JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

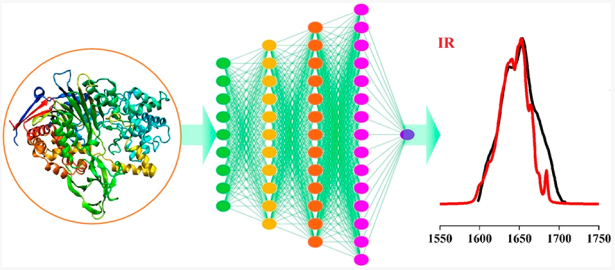
# A Machine Learning Protocol for Predicting Protein Infrared Spectra

Sheng Ye,[⊥] Kai Zhong,[⊥] Jinxiao Zhang,[⊥] Wei Hu, Jonathan D. Hirst, Guozhen Zhang, Shaul Mukamel, and Jun Jiang*

Cite This: J. Am. Chem. Soc. 2020, 142, 19071−19077
Read Online



**ABSTRACT:** Infrared (IR) absorption provides important chemical fingerprints of biomolecules. Protein secondary structure determination from IR spectra is tedious since its theoretical interpretation requires repeated expensive quantum-mechanical calculations in a fluctuating environment. Herein we present a novel machine learning protocol that uses a few key structural descriptors to rapidly predict amide I IR spectra of various proteins and agrees well with experiment. Its transferability enabled us to distinguish protein secondary structures, probe atomic structure variations with temperature, and monitor protein folding. This approach offers a cost-effective tool to model the relationship between protein spectra and their biological/chemical properties.






## INTRODUCTION

Understanding the function of proteins benefits enormously from knowledge of their atomistic structure. Infrared (IR) absorption spectroscopy, combined with atomic coordinates from first-principles simulations, offers an effective tool for probing the atomic-level structure of proteins.[1−3] The amide I region (1600−1700 $cm^{-1}$), dominated by the stretching vibration of the carbonyl group in the peptide bond, provides a fingerprint of protein structure and dynamics and has been the subject of extensive experimental and computational effort.[2,4−6] The theoretical interpretation of spectroscopic signals and connecting them with structural detail is an expensive task, which requires many electronic structure calculations at the quantum chemistry (QC) level for a large number (typically thousands) of representative configurations.

For decades, the map methods have been widely used,[7−10] to predict vibrational properties without large-scale QC calculations. Their basic philosophy is to compute (or predict) vibrational modes by using empirical polynomial functions in the local electric fields around the targeted molecules or amide group.[3,7] We have also developed several map models and employed them in a couple of studies on protein vibrational spectra.[11,12] However, the transferability of map methods is limited since a few-parameter fitting of observables to key structural parameters cannot account for the full versatility and complexity of proteins.[7,13] The biased parametrization might bring errors in spectroscopic simulations. Developing a cost-effective approach that has greater predictive power and transferability is called for.

There is a resurgence of interest, fueled by large data sets, advanced algorithms, and faster computers, in machine learning (ML), a class of artificial intelligence methods that gain predictive power from learning of data, as a powerful toolkit for modeling structure−property relationships in molecules and materials, such as predicting chemical reaction routes and accelerating discovery of materials.[14−18] In particular, neural networks (NN), a class of machine-learning algorithms, can establish the structure−property relationships by iteratively learning with a complex high-dimensional function.[19] NN has been proven useful for handling complex nonlinear problems and offers a transferable tool for simulating protein spectroscopy[20] and for predicting the frequency and transition dipole moments of the O−H stretch in water.[13] Gastegger et al. used NN to accelerate *ab initio* molecular dynamics (AIMD) to compute accurate IR spectra for materials,[21] and Ghosh et al. used Deep Neural Networks (DNN) to obtain spectra information directly from the molecular structure, which greatly accelerates the spectroscopic analysis of materials.[22]

It is always a worthy goal to realize first-principles predictions of IR spectra of proteins, despite its computationally prohibitive difficulties. In this study, we develop an ML protocol for predicting the amide I IR spectra of proteins with density functional theory (DFT) accuracy. The simulated fine structure of IR signals of various proteins from the trained ML model agrees well with experiment. Applications are presented for the identification of secondary structures, probing structural variations with temperature, and monitoring of protein folding.

**Received:** June 16, 2020
**Published:** October 30, 2020




https://dx.doi.org/10.1021/jacs.0c06530
J. Am. Chem. Soc. 2020, 142, 19071−19077

## THEORY AND COMPUTATION DETAIL

**Quantum-Mechanics Treatment for Amide I Vibration.** We adopt a divide-and-conquer strategy to treat the amide I vibrations of the whole protein. The protein vibrations are represented as a set of $n$ oscillators associated with each peptide bond in its backbone. The Frenkel exciton model is employed to construct a vibrational model Hamiltonian,[23] in which the diagonal elements are the frequency ($\omega_i$) of the $i$th amide I oscillator and the off-diagonal elements represent the coupling between two oscillators $i$ and $j$ (Figure 1). For a pair of non-neighboring oscillators, since the distances between oscillators are greater than their sizes, the coupling is calculated with the dipole approximation:[24]

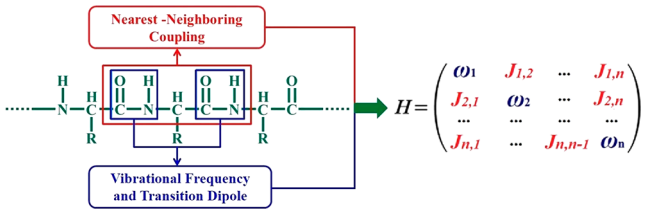



**Figure 1.** Model Hamiltonian for amide I vibrations in a protein.

$J_{ij} = \frac{1}{4\pi\epsilon_0}\left(\frac{\vec{\mu_i}\cdot\vec{\mu_j}}{r_{ij}^3} - 3\frac{(\vec{\mu_i}\cdot\vec{r_{ij}})(\vec{\mu_j}\cdot\vec{r_{ij}})}{r_{ij}^5}\right)$, where $\epsilon_0$ is the dielectric constant, $\vec{\mu}_i$ ($\vec{\mu}_j$) is the transition dipole of peptide bond $i$ ($j$), and $r_{ij}$ is the vector connecting dipole $i$ and $j$. For two neighboring oscillators, the couplings are computed directly using a dipeptide model.[10,25]

**Machine Learning Protocol for the Vibrational Hamiltonian Matrix.** The direct QC calculations of the necessary molecular quantities are time-consuming. Our aim is to predict the vibrational frequency ($\omega_i$), transition dipole ($\vec{\mu}_i$), and neighboring coupling ($J_{ij}$) parameters from an NN model. The *N*-methylacetamide (NMA) molecule (Figure S1) was taken as the model system for NN training. It represents the peptide bond moiety and has been widely used for generating parameters by the empirical map method.[11,26] For the vibrational couplings between two neighboring peptide bonds, we employed the *N*-acetyl-glycine-*N*′-methylamide (GLDP) molecule (Figure S1), also known as the glycine dipeptide. This molecule has been widely used to construct a map of the coupling as a function of the Ramachandran angles ($\phi$ and $\psi$) between the neighboring peptides.[10,27,28]

**Quantum Mechanical Calculations for Data Generation.** Configurations of NMA were extracted from AIMD simulations at 300 K in the NVT ensemble, conducted with the CP2K[29] program (details in Supporting Information). In order to sample relevant configurations, we have run seven independent AIMD simulations with different initial conformations (Figure S2). From 241.5 ps trajectories with a 0.5 fs time step a total of 9660 configurations were extracted at 25 fs intervals to avoid overly correlated configurations for machine learning training. For each configuration, we extracted the NMA molecule and surrounding water molecules within a 5 Å radius for the Hessian calculations using the Gaussian 16 package[30] at the B3LYP/cc-pVDZ level to generate data for machine learning training. The harmonic vibrational frequencies were scaled by 0.97.

A total of 5128 structures of GLDP molecules were generated with the Ramachandran angles $-180° \le \phi \le 180°$ and $-180° \le \psi \le 180°$ at 5° intervals for both angles for machine learning training (Figure S1). Then all Ramachandran angles were fixed, and the remaining coordinates were optimized.[10] The Hessian calculations were performed on the obtained structures, and solvation effects were modeled implicitly by the integral equation formalism polarizable continuum model, using the Gaussian16 package at the B3LYP/cc-pVDZ level. The local coupling of nearest neighbor amide I vibrational modes was calculated by the localizing normal modes scheme of Jacob and Reiher.[25,31]

**Data Analytics.** A total of 9660 NMA and 5128 GLDP conformations were generated as a training set to predict the vibrational frequency ($\omega_i$), transition dipole ($\vec{\mu}_i$), and neighboring coupling ($J_{ij}$). The calculated root-mean-square deviation (RMSD) of the extracted NMA (NMA molecule and surrounding water molecules within a 5 Å radius) and GLDP molecules indicate large conformational changes and low similarity (Figures S3 and S4), which mitigates issues originating from overcorrelation in training data. The broad distribution of training data ($\omega_i$, $\vec{\mu}_i$, $J_{ij}$) indicated that the sampling procedure adequately covered the ensemble of conformations (Figures S2 and S5), and the resulting data set is appropriate for establishing structure−property relationships via ML training.

**Neural Network Architecture and Descriptors.** Multilayer perceptron (MLP) with a supervised training scheme using a back-propagation algorithm implemented in TensorFlow[32] was used to predict the properties ($\omega_i$, $\vec{\mu}_i$, $J_{ij}$) from the geometric descriptors. We have chosen MLP for two reasons: (1) it handles regression problems well, which this work belongs to; (2) it is simple and easy to implement.[13,33,34] The NN consists of one input layer, three hidden layers, and one output layer. For each hidden NN layer we used the rectified linear unit activation function.[35] The number of hidden layer neurons are 32, 64, and 128, respectively. We adopted different learning rates of the Adam optimizer[36] in TensorFlow for the training process to avoid being trapped into local minima. The learning rate is set to be halved every 500 steps, and the initial learning rate is set to 0.0004. For the training, we added L2 regularization[37] to the architecture of the neural network to prevent overfitting. Hyperparameters were optimized by using a random search algorithm[38] in TensorFlow (including neurons for hidden layers, learning rate, and L2 regularization parameters) to create a reasonable ML protocol in this work.

In order to establish a structure−property relationship between geometry and optical properties of proteins, the ground state Coulomb matrix[39] (CM) of the NMA and some of the GLDP molecules (excluding hydrogens and the solvent molecule) was taken as a descriptor (Figure S6): $U_{ij} = \begin{cases} 0.5Q_i^{2.4} \ \forall\ i = j \\ Q_iQ_j/|R_i - R_j| \ \forall\ i \neq j \end{cases}$, where $i$ and $j$ are atomic indices, $|R_i - R_j|$ is the interatomic distance, and $Q_i$ represents nuclear charge. The merit of the CM lies in its simplicity and efficiency;[39] it is also a sufficient descriptor for the molecular spectra.[22] Internal coordinates were also tested for the ML training; we did not adopt it since it is less accurate and efficient than the Coulomb matrix (Figure S7). As a rotationally invariant descriptor, the CM lacks orientation information for predicting the vibrational transition dipole moment ($\vec{\mu}_i$). To remove the complexity of orientation dependence during the NN training for $\vec{\mu}_i$, a rotation matrix operation was applied on each NMA, to set the carbonyl C


https://dx.doi.org/10.1021/jacs.0c06530
J. Am. Chem. Soc. 2020, 142, 19071−19077

atom as the zero point in the *xyz* Cartesian coordinate, the C−O bond along the positive *y* axis, and the ∠OCN triangle in the *x*−*y* plane (Figure S8). Consequently, the NN prediction of the $\vec{\mu}_i$ for a new NMA also starts with a treatment of transferring it back to the original coordinate by using the inverse of the rotation matrix. The elements of the CM (NMA: 15; GLDP: 21) were then used as inputs (Figure S6) for NN training, and the output (size: 1) data are then compared with DFT calculations (Figure 2). A total of five ML models ($\omega_i$, $\vec{\mu}_i(x,y,z)$, $J_{ij}$) were obtained to construct the vibration model Hamiltonian.

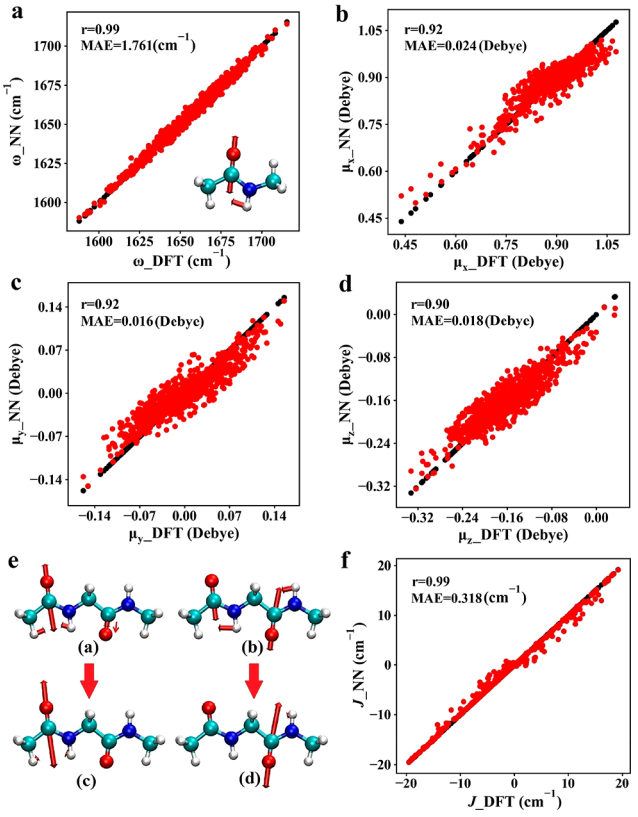



**Figure 2.** (a) Correlation between the DFT-computed ($\omega$_DFT) (black lines/dots) and NN-predicted ($\omega$_NN) (red lines/dots) amide I vibrational frequencies after cross-validation. (b−d) Comparison of the DFT-computed amide I vibrational transition dipole moment in the *x*, *y*, *z* direction ($\mu_{x,y,z}$_DFT) (black lines/dots) and NN-predicted ($\mu_{x,y,z}$_NN) (red lines/dots) after cross-validation. (e) Amide I vibrational normal modes (a, b) and local modes (c, d) of GLDP with DFT B3LYP/cc-pVDZ. (f) Comparison of DFT-computed (*J*_DFT) (black lines/dots) and NN-predicted (*J*_NN) (red lines/dots) coupling constants of nearest neighboring amide I modes after cross-validation.

**Machine Learning Model Evaluation.** The NN predictive accuracy is reported using the Pearson coefficient (*r*) and mean absolute deviation (MAE), and its robustness is verified by the standard cross-validation[40] procedure. All data sets were randomly and evenly distributed into 10 bins in this procedure. Each bin was used as a test set, while the remaining nine bins as training set. We have calculated the learning curves of the whole ML process in this work. The learning curves (Figure S10) indicate that the NN training for vibrational frequency and transition dipole moment converges with 6000 NMA samples, while that of coupling constant needs 4000 GLDP samples. Importantly, there is no significant overfit issues after adding the standard L2 regularization[37] treatment to the NN architecture (Figure S10). It is straightforward to predict the frequency and coupling constants because they mainly depend on the ground state structure. However, since the transition properties (e.g., vibration transition dipole moment) involve two different vibration states, it is expected to see more outliers because these quantities are more sensitive to structural changes. This phenomenon indeed poses a great challenge for NN training (Table S8). With the high Pearson coefficient ($r > 0.9$) and low MAE values (1.761 cm$^{-1}$) obtained in cross-validation, we have achieved accurate ML predictions for the vibration frequency ($\omega_i$), transition dipole ($\vec{\mu}_i$), and coupling constants ($J_{ij}$) in the exciton Hamiltonian (Figure 2).

## RESULTS AND DISCUSSION

**Machine Learning Prediction of IR Spectra.** The ML-predicted parameters were applied to construct the amide I band Hamiltonian using the protocol sketched in Figure 3. The

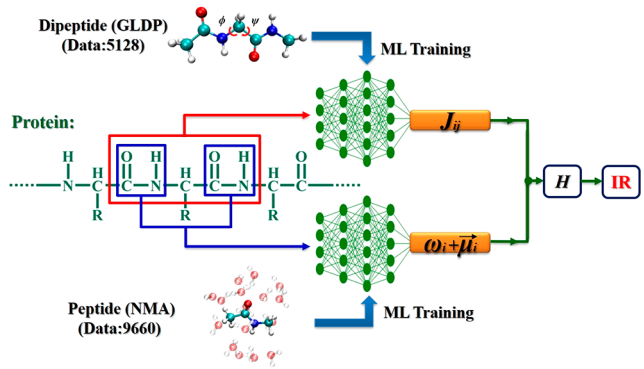



**Figure 3.** Machine learning protocol for predicting protein IR spectroscopy.

protein is split into individual peptide bonds and dipeptides. The $\omega_i$ and $\vec{\mu}_i$ values predicted from the NMA NN model are used to generate Hamiltonian diagonal elements and the off-diagonal elements arising from non-neighboring peptide couplings (computing $J_{ij}$ via the dipole approximation), respectively. The $J_{ij}$ values predicted from the GLDP NN model are used for generating off-diagonal elements owing to the nearest neighboring peptide couplings.

Finally, IR spectra were simulated with the model Hamiltonian, using the SPECTRON program developed by Mukamel and co-workers.[41] As Figure 2 and Figures S11 and S12 show, our ML model can reproduce the DFT data, and we also make this ML protocol[42,43] and simulation data[44] available online to provide rapid protein IR spectroscopy prediction, paving the way for a real-time operation of ultrafast experimental spectroscopy.

**IR Spectroscopic Assignment by ML for Protein Secondary Structures.** Then we applied the ML protocol to simulate the amide I IR spectra of 12 proteins (Figure 4 and Figure S13). The good agreement between our ML predictions and experimental spectra is evident from the high Spearman rank correlation coefficients ($\rho > 0.80$ for 11 cases, except one with 0.71 for 1DHR) (Table 1). This is a widely used measure for the agreement between the predicted and experimental spectra.[45,49−51] The structures of proteins are reflected by distinct spectral characteristics, such as the wavelength region


https://dx.doi.org/10.1021/jacs.0c06530
J. Am. Chem. Soc. 2020, 142, 19071−19077

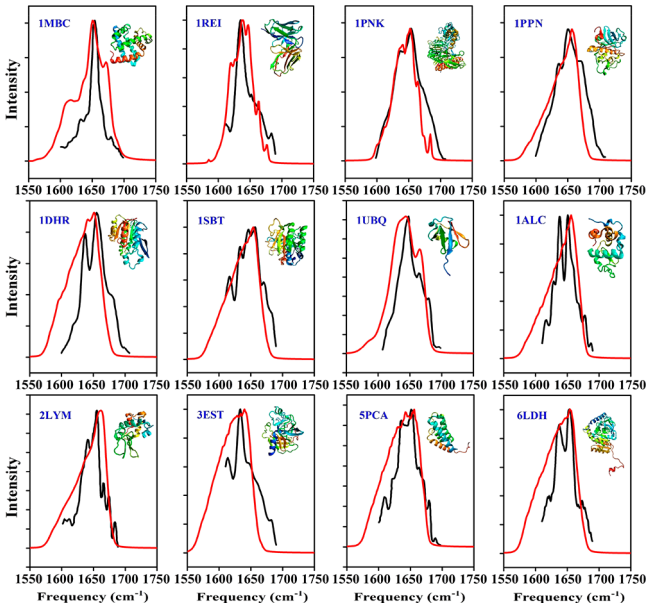



**Figure 4.** Good agreement (the quantitative agreement between the predicted and experimental spectra were measured by Spearman rank correlation coefficients;[45] see Table 1) is obtained between the experimental spectra of the proteins measured in $D_2O$ (black lines)[46−48] and the ML predictions based on 1000 MD configurations (red lines). Intensity is scaled to have the same maximum intensity for each panel.

for the dominant signal peak:[52] $\alpha$-helices: 1640−1650 $cm^{-1}$, $\beta$-sheets: 1620−1640 $cm^{-1}$ and 1680−1690 $cm^{-1}$, random coil: 1650−1660 $cm^{-1}$. As indicated by Table 1 and Figure 4, NN predictions can distinguish the $\alpha$-helix and $\beta$-sheet secondary structures. The $\alpha$-helical (PDB code: 1MBC) and $\beta$-sheet (PDB code: 1REI) proteins exhibit the major spectral peaks at 1650 $cm^{-1}$ and 1634 and 1680 $cm^{-1}$, respectively. Proteins containing both secondary structures ($\alpha + \beta$) show characteristic peaks for both motifs. Taking advantage of the speed of the ML model (Table 1), we can predict the IR spectra by averaging the NN-predicted signals of 1000 MD configurations (which would be prohibitively expensive via direct QC computations), so as to capture the fluctuating dynamics for each protein (details in Supporting Information). The essential features (both main peaks and lineshapes) of experimental spectra are successfully reproduced by the simulated spectra with high Spearman rank correlation coefficient (Figure 4 and Table 1). We have further investigated the amide I signals of $\lambda$-immunoglobulin (1REI) in different states, as shown in Figure S16; the dominant peak of spectra has a blue shift, which corresponds to the change of secondary structure content ($\beta$-turns and coil increased, while $\beta$-strands decreased (Table S4)). We have also predicted transient amide I spectra of 1REI at different moments based on MD trajectories. As Figure S17 shows, the main peak has a red shift accompanied by the decrease of $\beta$-turns and coil content and the increase in $\beta$-strand content (Table S4). The structural change is clearly captured by the change of spectra (Figures S16 and S17). This would be useful for tracking conformational changes of proteins.

**Map Method Calculates the IR Spectra.** To compare with the well-known map methods, we have calculated the amide I IR spectra of proteins using the electrostatic DFT map developed by Mukamel and co-workers (Figure S14).[11,12] As expected, due to the use of simple empirical polynomial functions, the map method is much faster (10−20 times) than our NN protocol. Roughly speaking, it is at least 5 orders of magnitude faster than the full DFT calculation (Table S1 and Table S2). Unfortunately, the map predictions can only explain the experimental spectra for four (1MBC, 1PPN, 3EST, 5PCA) out of 12 proteins (Figure S14 and Figure S15). Compared with experiment, map results have RMSE (root-mean-square error) values of 1.48 to 4.52 and Spearman rank correlation coefficients of 0.18 to 0.93 (normally a high Spearman coefficient, >0.6, is required for a good theoretical prediction). In contrast, NN results have RMSEs between 1.43 and 2.81 and Spearman rank correlation coefficients between 0.71 and 0.96 (Figure S15 and Table S9). The local electric potential/field used in a map depends on the quality of atomic charges in the chosen force field, and the empirical function of a map trained by a set of protein data set may not fit for other types of proteins. In short, the use of empirical parameters and force-field-dependent electric field values limits the transferability of the map method. We expect that the improvement of map results may require reparametrization the model for

**Table 1. ML Predicts IR Protein Spectra with the Root-Mean-Square Error (RMSE) and High Spearman Rank Correlation ($\rho$) Indicates the Quantitative Agreement with Experiment**[a]

| protein | PDB Code | secondary class | number of atoms | ML time (h) | RMSE | $\rho$ |
|---|---|---|---|---|---|---|
| carbonmonoxymyoglobin | 1MBC | $\alpha$ | 2459 | 4.68 | 2.73 | 0.94 |
| $\lambda$-Immunoglobulin | 1REI | $\beta$ | 3254 | 6.38 | 2.05 | 0.90 |
| penicillin amidohydrolase | 1PNK | $\alpha+\beta$ | 11 708 | 26.41 | 1.43 | 0.91 |
| papain | 1PPN | $\alpha+\beta$ | 3245 | 6.30 | 2.10 | 0.80 |
| dihydropteridine reductase | 1DHR | $\alpha+\beta$ | 3527 | 6.95 | 2.81 | 0.71 |
| subtilisin BPN | 1SBT | $\alpha+\beta$ | 3837 | 8.02 | 1.57 | 0.93 |
| ubiquitin | 1UBQ | $\alpha+\beta$ | 1231 | 3.00 | 2.26 | 0.88 |
| $\alpha$-lactalbumin | 1ALC | $\alpha+\beta$ | 1922 | 4.07 | 1.78 | 0.85 |
| egg white lysozyme | 2LYM | $\alpha+\beta$ | 1960 | 4.13 | 2.15 | 0.83 |
| native elastase | 3EST | $\alpha+\beta$ | 3584 | 7.39 | 2.12 | 0.91 |
| carboxypeptidase A$\alpha$ | 5PCA | $\alpha+\beta$ | 1881 | 8.52 | 2.07 | 0.92 |
| lactate dehydrogenase | 6LDH | $\alpha+\beta$ | 5156 | 9.65 | 2.12 | 0.96 |

[a]Structures of 12 proteins with different sizes were taken from the Protein Data Bank, representing a diverse range of secondary structure contents, i.e., different fractions of $\alpha$-helix and $\beta$-sheet. The IR spectrum of each protein was computed based on 1000 MD configurations. All reported calculation times refer to calculations on eight cores of an Intel(R) Xeon(R) CPU (E5-2683v4 @ 2.1 GHz).


https://dx.doi.org/10.1021/jacs.0c06530
J. Am. Chem. Soc. 2020, 142, 19071−19077

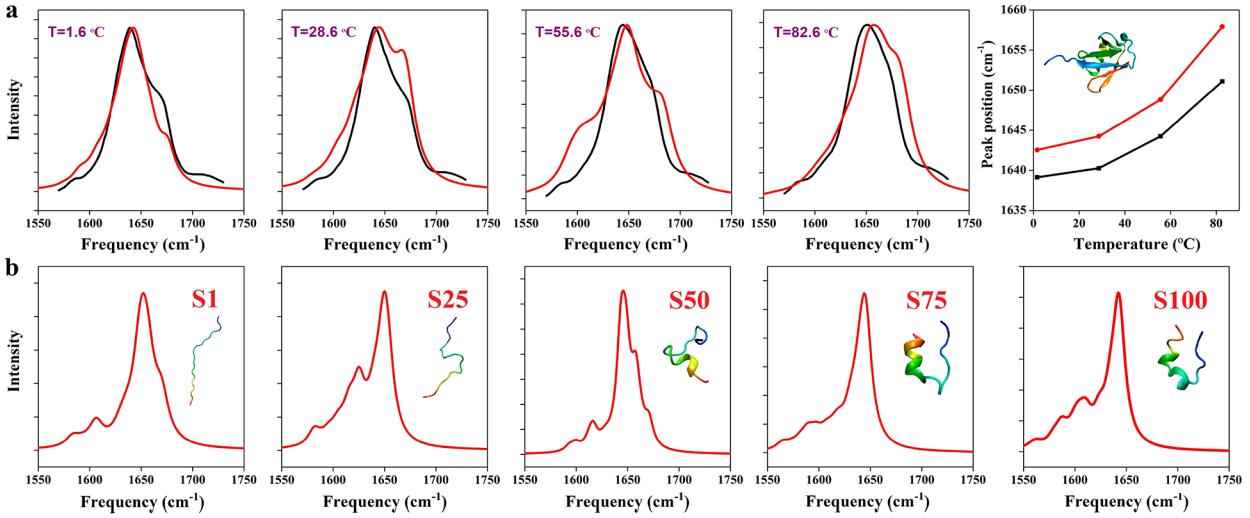



**Figure 5.** (a) From left to right: Simulated (red line) and experimental[53] (black line) IR spectra of ubiquitin at four different temperatures (1.6–82.6 °C) and the temperature variation of the dominant peak position. (b) ML-predicted IR spectra of the Trp-cage protein along its folding path (S1: the original unfolded strand structure; S25: slightly folded but retaining the coil structure; S50: folding rapidly with the emergence of helix elements; S75–S100: stably folded protein with helix structures forming a cage). All spectra are averaged over 100 (1000) MD snapshots for each state of the Trp-cage (ubiquitin).

specific proteins of interest or a more accurate force field for proteins.

**Probing Structure Variations of Ubiquitin with Temperature.** We have examined the ML transferability by simulating the IR spectra of ubiquitin (PDB code: 1UBQ) at different temperatures (1.6, 28.6, 55.6, 82.6 °C) (details in Supporting Information). Ubiquitin is a 76-residue protein that contains both $\alpha$ and $\beta$ secondary structures, which is frequently used as an exemplar of the folding/unfolding process. As the temperature rises in the range of interest, the dominant peak undergoes a blue shift from 1642 to 1657 $cm^{-1}$ (Table S5), accompanied by a broadening of bands and decrease in intensity (Figure 5a). This result is in line with experiment (Figure 5a and Figure S18).[53] The peak shift in ML prediction effectively reflects the effect on temperature, because temperature changes will lead to changes in protein structure, which can be well handled by the ML protocol, demonstrating good transferability of our ML model to varying external environment factors.

**Monitoring the Folding Path of Trp-Cage Protein.** We have verified the variation of amide I IR spectra across a protein folding path. Trp-cage (PDB code: 1L2Y) is a 20-residue miniprotein that has been widely used for studying folding dynamics. A total of 100 000 MD configurations along the Trp-cage folding pathway were retrieved from our previous study.[54] Five stages are taken to reflect the evolution from the unfolded strand (S1), slightly folding but retaining the coil structure (S25), rapid folding stage with a large amount of helical structures (S50), and the helix system folded like a cage (S75 and S100). The ML amide I IR spectra, predicted by averaging over 100 MD snapshots for each state, are depicted in Figure 5b. As the folding process proceeds (S1 → S100), the random coil content decreases followed by an increase in the helix content (Figure 5b and Table S6), leading to a 10 $cm^{-1}$ red shift (S1: 1652 $cm^{-1}$, S25: 1650 $cm^{-1}$, S50: 1646 $cm^{-1}$, S75: 1644 $cm^{-1}$, S100: 1642 $cm^{-1}$) of the dominant peak (Figure 5b and Figure S18 and Table S7). This is consistent with recent time-resolved IR experiments[55] and theoretical simulations.[56]

## SUMMARY

We have reported a machine learning protocol based on *ab initio* data for predicting the amide I IR spectra of a protein from its structure. It shows promise for providing IR spectra characterization of protein dynamics for different proteins under varying conditions, including secondary structure, temperature dependence, and folding status. It significantly boosts the speed of IR spectra simulation compared to conventional quantum chemistry approaches. We are currently improving the transferability of the model by increasing the size of the data set and consider the explicit solvent effect in the ML training to reduce ML model errors. This approach can be expanded to predict optical properties of proteins in other spectral regimes including UV, Raman, and other techniques including sum of frequency generation and multidimensional IR and UV spectroscopies.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.0c06530.

Computational details, molecular dynamics simulations, the machine learning protocol, structure of NMA and GLDP molecules, RMSD, data distribution, optimization steps, hyperparameter optimization, learning curves for the NN training, proteins of interest in this study, IR spectra of 12 proteins calculated by the map, IR spectra of 1REI with different configurations predicted by NN (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, People's Republic of China;* orcid.org/0000-0002-6116-5605; Email: jiangj1@ustc.edu.cn



https://dx.doi.org/10.1021/jacs.0c06530
J. Am. Chem. Soc. 2020, 142, 19071–19077

## Authors

**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, People's Republic of 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, People's Republic of China*

**Jinxiao 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, People's Republic of China*

**Wei Hu** − *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, People's Republic of China;* orcid.org/0000-0002-7467-4783

**Jonathan D. Hirst** − *School of Chemistry, University of Nottingham, Nottingham NG7 2RD, United Kingdom;* orcid.org/0000-0002-2726-0983

**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, People's Republic of China;* orcid.org/0000-0003-0125-9666

**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/jacs.0c06530


## Author Contributions

[⊥]S. Ye, K. Zhong, and J. Zhang contributed equally to this work.

## Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (2018YFA0208603, 2017YFA0303500, 2016YFA0400904) and the National Natural Science Foundation of China (22033007, 21633007, 21790350, 21703221). H.W. thanks the support from Young Taishan Scholar Program of Shandong Province (No. tsqn201909139). S.M. is grateful for the support of NSF (grant CHE-1953045). The numerical calculations have been carried out on the supercomputing system in the Supercomputing Center of the University of Science and Technology of China.

## ■ REFERENCES


(1) Pupeza, I.; Huber, M.; Trubetskov, M.; Schweinberger, W.; Hussain, S. A.; Hofer, C.; Fritsch, K.; Poetzlberger, M.; Vamos, L.; Fill, E. Field-resolved infrared spectroscopy of biological systems. *Nature* **2020**, *577* (7788), 52−59.

(2) Yang, H.; Yang, S.; Kong, J.; Dong, A.; Yu, S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat. Protoc.* **2015**, *10* (3), 382.

(3) Kim, H.; Cho, M. Infrared probes for studying the structure and dynamics of biomolecules. *Chem. Rev.* **2013**, *113* (8), 5817−5847.

(4) Kraack, J. P.; Hamm, P. Surface-sensitive and surface-specific ultrafast two-dimensional vibrational spectroscopy. *Chem. Rev.* **2017**, *117* (16), 10623−10664.

(5) Kratochvil, H. T.; Carr, J. K.; Matulef, K.; Annen, A. W.; Li, H.; Maj, M.; Ostmeyer, J.; Serrano, A. L.; Raghuraman, H.; Moran, S. D. Instantaneous ion configurations in the K+ ion channel selectivity filter revealed by 2D IR spectroscopy. *Science* **2016**, *353* (6303), 1040−1044.

(6) Reppert, M.; Tokmakoff, A. Computational amide I 2D IR spectroscopy as a probe of protein structure and dynamics. *Annu. Rev. Phys. Chem.* **2016**, *67*, 359−386.

(7) Ghosh, A.; Ostrander, J. S.; Zanni, M. T. Watching proteins wiggle: Mapping structures with two-dimensional infrared spectroscopy. *Chem. Rev.* **2017**, *117* (16), 10726−10759.

(8) Lin, Y.-S.; Shorb, J.; Mukherjee, P.; Zanni, M.; Skinner, J. Empirical amide I vibrational frequency map: application to 2D-IR line shapes for isotope-edited membrane peptide bundles. *J. Phys. Chem. B* **2009**, *113* (3), 592−602.

(9) Ham, S.; Kim, J.-H.; Lee, H.; Cho, M. Correlation between electronic and molecular structure distortions and vibrational properties. II. Amide I modes of NMA−n D 2 O complexes. *J. Chem. Phys.* **2003**, *118* (8), 3491−3498.

(10) la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J. Modeling the amide I bands of small peptides. *J. Chem. Phys.* **2006**, *125* (4), No. 044312.

(11) Hayashi, T.; Zhuang, W.; Mukamel, S. Electrostatic DFT map for the complete vibrational amide band of NMA. *J. Phys. Chem. A* **2005**, *109* (43), 9747−9759.

(12) 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.

(13) Kananenka, A. A.; Yao, K.; Corcelli, S. A.; Skinner, J. L. Machine Learning for Vibrational Spectroscopic Maps. *J. Chem. Theory Comput.* **2019**, *15* (12), 6850−6858.

(14) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555* (7698), 604−610.

(15) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559* (7714), 377−381.

(16) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **2018**, *140* (32), 10158−10168.

(17) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.

(18) Ma, S.; Huang, S.-D.; Liu, Z.-P. Dynamic coordination of cations and catalytic selectivity on zinc−chromium oxide alloys during syngas conversion. *Nat. Catal.* **2019**, *2* (8), 671−677.

(19) Hirst, J. D.; Sternberg, M. J. E. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry* **1992**, *31* (32), 7211−7218.

(20) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (24), 11612−11617.

(21) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8* (10), 6924−6935.

(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* (9), 1801367.

(23) Hamm, P.; Zanni, M. *Concepts and Methods of 2D Infrared Spectroscopy*; Cambridge University Press, 2011.

(24) Krimm, S.; Abe, Y. Intermolecular interaction effects in the amide I vibrations of $\beta$ polypeptides. *Proc. Natl. Acad. Sci. U. S. A.* **1972**, *69* (10), 2788−2792.
(25) Hanson-Heine, M. W.; Husseini, F. S.; Hirst, J. D.; Besley, N. A. Simulation of two-dimensional infrared spectroscopy of peptides using localized normal modes. *J. Chem. Theory Comput.* **2016**, *12* (4), 1905−1918.
(26) Wang, L.; Middleton, C. T.; Zanni, M. T.; Skinner, J. L. Development and validation of transferable amide I vibrational frequency maps for peptides. *J. Phys. Chem. B* **2011**, *115* (13), 3713−3724.
(27) Torii, H.; Tasumi, M. Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di-and tripeptides and considerations on the conformation of the extended helix. *J. Raman Spectrosc.* **1998**, *29* (1), 81−86.
(28) Hayashi, T.; Mukamel, S. Vibrational− Exciton couplings for the amide I, II, III, and a modes of peptides. *J. Phys. Chem. B* **2007**, *111* (37), 11032−11046.
(29) Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. cp2k: atomistic simulations of condensed matter systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (1), 15−25.
(30) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H. *Gaussian 16*; Gaussian, Inc.: Wallingford, CT, 2016.
(31) Jacob, C. R.; Reiher, M. Localizing normal modes in large molecules. *J. Chem. Phys.* **2009**, *130* (8), 084106.
(32) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. In *Tensorflow: A system for large-scale machine learning*; 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016; pp 265−283.
(33) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw* **1989**, *2* (5), 359−366.
(34) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15* (9), 095003.
(35) 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.
(36) Kingma, D. P.; Ba, J., Adam: A method for stochastic optimization. *arXiv* preprint arXiv:1412.6980, 2014.
(37) Ng, A. Y. In *Feature selection, L1 vs. L2 regularization, and rotational invariance*; Proceedings of the 21st international conference on Machine learning, 2004; p 78.
(38) Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach Learn Res.* **2012**, *13* (1), 281−305.
(39) 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.
(40) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9* (8), 3404−3419.
(41) Zhuang, W.; Abramavicius, D.; Hayashi, T.; Mukamel, S. Simulation protocols for coherent femtosecond vibrational spectra of peptides. *J. Phys. Chem. B* **2006**, *110* (7), 3362−3374.
(42) http://dcaiku.com:12880/platform/first.
(43) Jiang, J.; Ye, S. Machine learning protocol code [Data set]. Zenodo; 2020; DOI: 10.5281/zenodo.4106438.
(44) Jiang, J.; Ye, S. Machine learning simulation data [Data set]. Zenodo; 2020; DOI: 10.5281/zenodo.4106543.
(45) Besley, N. A.; Hirst, J. D. Theoretical Studies toward Quantitative Protein Circular Dichroism Calculations. *J. Am. Chem. Soc.* **1999**, *121* (41), 9636−9644.
(46) Watson, T. M.; Hirst, J. D. Calculating vibrational frequencies of amides: From formamide to concanavalin A. *Phys. Chem. Chem. Phys.* **2004**, *6* (5), 998−1005.
(47) Husseini, F. S.; Robinson, D.; Hunt, N. T.; Parker, A. W.; Hirst, J. D. Computing infrared spectra of proteins using the exciton model. *J. Comput. Chem.* **2017**, *38* (16), 1362−1375.
(48) Karjalainen, E.-L.; Ersmark, T.; Barth, A. Optimization of model parameters for describing the amide I spectrum of a large set of proteins. *J. Phys. Chem. B* **2012**, *116* (16), 4831−4842.
(49) Baumann, K.; Clerc, J. Computer-assisted IR spectra prediction—linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348* (1−3), 327−343.
(50) 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.
(51) Hirst, J. D.; Colella, K.; Gilbert, A. T. Electronic circular dichroism of proteins from first-principles calculations. *J. Phys. Chem. B* **2003**, *107* (42), 11813−11819.
(52) DeFlores, L. P.; Ganim, Z.; Nicodemus, R. A.; Tokmakoff, A. Amide I′− II′ 2D IR spectroscopy provides enhanced protein secondary structural sensitivity. *J. Am. Chem. Soc.* **2009**, *131* (9), 3385−3391.
(53) Waegele, M. M.; Gai, F. Power-law dependence of the melting temperature of ubiquitin on the volume fraction of macromolecular crowders. *J. Chem. Phys.* **2011**, *134* (9), 03B605.
(54) 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.
(55) Culik, R. M.; Serrano, A. L.; Bunagan, M. R.; Gai, F. Achieving Secondary Structural Resolution in Kinetic Measurements of Protein Folding: A Case Study of the Folding Mechanism of Trp-cage. *Angew. Chem., Int. Ed.* **2011**, *50* (46), 10884−10887.
(56) Lai, Z.; Preketes, N. K.; Mukamel, S.; Wang, J. Monitoring the folding of Trp-cage peptide by two-dimensional infrared (2dir) spectroscopy. *J. Phys. Chem. B* **2013**, *117* (16), 4661−4669.


https://dx.doi.org/10.1021/jacs.0c06530
J. Am. Chem. Soc. 2020, 142, 19071−19077

# Supporting Information

## A Machine Learning Protocol for Predicting Protein Infrared Spectra

Sheng Ye,[1,⊥] Kai Zhong,[1,⊥] Jinxiao Zhang,[1,⊥] Wei Hu,[1] Jonathan D. Hirst,[2] Guozhen Zhang,[1] Shaul Mukamel,[3] Jun Jiang[1,*]

[1] 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, P. R. China

[2] School of Chemistry, University of Nottingham, Nottingham, NG7 2RD

[3] Departments of Chemistry, and physics & astronomy, University of California, Irvine, CA 92697, USA

[⊥]These authors contribute equally to this work

*Corresponding author. E-mail: jiangj1@ustc.edu.cn

# Table of Contents

# Computational Details

## Molecular Dynamics Simulations

**NMA:** In order to sample relevant configurations, we have run *ab initio* molecular dynamics (AIMD) calculations seven times with different initial conformations (We obtained a total of seven different initial structures by retrieving one snapshot per 2 ps from a 14-ps trajectory file of AIMD simulated at 500 K in the NVT ensemble) (Figure S2h), which resulted in 9660 configurations used for machine learning training. Each simulation was performed using the CP2K[1] program with the QUICKSTEP method and M06L[2] meta-GGA with D3(0) vdW corrections.[3] Here M06L-D3(0) provides a significantly high density of water 1.25 g/cm$^3$ compared to any other methods.[4] So, it offers an extreme nature of hydration around the NMA molecules, which allows us to construct a diverse and robust dataset for frequency prediction. Meanwhile, accurate frequency calculations were done at a higher level of B3LYP. The triple-ζ valence basis set plus two sets of polarization functions (TZV2P) and Goedecker-Teter-Hutter PBE pseudopotentials was used for all C, N, H and O atoms. All the simulations were performed for one NMA molecule in a solution of 100 $D_2O$ (so all hydrogen atoms in NMA were deuterated) at 300 K in the NVT ensemble with the CSVR thermostat.[5] A cutoff of 10 Å was employed for the vdW interactions. The time step was 0.5 fs and generated trajectories of 241.5 ps, from which 9660 configurations were extracted at 25 fs intervals to avoid overly correlated configurations.

For each configuration, we extracted the NMA molecule and surrounding water molecules within a 5Å radius for the Hessian calculations. The optimization of the NMA molecule was performed in multiple stages to keep the relative orientation of the NMA molecule with the solvent molecules intact. The relative orientation of water molecules was fixed in space. Figure S4 shows various subgroups for the NMA structure. The optimization with B3LYP/cc-pVDZ. was performed for part 1 ($CH_3$-CO-), then for part 2 (-NH-$CH_3$), followed by part 3 (all atoms except methyl hydrogens) with Gaussian16 package.[6] Finally, the Hessian calculations with B3LYP/cc-pVDZ were carried out on the whole structure (an NMA molecule and surrounding water molecules within a 5 Å radius) with all deuterium atoms in NMA frozen except for D in ND bonds. The harmonic vibrational frequencies were scaled by 0.97.

**Proteins:** Molecular dynamics simulations were performed for 12 proteins with the GROMACS package[7] and the OPLS-AA force fields.[8] For each protein, periodic boundary conditions were imposed on a central cell containing one protein and [8571– 38707, depending on the protein] TIP3P water molecules. Electrostatic interactions were treated by the Particle mesh Ewald method and Coulomb interactions were truncated at 12.0 Å. Energy minimization was performed for 50,000 cycles for each protein. Thereafter an equilibration process in NPT ensemble with an integration timestep of 2 fs ran for 0.5 ns.[8] Production dynamics were performed for a period of 2 ns in the NPT ensemble at 300K while maintaining pressure at 1 atm. 1000 configurations were extracted with a 2 ps interval for calculating the IR spectra. The same MD parameters were used for Ubiquitin in conjunction with a set of different temperatures (1.6℃,28.6℃,55.6℃, 82.6℃)

In short, 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.[9-11]

**The Spearman rank correlation coefficients** ($\rho$) was computed with: $\rho = 1 - \frac{6\sum_i d_i^2}{n \cdot (n^2 - 1)}$, where $d_i$ is the difference between the ranks of $x_i$ (absorption intensities of experiment spectra) and $y_i$ (absorption intensities of predicted spectra) in their respective data set and $n$ the number of elements in each vector.

We computed ML-derived spectra of 1000 configurations for the purpose of comparison with the experiment. To predict the amide I IR spectra of 12 proteins, we performed a 2-ns MD simulation and obtained 1000 snapshots frames for each protein. For each frame, the amide I IR spectra was predicted by our ML protocol. Then we averaged data of all snapshots so as to capture the fluctuating dynamics for each protein. The obtained spectra characterize the averaged structure whose fine detail due to fluctuation has been cancelled out. Therefore, they are smooth and featureless.

## The Machine Learning Protocol

**Neural network architecture:** The NN consists of one input layer, three hidden layers and one output layer. The number of hidden layer neurons is 32, 64 and 128. The Rectified Linear Unit activation function[12] is used for each hidden layer. We added L2 regularization[13] to the neural network, so as to prevent overfitting. The training were subjected to a supervised training scheme using a back-propagation algorithm implemented in TensorFlow.[14]

**Feature engineering:** To avoid the use of raw variables with different range of values which may undermine the training process of NN, we normalized the input features i to reduce the dimensional inconsistency before input to the NN. The data was transformed with $x' = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$, where $x_i$ are input data, $x'$ are normalized data, and $x_{min}$ and $x_{max}$ are minimum and maximum values of the input data, respectively.

**Training and testing set and cross-validation.:** The accuracy and robustness of the ML prediction was verified by the cross-validation technique.[15] All sets of data were randomly and evenly distributed into 10 bins in this procedure. Each bin was used as a test set while the remaining nine bins as training set.

Pearson coefficient ($r$), mean absolute deviation ($MAE$) and cross-validation procedure were chosen to estimate the accuracy and robustness of NN model. The mean absolute deviation ($MAE$) was

computed with $MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|$, where $n$ is the number of samples, $A_t$ is the actual value and $F_t$ is the predicted value.

The mean absolute percentage error (*MAPE*) was computed with $MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$, where $A_t$ is the actual value and $F_t$ is the predicted value.

The root mean square error (*RMSE*) was computed with $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2}$, where $A_t$ is the actual value and $F_t$ is the predicted value.

**Learning curves:** We have calculated the learning curves of whole ML process in this work. The *MAE* converges as the size of sample in ML training exceeds 6000 for frequency and transition dipole moment and 4000 for vibrational coupling constants (Figure S7).

## Structure of NMA and GLDP molecules

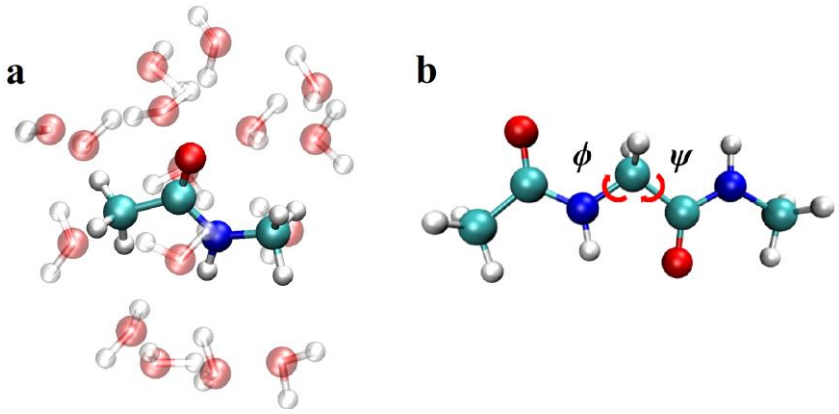



**Figure S1.** (a) The NMA molecule and surrounding solvent molecules within a 5 Å radius optimized at the B3LYP/cc-pVDZ level. (b) The GLDP molecule gas-phase geometry optimized at the B3LYP/cc-pVDZ level.

## The Root Mean Square Deviation (RMSD)

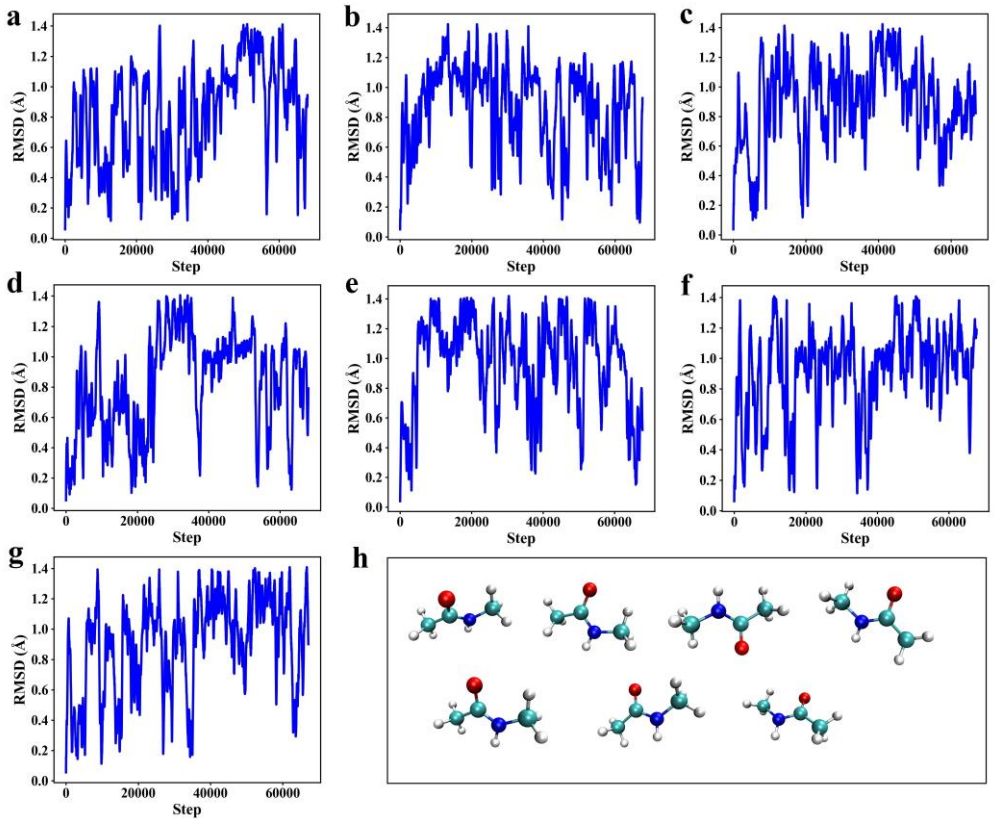



**Figure S2.** (a-g) The Root Mean Square Deviation (RMSD) of the AIMD trajectories and (h) the seven different initial conformations of NMA for the AIMD simulations.

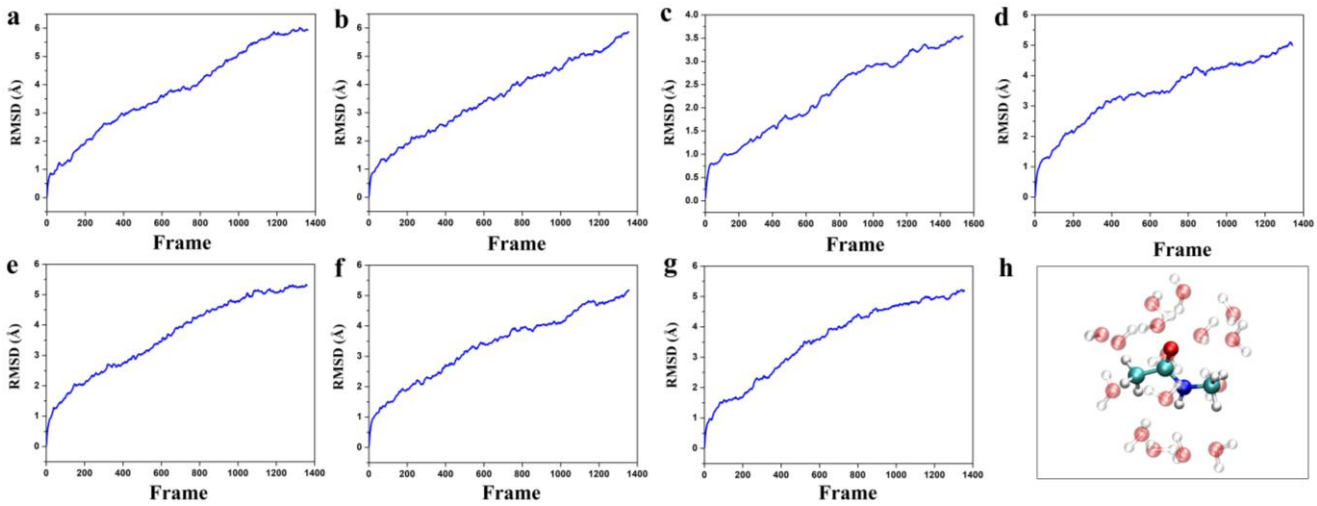



**Figure S3.** (a-g) The Root Mean Square Deviation (RMSD) of extracted NMA molecule and surrounding water molecules within a 5Å radius used to generate ML training data. (a:1360 frames; b:1354 frames; c:1534 frames; d:1342 frames; e:1359 frames; d:1356 frames; e: 1355 frames;) (h) NMA molecule and surrounding water molecules within a 5Å radius.

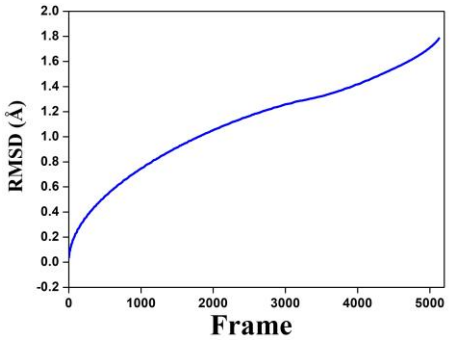



**Figure S4.** The Root Mean Square Deviation (RMSD) of 5128 frames GLDP molecular.

## Data distribution

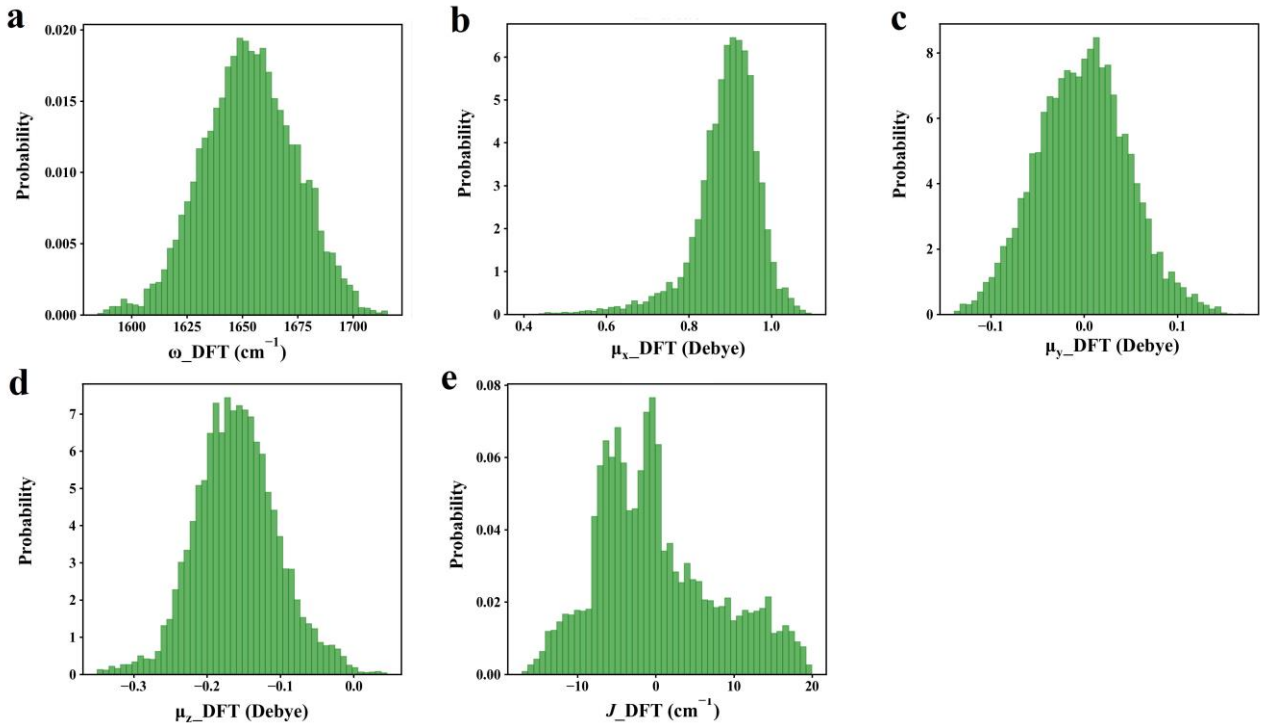



**Figure S5.** (a) Distribution of the frequency of the NMA. (b-d) Distribution of the amide I vibrational transition dipole moment in the x, y, and z direction. (e) Distribution of the coupling constants of the GLDP.

## Descriptors

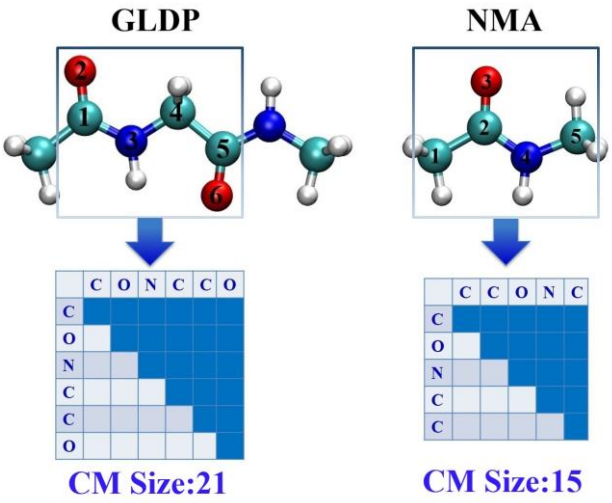
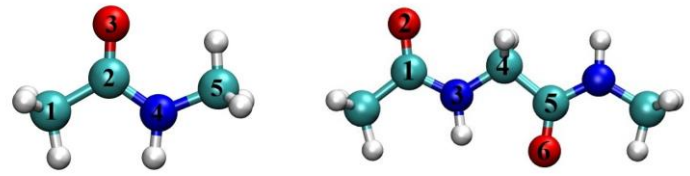



**Figure S6.** Molecular descriptors (ground state Coulomb Matrix (CM) excluding hydrogens and solvent molecule since the uncertainty of positions of hydrogen atoms and the inclusion of solvent molecules would significantly increase the complexity of NN training.) for machine learning.



Bond: 1-2; 2-3; 2-4; 4-5
Angle: 1-2-3; 1-2-4; 3-2-4; 2-4-5;
Dihedral: 4-2-1-3; 1-2-4-5;

Bond: 1-2; 1-3; 3-4; 4-5; 5-6
Angle: 1-1-3; 1-3-4; 3-4-5; 4-5-6;
Dihedral: 2-1-3-4; 1-3-4-5; 3-4-5-6;

| MAE | Internal coordinates | Coulomb matrix |
|---|---|---|
| $\omega$ | 1.80 | 1.71 |
| $\mu_x$ | 0.042 | 0.024 |
| $\mu_y$ | 0.033 | 0.021 |
| $\mu_z$ | 0.062 | 0.018 |
| $J$ | 0.737 | 0.318 |

**Figure S7.** Comparison of mean absolute deviation (*MAE*) of machine learning results based on internal coordinates and Coulomb matrix descriptors.

## Optimization steps

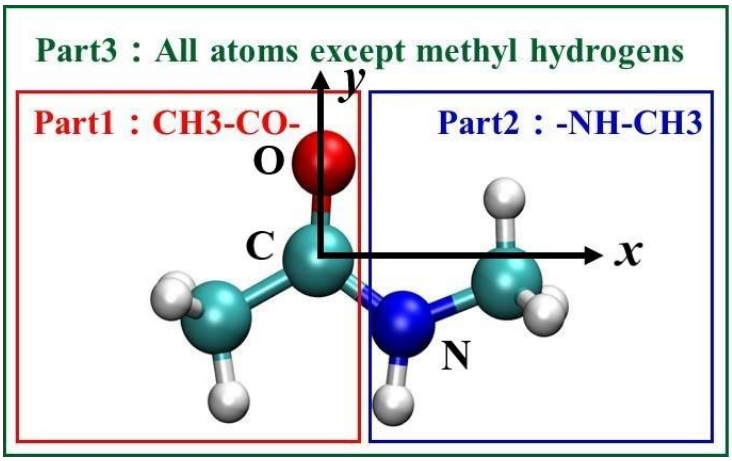



**Figure S8.** Optimization steps of NMA molecule.

.

## Hyperparameter optimization

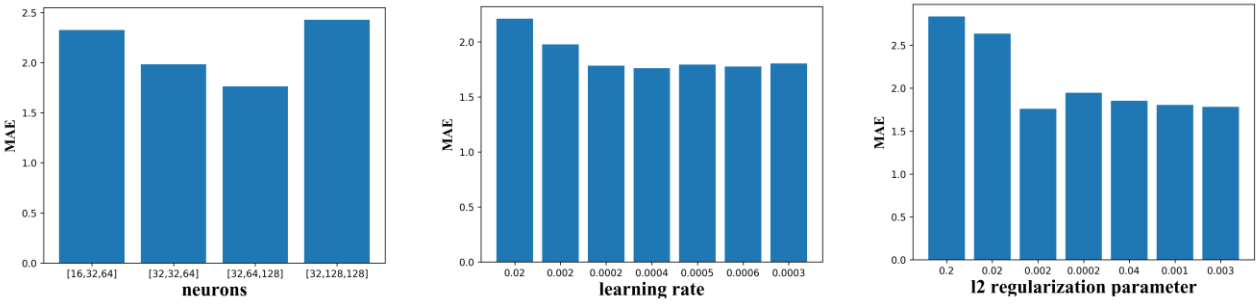



**Figure S9.** Hyperparameters optimized by using random search algorithm in tensorflow (including neurons for hidden layers, learning rate and l2 regularization parameter) to create a reasonable ML protocol.

## The Learning Curves for the NN training

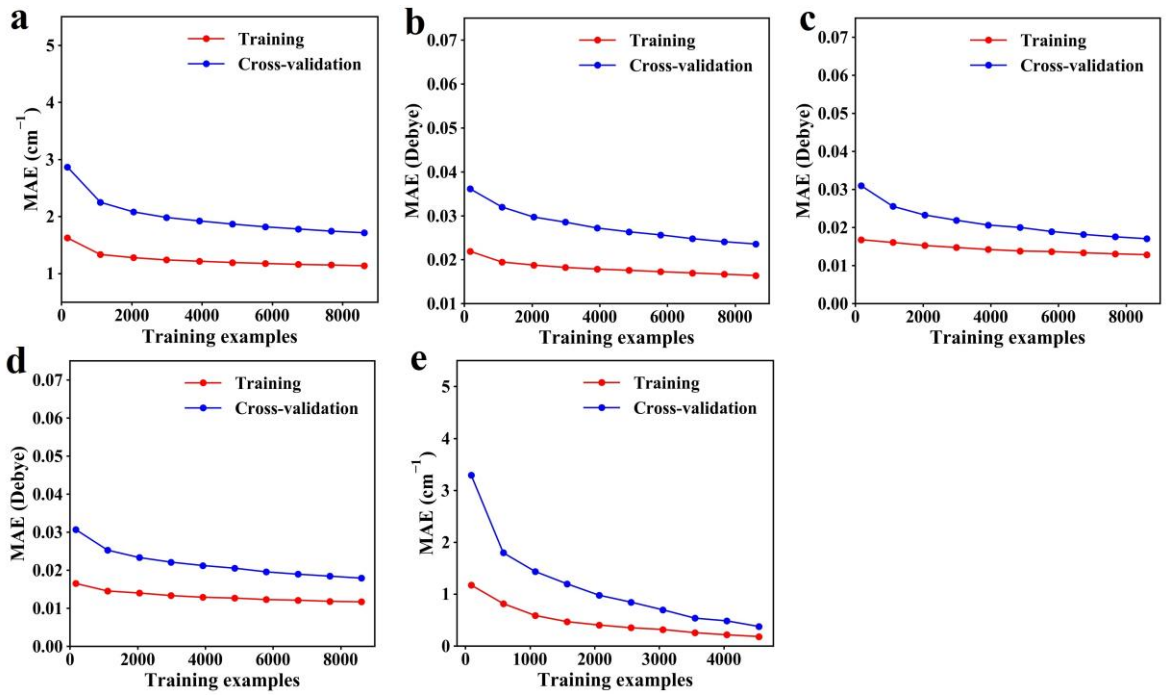



**Figure S10.** (a) The learning curves for predicting the frequencies of NMA. (b-d) The learning curves for predicting the transition dipole moment of NMA in the *x*, *y*, *z* direction. (e) The learning curves for predicting the coupling constants of GLDP.

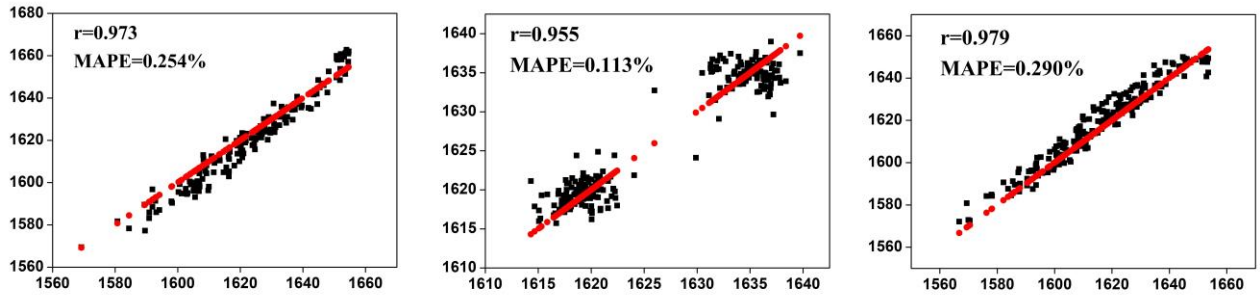



**Figure S11.** Comparison of DFT (B3LYP/cc-pVDZ) and ML on vibrational model Hamiltonian diagonal parameters (MAPE: mean absolute percentage error). From left to right: Hamiltonian diagonal parameters of 1MBC (DFT vs ML); Hamiltonian diagonal parameters of 1REI (DFT vs ML); Hamiltonian diagonal parameters of 1PPN (DFT vs ML)).

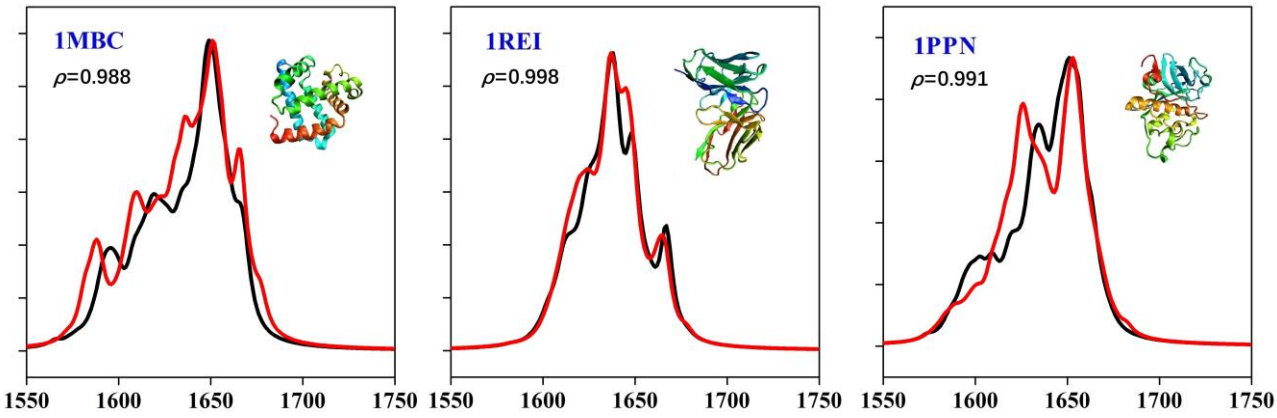



**Figure S12.** Comparison of DFT (B3LYP/cc-pVDZ； black line) and ML (red line) simulated amide I IR spectra of one snapshot based on Frenkel exciton model ($\rho$: Spearman rank correlation).

## Proteins of interest in this study

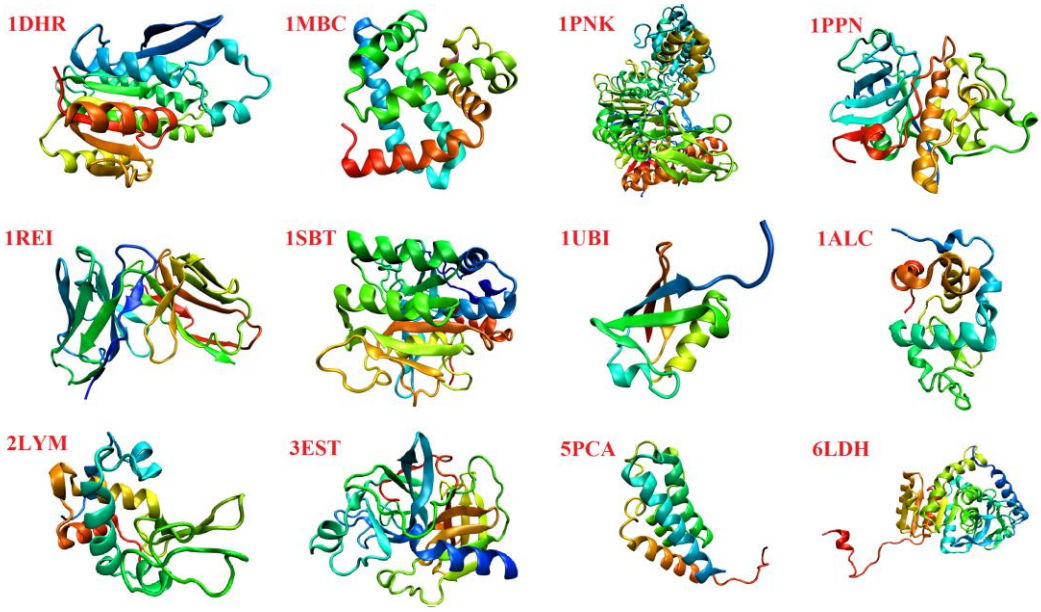



**Figure S13.** Proteins (labeled by their respective PDB code) of interest in this study.

## IR spectra of 12 proteins calculated by map

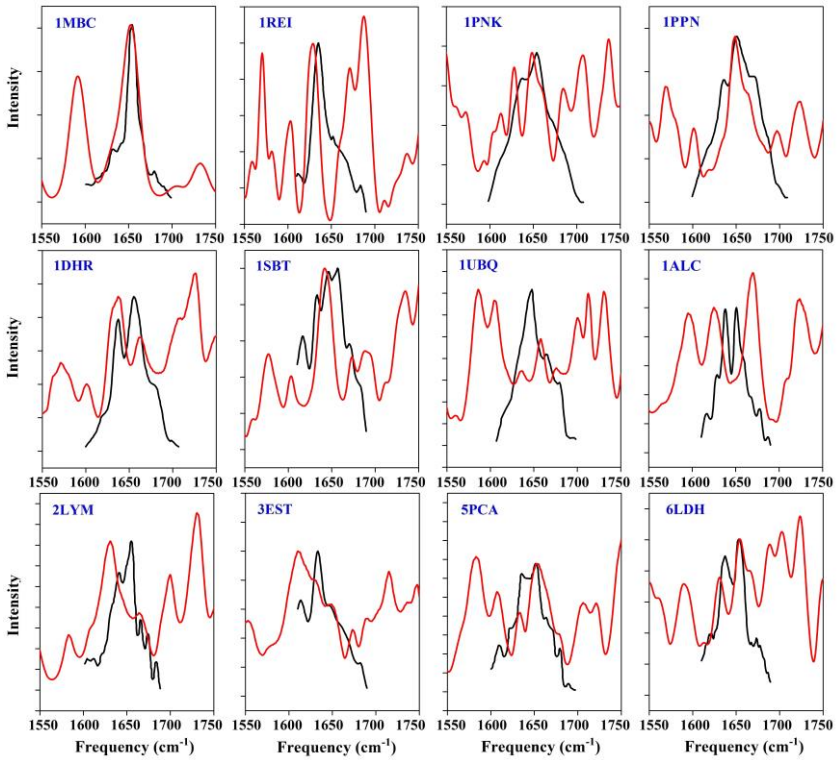



**Figure S14.** Amide I IR spectra of 12 proteins calculated by map method developed by Mukamel and co-works (after considering the deuteration effect of amide I mode).

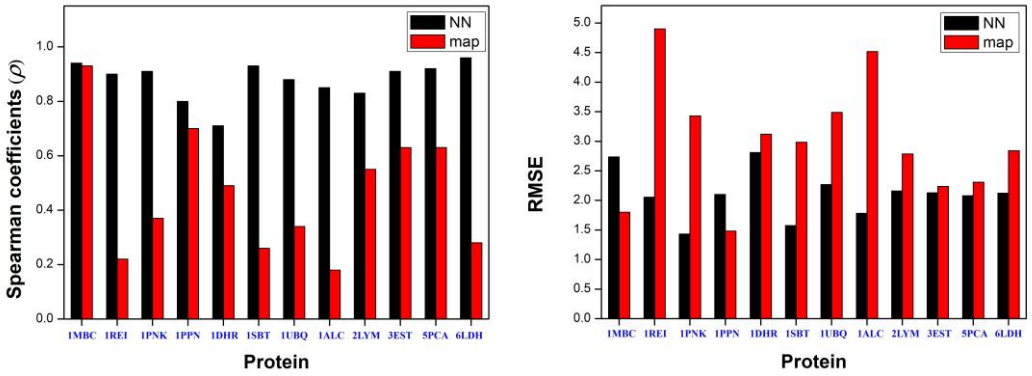



**Figure S15.** The spearman rank correlation coefficients ($\rho$), root mean square error (RMSE) of calculated amide I IR spectra by NN and map method (after considering the deuteration effect of amide I mode) using experimental spectra as references.

## IR spectra of 1REI with different configurations predicted by NN

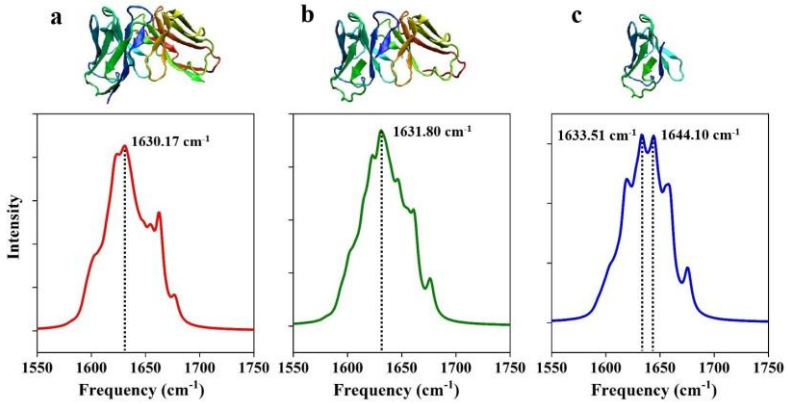



**Figure S16.** Amide I IR spectra of 1REI (red line (a)), and 1REI missing part of the skeleton (green (b) and blue (c) line). All the Lorentzian bandwidths were 4 $cm^{-1}$.

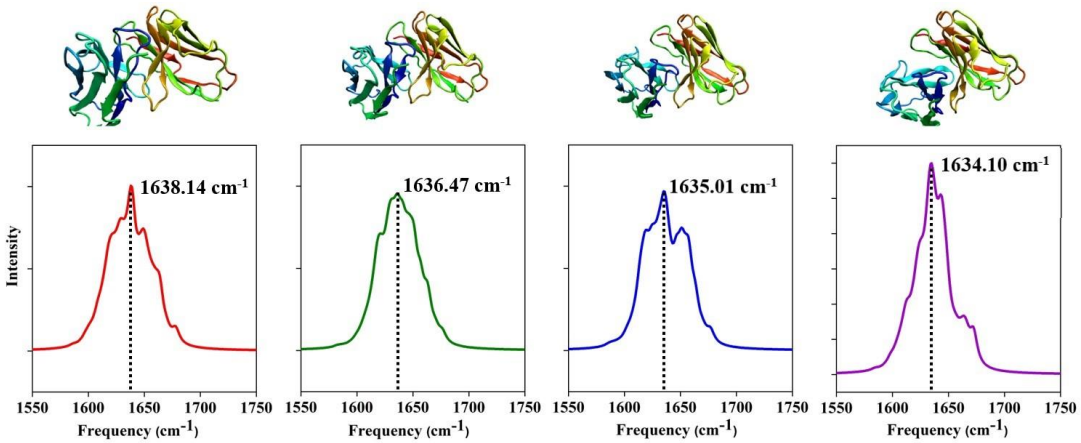



**Figure S17.** Amide I IR spectra of 1REI at different moments (t=500, 1000, 1500, and 2000 ps) predicted by NN, represented by red, green, blue, and purple lines, respectively. All the Lorentzian bandwidths were 4 $cm^{-1}$.

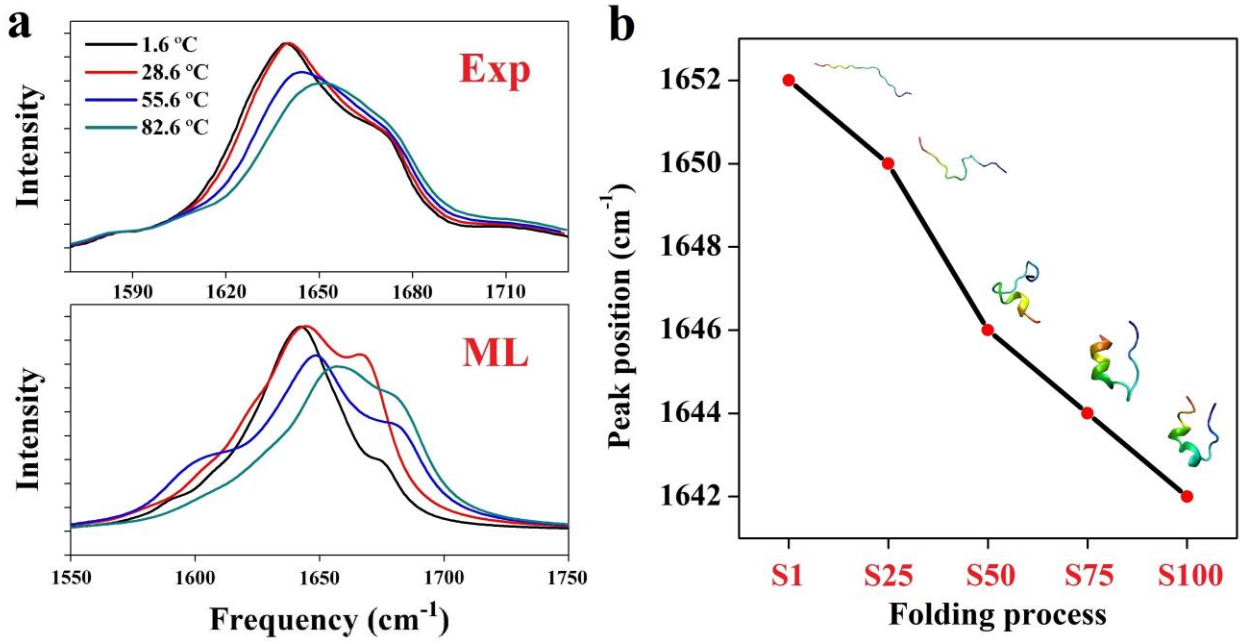



**Figure S18.** Machine learning predicted amide I vibration signals of (a) Ubiquitin revealing temperature transferability of the ML protocol and (b) the variation of the main peak position of the Trp-cage protein along with its folding process.

## Supporting information for Tables

**Table S1.** Time required for map method to calculate the amide I IR protein spectra with the root mean square error (RMSE) and Spearman rank correlation ($\rho$) compared to experiment. All reported times refer to calculations on an 8core of an Intel(R) Xeon(R) CPU (E5-2683v4 @ 2.1GHz).

| Protein | PDB Code | Secondary Class | Atom Size | RMSE | $\rho$ | Map(s) |
|---|---|---|---|---|---|---|
| Carbonmonoxymyoglobin | 1MBC | $\alpha$ | 2459 | 2.535 | 0.70 | 1.35 |
| λ-Immunoglobulin | 1REI | $\beta$ | 3254 | 3.384 | 0.24 | 1.91 |
| Penicillin Amidohydrolase | 1PNK | $\alpha+\beta$ | 11708 | 2.930 | 0.42 | 4.75 |
| Papain | 1PPN | $\alpha+\beta$ | 3245 | 2.710 | 0.54 | 1.83 |
| Dihydropteridine Reductase | 1DHR | $\alpha+\beta$ | 3527 | 2.540 | 0.65 | 2.23 |
| Subtilisin BPN | 1SBT | $\alpha+\beta$ | 3837 | 2.798 | 0.41 | 1.77 |
| Ubiquitin | 1UBQ | $\alpha+\beta$ | 1231 | 3.680 | 0.54 | 1.10 |
| $\alpha$-Lactalbumin | 1ALC | $\alpha+\beta$ | 1922 | 4.824 | 0.30 | 1.42 |
| Egg White Lysozyme | 2LYM | $\alpha+\beta$ | 1960 | 2.320 | 0.86 | 1.39 |
| Native Elastase | 3EST | $\alpha+\beta$ | 3584 | 2.348 | 0.68 | 2.17 |
| Carboxypeptidase A$\alpha$ | 5PCA | $\alpha+\beta$ | 1881 | 3.773 | 0.08 | 2.33 |
| Lactate Dehydrogenase | 6LDH | $\alpha+\beta$ | 5156 | 3.216 | 0.01 | 2.28 |

**Table S2.** Comparison of the time required for computing the amide I IR spectra (from the initial structure to the final amide I IR spectra) of various proteins (one snapshot) by DFT (B3LYP/cc-pVDZ) and ML based on Frenkel exciton model. All reported times refer to calculations on an 8core of an Intel(R) Xeon(R) CPU (E5-2683v4 @ 2.1GHz).

| Protein | PDB Code | Class | DFT (s) | ML (s) |
|---|---|---|---|---|
| Carbonmonoxymyoglobin | 1MBC | $\alpha$ | 291720 | 16.85 |
| λ-Immunoglobulin | 1REI | $\beta$ | 406920 | 22.95 |
| Penicillin Amidohydrolase | 1PNK | $\alpha+\beta$ | 1436040 | 95.06 |
| Papain | 1PPN | $\alpha+\beta$ | 405000 | 22.68 |
| Dihydropteridine Reductase | 1DHR | $\alpha+\beta$ | 451080 | 25.03 |
| Subtilisin BPN | 1SBT | $\alpha+\beta$ | 525960 | 28.86 |
| Ubiquitin | 1UBQ | $\alpha+\beta$ | 143880 | 10.80 |
| α-Lactalbumin | 1ALC | $\alpha+\beta$ | 232200 | 14.66 |
| Egg White Lysozyme | 2LYM | $\alpha+\beta$ | 245640 | 14.86 |
| Native Elastase | 3EST | $\alpha+\beta$ | 458760 | 26.59 |
| Carboxypeptidase A$\alpha$ | 5PCA | $\alpha+\beta$ | 587400 | 30.67 |
| Lactate Dehydrogenase | 6LDH | $\alpha+\beta$ | 1220040 | 34.74 |

**Table S3.** Lorentzian bandwidth ($cm^{-1}$) used for IR spectra of proteins (individual crystal structure).

| Protein | PDB Code | Class | Lorentzian bandwidth ($cm^{-1}$) |
|---|---|---|---|
| Carbonmonoxymyoglobin | 1MBC | $\alpha$ | 4 |
| λ-Immunoglobulin | 1REI | $\beta$ | 5 |
| Penicillin Amidohydrolase | 1PNK | $\alpha+\beta$ | 6 |
| Papain | 1PPN | $\alpha+\beta$ | 6 |
| Dihydropteridine Reductase | 1DHR | $\alpha+\beta$ | 2 |
| Subtilisin BPN | 1SBT | $\alpha+\beta$ | 5 |
| Ubiquitin | 1UBQ | $\alpha+\beta$ | 6 |
| $\alpha$-Lactalbumin | 1ALC | $\alpha+\beta$ | 2 |
| Egg White Lysozyme | 2LYM | $\alpha+\beta$ | 2 |
| Native Elastase | 3EST | $\alpha+\beta$ | 5 |
| Carboxypeptidase Aα | 5PCA | $\alpha+\beta$ | 4 |
| Lactate Dehydrogenase | 6LDH | $\alpha+\beta$ | 4 |

**Table S4.** The secondary structure content (computed using Stride[16]) of the λ-Immunoglobulin (1REI) in different states.

| Content / State | $\beta$-strands | $\beta$-turns | $\alpha$-helix | $3_{10}$-helices | Coil |
|---|---|---|---|---|---|
| a | 50.00% | 36.45% | 0% | 0% | 9.81% |
| b | 47.55% | 37.01% | 0% | 0% | 9.14% |
| c | 43.55% | 40.13% | 0% | 0% | 14.01% |
| t=500 (ps) | 48.30% | 33.70% | 0% | 0% | 13.50% |
| t=1000 (ps) | 49.80% | 32.60% | 0% | 0% | 13.10% |
| t=1500 (ps) | 50.70% | 35.30% | 0% | 0% | 9.30% |
| t=2000 (ps) | 51.50% | 32.40% | 0% | 0% | 12.10% |

**Table S5.** ML-predicted and experimental[17] main peak of Ubiquitin at four different temperatures (1.6 ° C ~ 82.6 ° C).

| Temperature (℃) | Main peak-Exp ($cm^{-1}$) | Main peak-ML ($cm^{-1}$) |
|---|---|---|
| 1.6 | 1639.14 | 1642.54 |
| 28.6 | 1640.26 | 1644.25 |
| 55.6 | 1644.25 | 1648.82 |
| 82.6 | 1651.12 | 1657.92 |

**Table S6.** The secondary structure content (computed using Stride[16]) of the Trp-cage protein at various stages in the simulated folding process.

| Content / State | $\beta$-strands | $\beta$-turns | $\alpha$-helix | $3_{10}$-helices | Coil |
|---|---|---|---|---|---|
| S0 | 0% | 0.9% | 0% | 0% | 99.1% |
| S25 | 0% | 25.0% | 0% | 0% | 73.0% |
| S50 | 0% | 38.5% | 8.2% | 3.7% | 49.7% |
| S75 | 0% | 29.9% | 19.7% | 1.6% | 48.9% |
| S100 | 0% | 16.6% | 37.1% | 8.7% | 37.6% |

**Table S7.** The ML-predicted main peak of the Trp-cage protein along its folding path (S1：the original unfolded strand structure; S25: slightly folded but retaining the coil structure; S50: folding rapidly with the emergence of helix elements; S75-S100: stably folded protein with helix structures forming a cage.)

| State | Main peak-ML (cm-1) |
|---|---|
| S0 | 1652 |
| S25 | 1650 |
| S50 | 1646 |
| S75 | 1644 |
| S100 | 1642 |

**Table S8.** Time required for the Neural Network to train vibrational frequency ($\omega_i$), transition dipole ($\vec{\mu_i}(x,y,z)$) and neighboring coupling ($J_{ij}$). All reported times refer to train on an 8core of an Intel(R) Xeon(R) CPU (E5-2683v4 @ 2.1GHz).

| Task for NN training | CPU Time (h) |
|---|---|
| $\omega_i$ | 0.91 |
| $\overrightarrow{\mu_x}$ | 2.23 |
| $\overrightarrow{\mu_y}$ | 1.89 |
| $\overrightarrow{\mu_z}$ | 2.62 |
| $J_{ij}$ | 0.86 |

**Table S9.** Comparison of the spearman rank correlation ($\rho$), root mean square error (RMSE) of the calculated amide I IR spectra by NN and map method using experimental spectra as references.

| Protein | PDB Code | Secondary Class | RMSE (NN) | RMSE (Map) | $\rho$ (NN) | $\rho$ (Map) |
|---|---|---|---|---|---|---|
| Carbonmonoxymyoglobin | 1MBC | $\alpha$ | 2.737 | 1.799 | 0.94 | 0.93 |
| λ-Immunoglobulin | 1REI | $\beta$ | 2.054 | 4.900 | 0.90 | 0.22 |
| Penicillin Amidohydrolase | 1PNK | $\alpha+\beta$ | 1.430 | 3.430 | 0.91 | 0.37 |
| Papain | 1PPN | $\alpha+\beta$ | 2.100 | 1.480 | 0.80 | 0.70 |
| Dihydropteridine Reductase | 1DHR | $\alpha+\beta$ | 2.810 | 3.120 | 0.71 | 0.49 |
| Subtilisin BPN | 1SBT | $\alpha+\beta$ | 1.572 | 2.984 | 0.93 | 0.26 |
| Ubiquitin | 1UBQ | $\alpha+\beta$ | 2.267 | 3.489 | 0.88 | 0.34 |
| $\alpha$-Lactalbumin | 1ALC | $\alpha+\beta$ | 1.780 | 4.517 | 0.85 | 0.18 |
| Egg White Lysozyme | 2LYM | $\alpha+\beta$ | 2.157 | 2.785 | 0.83 | 0.55 |
| Native Elastase | 3EST | $\alpha+\beta$ | 2.126 | 2.236 | 0.91 | 0.63 |
| Carboxypeptidase A$\alpha$ | 5PCA | $\alpha+\beta$ | 2.077 | 2.310 | 0.92 | 0.63 |
| Lactate Dehydrogenase | 6LDH | $\alpha+\beta$ | 2.122 | 2.842 | 0.96 | 0.28 |

**References**


1. Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J., cp2k: atomistic simulations of condensed matter systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014,** *4* (1), 15-25.
2. Zhao, Y.; Truhlar, D. G., A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006,** *125* (19), 194101.
3. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010,** *132* (15), 154104.
4. Ohto, T.; Dodia, M.; Xu, J.; Imoto, S.; Tang, F.; Zysk, F.; Kühne, T. D.; Shigeta, Y.; Bonn, M.; Wu, X.; Nagata, Y., Accessing the Accuracy of Density Functional Theory through Structure and Dynamics of the Water–Air Interface. *J. Phys. Chem. Lett.* **2019,** *10* (17), 4914-4919.
5. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007,** *126* (1), 014101.
6. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H., Gaussian 16. Gaussian, Inc. Wallingford, CT: 2016.
7. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005,** *26* (16), 1701-1718.
8. Husseini, F. S.; Robinson, D.; Hunt, N. T.; Parker, A. W.; Hirst, J. D., Computing infrared spectra of proteins using the exciton model. *J. Comput. Chem.* **2017,** *38* (16), 1362-1375.
9. 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.
10. Hirst, J. D.; Colella, K.; Gilbert, A. T., Electronic circular dichroism of proteins from first-principles calculations. *The Journal of Physical Chemistry B* **2003,** *107* (42), 11813-11819.
11. Baumann, K.; Clerc, J., Computer-assisted IR spectra prediction—linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997,** *348* (1-3), 327-343.
12. 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；p3, 2013; p 3.
13. Ng, A. Y. In *Feature selection, L1 vs. L2 regularization, and rotational invariance*, Proceedings of the 21st international conference on Machine learning, 2004; p 78.
14. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. In *Tensorflow: A system for large-scale machine learning*, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016; pp 265-283.
15. Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R., Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013,** *9* (8), 3404-3419.
16. Heinig, M.; Frishman, D., STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **2004,** *32* (suppl_2), W500-W502.

17. Waegele, M. M.; Gai, F., Power-law dependence of the melting temperature of ubiquitin on the volume fraction of macromolecular crowders. *J. Chem. Phys.* **2011,** *134* (9), 03B605.