Identifying signatures of proteolytic stability and monomeric propensity in O-glycosylated insulin using molecular simulation

Wei-Tse Hsu · Dominique A. Ramirez · Tarek Sammakia · Zhongping Tan · Michael R. Shirts

Received: 29 October 2021 / Accepted: 6 April 2022 / Published online: 4 May 2022
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

Abstract
Insulin has been commonly adopted as a peptide drug to treat diabetes as it facilitates the uptake of glucose from the blood. The development of oral insulin remains elusive over decades owing to its susceptibility to the enzymes in the gastrointestinal tract and poor permeability through the intestinal epithelium upon dimerization. Recent experimental studies have revealed that certain O-linked glycosylation patterns could enhance insulin’s proteolytic stability and reduce its dimerization propensity, but understanding such phenomena at the molecular level is still difficult. To address this challenge, we proposed and tested several structural determinants that could potentially influence insulin’s proteolytic stability and dimerization propensity. We used these metrics to assess the properties of interest from 10 μs aggregate molecular dynamics of each of 12 targeted insulin glyco-variants from multiple wild-type crystal structures. We found that glycan-involved hydrogen bonds and glycan-dimer occlusion were useful metrics predicting the proteolytic stability and dimerization propensity of insulin, respectively, as was part of the solvent-accessible surface area of proteolytic sites. However, other plausible metrics were not generally predictive. This work helps better explain how O-linked glycosylation influences the proteolytic stability and monomeric propensity of insulin, illuminating a path towards rational molecular design of insulin glycoforms.

Keywords Insulin · O-glycosylation · Stability · Protein degradation · Molecular dynamics

Introduction
Insulin has been widely used as a peptide drug to treat both type 1 and type 2 diabetes mellitus by promoting the absorption of glucose from the blood into the liver, fat, and skeletal muscle cells. Insulin is usually administered via subcutaneous injections, but excessive injections could lead to non-compliance by patients owing to injection pain and associated side effects, including trypanophobia, lipodystrophy, and peripheral hyperinsulinemia [1]. As such, there has been a growing interest in the development of an insulin drug for oral administration [1–3]. Orally ingested insulin would avoid the aforementioned side effects of subcutaneous administration. In addition, it has the advantage of reaching the liver at high concentration via the portal vein before reaching systemic circulation, which better mimics the physiology of endogenous secretion by the pancreas and provides a better glucose homeostasis [4, 5]. However, developing oral insulin remains a major challenge because of its susceptibility to proteases in the digestive system of the human body and poor permeability across the intestinal epithelium in oligomeric form, which leads to overall low oral bioavailability [6]. Therefore, proteolytic stability and dimerization propensity are two of the most important factors to address when developing oral insulin drugs. A wide array of strategies, including chemical modification [7, 8], nano-particulate carrier systems [9–11], and utilization of protease inhibitors [12], have been shown to have varying degrees of success in improving these two insulin properties.
Glycosylation is a chemical modification strategy that has shown significant promise in enhancing the proteolytic stability and monomeric propensity of various kinds of proteins [13–15]. While N-linked glycosylation is better-studied [16–18] due to better controllability and the ease of chemical synthesis, O-linked glycosylation has a larger diversity of carbohydrate structures and potentially a greater ability for tuning biophysical properties of glycoproteins. Guan et al. systematically investigated different insulin glycosylation patterns, including glycosylation sites, glycan sizes, and glycan structures [19]. Their study suggests that O-mannosylation of insulin B-chain Thr27 better enhances proteolytic stability and reduces dimerization propensity compared to other glycosylation patterns, all while retaining full biological activity. However, there is a lack of quantitative understanding of how insulin properties are influenced by the interactions between the attached sugar molecules and insulin residues, particularly at the molecular level.

Recent advances in molecular simulation methods and greater availability of computing resources allow much more comprehensive exploration of biomolecular systems at the atomistic level than was previously possible. A number of previous studies have shown molecular dynamics (MD) simulations to be a powerful tool for understanding the properties and dynamics of insulin. Mark et al. [20] and Zoete et al. [21] found that the insulin monomer is more flexible than the dimer. Zoete et al. [21] additionally confirmed the high flexibility in the B-chain C-terminus of insulin, which was consistent with the experimental data. Later, Yang [22] et al. used simulated annealing [23] to probe how PEGylation enhanced the stability and potency of insulin, discovering that an optimal chain length existed for PEGlyated insulin pharmaceuticals. More recently, with steered molecular dynamics [24] and replica-exchange umbrella sampling [25], Antoszewski et al. [26] revealed different pathways of insulin dimer dissociation and characterized relevant energetic and structural features of the process. Busto-Moner et al. [27] presented a pipeline integrating parallel tempering [28, 29], the integrated variational approach for conformational dynamics (IVAC) [30, 31], Markov state model (MSM) [32, 33], and Perron cluster analysis (PCCA) [34] for sampling the wild-type insulin ensemble and they concluded three important elements of disorder common in the insulin ensemble at a low pH value.

In this study, we aim to identify structural determinants from MD simulations of glycosylated insulin that lead to enhanced proteolytic stability against α-chymotrypsin or reduced dimerization propensity. We performed molecular dynamics simulations for each of the 12 insulin glycoforms (GFs) (Fig. 1) studied in the experimental work [19] done by Guan et al. Each of the glycoforms was built upon five different wild-type models resolved by different methods/groups to encapsulate a wider variety of initial configurations so as to sample the conformational ensemble more comprehensively. We proposed several potential metrics for assessing the proteolytic stability and dimerization propensity based on the proteolytic degradation mechanism by α-chymotrypsin [35] and the structural characteristics of insulin. These metrics include solvent-accessible surface area (SASA), secondary structures, the existence of specific glycosylation sites, and the conformational flexibility of the insulin dimer. By analyzing these metrics, we were able to identify key structural determinants that contribute to enhanced stability and reduced dimerization propensity.

![Fig. 1 Structures of a human insulin monomer and glycoforms studied in this work.](https://example.com/fig1.png)

Fig. 1 Structures of a human insulin monomer and glycoforms studied in this work. Two types of sugar moieties (N-acetylgalactosamine (GalNAc) and mannose (Man) in the α-anomer) with varying lengths (e.g. a mannose monomer, dimer, or trimer) were attached to five different glycosylation sites of insulin, including SerA9 (teal), SerA12 (purple), SerB9 (blue), ThrB27 (orange) and ThrB30 (green) all represented as beads. The glycosylation pattern of each glycoform is implied by its name—for example, GalNAcα (SerA9) represents the glycoform having an N-acetylgalactosamine (GalNAc) attached to the A chain Ser9 residue. Note that glycoforms having a mannose dimer or trimer as the glycan all have a α-1,2-linkage, so Tri-Manα (ThrB27) refers to the glycoform containing an α-1,2-linked tri-mannose at the B chain Thr27 residue.
of glycan-involved hydrogen bonds, and the occlusion between the glycan and the dimer interface. Notably, in preliminary studies, we found that the conformational rigidity of insulin with or without glycosylation was statistically indistinguishable. This is in contrast to a previous study [36] found that greater conformational rigidity of a carbohydrate binding module (CBM) protein could potentially disfavor reaction-capable transition states in proteolysis. Therefore, we instead adopted $\beta$-sheet propensity as a more precise measure of transition state accessibility, which specifically considered the structural features of the known $\alpha$-chymotrypsin transition state.

Importantly, both the proteolytic stability and dimerization propensity are related to configurational processes that require overcoming free energy barriers. The proteolytic stability is directly associated with the free energy that require overcoming free energy barriers. The proteolysis propensity are related to configurational processes corresponding to the transition state in the proteolytic degradation by $\alpha$-chymotrypsin. Similarly, dimerization propensity is strongly correlated with the free energy barrier to dimerization. While these free energy differences could serve as more direct measures for the two biophysical properties of interest, calculations of such free energy differences along transition pathways are far from trivial. The reason lies in the fact that the timescale of the binding/unbinding events between an insulin monomer and an $\alpha$-chymotrypsin or between two insulin monomers are prohibitively long. This necessitates the use of advanced sampling techniques to accelerate the sampling of the configurational space of the system of interest, such as umbrella sampling [37] or alchemical transformations [38, 39]. However, such free energy methods are usually much more complicated and computationally expensive than standard MD simulations of the same length. The increase in the system size of insulin/protease complex or insulin dimer calculations adopting either method could easily increase the computational cost significantly beyond the scope of what is possible to screen large numbers of insulin modifications. Advanced configurational sampling approaches can often be significantly difficult to set up and interpret, thus making easy screening impossible.

Therefore, instead of considering simulations of an insulin-$\alpha$-chymotrypsin complex and an insulin dimer, we worked to develop metrics for proteolytic stability and dimerization propensity based on standard MD simulations of monomers and their glycosylated analogs. The hypothesis is that the insulin monomer that participates in either a complex with $\alpha$-chymotrypsin or an insulin dimer should encode at least some important structural features of both the proteolytic stability and dimerization propensity. As long as we are able to sufficiently explore the configurational space of monomer-based structures, even if not completely, the conformational ensembles we get from the simulations should help us develop reasonable metrics for the two insulin properties.

All proposed metrics were measured for MD trajectories generated from each glycoform structure. They were later compared with the experimental work for assessing putative predictive potential. Notably, we found that at minimum $5 \times 2000$ ns = $10 \mu$s of simulation was necessary to capture the structural characteristics of the insulin ensembles. This aggregate length is, to our knowledge, one of the longest among all the computational studies of human insulin. In addition, our study is not only one of the very few studies that characterize conformations of insulin with a covalently attached moiety, but is also unique as the first study that correlates molecular dynamics simulations with the proteolytic stability of these proteins. Our investigation facilitates a better understanding of the underlying mechanism of how different $O$-glycosylation patterns influence insulin properties, which could be useful in guiding the design of insulin glyco-variants with better properties for oral delivery.

Methodology

Molecular dynamics simulations

To reduce sampling bias of our MD simulations and ensure that our analysis results were not dependent on the initial configuration of insulin, we built each of the 12 glycoforms on 5 wild-type (WT) structures resolved by different groups using different methods, whose PDB IDs are 4EYD, 4EY9, 4EY1, 3I3Z, and 2MVC, respectively (Fig. 2A). These initial models were chosen based on whether they were resolved in a complex or as a monomer. For example, 4EYD, 4EY9, and 4EY1 were all crystallized in complex with Zn$^{2+}$ as a hexamer composed of a dimer in 3-fold symmetry [40]. They represent high-resolution human insulin structures from pharmaceutical formulations. 3I3Z and 2MVC, on the other hand, are representative of monomeric insulin structures. 3I3Z was resolved by X-ray crystallography under low gravity conditions so that the asymmetric unit of insulin was a monomer [41]. 2MVC was resolved by NMR spectroscopy under acidic conditions, which is known to favor the monomeric form of insulin [42].

For those models resolved in the dimeric form, including 4EYD, 4EY9, and 4EY1, we extracted the insulin monomer from the dimer conformation. For 3I3Z, which was resolved as a monomer with the last residue on the B-chain mutated into an alanine, we used PyMOL [43] to mutate the residue back to a threonine to match the standard sequence of human insulin. Lastly, for the NMR-resolved 2MVC, we simply took the first model from the PDB file of the resolved ensemble. Notably, even if monomers extracted from a dimer conformation could potentially suffer from biases from
After extracting the monomer structure from each of the initial models, we used the H++ server (version 3.2) [44–46], which parameterized the protein with the AMBER ff14sb force field, to assign reasonable protonation states corresponding to the pH values adopted in experiments. Specifically, in the work [19] by Guan et al., the experiments of \( \alpha \)-chymotrypsin digestion (for estimating the \( \alpha \)-chymotrypsin half-life, hence the proteolytic stability) and analytical ultracentrifugation (for estimating the monomer/dimer population, hence the dimerization propensity) were conducted at pH 8.0 and pH 7.4, respectively. These two pH values respectively correspond to the typical pH values in the pancreatic tract (pH 7.5 to pH 8.0 [47]) and small intestine (pH 6 to pH 7.4 [48]) of human body. Around these pH values, the total charges of insulin were predicted to be -1 or -2, depending on the \( pK_{\text{a}} \) value of the two histidine residues (HisB5 and HisB10) of insulin. We then adjusted the external pH value in H++ of each wild-type structure to give total charges of -1. This value better corresponds to pH values ranging from 6.6 to 7.8 (see Supplemental Table S3), which overlaps with the pH ranges in both the small intestine and the pancreatic tract than a charge of -2 overlaps.

Simulated structures from 4EYD, 4EY1, and 3I3Z had exactly the same protonation state for each residue. 4EY9 and 2MVC, on the other hand, were found to have different histidines protonated compared to the ones in the other three models (see Supplemental Table S3). We chose to include this ensemble of protonation sites, as they were essentially attributable to the orientations of the histidine residues and their surroundings. In addition, these two histidine residues are far away from the residues involved in the hypotheses of our analysis methods (see Supplemental Fig. S6), making them less likely to have noticeable influences on the predictors we developed. We started from these parameterized wild-type insulin conformations with reasonable protonation states and used GLYCAM glycoprotein builder [49], which used the GLYCAM06j-1 force field, to build various glycoform structures by attaching different saccharide moieties to the corresponding glycosylation sites.

In this study, the simulations of all wild-type and glycoform structures were performed using GROMACS 2020.4 [50, 51] at 310.15 K, which was in agreement with experimental temperature. Outputs in AMBER formats generated by GLYCAM glycoprotein builder were converted into GROMACS formats using ACPYPE [52] to serve as the inputs of the simulations. Each structure was solvated in a dodecahedral box with 1.0 nm between the solute and the box edge. Sodium and chloride ions were added to neutralize the system and match the specified salinity of 0.15 M. The system was then energy minimized with the steepest descent algorithm until the maximum force was lower than 100.0 kJ/mol/nm. Subsequently, a 200 ps NVT equilibration followed by a 200 ps NPT equilibration was carried out, in which velocity rescaling [53] and the Berendsen barostat [54] were employed to maintain the reference temperature and pressure at 310.15K and 1 bar, respectively. Finally, an MD simulation was performed in an NPT ensemble, with the pressure
maintained at 1 bar by the Parinello-Rahman barostat [55, 56]. The cut-off distances for van der Waals interactions and Coulomb interactions were both set as 0.9 nm, with the switching of the van der Waals potential between 0.85 and 0.9 nm and an analytical correction for long-range dispersion. The particle mesh Ewald algorithm [57] was used with real-space switching of the potential between 0.89 nm and 0.9 nm. LINCS [58] was used to constrain hydrogen bonds. We extended all simulations up to 2000 ns, which we deemed necessary to capture major transitions between metastable states occurring on the time scale of 500–2000 ns, as observed from the time series of the pairwise RMSD calculations of the wild-type structures (see Supplemental Fig. S1). Our simulations captured the key disordered elements investigated by Busto-Moner et al. [27], though at different percentages. This discrepancy is not surprising given the pH difference between the two studies, with the differences elaborated in “Introduction” Section in the supporting information. All trajectories were stored every 250 ps, for a total of 8001 frames for analysis. All the input configurations and GROMACS mdp files are provided in the GitHub repository for this study.

Analysis techniques

Proteolytic degradation

**Metric 1: SASA of the scissile bonds** α-chymotrypsin is a common digestive protease secreted by the pancreas which performs proteolysis in the duodenum [59]. Experimental work [19] used α-chymotrypsin to assess the proteolytic stability by measuring the half-life of different insulin glycoforms. In proteolytic degradation of insulin by α-chymotrypsin, residues including TyrA14, TyrA19, TyrB16, PheB25, and TyrB26 serve as the cleavage sites, whose scissile bonds are on the carboxyl side [35]. Previous research into α-chymotrypsin [35] indicated that residues PheB25 and TyrB26 were the cleavage sites considered to be the most susceptible to proteolysis. We therefore hypothesized that the solvent exposure of the scissile bonds of these two residues was positively correlated with proteolytic susceptibility.

To quantify the solvent exposure of these sites, we used the double cubic lattice method (DCLM) [60] to calculate the SASA of the CONH atoms, which were the atoms sharing the same plane with the scissile bond, between PheB25 and TyrB26, and between TyrB26 and ThrB27. In DCLM, the accessible surface is defined by tracing the center of the probe sphere (with the radius as the van der Waals radius of water) as it rolls along the van der Waals surface of the solute. We calculated the mean of all SASA time series after discarding non-equilibrium region and decorrelation [61]. Finally, the reported value for each glycoform is the value averaged across the five different wild-type bases, with the uncertainty the standard deviation of the 5 runs. We then plotted the α-chymotrypsin half-life measured in the experimental work against these averaged SASA values in a correlation plot, with the Kendall’s tau correlation coefficient (τ) [62] and its uncertainty annotated. The uncertainty of the correlation coefficient was estimated from bootstrapping, as detailed in Section Using bootstrap methods to estimate uncertainties in correlation coefficients. Note that we adopted Kendall’s tau in favor of Pearson correlation coefficient to only assume a monotonic relationship because it was not clear if the relationship between the two variables should be linear.

**Metric 2: SASA of the P1 sites** In addition to the cleavage sites PheB25 and TyrB26, we hypothesized that their adjacent residues along the N-terminal direction were also important. These residues are denoted as the P1 residues according to Schechter-Berger nomenclature [63], which defines the residues N-terminal to the cleavage site as P1, P2, P3, ... etc., and the ones in the opposite direction as P1', P2', P3' ... etc. (See Fig. 3.) Our hypothesis stems from the importance of the hydrophobicity of the P1 residue in the molecular recognition by α-chymotrypsin [64]. Specifically, the deep hydrophobic pocket formed by Ser189, Gly216, and Gly226 in α-chymotrypsin requires the P1 residue to be hydrophobic as well [65]. Given the fact that the P1 residue itself needed to be sufficiently solvent-exposed to contact and fit in this hydrophobic pocket of α-chymotrypsin, we hypothesized that a glycoform would be more proteolytically stable if its P1 sites were less solvent accessible due to the steric hindrance by the glycan moiety. Therefore, we calculated the residue SASA of PheB24 and PheB25, which were the P1 sites corresponding to the cleavage sites PheB25 and TyrB26, respectively. We again employed DCLM to calculate the SASA time series, which was then truncated and decorrelated with the same method as in Metric 1. Similarly, the final reported value of each glycoform is the value averaged across simulations starting from the five different wild-type structures, with the uncertainty being the standard deviation of the 5 runs. The monotonicity of the relationship between the metric and the experimental data was assessed by Kendall’s tau correlation coefficient (τ) [62], with its uncertainty calculated using the method described in Section Using bootstrap methods to estimate uncertainties in correlation coefficients. Additionally, in the supporting information (Fig. S3), we examined the correlations between any two of the four SASA measures in Metric 1 and Metric 2, including the scissile bond SASA of B26–B27 and P1 site SASA of residues B24 and B25.

**Metric 3: β-sheet propensity of the P1–P3 region** Previous experimental studies [65–67] of α-chymotrypsin complexes
revealed an invariant feature that the polypeptide binding sites of α-chymotrypsin formed a short antiparallel β-sheet with the backbone atoms of the P1–P3 sites of the substrate. In the case of insulin as the substrate, we considered the P1 to P3 sites corresponding to the two most important cleavage sites (PheB25 and TyrB26), which included ArgB22, GlyB23, PheB24, and PheB25. We hypothesized that glycoforms whose ArgB22 to PheB25 residues had lower β-sheet propensity would destabilize the transition state in the protease-substrate binding event, hence decreasing the proteolytic susceptibility. This assumption also stems from the fact that drastic structural re-orientations or even secondary structure transformations of the insulin substrate costs free energy and disfavors proteolysis accordingly. To evaluate the β-sheet propensity of these sites, we calculated the ψ and φ angles of the residues. We then used MDAnalysis [68, 69] to locate all the frames in a Ramachandran plot [70] and defined the β-sheet propensity as the fraction of the points in the β-sheet region defined by Lovell et al [71]. Again, for each glycoform, we calculated the fractions averaged across different wild-type bases, with the uncertainty reported as the standard deviation of the mean over simulations with the five different initial structures. Finally, we plotted α-chymotrypsin half-life against these averaged β-sheet fractions in a correlation plot, where the Kendall’s tau correlation coefficient (τ) and its uncertainty were determined using the bootstrap method described in Section Using bootstrap methods to estimate uncertainties in correlation coefficients. The Ramachandran plots of each of the 4 residues of interest of each glycoform can be found in our supporting information (Figs. S8 to S27). Additionally, in the supporting information (Figs. S4 and S5), we examined the correlations between any two of the four measures (B22-B25) of the β-sheet propensity and the correlations between any SASA measure in Metric 1 or Metric 2 and any β-sheet propensity measure in Metric 3.

**Metric 4: Existence of glycan-involved hydrogen bonds**

In addition to the consensus conformations of the binding sites, some of the hydrogen bonds between α-chymotrypsin and the substrate were also found crucial for efficient substrate proteolysis [65]. These hydrogen bonds include the ones involved in the antiparallel β-sheet formed between (1) the carbonyl oxygen of Ser214 and the amide NH of the P1 site, (2) the amide NH of Gly216 and the carbonyl oxygen of the P3 site, and (3) the carbonyl oxygen of Gly216 and the amide NH of the P3 site. Additionally, the main-chain carbonyl oxygen of Phe41 forms a hydrogen bond with the amide NH of the P2′ site. These four hydrogen bonds in the P3′–P3 region are commonly believed to form the canonical hydrogen-bonding network between insulin and α-chymotrypsin. With this, we hypothesized that the oxygen atoms of the glycan could compete with the α-chymotrypsin residues (Ser214, Gly216, and Phe41) as the acceptors to form hydrogen bonds with insulin residues (the P1–P3 sites, and the P2′ site), hence disturbing the hydrogen-bonding network and potentially enhancing the proteolytic resistance of the structure. To see what additional interactions were formed due to the presence of the glycan, we examined all the glycan-involved hydrogen bonds, not just the ones involving the P1–P3 and P2′ residues. We used MDtraj [72] to identify hydrogen bonds between the glycan moiety (as

![Cleavage site: PheB25](image1)

**Fig. 3** The P- and P′-sites of the two most important cleavage sites, residues PheB25 and TyrB26. The cleavage sites are shown as orange sticks, while the remaining part of the protein is shown in cartoon representation. Schechter-Berger nomenclature for residues surrounding the cleavage sites are shown in varying shades of magenta and are labeled P1-3 and P1′-2′.
the acceptor) and any insulin residue (as the donor) according to the Baker-Hubbard criterion [73], which identified a hydrogen bond only if the angle formed between the donor, hydrogen atom and the acceptor was larger than 120 degrees and the distance between the hydrogen atom and the acceptor was less than 2.5 angstrom at least 10% of the time. For each trajectory, we calculated the fraction of the time each hydrogen bond existed and averaged the fractions of the hydrogen bonds of each glycoform across the 5 different wild-type bases. The ones whose fraction is larger than 5% are reported, with the uncertainty being the standard deviation.

**Dimerization propensity**

The dimerization propensity metric was compared to experimental dimerization data from Guan et al. [19]. Dimerization data exists only for wild type, GF 9, GF 10, and GF 13, and thus comparisons between experimental and computational data proved challenging. We focused only on residues GlyB23–TyrB26 because these are the dimer interface residues that form backbone hydrogen bonds with another insulin monomer (Fig. 2B) [41, 74, 75].

One potential signature of dimerization propensity is the presence of dimer-characteristic structure in the monomer ensemble, which would reduce the free energy of reorganization on dimerization. We examined secondary structure propensity of the dimerization residues and found no consistent differences in either the dimer or the monomer between the different wild-type insulin bases, or between glycoforms themselves. This was also true after examining additional monomer structures 1JCO [76], 1MHJ [77], and 2JV1 [78]. This suggests to us that structural features of insulin residues alone might not be useful for determining dimerization propensity, so we explored other metrics to assess this property.

**Metric 1: Glycan-dimer occlusion** We considered steric occlusion of the dimer interface by the glycans as a possible metric for dimerization. Several glycan moieties are attached close to the explicit dimerization region GlyB23–TyrB26 (Fig. 1), and there are several glycans large enough to sterically interfere with the dimerization interface (Fig. 4). We hypothesized that glycans that occupy space close to these residues, with high frequency in simulation time, would preclude dimerization.

We defined glycan-dimer occlusion as any instance when at least one atom of the glycan moiety comes within 5 angstroms of any atom in the dimer interface GlyB23–TyrB26, including the atoms in the backbone and in the side chains. We used the Python package ProDy [79] to calculate the total number of atom neighbor pairs between the glycan and dimer interface for each frame in the trajectories.

Atom neighbor pair autocorrelation lag times (in ns) were estimated by fitting autocorrelation data to an exponential decay function using SciPy [80, 81] and are presented in Supplemental Table S4. There are several glycoform models that

---

![Fig. 4 Classification of occlusion and no-occlusion states. Representative frames from four different glycoform trajectories show occlusion (GF 13, GF 10) and no occlusion (GF 4, GF 8) states. Insulin monomer is presented in translucent blue-white cartoon, dimerization residues presented in salmon sticks, and glycan moiety presented in yellow spheres. The 3I3Z dimer structure is provided as the reference.](image-url)
have no lag time ("NA" in Supplemental Table S4) and this is because for these trajectories, no occlusion was found. The lag times were used to estimate the independent occlusion states sampled throughout the simulation by dividing the total simulation time by the subsequent lag times.

To simplify the occlusion analysis, we calculated the proportion of simulation frames that contained a glycan-dimer atom neighbor pair out of the total simulation frames, and ignored the absolute number of atom neighbor pairs in each frame so as to treat any number of interactions as possible occlusion. Since larger glycans will have more possible neighbors than smaller glycans, this also served to normalize the data to prevent a bias for the larger glycans. The 95% confidence intervals for these binomial proportions were estimated using the independent occlusion samples and the Wilson score method, which produces bounded asymmetric intervals and does not require normal approximations for its use [82–84]. The proportion of frames with occlusion for each glycoform were only compared within its respective set, because no wild-type control could be included as the wild type is not glycosylated.

**Using bootstrap methods to estimate uncertainties in correlation coefficients**

In the correlation plots presented in both the main text and the supporting information, we performed bootstrapping to estimate the uncertainties of correlation coefficients. In each of the 500 bootstrap iterations we performed, for each variant, we drew five bootstrap samples for the value of the correlation metric from the set of five different wild-type models, and five bootstrap samples for the experimental reference from normal distributions centered at the experimental values. From each of these 500 bootstrap iterations, we calculated one Kendall’s tau correlation coefficient by correlating the sample mean of the two variables. Lastly, we calculated the standard deviation of these 500 values and reported it as the uncertainty of the correlation coefficient.

The same bootstrap approach was used to estimate the uncertainty of the Pearson correlation coefficients annotated in Supplemental Fig. S3 to S5, except that the bootstrap samples for both metric variables of interest were drawn from values based on different wild-type models.

**Molecular visualization**

Molecular visualization, specifically for Figs. 1, 2, 3, 4, and S6, were done using PyMOL version 2.4.1 [43].

---

### Results and discussions

#### Proteolytic degradation

**Metric 1: SASA of scissile bonds**

According to our hypothesis related to the peptide bond SASA of the cleavage sites, glycoforms whose scissile bonds are less solvent-exposed should have higher proteolytic stability. To examine this hypothesis, we plotted the $\alpha$-chymotrypsin half-life measured in the experimental work against the SASA value of each of the two scissile bonds, including the one between residues B25 and B26 (upper panel of Fig. 5A) and the one between residues B26 and B27 (lower panel of Fig. 5A). Ideally, a good computational metric should be able to reproduce consistent results as compared to the work by Guan et al. [19], leading to a Kendall’s tau correlation coefficient close to -1. A good metric should also be immune to the starting model bias, including the bias from wild-type models resolved using different methods or under different conditions, or the bias solely from different protonation states of the histidine residues.

As a result, we conclude that the SASA of the scissile bonds is a weak predictor for proteolytic stability. GF 13 and GF 10, which were experimentally found to be more proteolytically stable than the wild type, did have a lower SASA value than the wild type at both sites. Notably, as the most proteolytically stable structure, GF 13 also had the lowest SASA values at both sites. However, GF 10 had the second-longest $\alpha$-chymotrypsin half-life, but not the second-lowest SASA values at both sites. At the scissile bond between B26 and B27, most glycoforms had a lower SASA value than that of the wild type, but some of them were experimentally identified as less proteolytically stable than the wild type. Overall, the SASA values of either scissile bond have roughly the same efficacy given similar values of Kendall’s tau correlation coefficients. However, if we are only interested in the comparison between the wild type and any glycoform instead of between any two of the glycoforms, the SASA value of the scissile bond between B25 and B26 is marginally more indicative than the SASA value of the other site. Specifically, among all the glycoforms that had shorter $\alpha$-chymotrypsin half-lives than the wild type, GF 3, GF 4, GF 7, and GF 8 indeed had higher SASA values at this site than the wildtype. As for GF 6 and GF 11, which also had shorter $\alpha$-chymotrypsin half-lives than the wild type, the SASA value at the B25-B26 scissile bond failed to classify them as the less proteolytically stable variants. Notably, the errors in the metric are generally large, which might be attributable to the fact that the SASA of the scissile bond was calculated from
only 4 atoms (CONH atoms) and could lead to larger fluctuations in nature. Regardless of whether these large error bars are a direct result of the dependence of the initial bias, the high uncertainty undermined the predictiveness of this metric, making it less useful.

**Metric 2: SASA of the P1 sites**

Similar to Metric 1, Metric 2 is partially predictive and has relatively large error bars for some of the variants.

As seen on the lower panel of Fig. 5B, the SASA of residue B25 is less predictive than the SASA of the other site. Although partial agreement with experimental results can be seen from GF 13, which had the lowest SASA at residue B25 and was previously found to be the most proteolytically stable, other glycoforms, regardless of the \( \alpha \)-chymotrypsin half-life, all had a lower value compared to the wild type. The Kendall’s tau correlation coefficient, which was close to 0, shows that the SASA of residue B25 was almost uncorrelated with the \( \alpha \)-chymotrypsin half-life.

The SASA of residue B24 (the upper panel of Fig. 5B), on the other hand, had a higher correlation with the proteolytic stability of the structure. Specifically, GF 13 and GF 10, the two most proteolytically stable glycoforms, had significantly lower SASA value at residue B24. As for the glycoforms that had a lower SASA value at this site, except for GF 7, all had comparable proteolytic stability to the wild type. This higher correlation can also be seen from the magnitude of the corresponding Kendall’s tau correlation coefficient, which was higher than those of Metric 1.

Overall, these observations imply moderate predictiveness of the SASA of residue B24, since a low value at this site generally indicates that the structure is very likely to have more, or at least comparable proteolytic stability compared to the wild type. Notably, most proteolytically unstable glycoforms, including GF 2, GF 3, GF 4, GF 6, GF 7, and GF8, had the same level of SASA at residue B24 as the wild type. This could be attributable to the situation where the residue of the wild type is already very solvent-exposed, leaving considerably less space for the proteolytically unstable glycoforms to have an even higher SASA value. This
limitation is intrinsic to both Metric 1 and Metric 2, motivating us to look for more reliable predictors for proteolytic stability.

**Metric 3: β-sheet propensity of the P1–P3 region**

Our hypothesis suggests that glycoforms whose P1–P3 region has a lower β-sheet propensity should potentially have higher proteolytic stability. However, most of the residues, including B22, B23, and B24 did not show this trend, while residue B25 was predictive at roughly the same level as Metric 1 since it showed the lowest β-sheet propensity for the most proteolytically stable structures (GF 10 and GF 13). Notably, even if one of the four residues exhibited moderate predictiveness, the generally large error bars in the β-sheet propensity undermine its usage as a metric to capture structural determinants that influenced the proteolytic stability of the structure. These large error bars in the case of B22 to B24 also imply that the structure at these sites can adopt a wide range of orientations, contradicting our hypothesis that the glycoforms able to prevent proteolytic degradation should have common orientations at these sites due to high free energy costs of structural re-orientations or transformations. Similar to the first two metrics, we note that these large error bars are not necessarily the results of the starting model bias, since the β-sheet propensity was calculated from the orientations of just a few atoms, which naturally caused larger fluctuations of the results. Since the metric is insufficient to assess the proteolytic stability, whether or not the large error bars can be ascribed to the starting model bias is not relevant (Fig. 6).

**Metric 4: existence of glycan-involved hydrogen bonds**

We hypothesized that glycoforms with more glycan-involved hydrogen bonds, especially the ones that involved the P1–P3 or the P2′ sites, could interfere with the canonical hydrogen-bonding network formed between the protease and the substrate, thus enhancing the proteolytic stability of the

---

**Fig. 6** The correlation plot between the α-chymotrypsin half-life and the average β-sheet propensity of residues B22 to B25. The β-sheet propensity of the P1–P3 region does not clearly correlate with α-chymotrypsin half-life. Different regions were colored to indicate glycoforms having longer, comparable, or shorter half-life as compared with the wild type. Uncertainty in the Kendall’s tau correlation coefficient (\(r\)) was estimated as described in Section Using bootstrap methods to estimate uncertainties in correlation coefficients. The error bars of the α-chymotrypsin half-life were the standard deviations of three experimental repetitions, while the error bars of the computational metrics were calculated as the standard deviations across simulations from the five different wildtype models.
substrate. This metric does appear to be in large part predictive of proteolytic stability of the glycoforms.

Figure 7 shows the percentage of the time each kind of hydrogen bond existed in the MD simulations. As a result, proteolytically stable glycoforms such as GF 13 and GF 10 had the most glycan-involved hydrogen bonds. Glycoforms with comparable proteolytic stability as the wild type, including GF 5, GF 9, and GF 12 had at least one or two glycan-involved hydrogen bonds. Most of the remaining glycoforms, which were more susceptible to proteolytic degradation than the wild type, had no glycan-involved hydrogen bonds. Importantly, glycoforms having at least comparable proteolytic stability as the wild type typically have hydrogen bonds that involved residues PheB24 and ThrB27, which were the P1 site and the P2’ site corresponding to the cleavage site B25, respectively. In glycoforms that had an intermediate level of proteolytic stability (GF 5, and GF 9), hydrogen bonds involving these two residues were not present or at least not long-lasting in the MD trajectories (see Supplementary Table S5), indicating that this hydrogen bond might not be stable enough to consistently disturb the aforementioned hydrogen bonding network.

The fact that 4EYD-, 4EY1-, and 2MVC-based GF 6 all possess the same hydrogen bond is one of many examples showing that large error bars should be irrelevant to the initial modeling bias. Specifically, if there is a bias from the differences in the resolution methods or the histidine protonation states, glycoforms resolved by the same method (e.g. 4EYD-, 4EY9-, and 4EY1-based glycoforms) or having the same histidine protonation states (e.g. 4EYD-, 4EY1- and 3I3Z-based glycoforms) should exhibit similar hydrogen bond distributions at the examined site. However, this trend is absent from our results, suggesting that the metric in the name of each hydrogen bond, the index of the sugar moiety in the glycan is shown in the bracket following the name of the glycan. The atom type is shown in the parenthesis right after the residue name (see Supplemental Table S6 for more details of the atom types shown in the figure). For example, ThrB27(N)–Man[2](O6) means the hydrogen bond formed between the amide N atom of ThrB27 as the donor and one of the oxygen atoms of the second mannose as the acceptor. Texts for hydrogen bonds that involve any of the P1–P3 or the P2’ residues are colored in blue, which in our case only include PheB24 and ThrB27.


diagram here

in our study. Interestingly, GF 13 occasionally had a hydrogen bond formed between ThrB27 and the third mannose, which was absent from GF 10. This additional hydrogen bond that involved a P2’ site might further interfere with the hydrogen bonding network between insulin and α-chymotrypsin, making GF 13 even more proteolytically stable than GF 10.

Overall, the predictions made out of the glycan-involved hydrogen bonds analysis are relatively consistent with the experimental results. The only exception is GF 6, which was experimentally found to be less proteolytically stable than the wild type but had a hydrogen bond, ThrB27(N)–GalNAc[1](O3), that involved a P2’ site ThrB27. However, this hydrogen bond was only present in the MD trajectories of 4EYD-, 4EY1-, and 2MVC-based GF 6 (see Supplemental Table S5), indicating that this hydrogen bond might not be stable enough to consistently disturb the aforementioned hydrogen bonding network.

The fact that 4EYD-, 4EY1-, and 2MVC-based GF 6 all possess the same hydrogen bond is one of many examples showing that large error bars should be irrelevant to the initial modeling bias. Specifically, if there is a bias from the differences in the resolution methods or the histidine protonation states, glycoforms resolved by the same method (e.g. 4EYD-, 4EY9-, and 4EY1-based glycoforms) or having the same histidine protonation states (e.g. 4EYD-, 4EY1- and 3I3Z-based glycoforms) should exhibit similar hydrogen bond distributions at the examined site. However, this trend is absent from our results, suggesting that the metric
is independent of the starting wild-type models. Finally, we can conclude that long-lasting, stable glycan-involved hydrogen bonds, especially the ones that involve PheB24 and ThrB27, appear critical in enhancing the proteolytic stability of the structure. This conclusion also underpins the findings in the work by Guan et al. [19] that ThrB27 and ThrB30 were identified as the best two glycosylation sites, as the glycan attached to these sites are structurally closer to the important residues (ThrB27 and PheB24) found in this study.

**Dimerization propensity**

**Metric 1: glycan-dimer occlusion**

Based on our dimerization hypothesis, glycoforms whose glycans come in proximity to the dimer interface (GlyB23–PheB26) will be less prone to dimerize because of steric occlusion by the glycan. We found that this glycan-dimer occlusion might be a useful predictor of dimerization propensity.

To test this hypothesis, we calculated the glycan-dimer occlusion metric for each glycoform and considered those with low occlusion to have high dimerization propensity, and vice versa. Figure 4 shows representative frames from four 3I3Z trajectories to visually contrast the case where there is an occlusion event to its counterpart. In the figure, the light blue cartoon represents insulin, salmon sticks represent residues GlyB23–PheB26, and pale yellow spheres represent the glycan moiety.

Using the glycan-dimer occlusion metric, we calculated the proportion of simulation frames with occlusion for each glycoform and ordered them according to the frequency that an occlusion event occurred, as shown in Supplemental Table S7. Notably, this analysis method does not include the respective wild-type models for reference (which by default have no occlusion), which is a drawback of this metric.

There is an agreement in the order of glycoforms between model sets for this metric. While there are slight variations in the precise order between models, there are trends that sort the glycoforms into low-occlusion (GF 2, GF 3, GF 7, GF 8), medium-occlusion (GF 4, GF 6, GF 11, GF 12), and high-occlusion (GF 5, GF 9, GF 10, GF 13) batches which are independent of starting structure. This analysis consistently predicts GF 9, GF 10, and GF 13 as having the most occlusion, and therefore least dimerization propensity. This is in agreement with experimental data that these glycoforms form dimers less frequently than wild-type insulin [19]. None of the models accurately reproduce the correct experimental dimerization order, however (GF 10 > GF 13 > GF 9).

The proportion of frames with occlusion, and the associated uncertainty, were compared between glycoforms for each model (Supplemental Fig. S7). The Wilson score 95% confidence intervals for each proportion is shown in red and calculated from the occlusion autocorrelation lag times, which were used to estimate the number of independent occlusion configurations sampled. Differences in occlusion proportions, particularly between the low-occlusion batch and the high-occlusion batch, are statistically distinguishable and this finding is true for all 5 model sets. Differences within occlusion batches are not statistically differentiable.

To assess the dependence of these results on the starting structural model, for each glycoform we averaged the proportion of frames with occlusion across the five models and compared the means and associated standard deviations (Fig. 8). Occlusion proportions are consistent across the models for almost all the glycoforms, particularly those in the low- and high-occlusion batches. There is considerable variability in the glycoforms of the medium-occlusion batch (GF 6, GF 11) but considering the consistency in the other models, this might result more from initial glycoform minimization/optimization than the model itself.

These results suggest that the dimer occlusion metric is independent of the initial model of human insulin. Since we cannot include a wild type in this analysis, there is no negative control for each glycoform set nor is there a way to predict percentage dimerization relative to wild type. However, this metric still allows us to statistically differentiate results between glycoforms. In addition, the results are consistent with experimental data and support our hypothesis. The glycan-dimer occlusion metric, therefore, might be a
useful predictor of dimerization propensity. Interestingly, the occlusion data suggests a unifying theme in glycan placement which might preclude dimerization: glycans that are large and those that are attached close to the dimer interface, especially residues ThrB27 and ThrB30.

Conclusions

The primary goal of our study is to develop metrics for screening the proteolytic stability and dimerization propensity of insulin based on MD simulations of monomers and their glycosylated analogs. The use of monomer simulations eliminates the need of performing advanced sampling of more complicated systems, such as a protease-substrate complex or an insulin dimer. We also evaluated whether these metrics were independent of the initial configuration used.

We examined four metrics based on two overarching hypotheses for proteolytic stability. The first is that the glycan can sterically hinder the scissile bonds or part of the binding sites, especially the P1 site, to prevent proteolysis, while the second is that the existence of the glycan disfavors configurations required for proteolytic degradation. From the first hypothesis, we derived two metrics, the SASA of the scissile bonds and the P1 sites. Both metrics were found partly predictive for assessing the proteolytic stability. Based on the second hypothesis, we examined β-sheet propensity of the P1–P3 region and the existence of glycan-involved hydrogen bonds that compete with the hydrogen-bonding network present in α-chymotrypsin binding. The β-sheet propensity generally was not well correlated with proteolytic stability. In addition, the excessively large error bars makes the metric itself less useful. Notably, the involvement of only a few atoms in the calculation naturally makes these 3 metrics relatively uncertain and less valuable. While the first 3 metrics are either partially predictive or not predictive, long-lasting, stable glycan-involved hydrogen bonds, especially the ones involving residues PheB24 and ThrB27, were highly predictive in enhancing the proteolytic stability and independent initial wild-type model. In particular, the glycan attached to the glycosylation site ThrB27 is more likely to form hydrogen bonds with ThrB27 and PheB24 due to spatial proximity.

To assess dimerization propensity, we examined glycan occlusion of the dimer interface. This metric is consistent with the limited experimental data for dimerization and is free from starting model bias. There is no ability to include a wild-type model for comparison, which is a drawback of this metric. However, this metric showed clear, statistical differences between glycoforms with and without occlusion which translates to predictive differences in dimerization potential and is independent of starting model. Additionally, this metric suggests a generalizable glycosylation scheme that might preclude dimerization: large glycans and those attached near the dimerization interface (ThrB27, ThrB30).

The presence of relatively little experimental data (13 proteolytic stability data points, 3 dimerization data points) means that it is difficult to make firm statistical conclusions about these screening metrics. It is, however, the difficulty in generating experimental data that motivates the development of computational screening techniques. Our framework presented here is widely applicable and allows easy screening of large numbers of insulin glycoforms. To further validate this approach, we will in the future apply this framework in a blind challenge to more complicated insulin glycoforms that are under experimental investigation. The results presented in this paper suggest that we are likely to be able to differentiate the structures with high proteolytic stability and low dimerization propensity from others, as long as the configurational ensembles of the investigated structures are sufficiently sampled. Building more complicated statistical learning models at the present time, such as decision trees or multiple linear predictors is likely not viable, as the relative paucity of experimental data would cause high uncertainty of these weights and decision cuts. However, clearer metrics could possibly be identified in more sophisticated machine-learning approaches that essentially combine 3D coordinates of the systems of interest in a non-linear fashion. For example, recent deep learning frameworks [85] like DiffNets appear capable of identifying subtle structural signatures that could predict biophysical properties in future studies.

Supplementary Information The online version of this article (https://doi.org/10.1007/s10822-022-00453-6) contains supplementary material, which is available to authorized users.

Acknowledgements Research reported in this publication was primarily supported by the National Institute of Biomedical Imaging and Bio-engineering of the National Institutes of Health under award number R01EB025892. This work was also supported in part by the NIH/NC Molecular Biophysics Graduate Traineeship T32 GM065103 and the CAMS Innovation Fund for Medical Sciences (CIFMS 2021-1-12M-026). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions Contributions based on CRediT taxonomy: W.-T.H.: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Methodology, Investigation D.A.R.: Writing - Original Draft, Writing - Review & Editing, Methodology, Investigation T.S.: Conceptualization, Writing - Review & Editing, Funding Acquisition Z.T.: Conceptualization, Writing - Review & Editing, Funding Acquisition M.R.S.: Conceptualization, Writing - Review & Editing, Supervision, Funding Acquisition

Data availability As discussed in Section Analysis techniques, most values reported in our results are averaged over different wild-type crystal structure models. For individual analysis result of each insulin glycoform, please refer to our GitHub repository of the project. The repository also contains input configurations/MD parameters and
Python codes for data analysis. The outputs of the MD trajectories are too large to release as they are several terabytes in size and statistically representative outputs can be generated from the input files.

**Declarations**

**Disclosures** M.R.S. is an Open Science Fellow at and consults for Rovi-Vant Sciences.

**References**

1. Carino GP, Mathiowitz E (1999) Oral insulin delivery. Adv Drug Deliv Rev 35(2–3):249–257
2. Fonte P, Araújo F, Reis S, Sarmento B (2013) Oral insulin delivery: how far are we? J Diabetes Sci Technol 7(2):520–531
3. Gedawy A, Martinez J, Al-Salami H, Dass CR (2018) Oral insulin delivery: existing barriers and current counter-strategies. J Pharm Pharmacol 70(2):197–213
4. Hoffman A, Ziv E (1997) Pharmacokinetic considerations of new insulin formulations and routes of administration. Clin Pharmacokinet 33(4):285–301
5. Owens DR (2002) New horizons–alternative routes for insulin therapy. Nat Rev Drug Discovery 1(7):529–540
6. Bruno BJ, Miller GD, Lim CS (2013) Basics and recent advances in peptide and protein drug delivery. Ther Deliv 4(11):1443–1467
7. Hinds KD, Kim SW (2002) Effects of PEG conjugation on insulin properties. Adv Drug Deliv Rev 54(4):505–530
8. Clement S, Still JG, Kosutic G, McAllister R (2002) Oral insulin product hexyl-insulin monoconjugate 2 (HM2) in type 1 diabetes mellitus: the glucose stabilization effects of HM2. Diabetes Technol Ther 4(4):459–466
9. Deng W, Xie Q, Wang H, Ma Z, Wu B, Zhang X (2017) Selenium nanoparticles as versatile carriers for oral delivery of insulin: insight into the synergistic anti-diabetic effect and mechanism. Nanomed Nanotechnol Biomed 13(6):1965–1974
10. Bhattacharyya A, Mukherjee D, Mishra R, Kundu P (2017) Preparation of polyurethane-alginate/chitosan core shell nanoparticles for the purpose of oral insulin delivery. Eur Polym J 92:294–313
11. Zhou Y, Liu L, Cao Y, Yu S, He C, Chen X (2020) A Nano-composite vehicle based on metal-organic framework nanoparticle incorporated biodegradable microspheres for enhanced oral insulin delivery. ACS Appl Mater Interfaces 12(20):22581–22592
12. Agarwal V, Reddy IK, Khan MA (2000) Oral delivery of proteins: Effect of chicken and duck ovomucoid on the stability of insulin in the presence of α-chymotrypsin and trypsin. Pharm Pharmacol Commun 6(5):223–227
13. Kayser V, Chenamsetty N, Vojnov V, Forrer K, Helb K, Trout BL (2011) Glycosylation influences on the aggregation propensity of therapeutic monoclonal antibodies. Biotechnol J 6(1):38–44
14. Raju TS, Scallon BJ (2006) Glycosylation in the Fc domain of IgG increases resistance to proteolytic cleavage by papain. Biochem Biophys Res Commun 341(3):797–803
15. Russell D, Oldham NJ, Davis BG (2009) Site-selective chemical protein glycosylation protects from autoylation and proteolytic degradation. Carbohydr Res 344(12):1508–1514
16. Losey Y, Paul A, Frenkel-Pinter M, Abu-Hussein M, Khalaila I, Gazit E et al (2019) Novel model of secreted human tau protein reveals the impact of the abnormal N-glycosylation of tau on its aggregation propensity. Sci Rep 9(1):1–10
17. van Veen HA, Geerts ME, van Berkel PH, Nuijens JH (2004) The role of N-linked glycosylation in the protection of human and bovine lactoferrin against tryptic proteolysis. Eur J Biochem 271(4):678–684
18. Sareneva T, Pirhonen J, Cantell K, Julkunen I (1995) N-glycosylation of human interferon-γ: glycans at Asn-25 are critical for protease resistance. Biochem J 308(1):9–14
19. Guan X, Chaafey PK, Wei X, Gulbranson DR, Ruan Y, Wang X et al (2018) Chemically precise glycoengineering improves human insulin. ACS Chem Biol 13(1):73–81
20. Mark AE, Berendsen HJ, Van Gunsteren WF (1991) Conformational flexibility of aqueous monomeric and dimeric insulin: a molecular dynamics study. Biochemistry 30(45):10866–10872
21. Zoete V, Meuwly M, Karplus M (2004) A comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. J Mol Biol 342(3):913–929
22. Yang C, Lu D, Liu Z (2011) How PE Glylation enhances the stability and potency of insulin: a molecular dynamics simulation. Biochemistry 50(13):2585–2593
23. Pincus M (1970) Letter to the editor—a Monte Carlo method for the approximate solution of certain types of constrained optimization problems. Oper Res 18(6):1225–1228
24. Isravelowitz B, Gao M, Schulten K (2001) Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 11(2):224–230
25. Sugita Y, Kitao A, Okamoto Y (2000) Multidimensional replica-exchange method for free-energy calculations. J Chem Phys 113(15):6042–6051
26. Antoszewska A, Feng CJ, Vani BP, Thiede EH, Hong L, Weare J et al (2020) Insulin dissociates by diverse mechanisms of coupled unfolding and binding. J Phys Chem B 124(27):5571–5587
27. Busto-Moner L, Feng CJ, Antoszewska A, Tokmakoff A, Dinner AR (2021) Structural Ensemble of the Insulin Monomer. Biochemistry 60(42):3125–3136
28. Hansmann UH (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281(1–3):140–150
29. Earl DJ, Deem MW (2005) Parallel tempering: Theory, applications, and new perspectives. Phys Chem Chem Phys 7(23):3910–3916
30. Nuske F, Keller BG, Pérez-Hernández G, Mey AS, Noé F (2014) Variational approach to molecular kinetics. J Chem Theory Comput 10(4):1739–1752
31. Lorpaiboon C, Thiede EH, Webber RJ, Weare J, Dinner AR (2020) Integrated variational approach to conformational dynamics: A robust strategy for identifying eigenfunctions of dynamical operators. J Phys Chem B 124(42):9354–9364
32. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M et al (2011) Markov models of molecular kinetics: Generation and validation. J Chem Phys 134(7):174105
33. Bowman GR, Pande VS, Noé F (2013) An introduction to Markov state models and their application to long timescale molecular simulation, vol 797. Springer, New York
34. Schütte C, Fischer A, Huisenga W, Deuflhard P (1993) A direct variational approach to conformational dynamics and mechanical functions of proteins. Curr Opin Structural Biology 7(23):3910–3916
35. Antoszewska A, Feng CJ, Vani BP, Thiede EH, Hong L, Weare J et al (2020) Insulin dissociates by diverse mechanisms of coupled unfolding and binding. J Phys Chem B 124(27):5571–5587
36. Busto-Moner L, Feng CJ, Antoszewska A, Tokmakoff A, Dinner AR (2021) Structural Ensemble of the Insulin Monomer. Biochemistry 60(42):3125–3136
37. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M et al (2011) Markov models of molecular kinetics: Generation and validation. J Chem Phys 134(7):174105
simulating systems with rough energy landscape. Chem Phys Lett 329(3-4):261–270
39. Lyubartsev A, Martinsovski A, Shevkonov S, Vorontsov-Velyaminov P (1992) New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. J Chem Phys 96(3):1776–1783
40. Fávero-Retto MP, Palmieri LC, Souza TA, Almeida FC, Lima LMT (2013) Structural meta-analysis of regular human insulin in pharmaceutical formulations. Eur J Pharm Biopharm 85(3):1112–1121
41. Timofeev V, Chuprov-Netochin R, Samigina V, Bezuglov V, Miroshnikov K, Kuranova I (2010) X-ray investigation of gene-engineered human insulin crystallized from a solution containing polyalcoholic acid. Acta Crystallogr Sect F 66(3):259–263
42. Krňíková K, Veverka V, Maletínská L, Hektorová R, Brzozowski AM, Jiráček J et al (2014) Structural and functional study of the GlnB22–insulin mutant responsible for maturity-onset diabetes of the young. PLoS ONE 9(11):e112883
43. DeLano WL et al (2002) Pymol: an open-source molecular graphics tool. CCP4 Newsletter Protein Crystallogr 40(1):82–92
44. Anandakrishnan R, Aguilar B, Onufriev AV (2012) H++3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Res 40(W1):W537–W541
45. Myers J, Grothaus G, Narayanan S, Onufriev A (2006) A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. Proteins 63(4):928–938
46. Gordon JC, Myers JB, Folta T, Shova J, Heath LS, Onufriev A (2005) H++: a server for estimating pKas and adding missing hydrogens to macromolecules. Nucleic Acids Res 33(suppl-2):W368–W371
47. McQueen C (2017) Comprehensive toxicology. Elsevier, Amsterdam
48. Fallingborg J (1999) Intraluminal pH of the human gastrointestinal tract. Med Bull 46(3):183–196
49. Kirschner KN, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL et al (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates 29(4):622–655
50. Abraham MJ, Murtoila T, Schulz R, Pall S, Smith JC, Hess B et al (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1:19–25
51. Pall S, Abraham MJ, Kutzner C, Hess B, Lindahl E (2014) Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In: International Conference on Exascale Applications and Software. Springer, pp 3–27
52. Da Silva AWS, Vranken WF (2012) ACPYPE-Antechamber python parser interface. BMC Res Notes 5(1):1–8
53. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126(1):014101
54. Křížková K, Veverka V, Maletínská L, Hexnerová R, Brzozowski AM, Jiráček J et al (2014) Structural and functional study of the GlnB22–insulin mutant responsible for maturity-onset diabetes of the young. PLoS ONE 9(11):e112883
55. Parrinello M, Rahman A (1981) Polymorphic transitions in silicon crystals: a new molecular dynamics method. J Appl Phys 52(12):7182–7190
56. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103(19):8577–8593
57. Hess B, Bekker H, Berendsen HJ, Fraajje JG (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18(12):1463–1472
58. Wilcox P (1970) Chymotrypsinogen—chymotrypsins. In: Methods in enzymology, vol 19. Elsevier, pp 64–108
59. Winkel W, Huber R (1993) Natural protein kinase inhibitors and their interaction with proteinases. EJB Rev, pp 43–61
60. Eigenthaler, H, Lijina Z, Argos P, Sander C, Scharf M (1995) The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J Comput Chem 16(3):273–284
61. Chodera JD (2016) A simple method for automated equilibration detection in molecular simulations. J Chem Theory Comput 12(4):1799–1805
62. Perrin MG, et al (1948) The advanced theory of statistics. Vols. 1. The advanced theory of statistics, vol 1. (Ed. 4)
63. Schechter I, Berger A (1968) On the active site of proteinases. III. Mapping the active site of papain; specific peptide inhibitors of papain. Biochem Biophys Res Commun 32(5):898–902
64. Appel W (1986) Chymotrypsin: molecular and catalytic properties. Clin Biochem 19(6):317–322
65. Hedstrom L (2002) Serine protease mechanism and specificity. Chem Rev 102(12):4501–4524
66. Bode W, Huber R (1993) Natural protein kinase inhibitors and their interaction with proteinases. EJB Rev, pp 43–61
67. Coombs GS, Rao MS, Olson AJ, Dawse PE, Madson EL (1999) Revisiting catalysis by chymotrypsin family serine proteinases using peptide substrates and inhibitors with unnatural main chains. J Biol Chem 274(34):24074–24079
68. Hower RS, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, et al (2019) MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. Los Alamos National Lab.(LANL), Los Alamos, NM
69. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem 32(10):2319–2327
70. Ramachandran GNVS, Ramakrishnan C (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99
71. Lovell SC, Davis IW, Arendall WB III, De Bakker PIW, Word JM, Prisant MG et al (2003) Structure validation by Ca geometry: φ, ψ and Cβ deviation. Proteins 50(3):437–450
72. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX et al (2015) MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophys J 109(8):1528–1532
73. Baker EN, Hubbard RE (1984) Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 44(2):97–179
74. Harding MM, Hodgkin DC, Kennedy AF, O’Connor A, Weitzman PDM (1966) The crystal structure of insulin: II. An investigation of the hydrogen bonding in the insulin dimer interface*. J Biol Chem 241(10):684–699
75. Ramachandran GNVS, Ramakrishnan C (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99
76. Appel W (1986) Chymotrypsin: molecular and catalytic properties. Clin Biochem 19(6):317–322
77. Keller D, Clausen R, Josefsen K, Led JJ (2001) Flexibility and bioactivity of insulin: an NMR investigation of the solution structure and folding of an unusually flexible human insulin mutant with increased biological activity. Biochemistry 40(35):10732–10740
78. Jörgensen AMM, Olsen HB, Balschmidt P, Led JJ (1996) Solution structure of the superactive monomeric des-[Phe (B25)] human insulin: elucidation of the structural basis for the monomerization of des-[Phe (B25)] insulin and the dimerization of native insulin. J Mol Biol 257(3):684–699
79. Bocian W, Sitkowski J, Bednarek E, Tarnowska A, Kawecki R, Kozerski L (2008) Structure of human insulin monomer in water/acetonitrile solution. J Biomed NMR 40(1):55–64
80. Bokan A, Mieles LM, Bahar I (2011) ProDy: Protein dynamics inferred from theory and experiments. Bioinformatics 27(11):1575–1577
81. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D et al (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods 17:261–272
81. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D et al (2020) Array programming with NumPy. Nature 585:357–362
82. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22(158):209–212
83. Newcombe RG (1998) Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med 17(8):857–872
84. Wallis S (2013) Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. J. Quant. Linguistics 20(3):178–208
85. Ward MD, Zimmerman MI, Meller A, Chung M, Swamidass S, Bowman GR (2021) Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. Nat Commun 12(1):1–12