data set (Figure S6). In this work, we aim at obtaining accurate ground state and excited state properties for amino acid residues and peptide bonds, respectively, so we choose different combinations of the DFT method and basis set to handle different properties that they are best for. B3LYP is widely used to calculate the ground state dipole moment and

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

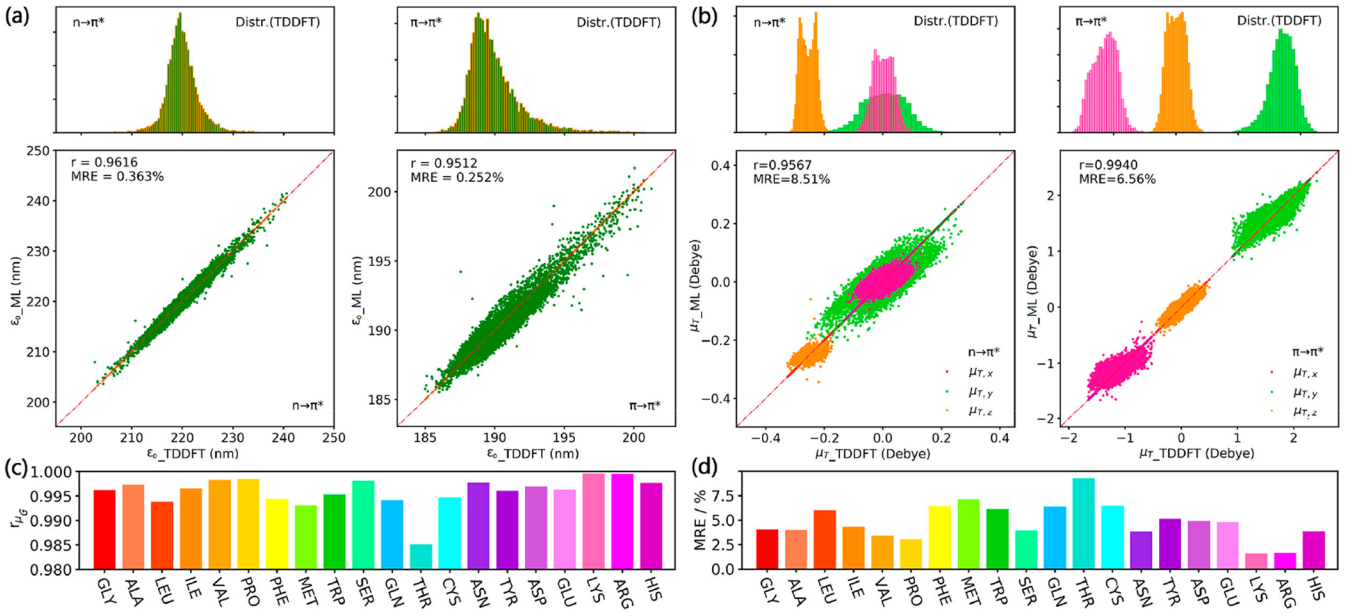


Figure 2. (a) From top to bottom: Data distribution and correlation plots of the TDDFT and ML predicted excitation energies of the n → π^* and $\pi \rightarrow \pi^*$ transitions of peptides. The diagonal orange lines/dots in the bottom column represent excitation energies calculated with TDDFT. (b) Same as part a but for transition dipole moments. The transition dipole moments in the *x*, *y*, and *z* directions are distinguished by pink, green, and orange. In the bottom column, the diagonal pink, green, and orange lines/dots represent the $\boldsymbol{\mu}_{T,x}$, $\boldsymbol{\mu}_{T,y}$, and $\boldsymbol{\mu}_{T,z}$ calculated with TDDFT, respectively. (c) Pearson correlation coefficients (*r*) of 20 amino acid residues. (d) The mean relative errors (MREs) of 20 amino acid residues.

tends to show a relatively small error.[36,37] The diffuse function is necessary for the calculation of ground state dipole moment, and 6-311++G** is a cost-effective choice. Therefore, we used the B3LYP/6-311++G** method to compute the ground state dipole moments of all amino acid residues. All DFT/TDDFT simulations are performed using the Gaussian 16 package.[38] (3) ML (neural network, NN) training. Internal coordinates, embedded density,[39] and converted Cartesian coordinates are chosen as the molecular descriptors for the input layer of a deep learning NN. Starting with the DFT/TDDFT data sets, we run the data-training process for the data set (80% for training and 20% for validation) to build the correlation between the descriptors and our prediction targets of ε_0, $\boldsymbol{\mu}_T$ (for peptide), and $\boldsymbol{\mu}_G$ (for residue). A deep learning protocol containing three hidden layers (with 32, 64, and 128 neurons, respectively) and L2 regularization is employed for data training of ε_0 for peptide and $\boldsymbol{\mu}_G$ for residue. Embedded atom neural networks (EANNs) which have been reported in our previous work have been employed for the prediction of $\boldsymbol{\mu}_T$ of peptide.[39] (4) ML prediction of the exciton Hamiltonian. Using the trained ML protocol, we input the geometry of a new protein outside the data set into the NN and predict ε_0, $\boldsymbol{\mu}_T$, and $\boldsymbol{\mu}_G$ parameters without additional quantum chemistry calculations. (5) The exciton Hamiltonian is diagonalized, and the FUV spectrum of the selected protein is calculated using the SPECTRON package.[40] More details can be found in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. Machine-Learning Prediction for Peptides and Residues. The peptide n → π^* and $\pi \rightarrow \pi^*$ transitions appear at ∼220 and ∼190 nm, respectively (Figure 1a). Four molecular descriptors, including internal coordinates, Coulomb matrix (CM),[41] bag of bonds (BOB),[42] and atom-centered symmetry functions (ACSFs),[43] have been tested for the prediction of ε_0, and the internal coordinates appear to be the best (Figure S5). We chose internal coordinates as the molecular descriptor for ε_0 because they directly reflect the fundamental structure−property relationship and only involve nine internal coordinates for each peptide bone. The Pearson relative coefficient (*r*) and the mean relative error (MRE) were used to estimate the accuracy and robustness of the trained ML model. TDDFT-based ε_0 and $\boldsymbol{\mu}_T$ and DFT-based $\boldsymbol{\mu}_G$ are used as reference values for ML training. ML gives excellent predicted ε_0 for both n → π^* and $\pi \rightarrow \pi^*$ transitions, with *r* being 0.9616 and 0.9512 and MRE being 0.363 and 0.252% (Figure 2a), respectively. For $\boldsymbol{\mu}_T$, we employed our previous proposed embedded atom neural network (EANN) in which we evaluate the density-like descriptors as the square of linear combination of Gaussian type atomic orbitals.[39] Similarly, we obtained a good prediction from the ML model (Figure 2b), as evident by *r* (>0.95) and MRE (<10%). $\boldsymbol{\mu}_T$ are more challenging for ML than ε_0 because they are vectors rather than scalars. For the $\boldsymbol{\mu}_G$ simulations, we reoriented all Cartesian coordinates to the same reference system before ML predictions. Again, for all types of residues, ML models give very good predictions (Figure 2c,d) with large *r* (>0.985) and small MRE (<10%). These results indicate the accuracy and robustness of our trained ML model.

3.2. Machine-Learning Prediction of FUV Spectra for Proteins. We now apply the ML models for ε_0, $\boldsymbol{\mu}_T$ of the peptide bond, and $\boldsymbol{\mu}_G$ for the residue, to predict these parameters for new proteins, construct the exciton Hamiltonian, and obtain FUV spectra. Note that the proteins presented in Figure 3 and Figure S8 are not included in the 1000 proteins employed for extracting of peptide and residues for ML training. We have first simulated FUV spectra of 12 proteins whose structures were randomly retrieved from the RCSB Protein Data Bank. Our ML-based approach agrees well with the DFT-based approach for all types of proteins (Figure

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

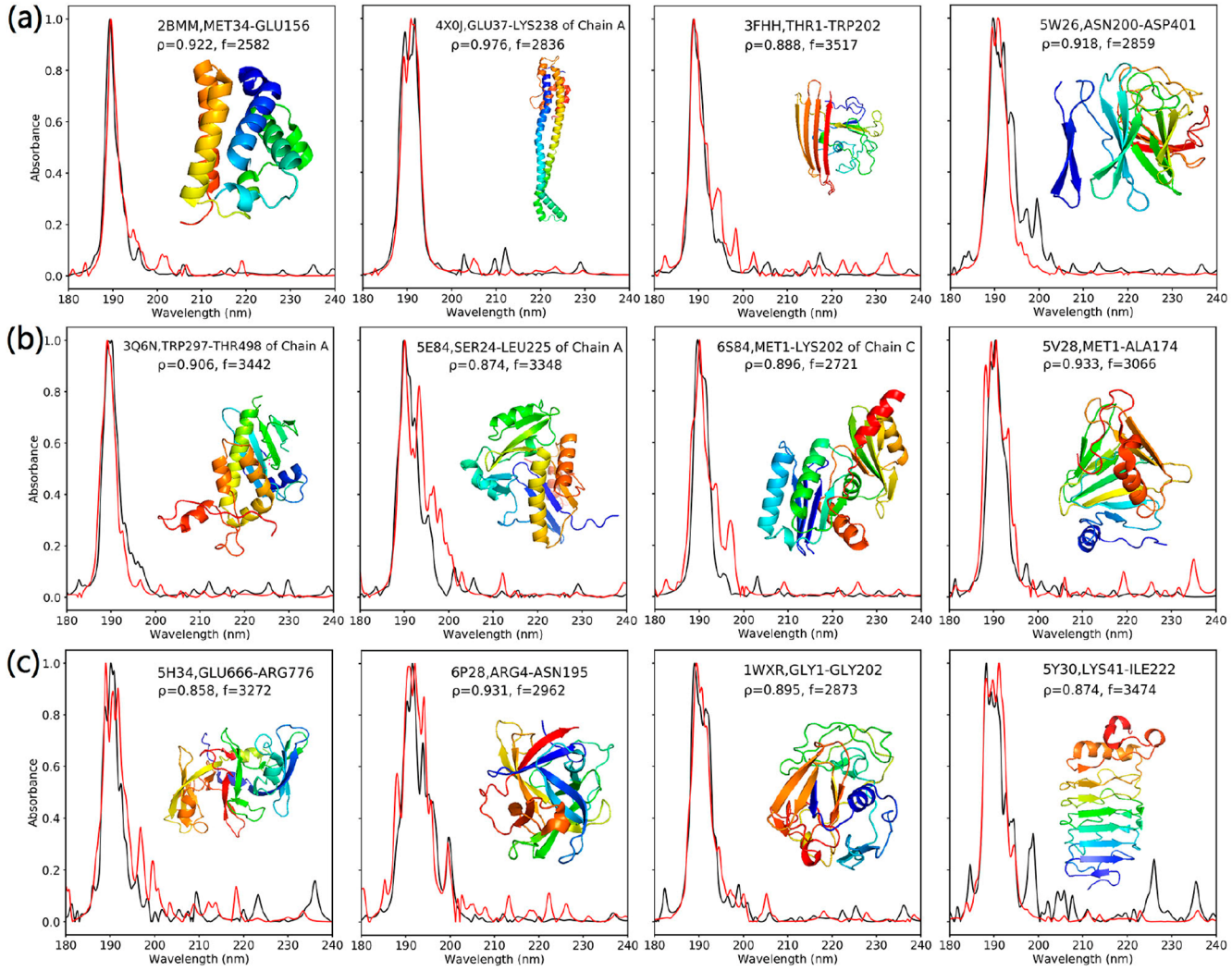


Figure 3. FUV spectra of 12 proteins (a: α-helix, β-sheet; b, c: α-helix + β-sheet) calculated with the DFT/TDDFT (black curves) and ML (red curves) methods. The excitation energy (ε_0), transition dipole moment ($\boldsymbol{\mu}_T$) of the peptide bond, and ground state dipole moment of each residue ($\boldsymbol{\mu}_G$) are first calculated with the DFT/ML methods. They are then employed as inputs for the construction of an exciton Hamiltonian and further diagonalized to acquire FUV spectra. ρ, Spearman rank correlation coefficients; f, time ratio (DFT/ML). Predicted FUV spectra of 230 proteins are plotted in Figure S8.

3) in terms of positions of peaks and line shapes. This was further demonstrated by high Spearman rank correlation coefficients[44] ($\rho > 0.80$), which reveals the quantitative agreement between the predicted and reference spectra. We then expanded the scope of investigation to a larger pool with 230 different proteins (Figure S8). For most proteins of interest, ML gives comparable FUV spectra to DFT, indicating the robustness and transferability of the trained ML model.

The good transferability can be rationalized as follows: (1) The large training data set covers different types of proteins, which ensures its diversity. (2) Each property has a favorable molecular descriptor after careful selection of descriptors. The domain parameters of FUV spectra are ε_0 and $\boldsymbol{\mu}_T$ of peptide. The internal coordinate is selected as a molecular descriptor for $\boldsymbol{\mu}_T$ and can construct the fundamental structure−property relationship very well. The EANN approach can effectively fit the value and orientation of the transition dipole vector. (3) We optimize the hyperparameters to create a unique NN for each property, which have showed excellent prediction power in our previous work. The present FUV spectra form the basis for predicting the more informative LD, CD, and 2DUV signals; extending the present protocol to LD, CD, and 2DUV spectra is a future direction. ML is significantly (3 orders of magnitude for most proteins with around 200 amino acid residues) faster than DFT in generating the model Hamiltonian needed for FUV spectra simulation (Table S1). For larger proteins containing more than 1000 amino acid residues (PDB: 3V03), the speed-up is even greater (4 orders of magnitude).

To take the fluctuating environment into account, we conducted classical MD simulations to generate trajectories for equilibrated protein structures. We had harvested 1000 MD conformations and computed their FUV spectra using our ML protocol. The averaged spectra are in good agreement with experiment (Figure 4a).[45,46] FUV spectroscopy is extremely sensitive to a variety of processes and excitation effects ($\sigma \rightarrow \sigma$ transitions, $\pi \rightarrow \pi^*$ electronic transitions, charge-transfer transitions, Rydberg transitions, electronic transfer and reaction process, etc.), solvents (water, oxygen, alkanes, and alcohols), and the surrounding environment (concentration,

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

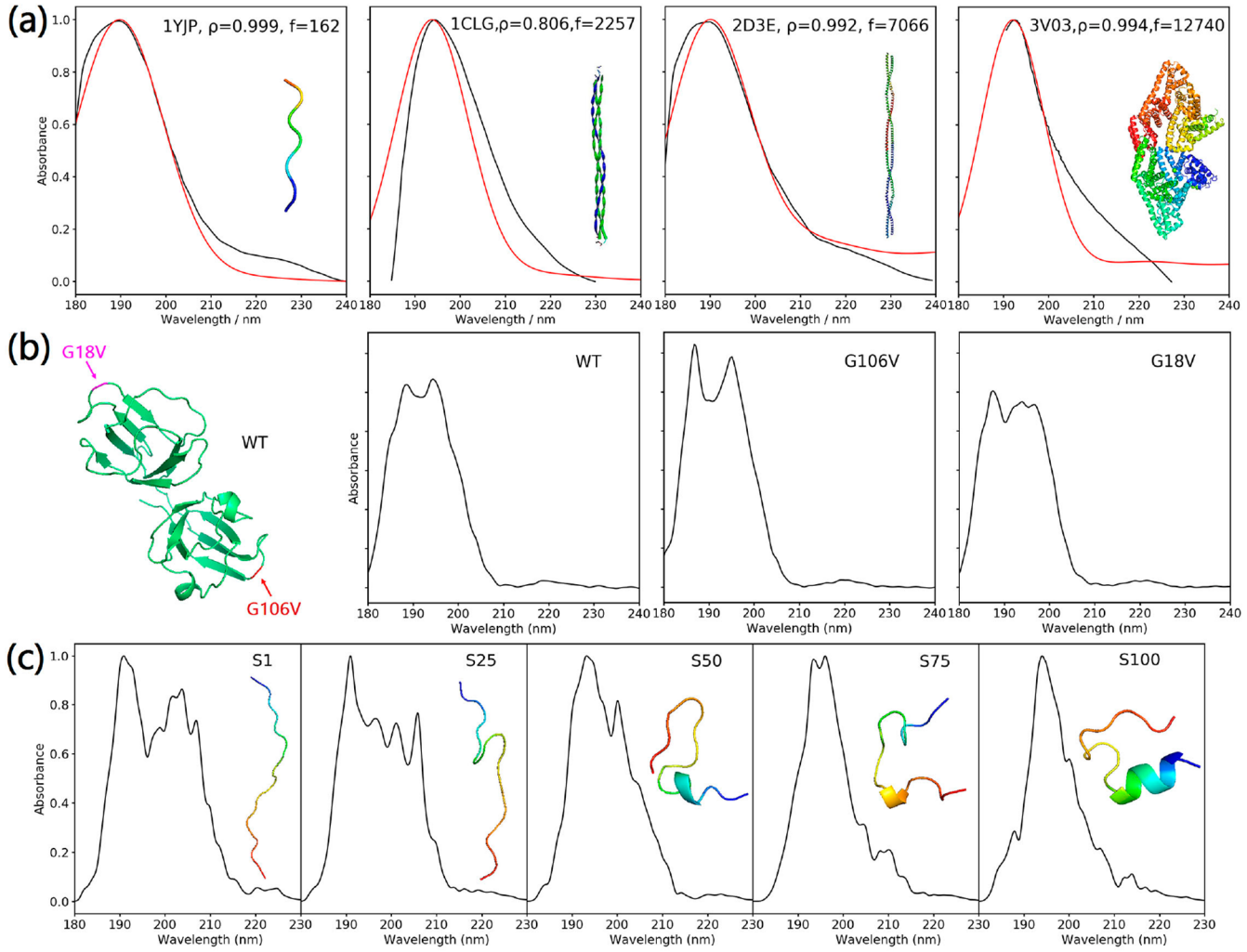


Figure 4. (a) Experimental (black curves) and ML predicted (red curves) FUV spectra based on 1000 MD conformations (β-sheet, α-helix, α-helix, α-helix). (b) The ML simulated FUV spectra based on 100 MD conformations for minoring the mutation of protein. The original structure wild type (WT), symmetry related G106V, and character related variant G18V. (c) The ML simulated FUV spectra of the Trp-cage protein along its folding path (S1, the initial unfolded structure; S100, the final folded structures). Each state is based on 100 MD conformations.

temperature, pH, etc.).[4,30,33,47] These result in broad peaks, which erodes its information. More distinguishable experimental FUV features can be obtained at low temperatures.

Protein aggregation is essential in some human diseases and functional amyloids.[48] Wild type (WT) γS-crystallin plays an important role in maintaining eye lens transparency, and its aggregation leads to cataract or opacification of the lens. There are two mutations in WT protein, including a symmetry related G106V variant and G18V variant which are associated with early onset cataract.[49] These minor structural mutations cause different aggregation tendencies in the order of WT < G106V < G18V. Figure 4b shows the FUV spectra of WT and its variants based on 100 MD conformations. As we can see, the G106V variant shows stronger absorbance and its double peaks become sharper compared with original WT. On the contrary, the G18V variant shows more blunt double peaks. The results suggest that FUV spectra are sensitive to minor structure variation. This offers a possible path to correlate the tendency of aggregation of different proteins with the specific structure factors that are responsible for the FUV signals. FUV is the foundation of CD and 2DUV spectra, which are more powerful in minoring mutations of proteins.[50]

To support real-time tracking of protein dynamics using time-resolved spectroscopy, we combined MD simulation and ML-based FUV spectra simulation to reveal the time-dependent evolution of FUV spectra of mini Trp-cage along its folding path.[51] Figure 4c is the ML predicted FUV spectra based on 100 MD conformations for five states along the folding path of Trp-cage (retrieved from our previous study).[52] Table 1 shows the averaged secondary structure contents and

Table 1. Averaged Secondary Structure Contents and Main Peaks of the Five States of the Mini Trp-Cage along Its Folding Process[a]

state	S1	S25	S50	S75	S100
coil (%)	99.1	73.9	49.7	48.9	37.6
turn (%)	0.9	25	38.5	29.9	16.6
α-helix (%)	0	0	8.2	19.7	37.1
3_{10}-helices (%)	0	0	3.7	1.6	8.7
bridge (%)	0	1.1	0	0	0
main peak (nm)	188.9	189	191.2	194	194.1

[a]Each state is based on 100 MD conformations.

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

main peaks of the corresponding five states. As we can see, the initial unfolded S1 state with a coil structure shows a double peak FUV spectrum. The S25 state is slightly folded along with the decrease of coil content, leading to the decrease of the shoulder peak of the FUV spectrum. The folding process becomes faster from S25 to S50 states and helical structures appear, accompanied by the narrowing of bandwidth. The protein becomes a cage in the S75 state with the rapid increase of α-helix, and the shoulder peak becomes even weaker and finally is merged into one narrower peak in the final folded structure (S100). It is worth noting that Trp-cage undergoes a decrease of coil structure and increase of helical content during its folding path, which results in a red shift of the dominant peak of the FUV spectra. These results show the potential of ML-based FUV simulations for monitoring the protein folding process.

4. CONCLUSIONS

In summary, we report an efficient and powerful ML protocol to predict the FUV spectra for proteins based on their structure descriptors. The ML model presented here shows good transferability and high performance for predicting protein UV signals. It can be used to interpret experimental spectra in solution, probe structural variations, and monitor protein folding. This protocol can be applied to other related electronic UV spectroscopies, such as ultraviolet resonant Raman, circular dichroism (CD), and two-dimensional UV spectroscopy. Spectral assignments, molecular interactions, and structure−property relationships will be investigated in a future study.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c03296.

Additional computational details, figures, and tables (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

Authors

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

Sheng Ye − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

Kai Zhong − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

Yaolong Zhang − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

Yuanyuan Chong − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

Luyuan Zhao − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

Huiting Zhou − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

Sibei Guo − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China*

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

Bin Jiang − *Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China;* orcid.org/0000-0003-2696-5436

Shaul Mukamel − *Departments of Chemistry and Physics & Astronomy, University of California, Irvine, California 92697, United States;* orcid.org/0000-0002-6015-3135

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c03296

Author Contributions

‖J.Z., S.Y., K.Z.: These authors contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Technology of People's Republic of China (2018YFA0208603, 2017YFA0303500, 2016YFA0400904) and the National Natural Science Foundation of China (21633006, 21633007, 21790350, 21703221). S.M. is grateful for the support of NSF (Grant: CHE-1953045), Natural Science Foundation of Guizhou Province (QHPT[2017]5790-01). The University of Science and Technology of China High Performance Computing Center is acknowledged for computing resources.

■ REFERENCES

(1) Wang, D. Using Genetics to Reveal Protein Structure. *Science* **2020**, *370*, 1269.

(2) Gabrieli, F.; Dooley, K. A.; Facini, M.; Delaney, J. K. Near-UV to Mid-IR Reflectance Imaging Spectroscopy of Paintings on the Macroscale. *Sci. Adv.* **2019**, *5*, No. eaaw7794.
(3) 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*, 3674−3682.
(4) Higashi, N. *Far- and Deep-Ultraviolet Spectroscopy*; Ozaki, Y., Kawata, S., Eds.; Springer: Tokyo, 2015.
(5) Demchenko, A. P. *Ultraviolet Spectroscopy of Proteins*; Springer Science & Business Media: New York, 2013.
(6) Prasad, S.; Mandal, I.; Singh, S.; Paul, A.; Mandal, B.; Venkatramani, R.; Swaminathan, R. Near UV-Visible Electronic Absorption Originating from Charged Amino Acids in a Monomeric Protein. *Chem. Sci.* **2017**, *8*, 5416−5433.
(7) Biter, A. B.; Pollet, J.; Chen, W.-H.; Strych, U.; Hotez, P. J.; Bottazzi, M. E. A Method to Probe Protein Structure from UV Absorbance Spectra. *Anal. Biochem.* **2019**, *587*, 113450.
(8) Rodger, A.; Dorrington, G.; Ang, D. L. Linear Dichroism as a Probe of Molecular Structure and Interactions. *Analyst* **2016**, *141*, 6490−6498.
(9) Li, Z.; Hirst, J. D. Computed Optical Spectra of SARS-CoV-2 Proteins. *Chem. Phys. Lett.* **2020**, *758*, 137935.
(10) Zhang, J.; Sharman, E.; Jiang, J. Two-Dimensional Ultraviolet Spectroscopy of Proteins. *Sci. China: Chem.* **2018**, *61*, 1099−1109.
(11) Seibert, J.; Bannwarth, C.; Grimme, S. Biomolecular Structure Information from High-Speed Quantum Mechanical Electronic Spectra Calculation. *J. Am. Chem. Soc.* **2017**, *139*, 11682−11685.
(12) 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*, 79−84.
(13) Tortorella, S.; Talamo, M. M.; Cardone, A.; Pastore, M.; De Angelis, F. Benchmarking DFT and Semi-Empirical Methods for a Reliable and Cost-Efficient Computational Screening of Benzofulvene Derivatives as Donor Materials for Small-Molecule Organic Solar Cells. *J. Phys.: Condens. Matter* **2016**, *28*, 074005.
(14) Ghosh, A.; Ostrander, J. S.; Zanni, M. T. Watching Proteins Wiggle: Mapping Structures with Two-Dimensional Infrared Spectroscopy. *Chem. Rev.* **2017**, *117*, 10726−10759.
(15) Kananenka, A. A.; Yao, K.; Corcelli, S. A.; Skinner, J. L. Machine Learning for Vibrational Spectroscopic Maps. *J. Chem. Theory Comput.* **2019**, *15*, 6850−6858.
(16) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.
(17) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336−2347.
(18) Lansford, J. L.; Vlachos, D. G. Infrared Spectroscopy Data- and Physics-Driven Machine Learning for Characterizing Surface Microstructure of Complex Materials. *Nat. Commun.* **2020**, *11*, 1513.
(19) 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*, 11612−11617.
(20) Legrain, F.; Carrete, J.; van Roekeghem, A.; Madsen, G. K. H.; Mingo, N. Materials Screening for the Discovery of New Half-Heuslers: Machine Learning Versus Ab Initio Methods. *J. Phys. Chem. B* **2018**, *122*, 625−632.
(21) Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2020**, DOI: 10.1021/acs.chemrev.0c00749.
(22) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Adv. Sci.* **2019**, *6*, 1801367.
(23) Unke, O. T.; Meuwly, M. Physnet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.
(24) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C. L.; Žídek, A.; Nelson, A. W. R.; Alex, B.; et al. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706−710.
(25) Frenkel, Y. On the Transformation of Light into Heat in Solids. *Phys. Rev.* **1931**, *37*, 17−44.
(26) Kasha, M.; Rawls, H.; El-Bayoumi, M. A. The Exciton Model in Molecular Spectroscopy. *Pure Appl. Chem.* **1965**, *11*, 371−392.
(27) Zhang, Y.; Luo, Y.; Zhang, Y.; Yu, Y. J.; Kuang, Y. M.; Zhang, L.; Meng, Q. S.; Luo, Y.; Yang, Y. J.; Dong, Z. C.; Hou, J. G. Visualizing Coherent Intermolecular Dipole-Dipole Coupling in Real Space. *Nature* **2016**, *531*, 623−627.
(28) 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*, 235−242.
(29) 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*, 8100−8107.
(30) De Silva, N.; Willow, S. Y.; Gordon, M. S. Solvent Induced Shifts in the UV Spectrum of Amides. *J. Phys. Chem. A* **2013**, *117*, 11847−11855.
(31) Brkljača, Z.; Mališ, M.; Smith, D. M.; Smith, A.-S. Calculating CD Spectra of Flexible Peptides: An Assessment of TD-DFT Functionals. *J. Chem. Theory Comput.* **2014**, *10*, 3270−3279.
(32) Nielsen, E. B.; Schellman, J. A. The Absorption Spectra of Simple Amides and Peptides. *J. Phys. Chem.* **1967**, *71*, 2297−2304.
(33) Sok, S.; Willow, S. Y.; Zahariev, F.; Gordon, M. S. Solvent-Induced Shift of the Lowest Singlet π-π^* Charge-Transfer Excited State of P-Nitroaniline in Water: An Application of the TDDFT/EFP1Method. *J. Phys. Chem. A* **2011**, *115*, 9801−9809.
(34) Leang, S. S.; Zahariev, F.; Gordona, M. S. Benchmarking the Performance of Time-dependent Density Functional Methods. *J. Chem. Phys.* **2012**, *136*, 104101.
(35) Jacquemin, D.; Mennucci, B.; Adamo, C. Excited-State Calculations with TD-DFT: From Benchmarks to Simulations in Complex Environments. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16987−98.
(36) Sarkar, R.; Boggio-Pasqua, M.; Loos, P.-F.; Jacquemin, D. Benchmarking TD-DFT and Wave Function Methods for Oscillator Strengths and Excited-State Dipole Moments. *J. Chem. Theory Comput.* **2021**, *17*, 1117−1132.
(37) Pereira, F.; Aires-De-Sousa, J. Machine Learning for the Prediction of Molecular Dipole Moments Obtained by Density Functional Theory. *J. Cheminf.* **2018**, *10*, 43.
(38) 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*, revision A.03; Gaussian, Inc.: Wallingford, CT, 2016.
(39) 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*, 4962−4967.
(40) Abramavicius, D.; Palmieri, B.; Voronine, D. V.; Šanda, F.; Mukamel, S. Coherent Multidimensional Optical Spectroscopy of Excitons in Molecular Aggregates; Quasiparticle Versus Supermolecule Perspectives. *Chem. Rev.* **2009**, *109*, 2350−2408.
(41) Rupp, M.; Tkatchenko, A.; Muller, K.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(42) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−31.
(43) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

(44) Besley, N. A.; Hirst, J. D. Theoretical Studies toward Quantitative Protein Circular Dichroism Calculations. *J. Am. Chem. Soc.* **1999**, *121*, 9636−9644.
(45) Ke, W.; Yu, D.; Feng, Y. The Analysis of UV Absorption Spectra and Raman Spectra in BSA Aqueous Solutions. *Spectrosc. Spec. Anal.* **1993**, *13*, 55−58.
(46) Bulheller, B. M.; Rodger, A.; Hicks, M. R.; Dafforn, T. R.; Serpell, L. C.; Marshall, K. E.; Bromley, E. H.; King, P. J.; Channon, K. J.; Woolfson, D. N.; et al. Flow Linear Dichroism of Some Prototypical Proteins. *J. Am. Chem. Soc.* **2009**, *131*, 13305−13314.
(47) Besley, N. A.; Hirst, J. D. Ab Initio Study of the Effect of Solvation on the Electronic Spectra of Formamide and N-Methylacetamide. *J. Phys. Chem. A* **1998**, *102*, 10791−10797.
(48) Nelson, T. J.; Zhao, J.; Stains, C. I. Chapter Three - Utilizing Split-Nanoluc Luciferase Fragments as Luminescent Probes for Protein Solubility in Living Cells. In *Methods Enzymol.*; Shukla, A. K., Ed.; Academic Press: 2019; Vol. *622*, pp 55−66.
(49) Jiang, J.; Golchert, K. J.; Kingsley, C. N.; Brubaker, W. D.; Martin, R. W.; Mukamel, S. Exploring the Aggregation Propensity of Gammas-Crystallin Protein Variants Using Two-Dimensional Spectroscopic Tools. *J. Phys. Chem. B* **2013**, *117*, 14294−14301.
(50) Brubaker, W. D.; Freites, J. A.; Golchert, K. J.; Shapiro, R. A.; Morikis, V.; Tobias, D. J.; Martin, R. W. Separating Instability from Aggregation Propensity in γS-Crystallin Variants. *Biophys. J.* **2011**, *100*, 498−506.
(51) Dobson, C. M. Protein Folding and Misfolding. *Nature* **2003**, *426*, 884−890.
(52) 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*, 1341−1346.

https://doi.org/10.1021/acs.jpcb.1c03296
J. Phys. Chem. B 2021, 125, 6171−6178

Supporting Information

A Machine-Learning Protocol for Ultraviolet Protein-backbone Absorption Spectroscopy under Environmental Fluctuations

Jinxiao Zhang[1‡], Sheng Ye[2‡], Kai Zhong[2‡], Yaolong Zhang[2], Yuanyuan Chong[2], Luyuan Zhao[2], Huiting Zhou[2], Sibei Guo[2], Guozhen Zhang[2], Bin Jiang[2], Shaul Mukamel[3], Jun Jiang[2*]

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

[2]Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China

[3]Departments of Chemistry and Physics & Astronomy, University of California, Irvine, California 92697, USA

[‡]These authors contributed equally.

*Correspondence author: Jun Jiang (Email:jiangj1@ustc.edu.cn)

Computational details

1. Theory of the Calculation of FUV Spectra

The far-ultraviolet (FUV) absorption spectra of a protein is mostly derived from it electronic excitations of the peptide backbone coupled with environmental fluctuations, which can be described by the Frenkel exciton model.[1-2]

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

The indices *m (n)* run over the peptide bonds and *a (b)* represents n → π* or π → π* transitions. $\hat{B}_{ma}^{\dagger}$ and $\hat{B}_{ma}$ are creation and annihilation operators of excitations of the *m th* peptide bond, respectively. The excitation energy ε_{ma} can be described by summing over the excitation energy of an isolated peptide ($\varepsilon_{0,ma}$) and environmental electrostatic interactions:

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

$\rho_{T,ma}$ and $\rho_{G,m}$ are the molecular charge density of the *a th* excited state and ground state of the peptide bond *m*, respectively. $\rho_{G,m}$ is the charge density of a residue *l* (*l* runs over all residues). The resonant coupling (*J*) between *a th* excited state of peptide *m* and *b th* excited state of peptides *n* can be written as:

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

Applying the dipole approximation that computes electrostatic interaction between two subjects with the product of their electric dipole moments,[3-4] we can calculate the excitation energy ε_{ma} by summing over the excitation energy of an isolated peptide ($\varepsilon_{0,ma}$) and its electrostatic interactions with surrounding environment.

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

$\boldsymbol{\mu}_{T,ma}$ and $\boldsymbol{\mu}_{G,l}$ are the electronic transition dipole moments of the peptide bond (*m, a*) and the ground state dipole moment of a residue *l* (*l* runs over all residues), respectively. Using the dipole approximation, the resonant coupling (*J*) between the excited states *a*, *b* of peptides *m*, *n* can be computed as:

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

The *J* couplings come from transitions of peptide bonds ($m \neq n$). The residues only contribute in Eq. (4) where they modify the excitation energies.

2. Data Preparation and Quantum Chemistry Simulations

In order to ensure the diversity of the dataset, 1000 PDB files cover almost all the typical types of proteins, including fibrous protein, globular protein, keratin, collagen, chaperone, myoglobin, hemoglobin and denaturation, are retrieved from RCSB Protein Data Bank.

The PDB ID can be found in Table S2. Afterwards, 50 000 peptides (NMA molecule is chosen as peptide model) and 200 000 amino acid residues are directly extracted in bulk from the downloaded PDB files with self-compiled codes. The numbers of peptides/residues randomly extracted from each proteins are roughly the same in order to ensure the diversity of the dataset. When extracting peptides, we also include the two connected C atoms, that is –C–CO–NH–C– rather than –CO–NH– since NMA molecule (–C – CO–NH–C–) is chosen as peptide model. The dangling bonds in extracted NMA molecules and residues are linked with hydrogen in bulk with Pymol package.[5] Therefore, the coordinates of both NMA molecules and residues are exactly the same as they are in original proteins configurations except that positions of hydrogen atoms are uncertain in PDB files. The structures of NMA molecules and residues can be found in Figure S4 and Figure S6, respectively.

We didn't carry out energy minimization for NMA molecules and residues for the following two reasons: (1) the coordinates of NMA molecules and residues are directly extracted from PDB files without any change, which are more consistent with the configurations of peptide bonds and amino acid residues in proteins. (2) The NMA molecules and residues directly extracted from proteins are unstable since they are dragged by each other when they are in proteins. The configurations of both NMA molecules and residues tend to be unified and the diversity will be reduced significantly after energy minimization, which would be adverse to machine learning training for structure-property relationships.

Time-dependent density functional theory (TDDFT) calculations at PBE0/cc-pVDZ level are employed to acquire the excitation energy (ε_0) and transition dipole moments ($\boldsymbol{\mu}_T$) of peptides. All the peptides are converted to the same coordinate before TDDFT calculations as shown in Figure S1 and the NOSYMM keyword is required to prevent structural reorientation during TDDFT calculations. The lowest 10 excitation states are calculated and phase correlation [6] is performed with Multiwfn code.[7] Density functional theory (DFT) calculations with B3LYP/6-311++G** method are performed to calculate the ground state dipole moments ($\boldsymbol{\mu}_G$) of amino acid residues. Polarizable continuum models (PCM) is used as solvent model and water is used as solvent for all the calculations. All the DFT/TDDFT simulations are carried out in Gaussian 16 package.[8]

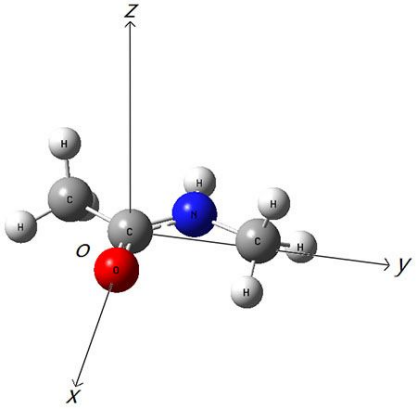


Figure S1. The peptide orientation after conversion of structures to the same Cartesian coordinate system.

3. Molecular dynamic simulations

Molecular dynamic (MD) simulations for seven proteins displayed in Figure 4a (1YJP, 1CLG, 2D3E, 3V03) and Figure 4b (WT, G106V, G18V) are carried out to acquire MD conformations. The MD conformations of Figure 4c are retrieved from our previous reported work.[9] The MD simulations in water of 2 ns with a time step of 2 fs are performed using OPLS-AA force field and TIP3P water at room temperature (300K) and pressure (1 atm) after NVT equilibration at 300 K in GROMACS package.[10] Periodic boundary conditions are employed during MD simulations. The short-range coulomb interactions and vdW forces are truncated at 1.2 nm. Particle-mesh Ewald is used to take long-range electrostatics into consideration.

4. Selection of molecular descriptors

Rational selection of molecular descriptor is crucial for us to create the structures-property relationship.[11] We carefully select different molecular descriptors for different properties.

Excitation energy of peptide: We compared four molecular descriptors, including internal coordinates, coulomb matrix (CM),[12] bad of bands (BOB),[13] atom-centered symmetry functions (ACSF),[14] in which the internal coordinates show the best result. Internal coordinates include the bond lengths, bond angles and dihedral angles of a molecule, which hold advantages in directly reflecting the fundamental structure-property relationship.

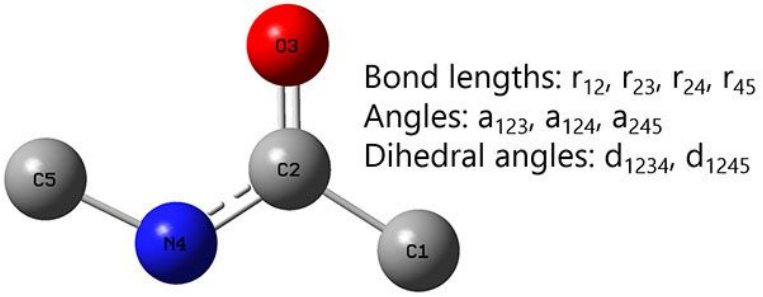


Figure S2. Internal coordinates of a peptide.

Coulomb matrix (CM) M is used to describe the local environment of a central atom k and can be written as follows:

$$M_{ij}(k) = \begin{cases} \frac{1}{2} Z_i^{2.4} \cdot f_{ik}^2 & i = j \\ \frac{Z_i Z_j}{||R_i - R_j||} f_{ik} f_{jk} f_{ij} & i \neq j \end{cases} \quad (6)$$

In which i, j and k indicate atom labels, Z is nuclear charge and R is coordinate, f_{ij} is a function used to describe long range effect:

$$f_{ij} = \begin{cases} 1 & ||R_i - R_j|| \leq r - \Delta r \\ \frac{1}{2}\left(1 + cos\left(\pi \frac{||R_i - R_j|| - r + \Delta r}{\Delta r}\right)\right) & r - \Delta r < ||R_i - R_j|| \leq r - \Delta r \\ 0 & ||R_i - R_j|| > r \end{cases} \quad (7)$$

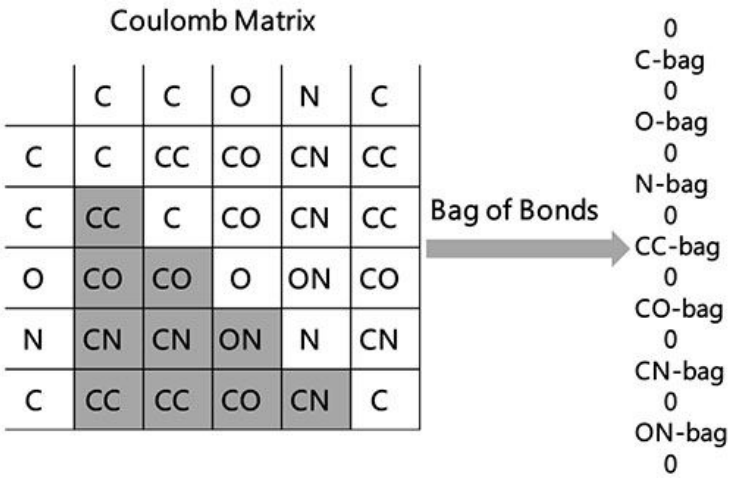


Figure S3. Bad of bags of a peptide.

Bad of bags (BOB) is an expansion of CM molecular descriptor which groups CM elements into bags based on unique atom pairs and sorts them by values. Each bag represents a particular bond type (e.g. 'C-C', 'C-O', 'C-N', etc) in BOB. The self-interactions part (e.g. 'C', 'O', 'N', etc) is constructed by diagonal CM elements:

$$\frac{1}{2}Z_i^{2.4} \quad (8)$$

The interaction between different atoms is created with the off-diagonal CM elements:

$$\frac{Z_i Z_j}{||R_i - R_j||} \quad (9)$$

where *Zi* and *Zj* are the nuclear charges, while *Ri* and *Rj* are the positions of the two atoms participating in a given bond.

Atom-centered symmetry functions (ACSF) employ a series of radial and angular symmetry functions to represent the local environment near a central atom to detect the structural features. The radial symmetry functions of a central atom *i* are given as:

$$f_c(R_{ij}) = \begin{cases} 0.5 \cdot \left[cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right] & R_{ij} \leq R_c \\ 0 & R_{ij} > R_c \end{cases} \quad (10)$$

$$G_i^1 = \sum_j f_c(R_{ij}) \quad (11)$$

$$G_i^2 = \sum_j e^{-\eta(R_{ij} - R_s)^2} \cdot f_c(R_{ij}) \quad (12)$$

In which $f_c(R_{ij})$ is the cutoff function, R_{ij} is the distance between atoms i and j, R_c is the cutoff radius. G_i^1 radial symmetry function is the sum of cutoff functions. G_i^2 radial symmetry function is the product of a Gaussians and cutoff function, which can be used to describe a spherical shell of a central atom. η and R_s can are employed to describe the width and shift of Gaussians. The G_i^4 angular function can be written as:

$$G_i^4 = 2^{1-\xi} \sum_{j,k \neq i}^{all} (1 + \lambda cos\theta_{ijk})^{\xi} \quad \cdot e^{-\eta\left(R_{ij}^2 + R_{ik}^2 + R_{jk}^2\right)^2} \cdot f_c(R_{ij}) \cdot f_c(R_{ik}) \cdot f_c(R_{jk}) \quad (13)$$

ξ is used to modify the distribution of angles centered of a reference atom. All the parameters of the equation describe above are directly derived from the best performing parameters reported by Marquetand et al.[15]

For transition dipole moment of peptide, we employ an embedded atom neural network (EANN) approach reported in our previous works, which introduces a Gaussian-type orbital based density vector into empirical embedded atom method to describe the complex relationship between the embedded density vector and atomic energy by neural networks. The successful construction of relationship between embedded density and atomic energy of EANN inspires us to extend its application to the prediction of dipole moment. The ground state dipole moment of $\boldsymbol{\mu}_G$ of can be described as the sum of atomics contributions:

$$\boldsymbol{\mu}_G = \sum_{i=1}^{N} q_i \boldsymbol{r}_i \quad (14)$$

Where N is the atom number, q_i is the atomic effective charge and can be easily fit with NN method. The coordinate vector $\boldsymbol{r}_i = (x_i, y_i, z_i)^T$ is oriented from the center of mass of a molecule to atom i. $\boldsymbol{\mu}_G$ can be easily acquired by multiplying q_i and $\boldsymbol{r}_i$. However, the transition dipole moment ($\boldsymbol{\mu}_T$) can be perpendicular to the molecular plane since it is associated with the transition between two states, which it different from that of $\boldsymbol{\mu}_G$ which is corresponding to only one state. For example, if a molecular plane is on xy plane and $\boldsymbol{\mu}_T$ is perpendicular to xy plane (on z plane) and only μ_z is be nonzero. While Eq. (14) will give a zero μ_z since r_z is zero, which is apparently unreasonable. Here we define the transition dipole moment of peptide as follows to tackle this problem:

$$\boldsymbol{\mu}_T^j = \sum_{i=1}^{N} q_i^j \boldsymbol{r}_i \quad (15)$$

Where j=1, 2 and q_i^j is the different output of atomic NN. Two vectors $\boldsymbol{\mu}_T^1$ and $\boldsymbol{\mu}_T^2$ can define a plane (molecular plane), and the plane which perpendicular to the molecular plane can be defined as:

$$\boldsymbol{\mu}_T^3 = \sum_{i=1}^{N} q_i^3 (\boldsymbol{\mu}_T^1 \times \boldsymbol{\mu}_T^2) \quad (16)$$

Finally, the transition dipole moment $\boldsymbol{\mu}_T$ can be written as the linear combination of the three transition dipole moment vectors:

$$\boldsymbol{\mu}_T = \boldsymbol{\mu}_T^1 + \boldsymbol{\mu}_T^1 + \boldsymbol{\mu}_T^3 \quad (17)$$

As a result, $\boldsymbol{\mu}_T$ can still be rationally described even when it is perpendicular to the molecular plane.

For the ground state dipole moment of the residue which is employed for the calculation of the perturbation term of excitation energy, we employ converted Cartesian coordinates which can directly reflect the structural feather and orientation of a molecule.

5. Machine learning (neural network)

Neural network (NN) implemented in TensorFlow is employed for all the training procedure. We manually optimize the hyperparameters of NN, including hidden layers, neurons of each hidden layers, activation functions, algorithms to against overfitting and the corresponding regularization coefficient, and learning rate, to create a favorable ML protocol. The hyperparameters of NN are determined after carefully optimization. NN with three hidden layers are used in all the NN training process. For excitation energy of peptide and the ground state dipole moment of residues, the neurons of three hidden layers are 32, 64 and 128, respectively. The Rectified Linear Unit is employed as activation function for each NN layer to speed up the NN training and resist the gradient disappearing. L2 regularization with a coefficient of 0.01 is used to mitigate overfitting. Adam algorithm with an exponentially decaying learning rate, which employs an initial learning rate of 0.001 and lets the learning rate decreased by 80% every 500 steps during the NN training, is used to avoid the local minima during the NN training. For transition dipole moments of peptides, we used an Embedded atom neural network (EANN) model which has been reported in our previous work.[16] In this model, a neural network with 3 hidden layers (33, 30, 30), early stopping in which the training will stop if the validation loss shows a consecutive increase in 6 epoch to prevent overfitting, and Levenberg-Marquardt algorithm are employed for ML protocol.

For the validation procedure, we randomly select 80% of the peptide/residues data extracted from 1000 proteins for machine learning training and remaining 20% for validation. Our NN training results reveal that the NN model show favorable accuracy and transferability (Figure 2 & Figure S6). All the data are normalized with the following equation before NN training to avoid remarkably different range of raw input values:

$$\boldsymbol{\mu}_T = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \quad (18)$$

The mean relative error (*MRE*) is defined as follows:

$$\boldsymbol{\mu}_T = \frac{100\%}{n}\sum_{i=1}^{N}\left|\frac{(R_i - P_i)}{R_i}\right| \quad (19)$$

Where R_i and P_i are the reference values and predicted values of molecule *i*, respectively. N is the molecular number.

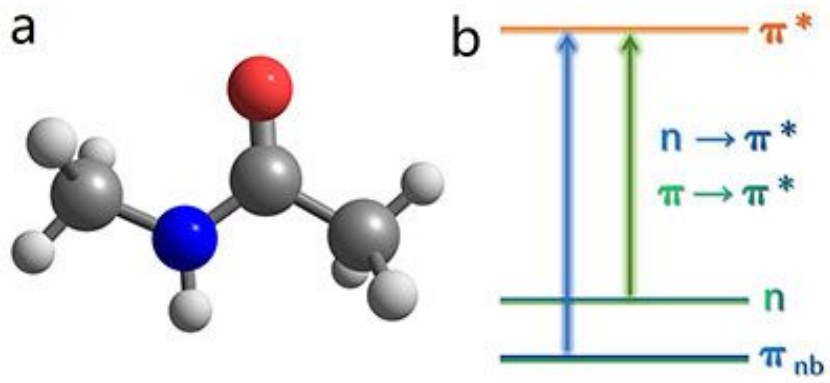


Figure S4. Peptide model and the corresponding electronic transitions. (a) The molecular structures of peptide model (N-methylacetamide, NMA). (b) n → π* transition mainly distributes in 220 nm and π → π* transition mainly locates in 190 nm in peptide.

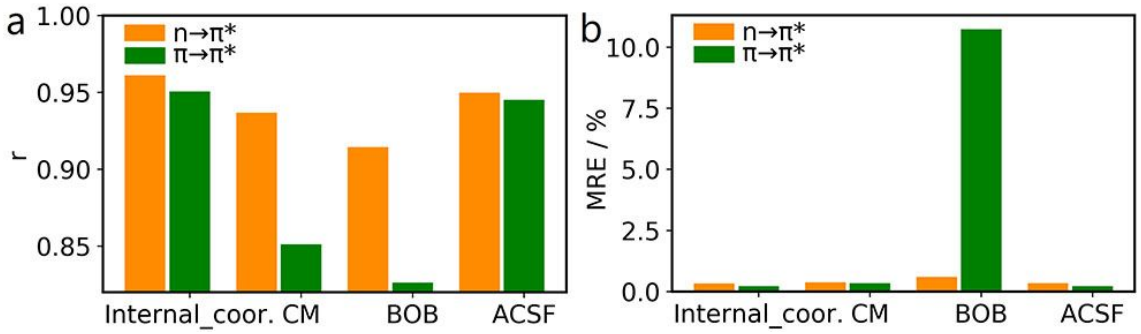
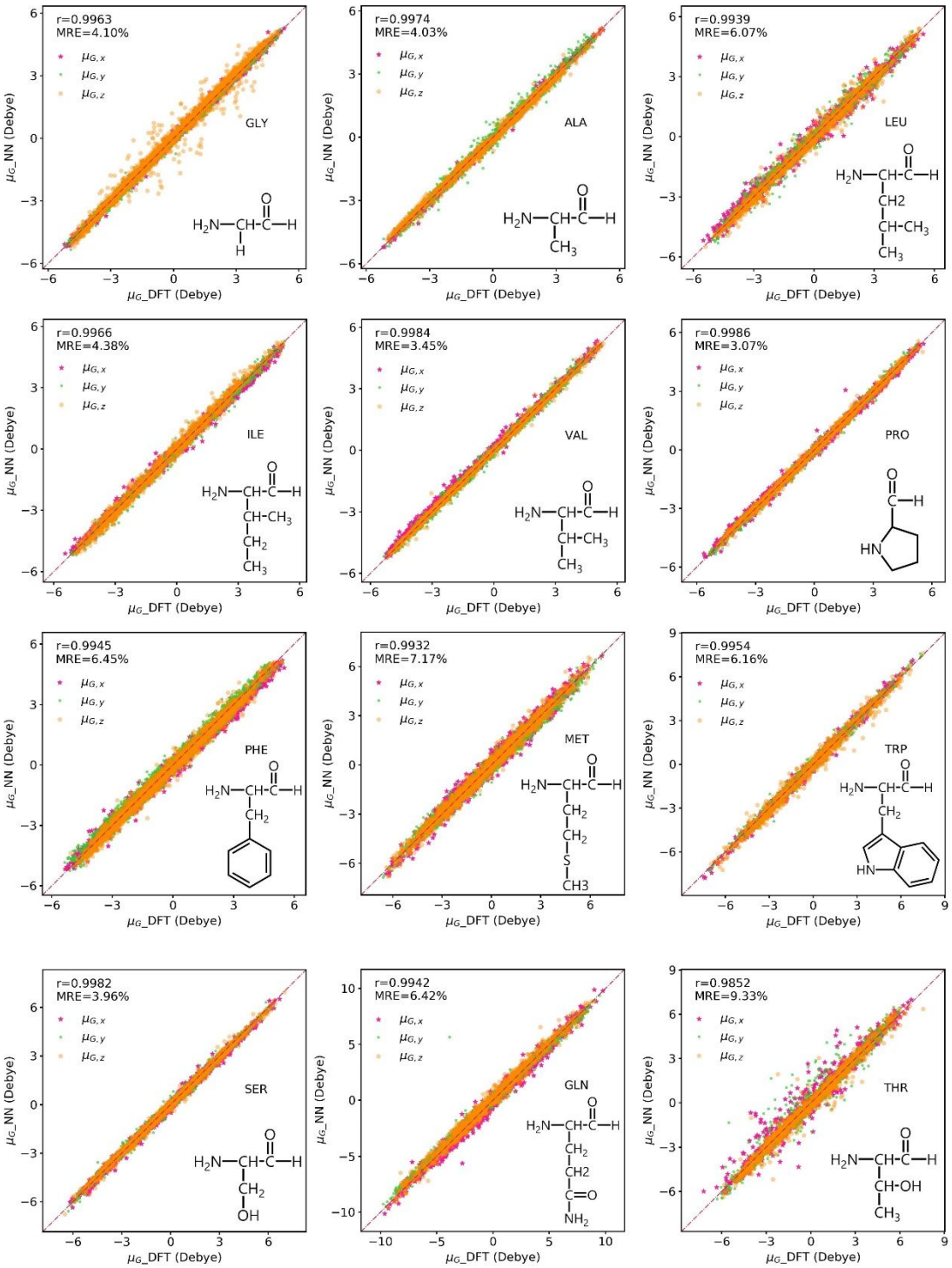


Figure S5. Machine learning results of the excitation energies of peptide based on internal coordinates, coulomb matrix (CM), bad of bonds (BOB) and atom-centered symmetry functions (ACSF). Comparison of (a) the Pearson correlation coefficient (*r*) and (b) mean relative error (*MRE*) of the n → π* and π → π* transitions based on the four molecular descriptors described above. Internal coordinates exhibit the largest *r* and the smallest *MRE*.

Figure S6. ML prediction of ground state dipole moments ($\boldsymbol{\mu}_G$) of twenty amino acid residues. $\boldsymbol{\mu}_G$_DFT was performed at the B3LYP/6-311G++(d,p) level. The purple star, green dot and orange pentagon represent the μ_NN in the x, y and z directions, respectively.



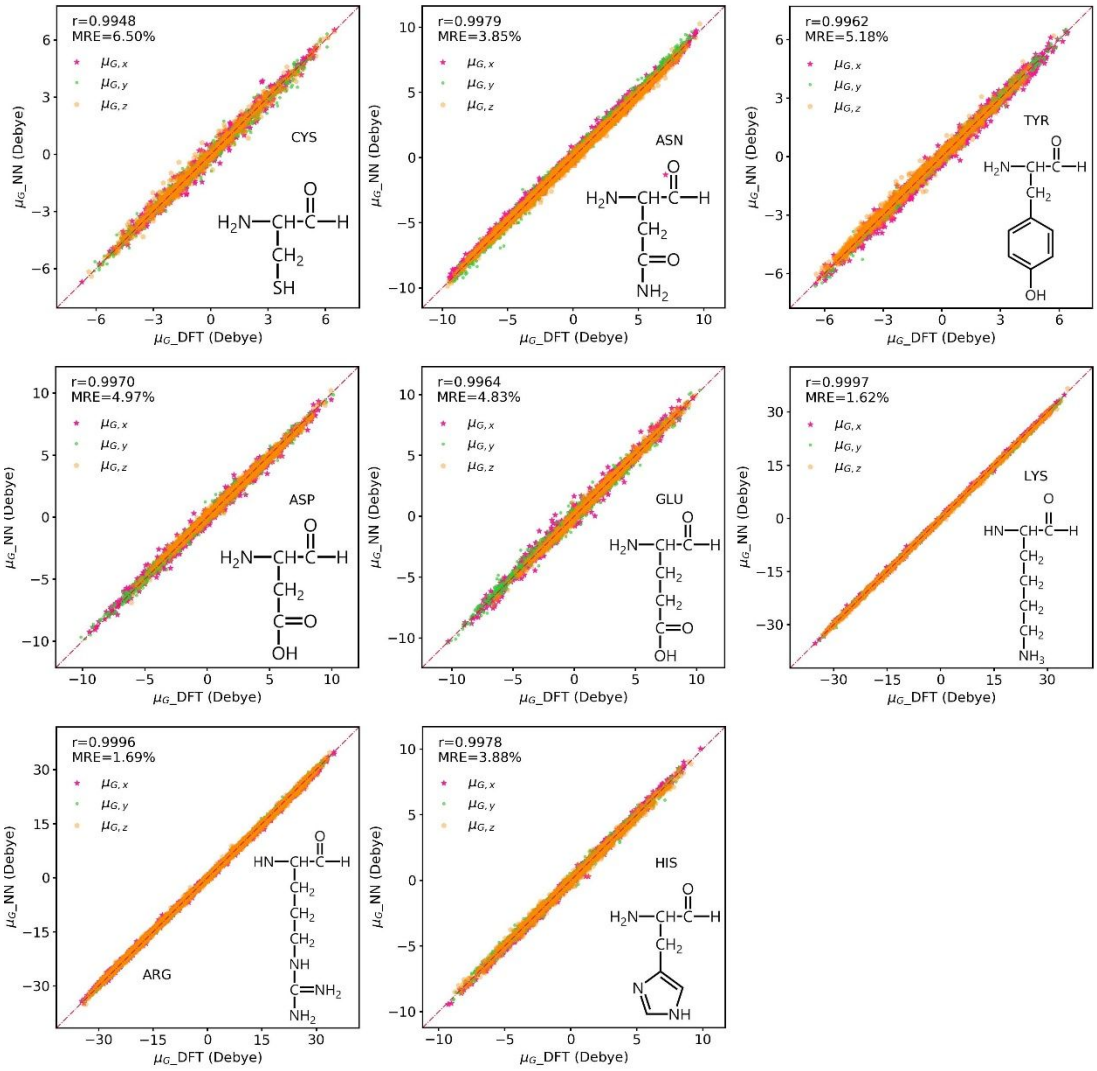

r=0.9948
MRE=6.50%
CYS
r=0.9979
MRE=3.85%
ASN
r=0.9962
MRE=5.18%
TYR
r=0.9970
MRE=4.97%
ASP
r=0.9964
MRE=4.83%
GLU
r=0.9997
MRE=1.62%
LYS
r=0.9996
MRE=1.69%
ARG
r=0.9978
MRE=3.88%
HIS
$\mu_{G,x}$
$\mu_{G,y}$
$\mu_{G,z}$
μ_G_NN (Debye)
μ_G_DFT (Debye)

Table S1. Comparison of simulation time between DFT and ML methods for proteins in Figure 3 and Figure 4a.

PDB ID	DFT / s	ML / s	DFT / ML	ρ
2BMM, MET34-GLU156	15490	6	2582	0.922
4X0J, GLU37-LYS238 of Chain A	22689	8	2836	0.976
3FHH, THR1-TRP202	28135	8	3517	0.888
5W26, ASN200-ASP401	22874	8	2859	0.918
3Q6N, TRP297-THR498 of Chain A	24094	7	3442	0.906
5E84, SER24-LEU225 of Chain A	26781	8	3348	0.874
6S84, MET1-LYS202 of Chain C	21765	8	2721	0.896
5V28, MET1-ALA174	21460	7	3066	0.933
5H34, GLU666-ARG776	26179	8	3272	0.858
6P28, ARG4-ASN195	23697	8	2962	0.931
1WXR, GLY1-GLY202	22982	8	2873	0.895
5Y30, LYS41-ILE222	27789	8	3474	0.874
1YJP	810	5	162	0.999
1CLG	11284	5	2257	0.806
2D3E	70664	10	7066	0.992
3V03	152880	12	12740	0.994

Figure S7. Comparison of timing plot between DFT and ML methods for the four proteins in Figure 4a and Table S1. As we can see, the simulation time of a DFT calculation increases rapidly with the increases of residues number, while it remains almost the same for ML methods

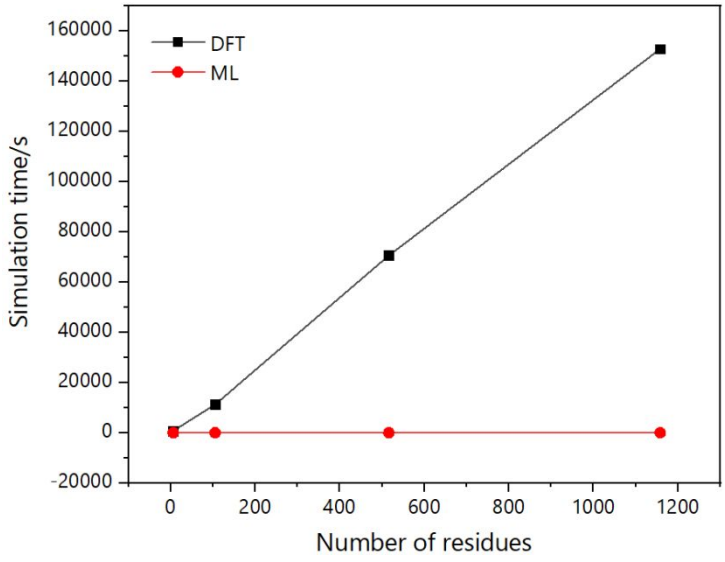
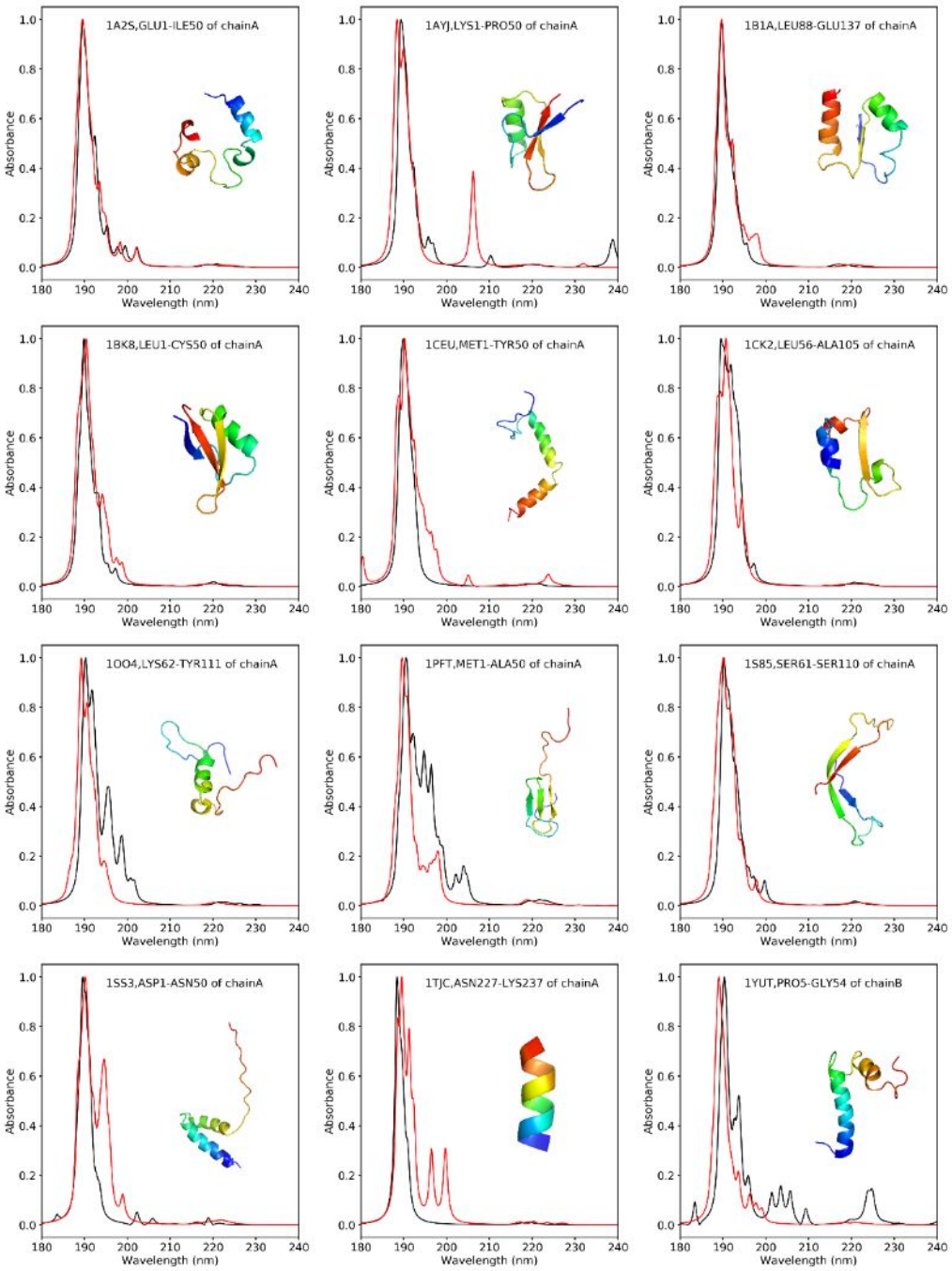
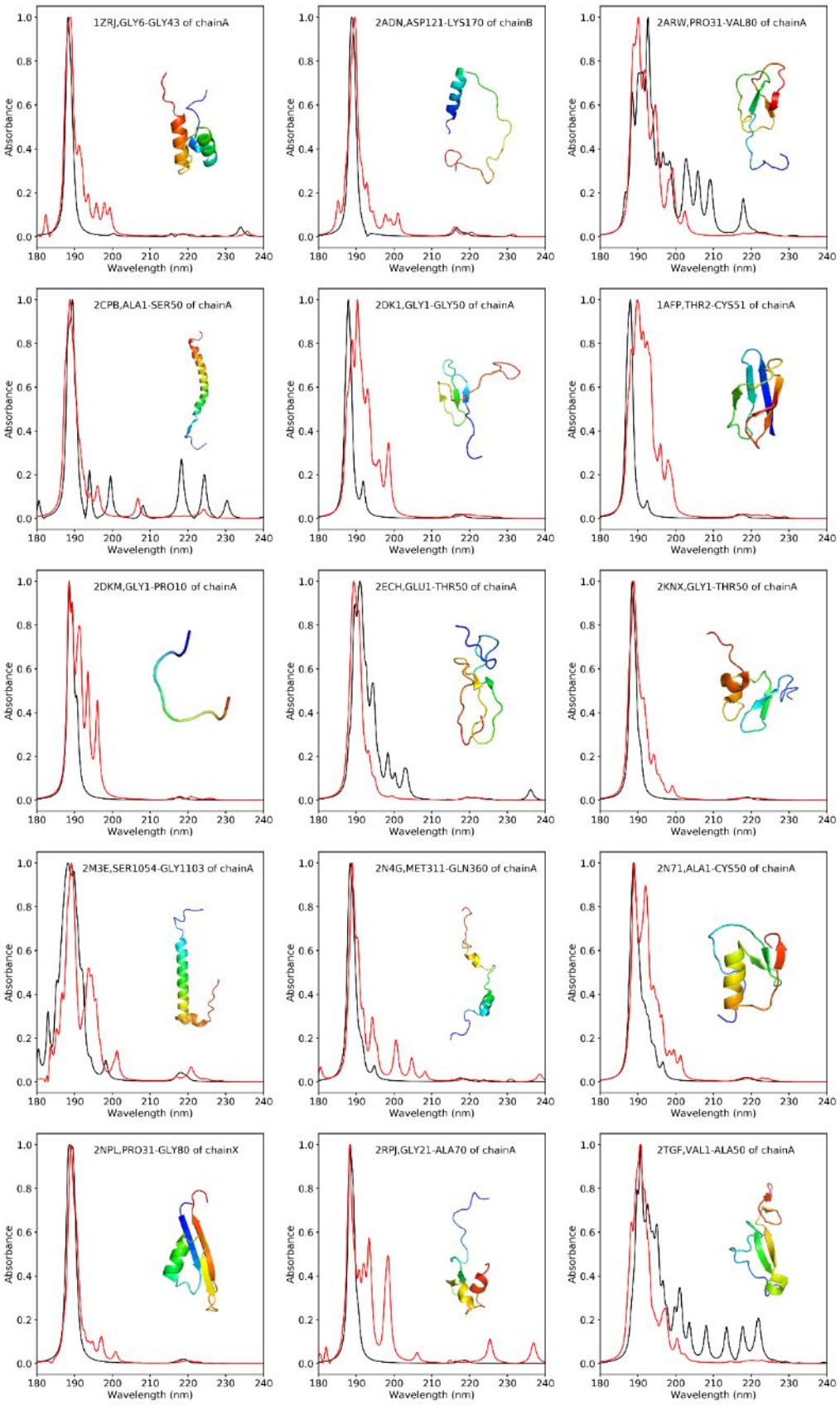
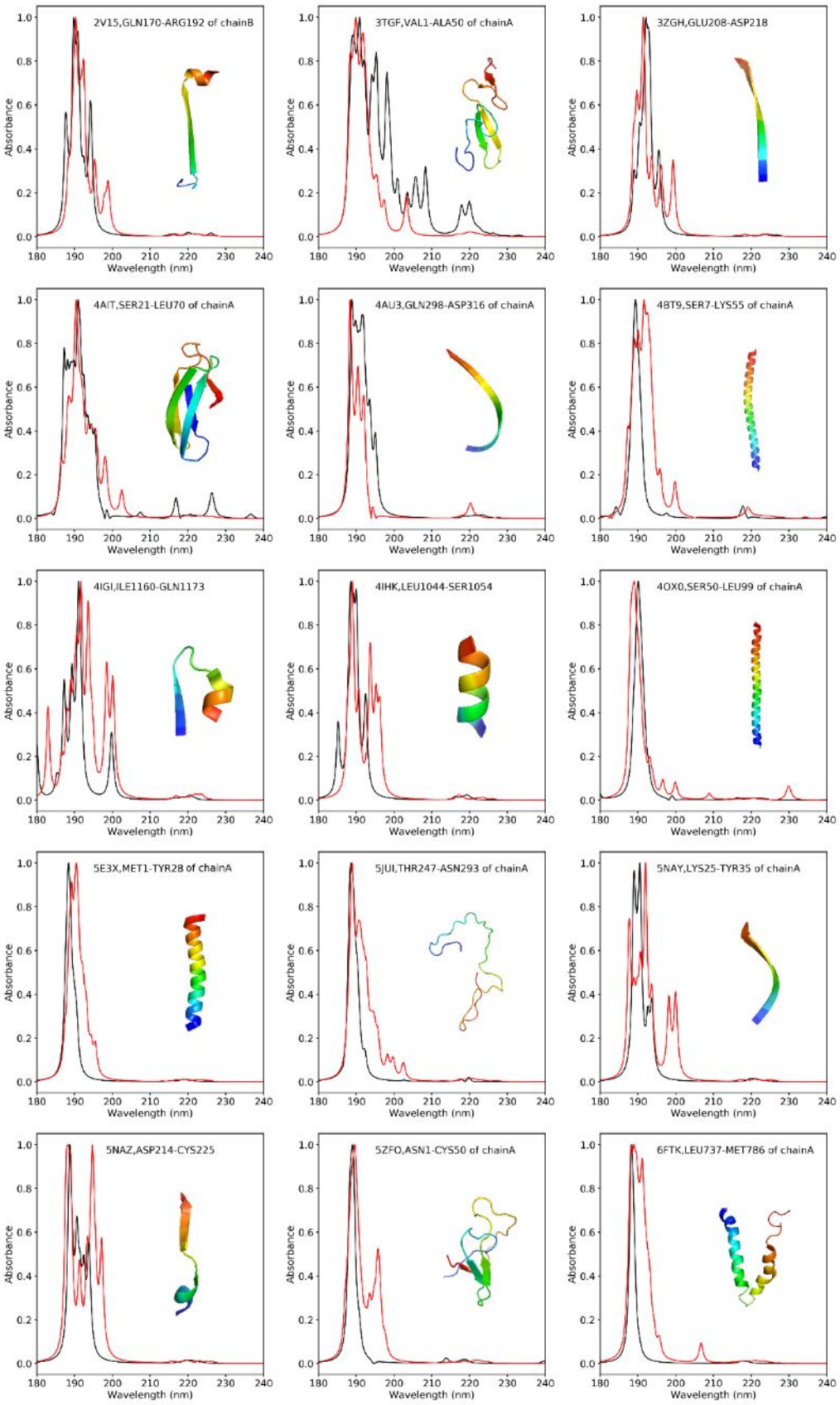
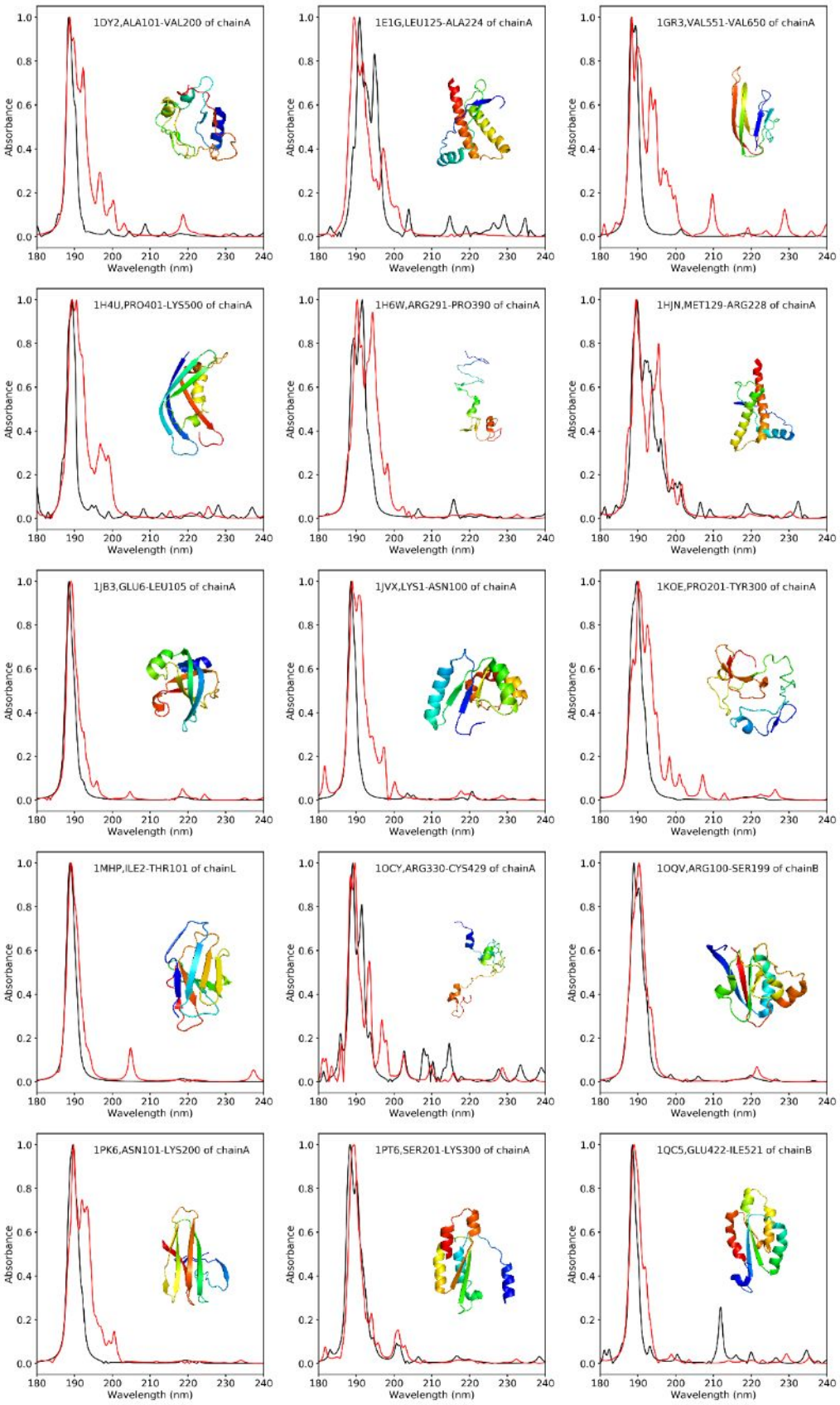


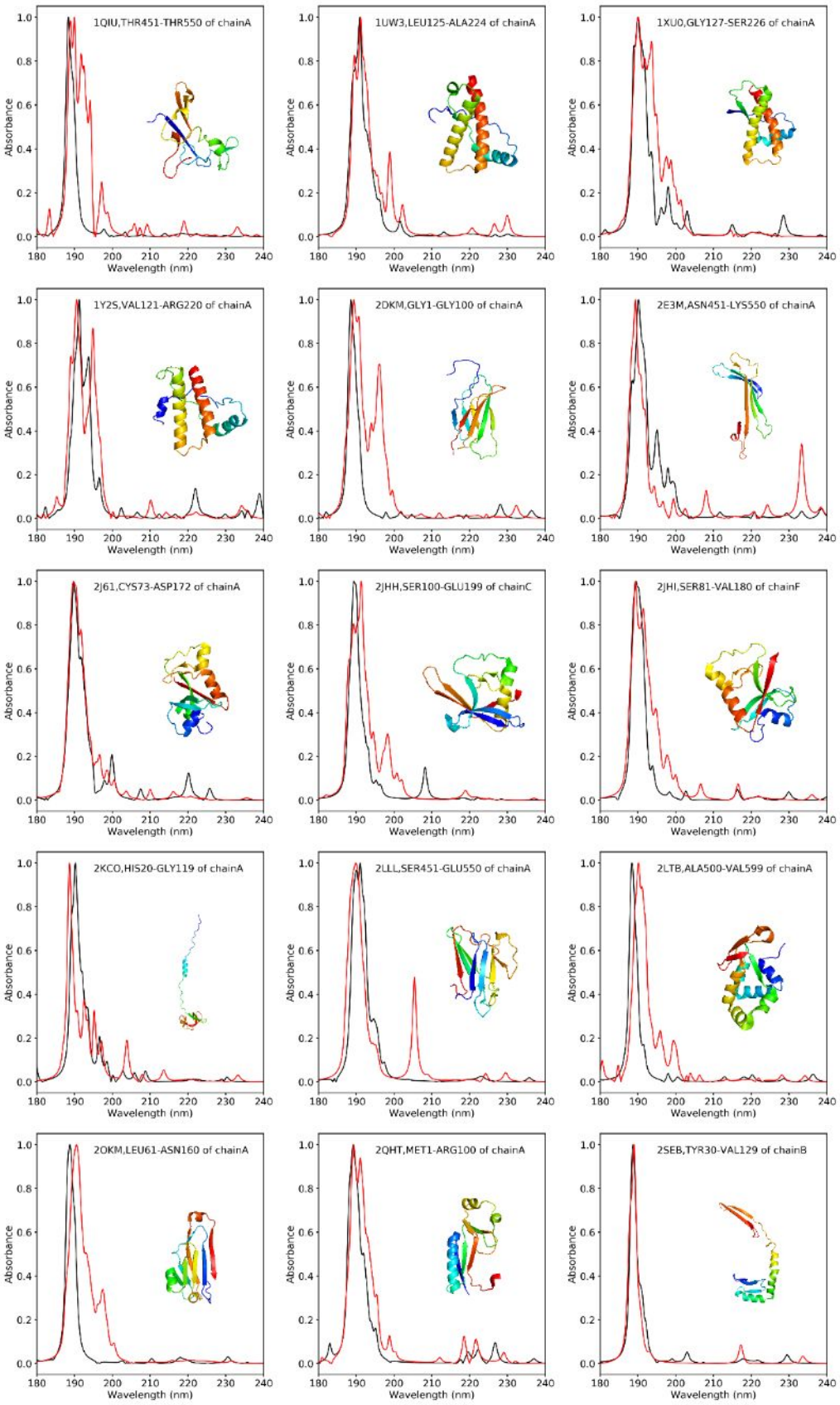
Figure S8. Far-ultraviolet spectra of 230 proteins calculated with DFT (black curves) and NN (red curves)**.**

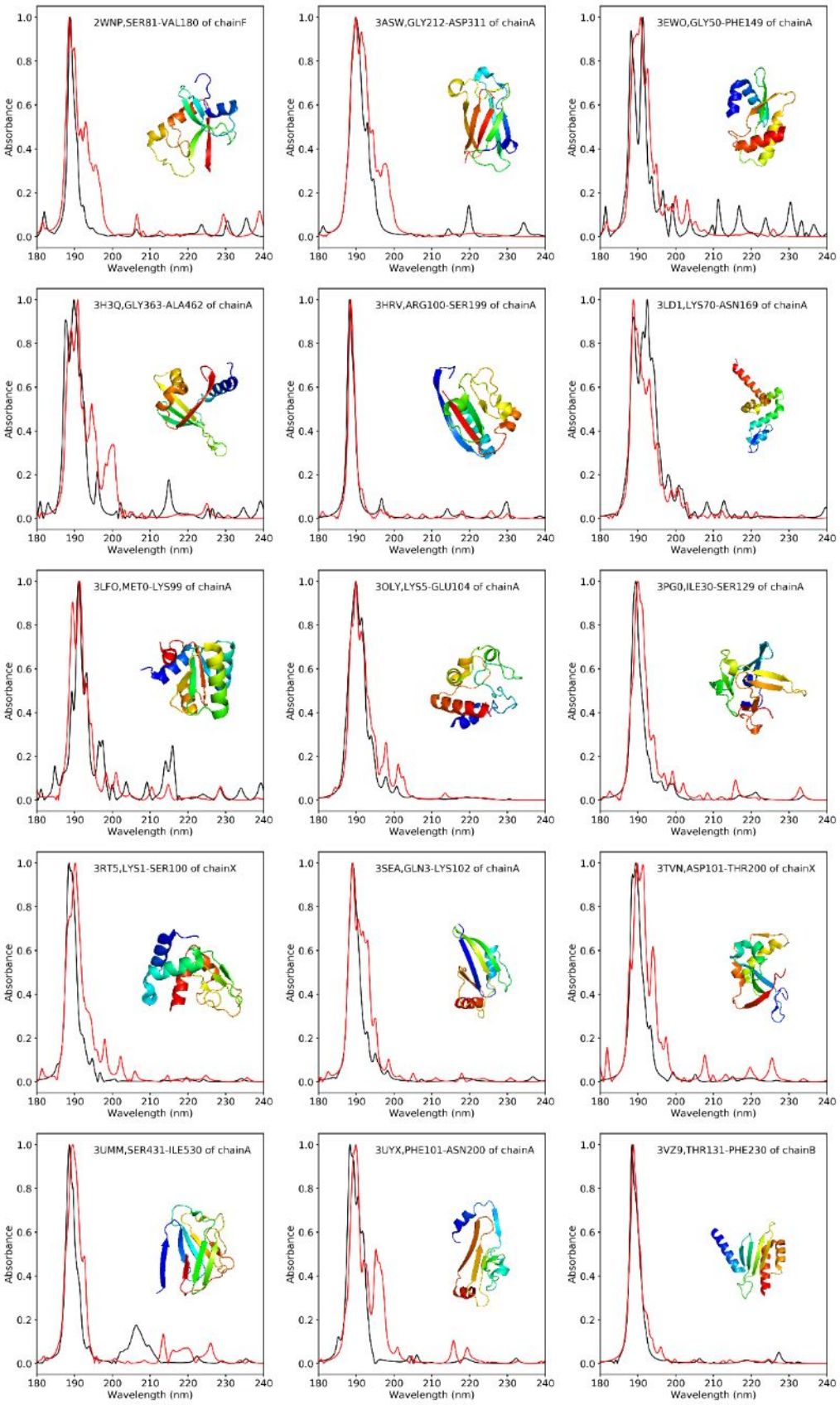


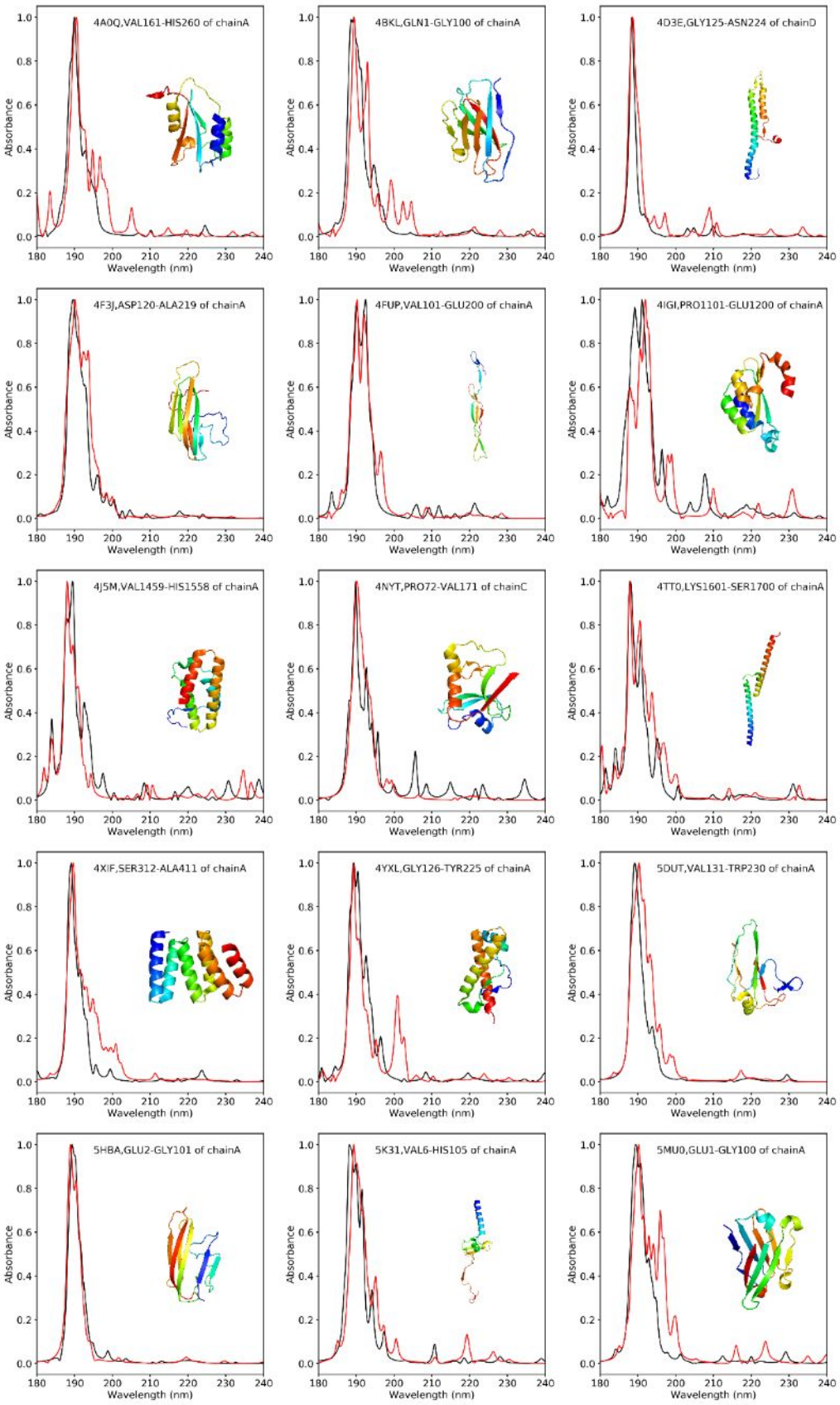

1ZRJ,GLY6-GLY43 of chainA
2ADN,ASP121-LYS170 of chainB
2ARW,PRO31-VAL80 of chainA
2CPB,ALA1-SER50 of chainA
2DK1,GLY1-GLY50 of chainA
1AFP,THR2-CYS51 of chainA
2DKM,GLY1-PRO10 of chainA
2ECH,GLU1-THR50 of chainA
2KNX,GLY1-THR50 of chainA
2M3E,SER1054-GLY1103 of chainA
2N4G,MET311-GLN360 of chainA
2N71,ALA1-CYS50 of chainA
2NPL,PRO31-GLY80 of chainX
2RPJ,GLY21-ALA70 of chainA
2TGF,VAL1-ALA50 of chainA
Absorbance
Wavelength (nm)

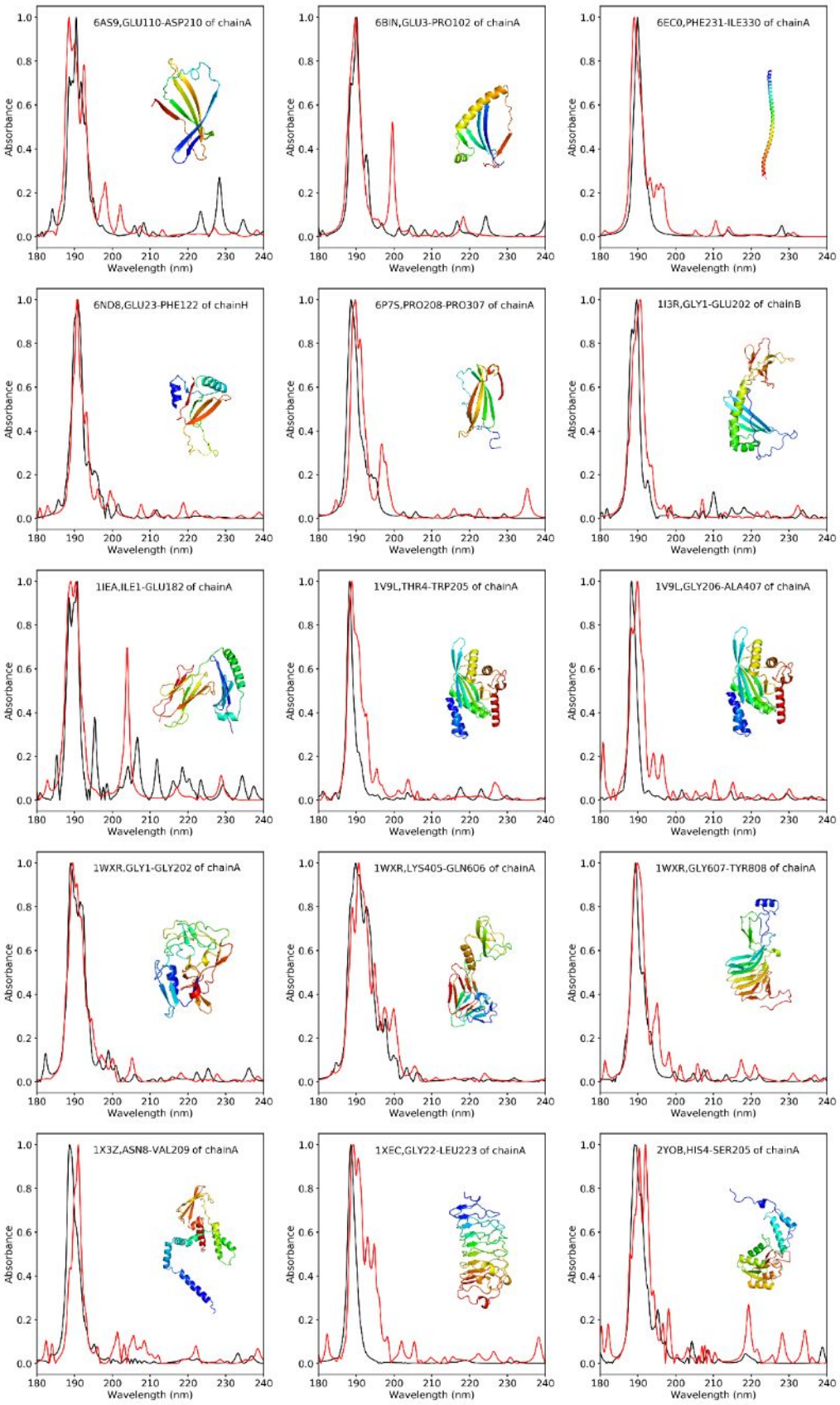

2V15,GLN170-ARG192 of chainB
3TGF,VAL1-ALA50 of chainA
3ZGH,GLU208-ASP218
4AIT,SER21-LEU70 of chainA
4AU3,GLN298-ASP316 of chainA
4BT9,SER7-LYS55 of chainA
4IGI,ILE1160-GLN1173
4IHK,LEU1044-SER1054
4OX0,SER50-LEU99 of chainA
5E3X,MET1-TYR28 of chainA
5JUI,THR247-ASN293 of chainA
5NAY,LYS25-TYR35 of chainA
5NAZ,ASP214-CYS225
5ZFO,ASN1-CYS50 of chainA
6FTK,LEU737-MET786 of chainA
Absorbance
Wavelength (nm)

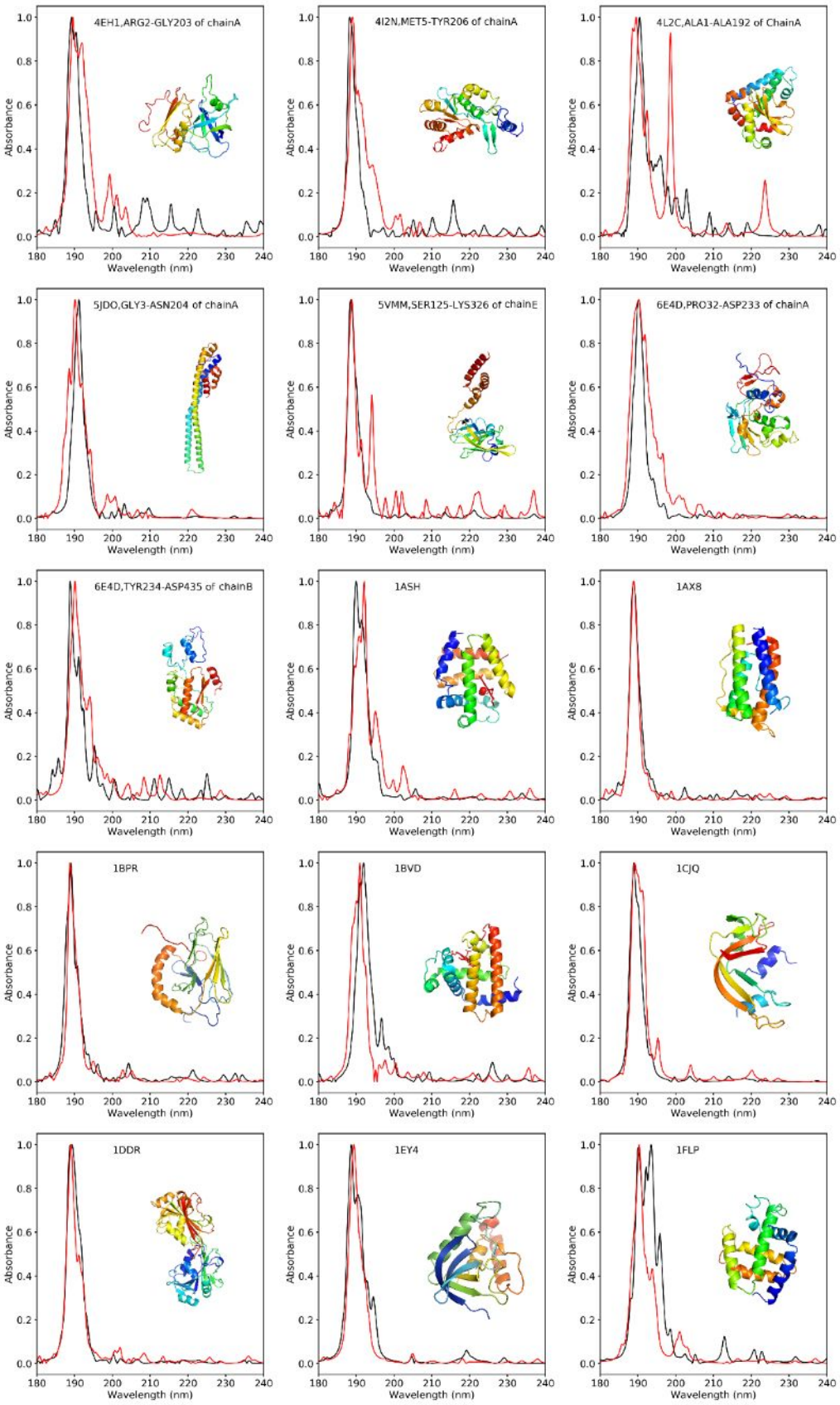

1DY2,ALA101-VAL200 of chainA
1E1G,LEU125-ALA224 of chainA
1GR3,VAL551-VAL650 of chainA
1H4U,PRO401-LYS500 of chainA
1H6W,ARG291-PRO390 of chainA
1HJN,MET129-ARG228 of chainA
1JB3,GLU6-LEU105 of chainA
1JVX,LYS1-ASN100 of chainA
1KOE,PRO201-TYR300 of chainA
1MHP,ILE2-THR101 of chainL
1OCY,ARG330-CYS429 of chainA
1OQV,ARG100-SER199 of chainB
1PK6,ASN101-LYS200 of chainA
1PT6,SER201-LYS300 of chainA
1QC5,GLU422-ILE521 of chainB
Absorbance
Wavelength (nm)

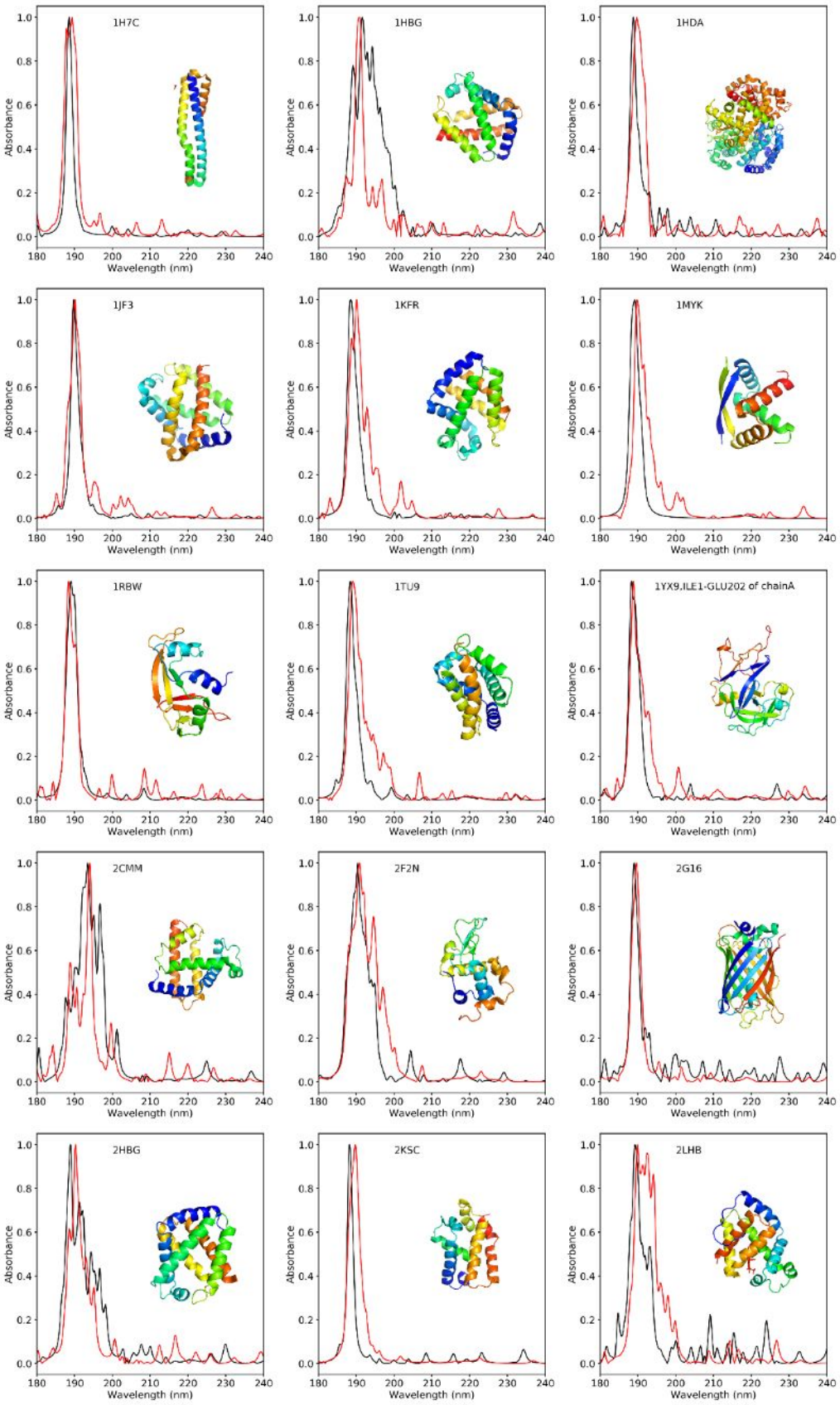

1QIU,THR451-THR550 of chainA
1UW3,LEU125-ALA224 of chainA
1XU0,GLY127-SER226 of chainA
1Y2S,VAL121-ARG220 of chainA
2DKM,GLY1-GLY100 of chainA
2E3M,ASN451-LYS550 of chainA
2J61,CYS73-ASP172 of chainA
2JHH,SER100-GLU199 of chainC
2JHI,SER81-VAL180 of chainF
2KCO,HIS20-GLY119 of chainA
2LLL,SER451-GLU550 of chainA
2LTB,ALA500-VAL599 of chainA
2OKM,LEU61-ASN160 of chainA
2QHT,MET1-ARG100 of chainA
2SEB,TYR30-VAL129 of chainB
Absorbance
Wavelength (nm)

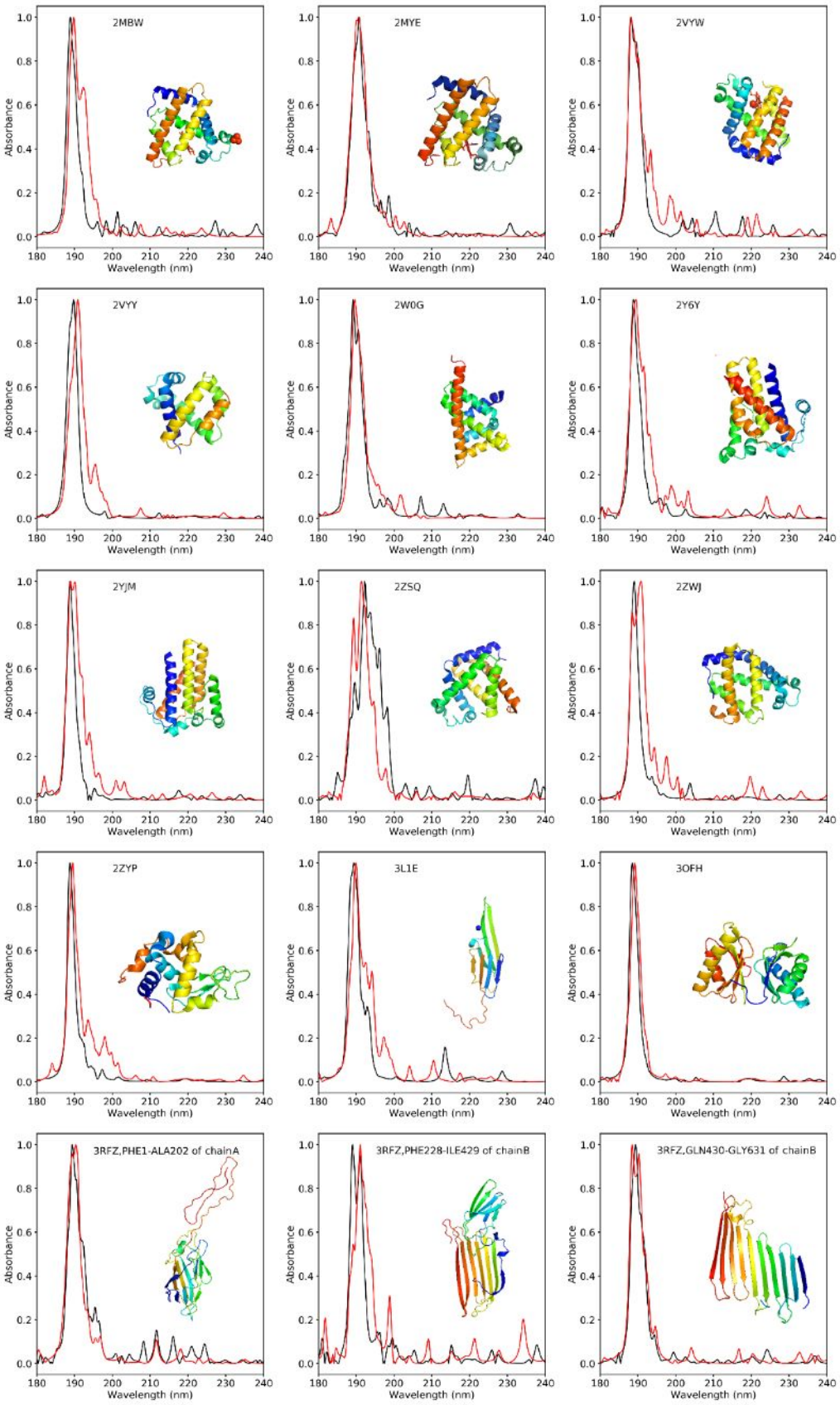

2WNP,SER81-VAL180 of chainF
3ASW,GLY212-ASP311 of chainA
3EWO,GLY50-PHE149 of chainA
3H3Q,GLY363-ALA462 of chainA
3HRV,ARG100-SER199 of chainA
3LD1,LYS70-ASN169 of chainA
3LFO,MET0-LYS99 of chainA
3OLY,LYS5-GLU104 of chainA
3PG0,ILE30-SER129 of chainA
3RT5,LYS1-SER100 of chainX
3SEA,GLN3-LYS102 of chainA
3TVN,ASP101-THR200 of chainX
3UMM,SER431-ILE530 of chainA
3UYX,PHE101-ASN200 of chainA
3VZ9,THR131-PHE230 of chainB
Absorbance
Wavelength (nm)

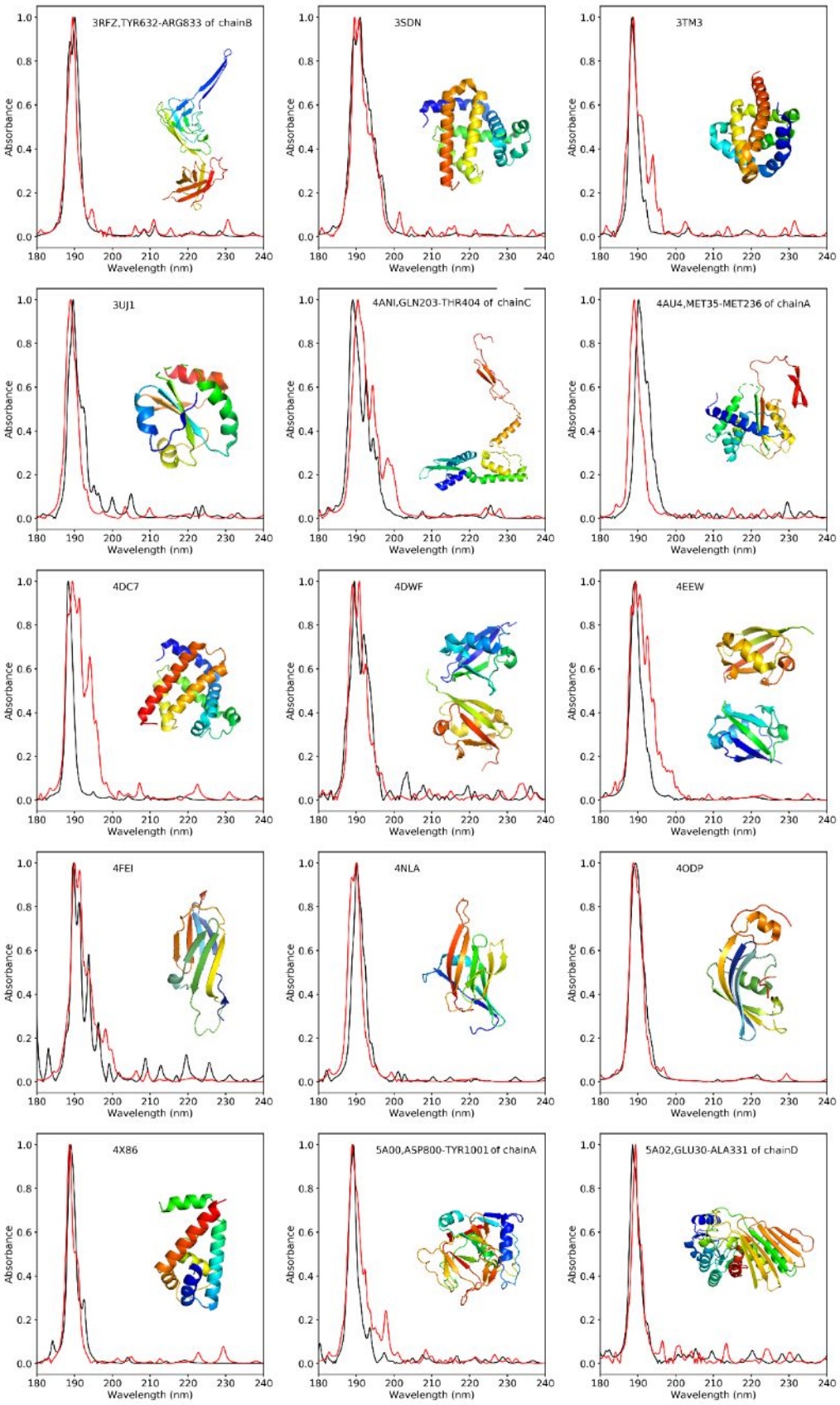

4A0Q,VAL161-HIS260 of chainA
4BKL,GLN1-GLY100 of chainA
4D3E,GLY125-ASN224 of chainD
4F3J,ASP120-ALA219 of chainA
4FUP,VAL101-GLU200 of chainA
4IGI,PRO1101-GLU1200 of chainA
4J5M,VAL1459-HIS1558 of chainA
4NYT,PRO72-VAL171 of chainC
4TT0,LYS1601-SER1700 of chainA
4XIF,SER312-ALA411 of chainA
4YXL,GLY126-TYR225 of chainA
5DUT,VAL131-TRP230 of chainA
5HBA,GLU2-GLY101 of chainA
5K31,VAL6-HIS105 of chainA
5MU0,GLU1-GLY100 of chainA
Absorbance
Wavelength (nm)

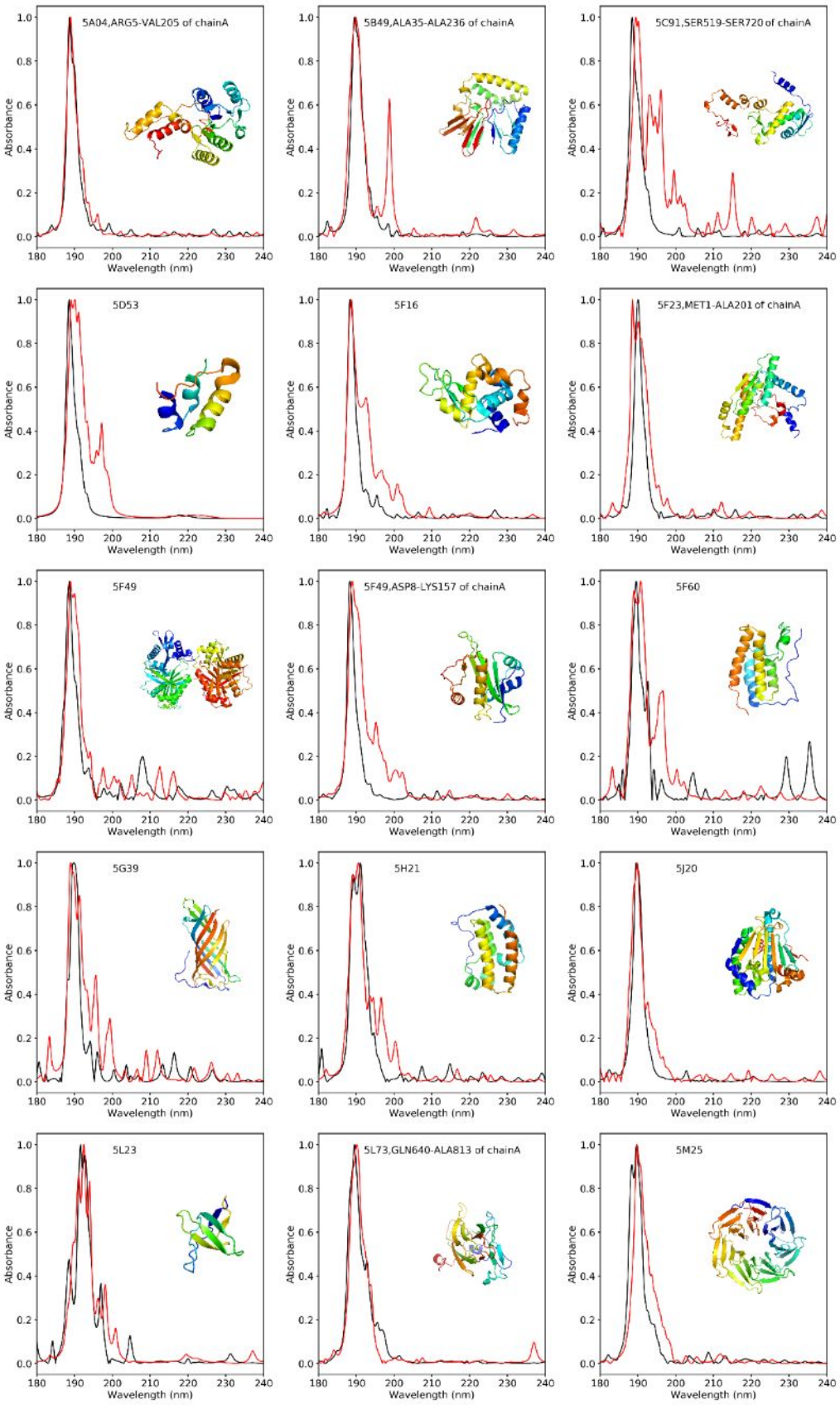

6AS9,GLU110-ASP210 of chainA
6BIN,GLU3-PRO102 of chainA
6EC0,PHE231-ILE330 of chainA
6ND8,GLU23-PHE122 of chainH
6P7S,PRO208-PRO307 of chainA
1I3R,GLY1-GLU202 of chainB
1IEA,ILE1-GLU182 of chainA
1V9L,THR4-TRP205 of chainA
1V9L,GLY206-ALA407 of chainA
1WXR,GLY1-GLY202 of chainA
1WXR,LYS405-GLN606 of chainA
1WXR,GLY607-TYR808 of chainA
1X3Z,ASN8-VAL209 of chainA
1XEC,GLY22-LEU223 of chainA
2YOB,HIS4-SER205 of chainA
Absorbance
Wavelength (nm)

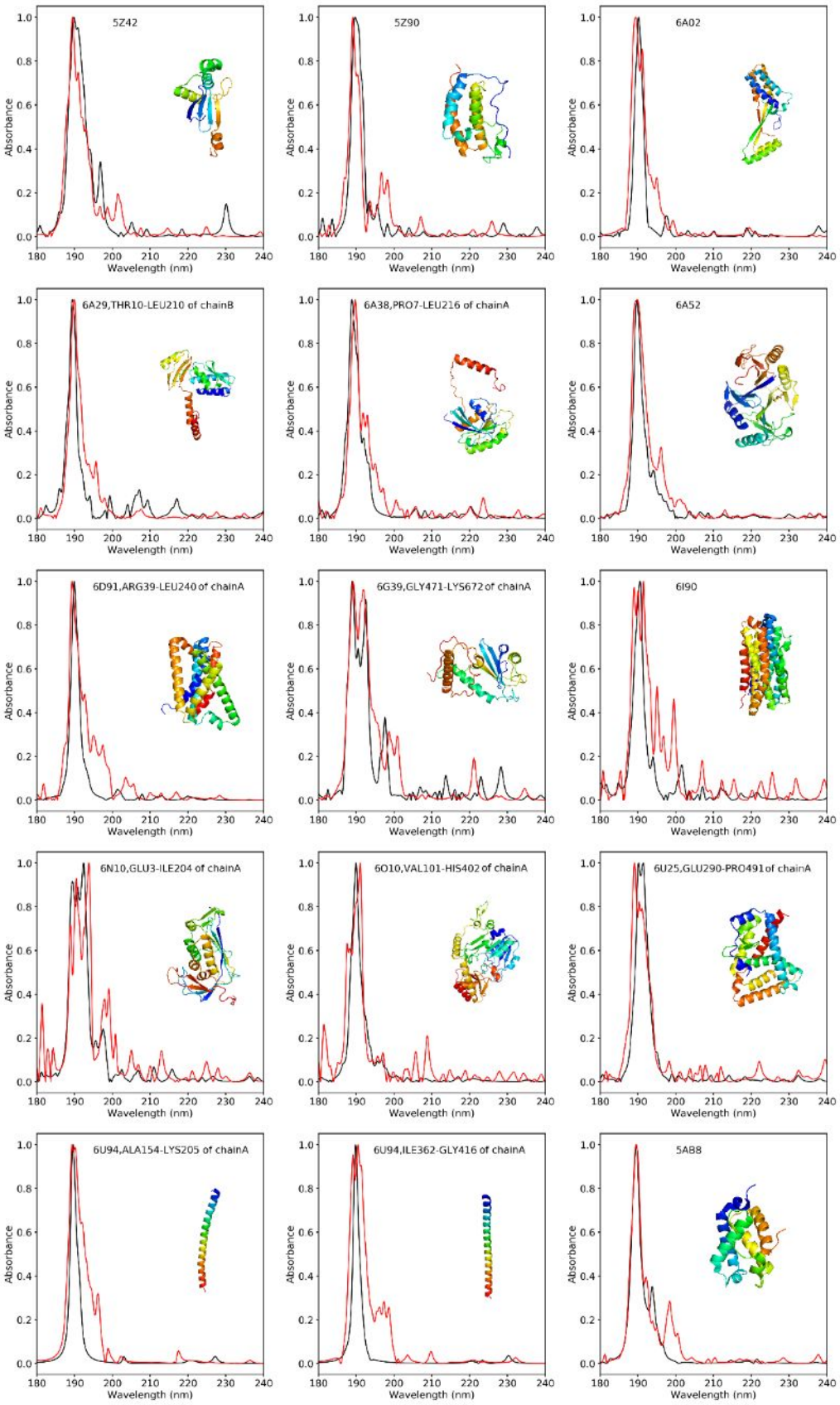

4EH1,ARG2-GLY203 of chainA
4I2N,MET5-TYR206 of chainA
4L2C,ALA1-ALA192 of ChainA
5JDO,GLY3-ASN204 of chainA
5VMM,SER125-LYS326 of chainE
6E4D,PRO32-ASP233 of chainA
6E4D,TYR234-ASP435 of chainB
1ASH
1AX8
1BPR
1BVD
1CJQ
1DDR
1EY4
1FLP
Absorbance
Wavelength (nm)

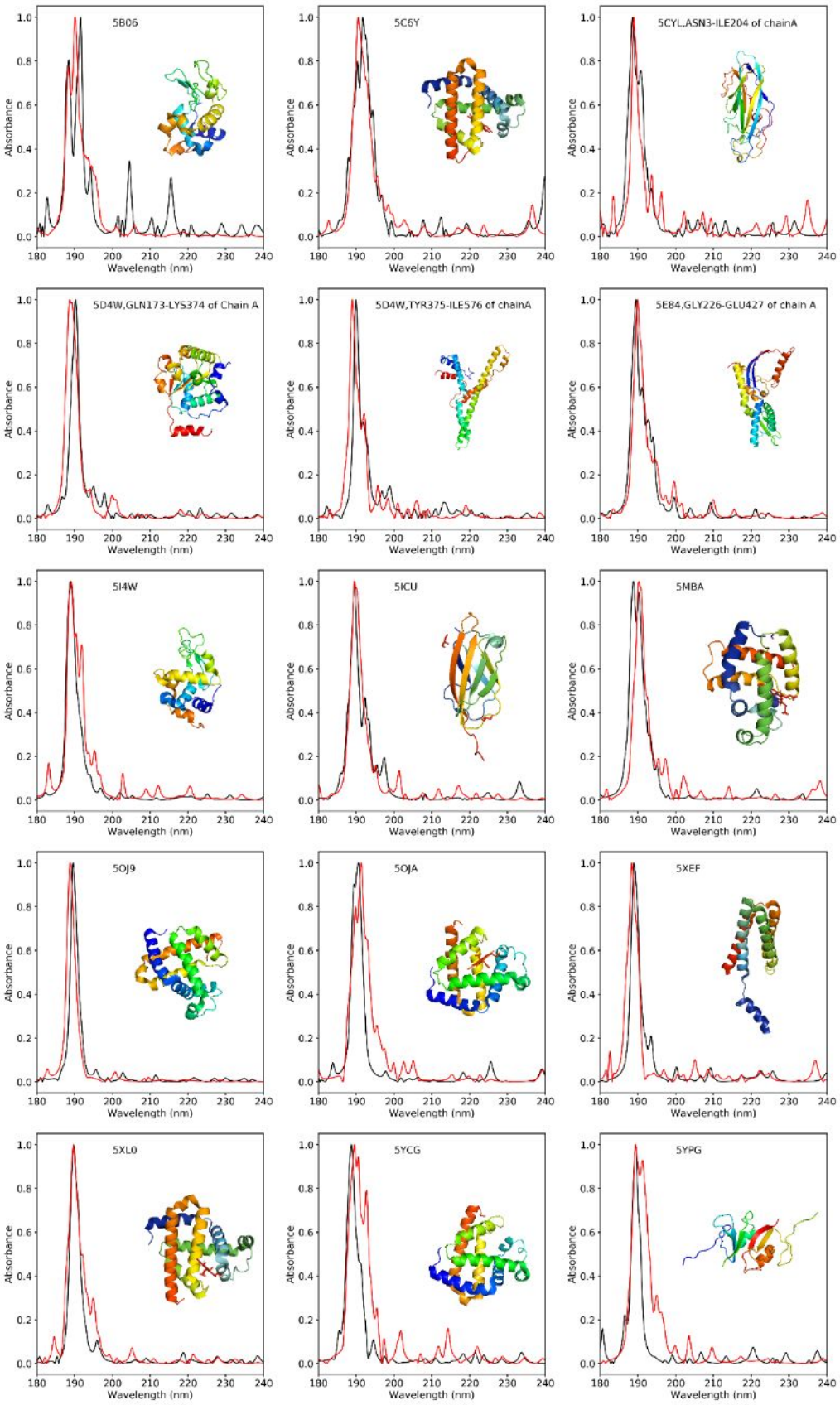

1H7C
1HBG
1HDA
1JF3
1KFR
1MYK
1RBW
1TU9
1YX9,ILE1-GLU202 of chainA
2CMM
2F2N
2G16
2HBG
2KSC
2LHB
Absorbance
Wavelength (nm)


2MBW
2MYE
2VYW
2VYY
2W0G
2Y6Y
2YJM
2ZSQ
2ZWJ
2ZYP
3L1E
3OFH
3RFZ,PHE1-ALA202 of chainA
3RFZ,PHE228-ILE429 of chainB
3RFZ,GLN430-GLY631 of chainB
Absorbance
Wavelength (nm)


3RFZ,TYR632-ARG833 of chainB
3SDN
3TM3
3UJ1
4ANI,GLN203-THR404 of chainC
4AU4,MET35-MET236 of chainA
4DC7
4DWF
4EEW
4FEI
4NLA
4ODP
4X86
5A00,ASP800-TYR1001 of chainA
5A02,GLU30-ALA331 of chainD
Absorbance
Wavelength (nm)


5A04,ARG5-VAL205 of chainA
5B49,ALA35-ALA236 of chainA
5C91,SER519-SER720 of chainA
5D53
5F16
5F23,MET1-ALA201 of chainA
5F49
5F49,ASP8-LYS157 of chainA
5F60
5G39
5H21
5J20
5L23
5L73,GLN640-ALA813 of chainA
5M25
Absorbance
Wavelength (nm)


5Z42
5Z90
6A02
6A29,THR10-LEU210 of chainB
6A38,PRO7-LEU216 of chainA
6A52
6D91,ARG39-LEU240 of chainA
6G39,GLY471-LYS672 of chainA
6I90
6N10,GLU3-ILE204 of chainA
6O10,VAL101-HIS402 of chainA
6U25,GLU290-PRO491 of chainA
6U94,ALA154-LYS205 of chainA
6U94,ILE362-GLY416 of chainA
5AB8
Absorbance
Wavelength (nm)


5B06
5C6Y
5CYL,ASN3-ILE204 of chainA
5D4W,GLN173-LYS374 of Chain A
5D4W,TYR375-ILE576 of chainA
5E84,GLY226-GLU427 of chain A
5I4W
5ICU
5MBA
5OJ9
5OJA
5XEF
5XL0
5YCG
5YPG
Absorbance
Wavelength (nm)

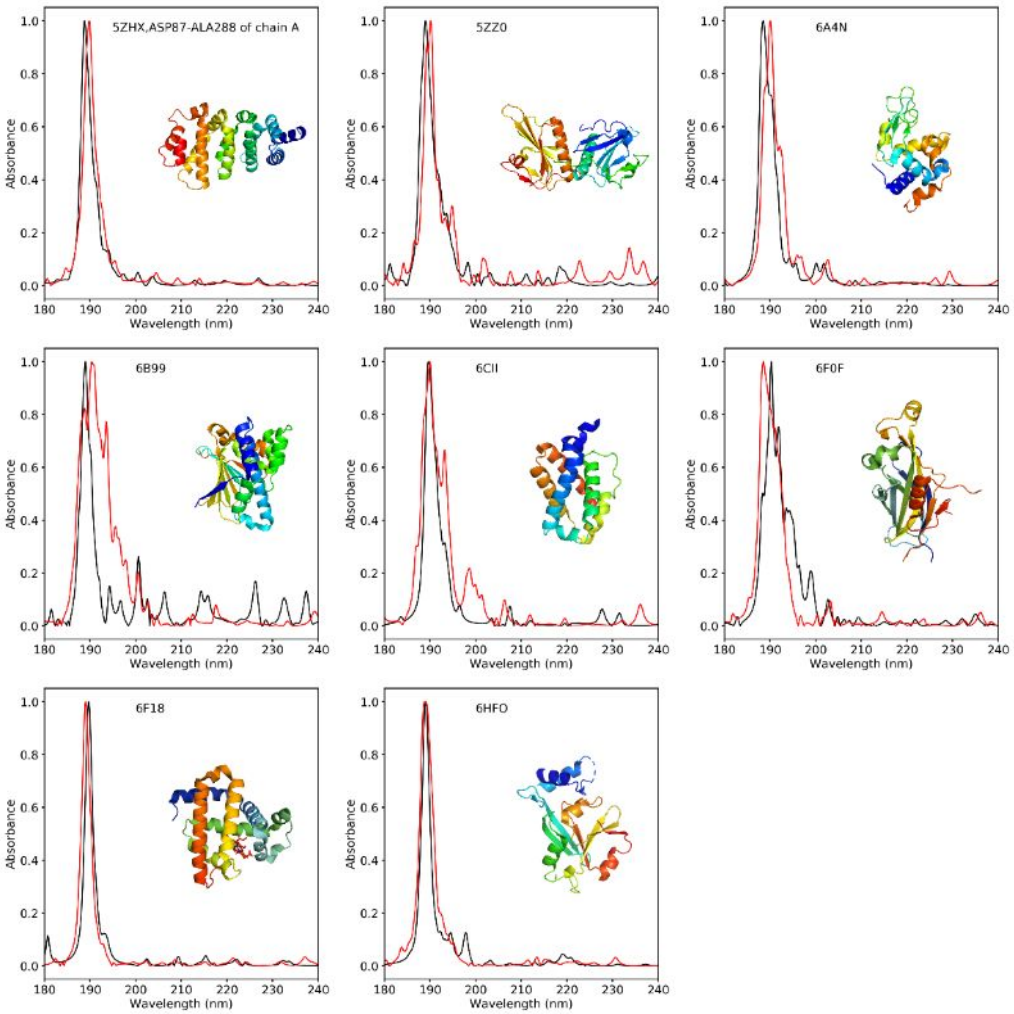

5ZHX,ASP87-ALA288 of chain A
5ZZ0
6A4N
6B99
6CII
6F0F
6F18
6HFO
Absorbance
Wavelength (nm)

Table S2. PDB ID of 1000 proteins downloaded from RCSB Protein Data Bank for extracting 50 000 peptides and 200 000 residues (20 types of residues and 10 000 structures of each residue). The downloaded proteins cover different types of proteins, including fibrous protein, globular protein, keratin, collagen, chaperone, myoglobin, hemoglobin and denaturation.

1A00	1A01	1A0U	1A3O	1A4F	1A6G	1A6M	1ABY	1AH6	1AH8	1AJ9	1AMX
1ANB	1AOX	1B0B	1B86	1B9Q	1BBB	1BF8	1BIJ	1BKV	1BUW	1BUY	1BVC
1C40	1CBL	1CG5	1CG8	1CH4	1CK7	1CLG	1CMY	1CN4	1CO9	1COH	1CPZ
1DG4	1DGF	1DGH	1DKE	1DKG	1DKX	1DKY	1DLW	1DM1	1DXU	1DY2	1DZI
1ECD	1EER	1EZU	1F4J	1FAW	1FCS	1FDH	1FHJ	1FM1	1FSZ	1FUJ	1G08
1G0A	1G3J	1GCV	1GJN	1GR3	1GVL	1GXD	1GZX	1H1X	1HAB	1HBA	1HBH
1HBS	1HCO	1HGA	1HGB	1HGC	1HK7	1HX1	1HYL	1I6Z	1I7X	1IBE	1IRD
1IWH	1J14	1J3Z	1J52	1J7W	1J7Y	1JBK	1JJ9	1JWN	1JY7	1JZK	1JZL
1JZM	1K0V	1K0Y	1K9O	1KD2	1KHY	1KIU	1KKE	1KR7	1LFL	1LFQ	1LFT
1LFV	1LI1	1M3D	1M9P	1MBA	1MBD	1MBN	1MBO	1MBS	1MGN	1MKO	1MOH
1MWB	1MYH	1MYI	1MYM	1MYZ	1MZ0	1N9X	1NEJ	1NIH	1NPF	1NPG	1NQP
1NWI	1NWN	1O1I	1O1K	1O1N	1O91	1P9H	1PBX	1PMB	1Q5L	1Q7D	1QI8
1QPW	1QQW	1QUN	1QVR	1QXD	1R1X	1R1Y	1ROC	1RPS	1RTX	1RVW	1S5Y
1S69	1S6A	1SB6	1SDK	1SDL	1SHR	1SI4	1SLU	1SPG	1SS8	1SWM	1T08
1T60	1T7S	1THB	1U5M	1U7S	1U97	1UIW	1UMK	1US7	1USU	1UVY	1V4U
1V4W	1V4X	1V8X	1V9Q	1W09	1W0A	1W0B	1WG3	1WVP	1WXV	1X46	1X9F
1XUC	1XXT	1XYE	1XZ2	1XZY	1Y01	1Y09	1Y4P	1Y5J	1Y8H	1Y8I	1YCA
1YDZ	1YEO	1YEQ	1YGF	1YHU	1YIE	1YJP	1YKT	1YMB	1YOU	1YVQ	1YVT
1YZI	1Z2G	1Z8U	1ZAV	1ZE3	1ZTQ	1ZWH	2A3G	2AA1	2AKP	2AV0	2B7H
2BPR	2BRC	2BRE	2BW9	2BWH	2C0K	2CG9	2CGE	2D1N	2D2M	2D3E	2D5X
2D5Z	2D60	2D6C	2DHB	2DN1	2DN2	2DN3	2DXM	2E2D	2E2Y	2E3O	2E3R
2EKU	2EVP	2F6A	2FAM	2FRF	2FRJ	2FSE	2FXS	2G0S	2G12	2GTL	2H35
2H8D	2H8F	2HBC	2HBD	2HBF	2HBS	2HCO	2HHB	2HHD	2HHE	2HP8	2HUE
2HZ1	2IDC	2IN4	2IW2	2IWS	2JHO	2KHO	2LKV	2LLL	2LLP	2LM1	2LWP
2LYJ	2LYK	2LYL	2LYP	2LYQ	2LYR	2LYS	2M0M	2M6Z	2M8S	2MB5	2MGO
2MIQ	2MZE	2MZI	2N8R	2NB0	2ND2	2ND3	2ND5	2NRL	2NSR	2NX0	2O5L
2O5Q	2O5S	2OHB	2OJ5	2OKN	2PEI	2PEO	2PEQ	2PGH	2QIF	2QLS	2QSP
2QSS	2QU0	2R1H	2R80	2R9Y	2RAO	2SEB	2UUR	2V1E	2V1F	2V1I	2V1K
2V53	2V7Y	2VLY	2VW5	2W6V	2W6W	2W72	2XD6	2XI6	2XIF	2XIL	2XJ6
2XKI	2XX4	2YRS	2Z44	2Z46	2Z6S	2Z6T	2Z85	2Z9Y	2Z9Z	2ZLV	2ZLW
2ZLX	2ZSP	2ZSS	2ZSY	3A0G	3A2G	3A59	3AEH	3AK5	3AQ5	3ASE	3ASW
3B75	3BJ1	3BWU	3C11	3CIU	3D17	3D1K	3D7O	3DHR	3DLL	3DPO	3DPQ
3DUT	3EDA	3EJH	3ELM	3EOK	3EU1	3FH9	3FP8	3FS4	3FZH	3FZK	3GKV
3GLN	3GOU	3GQG	3GQP	3GYS	3H0X	3H3T	3HC9	3HF4	3HQV	3IA3	3IC0
3IC2	3IUC	3K8B	3KEK	3LDL	3LDN	3LDO	3LDP	3LDQ	3LJZ	3LQD	3LR7
3LW2	3M0B	3M38	3M3B	3MBA	3MJP	3MJU	3MVF	3N3F	3NL7	3NML	3O2X
3ODQ	3OGB	3OVU	3PEL	3PI8	3PI9	3QJE	3QL1	3QZL	3QZM	3QZN	3QZO
3RIK	3RJR	3RTL	3RUR	3S48	3S5C	3S5H	3S5K	3SDH	3SZK	3TFB	3TNU
3TVC	3UHI	3UT2	3V03	3V2V	3VFE	3VM5	3VM9	3VND	3VNW	3VQK	3VQL
3VQM	3W6L	3WFT	3WHM	3WI8	3WTG	3WV1	3WVL	3WYO	3ZHC	3ZHD	3ZHK
3ZHL	4A7B	4AIX	4AIZ	4AJ0	4AU2	4B2T	4B9Q	4BB2	4BJ3	4BKL	4BNR
4C0N	4C44	4CTD	4CUD	4CUE	4CUF	4D0E	4D2U	4D8N	4DC5	4DF3	4DOU

4EO5	4EZN	4EZO	4EZP	4EZR	4EZW	4EZX	4F01	4F4O	4F68	4FC3	4FCT
4FCW	4FVL	4FWZ	4GR7	4H32	4HRR	4HRT	4HSE	4HWC	4I0C	4I0Y	4I1E
4I2S	4I37	4I3N	4I96	4IJ2	4JA7	4JA9	4JB0	4JB2	4JSD	4JSO	4K07
4K5Q	4K6G	4K6H	4K6K	4KJT	4L2A	4L2D	4LJ6	4LJA	4M4B	4M56	4M8U
4MA7	4MBN	4MKF	4MKG	4MKH	4MPB	4MPR	4MQK	4MTH	4N79	4N7P	4N8W
4NI0	4NSM	4NWE	4NWH	4O4T	4O4Z	4OF9	4OJ0	4OOD	4OW4	4PNJ	4QBY
4R1E	4RMB	4RRP	4RX9	4TQL	4TYU	4U3H	4U5T	4U8U	4UOS	4UOT	4UOX
4UOY	4URG	4URQ	4URS	4UZV	4W68	4W70	4W81	4W94	4WJG	4WUY	4XIF
4XS0	4Y00	4YU3	4YU4	4Z3V	4ZLY	4ZRY	5AKS	5AO6	5AQG	5AQI	5AQO
5AQT	5AUY	5AZQ	5B5O	5B85	5BOY	5BX0	5CE5	5CJB	5CMV	5CN5	5CNC
5CTD	5CTI	5CVA	5CVB	5D0Q	5D5R	5E83	5E84	5E85	5EII	5EIV	5F2R
5FFO	5FQD	5FWL	5FWP	5FXP	5GHU	5GW4	5GW5	5HCL	5HLY	5HQ3	5HY8
5IAT	5IAX	5IKS	5ILM	5ILP	5ILR	5J3P	5J3S	5J3Z	5JG9	5JHI	5JI4
5JOM	5KA0	5KER	5KI0	5KKK	5KRR	5KSI	5KSJ	5KVN	5KWX	5KWZ	5KX0
5KX1	5KX2	5M4G	5M4J	5M4L	5M9M	5MBY	5MC1	5MU0	5MV3	5MZU	5N30
5N4H	5NAX	5NI1	5NIR	5NJX	5NRO	5NX3	5O4P	5OBU	5OCX	5OFO	5OMP
5OMY	5OPW	5OPX	5OU8	5OU9	5OWI	5OWJ	5PKC	5Q5Z	5QEH	5R4J	5SV3
5SV7	5SXD	5THP	5TU7	5TU8	5TU9	5U2L	5U2U	5UCB	5UCU	5UE2	5UE5
5UEA	5UEK	5URC	5UT7	5UT9	5UWK	5UYX	5V4M	5V4N	5VPN	5VQP	5VY8
5VY9	5VZN	5VZO	5VZP	5VZQ	5W0S	5WOG	5X2R	5X2S	5XKV	5Y45	5YAN
5YCE	5YP8	5YPB	5YUP	5YZF	5Z5O	5ZBA	5ZHB	5ZUI	5ZYK	5ZZF	5ZZG
5ZZT	5ZZY	6A06	6A0H	6A0V	6A0Y	6A19	6A1W	6A23	6A2U	6A32	6A39
6A3C	6AHF	6AIT	6ASY	6AXB	6BB5	6BIN	6BJR	6BNR	6BWU	6CD2	6CF0
6CQG	6CQV	6D45	6D6S	6DDK	6DFM	6DJU	6DL9	6DTC	6E14	6E15	6E0F
6E0G	6E2J	6E7G	6E7H	6EC0	6ED3	6EOF	6F0Y	6F17	6F25	6FQF	6FSE
6FZW	6G5A	6G5T	6GCQ	6GZD	6H2P	6H2Q	6HAL	6HBI	6HBW	6HG7	6HV2
6IHX	6II1	6IWK	6J0A	6J81	6JBX	6JP1	6M8F	6MV0	6N02	6N8V	6N8Z
6NBC	6NBD	6ND8	6NDH	6O5V	6O69	6OG3	6QFF	6QFH	6QH9	6QI8	6REU
6S0F	6TSZ	6UUV	6VGK	6W75	6XV4	6Y6W	7ABP	7ACN	7AHL	7AME	7API
7BNA	7CA2	7CCP	7CEI	7CEL	7CGT	7DFR	7FAB	7FD1	7GAT	7GCH	7HSC
7HVP	7ICD	7ICE	7ICF	7ICN	7ICO	7ICQ	7ICR	7ICV	7INS	7KME	7LPR
7LYZ	7LZM	7MHT	7MSF	7NN9	7NSE	7PAZ	7PCK	7PTD	7R1R	7REQ	7RSA
7RXN	7STD	7TIM	7TLN	7WGA	7XIM	7YAS	7ZNF	821P	830C	8A3H	8AAT
8ABP	8ACN	8ADH	8AME	8API	8AT1	8ATC	8BNA	8CA2	8CAT	8CGT	8CHO
8CPA	8CPP	8DFR	8DRH	8EST	8FAB	8GCH	8GEP	8GPB	8GSS	8HVP	8I1B
8ICA	8ICZ	8JDW	8KME	8LDH	8LPR	8LYZ	8MHT	8MSI	8NSE	8OHM	8PAZ
8PCH	8PRK	8PRN	8PSH	8PTI	8RAT	8RNT	8RSA	8RUC	8RXN	8TFV	8TIM
8TLI	8TLN	8XIA	8XIM	9ABP	9AME	9ANT	9ATC	9CA2	9CGT	9DNA	9EST
9GAA	9GAC	9GAF	9GPB	9GSS	9HVP	9ICA	9ICC	9ICE	9ICH	9ICJ	9ICK
9ICM	9ICO	9ICQ	9ICS	9ICU	9ICV	9ICY	9ILB	9INS	9JDW	9LDB	9LDT
9LPR	9LYZ	9MHT	9MSI	9NSE	9PAI	9PAP	9PCY	9PTI	9RAT	9RNT	9RSA
9RUB	9WGA	9XIA	9XIM								

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*, 79-84.

(2) Frenkel, Y., On the Transformation of Light into Heat in Solids. I. *J. Phys. ReV.* **1931**, *37*, 17-44.

(3) Kasha, M.; Rawls, H.; El-Bayoumi, M. A., The Exciton Model in Molecular Spectroscopy. *Pure Appl. Chem.* **1965**, *11*, 371-392.

(4) Zhang, Y.; Luo, Y.; Zhang, Y.; Yu, Y. J.; Kuang, Y. M.; Zhang, L.; Meng, Q. S.; Luo, Y.; Yang, Y. J.; Dong, Z. C.; Hou, J. G., Visualizing Coherent Intermolecular Dipole-Dipole Coupling in Real Space. *Nature* **2016**, *531*, 623-627.

(5) Schrödinger, L. & Delano, W., 2020. Pymol, Available At: http://www.pymol.org/pymol.

(6) 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*, 8100-8107.

(7) Lu, T.; Chen, F., Multiwfn: A Multifunctional Wavefunction Analyzer. *J. Comput. Chem.* **2012**, *33*, 580-592.

(8) 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, Revision A.03, Gaussian, Inc., Wallingford CT, **2016**.

(9) 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*, 1341-1346.

(10) 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*, 1701-1718.

(11) Li, X.; Chiong, R.; Hu, Z.; Cornforth, D.; Page, A. J., Improved Representations of Heterogeneous Carbon Reforming Catalysis Using Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 6882-6894.

(12) Rupp, M.; Tkatchenko, A.; Muller, K.; Von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(13) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326-2331.

(14) Behler, J., Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(15) Gastegger, M.; Schwiedrzik, L.; Bittermann, M.; Berzsenyi, F.; Marquetand, P., wACSF-Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241709.

(16) 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*, 4962-4967.