COMMUNICATION

Theoretical study of the optical spectra of SARS-CoV-2 proteins†

Zhuo Li[a] and Jonathan D. Hirst*[b]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Treatment for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes Covid-19, may well be predicated on knowledge of the structures of protein of this virus. However, often these cannot be determined easily or quickly. Herein, we provide calculated circular dichroism (CD) spectra in the far- and near-UV, and infra-red (IR) spectra in the amide I region for experimental structures and computational models of SARS-CoV-2 proteins. The near-UV CD spectra offer greatest sensitivity in assessing the accuracy of models.

Since the outbreak of SARS-CoV-2 at the beginning of 2020, scientists have been seeking insights that will underpin solutions to this new pandemic. Structural characterization of the proteins of this virus is crucial in terms of understanding their biological functions, finding inhibitors and designing vaccines. Currently, structures of more than half of the SARS-CoV-2 proteins have not yet been determined experimentally. Prediction of protein structures using computational methods is an alternative approach. Computational researchers from around the world are sharing the models of SARS-CoV-2 proteins that they obtained with different prediction algorithms, either from a homology template or de novo, in order to facilitate drug design or functional studies.[1] However, the predicted protein structures need to be validated via experimental methods. Optical spectroscopy may provide valuable information in this respect.

In this study, we have collected SARS-CoV-2 protein coordinates either from experimental determination as reported in the Protein Data Bank (PDB) or from computational modelling[2-4] in order to calculate their optical spectra (Table S1 in the ESI†). Eight of the proteins have more than one PDB entry. Multiple PDB structures of the same protein were used if they were determined under similar experimental conditions. Other PDB entries of the considered proteins were not included, because of the variation of the experimental conditions, bound ligands and complex composition. For the spike glycoprotein we only calculated its near-UV CD spectrum, due to its large size. Protein structures are flexible under physiological conditions. Although informative, a crystal structure or a modelled structure can only represent a fixed stable state. Thus, we also calculated near-UV CD spectra using snapshots from MD trajectories shared by D. E. Shaw Research[5] of five proteins to take into account the conformational flexibility of the proteins in solution. Simulations of spike protein RBD-ACE2 (6M17), spike glycoprotein (6VXX and 6VYB) and RNA polymerase-nsp7-nsp8 (6M71) had been performed for 10μs. We extracted one snapshot every 20 frames, corresponding to a time interval of 1.2 ns between snapshots. For each protein, a total of 425 snapshots were used in the calculation of the near-UV CD spectra. The simulation of the main protease (6Y84) was over a period of 100 μs. We extracted one snapshot every 200 frames and in this case 500 snapshots at time intervals of 1 ns were used in the near-UV CD calculation.

We considered 23 computational models of 17 protein or protein fragments from three sources (Table S1). The SWISS-MODEL server has provided the full SARS-CoV-2 proteome including 25 protein models and five hetero-oligomeric complexes.[2] In this study, we selected the models where there are no experimental coordinates available and where their QMEAN quality estimates[6] indicate high quality/confidence. Thus, we computed spectra of seven SwissModel SARS-CoV-2 structures (Table S1). There are proteins where no homology model is available, due to a lack of a suitable template. Six proteins have also been studied with the 'free modelling' method, AlphaFold, which is a deep learning system to predict the structure of proteins for which there is no similar template on which to base a model.[3] Models of these proteins and another four were refined by Heo and Feig[4] via their pipeline including inter-residue distance prediction with trRosetta,[7] lowest energy model selection and molecular dynamics (MD)

[a] School of Pharmaceutical Sciences, Jilin University, Changchun 130021, China.
[b] School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom.

†Electronic Supplementary Information (ESI) available: [Table of the SARS-CoV-2 protein structures and models studied and a complete set of computed spectra]. See DOI: 10.1039/x0xx00000x

simulation-based refinement. The secondary structure content of the proteins shown in Table S1 was calculated with the DSSP server.[8,9]

We calculate three types of optical spectra of SARS-CoV-2 proteins: far-ultraviolet (UV) electronic circular dichroism (CD), near-UV CD and infrared (IR) spectra in the amide I region (Fig. 1-3 and Fig. S1-S12). Although these techniques do not provide atomic level structural information, they reflect the conformational flexibility of proteins in solution (and by extension under physiological conditions). All three spectroscopic methods can monitor conformational changes perhaps induced by changes in experimental conditions or by ligand/inhibitor binding. Thus, comparing the experimental spectra with the calculated spectra may provide information about the quality of the structural model.

Far-UV CD spectroscopy is sensitive to chiral features of the protein backbone which are reflected by the excitation bands arising from the peptide bond. Each type of protein secondary structure has its unique far-UV CD features[10] and the far-UV CD spectrum can be deconvolved as a combination of each secondary structure component.[11] It has been used in determining the secondary structures of the papain-like protease of MERS[12] and the envelope protein of SARS.[13] The far-UV CD has also been used to measure the interaction between the SARS-CoV-2 spike protein receptor binding domain (RBD) and heparin.[14] Calcium binding by the SARS 3a protein has been monitored using CD.[15] The far-UV CD spectrum showed the increase of α-helicity when mixing peptides from two heptad repeat regions of the SARS spike protein which suggested the formation of a complex.[16,17] This fusogenic mechanism is also evident in SARS-CoV-2 and has been used to design fusion inhibitors.[18] In this study, we employed the PDB2CD server,[19] an empirical method, to predict the far-UV CD spectra of the SARS-CoV-2 proteins. A reference dataset, SMP180 (soluble + membrane), is used to search structural similarity of the query protein and proteins in the dataset and the CD spectra of the reference proteins are used to create the CD spectrum.

The amide I IR spectrum of a protein also provides information about the backbone conformation. The signal arises mainly from the C=O stretch mode and the N-H bond wagging and bending. Different secondary structures have different peak positions between 1620 and 1690 cm^{-1}.[20] The amide I band was calculated with Coupled Oscillator Model Spectrum Simulator (COSMOSS).[21] All settings were used at their default values. COSMOSS constructs a vibrational exciton Hamiltonian. The coupling between the local amide I modes was modelled was using transition dipole coupling with the nearest neighbour coupling corrected by Jansen's (ϕ, ψ) angle map.[22] The local mode frequencies of the amide I vibrations are also calculated using a nearest-neighbour frequency shift.[22] The calculation of the IR amide I band was based on a single structure, but the influence of conformational dynamics on the inhomogeneous broadening of the bands is approximated by adding some random disorder to the elements of the exciton Hamiltonian: 20 cm^{-1} to the diagonal terms and 5 cm^{-1} to the off-diagonal elements, corresponding to the magnitude of fluctuations that might be anticipated in solution under ambient conditions. Homogeneous broadening is approximated by convolving the computed line spectra with a 10 cm^{-1} linewidth.

Near-UV CD spectra of proteins, on the other hand, contain information regarding the orientations of the aromatic side chains as well as their interactions with the surrounding environment.[11] Near-UV CD spectra were calculated with the DichroCalc server,[23] which uses a matrix formulation to represent the exciton coupling of the chromophores in the protein. The electronic transitions of the aromatic side chain chromophores, phenylalanine (Phe), tyrosine (Tyr) and tryptophan (Trp), were described via *ab initio* calculated parameters[24] extended to incorporate the vibrational structure under the electronic bands.[25] Backbone chromophores were modelled with an *ab initio* parameter set[26] and two transitions were employed for the peptide bond. The calculated near-UV CD line spectra were convolved with a Gaussian function with a 4 nm bandwidth. Calculated intensities from each of the MD snapshots were averaged with equal weighting to give the final spectrum.

Fig. 1 compares the spectroscopic features of the four structures of the main protease in its apo form. All four structures have almost identical far-UV CD spectra with one moderate positive peak at 190 nm, one moderate negative band at 208 nm and an unresolved shoulder at 220 nm.

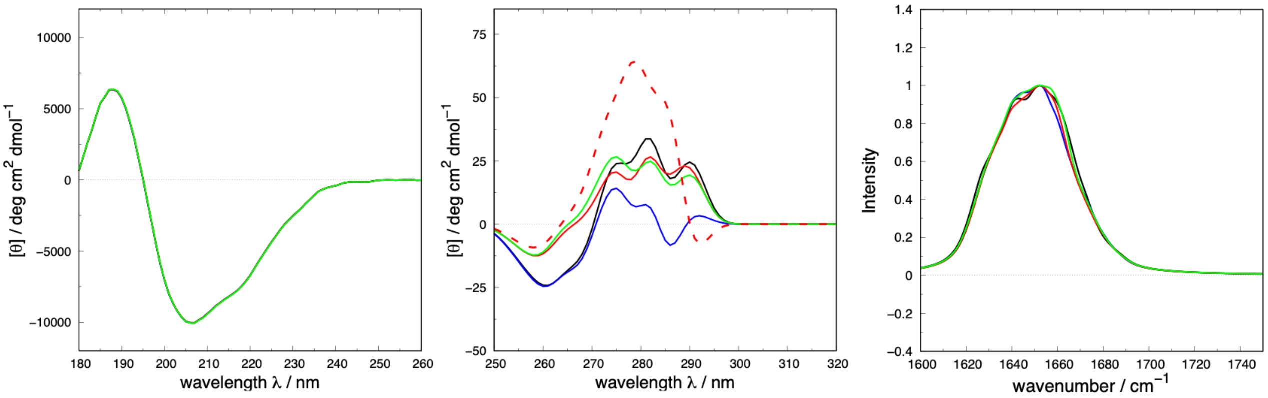

Fig. 1 Calculated spectra with four experimentally determined structures of the apo-form SARS-CoV-2 main protease. Left: far-UV CD; middle: near-UV CD; right: IR. PDB ID and colour code: 6M03 (black), 6Y2E (blue), 6Y84 (red), 5R8T (green). The spectrum depicted with dashed lines in the middle panel is calculated with MD simulation snapshots.

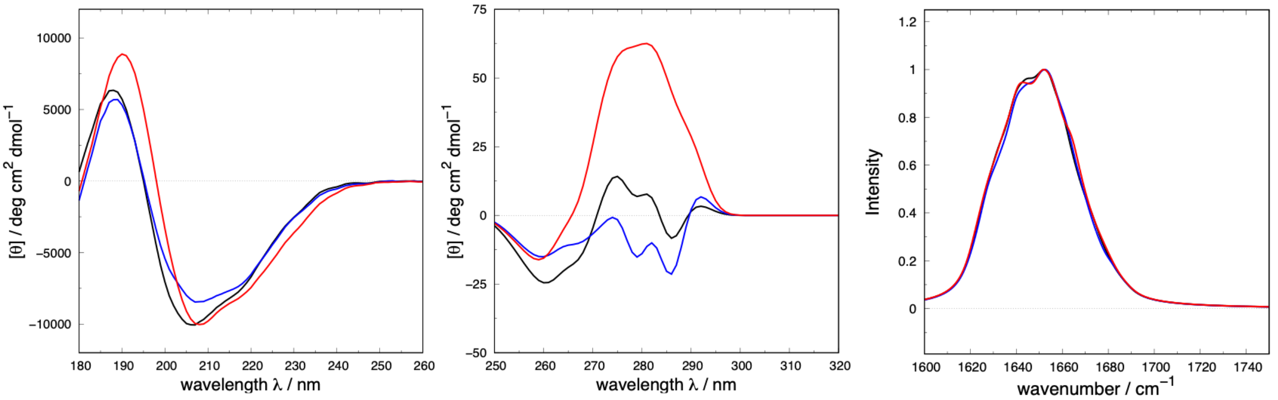

Fig. 2 Calculated spectra with crystal structures of SARS-CoV-2 main protease with (red and blue) or without (black) alpha-ketoamide inhibitors binding. Left: far-UV CD; middle: near-UV CD; right: IR. PDB ID and colour code: 6Y2E (black), 6Y2F (blue), 6Y2G (red).

This feature matches the mixed α-helical and β-strand content of the secondary structure of this protein (Table S1). The calculated amide I IR bands of these four structures are very similar as well. The near-UV CD spectra, however, show obvious differences. Calculation with the 6Y2E structure leads to less intense signals in the tyrosine and tryptophan region (270 – 290 nm), while the other three structures show similar bands in terms of the band shape and the intensity. 6Y2E and 6M03 have a stronger negative bands in the phenylalanine region (250 – 268 nm) than the other two 3C-like proteinase structures. This example illustrates that crystal structures from different experiments show similar secondary structure content, but the conformations of the aromatic side chains may be different. Near-UV CD spectrum is sensitive to the different orientation of the aromatic side chains. Calculated spectra of the methyltransferase-nsp10 complex (Fig. S1), ADP ribose phosphatase (Fig. S2) and RNA polymerase-nsp7-nsp8 complex (Fig. S6) also illustrate that the near-UV CD is sensitive to the local conformational differences of the aromatic residues. The other three proteins, nsp9 (Fig. S4), nucleocapsid phosphoprotein (Fig. S5) and the spike protein RBD and ACE2 complex (Fig. S7), have similar band shapes for all three types of calculated spectra.

We have calculated the near-UV CD spectra with the snapshots extracted from the trajectories provided by D. E. Shaw Research[5] (Fig. 1, middle panel), since theoretical prediction of the spectrum can sometimes be enhanced by using a set of conformations sampled from an equilibrium ensemble.[27] The calculated spectra with snapshots show similar band shapes for the tyrosine and phenylalanine region with varying magnitude. However, the tryptophan peak is oppositely signed.

Small molecule inhibitors are one possible solution to prevent the function of SARS-CoV-2 main protease in processing RNA translated proteins. Fig. 2 compares the calculated spectra of the main protease with and without the inhibitor binding. Compared to the apo form (6Y2E), inhibitor binding structures (6Y2F and 6Y2G) have slightly lower helical content and a slightly higher amount of turn (Table S1). 6Y2G has the same β-strand content as the apo form, whereas 6Y2E has less β-strand in its structure. These secondary structure changes are reflected in the far-UV CD spectra (Fig. 2 left panel). In the near-UV CD spectra (Fig. 2 middle panel), binding to inhibitors induced opposite spectroscopic changes. 6Y2F shows a weak band with opposite sign in the tyrosine region (270 – 286 nm) compared to the apo structure (6Y2E), while 6Y2G gives a much stronger positive spectral signal from the tyrosine and tryptophan residues. The influence of inhibitor binding on the near-UV CD arises from the overall conformation changes of the protein rather than direct electronic coupling with the aromatic side chain, since only one phenylalanine is within 6 Å of the inhibitor (Fig. S13).

Fig. 3 shows examples of calculated spectra with different models of the same proteins, nsp6, M-protein and nsp4. The calculated far-UV CD features (Fig. 3) consist of an intense positive peak at 190 nm and negative peaks at 208 and 220 nm, which correspond well to this secondary structure composition, namely high helix content (Table S1). The intensities of peaks decrease with the reduction of the helical content of the protein. The two models of nsp6 show a similar band shape in the near-UV CD spectrum below 290 nm, but the Feig model gives weaker calculated signals. The tryptophan in the Feig model leads to a negative signal, while no such band was observed in the AlphaFold structure. For M-protein and nps4, models from the two groups have opposite signed peaks in their calculated near-UV CD spectra. We are still investigating the origin of this. However, it is hard to dissect the near-UV CD signals especially for proteins with a large number of aromatic residues.

ORF10 and Protein 7a have no helical structure. They contain 26.3% and 60.3% β-strand content, respectively. As shown in Figure S10, they have very different far-UV CD spectra. This is mainly due to the nature of the β structures. Regular β sheets have a positive band at around 195 nm and a negative band with comparable magnitude near 216 nm. A spectrum with a positive band near 190 nm and a minimum at about 200 nm indicates irregular β structures in the protein.[28] Thus, Protein 7a shows regular β–sheet features and ORF10 has more β–bulges and irregular strands in its structure.

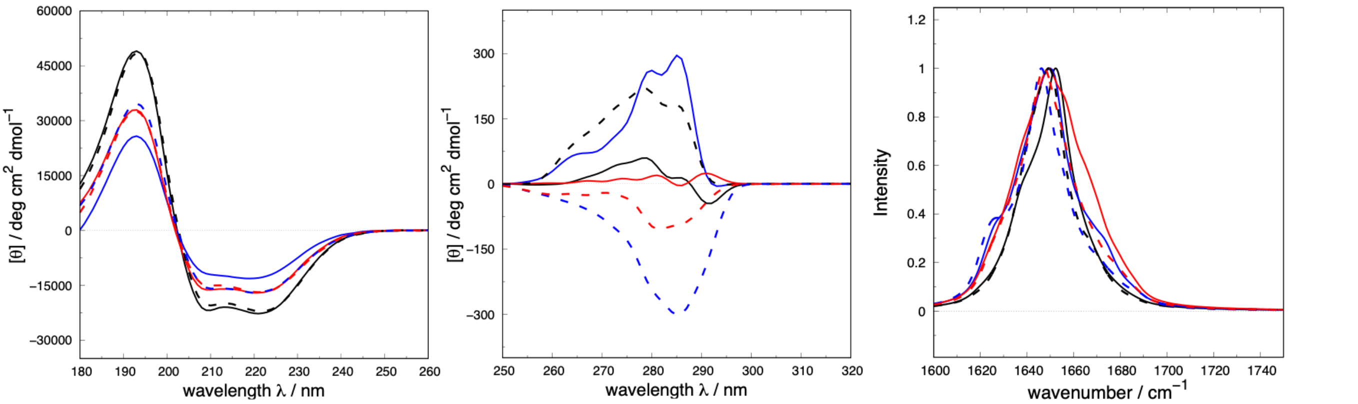

Fig. 3 Calculated spectra of SARS-CoV-2 proteins with models from the Feig group (solid line) or AlphaFold (dashed line). Left: far-UV CD; middle: near-UV CD; right: IR. Colour code: nsp6 (black), M-protein (blue) and nsp4 (red).

The experimentally determined protein structures share similar secondary structures among different PDB entries. These generated similar calculated far-UV CD and amide I IR spectra. The near-UV CD spectrum is sensitive to the local conformational changes of aromatic residues. Thus, it can distinguish subtly varying tertiary structure features in different protein models. When combined with an ensemble of MD snapshots, calculated spectra may correspond directly to the actual conformational distribution of the protein in solution. Furthermore, the calculated spectra are sensitive to ligand binding. They provide a direct connection between models and experimental observables and they should be useful in assessing the accuracy of the computational models when compared with experimentally measured spectra of the proteins. The above remarks, of course, apply generally to proteins. In the context of the intense current interest in the SARS-CoV-2 proteins, we suggest that measurement of the optical spectra, particularly of the near-UV CD spectra, would be a valuable complement to the ongoing associated structural, simulation and modelling studies.

Conflicts of interest

There are no conflicts to declare.

Notes and references

1 R. E. Amaro and A. J. Mulholland, *J. Chem. Inf. Model.*, 2020, in press. https://pubs.acs.org/doi/10.1021/acs.jcim.0c00319.

2 https://swissmodel.expasy.org/repository/species/2697049.

3 J. Jumper, K. Tunyasuvunakool, P. Kohli, D. Hassabis and the AlphaFold team. DeepMind website, March 5, 2020. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.

4 L. Heo and M. Feig, *bioRxiv* 2020, https://doi.org/10.1101/2020.03.25.008904.

5 D. E. Shaw Research, "Molecular Dynamics Simulations Related to SARS-CoV-2," D. E. Shaw Research Technical Data, 2020. http://www.deshawresearch.com/resources_sarscov2.html.

6 P. Benkert, M. Biasini and T. Schwede, *Bioinformatics*, 2011, **27**, 343.

7 J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov and D. Baker, *Proc. Natl. Acad. Sci. U S A,* 2020, **117**, 1496.

8 W. Kabsch and C. Sander, *Biopolymers,* 1983, **22**, 2577.

9 W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten and G. Vriend, *Nucleic Acids Res.,* 2015, **43**, D364.

10 Y. H. Chen, J. T. Yang and H. M. Martinez, *Biochemistry,* 1972, **11**, 4120.

11 R. W. Woody, *Methods Enzymol*., 1995, **246**, 34.

12 M.-H. Lin, S.-J. Chuang, C.-C. Chen, S.-C. Cheng, K.-W. Cheng, C.-H. Lin, C.-Y. Sun and C.-Y. Chou, *J. Biomed. Sci*., 2014, **21**, 54.

13 Y. Li, W. Surya, Claudine and J. Torres, *J. Biol. Chem.*, 2014, **289**, 12535.

14 C. Mycroft-West, D. Su, S. Elli, S. Guimond, G. Miller, J. Turnbull, E. Yates, M. Guerrini, D. Fernig, M. Lima and M. Skidmore, *bioRxiv* 2020, https://doi.org/10.1101/2020.02.29.971093.

15 R. Minakshi, K. Padhan, S. Rehman, I. M. Hassan and F. Ahmad, *Virus Res.*, 2014, **191**, 180.

16 S. Liu, G. Xiao, Y. Chen, Y. He, J. Niu, C. R. Escalante, H. Xiong, J. Farmar, A. K. Debnath, P. Tien and S. Jiang, *Lancet*, 2004, **363**, 938.

17 B. Tripet, M. W. Howard, M. Jobling, R. K. Holmes, K. V. Holmes and R. S. Hodges, *J. Biol. Chem.*, 2004, **279**, 20836.

18 S. Xia, M. Liu, C. Wang, W. Xu, Q. Lan, S. Feng, F. Qi, L. Bao, L. Du, S. Liu et al., *Cell Res.*, 2020, **30**, 343.

19 L. Mavridis and R. W. Janes, *Bioinformatics,* 2017, **33**, 56.

20 H. Yang, S. Yang, J. Kong, A. Dong and S. Yu, *Nat. Protoc*. 2015, **10**, 382.

21 J.-J. Ho, COSMOSS: Coupled OScillator MOdel Spectrum Simulator. https://github.com/JJ-Ho/COSMOSS Retrieved April 28, 2020.

22 T. L. C. Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst and J. Knoester, *J. Chem. Phys.*, 2006, **125**, 44312.

23 B. M. Bulheller and J. D. Hirst, *Bioinformatics,* 2009, **25**, 539.

24 D. M. Rogers and J. D. Hirst, *Biochemistry*, 2004, **43**, 11092.

25 Z. Li and J. D. Hirst, *Chem. Sci.* 2017, **8**, 4318.

26 N. A. Besley and J. D. Hirst, *J. Am. Chem. Soc*., 1999, **121**, 9636.

27 J. D. Hirst, S. Bhattacharjee and A. V. Onufriev, *Faraday Discuss.*, 2003, **122**, 253.

28 P. Manavalan and W. C. Johnson Jr., *Nature*, 1983, **305**, 831.