AI-based Spectroscopic Monitoring of Real-time Interactions between SARS-CoV-2 and Human ACE2

Sheng Ye,[1,⊥] Guozhen Zhang,[1,⊥] 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

[⊥]These authors contribute equally to this work.

ABSTRACT: The novel coronavirus (SARS-CoV-2) invades a human cell via human angiotensin-converting enzyme 2 (hACE2) as the entry, causing the severe coronavirus disease (COVID-19). The interactions between hACE2 and the spike glycoprotein (S protein) of SARS-CoV-2 hold the key to understand molecular mechanism to develop treatment and vaccines, yet the dynamic nature of these interactions in a fluctuating surrounding is very challenging to probe by those structure determination techniques requiring the structures of samples fixed. Here we demonstrate by a proof-of-concept simulation of IR spectra of S protein and hACE2, that a time-resolved spectroscopy may monitor the real-time information of the protein-protein complexes of interest with the help of machine learning. We expect our machine learning protocol would accelerate the development of real-time spectroscopy study of protein interactions.

The ongoing pandemic of COVID-19, a highly infectious disease caused by SARS-CoV-2, has posed tremendous threat to human health and well-being by having affected several millions of people and killed hundreds of thousands of those who were affected in just a couple of months.[1] It has spurred enormous effort in biological and biomedical research to search for solution of this fatal disease, which rapidly advances our knowledge about it, including the identity of pathogen (i.e. SARS-CoV-2), genome sequence of the virus, and the structural basis for coronavirus recognition and infection.[2-5] SARS-CoV-2 recognizes hACE2 as the entry receptor to host cells using its surface spike glycoprotein (S protein).[1] The interactions of S protein with hACE2 have been subjected to intensive investigations by several groups,[6-10] which laid the foundation for comprehensive understanding on the invasion of SARS-CoV-2 into human body at the atomic scale,[11] helps the search of intermediate hosts of the coronavirus,[12] and will guide the design of therapeutics and vaccines.[11, 13] Since the physiological environment in which S protein and hACE2 interact is always fluctuated due to the dynamic nature of water, a dynamic picture of the interactions between them is needed for precise mechanistic understanding that will inspire modulation and application. Unfortunately, such information relies on real-time tracking of proteins conformations, which cannot be achieved by powerful structure characterization techniques with atomic precision like X-ray diffraction and cryo-electron microscopy because they require fixed structures in samples. It motivates us to develop alternative approaches to resolve the issue.

Recently, time-resolved Infrared (IR) spectroscopy techniques have realized successful monitoring of changes of secondary structure with time,[14] signaling the feasibility of real-time observation of protein dynamics in ambient conditions using spectroscopy. However, to facilitate the monitoring of specific peptide fragments in a secondary structure, it typically needs isotope labelling (e.g. C=O in the amide of protein backbone is replaced with $^{13}C=O$ or $C=^{18}O$) in the preparation of samples, which is unfortunately tedious and expensive for systematic investigation on conformation changes in protein dynamics. Therefore, it is desirable to develop isotope labelling-free spectroscopy to accelerate structure study of proteins for biological and biomedical sciences. To achieve this goal, one needs to employ quantum chemistry calculations to complete spectra signals assignment and structures determination. In fact, it relies on computer simulations of various possible conformers to nail the job, which is unfortunately very expensive for macromolecules like proteins. Thus, developing a cost-effective spectra simulation protocol becomes a pressing task to advance real-time spectroscopy study of protein structures.

Machine learning (ML), a collection of statistics-based methods which gain prediction power from learning of data, has emerged as a powerful toolkit to reduce the barrier of revealing structure-property relationship.[15] It has been increasingly popular in study of molecules and materials, such as predicting chemical reaction routes[16] and accelerating discovery of materials.[17] Especially, neural networks (NN), a subclass of machine-learning algorithms, are well-recognized for handling complex non-linear problems. NN creates the structure-property relationship by iterative learning using a complex high-dimensional function in a much larger, essentially unlimited parameter space. These make it a more adaptable and transferrable tool for the simulation of spectroscopy of protein.[18]

In this communication, we show that our recently developed machine learning protocol will facilitate a real-time prediction of the IR spectra of S protein of SARS-CoV-2. The efficient simulation of IR signals of different states of the protein concerted with the changes in its secondary structure is very encouraging for studying dynamic interactions between S protein of SARS-CoV-2 and human ACE2 with the help of ML techniques. Machine learning should provide a cost-effective tool for simulating optical properties of SARS-CoV-2.

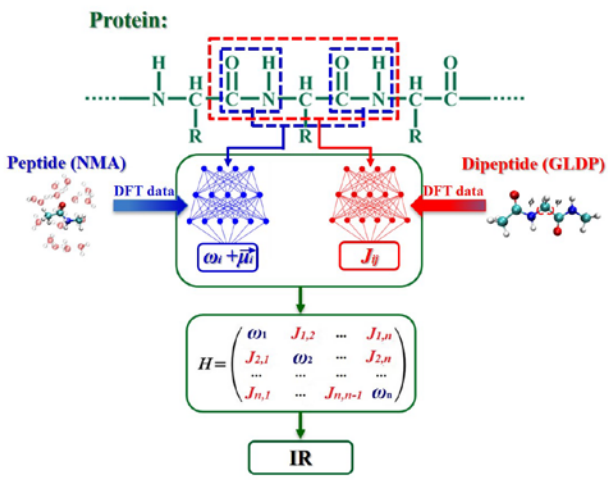

Figure 1. Machine learning protocol for the IR spectra of proteins based on vibrational exciton model.

The technique detail of this ML protocol has been elaborated elsewhere (paper under review). Here we just sketch the basic idea of the framework (Fig. 1). We adopt a divide-and-conquer strategy to treat the amide I vibrations of the whole protein. The vibration of a protein is 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,[19] in which the diagonal elements are the frequency (ω_i) of the ith amide I oscillator, and the off-diagonal elements include the coupling coefficient (J_{ij}) between two oscillators i and j (Fig. 1). To obtain these matrix elements, a protein is split into individual peptide bonds and dipeptides. The values of ω_i and $\vec{\mu}_i$ are predicted from an NN model of peptide, i.e. *N*-methylacetamide (NMA)[20-21]. For off-diagonal elements, there are two scenarios: those coupling coefficients between two neighboring oscillators are computed using a NN model of dipeptide, i.e. *N*-acetyl-glycine-*N*'-methylamide (GLDP);[22-23] those between a pair of non-neighboring oscillators are calculated with the dipole approximation[24] assuming that given the distances between oscillators are greater than their sizes: $J_{ij} = \frac{1}{4\pi\varepsilon_0}\left(\frac{\vec{\mu}_i \cdot \vec{\mu}_j}{r_{ij}^3} - 3\frac{(\vec{\mu}_i \cdot \vec{r}_{ij})(\vec{\mu}_j \cdot \vec{r}_{ij})}{r_{ij}^5}\right)$, 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. After all matrix elements of the model Hamiltonian are obtained, IR spectra are simulated using the SPECTRON program developed by Mukamel and co-workers.[25] We also make this ML protocol online to provide rapid protein IR spectroscopy prediction, paving the way for a real-time operation of ultrafast experimental spectroscopy.[26]

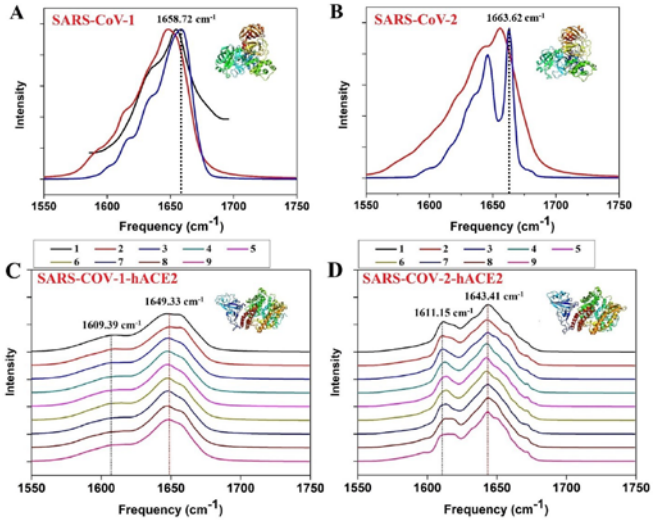

Figure 2. (A) Comparison of experimental[27] (black line) and ML predicted (red line: single crystal structure, black line: average of 1000 configurations) spectra of SARS-CoV-1. (B) ML-predicted IR spectra of SARS-CoV-2 based on a single crystal structure (red lines) and 2000 MD configurations (blue lines). (C) ML-predicted IR spectra of SARS-CoV-1-ACE2 during 10us MD simulation (contains 9 trajectories,1000 snapshots for No. 1-8 trajectories, 334 snapshots for No.9 trajectory). (D) same as (C) but for SARS-CoV-2-ACE2. Intensity is scaled to have the same maximum intensity for each panel.

We first simulated the amide I IR spectra of SARS-CoV-1 and SARS-CoV-2 were simulated using the ML protocol described in Fig. 1 by average 1000 and 2000 snapshots for each other (which would be prohibitively expensive via direct QM computations). The structures and trajectories of SARS-CoV-1 and SARS-CoV-2 are obtained from MD simulations of ourselves (MD simulation details in Supporting Information) and Komatsu and co-workers.[28] The good agreement of SARS-CoV-1 between our ML predictions (average 1000 snapshots) and experimental spectra[27] is evident from the high Spearman rank correlation coefficients[29] (ρ=0.93) (Fig. 2), which was widely used to measure the agreement between the predicted and experimental spectra. Then we predicted the amide I IR spectra of the SARS-CoV-2 with this ML protocol (average 2000 snapshots). As shown in Fig. 2, the dominant peak of SARS-COV-2 has a 5 cm^{-1} blue-shift compared with SARS-COV-1 (SARS-COV-1:1658.72 cm^{-1}, SARS-COV-2:1663.62 cm^{-1}). This may be accounted by that SARS-COV-2 has a larger portion of the β-turns content than SARS-COV-1 (Table 1) and β-turns possess an amide IR signal of higher frequency. Importantly, our ML protocol not just identifies the fine difference in amid I IR spectra associated with the difference between their secondary structures, but also is four orders of magnitude faster than conventional quantum chemistry calculations.

Then we simulated the amide I IR spectra of SARS-CoV-1-ACE2 (hACE2 in complex with the receptor binding domain of spike protein from SARS-CoV-1) and SARS-CoV-2-ACE2 (hACE2 in complex with the receptor binding domain of spike protein from SARS-CoV-2) by average 8334 snapshots with our ML protocol (Fig. 2). These MD simulation data were retrieved from the website of D. E. Shaw research.[30] Each MD simulation is 10 μs and contains 9 trajectories (1000 snapshots for No. 1-8 trajectories, 334 snapshots for No.9 trajectory). We also chose

Table 1. Average secondary structure content (computed by Stride program) of various coronavirus and comparison of the time required for computing IR spectra of single structures by DFT and our ML model based on vibrational exciton model. All reported times refer to calculations on an 8core of an Intel(R) Xeon(R) CPU (E5-2683v4 @ 2.1GHz).

	β-strands	β-turns	α-helix	310-helices	Coil	Bridge	DFT (s)	ML (s)
SARS-COV-1	30.1%	19.9%	23.9%	2.5%	21.0%	2.5%	1165320	70.69
SARS-COV-2	28.3%	25.5%	20.3%	2.6%	20.4%	2.9%	1173000	72.68
SARS-CoV-1-ACE2	7.6%	23.2%	45.2%	3.9%	18.0%	2.2%	1482120	100.80
SARS-CoV-2-ACE2	7.0%	21.2%	45.6%	3.2%	21.8%	1.2%	1474440	98.68
Trimeric SARS-CoV-2 spike glycoprotein (closed state)	30.5%	25.3%	17.9%	1.8%	22.7%	1.7%	6060720	960.33
Trimeric SARS-CoV-2 spike glycoprotein (open state)	30.2%	22.9%	17.9%	2.2%	24.4%	1.4%	6060720	960.33
RBD/hACE2 binding (S1 state)	32.3%	22.1%	9.4%	7.8%	27.7%	0.8%	370440	20.64
RBD/hACE2 binding (S2 state)	31.8%	21.5%	12.1%	6.2%	27.3%	1.2%	370440	20.64
RBD/hACE2 binding (S3 state)	33.5%	25.5%	12.1%	6.2%	21.5%	1.2%	370440	20.64
RBD/hACE2 binding (S4 state)	33.0%	21.4%	9.4%	7.8%	27.3%	1.2%	370440	20.64
RBD/hACE2 binding (S5 state)	33.0%	21.9%	11.6%	4.7%	27.6%	1.2%	370440	20.64

the averaged IR spectra of the first trajectory (1th: 1200 ns which contain 1000 snapshots) for comparison. From the average secondary structure content analysis (by average 1000 snapshots from No.1 trajectory) by Stride program,[31] the random coil content of RBD2-hACE2 was higher than RBD1-hACE2, and the β-turn content was lower than RBD1-hACE2, which leading to a 6 cm^{-1} red-shift of the dominant peak (RBD1-hACE2: 1649.33 cm^{-1}, RBD2-hACE2:1643.41cm^{-1}) (Table 1). Again, the difference of secondary structures between RBD1-hACE2 and RBD2-hACE2 is clearly characterized by our ML-based IR spectra simulation.

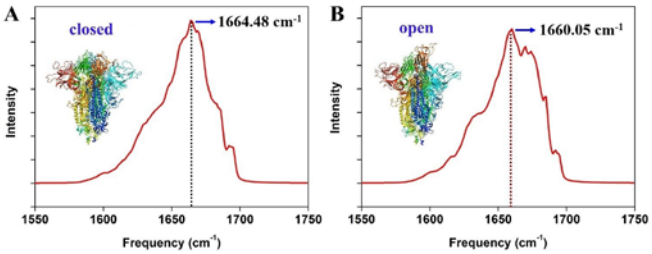

Figure 3. (A) ML-predicted I R spectra of Trimeric SARS-CoV-2 spike glycoprotein in (A) closed state and (B) open state.

Recently research shows that the trimeric SARS-CoV-2 spike glycoprotein has two distinctive states: closed state and open state.[6] Intriguingly, they have substantially different secondary structures. We simulated the IR spectra by averaging 200 snapshots (10 μs simulation MD data were retrieved from the website of D. E. Shaw research; 200 frames were selected in first 240ns) of closed and open state with this ML protocol. It is noticed that the dominant peak of the Trimeric SARS-CoV-2 spike glycoprotein in open state has a 4.4 cm^{-1} red shift compared with closed state which coincide the secondary structure content difference (the β-turns of the open state is lower but the coil content is higher than closed state) (Fig. 3 and Table 1).

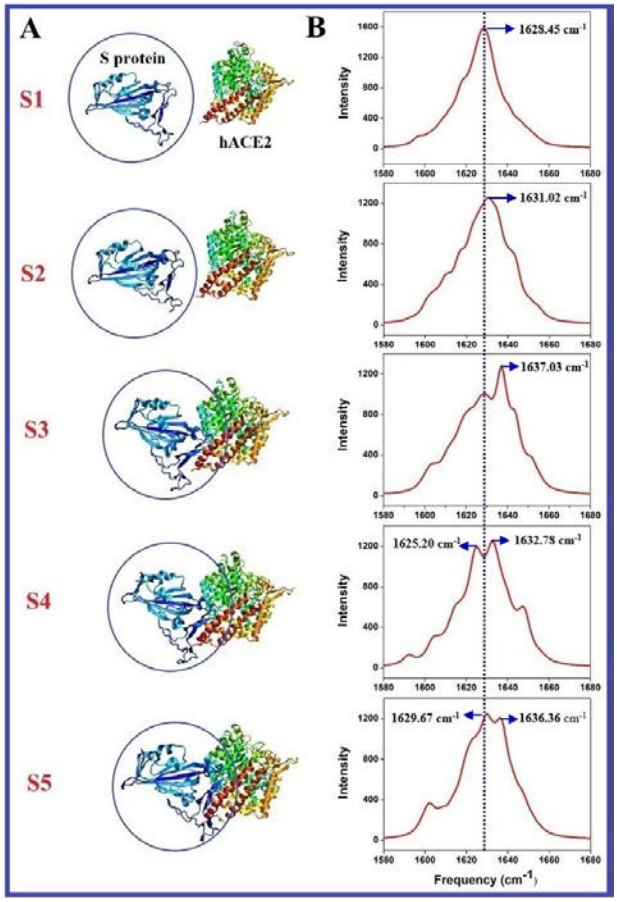

Figure 4. Five representative states of the receptor-binding domain (RBD) of the SARS-CoV-2 spike (S protein) and the human ACE2 (hACE2) receptor were selected from the combination trajectory.

Finally, we investigate the dynamics of S protein of SARS-CoV-2 interacting with hACE2 interaction using our ML protocol. Five representative structures were selected from the D. E. Shaw research.[30] We predicted the IR spectra of S protein in different states during the combination process by ML and calculated the average secondary structure components in each state (Fig. 4 and Table 1). From the S1 to S2 state, the IR spectra has a 2.57 cm^{-1} blue shift. The analysis of the average secondary structure content showed that the main change from S1 to S2 was the increased content of α-helix which lead a blue-shift. From S2 to S3, the IR

spectra also has a 6 cm^{-1} blue-shift correspond with averaged secondary structure content change (S2 to S3: β-turns increased while coil decreased). From S3 to S4, the IR spectra has a 5cm^{-1} red-shift which caused by the β-turns and α-helix decreased while coil content increased. From S4 to S5, the IR spectra has a 4 cm^{-1} blue shift which caused by β-turns and α-helix increased. Since the changes in the IR spectra of the S protein under different states associated with the changes in the secondary structure are correctly captured by our ML protocol, we think our method provides a promising route for studying real-time dynamics regarding to the interactions of SARS-CoV-2 and human ACE2.

In conclusion, we proposed a cost-effective machine learning protocol for predicting amide I IR spectra of SARS-COV-2 spike protein. Compared to conventional quantum chemistry approaches, it significantly accelerates the simulation of IR spectra of protein complexes, which is crucial for developing time-resolved IR spectroscopy techniques for studying dynamic protein-protein interactions.

ASSOCIATED CONTENT

Supporting Information

Molecular Dynamics Simulations.

AUTHOR INFORMATION

Corresponding Author: jiangj1@ustc.edu.cn

ACKNOWLEDGMENTS

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of University of Science and Technology of China.

References

1. Zhou, P., et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature* **2020,** *579* (7798), 270-273.
2. Zhu, N., et al., A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* **2020**.
3. Lu, R., et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **2020,** *395* (10224), 565-574.
4. Wu, F., et al., A new coronavirus associated with human respiratory disease in China. *Nature* **2020,** *579* (7798), 265-269.
5. Wrapp, D., et al., Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020,** *367* (6483), 1260-1263.
6. Yan, R., et al., Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **2020,** *367* (6485), 1444-1448.
7. Lan, J., et al., Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **2020,** *581* (7807), 215-220.
8. Shang, J., et al., Structural basis of receptor recognition by SARS-CoV-2. *Nature* **2020,** *581* (7807), 221-224.
9. Walls, A. C., et al., Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **2020**.
10. Wang, Q., et al., Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **2020**.
11. Zhang, H., et al., Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive care medicine* **2020,** *46* (4), 586-590.
12. Liu, Z., et al., Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *Journal of medical virology* **2020,** *92* (6), 595-601.
13. Zheng, M.; Song, L., Novel antibody epitopes dominate the antigenicity of spike glycoprotein in SARS-CoV-2 compared to SARS-CoV. *Cellular & molecular immunology* **2020,** *17* (5), 536-538.
14. Seo, J., et al., An infrared spectroscopy approach to follow β-sheet formation in peptide amyloid assemblies. *Nature chemistry* **2017,** *9* (1), 39-44.
15. Butler, K. T., et al., Machine learning for molecular and materials science. *Nature* **2018,** *559* (7715), 547-555.
16. Segler, M. H., et al., Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018,** *555* (7698), 604-610.
17. Ma, S., et al., Dynamic coordination of cations and catalytic selectivity on zinc–chromium oxide alloys during syngas conversion. *Nat. Catal.* **2019,** *2* (8), 671-677.
18. Ye, S., et al., A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U.S.A.* **2019,** *116* (24), 11612-11617.
19. Hamm, P.; Zanni, M., *Concepts and methods of 2D infrared spectroscopy*. Cambridge University Press: 2011.
20. Wang, L., et al., Development and validation of transferable amide I vibrational frequency maps for peptides. *J. Phys. Chem. B* **2011,** *115* (13), 3713-3724.
21. Hayashi, T., et al., Electrostatic DFT map for the complete vibrational amide band of NMA. *J. Phys. Chem. A* **2005,** *109* (43), 9747-9759.
22. la Cour Jansen, T., et al., Modeling the amide I bands of small peptides. *J. Chem. Phys.* **2006,** *125* (4), 044312.
23. 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.
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. Zhuang, W., et al., Simulation protocols for coherent femtosecond vibrational spectra of peptides. *J. Phys. Chem. B* **2006,** *110* (7), 3362-3374.
26. http://dcaiku.com:12880/platform/first.
27. Surya, W., et al., Structural and functional aspects of viroporins in human respiratory viruses: respiratory syncytial virus and coronaviruses. *Respiratory Disease and Infection-A New Insight; Vats, M., Ed* **2013**, 47-76.
28. KOMATSU, Teruhisa S.; KOYAMA, Yohei M.; OKIMOTO, Noriaki; MORIMOTO, Gentaro; OHNO, Yousuke; TAIJI, Makoto (2020), "COVID-19 related trajectory data of 10 microseconds all atom molecular dynamics simulation of SARS-CoV-2 dimeric main protease", Mendeley Data, v1 http://dx.doi.org/10.17632/vpps4vhryg.1.
29. Besley, N. A.; Hirst, J. D., Theoretical Studies toward Quantitative Protein Circular Dichroism Calculations. *J. Am. Chem. Soc.* **1999,** *121* (41), 9636-9644.
30. D. E. Shaw Research, "Molecular Dynamics Simulations Related to SARS-CoV-2," D. E. ShawResearch Technical Data, 2020. http://www.deshawresearch.com/resources_sarscov2.html.
31. 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.