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.

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 (ML) 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. Its basic philosophy is to compute (or predict) vibrational modes by using empirical polynomial function 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 parameterization 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 datasets, 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 modelling 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 non-linear problems, and offers a transferrable 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 MD (AIMD) to compute accurate IR spectra for materials[21] and Ghosh et al. used 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 proteins IR spectra, despite of 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.

Theory and Computation Detail

Quantum mechanics treatment for amide I vibration. We adopt a divide-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 (ω_i) of the *i-th* amide I oscillator, and the off-diagonal elements represents the coupling between two oscillators i and j (Fig. 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]

$J_{ij} = \frac{1}{4\pi\epsilon_0}\left(\frac{\vec{\mu_i}\cdot\vec{\mu_j}}{r_{ij}^3} - 3\frac{(\vec{\mu_i}\cdot\vec{r_{ij}})(\vec{\mu_j}\cdot\vec{r_{ij}})}{r_{ij}^5}\right)$, where ε_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]

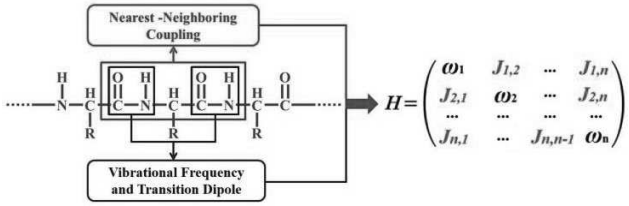


Figure 1. Model Hamiltonian for amide I vibrations in a protein.

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 (ω_i), transition dipole ($\vec{\mu_i}$), and neighboring coupling (J_{ij}) parameters from a NN model. The *N*-methylacetamide (NMA) molecule (Fig. 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 (Fig. S1), also known as the glycine dipeptide. This molecule has been widely used to construct a map of the coupling as function of the Ramachandran angles (ϕ and ψ) between the neighboring peptides.[10, 27-28]

Quantum mechanical calculations for data generation. Configurations of NMA were extracted from *ab initio* molecular dynamics (AIMD) simulation 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 (Fig. S2). From 241.5 ps trajectories with a of 0.5 fs, time step a total of 9660 configurations were extracted at a 25-fs interval 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° $\leqslant\phi\leqslant$ 180° and -180° $\leqslant\psi\leqslant$ 180° at 5° intervals for both angles for machine learning training (Fig. 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. 9660 NMA and 5128 GLDP conformations were generated as training set to predict the vibrational frequency (ω_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 indicating large conformational changes and low similarity (Fig. S3 and S4), which mitigates issues originated from over correlation 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 (Fig. S2 and S5), and the resulting data set is appropriate for establishing structure–property relationships via ML training.

Neural networks architecture and descriptors. Multilayer Perceptron (MLP) with supervised training scheme using a back-propagation algorithm implemented in TensorFlow[32] 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 random search algorithm[38] in TensorFlow (including neurons for hidden layers, learning rate and l2 regularization parameter) to create a reasonable ML protocol in this work.

In order to establish structure–property relationship between geometry and optical properties of proteins, the ground state Coulomb Matrix[39] (CM) of the NMA and part of GLDP molecules (excluding hydrogens and solvent molecule) was taken as descriptor (Fig. S6): $U_{ij} = \begin{cases} 0.5Q_i^{2.4} & \forall i = j \\ Q_iQ_j / \left|R_i - R_j\right| & \forall i \neq j \end{cases}$, where i and j are atomic indices, $\left|R_i - R_j\right|$ is the interatomic distance, and Q_i represents nuclear charge. The merit of 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 the less accurate and efficient than the Coulomb Matrix (Fig. S7). As a rotationally invariant descriptor, the CM lacks of 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 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 (Fig. S8). Consequently, the NN prediction of the $\vec{\mu}_i$ for a new NMA also starts with a treatment of transferring back it to the original coordinate by using the inverse of rotation matrix. The elements of CM (NMA:15; GLDP:21) were then used as inputs (Fig. S6) for NN training and the output (size:1) data are then compared with DFT calculations (Fig. 2). A total of five ML models (ω_i, $\vec{\mu}_i$ (*x,y,z*), J_{ij}) were obtained to construct the vibration model Hamiltonian.

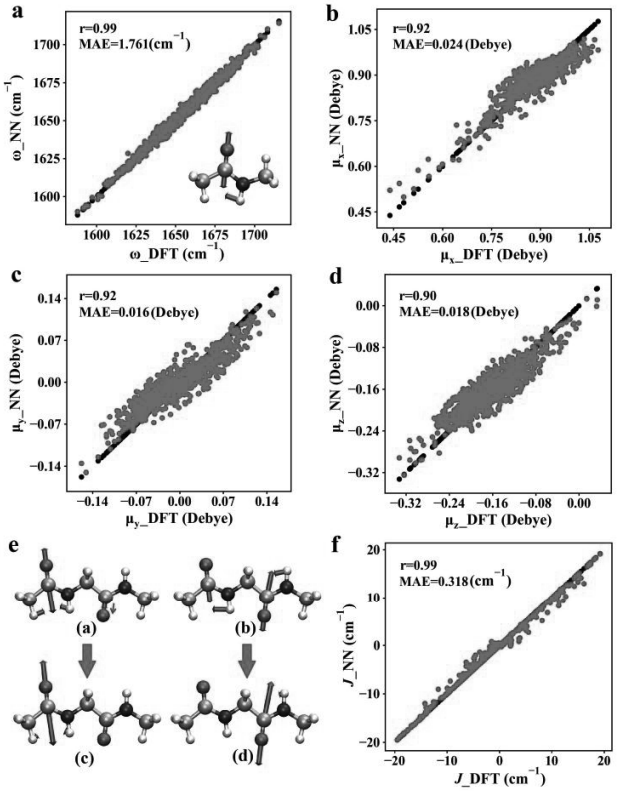


Figure 2. (a) Correlation between the DFT-computed (ω_DFT) (black lines/dots) and NN-predicted (ω_NN) (red lines/dots) amide I vibrational frequencies after cross-validation. (b-d) Comparison of the DFT-computed amide I vibrational transition dipole moment in the *x*, *y*, *z* direction ($\mu_{x,y,z}$_DFT) (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 whole ML process in this work. The learning curves (Fig. S10) indicates 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 (Fig. 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 (ω_i), transition dipole ($\vec{\mu}_i$), and coupling constants (J_{ij}) in the exciton Hamiltonian (Fig. 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 Fig. 3. The protein is split into individual peptide bonds and dipeptides. The ω_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 peptides couplings (computing J_{ij} via the dipole approximation), respectively. The J_{ij} values predicted from GLDP NN model are used for generating off-diagonal elements owing to the nearest neighboring peptide couplings.

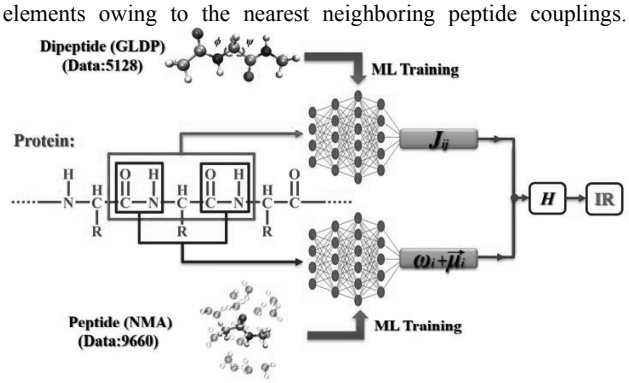


Figure 3. Machine learning protocol for predicting protein IR spectroscopy.

Finally, IR spectra were simulated with the model Hamiltonian, using the SPECTRON program developed by Mukamel and co-workers[41]. As Fig2, Fig S11 and S12 shows, 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.

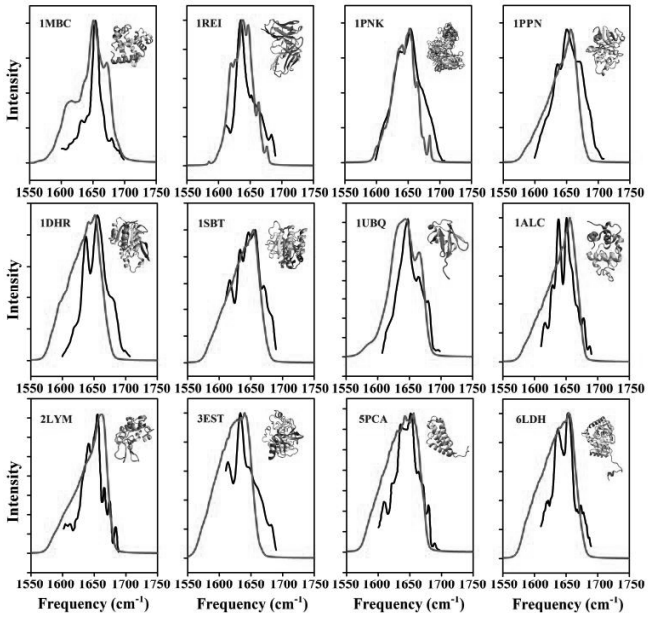


Figure 4. Good agreement (the quantitative agreement between the predicted and experimental spectra were measured by Spearman rank correlation coefficients[45], see Table 1) is obtained between the experimental spectra of the proteins measured in D_2O (black lines)[46-48] and the ML predictions based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

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 (Fig4 and Fig S13). The good agreements 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 the 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 for the dominant signal peak[52]: α-helices: 1640~1650 cm^{-1}, β-sheets: 1620~1640 cm^{-1} & 1680~1690 cm^{-1}, random coil: 1650~1660 cm^{-1}. As indicated by Table 1 and Fig. 4, NN predictions can distinguish the α-helix and β-sheet secondary structures. The α-helical (PDB code: 1MBC) and β-sheet (PDB code: 1REI) proteins exhibit the major spectral peaks at 1650 cm^{-1}, 1634 and 1680 cm^{-1}, respectively. Proteins containing both secondary structures (α+β) show characteristic peaks for both motifs. Taking the 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 (Fig. 4 and Table 1). We have further investigated the amide I signals of λ-Immunoglobulin (1REI) in different states, as shown in Fig. S16, the dominant peak of spectra has a blue shift which corresponds to the change of secondary structure content (β-turns and coil increased while β-strands decreased (Table S4)). We have also predicted transient amide I spectra of 1REI at different time moments based on MD trajectories. As Fig. S17 shows, the main peak has a red shift accompanied with the decrease of β-turns and coil content, and the increase of β-strands content (Table S4). The structural change is clearly captured by the change of spectra (Fig. S16 and S17). This would be useful for tracking conformation 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-works (Fig. 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 neural network (NN) protocol. Roughly speaking, It is at least five 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 4 (1MBC, 1PPN, 3EST, 5PCA) out of 12 proteins (Fig. S14 and Fig S15). Compared with experiment, map results have the RMSE (root mean square error) values of 1.48 to 4.52, and the 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 the Spearman rank correlation coefficients between 0.71 and 0.96 (Fig S15 and Table S9). The local electric potential/field used in map dependens on the quality of atomic charges in the chosen force field. And the empirical function of map trained by a set of protein dataset 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 map method. We expect that the improvement of map results may require re-parameterization the model for 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℃) (Details in Supporting Information). Ubiquitin is a 76-residue protein which contains both α and β 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 cm^{-1} to 1657 cm^{-1} (Table S5), accompanied by a broadening of bands and decrease in intensity (Fig. 5a). This result is in line with experiment (Fig. 5a and Fig. 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 ML protocol, demonstrating good transferability of our ML model to varying external environment factors.

Monitoring 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 mini-protein which has been widely used for studying folding dynamics. 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 un-folded strand (S1), slightly folding but

Table 1. ML predicts IR protein spectra with the root mean square error (RMSE) and high Spearman rank correlation (ρ) indicates the quantitative agreement with experiment. 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 α-helix and β-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 @ 2.1GHz).

Protein	PDB Code	Secondary Class	Number of atoms	ML time (h)	RMSE	ρ
Carbonmonoxymyoglobin	1MBC	α	2459	4.68	2.73	0.94
λ-Immunoglobulin	1REI	β	3254	6.38	2.05	0.90
Penicillin Amidohydrolase	1PNK	$\alpha+\beta$	11708	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
α-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 Aα	5PCA	$\alpha+\beta$	1881	8.52	2.07	0.92

Lactate Dehydrogenase	6LDH	$\alpha+\beta$	5156	9.65	2.12	0.96

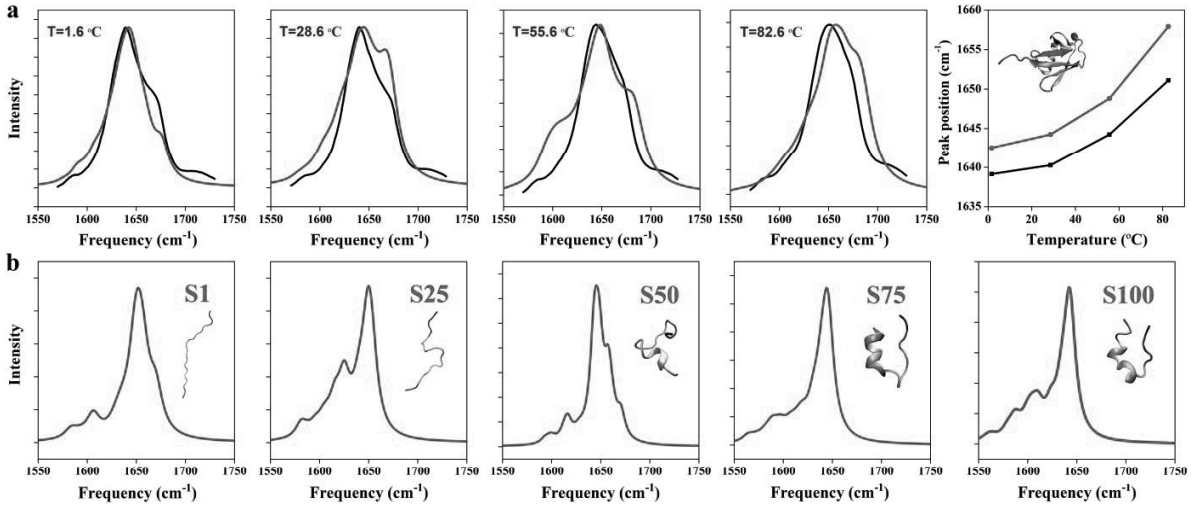


Figure 5. (a) From left to right : Simulated (red line) and Experimental[53] (black line) IR spectra of Ubiquitin at four different temperatures (1.6 ° C ~ 82.6 ° C) and the temperature variation of the dominant peak position. (b) The ML-predicted IR spectra of the Trp-cage protein along its folding 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 Trp-cage (Ubiquitin).

retaining the coil structure (S25), rapid folding stage with a large amount of helical structures (S50), and to the folded helix system 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 Fig. 5b. As the folding process proceeds (S1→S100), the random coil content decreases followed by an increase in the helix content (Fig. 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 (Fig. 5b and Fig. 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 a 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 data set and consider 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 (SFG), and multi-dimensional IR and UV spectroscopies.

ASSOCIATED CONTENT

Supporting Information: Computational details, Molecular Dynamics Simulations, The Machine Learning Protocol, Structure of NMA and GLDP molecules, the root mean square deviation (RMSD), Data distribution, Optimization steps, Hyperparameter optimization, the learning curves for the NN training, Proteins of interest in this study, IR spectra of 12 proteins calculated by map, IR spectra of 1REI with different configurations predicted by NN.

AUTHOR INFORMATION

Corresponding Author:jiangj1@ustc.edu.cn

Acknowledgements

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 (21633006, 21633007, 21790350, 21703221). S.M is grateful to 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. *The Journal of Physical Chemistry 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), 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 β 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. *The Journal of Physical Chemistry 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. *The Journal of Physical Chemistry 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*, Proc. icml, 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, L 1 vs. L 2 regularization, and rotational invariance*, Proceedings of the twenty-first 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. *The Journal of Physical Chemistry B* **2006,** *110* (7), 3362-3374.

42. http://dcaiku.com:12880/platform/first.

43. http://doi.org/10.5281/zenodo.4106438.

44. http://doi.org/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. *PCCP* **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. *The Journal of Physical Chemistry 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. *The Journal of Physical Chemistry 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.

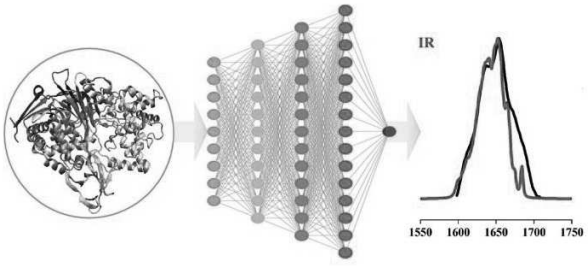
53. Waegele, M. M.; Gai, F., Power-law dependence of the melting temperature of ubiquitin on the volume fraction of macromolecular crowders. *J. Chem. Phys.* **2011,** *134* (9), 03B605.

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. *Angewandte Chemie International Edition* **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. *The Journal of Physical Chemistry B* **2013,** *117* (16), 4661-4669.

SYNOPSIS TOC


IR
1550
1600
1650
1700
1750