


Dynamic structural analysis-based epitope prediction of Exendin-4 in aqueous solutionJianfeng He ^{*}*School of Physics, Beijing Institute of Technology, Beijing 100081, People's Republic of China*

Jing Li

Research and Development Center, Beijing Genetech Pharmaceutical Co., Ltd., Beijing 102200, People's Republic of China

Kingsley Leung

Uni-Bioscience Pharm Company Limited, Hong Kong, People's Republic of China

(Received 21 March 2023; accepted 22 July 2023; published 18 August 2023)

The study of epitopes has a broad range of applications in drug discovery, vaccine design, and immunotherapy. In this study, an epitope prediction method was developed based on the dynamic structure of protein antigens. Solvent accessible surface area, charge, and root mean square fluctuation were introduced as the key residue property parameters. The epitope prediction algorithm was established by constructing a three-parameter complex metrics of seven-peptide groups. The method was applied to predict the epitopes of Exendin-4, an effective antidiabetic drug. The epitopes of both the natural and C-terminal amidated forms of Exendin-4 were predicted and compared in their folded and intermediate states. In the folded state, the epitopes of natural Exendin-4 (His1-Phe6 and Asp9-Val19) were found to be nearly identical to the epitopes of C-terminal aminated Exendin-4 (His1-Thr7 and Asp9-Val19). In the intermediate state, however, the epitopes of natural Exendin-4 (His1-Gly4, Phe6 and Lys12-Arg20) covered fewer amino acids than the epitopes of C-terminal aminated Exendin-4 (His1-Gly4, Phe6, Asp9-Val19 and Trp25-Lys27). The comparison with the results from other prediction tools demonstrates the reliability of our predicted epitopes of Exendin-4.

DOI: [10.1103/PhysRevE.108.024403](https://doi.org/10.1103/PhysRevE.108.024403)**I. INTRODUCTION**

An antigen is a substance that stimulates an immune response to fight diseases in a living organism. It initiates an immunoreaction by binding to secreted antibodies or antigen-specific membrane receptors on lymphocytes [1]. Epitopes, or antigenic determinants, are the critical pieces of an antigen that are specifically recognized by antibodies and receptors on lymphocytes. The epitopes of protein antigens can be categorized into linear and conformational epitopes [2]. Linear epitopes are composed of continuous residues along a polypeptide chain. Conformational epitopes are composed of discontinuous residues that are assembled together through the folding of protein antigen into its 3D structure [1,3]. As the molecular basis of the immune response, both types of epitopes play a vital role in drug discovery, vaccine design, and immunotherapy [4–10].

Experimental methods for the identification of epitopes have been developed, such as x-ray crystallography [11], overlapping peptide scan [12], high-throughput shotgun mutagenesis [13], mass spectrometry [14], and random peptide phage libraries [15]. With advances in epitope mapping technologies and databases, computer-aided methods have been exploited for epitope prediction. Algorithms based on machine learning, such as the hidden Markov model [16], neural

network [17], and support vector machine [18] have been proposed to improve prediction performance. There are available tools for predicting linear and conformational epitopes, including BepiPred [16], ABCpred [17], DiscoTope [19], ElliPro [20], NetMHCpan [21], and NN-align [22]. Due to the advantages of speed, efficiency, and cost effectiveness, the computer-aided prediction has become a valuable means of epitope identification.

Research has suggested a correlation between epitope localization and the physicochemical properties of protein antigens [23]. Parameters related to the sequence and structure of protein antigens have been introduced into epitope prediction (e.g., hydrophilicity, solvent accessibility, antigenicity, surface exposure, etc.). It has been demonstrated that prediction methods based on a single residue property yield poor results [24]. Combining two or more residue properties has become a trend to improve prediction performance [25]. On the other hand, existing prediction methods can be divided into sequence-based and structure-based methods according to the input information. Since 3D structures are more conserved than the sequences of protein antigens, structure-based methods often outperform sequence-based methods [26]. However, prediction methods based on spatial structure are still limited and require further development.

In this work, a method for epitope prediction is proposed by using the dynamic structure of protein antigens. The innovative concept involves using the dynamic structure obtained from molecular dynamics simulations in an aqueous

^{*}Corresponding author: hjf@bit.edu.cn

solution as the input to the method, rather than the traditional PDB file. To enhance performance, three residue properties are utilized as prediction parameters: solvent accessibility, charge, and flexibility. Most current structure-based methods predict epitopes based on static crystal structures [11,19,20]. In comparison, our dynamic structure-based method can more realistically reflect the physicochemical properties of protein antigens in solution. Additionally, intermediate states have been observed during the protein folding [27–29], which may interact with antibodies and receptors on lymphocytes. Hence, it is imperative to study the epitopes present in the intermediate state of protein antigens to gain a more comprehensive understanding of epitopes.

Exendin-4 is a polypeptide hormone composed of 39 amino acid residues [30]. It has the ability to effectively stimulate glucose-dependent insulin secretion and inhibit glucagon secretion [31,32]. As an antidiabetic drug, Exendin-4 holds great potential for the clinical treatment of type 2 diabetes mellitus [33]. It is noteworthy that Exendin-4 also emerges as an exogenous antigen upon entry into the organism. It can trigger an immune response like any other antigen, thereby attenuating its hypoglycemic effect. In this work, the epitopes of Exendin-4 are identified by the dynamic structure-based method. This is of practical value to understand the immunogenicity of Exendin-4 as well as the resistance of organisms to Exendin-4.

In practice, Exendin-4 is mainly prepared through chemical synthesis and recombinant DNA [34,35]. The two species of Exendin-4 share the same sequence, but the C-terminus of the synthesized Exendin-4 is amidated. It has been discovered that the biological properties of the recombinant Exendin-4 are almost identical to those of the natural Exendin-4. However, the synthesized Exendin-4 has different properties, including its immunogenicity. This motivates us to investigate the differences in the epitopes between the natural and C-terminal amidated forms of Exendin-4.

This paper is organized as follows. Section II describes the molecular dynamics simulation, the steered molecular dynamics simulation, the selection of residue property parameters, and the algorithm of epitope prediction. Section III presents the results of our epitope prediction for Exendin-4. Both the folded and intermediate states of Exendin-4 are considered in our predictions. We compare the results for two species of Exendin-4: the natural and C-terminal amidated forms. Finally, Sec. IV summarizes the results of our work.

II. METHODS

A. Molecular dynamics simulation

Molecular dynamics (MD) simulations were employed to obtain the dynamic structure of Exendin-4 for epitope identification. The NMR structures of natural and C-terminal amidated Exendin-4 were downloaded from the PDB as initial structures (PDB entries 1JRJ and 7MLL) [34,35]. All simulations were carried out using GROMACS 5.0.7 and PLUMED 2.3 [36,37] with the CHARMM36 force field [38]. Exendin-4 was placed in a $65 \times 65 \times 65 \text{ \AA}^3$ cubic box under periodic boundary conditions. The box was filled with explicit water in the TIP3P model [39]. Counterions were added to neutral-

ize the net charge of Exendin-4, and the salt concentration was set to 100 mM NaCl. The particle mesh Ewald (PME) algorithm was used to calculate the long-range electrostatic interactions [40]. The short-range electrostatic and van der Waals interactions were truncated at a distance of 10 Å. An energy minimization of 10 ps was performed using the steepest descent algorithm [41]. The modified Berendsen thermostat and Parrinello-Rahman barostat were used for temperature and pressure coupling, respectively [42,43]. A 100 ps NVT equilibration followed by a 100 ps NPT equilibration was performed to stabilize the system at 310 K and 1 bar. The MD simulations were run for 100 ns. The dynamic structure was extracted from the output trajectory for epitope identification of Exendin-4 in its folded state.

B. Steered molecular dynamics simulation

The natural and C-terminal amidated Exendin-4 exhibit stable intermediate states. Structures in these intermediate states are partially folded. The epitopes of intermediate states also need to be identified for natural and C-terminal amidated Exendin-4. Steered molecular dynamics (SMD) simulations were used to pull Exendin-4 from the folded state to the intermediate state [44–48]. In SMD, a harmonic potential was introduced into the Hamiltonian to mimic the pull exerted on Exendin-4. The time-dependent harmonic potential was defined as,

$$V_\lambda(\vec{q}, t) = \frac{1}{2} \kappa(t) (\xi(\vec{q}) - \lambda(t))^2. \quad (1)$$

Here, κ was the spring constant of harmonic potential, $\xi(\vec{q})$ represented the reaction coordinate, \vec{q} referred to the coordinate vector of heavy atoms, and $\lambda(t)$ was the control parameter of the drag path. The reaction coordinate $\xi(\vec{q})$ varied along the control parameter $\lambda(t)$ or a path defined by $\xi_0(\vec{q})$ to implement the drag process.

One or more collective variables (CVs) can be selected as reaction coordinates, such as the fraction of native contacts, the fraction of hydrogen bonds, the radius of gyration (Rg), root mean square deviation (RMSD), etc. The fraction of native contacts (Q_N) and helix content (S_H) of Exendin-4 were chosen based on the structural characteristics. Q_N was defined as [49]

$$Q_N = \frac{1}{N} \sum_i \sum_j \frac{1}{1 + \exp(\beta(r_{ij} - \lambda r_{ij}^0))}. \quad (2)$$

Here the sum ran over the N atom pairs of native contact. r_{ij} was the distance between two heavy atoms, i and j . r_{ij}^0 was the reference distance, chosen to be the distance between the two heavy atoms in the folded structure. Only nonadjacent residues were considered when calculating folded contacts. The parameters were configured based on the settings in literature [49]. Specifically, the smoothing parameter β was set to 50 nm^{-1} , and the factor λ was assigned a value of 1.8 for the all-atom model. S_H was defined as a differentiable function of the atomic coordinates in the following manner [50]:

$$S_H = \sum_\alpha n[\text{RMSD}(\{R_i\}_{i \in \Omega_\alpha}, \{R^0\})], \quad (3)$$

$$n(\text{RMSD}) = \frac{1 - (\text{RMSD}/0.12)^8}{1 - (\text{RMSD}/0.12)^{12}}. \quad (4)$$

Here, n was a function switching smoothly between zero and one. $\{R_i\}_{i \in \Omega_\alpha}$ were the atomic positions of a set Ω_α of six consecutive residues in Exendin-4. $\{R^0\}$ were the corresponding atomic positions of an ideal α -helix structure. The sum traversed all six-residue sets in Exendin-4. The determination of function $n(\text{RMSD})$ involved computing the RMSD between the six-residue fragment and the ideal α helix, with RMSD measured in nm. During the RMSD calculations, only the N, C_α , C, O, and C_β atoms were considered.

The natural Exendin-4 exhibits an intermediate state with Q_N and S_H values of 0.82 and 11.1, respectively. The C-terminal aminated Exendin-4 has Q_N and S_H values of 0.54 and 13.3, respectively, in its intermediate state. In SMD simulations, the Q_N and S_H as reaction coordinates changed along a linear path towards the values of intermediate state. The spring constant was set to 30 000 kJ/mol. Each atom experienced a “restoring force” of approximately 1.6 kJ/mol due to the changes in the collective variables in our SMD simulations. This selection of the spring constant enabled efficient pulling without compromising the structural integrity. It first took 20 ns to drag Exendin-4 from the folded state to the intermediate state. An equilibrium simulation was then continued for 80 ns to generate a trajectory in the intermediate state. From this trajectory, the dynamic structure of Exendin-4 in the intermediate state was extracted and used for epitope prediction.

C. Selection of residue property parameter

Epitopes are often found on the outer surface of protein antigens. This makes them easily accessible to antibodies and receptors on lymphocytes. Previous investigations have shown that the flexibility, charge, and high exposure to the solvent are important features of epitopes. Therefore, the three residue properties were selected for the epitope prediction. These properties were described by the solvent accessible surface area (SASA), charge of each residue (Qg), and root mean square fluctuation of C_α atoms (RMSF), respectively.

The value of SASA represents the area of a protein antigen that is accessible to the solvent. The residue-specific SASA can distinguish whether a residue is located on the outer surface of protein antigen [51,52]. In this method, the average SASA of each residue was calculated using the SASA program of Gromacs, based on the dynamic structure. The rolling ball algorithm was employed in the calculation of SASA [53]. The radius of ball was set to 1.4 Å.

Since acidic/alkaline residues may lose/gain protons in the solvent, the surface of protein antigens usually exhibits a charge distribution. These charges can electrostatically interact with those on antibodies or receptors on lymphocytes. As is known, the electrostatic interaction is one of the main protein-protein interactions. Previous prediction methods rarely considered the charge distribution. In our method, the charge carried by each residue was explicitly introduced as a residue property parameter. The quantity of charge per residue (Qg) was evaluated using the topology file of MD simulation.

The fragment flexibility is indicative of epitopes in protein antigens [54]. It is often used as an epitope prediction parameter [51,52,54,55]. Typically, Karplus and Schulz estimated the flexibility of fragments using individual atomic temperature

factors, i.e., B values [55]. In our method, the fragment flexibility was analyzed by RMSF of C_α atoms in the dynamic structure. Since RMSF describes the freedom of movement of heavy atoms, it can better reflect the local flexibility of protein antigens.

In all, SASA effectively discerns the exposed residues on the antigen’s surface, while Qg provides valuable information about the antigen’s surface charge, and RMSF sheds light on the local flexibility of the epitope region. To enhance the reliability of our prediction method, we undertook a comprehensive consideration of these three property parameters, as they collectively captured diverse aspects pertinent to epitopes. We noted that their individual contributions may vary, and this can be appropriately addressed by applying suitable weighting to each parameter. However, our overarching aim was to devise a straightforward yet powerful framework. Hence, we directly aggregated them through linear summation to yield a three-parameter complex metric.

D. Algorithm of epitope prediction

The procedure for predicting potential epitopes was implemented as follows. First, the dynamic structure of the folded or intermediate state of protein antigens was obtained using either MD or SMD simulation. Three property parameters of each residue, namely SASA, Qg and RMSF, were evaluated based on the dynamic structure. Second, ignoring the first and last residues, the SASAs, Qgs, and RMSFs of (N-2) residues were averaged (where N is the number of residues). These three averages served as a baseline of 1.0. The three property parameters of each residue were then divided by their corresponding averages and used as their rescaled values (denoted by SASA' , Qg' , and RMSF'). We selected different lengths of consecutive residue fragments, including one to nine residues, and evaluated them using the property parameters. Among these, the seven-consecutive-residue fragment exhibited the optimal average sequence attributes. Therefore, in the third step, a series of seven-consecutive-residue fragments (called seven-peptide groups) were extracted from the N to C terminus of protein antigens. For each seven-peptide group, the SASA' , Qg' , and RMSF' of its seven residues were averaged. And these averages were assigned to its central residue as the epitope prediction metrics (denoted by SASA'' , Qg'' , and RMSF''). Finally, the three single metrics were linearly added to construct three-parameter complex metrics. The profile of the three-parameter complex metrics was drawn and analyzed to identify potential epitopes of protein antigens.

III. RESULTS AND DISCUSSION

A. Epitopes of natural Exendin-4 in the folded state

Figure 1(a) depicts the typical structure of natural Exendin-4 in the folded state. The His1-Thr5 fragment exhibits the preference of a soft loop. The Phe6-Asn28 and Gly29-Ser33 fragments form an α helix and a 3_{10} helix, respectively. The Leu21-Pro38 fragment folds into a compact hydrophobic core (Trp cage), where the sidechains of hydrophobic residues wraps around the sidechain of Trp25 to prevent its exposure to the solvent.

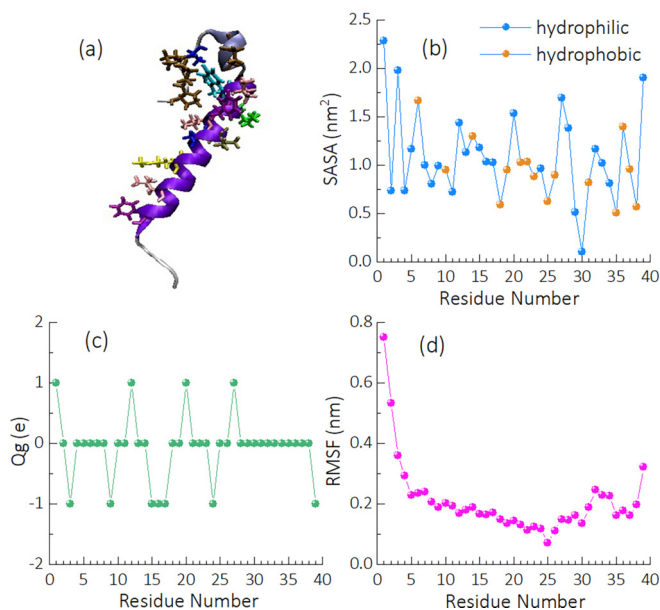


FIG. 1. The structure and three property parameters of natural Exendin-4 in the folded state. (a) Cartoon representation of the typical structure. [(b)–(d)] Values of SASA, Qg, and RMSF for each residue.

Figures 1(b)–1(d) present the average solvent accessible surface area (SASA), charge (Qg), and root mean square fluctuation (RMSF) of C_{α} atoms for each residue. The average SASAs of His1, Glu3, Thr5-Phe6, Lys12-Glu17, Arg20-Phe22, Lys27-Asn28, Ser32-Ser33, Pro36, and Ser39 are greater than 1.0 nm^2 . Gly30 has the smallest SASA due to having only one hydrogen atom in its sidechain. Clearly, the His1-Arg20 fragment has more hydrophilic residues (blue spheres), whereas the Leu21-Ser39 fragment has more hydrophobic residues (orange spheres). In Trp cage, the average SASAs of Trp25, Ala35, and Pro38 are significantly lower than those of other hydrophobic residues. The charge analysis of natural Exendin-4 in a $\text{pH} = 5.5$ solvent environment shows that His1, Lys12, Arg20, and Lys27 are positively charged, whereas Glu3, Asp9, GLU15-17, Glu24, and Ser39 are negatively charged. It can be found that the Lys12-Arg20 fragment has a relatively dense charge distribution. The RMSFs of His1-Thr7, Ser32-Gly34, and Ser39 are greater than 0.2 nm . This suggests that these fragments have better flexibility, particularly His1-Gly4 and Ser39.

Figure 2 shows the profiles of SASA'' , Qg'' , RMSF'' , and the three-parameter complex metrics. In the SASA'' profile, we can observe several local maxima located at indices 4-6, 9, 12-15, 17, and 25. Due to the hydrophobic collapse in Trp cage, the SASA'' values at indices 20–38 are relatively small. In the Qg'' profile, a main peak is formed in the 12–19 index region. This derives from the high charge distribution in the Lys12-Arg20 fragment. Two peaks in the RMSF'' profile can be found at indices 4–10 and 32–36, respectively. In the three-parameter complex profile, there are two local maxima at indices 4–6 and 9–19. This reflects the compound effect of three single metrics.

Figure 3 displays the predicted epitopes of natural Exendin-4 in the folded state. The epitopes contain the

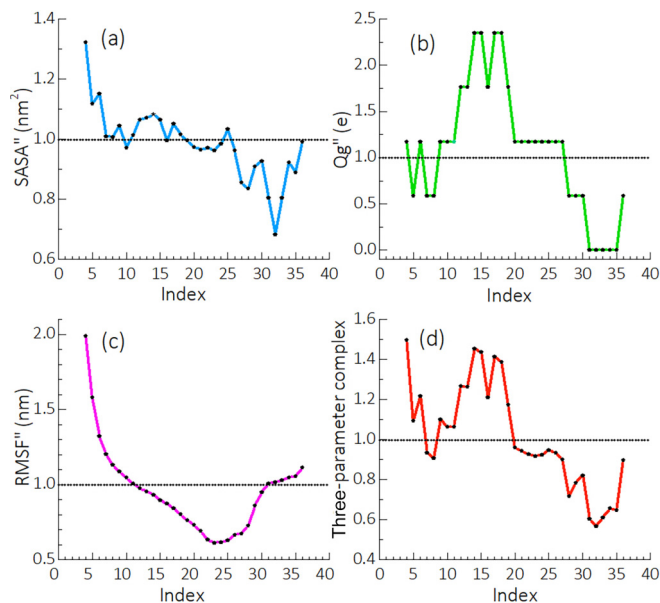


FIG. 2. Profiles of single and three-parameter complex metrics of natural Exendin-4 in the folded state: SASA'' (a), Qg'' (b), RMSF'' (c), and the three-parameter complex (d).

key residues His1-Phe6 and Asp9-Val19. Among them, the His1-Phe6 fragment has the good solvent accessibility and structural flexibility. And the Asp9-Val19 fragment has the good solvent accessibility and high charge distribution. Due to the formation of Trp cage, the Leu21-Pro38 fragment is unlikely to be an epitope. Figure 3 also shows the epitopes identified by three single metrics. When the SASA'' is used as a metric, the epitopes consist of His1-Phe6, Asp9, Lys12-Glu15, Glu17, and Trp25. The Lys12-Val19 fragment is the epitope measured by the Qg'' . From the RMSF'' metrics, the epitopes include the residues His1-Leu10 and Ser32-Ser39. It can be found that the epitope prediction of three-parameter complex metrics covers and highlights the common features of the three single metrics.

B. Epitopes of natural Exendin-4 in the intermediate state

The potential epitopes of natural Exendin-4 are further identified in the intermediate state. The dynamic structure of intermediate state is obtained by the SMD simulation described in Sec. II. Figure 4(a) shows the typical structure of the intermediate state. The His1-Phe6 and Lys12-Glu15

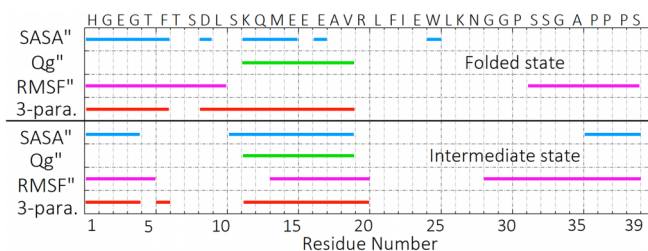


FIG. 3. Epitope predictions of natural Exendin-4 using different metrics: SASA'' , Qg'' , RMSF'' , and the three-parameter complex (3-para.).

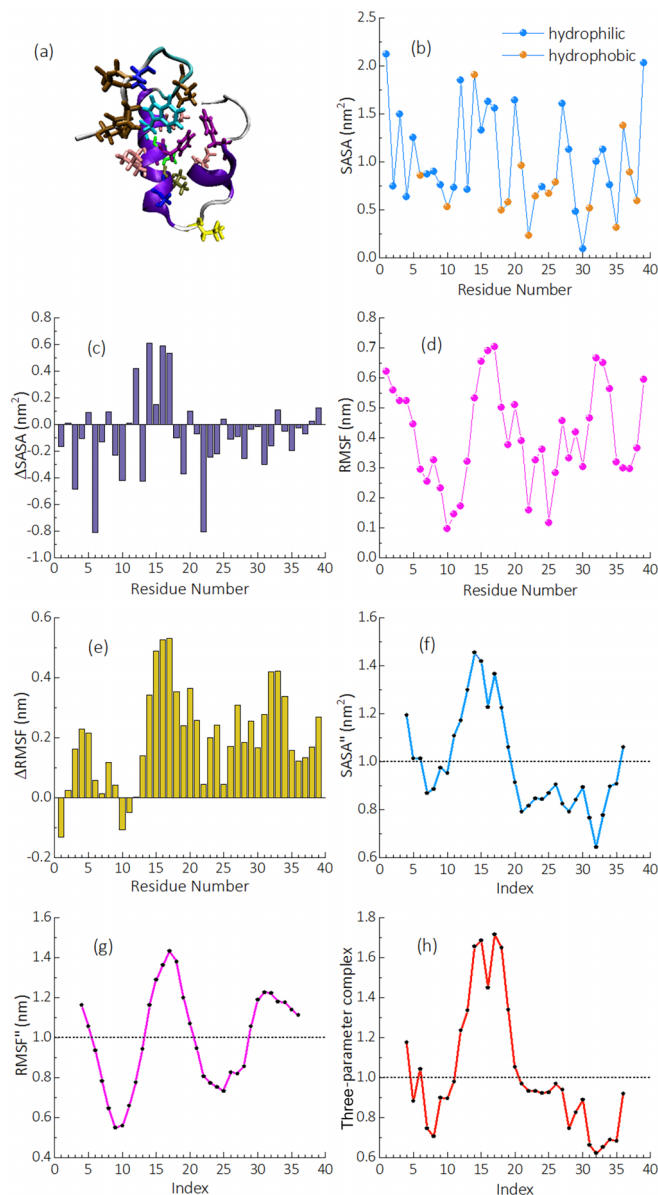


FIG. 4. The structure and properties of natural Exendin-4 in the intermediate state. (a) The typical structure of intermediate state. [(b)–(e)] SASA and RMSF for each residue and their changes compared to those in the folded state: SASA (b), Δ SASA (c), RMSF (d), Δ RMSF (e). [(f)–(h)] Profiles of SASA'', RMSF'', and the three-parameter complex metrics.

fragments are two loops. The Thr7-Ser11 and Glu16-Asn28 fragments form two α helices. In comparison to the structure observed in the folded state, the Leu21-Pro38 fragment maintains the Trp cage structure. However, the long α helix is broken into two helices in the intermediate state.

Figures 4(b)–4(e) show the SASA and RMSF for each residue, as well as the differences between these values and those observed in the folded state (Δ SASA, Δ RMSF). The average SASAs of His1, Glu3, Thr5, Lys12, Met14-Glu17, Arg20, Lys27-Asn28, Ser32-Ser33, Pro36, and Ser39 are greater than 1.0 nm^2 . Compared with the folded state, the SASAs of Lys12 and Met14-Glu17 increase, while the SASAs

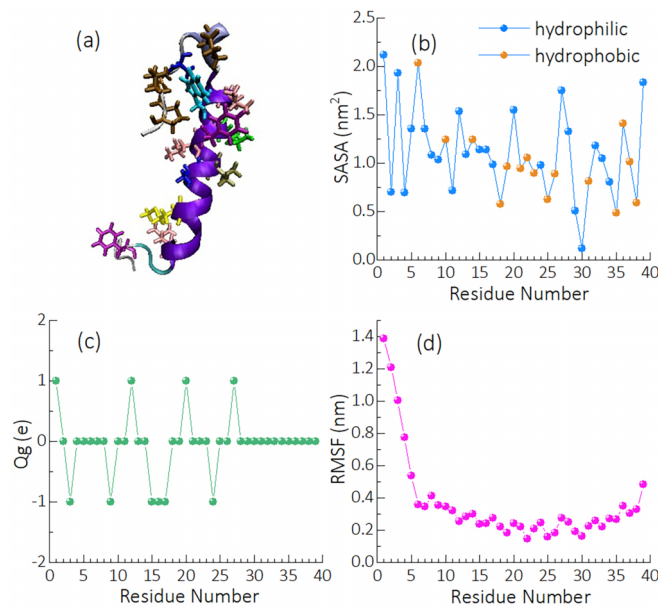


FIG. 5. The structure and three property parameters of C-terminal amidated Exendin-4 in the folded state. (a) Cartoon representation of the typical structure. (b)–(d) Values of SASA, Qg, and RMSF for each residue.

of Glu3, Phe6, Asp9-Leu10, Gln13, Val19, Phe22-Glu24, Asn28, and Pro31 decrease significantly. This is likely due to the bending of the His1-Asn28 fragment into a U-shaped structure. There are 34 residues (His1-Asp9, Gln13-Leu21, Ile23-Glu24, and Leu26-Ser39) with RMSF greater than 0.2 nm . In general, the RMS fluctuations in the intermediate state are stronger than those in the folded state.

Figures 4(f)–4(h) depict the profiles of SASA'', RMSF'', and the three-parameter complex metrics in the intermediate state of natural Exendin-4. The SASA'' profile shows maxima in the regions of 4, 11–19, and 36. In the RMSF'' profile, there are three peaks located in indices 4–5, 14–20, and 29–36. The Qg'' profile of the intermediate state is the same as that of the folded state, due to the same charge distribution. Three maxima at indices 4, 6, and 12–20 can be observed in the profile of the three-parameter complex.

The predicted epitopes of the intermediate state are illustrated in Fig. 3. Based on the three-parameter complex metrics, the potential epitopes are composed of His1-Gly4, Phe6, and Lys12-Arg20. It is apparent that the epitopes in the intermediate state consist of almost the same residues as those in the folded state. However, the difference lies in the fact that the Lys12-Arg20 epitope in the intermediate state essentially covers all the predictions of the three single metrics.

C. Epitopes of C-terminal amidated Exendin-4 in the folded state

The epitopes of C-terminal amidated Exendin-4 are investigated to reveal the differences in immunogenicity compared to natural Exendin-4. The dynamic structure in the folded state is obtained by the MD simulation. Figure 5(a) shows the typical structure of the folded state. This structure mainly consists of a long α helix (Leu10-Asn28) and a Trp cage (Leu21-Pro38),

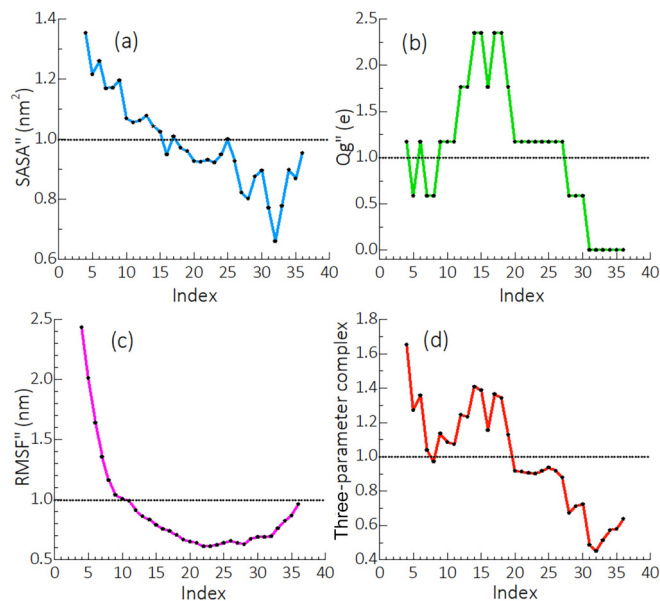


FIG. 6. Profiles of single and three-parameter complex metrics of C-terminal amidated Exendin-4 in the folded state: $SASA''$ (a), Qg'' (b), $RMSF''$ (c), and the three-parameter complex (d).

which is similar to the folded structure of natural Exendin-4. However, C-terminal amidated Exendin-4 has a longer N-terminal loop (His1-Asp9) than natural Exendin-4.

The SASA, Qg , and RMSF of each residue are shown in Figs. 5(b)–5(d). The average SASAs of His1, Glu3, Thr5-Leu10, Lys12-Glu16, Arg20, Phe22, Lys27-Asn28, Ser32-Ser33, Pro36-Pro37, and Ser39 are greater than 1.0 nm^2 . Since Ser39 is amidated, it is uncharged. Thus, C-terminal amidated Exendin-4 has the same charge distribution as natural Exendin-4, except for Ser39. Most residues have an RMSF greater than 0.2 nm . The His1-Thr5 and Ser39 fragments exhibit particularly strong fluctuations. Overall, the RMS fluctuation of C-terminal amidated Exendin-4 is stronger than that of natural Exendin-4. This indicates that C-terminal amidated Exendin-4 has better flexibility than natural Exendin-4.

Figure 6 shows the profiles of $SASA''$, Qg'' , $RMSF''$, and the three-parameter complex metrics. The $SASA''$ profile displays a major peak in indices 4–15. Except for index 36, the Qg'' profile is nearly identical to that of natural Exendin-4 in the folded state. That is, a major peak occurred in indices 12–19. The $RMSF''$ profile shows a clear peak at indices 4–9. From the profile of the three-parameter complex, two peaks can be observed in indices 4–7 and 9–19. This profile is similar to that of natural Exendin-4 in the folded state. However, it should be noted that the first peak has relatively high values of the three-parameter complex metrics.

Figure 7 displays the predicted epitopes of C-terminal amidated Exendin-4 in the folded state. The epitopes are determined to be composed of the His1-Thr7 and Asp9-Val19 fragments, according to the three-parameter complex metrics. Combined with the analysis of the three single metrics, the N-terminal epitope (His1-Thr7) is found to possess good solvent accessibility and flexibility. The main characteristic of the middle epitope (Asp9-Val19) is its charge-dense distribution. In the folded state, the middle epitope of C-terminal

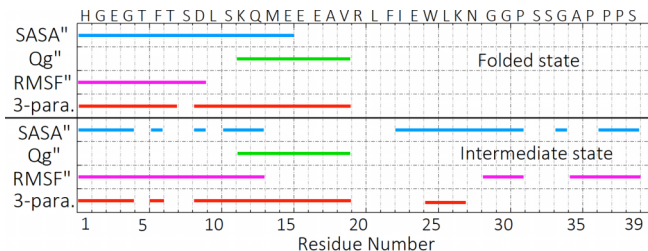


FIG. 7. Epitope predictions of C-terminal amidated Exendin-4 using different metrics, including $SASA''$, Qg'' , $RMSF''$, and the three-parameter complex (3-para.).

amidated Exendin-4 contains the same residues as that of natural Exendin-4. However, the fragment of its N-terminal epitope is longer than that of natural Exendin-4. Essentially, this comes from C-terminal amidated Exendin-4 which has a longer N-terminal loop.

D. Epitopes of C-terminal amidated Exendin-4 in the intermediate state

We continue our exploration of the epitopes of C-terminal amidated Exendin-4 in the intermediate state. Figure 8(a) shows the typical structure of the intermediate state. This structure is clearly distinct from the folded structure of C-terminal amidated Exendin-4. In the intermediate state, the Phe6-Leu10 and Met14-Leu26 fragments form two α helices. The His1-Thr5 and Ser11-Gln13 fragments are flexible loops. It's worth noting that the Leu21-Pro38 fragment does not fold into a Trp cage structure. This results in the exposure of the Trp25 sidechain to the solvent.

Figures 8(b)–8(e) show the SASA and RMSF of each residue, as well as their differences from the corresponding values in the folded state ($\Delta SASA$, $\Delta RMSF$). These residues, which include His1, Glu3, Phe6, Asp9-Leu10, Lys12, Met14-Glu16, Val19-Arg20, Phe22, Trp25-Asn28, Pro31, Ser33, Pro36-Pro37, and Ser39, have relatively large SASA. The SASAs of Asp9-Leu10, Trp25-Leu26, Asn28-Pro31, Ser33, and Pro37 increase, compared to the folded state. The RMSFs of His1-Thr7, Asp9-Glu15, Lys27-Gly34, and Pro36-Ser39 are greater than 0.2 nm . The fluctuations of His1-Thr5 become weak in the intermediate state.

Figures 8(f)–8(h) show the profiles of $SASA''$, $RMSF''$, and the three-parameter complex metrics. The $SASA''$ profile reveals six peaks at indices 4, 6, 9, 11–13, 23–31, and 34. In the $RMSF''$ profile, three peaks appear at indices 4–13, 29–31, and 35–36. The Qg'' profile of the intermediate state is identical to that of the folded state. After compounding three single metrics, four peaks can be observed, located at indices 4, 6, 9–19, and 25–27, respectively.

Figure 7 depicts the predicted epitopes of C-terminal amidated Exendin-4 in the intermediate state. The analysis of the three-parameter complex metrics reveals that the epitopes comprise of His1-Gly4, Phe6, Asp9-Val19, and Trp25-Lys27. The N terminal and middle epitopes are almost identical to those in the folded state of C-terminal amidated Exendin-4. However, the intermediate state has one additional epitope (Trp25-Lys27) compared to the folded state. Moreover, this epitope is absent for natural Exendin-4. The reason for this

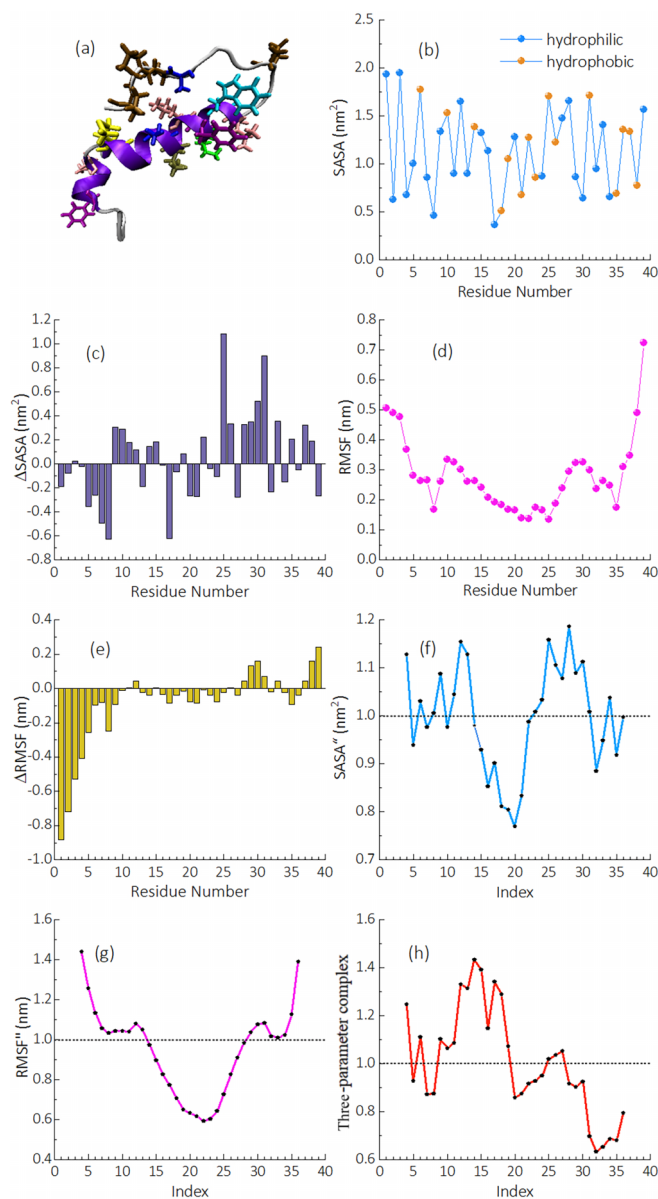


FIG. 8. The structure and properties of C-terminal amidated Exendin-4 in the intermediate state. (a) The typical structure of intermediate state. (b)–(e) SASA and RMSF for each residue and their changes compared to those in the folded state: SASA (b), Δ SASA (c), RMSF (d), Δ RMSF (e). (f)–(h) Profiles of SASA'', RMSF'', and the three-parameter complex metrics.

difference is that there is no Trp cage in the intermediate state of C-terminal amidated Exendin-4. This brings about high SASA'' values for the Trp25-Lys27 fragment.

E. Comparison with other computational prediction tools

The validity of our prediction method can be confirmed by comparing our results to those of other computational prediction tools [51,52,55]. In Fig. 9, we present the epitopes of Exendin-4 identified by various prediction tools. Additionally, our predictions for natural and C-terminal amidated Exendin-4 in their folded states are included in the plot for comparison. The first six methods are sequence-based and only require

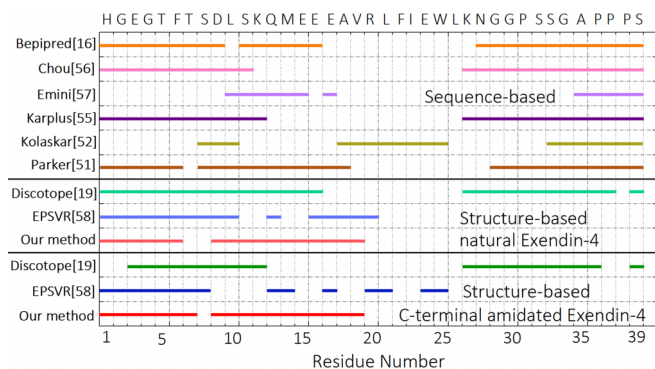


FIG. 9. Comparison of Exendin-4 epitopes predicted by different methods: sequencebased (from Bepipred to Parker), structure-based (Discotope, EPSVR, and our method). Epitopes predicted based on the structure of natural Exendin-4 are shown in lines 7–9, while epitopes predicted based on the structure of C-terminal amidated Exendin-4 are presented in lines 10–12.

the FASTA files of antigens as input data [16,51,52,55–57]. However, they could not distinguish between the natural and C-terminal amidated forms of Exendin-4 due to their identical sequence. These predicted epitopes are located in three regions of Exendin-4: the N-terminal loop, the middle helix, and the C-terminal extension (Gly29-Ser39). Our prediction for the N-terminal epitope (His1-Phe6 or His1-Thr7) is supported by the predictions from four tools. The epitopes predicted by the six tools for the middle helix have partial overlap with our predicted middle epitope (Asp9-Val19). Our predictions do not include the C-terminal epitope of Exendin-4 predicted by these tools. It's worth noting that the epitopes predicted by our method agree reasonably well with the N-terminal and middle epitopes predicted by Parker's method, which utilizes hydrophilic high-performance liquid chromatography parameters.

Discotope and EPSVR are structure-based prediction methods [19,58]. The epitopes for natural and C-terminal amidated Exendin-4 are predicted by the two tools, using PDB files (1JRJ.pdb and 7MLL.pdb) as input. For natural Exendin-4, the N-terminal epitope predicted by our method is fully covered by the predictions from Discotope and EPSVR. The middle epitope predicted by the two tools partially overlaps with our prediction. Similar results can be seen in the predicted epitopes for C-terminal amidated Exendin-4. Notably, EPSVR also does not predict the C-terminal epitope for both natural and C-terminal amidated Exendin-4. An examination of the folded structures of the two peptides reveals that the hydrophobic sidechains in their C-terminal extensions are inwardly buried, forming a hydrophobic cluster. This reduces the interaction between the C-terminal extensions and antibodies, decreasing the likelihood of the C-terminal extensions serving as epitopes. Thus, it is more plausible to conclude that the epitopes of Exendin-4 in the folded state do not include the residues of the C-terminal extension.

The comparison and analysis of Exendin-4 epitope predictions using several tools demonstrate the accuracy and reliability of our dynamic structure-based prediction method. In contrast to sequence-based methods, our approach effectively captures the differences in epitopes between natural

Exendin-4 and C-terminal amidated Exendin-4. Unlike static structure-based methods, our method emphasizes the significance of conformational flexibility and dynamics in epitope recognition. Additionally, our method outperforms Discotope and EPSVR by not being constrained by the richness of the epitope database and employing a more concise set of evaluation parameters (EPSVR uses six scoring terms). Importantly, our method goes beyond predicting folded-state epitopes and is capable of identifying epitopes in any intermediate state.

IV. CONCLUSION

The identification of epitopes is crucial for drug discovery, vaccine design, and immunotherapy. To improve the efficiency of epitope recognition, we have developed a computer-aided prediction method based on the dynamic structure of protein antigens. The method has been applied to predict the epitopes of Exendin-4. Our algorithm incorporates three residue properties (SASA, Qg, and RMSF) as prediction parameters. The results confirm that the epitopes predicted using the three-parameter complex metrics accurately highlight the key epitope residues predicted by single parameter metrics. In addition to the folded state of Exendin-4, our

method can also analyze the epitopes in the intermediate state. We have found that the epitopes in the intermediate state are different from those in the folded state when the structure of the intermediate state undergoes significant changes.

Our study has revealed the differences in epitopes between the natural and C-terminal aminated forms of Exendin-4. In the folded state, the epitopes of natural Exendin-4 (His1-Phe6 and Asp9-Val19) are nearly identical to those of C-terminal aminated Exendin-4 (His1-Thr7 and Asp9-Val19). However, in the intermediate state, the epitopes of natural Exendin-4 (His1-Gly4, Phe6, and Lys12-Arg20) cover fewer residues clearly when compared to those of C-terminal aminated Exendin-4 (His1-Gly4, Phe6, Asp9-Val19, and Trp25-Lys27). As a result, the difference in epitopes results in lower antibody affinity for natural Exendin-4, leading to lower immunogenicity and weaker resistance to drugs compared to C-terminal aminated Exendin-4.

ACKNOWLEDGMENT

We thank Prof. Xubiao Peng of the Beijing Institute of Technology for helpful discussions. J.H. is grateful for financial support from the Beijing Genetech Pharmaceutical Co., Ltd.

-
- [1] R. A. Goldsby, T. J. Kindt, B. A. Osborne, and J. Kuby, *Immunology*, 5th ed. (W. H. Freeman, New York, 2002).
- [2] J. Huang and W. Honda, *BMC Immunol.* **7**, 7 (2006).
- [3] D. J. Barlow, M. S. Edwards, and J. M. Thornton, *Nature (London)* **322**, 747 (1986).
- [4] B. Deng, C. Lento, and D. J. Wilson, *Anal. Chim. Acta* **940**, 8 (2016).
- [5] G. R. Masson, M. L. Jenkins, and J. E. Burke, *Expert Opin. Drug Discov.* **12**, 981 (2017).
- [6] S. Parvizpour, M. Pourseif, J. Razmara, M. Rafi, and Y. Omid, *Drug Discov. Today* **25**, 1034 (2020).
- [7] H. Matsuo, K. Kohno, H. Niihara, and E. Morita, *J. Immunol.* **175**, 8116 (2005).
- [8] G. Nybakken, T. Oliphant, S. Johnson, S. Burke, M. Diamond, and D. Fremont, *Nature (London)* **437**, 764 (2005).
- [9] G. Adams and L. Weiner, *Nat. Biotechnol.* **23**, 1147 (2005).
- [10] L. Conforti, *Clin. Immunol.* **142**, 105 (2012).
- [11] L. Potocnakova, M. Bhide, and L. Pulzova, *J. Immunol. Res.* **2016**, 6760830 (2016).
- [12] M. Linnebacher, P. Lorenz, C. Koy, A. Jahnke, N. Born, F. Steinbeck, J. Wollbold, T. Latzkow, H. Thiesen, and M. Glocker, *Anal. Bioanal. Chem.* **403**, 227 (2012).
- [13] E. Davidson and B. Doranz, *Immunology* **143**, 13 (2014).
- [14] H. Sun, L. Ma, L. Wang, P. Xiao, H. Li, M. Zhou, and D. Song, *Anal. Bioanal. Chem.* **413**, 2345 (2021).
- [15] L. Wang and M. Yu, *Curr. Drug Targets* **5**, 1 (2004).
- [16] J. Larsen, O. Lund, and M. Nielsen, *Immunome Res.* **2**, 2 (2006).
- [17] S. Saha and G. Raghava, *Proteins* **65**, 40 (2006).
- [18] Z. Chen, J. Li, and L. Wei, *Artif. Intell. Med.* **41**, 161 (2007).
- [19] P. H. Andersen, M. Nielsen, and O. Lund, *Protein Sci.* **15**, 2558 (2006).
- [20] J. Ponomarenko, H. Bui, W. Li, N. Fussedder, P. Bourne, A. Sette, and B. Peters, *BMC Bioinf.* **9**, 514 (2008).
- [21] I. Hoof, B. Peters, J. Sidney, L. Pedersen, A. Sette, O. Lund, S. Buus, and M. Nielsen, *Immunogenetics* **61**, 1 (2009).
- [22] M. Nielsen and O. Lund, *BMC Bioinf.* **10**, 296 (2009).
- [23] J. L. Pellequer, E. Westhof, and M. H. Regenmortel, *Method. Enzymol.* **203**, 176 (1991).
- [24] M. J. Blythe and D. R. Flower, *Protein Sci.* **14**, 246 (2005).
- [25] H. W. Wang and T. W. Pai, in *Immunoinformatics*, 2nd ed., edited by R. K. De and N. Tomar (Humana Press, New Jersey, 2014), Vol. 1184, pp. 217–236.
- [26] J. V. Kringelum, C. Lundegaard, O. Lund, and M. Nielsen, *PLoS Comput. Biol.* **8**, e1002829 (2012).
- [27] R. Rakhit, J. Robertson, C. V. Velde, P. Horne, D. M. Ruth, J. Griffin, D. W. Cleveland, N. R. Cashman, and A. Chakrabarty, *Nat. Med.* **13**, 754 (2007).
- [28] S. Yamada, N. D. Ford, G. E. Keller, W. C. Ford, H. B. Gray, and J. R. Winkler, *Proc. Natl. Acad. Sci. USA* **110**, 1606 (2013).
- [29] C. Cecconi, E. A. Shank, C. Bustamante, and S. Marqusee, *Science* **309**, 2057 (2005).
- [30] J. Eng, W. A. Kleinman, L. Singh, G. Singh, and J. P. Raufman, *J. Biol. Chem.* **267**, 7402 (1992).
- [31] A. A. Young, B. R. Gedulin, S. Bhavsar, N. Bodkin, C. Jodka, B. Hansen, and M. Denaro, *Diabetes* **48**, 1026 (1999).
- [32] O. G. Kolterman, J. B. Buse, M. S. Fineman, E. Gaines, S. Heintz, T. A. Bicsak, K. Taylor, D. Kim, M. Aisporna, Y. Wang, and A. D. Baron, *J. Clin. Endocrinol. Metab.* **88**, 3082 (2003).
- [33] R. A. DeFronzo, R. E. Ratner, J. Han, D. D. Kim, M. S. Fineman, and A. D. Baron, *Diabetes Care* **28**, 1092 (2005).

- [34] J. W. Neidigh, R. M. Fesinmeyer, K. S. Prickett, and N. H. Andersen, *Biochemistry* **40**, 13188 (2001).
- [35] S. H. Mishra, S. Bhavaraju, D. R. Schmidt, and K. L. Carrick, *J. Pharm. Biomed. Anal.* **203**, 114136 (2021).
- [36] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, *SoftwareX* **1-2**, 19 (2015).
- [37] M. Bonomia, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello, *Comput. Phys. Commun.* **180**, 1961 (2009).
- [38] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha *et al.*, *J. Phys. Chem. B* **102**, 3586 (1998).
- [39] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- [40] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *J. Chem. Phys.* **103**, 8577 (1995).
- [41] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge, UK, 2007).
- [42] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, *J. Chem. Phys.* **81**, 3684 (1984).
- [43] M. Parrinello and A. Rahman, *J. Appl. Phys.* **52**, 7182 (1981).
- [44] G. Bussi, A. Laio, and M. Parrinello, *Phys. Rev. Lett.* **96**, 090601 (2006).
- [45] A. Laio and F. L. Gervasio, *Rep. Prog. Phys.* **71**, 126601 (2008).
- [46] H. Grubmüller, B. Heymann, and P. Tavan, *Science* **271**, 997 (1996).
- [47] C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1997).
- [48] R. Casasnovas, V. Limongelli, P. Tiwary, P. Carloni, and M. Parrinello, *J. Am. Chem. Soc.* **139**, 4780 (2017).
- [49] R. B. Best, G. Hummer, and W. A. Eaton, *Proc. Natl. Acad. Sci. USA* **110**, 17874 (2013).
- [50] F. Pietrucci and A. Laio, *J. Chem. Theory Comput.* **5**, 2197 (2009).
- [51] J. M. Parker, D. Guo, and R. S. Hodges, *Biochemistry* **25**, 5425 (1986).
- [52] A. S. Kolaskar and P. C. Tongaonkar, *FEBS Lett.* **276**, 172 (1990).
- [53] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf, *J. Comput. Chem.* **16**, 273 (1995).
- [54] E. Westhof, D. Altschuh, D. Moras, A. C. Bloomer, A. Mondragon, A. Klug, and M. Regenmortel, *Nature (London)* **311**, 123 (1984).
- [55] P. A. Karplus and G. E. Schulz, *Naturwissenschaften* **72**, 212 (1985).
- [56] P. Chou and G. D. Fasman, *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**, 45 (1978).
- [57] E. A. Emini, J. V. Hughes, D. S. Perlow, and J. Boger, *J. Virol.* **55**, 836 (1985).
- [58] S. Liang, D. Zheng, C. Zhang, and M. Zacharias, *BMC Bioinf.* **10**, 302 (2009).