Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Chemical Physics Letters 800 (2022) 139663

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Frontiers article

Constructing high-accuracy theoretical Raman spectra of SARS-CoV-2 spike proteins based on a large fragment method

Shuang Ni [a], Qiang Yang [b], Jinling Huang [a], Minjie Zhou [a,*], Lai Wei [a], Yue Yang [a], Jiaxin Wen [a,c], Wenbo Mo [a,c], Wei Le [a], Daojian Qi [a], Lei Jin [a], Bo Li [a], Zongqin Zhao [a], Kai Du [a]

[a] Laser Fusion Research Center, China Academy of Engineering Physics, 621900 Mianyang, China
[b] China Academy of Engineering Physics, 621900 Mianyang, China
[c] Department of Engineering Physics, Tsinghua University, 100084 Beijing, China

ARTICLE INFO

Keywords:
Raman spectra
Spike protein
SARS-CoV-2
Large fragment

ABSTRACT

In order to control COVID-19, rapid and accurate detection of the pathogenic, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is an urgent task. The target spike proteins of SARS-CoV-2 have been detected experimentally via Raman spectroscopy. However, there lacks high-accuracy theoretical Raman spectra of the spike proteins to as a standard reference for the clinic diagnostic purpose. In this paper, we propose a large fragment method to construct the high-precision Raman spectra for the spike proteins. The large fragment method not only reduces the calculation error but also improves the accuracy of the protein Raman spectra by completely calculating the interactions within the large fragment. The Pearson correlation coefficient of theoretical Raman spectra is greater than 0.929 or more. Compared with the experimental spectra, the characteristic patterns are easily visible. This work provides a detection standard for the spike proteins which shall bring a step closer to the fast recognition of SARS-CoV-2 *via* Raman spectroscopy method.

1. Introduction

During the past several years, COVID-19 has caused great damage to global public health [1,2] and thus calling for the prompt development of rapid and accurate test methods for this lethal virus. Currently, real-time polymer chain reaction (RT-PCR) analysis [3] is the golden standard method for SARS-CoV-2 [4–6] infection test but is costly and time-consuming. It is urgent to develop new methods for rapid and highly sensitive detection of SARS-CoV-2. Among all possible ways, Raman spectroscopy, which is a fast and non-destructive analytic method, has been developed to be a new technique for rapid detection of SARS-CoV-2 [7,8]. Due to a large number of spike proteins [4,9,10] covering the SARS-CoV-2 protein, the signals of spike proteins dominate Raman spectra of SARS-CoV-2 and the Raman spectra of spike proteins can be used to identify of SARS-CoV-2 [8]. Although the Raman spectra of spike protein have been measured experimentally, there lacks high-precision theoretical Raman spectra of spike protein a standard control for the experimental spectra.

The Raman spectra from first-principles calculations[11–13] can provide accurate theoretical spectra, but currently can-not be directly applied to SARS-CoV-2 spike protein due to high computational costs since each peptide chain contains 1208 amino acid residues and more than 7300 non-hydrogen atoms per peptide chain [14]. Alternatively, an useful strategy is the divide-and-conquer method [15,16], where the protein is cut into fragments and the interactions between the fragments can be calculated at relatively a high accurate level (Fig. s1). Considering the computational cost and rate, previous work has been reported within 100 atoms as fragment by using divide-and-conquer method [17–19]. However, interaction between fragments needs further modifications in this method. The interactions between small fragments are calculated by using various approximation methods [17–20] and cannot be fully compensated because of the breaking of chemical bonds. It will be more accurate if larger fragments, for example, more than 400 atoms for each fragment, can be calculated precisely [21]. If the fragment is large enough, ignoring the interaction between fragments shall have little impact on the theoretic Raman spectra.

Here we propose a new method by using large fragments, more than 400 atoms for each fragment, to construct high-precision Raman spectra of spike protein. That is, the spike protein is cut into large fragments, and Raman spectra of spike protein obtained from the sum of these large

* Corresponding author.
E-mail address: mjzhou@ustc.edu (M. Zhou).

https://doi.org/10.1016/j.cplett.2022.139663
Received 11 January 2022; Received in revised form 27 March 2022; Accepted 26 April 2022
Available online 30 April 2022

Fig. 1. Diagram of SARS-CoV-2 spike protein divides into large fragments, each colored region represents a large fragment.

fragments spectra. Compared to previous method within 100 atoms for each fragment, the interaction between fragments accounts for a large proportion of total energy by using small fragment method while the interactions between large fragments accounts for few proportion of total energy. Besides the improvement of accuracy, this method is highly parallel, so that it decrease computational time therefore the whole spike protein Raman spectra can be obtained in short time.

2. Methods

In order to balance precision and efficiency, approximately 25 amino acid residues, more than 400 atoms, are chosen to calculate spike protein Raman spectra. The whole protein is divided into 38 large fragments (Fig. 1) and the spike protein Raman spectra are obtained from the sum of these large fragments spectra. The containing atoms in each large fragment are shown in Table.s1. All fragment structures are based on experimental structure (pdb code: 6vsb) [14].

Since the experimental structure does not contain hydrogen atoms, hydrogen atoms were added to all fragments which are optimized by performing the Gaussian 09 package [22] at the hybrid B3LYP [23] level with 6-31G(d) basis sets. Then, the frequency analysis and Raman calculations are also performed.

The spectra are obtained by the following formula [24]:

$$f(\omega) = \sum_{i=1,n} \frac{f_{1,n}}{(\omega - \Omega_{1,n})^2 + \gamma^2} \tag{1}$$

where $f_{1,n}$, $\Omega_{1,n}$ represent the intensity and frequency of the oscillator, and γ represents the spreading width. For comparison with the experimental resutlts, γ parameter is set to be 4 cm^{-1} and Infrared spectra are calculated at the same theoretical calculation level.

To analyze the precision of large fragments method, we used Pearson correlation coefficient (r) as an evaluation parameter, the Pearson correlation coefficient is defined as:

$$r = \frac{\sum_{i=1,n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1,n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1,n}(y_i - \bar{y})^2}} \tag{2}$$

where, n *is* the number of samples and x_i, y_i are two variables. The greater the r is, the stronger the correlation between two variables. In our calculation, the Pearson correlation coefficients of the two curves are calculated in the range of 500–2000 cm^{-1} respectively. 1714 points are taken evenly on the curve as samples in the range of 500–2000 cm^{-1}.

3. Results and discussion

In order to evaluate the precision of our method, a fragment with 25 amino acid residues is chosen, as shown in Fig. 2, which is labeled as frag-(1), and it is divided into smaller fragments with different sizes. Fig. 2 (b) shows that the frag-(1) is divided into two pieces with

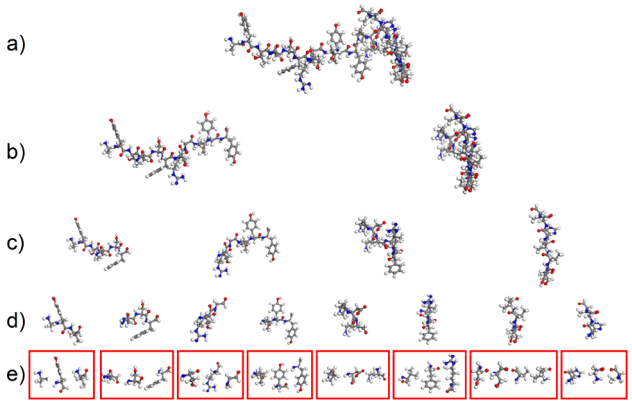

Fig. 2. Schematic diagram of large fragment of spike protein divided into small fragments with different sizes. The average residue number of fragments is a) 25.0; b) 12.5; c) 6.25; d) 3.125; e) 1.0.

complete amino acid residue and roughly equal atom numbers, named frag-(1, 1) and frag-(1, 2). In the same way, each fragment in Fig. 2 (b) can be divided into two pieces, labeled as frag-(1, 1, 1), frag-(1, 1, 2), frag-(1, 2, 1), frag-(1, 2, 2). Similarly, each fragment in Fig. 2 (c) can be divided into two pieces, named frag-(1, 1, 1, 1), frag-(1, 1, 1, 2), frag-(1, 1, 2, 1), frag-(1, 1, 2, 2), frag-(1, 2, 1, 1), frag-(1, 2, 1, 2), frag-(1, 2, 2, 1), frag-(1, 2, 2, 2). As shown in Fig. 2 (e), each fragment in Fig. 2 (d) is divided into one amino acid residue, labeled as frag-(1, 1, 1, 1, 1), frag-(1, 1, 1, 1, 2), frag-(1, 1, 1, 1, 3) and so on. The containing non-hydrogen atoms in each fragment are shown in Table.s2.

It was assumed that the final target protein, frag-(1), is composed of the smallest fragments step-by-step (Fig. 2 (e)-Fig. 2 (a)). We have firstly calculated the Raman spectra of frag-(1, 1, 1, 1) and the sum of Raman spectra of frag-(1, 1, 1, 1, 1), frag-(1, 1, 1, 1, 2), frag-(1, 1, 1, 1, 3). The

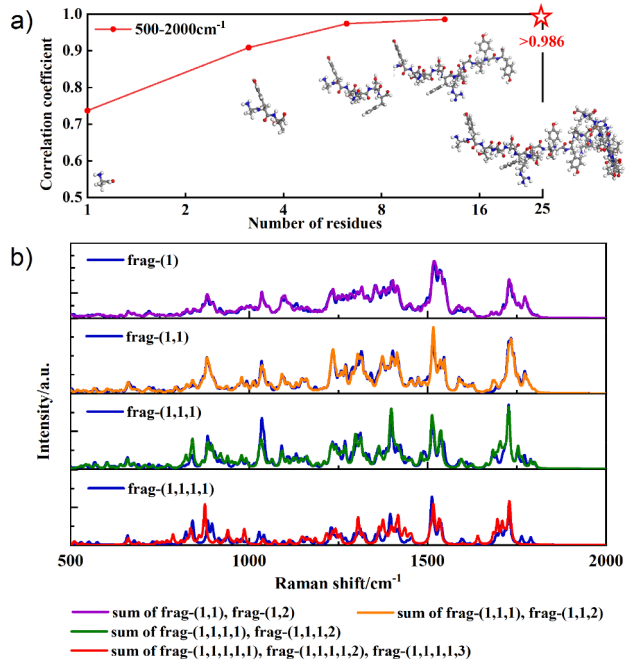

Fig. 3. The precision of each step that constructing frag-(1) spectra from small fragments. a) Relationship between Pearson correlation coefficients and average residue number of fragments. b) The Raman spectra of frag-(1), frag-(1, 1), frag-(1, 1, 1), frag-(1, 1, 1, 1) (blue curve) and the sum of Raman spectra of smaller fragments consist of them (violet, orange, olive and red curve) (wavenumber range, 500–2000 cm^{-1}). In all structures, white dots, blue dots, red dots and gray dots represent H atoms, N atoms, O atoms and C atoms, respectively.

comparison of the two spectra is shown in Fig. 3 (b) (red curve). The spectra shown in the figure have many peaks with different peak positions. The Pearson correlation coefficient is only 0.737, which indicate that the two spectra is poorly correlated. Then we have calculated the Raman spectra of frag-(1, 1, 1) and the sum of Raman spectra of frag-(1, 1, 1, 1), frag-(1, 1, 1, 2). As can be seen from the figure (olive curve), there are some peaks within 1600–1800 cm^{-1} have different intensities. The Pearson correlation coefficient is 0.909, which is improved considerably albeit is still not good enough. Lastly, we have calculated the Raman spectra of frag-(1, 1), frag-(1) and the sum of Raman spectra of corresponding smaller fragments (orange and violet curve). The Pearson correlation coefficient now is 0.974 and 0.986 respectively, which indicate that the two sum Raman spectra have a much higher precision.

The relationship between Pearson correlation coefficients and average residue number of fragments are summarized in Fig. 3 (a) which compares Raman spectra of frag-(1), frag-(1, 1), frag-(1, 1, 1), frag-(1, 1, 1, 1) with the sum of Raman spectra of smaller fragments. As shown in the figure, the larger the fragments provides the higher the Pearson correlation coefficients. We have also tested the Raman spectra of frag-(1), frag-(1, 2), frag-(1,2, 2), frag-(1, 2, 2, 2) with the sum of Raman spectra of smaller fragments. Similar conclusions can be seen from Fig. s2. Thereby, the Pearson correlation coefficient is expected to be greater than 0.986 when the average residues number of fragments is larger than 25 (Fig. 3 (a) red star).

It is important to estimate the precision of constructing spectra with this step-by-step method. Multiplying the Pearson correlation coefficients of each step to measure the overall precision, the precision of constructing frag-(1) using smaller fragments (Fig. 2 (e)-Fig. 2 (b)) is approximate 0.737*0.909*0.974*0.986 = 0.643, 0.909*0.974*0.986 = 0.873, 0.974*0.986 = 0.960, and 0.986, respectively. The higher precision of constructing frag-(1) using large fragments stems from precision improvement of each step and a decrease of step numbers. In the similar process, the precision of constructing spike protein spectra with this method have been estimated and the estimate precision is greater than 0.986 $\exp(\log_2 38) = 0.929$, as the whole spike protein is divided into 38 fragments.

The true precision of constructing frag-(1) directly with fragments of different size is estimated and the comparison between the sum spectra and the frag-(1) spectra is shown in Fig. 4 (b)-Fig. 4 (d). We have calculated the sum spectra of fragments in Fig. 2. (b), Fig. 2. (c), Fig. 2. (d) and Fig. 2. (e), separately. As shown in Fig. 4 (b) (red curve), the two curves differ a lot within 1600–1800 cm^{-1}. The Pearson correlation coefficient is 0.772, being greater than the estimate Pearson correlation coefficient (0.643) of the step-by-step method (Fig. 3. (a)). This can be rationalized by the fact that with the increase of protein atoms, the peaks become dense, and the difference of peak position is overweighed by the difference of peak intensity. As shown in Fig. 4. (b) (olive curve), the Pearson correlation coefficient of the two curves is 0.936, being also greater than the estimate Pearson correlation coefficient (0.873) of the step-by-step method. The same result can be seen in Fig. 4 (b) (orange curve).

The comparison between the estimate precision and the true precision of constructing frag-(1) from smaller fragments is shown in Fig. 5. The sizes of the fragments are illustrated by the structures in the figures. From the figure, we can see that all estimate precision is lower than the true precision except the last point which represents the same process. Based on this result, one can infer that the precision of constructing spike protein spectra is underestimated. The Pearson correlation coefficient of constructing spike protein spectra should be greater than 0.929.

Following the above large fragments method, the theoretical spectrum of SARS-CoV-2 spike protein can be constructed easily and quickly. The whole protein is divided into 38 large fragments (Fig. 1), and the spectra (Fig. 6) are constructed by summing the spectra of these large fragments. In order to compare with our experimental Raman spectra [7], we have normalized them and then compared the two spectra using

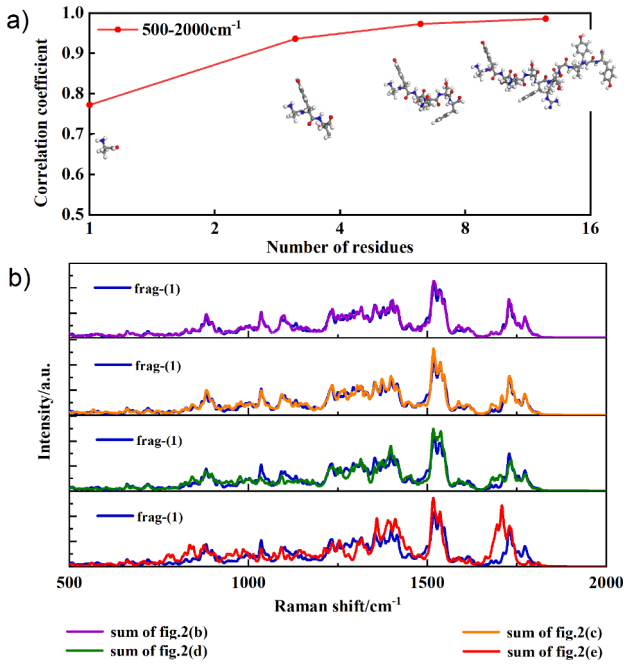

Fig. 4. The true precision of constructing frag-(1) spectra from smaller fragments directly. a) Relationship between Pearson correlation coefficients and average residue number of fragments. b) The Raman spectra of frag-(1) (blue curve) and the sum of Raman spectra of fragments of different sizes of which the average residue number is 12.5 (violet curve), 6.25 (orange curve), 3.125 (olive curve), 1.0 (red curve) (wavenumber range, 500–2000 cm-1), respectively. In all structures, white dots, blue dots, red dots and gray dots represent H atoms, N atoms, O atoms and C atoms, respectively.

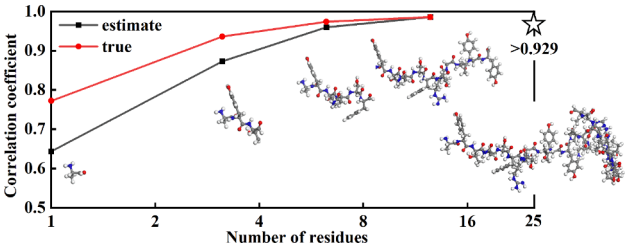

Fig. 5. The comparison between the estimate precision and the true precision of constructing frag-(1) from smaller fragments. In all structures, white dots, blue dots, red dots and grey dots represent H atoms, N atoms, O atoms and C atoms, respectively.

a correction factor of 0.96, It can be seen from the Fig. 6 that that the main Raman bands of both spectra match well with each other. For better comparison purpose, theoretical IR spectra have also been calculated in Fig. 6. Compared with our experimental spectra, characteristic patterns are maintained in the theoretical ones although the intensities may have some differences. The main characteristic bands of Spike protein such as 1655 cm^{-1} (amide I), 1616 cm^{-1} (Tryptophan), 1552 cm^{-1} (amide II), 1449 cm^{-1} (CH_2 bending vibration), 1323 cm^{-1} (CH deformation vibration), 1240 cm^{-1} (amide III), 1003 cm^{-1}(Phenylalanine), 642 cm^{-1} (C-S stretch vibration), 621 cm^{-1}(amide IV) are in a good agreement with experimental ones. The detailed summarization of these bands can be seen in Table.s3. Based on these results, all the theoretical results are well-supported by the experimental ones and the large fragment method can be used to construct theoretical Raman spectra with clear characteristic patterns.

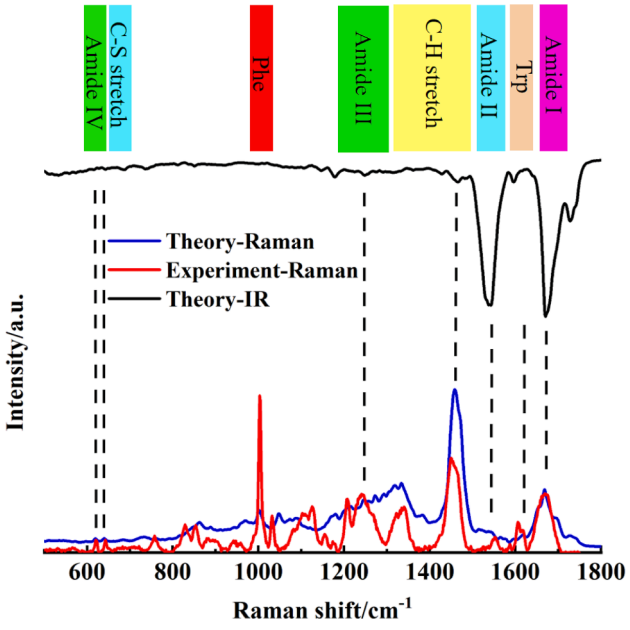

Fig. 6. Comparison between the DFT calculated Raman and IR spectra using large fragments method and experimental Raman spectra of SARS-CoV-2 spike protein.

4. Conclusion

In summary, we propose a method for constructing high-precision protein Raman spectra based on large fragments method and evaluate the precision of this method. Based on the large fragments method, we theoretically constructs a high-precision standard Raman spectrum of SARS-CoV-2 spike protein with clear characteristic patterns. Based on this method and the spectra, we can further simulate the Raman spectra of actual detection target and environment theoretically. Combining with machine learning algorithm, we can simulate the process of detection of spike protein theoretically and then explain the physical basis of the machine learning determination point and spike protein detection limit. The work in this paper will promote the development of SARS-CoV-2 Raman recognition technology.

CRediT authorship contribution statement

Shuang Ni: Conceptualization, Methodology. **Qiang Yang:** Data curation. **Jinling Huang:** Data curation. **Minjie Zhou:** Formal analysis, Writing – original draft. **Lai Wei:** Data curation. **Yue Yang:** Data curation. **Jiaxin Wen:** Visualization. **Wenbo Mo:** Visualization. **Wei Le:** Investigation. **Daojian Qi:** Investigation. **Lei Jin:** Investigation. **Bo Li:** Project administration. **Zongqin Zhao:** Project administration. **Kai Du:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially supported by the National Key Reserch and Development Program of China (2017YFA0206001) and the National Natural Science Foundation of China (No.11805176). Thanks Dr. Lei for discussion.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cplett.2022.139663.

References

[1] A.S. Al-Mandhari, R.J. Brennan, A. Abubakar, R. Hajjeh, Lancet 396 (2020) 1786–1788.
[2] J. Silver, T.C. Smith, C. Wenham, et al., Lancet 396 (2020) 1800–1801.
[3] V.M. Corman, O. Landt, M. Kaiser, et al., Eurosurveillance 25 (2020). No. 2000045.
[4] F. Wu, S. Zhao, B. Yu, et al., Nature 579 (2020) 265–269.
[5] Q. Wang, Y. Zhang, L. Wu, et al., Cell 181 (2020) 894–904.
[6] H. Yao, Y. Song, Y. Chen, et al., Cell 183 730–738 (2020), e13.
[7] J. Huang, J. Wen, M. Zhou, et al., Anal. Chem. 93 (2021) 9174–9182.
[8] Y. Yang, Y. Peng, C. Lin, et al., Nano-Micro Lett. 13 (2021) 109.
[9] Z. Ke, J. Oton, K. Qu, et al., Nature 588 (2020) 498–502.
[10] B. Turonová, M. Sikora, C. Schürmann, et al., Science 370 (2020) 203–208.
[11] H. Ren, J.D. Biggs, S. Mukamel, J. Raman Spectrosc. 44 (2013) 544–559.
[12] B. Tian, S. Li, S. Lei, et al., Chin. Chem. Lett. 32 (2021) 2469–2473.
[13] H. Ren, Z. Wang, S. Guo, et al., J. Chem. Phys. 155 (2021), 174301.
[14] D. Wrapp, N. Wang, K.S. Corbett, et al., Science 367 (2020) 1260–1263.
[15] W. Yang, Phys. Rev. Lett. 66 (1991) 1438.
[16] W. Yang, T. Lee, J. Chem. Phys. 103 (1995) 5674–5678.
[17] H. Torii, M. Tasumi, J. Chem. Phys. 96 (1992) 3379.
[18] J. Kessler, J. Kapitán, P. Bour, J. Phys. Chem. Lett. 6 (2015) 3314–3319.
[19] F.S. Husseini, D. Robinson, N.T. Hunt, et al., J. Comput. Chem. 38 (2017) 1362–1375.
[20] S. Ye, K. Zhong, J. Zhang, et al., J. Am. Chem. Soc. 142 (2020) 19071–19077.
[21] D.J. Cole, N.D.M. Hine, J. Phys-Condens, Mat. 28 (2016), 393001.
[22] M.J. Frisch, G.-W. Trucks, H.B. Schlegel, et al., Gaussian 16. Revision A 3 (2016).
[23] P.J. Stephens, F.-J. Devlin, C.F. Chabalowski, et al., j. Phys. Chem. 98 (1994) 11623–11627.
[24] S. Ye, H.u. Wei, X. Lia, et al., PNAS 116 (2019) 11612–11617.