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

Wei-Tse Hsu[1]**, Dominique A. Ramirez**[2]**, Tarek Sammakia**[3]**, Zhongping Tan**[4]**, Michael R. Shirts**[1]

[1]Department of Chemical & Biological Engineering, University of Colorado Boulder, Boulder, CO, USA 80309; [2]Department of Biochemistry, University of Colorado Boulder, Boulder, CO, USA 80309; [3]Department of Chemistry, University of Colorado Boulder, Boulder, CO, USA 80309; [4]Institute of Materia Medica, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, 100050, China

*For correspondence:
michael.shirts@colorado.edu (MRS); zhongping.tan@imm.pumc.edu.cn (ZT)

Abstract   Insulin has been commonly adopted as a peptide drug to treat diabetes given its ability to facilitate 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 the understanding of such phenomena at the molecular level is still evasive. To address this challenge, we propose and test several structural determinants that could potentially influence insulin's proteolytic stability and dimerization propensity. We used these as the 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, as was in part the solvent accessible surface area of proteolytic sites, while 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.

1 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. While it is usually administered via subcutaneous injections, excessive injections could lead to non-compliance by the 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 not only avoids the aforementioned side effects of subcutaneous administration, but also 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 upon dimerization, which leads to overall low absorption efficiency [6]. Therefore, proteolytic stability and dimerization propensity are two of the most important controlling factors when developing oral insulin drugs. During past years, a wide array of strategies, including chemical modifications [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 one of these chemical modifications strategies and has shown significant promise in enhancing the proteolytic stability and monomeric propensity of various kinds of proteins [13–15]. While N-linked glycosylation has been more 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 glycosylation patterns, including glycosylation sites, glycan sizes, and glycan structures [19]. Their study suggested the superiority of O-mannosylation of insulin B-chain Thr27 to other studied glycosylation patterns in increasing the proteolytic stability against α-chymotrypsin and decreasing the dimerization propensity of insulin, while still maintaining the 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 PEGglyated insulin pharmaceuticals. More recently, with steered molecular dynamics [24] and replica-exchange umbrella sampling [25], Antoszewski et al. [26] determined the relative energies of unfolding and identified different unfolding pathways of insulin.

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. Specifically, we performed molecular dynamics simulations for each of the 12 insulin glycoforms (GFs) (Figure 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. Based on the proteolytic degradation mechanism by α-chymotrypsin [27] and the structural characteristics of insulin, we proposed several potential metrics for assessing the two insulin properties, including solvent-accessible surface area (SASA), secondary structures, existence of glycan-involved hydrogen bonds, and the occlusion between the glycan and the dimer interface. A previous study [28] found that greater conformational rigidity of a carbohydrate binding module (CBM) protein was associated with larger proteolytic stability of glycoforms, thus disfavoring reaction-capable transition state. However, overall conformational rigidity of insulin is significantly different than CBM, and thus more specific measures of transition state accessibility were used, such as the β-sheet propensity, which more specifically takes the structural features of the known α-chymotrypsin transition state into account.
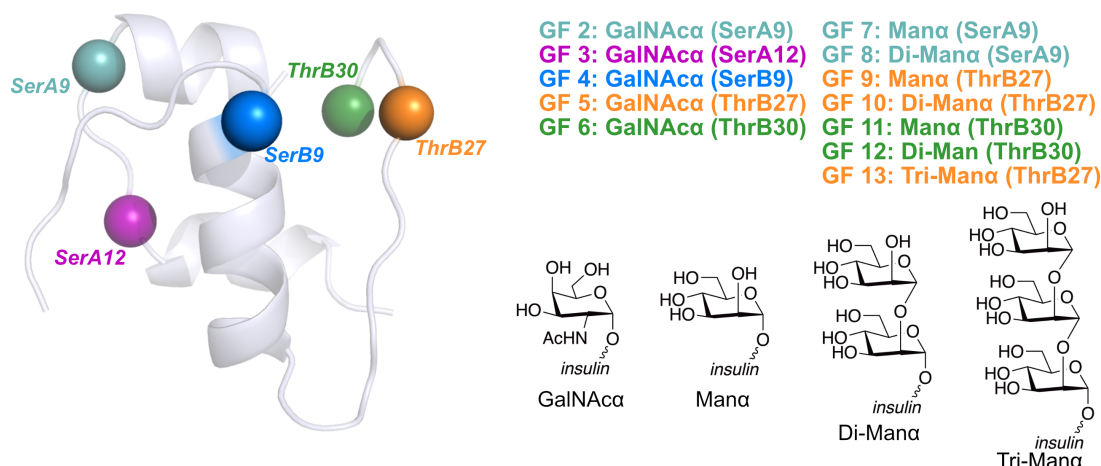
Figure 1. Structures of a human insulin monomer and glycoforms studied in this work. Two kinds of sugar moieties (N-acetylgalactosamine (GalNAc) and mannose (Man) in the α-anomer) with varying lengths (e.g. a mannose monomer, dimer, or trimer) were attached to five different glycosylation sites of insulin, 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 for glycoforms having a mannose dimer or trimer as the glycan, the linkages between monomers were all α-1,2-linkages, so Tri-Manα (ThrB27) refers to the glycoform containing an α-1,2-linked tri-mannose at the B chain Thr27 residue.

Notably, both the proteolytic stability and dimerization propensity are related to transformation processes that require overcoming free energy barriers. The proteolytic stability can be directly associated with the free energy barrier of the formation of the conformational ensemble corresponding to the transition state in the proteolytic degradation by α-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 are far from trivial. The reason lies in the fact that the timescale of the binding/unbinding events between an insulin monomer and an α-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 [29] or alchemical transformations [30, 31]. 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-α-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 α-chymotrypsin or an insulin dimer should encode at least some important structural insights into both the proteolytic stability and the dimerization propensity. As long as we are able to explore the configurational space of monomer-based structures sufficiently, the conformational ensembles we get from the simulations should shed light on developing reasonable metrics for the two insulin properties.

Given the output MD trajectories, all metrics were measured for each glycoform structure. They were later compared with the experimental work for assessing the efficacy. Notably, we found that at minimum 5×2000 ns = 10 μs of simulation was necessary to capture the structural characteristics of the insulin ensembles. This aggregate length is, to our knowledge, 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 assesses proteolytic stability of a protein by molecular dynamics. 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.

2 Methodology

2.1 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, in our study, we built each of the 12 glycoforms on 5 wild-type (WT) structures resolved by different groups using different methods, whose PDB IDs were 4EYD, 4EY9, 4EY1, 3I3Z, and 2MVC, respectively (Figure 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 [32]. They represented high-resolution human insulin structures from pharmaceutical formulations. 3I3Z and 2MVC, on the other hand, were 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 [33]. 2MVC was resolved by NMR spectroscopy under acidic conditions, which were known to favor the monomeric form of insulin [34].

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 [35] 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 experimental conditions that favor dimer resolution, as later shown, these biases did not have noticeable influences on the consistency of our methods.
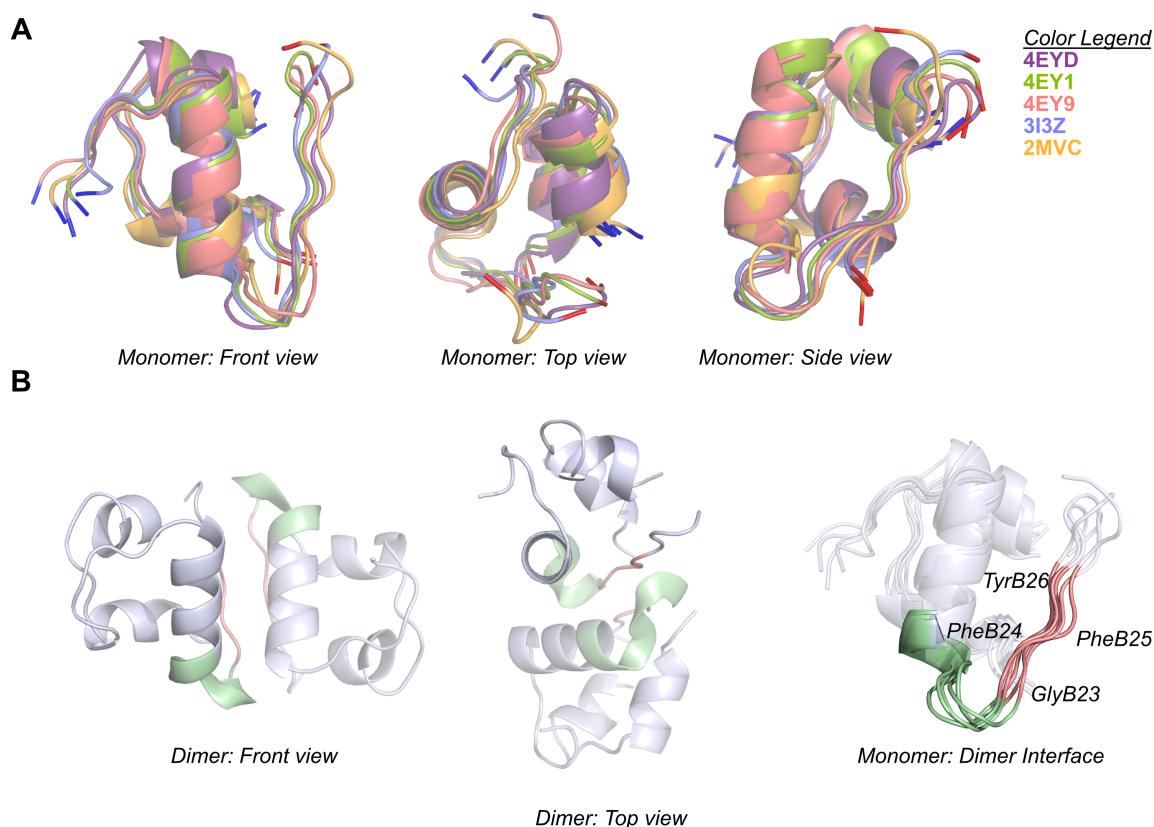
Figure 2. Structures of wild-type insulin models used in this study. (A) The initial monomer structures, after equilibration and before production simulation, are superimposed and are shown from different views. (B) Representative dimer structure illustrating the dimerization interface. The 3I3Z crystal structure was used to reconstruct an insulin dimer in the first two images. Residues GlyB23–TyrB26 (salmon) are highlighted. The last image shows the superimposed equilibrated wild-type structures with labels for the dimer interface.

After extracting the monomer structure from each of the initial models, we used the H++ server (version 3.2) [36–38] to assign reasonable protonation states under the pH value (8.0) adopted in experiments. At pH 8.0, the total charges of insulin were predicted to be -1 or -2, depending on the $pK_{1/2}$ value of the two histidine residues (HisB5 and HisB10) of insulin. We therefore adjusted the external pH value in H++ of each wild-type structure to make sure all of them had total charges of -2. 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 protonation states for the histidine residues different from the ones in the other three models (see Supplemental Table S1). 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 were far away from the residues involved in the hypotheses of our analysis methods (see Supplemental Figure S1), 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 [39], which utilized 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 [40, 41] 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 [42] 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 the Berendsen barostat [43] and velocity rescaling [44] 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 [45, 46]. The cut-off distances for van der Waals interactions and Coulomb interactions were both set as 0.9 nm, with swiching 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 [47] was used with real-space switching of the potential between 0.89 nm and 0.9 nm. LINCS [48] was used to constrain hydrogen bonds. All the simulations were extended up to 2000 ns, which we deemed necessary to capture insulin dynamics in which the major transitions between metastable states occurred on the time scale of 500–2000ns, as interpreted from time series of the pairwise RMSD calculations of the wild-type structures (see Supplemental Figure S2). 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 repositoryhttps://github.com/shirtsgroup/Glycoinsulin_project for this study.

2.2 Analysis techniques

2.2.1 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 [49]. 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 [27]. Previous research into α-chymotrypsin [27] indicated that residues PheB25 and TyrB26 were the cleavage sites considered to be the most susceptible to proteolytic stability. We therefore hypothesized that the exposure of the scissile bonds of these two residues to the solvent was positively correlated with proteolytic susceptibility.

To quantify the solvent exposure of these sites, we used the double cubic lattice method (DCLM) [50] 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. The obtained SASA time series was truncated to discard the non-equilibrium region and was then decorrelated [51]. The mean and the standard deviation of the time series were calculated for each glycoform. Finally, the reported value for each glycoform is the value averaged across the five different wild-type bases, with the uncertainty being the propagated error. 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) [52] and its corresponding two-tailed p-value determined. Kendall's tau was used as it was not clear that the relationship should be linear, so we are assuming only a monotonic relationship.

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, namely, the P1 residues according to Schechter-Berger nomenclature [53] were also important. Specifically, the hydrophobicity of the P1 residue was found to be important in the molecular recognition of α-chymotrypsin [54], as the deep hydrophobic pocket formed by Ser189, Gly216, and Gly226 in α-chymotrypsin requires the P1 residue to be hydrophobic as well [55]. 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. DCLM was used to generate the SASA time series, which was then truncated and decorrelated with the same method as the one used 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 propagated error. The monotonicity of the relationship between the metric and the experimental data was assessed by the Kendall's tau correlation coefficient (τ) [52], with the corresponding two-tailed p-value calculated.

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

Previous experimental studies [55–57] 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, considering the P1 to P3 sites of both important cleavage sites (PheB25 and TyrB26) leads to residues including ArgB22, GlyB23, PheB24, and PheB25. As such, 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 stemmed from the fact that drastic structural re-orientations or even secondary structure transformations of the insulin substrate cost free energy and disfavored 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 [58, 59] to locate all the frames in a Ramachandran plot [60] and defined the β-sheet propensity as the fraction of the points in the β-sheet region defined by Lovell et al. [61] Again, for each glycoform, we report the fractions averaged across different wild-type bases, with the uncertainty the standard deviation over runs with the five different initial structures. Finally, the α-chymotrypsin half-life was plotted against these averaged β-sheet fractions in a correlation plot, where the Kendall's tau correlation coefficient (τ) and its corresponding two-tailed p-value were determined. The Ramachandran plot of each of the 4 residues of interest of each glycoform can be found in our GitHub repository.

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 hydrolysis [55]. 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 also 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 in mind, 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 [62] to identify hydrogen bonds between the glycan moiety (as the acceptor) and any insulin residue (as the donor) according to the Baker-Hubbard criterion [63], 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.

2.2.2 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 (Figure 2B) [33, 64, 65].

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. However, we found no consistent secondary structure propensity differences in either the dimer or the monomer between glycoforms or between glycoforms and the wild type. This was also true after examining additional monomer structures 1JCO, 1MHJ, and 2JV1. This suggests to us that such structural analysis alone of glycoforms will not allow us to determine dimerization propensity, and so we explored different metrics to assess dimerization.

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 (Figure 1), and there are several glycans large enough to sterically interfere with the dimerization interface (Figure 3). We hypothesized that glycans that occupy space close to these residues, with high frequency in simulation time, will preclude dimerization.

We used the same converted trajectories described above to calculate this metric. 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 [66], to calculate the total number of atom neighbor pairs between the glycan and dimer interface for each frame in the converted trajectories.

Atom neighbor pair autocorrelation lag times (in ns) were estimated by fitting autocorrelation data to an exponential decay function using SciPy [67, 68] and are presented in Supplemental Table S2. There are several glycoform models that have no lag time ("NA" in Supplemental Table S2) 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 contain 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 [69–71]. 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.
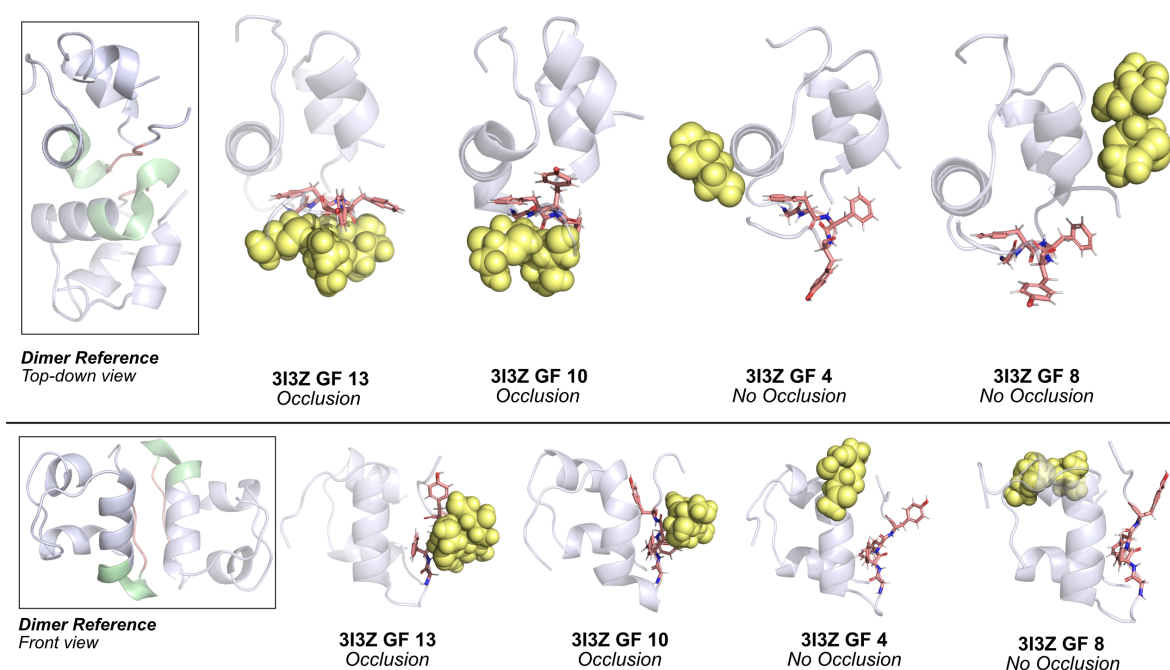
Figure 3. Classification of occlusion and no-occlusion states. Representative frames from four different glycoform trajectories show occlusion (GF 13, GF 10) and no occlusion (GF 4, GF 8) states. Insulin monomer presented in translucent blue-white cartoon, dimerization residues presented in salmon sticks, and glycan moiety presented in yellow spheres.

2.3 Molecular Visualization

Molecular visualization, specifically for Figure 1, 2, 3, and S1, were done using PyMOL version 2.4.1 [35].

3 Results and Discussions

3.1 Proteolytic degradation

Metric 1: SASA of scissile bonds

According to our hypothesis of 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 α-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 Figure 4A) and the one between residues B26 and B27 (lower panel of Figure 4A). 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 concluded that although the SASA of the scissile bonds was free from inital structure bias, it was only 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 α-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. For glycoforms having lower permeability than the wild type, such as GF 3, GF 4, GF 7, and GF 8, the SASA values were indeed higher than the wild type. Still, some of its predictions provided opposite results as compared to the experimental ones (GF 6 and GF 11). Importantly, all observations are independent of different initial wild-type structures, which is reflected by the small error bars of the SASA values.
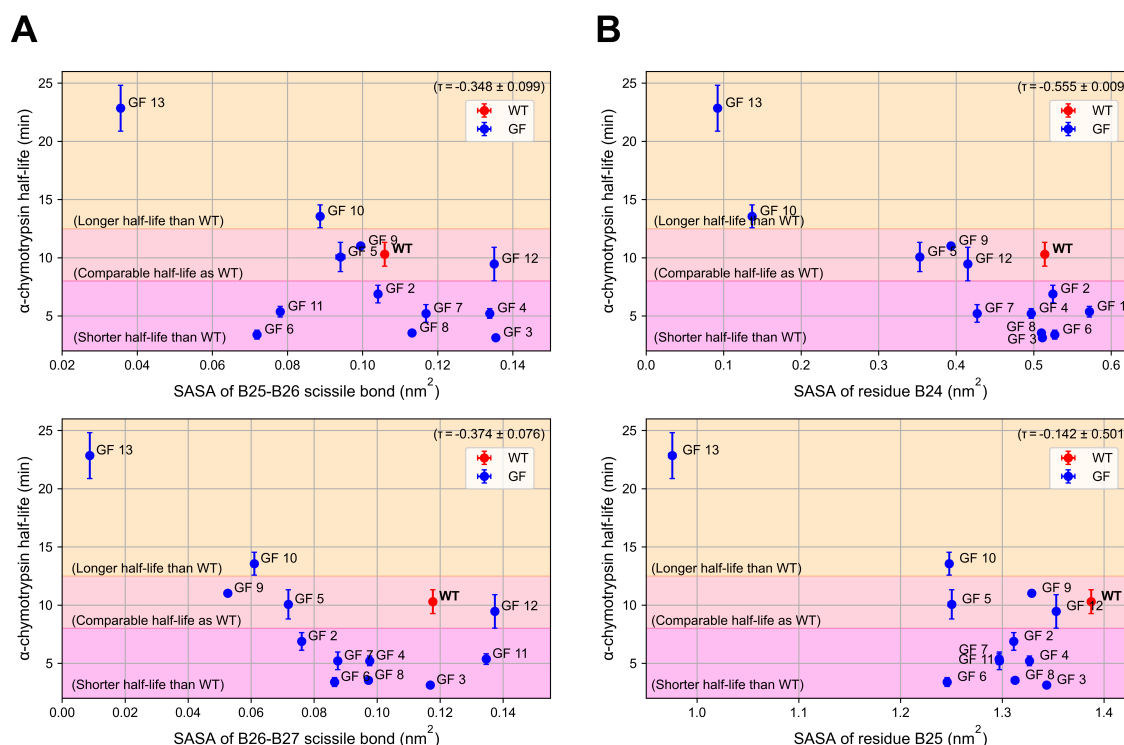

Figure 4. The SASA of both the scissile bonds and the P1 sites are weakly correlated with α-chymotrypsin half-life, implying moderate predicitveness for the proteolytic stability. (A) The correlation plot between the α-chymotrypsin half-life and the average SASA of the scissile bonds in the glycoforms (GF). (B) The correlation plot between the α-chymotrypsin half-life and the average SASA of the P1 sites in the glycoform (GF). Different regions were colored to indicate glycoforms having longer, comparable or shorter half-life as compared with the wild type. The Kendall's tau correlation coefficient (τ) and its two-tailed p-value were calculated. Error bars of both variables are shown, but the error bars of the SASA metrics are generally very small and invisible.

Metric 2: SASA of the P1 sites

Similarly to Metric 1, Metric 2 is independent of the initial wild-type structures, as evidenced by the small error bars of the SASA values. However, we again concluded that it was partially predictive.

In the lower panel of Figure 4B, the SASA of residue B25 is less predictive than the SASA of the other site. Although 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 α-chymotrypsin half-life, all had a lower value compared to the wild type. The Kandall's tau correlation coefficient, which was close to 0, showing that the SASA of residue B25 was almost uncorrelated with the α-chymotrypsin half-life.

The SASA of residue B24 (the upper panel of Figure 4B), 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, from Figure 5, it can be seen that none of the examined residues showed this trend and all the Kendall's tau correlation coefficients were very low and highly uncertain, indicating that Metric 3 was inadequate in capturing structural determinants that influenced the proteolytic stability of the structure. This is also reflected by the large error bars in the β-sheet propensity of most of the residues of interest, including residues B22, B23, and B24. The large error bars in β-sheet propensity imply that the structure at these sites (residues B22 to B24) 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. 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. We note the only site that showed small error bars in the β-sheet propensity was residue B25, which might be related to the unique role of residue B25 as a P1 site and a cleavage site at the same time. However, B25 essentially has high β-sheet propensity in all glycoforms, so variability at this site cannot be used.
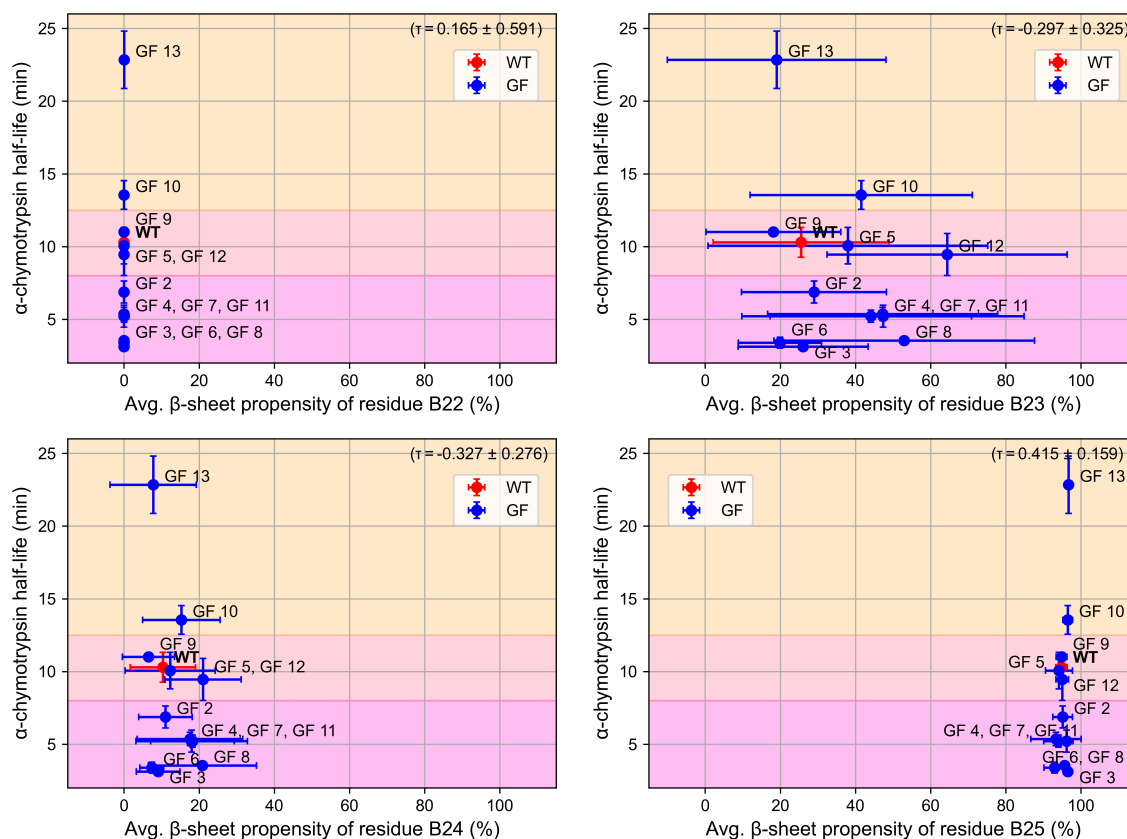
Figure 5. The correlation plot between the α-chymotrypsin half-life and the average β-sheet propensity of residues B22 to B25. The β-sheet propensity of the P1-P3 region does not correlate with α-chymotrypsin half-life. Different regions were colored to indicate glycoforms having longer, comparable or shorter half-life as compared with the wild type. The Kendall's tau correlation coefficient (τ) and its two-tailed p-value were calculated. Error bars of both variables are shown.

Metric 4: Existence of glycan-involved hydrogen bonds

We hypothesized that glycoforms with more glycan-involved hydrogen bonds, especially the ones that involve 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 substrate. This metric does turn out to be in large part predictive of proteolytic stability of the glycoforms.

Figure 6 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, GF9, 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 or 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 Supplemental Table S3), as reflected by the large errors bars that almost reached the bottom of the graph. On the other hand, in the proteolytically stable glycoforms, GF 13 and GF 10, hydrogen bonds ThrB27(N)–Man[2](O6) and PheB24(N)–Man[1](O3) were found in all GF 13/GF 10 MD trajectories, regardless of which wild-type model the glycoform was built on. We concluded these two kinds of hydrogen

bonds (ThrB27(N)–Man[2](O6) and PheB24(N)–Man[1](O3)), the longest-lasting glycan-involved hydrogen bonds in the most proteolytically stable structures, were the most important hydrogen bonds influencing the proteolytic stability 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 S3, indicating that this hydrogen bond might not be stable enough to consistently disturb the aforementioned hydrogen bonding network. Importantly, the fact that 4EYD-, 4EY1-, and 2MVC-based GF 6 all possess the same hydrogen bond is one of many examples showing the 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 upderpins 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.
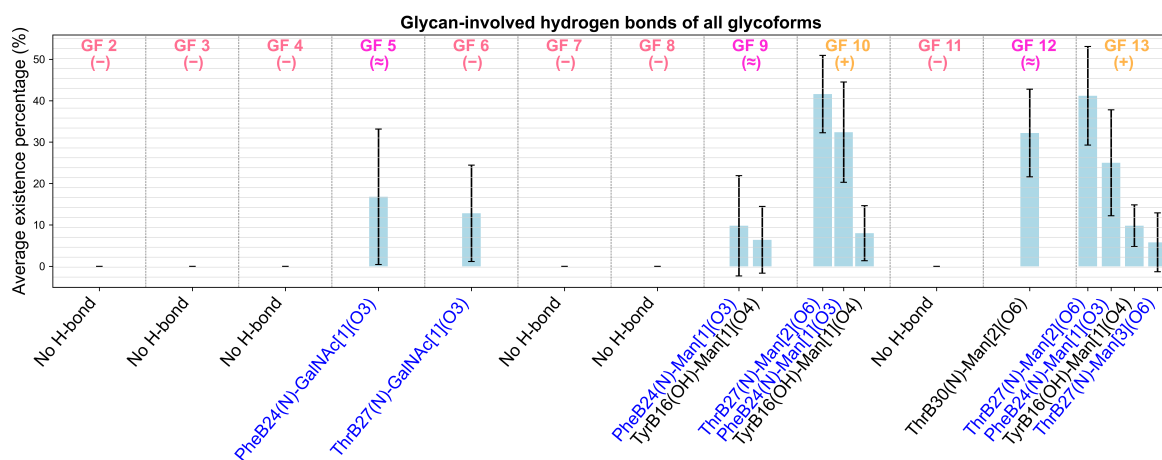

Figure 6. The most proteolytically stable glycoforms tend to have more glycan-involved hydrogen bonds, especially the ones involving the residues PheB24 and ThrB27. The figure shows the average percentage of the time of each hydrogen bond existed in the MD simulations. The experimental results are summarized such that +(yellow), −(red), and ≈(magenta) respectively indicate that the corresponding glycoform was experimentally found to have higher, lower and comparable proteolytic stability as compared to the wild type. Note that in the name of each hydrogen bond, the 1-based index of 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 S4 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. Text 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.

3.2 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 3 shows representative frames from four 3I3Z trajectories to visually demonstrate occlusion versus no-occlusion, where 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 from least occlusion to most occlusion, shown in Supplemental Table S5. We must carefully compare these results to experimental dimerization data [19] because this analysis method cannot include the respective wild-type models for reference (which by default have no occlusion), which is a drawback of this metric.

There is 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* (GF 2, GF 3, GF 7, GF 8), *medium* (GF 4, GF 6, GF 11, GF 12), and *high* (GF 5, GF 9, GF 10, GF 13) occlusion 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 dimer propensity, of the glycoforms. This is in agreement with experimental data that these glycoforms form dimers less frequently than wild-type insulin [19]. None of the models accurately reproduced 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 Figure S3). 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.
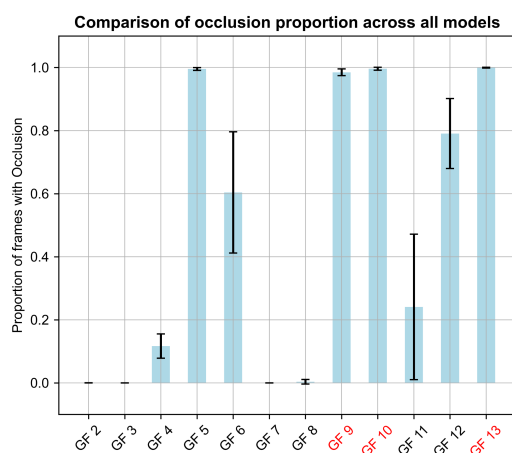

Figure 7. Average occlusion analysis distinguishes glycoforms with low, medium, and high occlusion proportion. All five models for each glycoform were averaged for occlusion proportion, and are plotted with standard deviation represented in black bars. Axis labels in red indicate glycoforms with experimental dimerization data.

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 associ-

ated standard deviations (Figure 7). 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. But despite those drawbacks, this metric allows us to statistically differentiate results between glycoforms and the results are consistent with experimental data, and supports 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.

4 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 removes 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 initial configuration used.

We examined four metrics based on two overarching hypotheses for proteolytic stability. The first is that the glycan could 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 independent of the initial wild-type models, and were 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 was not correlated with proteolytic stability. However, 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, using 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 was 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 are applying our 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. Finally, it is possible that clearer metrics could be identified in more sophisticated machine-learning based approaches, such as recent deep learning frameworks [72] that appear capable of identifying subtle structural signatures that predict biophysical properties, and such techniques will be tested in the future.

5 Data Availability

As discussed in Section 2.2, most values reported in our results were values averaged over different bases of wild-type 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.

6 Acknowledgements

7 Author and 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

8 Disclosures

MRS is an Open Science Fellow at and consults for Roivant Sciences.

References

[1] **Carino GP**, Mathiowitz E. Oral insulin delivery. Advanced drug delivery reviews. 1999; 35(2-3):249–257. https://10.1016/S0169-409X(98)00075-1.

[2] **Fonte P**, Araújo F, Reis S, Sarmento B. Oral insulin delivery: how far are we? Journal of diabetes science and technology. 2013; 7(2):520–531. https://doi.org/10.1177/193229681300700228.

[3] **Gedawy A**, Martinez J, Al-Salami H, Dass CR. Oral insulin delivery: existing barriers and current counter-strategies. Journal of pharmacy and pharmacology. 2018; 70(2):197–213. https://doi.org/10.1111/jphp.12852.

[4] **Hoffman A**, Ziv E. Pharmacokinetic considerations of new insulin formulations and routes of administration. Clinical pharmacokinetics. 1997; 33(4):285–301. https://doi.org/10.2165/00003088-199733040-00004.

[5] **Owens DR**. New horizons—alternative routes for insulin therapy. Nature Reviews Drug Discovery. 2002; 1(7):529–540. https://doi.org/10.1038/nrd836.

[6] **Bruno BJ**, Miller GD, Lim CS. Basics and recent advances in peptide and protein drug delivery. Therapeutic delivery. 2013; 4(11):1443–1467. https://0.4155/tde.13.104.

[7] **Hinds KD**, Kim SW. Effects of PEG conjugation on insulin properties. Advanced drug delivery reviews. 2002; 54(4):505–530. https://doi.org/10.1016/S0169-409X(02)00025-X.

[8] **Clement S**, Still JG, Kosutic G, McAllister R. Oral insulin product hexyl-insulin monoconjugate 2 (HIM2) in type 1 diabetes mellitus: the glucose stabilization effects of HIM2. Diabetes technology & therapeutics. 2002; 4(4):459–466. https://doi.org/10.1089/152091502760306544.

[9] **Deng W**, Xie Q, Wang H, Ma Z, Wu B, Zhang X. Selenium nanoparticles as versatile carriers for oral delivery of insulin: insight into the synergic antidiabetic effect and mechanism. Nanomedicine: nanotechnology, biology and medicine. 2017; 13(6):1965–1974. https://doi.org/10.1016/j.nano.2017.05.002.

[10] **Bhattacharyya A**, Mukherjee D, Mishra R, Kundu P. Preparation of polyurethane–alginate/chitosan core shell nanoparticles for the purpose of oral insulin delivery. European Polymer Journal. 2017; 92:294–313. https://doi.org/10.1016/j.eurpolymj.2017.05.015.

[11] **Zhou Y**, Liu L, Cao Y, Yu S, He C, Chen X. A Nanocomposite Vehicle Based on Metal–Organic Framework Nanoparticle Incorporated Biodegradable Microspheres for Enhanced Oral Insulin Delivery. ACS applied materials & interfaces. 2020; 12(20):22581–22592. https://doi.org/10.1021/acsami.0c04303.

[12] **AGARWAL V**, REDDY IK, KHAN MA. Oral delivery of proteins: Effect of chicken and duck ovomucoid on the stability of insulin in the presence of α-chymotrypsin and trypsin. Pharmacy and Pharmacology Communications. 2000; 6(5):223–227. https://doi.org/10.1211/146080800128735935.

[13] **Kayser V**, Chennamsetty N, Voynov V, Forrer K, Helk B, Trout BL. Glycosylation influences on the aggregation propensity of therapeutic monoclonal antibodies. Biotechnology journal. 2011; 6(1):38–44. https://doi.org/10.1002/biot.201000091.

[14] **Raju TS**, Scallon BJ. Glycosylation in the Fc domain of IgG increases resistance to proteolytic cleavage by papain. Biochemical and biophysical research communications. 2006; 341(3):797–803. https://doi.org/10.1016/j.bbrc.2006.01.030.

[15] **Russell D**, Oldham NJ, Davis BG. Site-selective chemical protein glycosylation protects from autolysis and proteolytic degradation. Carbohydrate research. 2009; 344(12):1508–1514. https://doi.org/10.1016/j.carres.2009.06.033.

[16] **Losev Y**, Paul A, Frenkel-Pinter M, Abu-Hussein M, Khalaila I, Gazit E, Segal D. Novel model of secreted human tau protein reveals the impact of the abnormal N-glycosylation of tau on its aggregation propensity. Scientific reports. 2019; 9(1):1–10. https://doi.org/10.1038/s41598-019-39218-x.

[17] **van Veen HA**, Geerts ME, van Berkel PH, Nuijens JH. The role of N-linked glycosylation in the protection of human and bovine lactoferrin against tryptic proteolysis. European Journal of Biochemistry. 2004; 271(4):678–684. https://doi.org/10.1111/j.1432-1033.2003.03965.x.

[18] **Sareneva T**, Pirhonen J, Cantell K, Julkunen I. N-glycosylation of human interferon-γ: glycans at Asn-25 are critical for protease resistance. Biochemical Journal. 1995; 308(1):9–14. https://doi.org/10.1042/bj3080009.

[19] **Guan X**, Chaffey PK, Wei X, Gulbranson DR, Ruan Y, Wang X, Li Y, Ouyang Y, Chen L, Zeng C, et al. Chemically precise glycoengineering improves human insulin. ACS chemical biology. 2018; 13(1):73–81. https://doi.org/10.1021/acschembio.7b00794.

[20] **Mark AE**, Berendsen HJ, Van Gunsteren WF. Conformational flexibility of aqueous monomeric and dimeric insulin: a molecular dynamics study. Biochemistry. 1991; 30(45):10866–10872. https://doi.org/10.1021/bi00109a009.

[21] **Zoete V**, Meuwly M, Karplus M. A comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. Journal of molecular biology. 2004; 342(3):913–929. https://doi.org/10.1016/j.jmb.2004.07.033.

[22] **Yang C**, Lu D, Liu Z. How PEGylation enhances the stability and potency of insulin: a molecular dynamics simulation. Biochemistry. 2011; 50(13):2585–2593. https://doi.org/10.1021/bi101926u.

[23] **Pincus M**. Letter to the editor—a Monte Carlo method for the approximate solution of certain types of constrained optimization problems. Operations research. 1970; 18(6):1225–1228. https://doi.org/10.1287/opre.18.6.1225.

[24] **Isralewitz B**, Gao M, Schulten K. Steered molecular dynamics and mechanical functions of proteins. Current opinion in structural biology. 2001; 11(2):224–230. https://doi.org/10.1016/s0959-440x(00)00194-9.

[25] **Sugita Y**, Kitao A, Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. The Journal of Chemical Physics. 2000; 113(15):6042–6051. https://doi.org/10.1063/1.1308516.

[26] **Antoszewski A**, Feng CJ, Vani BP, Thiede EH, Hong L, Weare J, Tokmakoff A, Dinner AR. Insulin dissociates by diverse mechanisms of coupled unfolding and unbinding. The Journal of Physical Chemistry B. 2020; 124(27):5571–5587. https://doi.org/10.1021/acs.jpcb.0c03521.

[27] **Schilling RJ**, Mitra AK. Degradation of insulin by trypsin and alpha-chymotrypsin. Pharmaceutical research. 1991; 8(6):721–727. https://doi.org/10.1023/a:1015893832222.

[28] **Chaffey PK**, Guan X, Chen C, Ruan Y, Wang X, Tran AH, Koelsch TN, Cui Q, Feng Y, Tan Z. Structural insight into the stabilizing effect of O-glycosylation. Biochemistry. 2017; 56(23):2897–2906. https://doi.org/10.1021/acs.biochem.7b00195.

[29] **Torrie GM**, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics. 1977; 23(2):187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

[30] **Sugita Y**, Okamoto Y. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. Chemical Physics Letters. 2000; 329(3-4):261–270. https://doi.org/10.1016/S0009-2614(00)00999-4.

[31] **Lyubartsev A**, Martsinovski A, Shevkunov S, Vorontsov-Velyaminov P. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. The Journal of chemical physics. 1992; 96(3):1776–1783. https://doi.org/10.1063/1.462133.

[32] **Fávero-Retto MP**, Palmieri LC, Souza TA, Almeida FC, Lima LMT. Structural meta-analysis of regular human insulin in pharmaceutical formulations. European Journal of Pharmaceutics and Biopharmaceutics. 2013; 85(3):1112–1121. https://doi.org/10.1016/j.ejpb.2013.05.005.

[33] **Timofeev V**, Chuprov-Netochin R, Samigina V, Bezuglov V, Miroshnikov K, Kuranova I. X-ray investigation of gene-engineered human insulin crystallized from a solution containing polysialic acid. Acta Crystallographica Section F: Structural Biology and Crystallization Communications. 2010; 66(3):259–263. https://doi.org/10.1107/S1744309110000461.

[34] **Křížková K**, Veverka V, Maletínská L, Hexnerová R, Brzozowski AM, Jiráček J, Žáková L. Structural and functional study of the GlnB22-insulin mutant responsible for maturity-onset diabetes of the young. PloS one. 2014; 9(11):e112883. https://doi.org/10.1371/journal.pone.0112883.

[35] **DeLano WL**, et al. Pymol: An open-source molecular graphics tool. CCP4 Newsletter on protein crystallography. 2002; 40(1):82–92.

[36] **Anandakrishnan R**, Aguilar B, Onufriev AV. H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic acids research. 2012; 40(W1):W537–W541. https://doi.org/10.1093/nar/gks375.

[37] **Myers J**, Grothaus G, Narayanan S, Onufriev A. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. Proteins: Structure, Function, and Bioinformatics. 2006; 63(4):928–938. https://doi.org/10.1002/prot.20922.

[38] **Gordon JC**, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. H++: a server for estimating p K as and adding missing hydrogens to macromolecules. Nucleic acids research. 2005; 33(suppl_2):W368–W371. https://doi.org/10.1093/nar/gki464.

[39] **Kirschner KN**, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, Woods RJ. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. Journal of computational chemistry. 2008; 29(4):622–655. https://doi.org/10.1002/jcc.20820.

[40] **Abraham MJ**, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015; 1:19–25. https://doi.org/10.1016/j.softx.2015.06.001.

[41] **Páll S**, Abraham MJ, Kutzner C, Hess B, Lindahl E. Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In: *International conference on exascale applications and software* Springer; 2014. p. 3–27. https://doi.org/10.1007/978-3-319-15976-8_1.

[42] **Da Silva AWS**, Vranken WF. ACPYPE-Antechamber python parser interface. BMC research notes. 2012; 5(1):1–8. https://doi.org/10.1186/1756-0500-5-367.

[43] **Berendsen HJ**, Postma Jv, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. The Journal of chemical physics. 1984; 81(8):3684–3690. https://doi.org/10.1063/1.448118.

[44] **Bussi G**, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. The Journal of chemical physics. 2007; 126(1):014101. https://doi.org/10.1063/1.2408420.

[45] **Parrinello M**, Rahman A. Crystal structure and pair potentials: A molecular-dynamics study. Physical review letters. 1980; 45(14):1196. https://doi.org/10.1103/PhysRevLett.45.1196.

[46] **Parrinello M**, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied physics. 1981; 52(12):7182–7190. https://doi.org/10.1063/1.328693.

[47] **Essmann U**, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. The Journal of chemical physics. 1995; 103(19):8577–8593. https://doi.org/10.1063/1.470117.

[48] **Hess B**, Bekker H, Berendsen HJ, Fraaije JG. LINCS: a linear constraint solver for molecular simulations. Journal of computational chemistry. 1997; 18(12):1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[49] **Wilcox P**. [5] Chymotrypsinogens—chymotrypsins. In: *Methods in enzymology*, vol. 19 Elsevier; 1970.p. 64–108. https://doi.org/10.1016/0076-6879(70)19007-0.

[50] **Eisenhaber F**, Lijnzaad P, Argos P, Sander C, Scharf M. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. Journal of computational chemistry. 1995; 16(3):273–284. https://doi.org/10.1002/jcc.540160303.

[51] **Chodera JD**. A simple method for automated equilibration detection in molecular simulations. Journal of chemical theory and computation. 2016; 12(4):1799–1805. https://doi.org/10.1021/acs.jctc.5b00784.

[52] **Kendall MG**, et al. The advanced theory of statistics. Vols. 1. The advanced theory of statistics Vols 1. 1948; 1(Ed. 4). https://www.cabdirect.org/cabdirect/abstract/19501403039.

[53] **Schechter I**, Berger A. On the active site of proteases. III. Mapping the active site of papain; specific peptide inhibitors of papain. Biochemical and biophysical research communications. 1968; 32(5):898–902. https://doi.org/10.1016/0006-291X(68)90326-4.

[54] **Appel W**. Chymotrypsin: molecular and catalytic properties. Clinical biochemistry. 1986; 19(6):317–322. https://10.1016/s0009-9120(86)80002-9.

[55] **Hedstrom L**. Serine protease mechanism and specificity. Chemical reviews. 2002; 102(12):4501–4524. https://doi.org/10.1021/cr000033x.

[56] **Bode W**, Huber R. Natural protein proteinase inhibitors and their interaction with proteinases. EJB Reviews. 1993; p. 43–61. https://doi.org/10.1111/j.1432-1033.1992.tb16654.x.

[57] **Coombs GS**, Rao MS, Olson AJ, Dawson PE, Madison EL. Revisiting catalysis by chymotrypsin family serine proteases using peptide substrates and inhibitors with unnatural main chains. Journal of Biological Chemistry. 1999; 274(34):24074–24079. https://doi.org/10.1074/jbc.274.34.24074.

[58] **Gowers RJ**, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, Domanski J, Dotson DL, Buchoux S, Kenney IM, et al. MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2019.

[59] **Michaud-Agrawal N**, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. Journal of computational chemistry. 2011; 32(10):2319–2327. https://doi.org/10.1002/jcc.21787.

[60] **G N Ramachandran VS C Ramakrishnan**. Stereochemistry of polypeptide chain configurations. J mol Biol. 1963; 7:95–99. https://doi.org/10.1016/s0022-2836(63)80023-6.

[61] **Lovell SC**, Davis IW, Arendall III WB, De Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins: Structure, Function, and Bioinformatics. 2003; 50(3):437–450. https://doi.org/10.1002/prot.10286.

[62] **McGibbon RT**, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang LP, Lane TJ, Pande VS. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophysical journal. 2015; 109(8):1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015.

[63] **Baker EN**, Hubbard RE. Hydrogen bonding in globular proteins. Progress in biophysics and molecular biology. 1984; 44(2):97–179. https://doi.org/10.1016/0022-2836(92)91058-w.

[64] **Harding MM**, Hodgkin DC, Kennedy AF, O'Connor A, Weitzmann PDJ. The crystal structure of insulin: II. An investigation of rhombohedral zinc insulin crystals and a report of other crystalline forms. Journal of Molecular Biology. 1966 Mar; 16(1):212–IN30. https://doi.org/10.1016/S0022-2836(66)80274-7.

[65] **Antolíková E**, Žáková L, Turkenburg JP, Watson CJ, Hančlová I, Šanda M, Cooper A, Kraus T, Brzozowski AM, Jiráček J. Non-equivalent Role of Inter- and Intramolecular Hydrogen Bonds in the Insulin Dimer Interface*. Journal of Biological Chemistry. 2011 Oct; 286(42):36968–36977. https://doi.org/10.1074/jbc.M111.265249.

[66] **Bakan A**, Meireles LM, Bahar I. ProDy: Protein Dynamics Inferred from Theory and Experiments. Bioinformatics. 2011 Jun; 27(11):1575–1577. https://doi.org/10.1093/bioinformatics/btr168, publisher: Oxford Academic.

[67] **Virtanen P**, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020; 17:261–272. doi: 10.1038/s41592-019-0686-2.

[68] **Harris CR**, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Fernández del Río J, Wiebe M, Peterson P, Gérard-Marchant P, et al. Array programming with NumPy. Nature. 2020; 585:357–362. https://doi.org/10.1038/s41586-020-2649-2.

[69] **Wilson EB**. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association. 1927 Jun; 22(158):209–212. https://doi.org/10.1080/01621459.1927.10502953, publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1927.10502953.

[70] **Newcombe RG**. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine. 1998; 17(8):857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.

[71] **Wallis S**. Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. Journal of Quantitative Linguistics. 2013 Aug; 20(3):178–208. https://doi.org/10.1080/09296174.2013.799918.

[72] **Ward MD**, Zimmerman MI, Meller A, Chung M, Swamidass S, Bowman GR. Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. Nature communications. 2021; 12(1):1–12. https://doi.org/10.1038/s41467-021-23246-1.