A Reliable Computational Method for Obtaining the Conformational Ensemble of Tripeptide*

LU Xiya, RU Xiao, LIN Zijing†

Department of Physics, University of Science and Technology of China, Hefei 230026

Received date: 2023-08-29; accepted date: 2023-09-04

【Abstract】 Tripeptide is a basic building block of proteins and has important physiological functions. Analyzing the structure is of high significance for the study of larger peptide/protein systems. An *ab initio* method for tripeptide structure prediction based on the combination rule of dihedral angles combined with bond rotations is developed. The method is used to conduct thorough searches on the potential energy surfaces of 8 representative tripeptides. The results are compared with a few previous methods and that deduced from the Protein Data Bank (PDB). The comparison shows that the new method provides the most complete conformational ensembles of the tripeptides. It is also demonstrated that the results from the PDB are unreliable and miss a large portion of conformations in the low and middle energy regions. The newly obtained tripeptide conformations are further used in IR spectra study and provide a much improved agreement with the experimental data.

Keywords: Peptide structure prediction, Tripeptide, Quantum chemical calculation, Conformational ensemble, IR spectrum

PACS: 87.14.ef, 87.15.ag, 87.15.bd

DOI: 10.13380/j.ltpl.2023.04.001

Reference method: LU Xiya, RU Xiao, LIN Zijing, Low. Temp. Phys. Lett. **45**, 0183 (2023)

一种获取三肽构象系综的可靠量子化学方法*

陆茜雅，汝啸，林子敬†

中国科学技术大学物理系，合肥 230026

收稿日期：2023-08-29；接收日期：2023-09-04

【摘要】 三肽是蛋白质的基本组成模块，具有重要生理功能. 解析其结构对于更大的肽及蛋白质研究具有重要意义. 基于二面角组合规则，结合键旋转手段，提出一种获取三肽构象系综的从头算量子化学方法. 使用该方法对八个目标三肽的势能面进行彻底搜索，将所得结果与以前的预测方法及蛋白质数据库(PDB)提取的结构进行比较. 结果表明，新方法搜索到了最完整的三肽构象系综. 此外证明了 PDB 结构存在缺陷，遗漏了大部分中低能区的重要构象. 将新获取的三肽结构用于红外光谱研究，理论结果与实验数据符合得更精准.

关键词：肽结构预测，三钛，量子化学计算，构象系综，红外光谱

PACS: 87.14.ef, 87.15.ag, 87.15.bd

DOI: 10.13380/j.ltpl.2023.04.001

* Project supportedby the National Natural Science Foundation of China (Grant Nos. 12074362&12374017).

† zjlin@ustc.edu.cn

1 Introduction

Peptides are important biomoleculesinvolved in various fields of human hormones, nerves, cell growth and reproduction. It can transmit information between various systems and cells, participate in the regulation of various physiological functions, and assist in protein localization and functional characterization, such as signal peptide[1], vasoactive intestinal peptide[2], binding peptide[3], calcitonin Peptide[4], and so on. As important human physiological regulators, peptide drugs such as antibacterial peptides[5] and antibiotic peptides[6] with antiviral and anticancer activities are widely used in the field of disease diagnosis and treatment. Compared with small molecule drug candidates, peptides have the advantages of low toxicity and strong binding specificity[3], and are important sources of drugs.

The activity and function of peptideare are determined by its three-dimensional structure. Analyzing the peptide structure is the basis for further research and application of peptides. Experimental methods such as X-ray diffraction are not suitable for oligopeptide systems, not only because of the high cost but also due to the difficulty of sample preparation. It has become an inevitable choice to resort to more convenient and faster computational methods for peptide structure determination. However, there is no theoretical method to perfectly solve the problem of peptide structure prediction. For instance, the systematic search method is reliable, but the amount of computations increases exponentially with the increased size of peptide. Stochastic methods such as Monte Carlo can be used for large systems, but the quality of the search results is quite limited. Molecular dynamics simulation is quite dependent on the selection and accuracy of the force field potential function[7], and the simulation time will skyrocket with the increase of the system[8]. Machine learning (ML) methods such as AlphaFold2 are quite powerful, but the prediction is often based on the co-evolution information in structural databases such as the PDB[9]. When the homologous structures in the database are lacking, the accuracy of the ML results cannot be guaranteed[10]. That is the case for most of the highly recognized and widely used peptide structure research methods[11-14]. Moreover, there are defects in the structures of PDB[15], such as the loss of side chain information and the difference in resolution, which will lead to the deterioration of relevant theoretical model parameters[16].

Notice that peptide generally appears as a multi-conformation ensemble. The conformational ensemble has direct significance in molecular docking and experimental prediction. For example, the active peptide conformation is usually not the lowest energy conformation. A conformational ensemble is required in the set docking method[17] to solve the problem of molecular docking flexibility. Furthermore, the ensemble combined with the Boltzmann distribution can assist in the analysis and prediction of experimental results such as NMR spectra[18].

It has been shown that using 3-9 residues as small fragments for assembly[19,20] has the best structural correlation with the protein structure. The active sites of interest in disease treatment are usually three or four residues[21], which play important functions in physiological activities such as recognizing binding targets and activating receptors. It can then be seen that, as the basic fragment for the transition from amino acid to protein, the tripeptide is currently the most diverse oligopeptide with important research and application value. Consequently, this study focuses on the structures of tripeptide.

Considering the importance of tripeptide structures, a new method to thoroughly search the conformational ensemble of tripeptides is developed here. The potential energy surfaces of the eighttripeptides were thoroughly searched

through high precision calculation and the obtained conformations were compared with that of previous methods and PDB. The theoretical and experimental infrared spectra of the other five tripeptides were further compared to verify the reliability of the new method. Moreover, the HOMO-LUMO energy gaps for the tripeptide conformations are calculated and added to the molecular orbital databases to facilitate catalyst search, molecular assembly docking, peptide drug design and related machine learning training.

2 Method

2.1 Construction of the tripeptide conformations

The acquisition of the tripeptide structures is completed by splicing amino acid and dipeptide low energy conformations. Similarly, the dipeptide structure ensemble is obtained by splicing two amino acid structure ensembles[22]. The splicing cuts the increase of the computational cost from exponential to linear and greatly improves the calculation efficiency[23].

To obtain the high precision conformational ensemble of amino acids, the initial structure was generated by the law of dihedral combination. Then the single-point energy was calculated at the B97-D3/6-31G *[24-27] level after DFTB[28,29] optimization. The structure of the first 10 kcal/mol was extracted for main Chain dihedral rotation and finally optimized at B97-D3/6-31G * level. Repeat the aforementioned rotation and interception optimization steps one by one for the main chain dihedral angles to obtain the final conformation. The law of dihedral combination[30] is extracted and summarized through a large number of calculations and tests which can eliminate unnecessary combinations of dihedral angles in the initial structure and reduce the sampling PES space. Therefore, the entire PES space is divided into subspaces reflecting the basic characteristics of low-energy conformation sand the sampling efficiency of the system is improved. For a single PES subspace, bond rotation thoroughly searches all possible combinations of rotational degrees of freedom to ensure the quality of the structure. Bonding rotation is an effective means to supplement important conformations[31,32]. The rotation of the specific dihedral angle is according to the summarized rules, for φ with 3 degrees of freedom and Ψ with 4. Geometry optimization can form new interactions between residues, resulting in new low-energy distributions.

The dipeptide conformational ensemble is obtained by splicing the amino acid ensemble obtained by the above process. The first 14 kcal/mol conformations of the two amino acids are directly connected to obtain the first batch of tentative conformations. The connection follows the principle of dehydration condensation. The amino nitrogen atom at the C-terminal covers the position of the hydroxyl oxygen atom at the N-terminal. The N—H of the newly formed peptide bond bisects the original amino H—N—H angle at the C-terminus. Then optimize the tentative conformations, calculate the single-point energy and intercept the low-energy conformations for bond rotation. In addition to the main chain, it needs to rotate the Cα—C bond at the N-terminal junction to deal with the isomers of the C7eq/C7ax structure. In the second batch of tentative conformations obtained by rotation, the first 12 kcal/mol conformations were selected for optimization. The additional important conformers were generated which were combined with the first batch of optimization results to obtain the final dipeptide ensemble. Each bond rotates with 3 degrees of freedom and the rotation of the side chain does not exceed 4. Since the low-energy isomers of amino acids are used to generate initial structures, the rotational degrees of bonds are reduced and the search space is compressed.

Similarly, tripeptide can be obtained by splicing dipeptide and single amino acid with bond rotation. We calculated Alanine-Asparticacid-Ala-

nine (Ala-Asp-Ala, or ADA), Asparticacid-Alanine-Asparticacid (Asp-Ala-Asp, or DAD), Serine-Valine-Serine (Ser-Val-Ser, or SVS), Valine-Serine-Valine (Val-Ser-Val, or VSV), Asparagine-Asparagine-Asparagine (Asn-Asn-Asn, or NNN), Phenylalanine-Phenylalanine-Phenylalanine (Phe-Phe-Phe, FFF), Asparagine-Glutarnine-Asparagine (Asn-Gln-Asn, or NQN), Glutarnine-Asparagine-Glutarnine (Gln-Asn-Gln, or QNQ) eight tripeptide conformations, and obtained their conformational ensembles. Each tripeptide is spliced in two ways. For example, Ala-Asp-Ala using the dipeptide Ala-Asp and Ala splicing; Ala and Asp-Ala splicing. See Figure 1 for details.

The results show that there is no obvious advantage or disadvantage between the two ways, both of which have searched for new important low-energy conformations. It is worth noting that the structure loss rate in the low-to-medium energy region is relatively high with only one splicing way. Therefore, to obtain more comprehensive and multiple conformations, the results of the two splicing ways should be combined as the final ensemble of the tripeptide. This method for tripeptide structure prediction is referred to below as the comprehensive segment assembly (CSA).

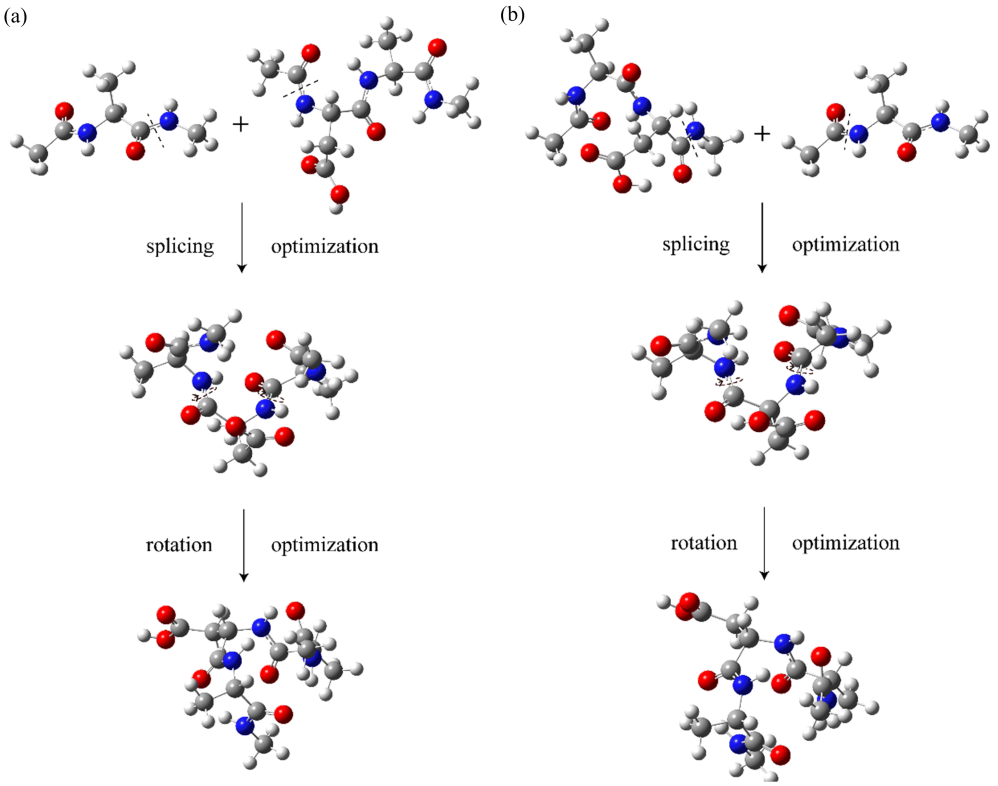

Figure 1 The specific splicing process of the tripeptide. Take Ala-Asp-Ala as an example. The splicing process of (a) Ala and Asp-Ala and (b) for Ala-Asp and Ala. The connection part is marked with a black dotted line.

2.2 Calculation of vibration frequency

Using the CSA method to generate the conformational ensemble of Tryptophan-Glycine-Glycine (Trp-Gly-Gly, or WGG), Phenylalanine-Glycine-Glycine (Phe-Gly-Gly, or FGG), Alanine-Alanine-Phenylalanine (Ala-Ala-Phe, or AAF), Alanine-Phenylalanine-Alanine (Ala-Phe-Ala, or AFA), Phenylalanine-Alanine-Alanine (Phe-Ala-Ala, or FAA), sorted after calculating the Gibbs free energy. The obtained vibrational spectrum was compared with the experiment and other computations. All vibration analyses were per-

formed using ORCA at a temperature of 298.15 K and standard atmospheric pressure. The calculated spectra are represented by the Lorentz function, and the FWHM broadening is taken as 8 cm^{-1}.

Due to the excessive number of conformers, a hierarchical energy cutoff was taken. Firstly we selected the first 10 kcal/mol conformers for DFTB optimization and then chose the first 8 kcal/mol to optimize at the RI-B3LYP-D3/6-31+G* level[33]. Single-point energy was calculated at the RI-B3LYP-D3/6-311+G** [34] level, and the first 4.5 kcal /mol conformations were intercepted for vibrational analysis at RI-B3LYP-D3/def2-TZVP level[35]. The RI auxiliary basis is def2[36]. According to different functional basis sets, different scaling factors are used to scale the obtained frequencies. The scale factor is obtained from the ratio of the theoretical and experimental peaks[37], such as the terminal hydroxyl group of the neutral system FGG that does not participate in the bond.

All conformations in this research were generated by our inner developed program and calculated by Guassian09[38] and ORCA4.0[39] software package. The conformations are put into the PPbank database (https://ppbank.snquantum.com/#/index) which we exploit for oligopeptide.

3 Results and Discussion

The selected and calculated amino acids A, D, S, V, N, F, Q belong to 20 common amino acids, covering all basic types of amino acids, to verify whether the new method is universal. Select tripeptides of different types and complexities, search the potential energy surfaces of the four groups of tripeptides ADA, DAD, SVS, VSV, NNN, FFF, NQN, QNQ, and compare the conformations with the PDB database. Then five tripeptides were calculated and the infrared spectra were compared with the experiment and other calculation results to further verify the quality of the obtained conformational ensemble. To ensure the objective and accurate comparison of results, the calculation software, optimization method, function and basis set used are completely consistent.

3.1 Comparison with PDB

To verify the quality of the obtained ensemble, all the structural information including the selected tripeptide fragments were extracted from the currently most recognized PDB database (185,541 structures for 2021 year), and the results were compared after unified optimization using the same process. The specific numbers of conformation are shown in Figure 2.

For the characteristics of the active conformation, we expanded the energy range of the ensemble. Affinity is an important index to measure molecular interactions. It can be seen from the formula $\Delta G = -1.42 \log K_i$ that if the energy deviation between the compound conformer and the lowest energy conformation reaches 1.42 kcal/mol, the affinity of the compound will decrease by an order of magnitude. So the molecular conformation with high affinity tends to have lower energy[40]. Considering that the active peptide conformation is usually not the lowest energy conformation, further expanding the upper limit of the energy cutoff of the conformational ensemble, from 3 to 8 kcal/mol[15,41,42]. The promotion can in principle include almost all active conformations. Combined with the ensemble docking method[17], it provides more possibilities to solve the problem of molecular docking flexibility.

It can be seen from Figure 2 that PDB performs best for ADA and DAD groups and has a new conformer at 2—3 kcal/mol. But it is still inferior to CSA in the low-energy range and the disadvantages in the middle-high energy range are intensified. The number of low-energy conformations in the PDB of the SVS and VSV groups is small, and the number of new conformations searched by CSA in the low and middle energy range is much higher than that of PDB. FFF carries a benzene ring. It may be due to the

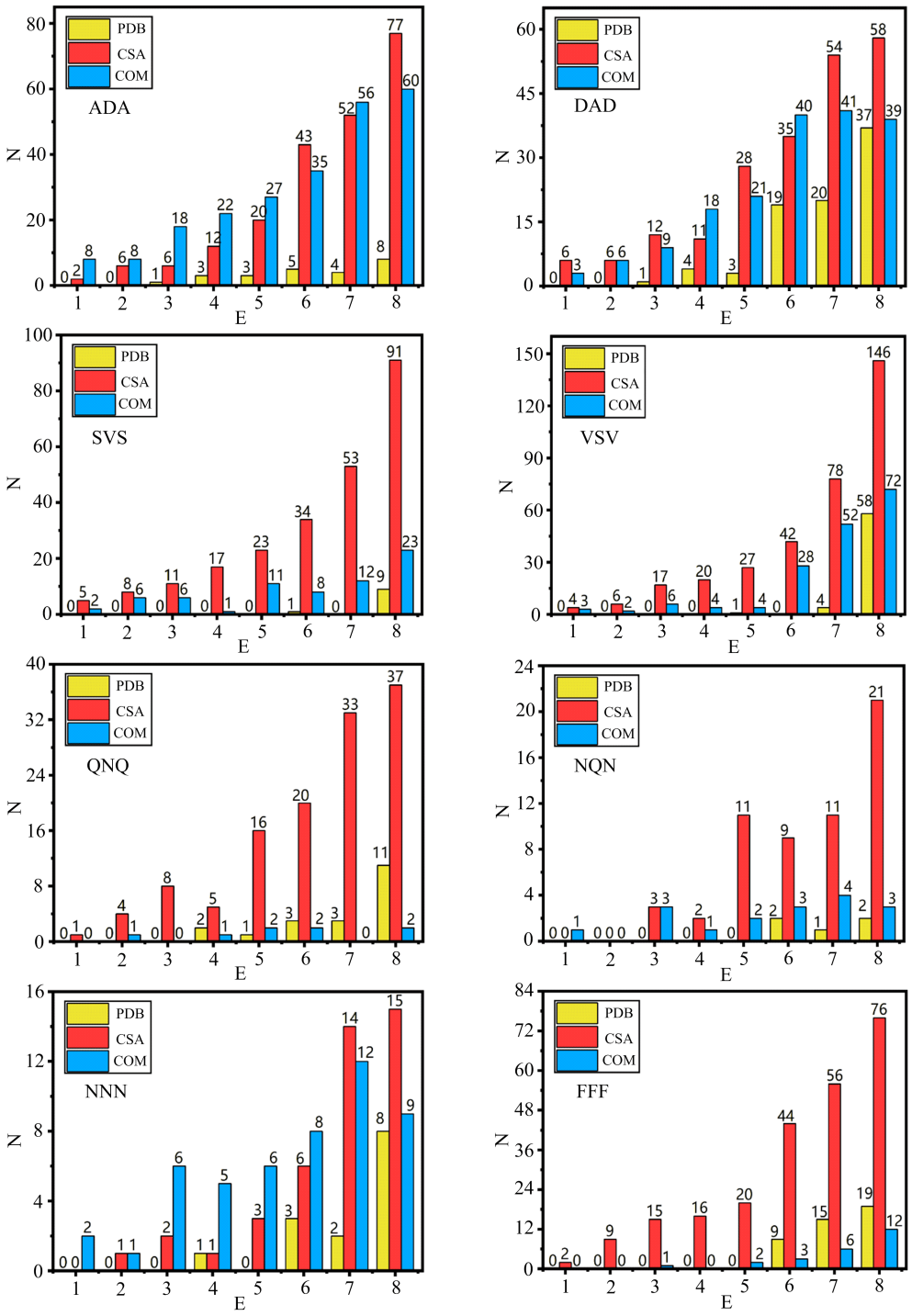

Figure 2 Comparison of the number of CSA and PDB conformers. The four groups of tripeptides correspond to ADA, DAD, SVS, VSV, QNQ, NQN, NNN and FFF from left to right. The abscissa E represents energy, in unit of kcal/mol; the ordinate N is the number of conformers. The red column represents the number of conformations newly found by the CSA method but not in the PDB; the yellow represents the unique conformation number of the PDB and the blue is the number of commonly searched conformers.

lack of substituents on the benzene ring which is inconvenient to analyze and determine, and the mutual repulsion stability between the benzene rings is low. There are only a few conformers of FFF in PDB and the lowest energy conformation differs from CSA by 3 kcal/mol. So it is speculated

that there is still large room for increasing the amount of PDB data for the aromatic. For the three conformations of SVS, VSV, NNN, FFF, NQN and QNQ in the interval of 0—3 kcal/mol, CSA not only accurately found all the conformers in the PDB, but also found many new conformations. The new method has obvious advantages in FFF, NQN, QNQ, SVS, VSV and the number of low-energy conformers directly increases by an order of magnitude.

To explore the impact of the existing structure of the database on the relevant results, the performance of the data in the PDB is discussed separately. Comparing the quality of tripeptides in the PDB, it was found that the ADA and DAD performed best. It is reasonable to speculate from Table 1 that due to the low complexity of these two fragments, the charged polar side chain in aspartic acid has multiple functions and is easy to determine, and it can form hydrogen bonds with various nitrogen and oxygen atoms and is relatively stable. A large amount of structural data covers more important characteristic conformations, and the ensemble is relatively more complete. There are also many SVS and VSV conformations, but due to the external hydrophobic residues of VSV, most of them are concentrated in the higher energy region, resulting in poor performance. The number of conformers in the remaining four systems is less and the comparison results are worse. So more important conformers need to be supplemented. The correlation between the prediction results of conformers and the amount of data and the properties of specific fragments is instructive for the method using the database information.

In summary, for the selected four groups of tripeptides, whether in the low or middle energy region, the number of newly searched conformations is far better than that of PDB. The advantage of the low energy interval is particularly obvious, not only covering all the conformations in the PDB but also supplementing a large number of important low-energy conformations that may be missed. For related methods that rely on databases, a richer and more accurate ensemble of conformers can improve the quality of results to a certain extent. In the fields of enzyme catalysis research, drug design combined with machine learning and structure prediction[43-45], the CSA method, which is not limited by experiments, provides a powerful help for various database-based methods.

Table 1 Number of specified fragments in the PDB database.

Seq	PDB num
ADA	54101
DAD	34580
SVS	42071
VSV	41915
NNN	11291
FFF	6466
NQN	8273
QNQ	7322

3.2 IR spectra comparison

3.2.1 WGG

Figure 3 shows the important conformers of WGG. The N—H···O hydrogen bond formed between the hydroxyl oxygen of the terminal carboxyl group and the nitrogen-containing ring in the tryptophan side chain greatly limits the freedom of the main chain. The low-energy conformation follows the close structure of end-to-end. The vibrational calculations show that the theoretical spectra of wgg01 and wgg10 are in good agreement with the experiments. The specific vibrations can be assigned to the carboxyl C=O stretching (CO_{carb}), the peptide C=O stretching (CO_{pep}), the peptide N—H in-plane bending (NH_{ipb}) and the carboxylic O—H in-plane bending (OH_{ipb}) vibrations. Figure 4 gives more details.

Compared with the calculated spectrum of Cerný[46], the number of important low-energy conformations of WGG generated by the new method is more. It is remarkable that the experi-

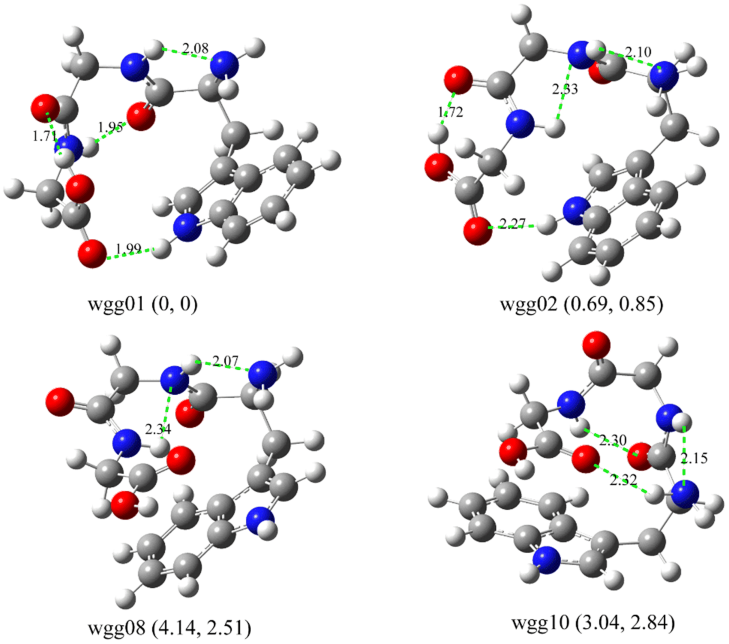

Figure 3 The structure and hydrogen bond information of the important low free-energy conformers of WGG. It is numbered according to Gibbs free energy. Relative electronic energy and free energy(in kcal/mol) are given in parentheses respectively. The green dotted line represents the hydrogen bond. The bond length information is attached and the unit is Å. The other conformational diagrams below have the same meaning.

mental spectrum completely corresponds to the conformation with the lowest free energy ranking and the peak position is more accurate. See Table 2 for details. For the WGG, the obtained ensemble is more complete and accurate, giving more important potential conformations. Furthermore, wgg02 corresponds to the most stable conformer in W. Yu[23]. Our newly obtained wgg01 has one more hydrogen bond, a tighter combination and a lower energy of 0.69 kcal/mol.

3.2.2 FGG

The calculation steps of FGG are the same as those of WGG. Since phenylalanine contains a benzene ring, the temperature has a great influence on energy. Therefore the energy cut-off range is appropriately relaxed, and the first 5 kcal/mol conformations are intercepted for vibration analysis after single-point energy calculation. Figure 5 shows the important conformation and hydrogen bond information of FGG. Figure 6 shows the vibrational spectrum comparison.

The hydroxyl group in other conformers

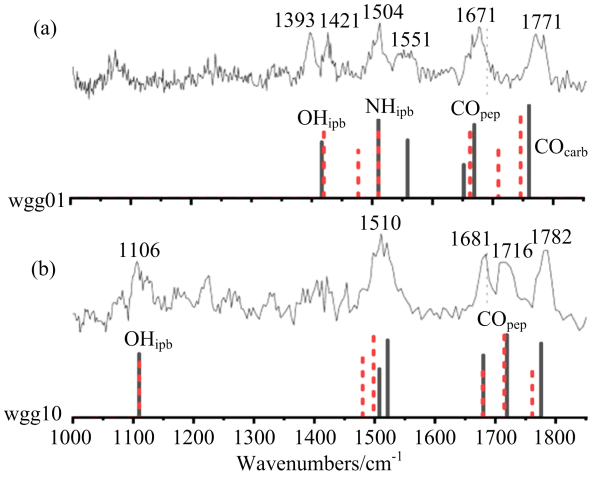

Figure 4 Comparison of WGG infrared spectra with experiments[47] and other theoretical calculations. The abscissa is wavenumber (cm^{-1}). The upper curve is the experimental spectrum. The black line represents our calculation results, and the red dotted line represents the previous calculation results[46]. Subsequent figures have the same meaning. The conformer wgg01 corresponds to the first experimental spectral conformation and wgg10 corresponds to the second. Scale uniformly using a scale factor of 0.979.

forms a strong hydrogen bond with the second peptide bond oxygen, which is inconsistent with the experimentally observed structure. From the bond length, it can be seen that the N1—H···N formed by the peptide bond N and the N atom of the amino group of the side chain of phenylalanine is the most compact, and the N—H ··· O formed by the carbonyl oxygen of the terminal carboxylic acid and the amino group is relatively loose. The hydrogen bonds of fgg06 between the peptide bonds are especially strong, so even though the benzene ring is stretched, its energy is still low. Unexpectedly, no interaction with the phenyl ring π was observed significantly in any of the low-energy important conformations. The fgg01 was also not observed in the experiment. The folding degree of the conformational main chain mainly depends on N—H···O and the N2-H ··· O formed between the peptide bonds inside the molecule, especially the N—H··· O connecting the head and tail carboxyl and amino groups. Generally, the tighter the benzene ring surrounds the main chain, the lower the conformation has energy.

Table 2 Spectroscopic IR data of the mainconformers of WGG.

Conformer	OH_{ipb}	NH_{ipb}	NH_{ipb}	CO_{pep}	CO_{pep}	CO_{carb}	RMSE
exp(a)	1421	1504	1551	—	1671	1771	—
wgg01	1416	1510	1559	1651	1669	1760	7.07
ref[46]	1420	1477	1510	1662	1709	1746	29.93
exp(b)	1106	—	1510	1681	1716	1782	—
wgg10	1110	1508	1521	1680	1719	1776	6.05
ref[46]	1110	1480	1498	1679	1715	1761	11.01

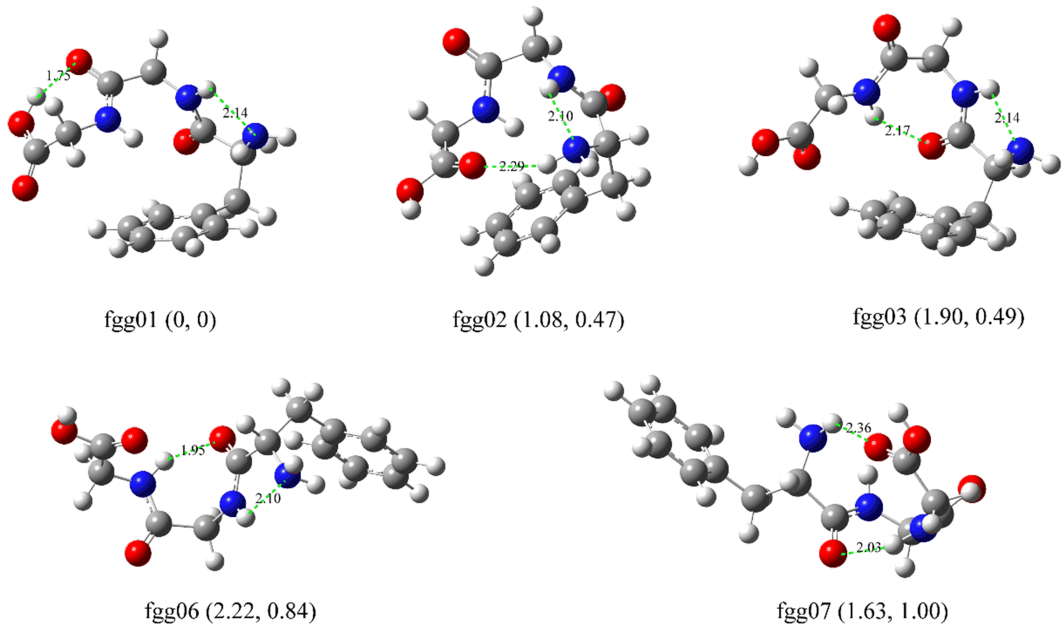

Figure 5 Important low-energy conformation and hydrogen bond information of FGG.

Figure 6 shows that the calculated spectrum of FGG is consistent with the experiment. The hydroxyl groups of carboxylic acids do not interact with other groups, so the stretching peak p(OH) is located around 3600 cm^{-1}. The vibrational stretching peaks p(N1H) in N1—H ··· N and p

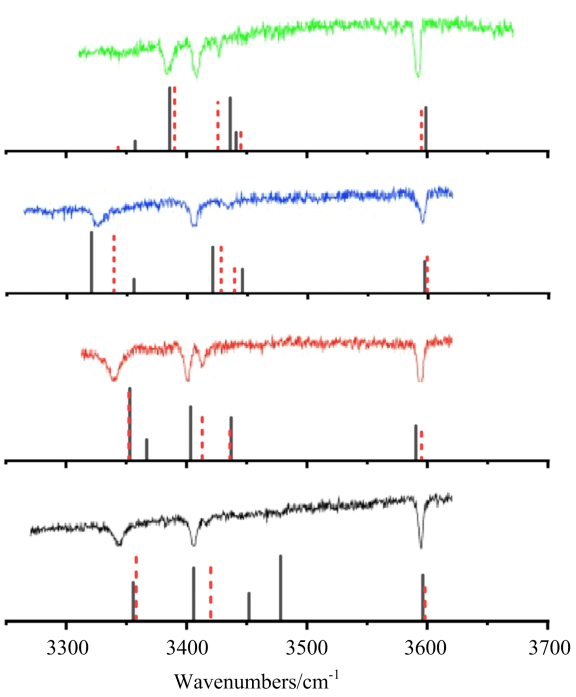

Figure 6 Comparison of FGG calculated IR spectra with experiment. From top to bottom are the vibrational spectra of fgg03, fgg06, fgg07 and fgg02. Experimental spectra and calculated results with dashed lines are from D. Řeha[48].

(N2H) in N2—H⋯O were successfully observed in the four important basic conformation calculation spectra, which are low-energy isomers of FGG important structure.

The conformer fgg03 is consistent with the most stable conformation given by W. Yu. The vibrational peaks of FGG are basically consistent with the calculation results of the method specially developed by D. Řeha (ref[48]). Using the improved method, we obtained more low-energy conformations and the energy intervals of the four important conformations corresponding to the experiment were smaller. Due to differences in functional basis sets and experimental limitations, some theoretical peaks were not observed in experiments. In addition, they proposed that DFT is not suitable for computing aromatic oligopeptides. However, our results of FGG showed that the DFT method considering the dispersion can also describe the fine structure information of oligopeptides containing benzene ringsaccurately. To a certain extent, it illustrates the universality of the CSA method of tripeptide.

3.2.3 The capping system of AAF, AFA and FAA

The following discusses the capping system in which an acetyl group (Ac) is added to the head of the peptide chain and an amino group is added to the end. For example, the structural expression of FAA is Ac-Phe-Ala-Ala-NH_2. For AAF, AFA and FAA, conformers were difficult to converge with previous high precision basis set optimization. So the following steps were used instead for optimization and vibration analysis. DFTB optimization was first performed to intercept the conformations of the first 8kcal/mol, followed by optimization and vibration analysis at the RI-B3LYP-D3/6-31＋G＊ level. Then selected the first 5 kcal/mol conformers to calculate the zero-point energy at RI-B3LYP-D3/6-311＋＋G＊＊ level, and superimposed the energy optimized in the previous step to obtain the final vibration spectrum.

Figure 7 shows the important low-energy conformations of the three tripeptides. See Table 3 for specific spectral peak information. For the capping systems of the tripeptide, hydrogen bonds determine the types of secondary structures, such as γ and β-turns (C7), 3_{10}-(C10) and α-helix (C13). To express the hydrogen bond structure information more clearly, starting from the Ac group, the N atoms are numbered N1—N4, and the sequence of peptide bonds is also in this direction, the same as below.

Our lowest energy conformation aaf01 fits the experimental spectrum[49]. The C7—C10 structure composed of the lowest frequency large red shift and high-frequency β-turn was observed in aaf01, and the calculated spectrum well confirmed this point in Figure 8 (AAF). There is no obvious π bond formation observed with the benzene ring, but both have a folding tendency to ensure confor-

mational stability. There is a weak N3—H···π in aaf01; N1—H ··· π in aaf02 and aaf03. These prevent the formation of a second C10 structure, which together with C7 further energetically stabilizes the conformation. According to the comparison with the experimental spectrum, the symmetrical s(NH_2) and asymmetric a(NH_2) stretching peaks of the terminal amino group correspond precisely. Due to the hydrogen bond in C7 and the weak π bond formed with the benzene ring, the stretching p(N2H) and p(N3H) peaks have a red shift, of about 12 cm^{-1}.

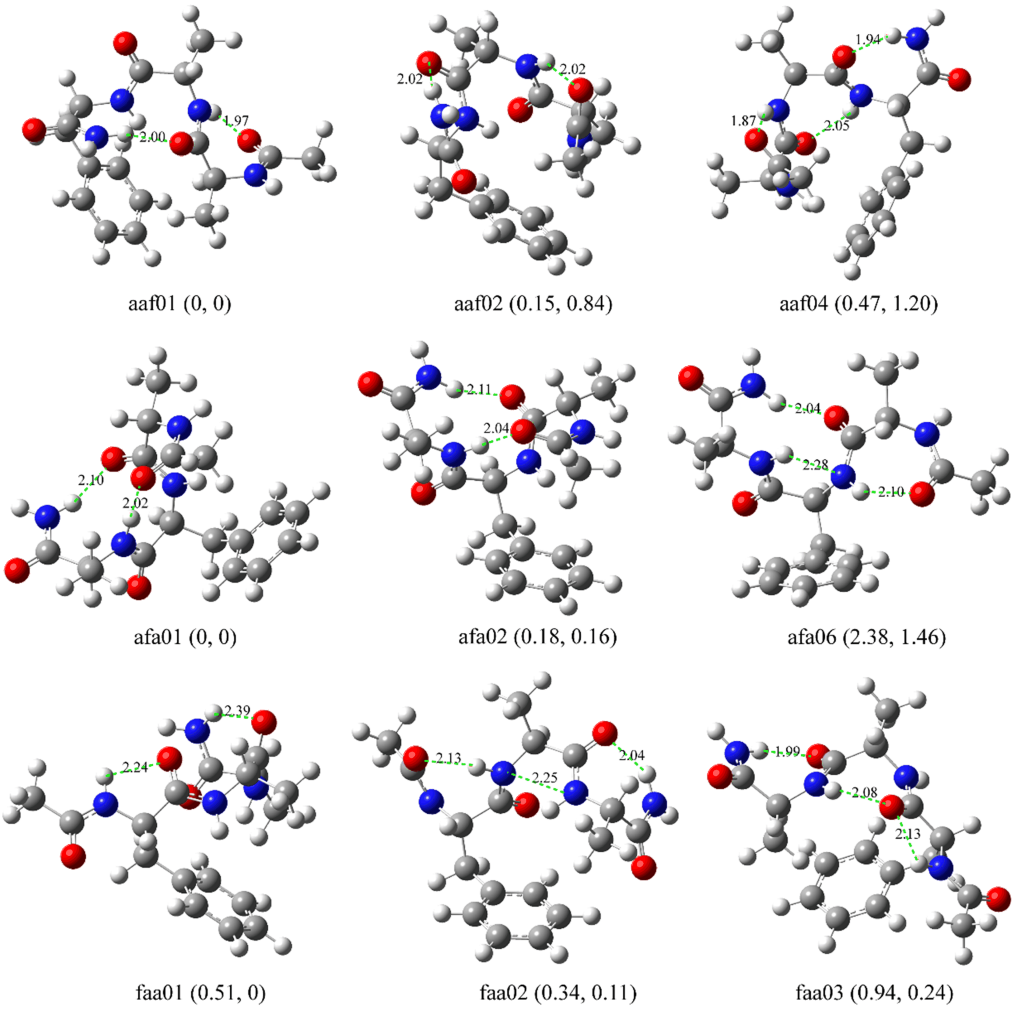

Figure 7 Important conformers and hydrogen bond information of AAF, AFA and FAA capping systems.

Table 3 Peaks and related vibration information of important conformers of AAF, AFA and FAA.

Conformer	p(N2H)	s(NH_2)	p(N3H)	p(N1H)	a(NH_2)	RMSE
exp	3303	3389	3440	—	3524	—
aaf01	3316	3390	3438	3460	3529	7.05
ref[50]	3314	3387	3437	3464	3522	5.87
exp	3429	3408	3374	—	3534	—
afa01	3434	3409	3379	3458	3537	3.87
ref[50]	3422	3412	3378	3461	3542	6.02
exp	3338	3363	3423	3447	3522	—
faa02	3334	3357	3422	3439	3520	4.92
ref[50]	3337	3360	3414	3447	3516	5.04

The first line represents the vibration mode and specific atomic information. The "exp" is the experimental result from ref[49]. The value is the wave number and the unit is cm^{-1}. Other theoretical calculations are from Bouteiller and scaling was performed using the RI-DFT-D scale factor from this research[50]. The a(NH_2) is 0.9623; p(N1H) is 0.9570; p(N3H) is 0.9570; s(NH_2) is 0.9623; p(N2H) is 0.9570.

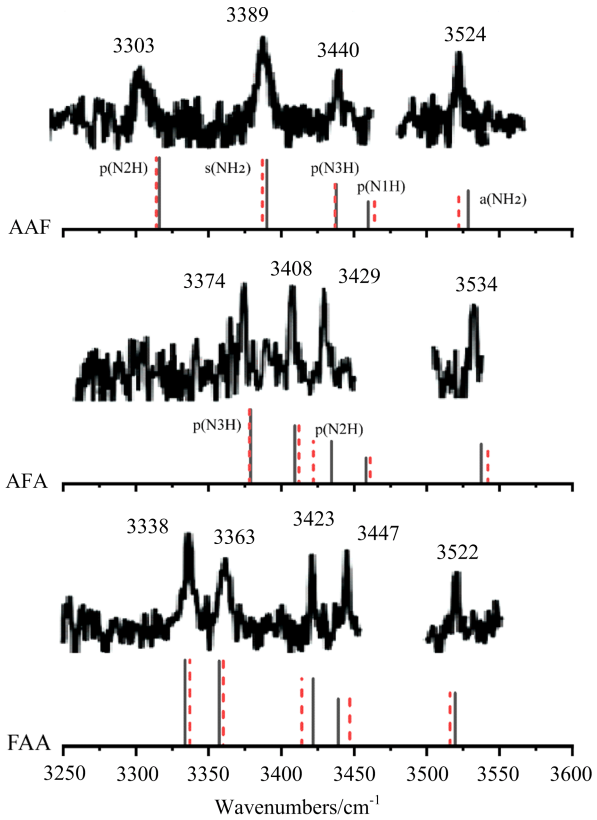

Figure 8 Comparison of theoretical and experimental infrared spectroscopy of AAF, AFA and FAA. In the figure are the calculated spectra of aaf01, afa01 and faa02. The dotted line calculation part comes from ref[50].

The important low-energy conformation of AFA is the C10-C10 helical structure. One of the C10 is N3—H···O formed by the oxygen at the c-terminal of Ac and N3, and the other is N4—H···O formed by the oxygen of the second peptide bond and the terminal amino group. There is still no obvious π bond observed. In addition, a very weak N2—H···π interaction in afa01 leads to a very small red shift of p(N3H) and p(N2H), as shown in Figure 8 (AFA). However, we did not observe the result in ref[49] that the blue shift of the amino anti-symmetric band due to the helical structure. The terminal amino group of afa01 forms a helical folded C10 structure, but the highest frequency a(NH_2) peaks have no blue shift. Combining the comparison with the experimental and other theoretical calculation results, we believe that this conclusion about the blue shift of the antisymmetric band is wrong. In all important low energy conformers of AFA, aliphatic-aromatic interactions between adjacent Ala and Phe residues in Belén[51] were successfully observed.

There are three secondary structures in the important low-energy configuration of FAA. The first and second peptide bonds form C5 structure N1-H ··· O (existing in faa01 and faa03); C7 structure N2-H···O (in faa02); the third peptide bond oxygen and the terminal amino hydrogen form the C7 structure N4-H···O. According to the bond length and frequency, the strength of C7 is higher than C5. All theoretical results have not observed the interaction with the benzene ring effect. Figure 8 (FAA) is the calculated spectrum of faa02. Compared with the experiment, it can be seen that due to the formation of intramolecular hydrogen bonds and the average of the chemical bond electron cloud, the corresponding peaks have different degrees ofblue shift. From the perspective of atomic distance, for the low-energy FAA conformers, the interaction between the oxygen on the fourth peptide bond at the end and the hydrogen on the benzene ring is more obvious than the possible N—H···π bond with the benzene ring (faa01 is N3). The conformers are all tightly folded end to end (faa01-faa03), which is not observed in previous studies, possibly due to the small number of conformers.

In summary, for the AAF and AFA systems, the infrared spectra calculated by the conformational ensemble generated by CSA are roughly the same as those analyzed in previous

studies, and more accurately interpret the experimental data. The structure of the AFA system overturned the conclusion in ref[49] that the C10 helical structure caused the blue shift of the antisymmetric band. In addition, the local scaling factor method in ref[50] is recommended for vibrational analysis of small peptides. Most importantly, there has been a substantial increase in the number of conformations. More comprehensive and multiple conformational ensembles deepen the understanding of existing experimental results. Especially for the FAA system, the additional new conformations have observed new structural rules that may exist in the low-energy conformation.

3.3 Related application fields and significance

The proposed method provides a new approach to address the issue of peptide flexibility during docking. As the conformational space of the ligand is explored, the ligand flexibility is represented by docking a pre-generated conformational ensemble, thereby eliminating computational costs. Not only that, the energy corresponding to the ensemble accelerates the process of calculating high-precision energy methods. Unlike most structure databases based on experimental structures, our obtained database does not use any experimental results. This makes up for the limitations of experiments from a theoretical perspective.

In addition, we further calculated the energy gap of the molecule. Combined with the HOMO-LUMO gap, the conformers with energy can quickly assist in judging the reaction mechanism of the enzyme-catalyzed process[52], provide a more accurate training set for machine learning[53], promote the process of drug design[54], and assist in template-free structure prediction, providing lower cost and broader prospects for related application fields.

4 Conclusions

Based on previous studies, we have developed a reliable computational method for the prediction of the conformational ensemble of tripeptides. The comparison with the PDB showed that the number of newly searched conformers has significantly increased, especially in the low and middle energy regions of interest for medical applications. In addition, it is also confirmed that there is a connection between the amount of data in the PDB database and the quality of the deduced results. The comparison between the theory and experiment of infrared spectroscopy explained the experimental results more accurately and gave a deeper understanding of the experimental structure. The reliability of the new method is proved both statistically and experimentally.

参 考 文 献

[1] C. Savojardo, P. L. Martelli, P. Fariselli, R. Casadio, *Bioinformatics*, **34** (2017), 1690

[2] D. Pozo, M. Delgado, C. Martínez, J. M. Guerrero, J. Leceta, R. P. Gomariz, J. R. Calvo, *Immunol. Today*, **21** (2000), 7

[3] C. Yang, S. Zhang, Z. Bai, S. Hou, D. Wu, J. Huang, P. Zhou, *Mol. BioSyst.*, **12** (2016), 1201

[4] F. A. Russell, R. King, S.-J. Smillie, X. Kodji, S. D. Brain, *Physiol. Rev.*, **94** (2014), 1099

[5] N. Y. Yount, A. S. Bayer, Y. Q. Xiong, M. R. Yeaman, *Pept. Sci.*, **84** (2006), 435

[6] R. E. W. Hancock, D. S. Chapple, *Antimicrob. Agents Chemother.*, **43** (1999), 1317

[7] A. D. Mackerell Jr., M. Feig, C. L. Brooks III, *J. Comput. Chem.*, **25** (2004), 1400

[8] J. D. Durrant, J. A. McCammon, *BMC Biol.*, **9** (2011), 71

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.*, **28** (2000), 235

[10] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, *Proc. Natl. Acad. Sci.*, **117** (2020), 1496

[11] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature*, **596** (2021), 583

[12] M. Baek, F. DiMaio, I. V. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. M. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, *Science*, **373** (2021), 871

[13] A. Roy, A. Kucukural, Y. Zhang, *Nat. Protoc.*, **5** (2010), 725

[14] A. Šali, T. L. Blundell, *J. Mol. Biol.*, **234** (1993), 779

[15] S. Zivanovic, F. Colizzi, D. Moreno, A. Hospital, R. Soliva, M. Orozco, *J. Chem. Theory Comput.*, **16** (2020), 6575

[16] F. Stanzione, I. Giangreco, J. C. Cole, (Witty D R, Cox B ed) *Prog. Med. Chem.*, (2021), p273

[17] Y. Yan, D. Zhang, S.-Y. Huang, *J. Cheminf.*, **9** (2017), 59

[18] A. Rayan, H. Senderowitz, A. Goldblum, *J. Mol. Graph. Model.*, **22** (2004), 319

[19] X. Pan, T. Kortemme, *J. Biol. Chem.*, **296** (2021), 100558

[20] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, D. Baker, *Nature*, **491** (2012), 222

[21] M. Garton, S. Nim, A. Stone Tracy, E. Wang Kyle, M. Deber Charles, M. Kim Philip, *Proc. Natl. Acad. Sci.*, **115** (2018), 1505

[22] B. Yang, Z. Lin, *Comput. Theor. Chem.*, **1108** (2017), 40

[23] W. Yu, Z. Wu, H. Chen, X. Liu, A. D. MacKerell, Jr., Z. Lin, *J. Phys. Chem. B*, **116** (2012), 2269

[24] D. Robert, J. H. Warren, A. P. John, *J. Chem. Phys.*, **52** (1970), 5001

[25] G. Stefan, A. Jens, E. Stephan, K. Helge, *J. Chem. Phys.*, **132 15** (2010), 154104

[26] S. Grimme, S. Ehrlich, L. Goerigk, *J. Comput. Chem.*, **32** (2011), 1456

[27] S. Grimme, *J. Comput. Chem.*, **27** (2006), 1787

[28] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Deshaye, T. Dumitrică, A. Dominguez Garcia, S. Ehlert, M. Elstner, T. Heide, J. Hermann, S. Irle, J. M. Julián, C. Köhler, T. Kowalczyk, T. Kubař, T. Frauenheim, *J. Chem. Phys.*, **152** (2020), 124101

[29] M. Gaus, Q. Cui, M. Elstner, *J. Chem. Theory Comput.*, **7** (2012), 931

[30] X. Ru, C. Song, Z. Lin, *J. Phys. Chem. B*, **121** (2017), 2525

[31] A. J. Campbell, M. L. Lamb, D. Joseph-McCarthy, *J. Chem. Inf. Model.*, **54** (2014), 2127

[32] O. Korb, T. S. G. Olsson, S. J. Bowden, R. J. Hall, M. L. Verdonk, J. W. Liebeschuetz, J. C. Cole, *J. Chem. Inf. Model.*, **52** (2012), 1262

[33] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.*, **98** (1994), 11623

[34] R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople, *J. Chem. Phys.*, **72** (2008), 650

[35] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.*, **7 18** (2005), 3297

[36] F. Weigend, *J. Comput. Chem.*, **29** (2008), 167

[37] H. Valdés, D. Řeha, P. Hobza, *J. Phys. Chem. B*, **110** (2006), 6385

[38] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian* 09 *Rev. C*.01,(Wallingford, CT) (2016)

[39] F. Neese, *WIREs Comput. Mol. Sci.*, **8** (2018), e1327

[40] W. Sippl, (Wermuth C G ed) *The Practice of Medicinal Chemistry* (*Third Edition*) (New York: Academic Press), (2008), p572

[41] H. H. Avgy-David, H. Senderowitz, *J. Chem. Inf. Model.*, **55** (2015), 2154

[42] E. Freire, *Drug Discov. Today*, **13** (2008), 869

[43] B. Kuhlman, P. Bradley, *Nat. Rev. Mol. Cell Biol.*, **20** (2019), 681

[44] Q. Bai, J. Ma, S. Liu, T. Xu, A. J. Banegas-Luna, H.

Pérez-Sánchez, Y. Tian, J. Huang, H. Liu, X. Yao, *Comput. Struct. Biotechnol. J.*, **19** (2021), 3573

[45] P. M. Szell, S. Zablotny, D. L. Bryce, *Nat. Commun.*, **10** (2019), 1

[46] J. Černý, P. Jurečka, P. Hobza, H. Valdés, *J. Phys. Chem. A*, **111** (2007), 1146

[47] J. M. Bakker, C. Plützer, I. Hünig, T. Häber, I. Compagnon, G. von Helden, G. Meijer, K. Kleinermanns, *ChemPhysChem*, **6** (2005), 120

[48] D. Řeha, H. Valdés, J. Vondrášek, P. Hobza, A. Abu-Riziq, B. Crews, M. S. de Vries, *Chem. Eur. J.*, **11** (2005), 6803

[49] W. Chin, F. Piuzzi, J.-P. Dognon, I. Dimicoli, B. Tardivel, M. Mons, *J. Am. Chem. Soc.*, **127** (2005), 11900

[50] Y. Bouteiller, J. C. Poully, C. Desfrançois, G. Grégoire, *J. Phys. Chem. A*, **113** (2009), 6301

[51] B. Hernández, F. Pflüger, S. G. Kruglik, M. Ghomi, *J. Mol. Graph. Model.*, **102** (2021), 107790

[52] Y. Zou, S. Yang, J. N. Sanders, W. Li, P. Yu, H. Wang, Z. Tang, W. Liu, K. N. Houk, *J. Am. Chem. Soc.*, **142** (2020), 20232

[53] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.*, **9** (2017), 513

[54] D. Gogoi, A. K. Chaliha, D. Sarma, B. B. Kakoti, A. K. Buragohain, *Biomed. Pharmacother.*, **85** (2017), 646