Supporting Information: 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)

1 ~~Supplemental Tables~~

1 Characterization of disordered elements

To examine whether our simulations captured important structural features of wild-type insulin, we compared our simulations with the wild-type insulin ensemble investigated in the paper [1] by Busto-Moner et al., which presents a pipeline for sampling wild-type insulin ensemble at room temperature. The pipeline integrates parallel tempering [2, 3], the integrated variational approach for conformational dynamics (IVAC) [4, 5], Markov state model (MSM) [6, 7], and Perron cluster analysis (PCCA) [8]. Based on the clusters generated by PCCA and the first three tICs obtained in IVAC, the authors concluded that 60% of insulin structures exhibited at least one of the following elements of disorder: melting of A-chain N-terminus helix (A1-A9), detachment of B-chain N-terminus (B1-B7) and detachment of B-chain C-terminus (B20-B30).

We note that the results from that study are not directly comparable to our study, as the two studies look at significantly different pH ranges. Specifically, Busto-Moner et al. investigated the insulin ensemble at pH 2.5, which is very different from the neutral pH values adopted in our study. Since insulin is a pH-sensitive protein, ensembles at distinct pH values are not directly comparable. However, we repeated some of their analysis, as there is likely to be a partial agreement in some of the metrics.

To compare our wild-type insulin simulations with the referred paper, we examined our simulations to see if the above-mentioned elements of disorder were also present. We adopted the same definitions of the disordered elements used by Busto-Moner et al, as summarized below.

(1) Melting of A-chain N-terminus helix (A1-A9)

Melting of A-chain N-terminus helix is characterized by the sum of helicity of four segments along residues A1-A9, each of which contains six contiguous residues (A1-A6, A2-A7, A3-A8, and A4-A9). By definition, the helicity of a segment can be defined by a switching function:

$$H = \sum_i \frac{1-(r/0.08)^8}{1-(r/0.08)^{12}} \tag{1}$$

In the equation, r is the RMSD with respect to an idealized helix model. Note that such a helix model was not provided by the referred paper, so we constructed our own model using cg_openmm with idealized helical parameters (helix radius $r = 2.3$Å, pitch $p = 5.5$Å, and distance between C_α atoms $d = 3.8$Å) [9, 10]. With an idealized helix model, the A-chain N-terminus helix is considered melted if $H_{A1\text{-}A9} = H_{A1\text{-}A6} + H_{A2\text{-}A7} + H_{A3\text{-}A8} +$

(2) Detachment of B-chain N-terminus (B1-B7)
Detachment of B1-B7 is characterized by two angles: θ, which is formed by the C_α atoms of residues B3, B9, and B20; ψ (dihedral), which is formed by the C_α atoms of residues B3, A15, B18, and B15. Structures with $\theta > 85°$ and $\psi > 10°$ are considered as having B1-B7 detachment.

(3) Detachment of B-chain C-terminus (B20-B30)
Structures with the dihedral formed by the C_α atoms of residues A13, A19, B19, and B25 smaller than 0° are considered as having B20-B30 detachment.

As summarized in (Supplemental Table S1), our simulations of wild-type insulin structures did capture the 3 elements of disorder described in the referred paper, but with different percentages compared to the values reported by Busto-Moner et al, with the differences in pH being the most likely, though not only, explanation. Notably, 2MVC exhibits a very different percentage of B1-B7 detachment from others, which is probably associated with the fact that it was the only model resolved by NMR at acidic conditions. However, as investigated in the main text, even if the method for solving the structures does have influences on our simulations, it does not have obvious impacts on the efficacy of our metrics.

Model	A1-A9 helix melting	B1-B7 detachment	B20-B30 detachment
4EYD	38%	5.4%	100%
4EY9	35%	0.49%	96%
4EY1	36%	3.2%	99%
3I3Z	42%	1.6%	92%
2MVC	40%	51%	99%
Total	38%	12%	97%
Reference	24%	44%	15%

Table S1. The percentage of each disordered element calculated from the wild-type insulin simulations compared with the reported values in the paper by Busto-Moner et al. as the reference.

Overall, our simulations agree qualitatively with the Busto-Moner study with the presence of the three key areas of disorder. However, a quantitative agreement is not expected because their simulations were carried out at pH 2.5, unlike the neutral pH of our study."

2 Influence of transitions between states on the ranges of our metrics

As shown in Supplemental Figure S1, we examined the transition between states for each of the wild-type models by calculating the pairwise RMSD between any two configurations in the trajectory. As a result, at least one major transition occurred for each of the wild-type models except for 4EYD, which did not have any clear transition between states. To examine whether the ranges of metrics vary significantly with these transitions between different states, we first used the pruned exact linear time (PELT) algorithm [11] to identify the time frame where the transition occurred, assuming only 1 major transition. As a result, the change points in the pairwise RMSD of 4EY9, 4EY1, 3I3Z, and 2MVC were at 621.25 ns, 1120.0 ns, 652.5 ns, and 1477.5 ns, respectively.

As such, for these 4 wild-type models, we repeated data analyses that did not involve the glycan moiety, for the states before and after the major transition in pairwise RMSD. These analyses included the calculations of 8 measures involved in the first 3 metrics for the proteolytic stability, which are the scissile bond SASA of B25-B26 and B26-B27, residue SASA of B24 and B25, and the β-sheet propensity of residues B22 to B25. As shown in Supplemental Table S2, SASA measures generally did not vary significantly in their values upon transition between states, while at least one of the β-sheet propensity measure (e.g. the β-sheet propensity at residue B25) show a large change after the major transition occurs. This implies that our simulation might not comprehensively sample the configurational space of wild-type insulin, which is also reflected by the fact observed from the pairwise RMSD that the system did not have frequent major transitions back and forth between different states. However, frequent minor transitions between states shown in

pairwise RMSD suggest that the simulation still captures a large amount of configurational diversity and long-timescale events. We emphasize that the goal of our study is not to comprehensively sample the whole configurational space of insulin and its glyco-variants but to develop reasonable metrics that work with MD simulations requiring manageable computational cost and distinguish variants with certain properties from their counterparts.

		Metric 1: Scissile bond SASA		Metric 2: P1 site SASA		Metric 3: β-sheet propensity			
		B25-B26	B26-B27	B24	B25	B22	B23	B24	B25
4EY9	Before transition	0.07 ± 0.02	0.12 ± 0.03	0.61 ± 0.05	1.58 ± 0.09	0%	0%	98%	89%
	After transition	0.15 ± 0.02	0.10 ± 0.02	0.51 ± 0.04	1.49 ± 0.05	0%	0%	96%	33%
	Overall	0.13 ± 0.02	0.11 ± 0.01	0.54 ± 0.04	1.52 ± 0.05	0%	0%	97%	50%
4EY1	Before transition	0.12 ± 0.02	0.11 ± 0.02	0.51 ± 0.05	1.40 ± 0.08	29%	13%	95%	51%
	After transition	0.07 ± 0.02	0.11 ± 0.02	0.57 ± 0.06	1.30 ± 0.08	47%	20%	89%	84%
	Overall	0.10 ± 0.02	0.11 ± 0.01	0.54 ± 0.04	1.36 ± 0.06	37%	16%	92%	65%
3I3Z	Before transition	0.14 ± 0.02	0.11 ± 0.02	0.52 ± 0.06	1.46 ± 0.09	0%	0%	96%	35%
	After transition	0.06 ± 0.01	0.08 ± 0.02	0.59 ± 0.08	1.24 ± 0.08	0%	0%	94%	84%
	Overall	0.09 ± 0.02	0.09 ± 0.02	0.56 ± 0.06	1.31 ± 0.06	0%	0%	95%	68%
2MVC	Before transition	0.04 ± 0.01	0.06 ± 0.01	0.54 ± 0.06	1.17 ± 0.05	58%	20%	96%	96%
	After transition	0.11 ± 0.03	0.12 ± 0.02	0.50 ± 0.05	1.36 ± 0.10	70%	24%	95%	47%
	Overall	0.06 ± 0.01	0.07 ± 0.01	0.53 ± 0.04	1.22 ± 0.05	61%	21%	96%	83%

Table S2. Data analysis of the first three metrics for the proteolytic stability for each state of each wild-type model. Note that the data of SASA are in the units of nm^2. The data analysis repeated here for each state follows the same method in Section 2.

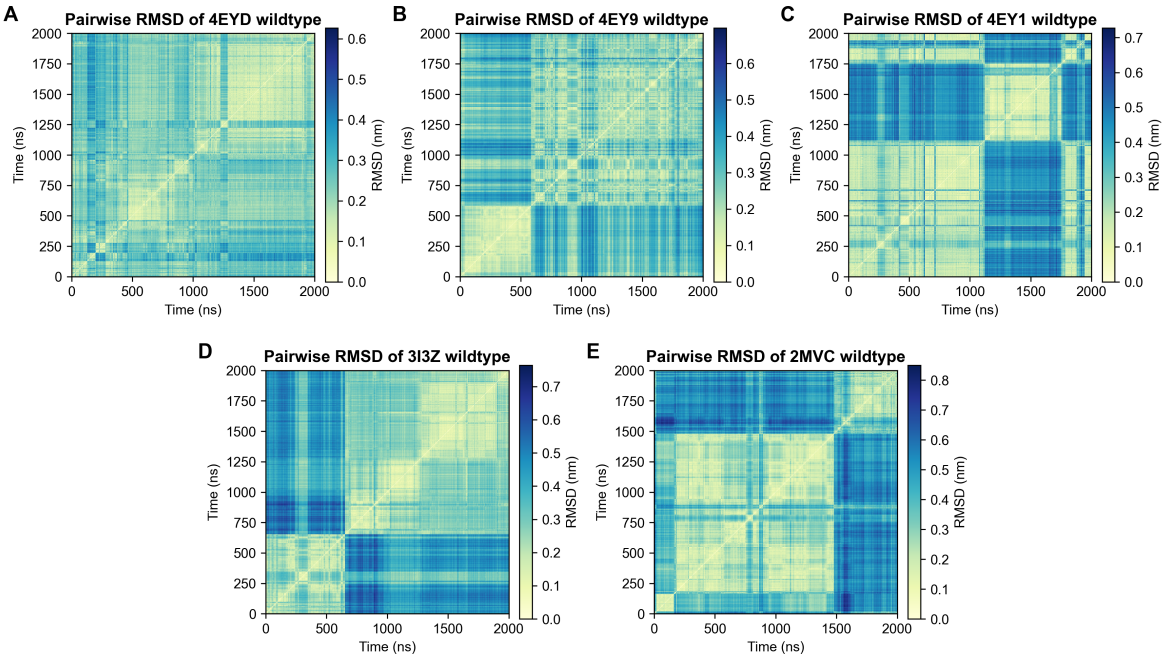

Figure S1. The pairwise RMSD of calculated from the 250ps-spaced MD trajectory of each wild-type model, including 4EYD (A), 4EY9 (B), 4EY1 (C), 3I3Z (D), and 2MVC (E). As shown in the figure, the major transitions occur around 500–1500 ns, we therefore concluded that at least 2000 ns was required to sample the configurational ensemble of insulin.

3 Metric distributions

As supplemental statistics of our metrics, in Supplemental Figure S2, we provide the distributions of the SASA of the insulin scissile bonds and P1 sites. Note that we present kernel density estimation (KDE) [12, 13]

of the distributions instead of histograms so that the highly overlapped data would not obscure each other. In all the KDE plots, Scott's Rule [14] was used to automatically determine the smoothing bandwidth such that the KDE plots were visually consistent with the histograms of the data. As shown in the figure, the distributions of the scissile bond SASA all have multiple peaks, which reflects the fact that the SASA of the scissile bond is largely influenced by the orientation of the two adjacent residues. On the other hand, the two distributions of the P1 site SASA only have one peak. Importantly, in all metrics presented in the figure, GF 13 peaks at apparently lower SASA values compared to other variants, which is consistent with our finding mentioned in the main text that Metric 1 and Metric 2 had better predictiveness for the more proteolytically stable variants. Notably, we do not plot distributions for the other three metrics (β-sheet propensity, fraction of glycan-involved hydrogen bonds, and glycan-dimer occlusion fraction) presented in the main text because these metrics themselves are not time series but counts of occurrences.

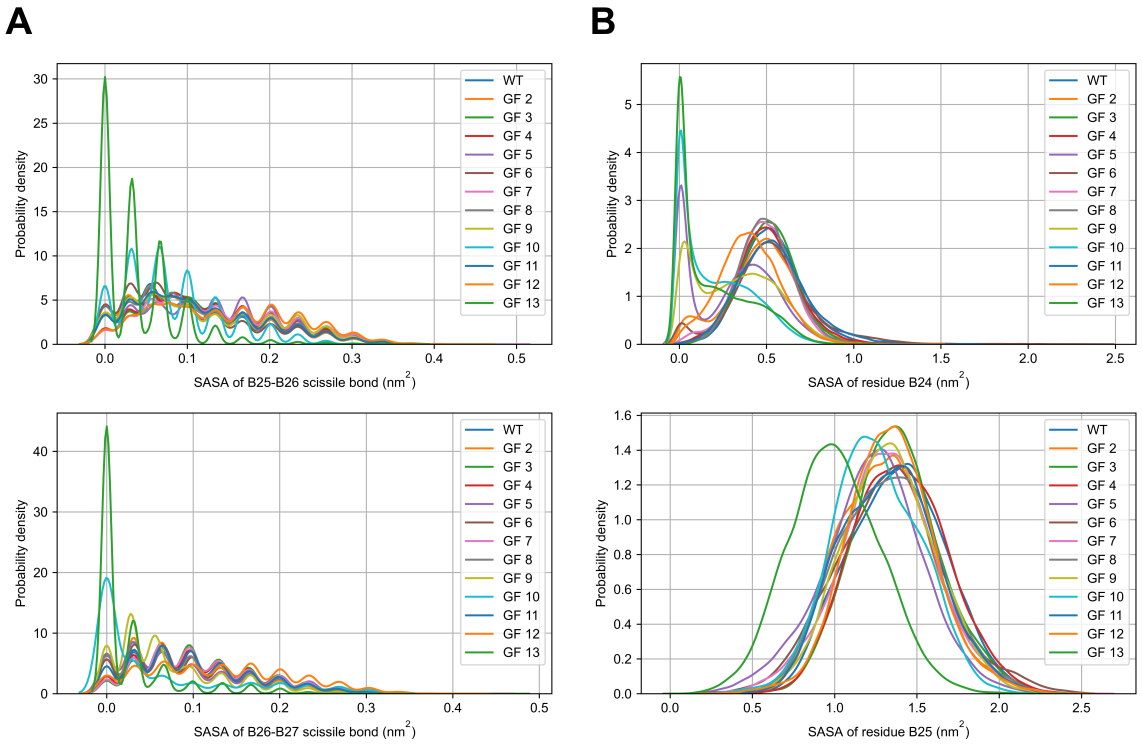

Figure S2. The kernel density estimation (KDE) of distributions of the SASA at the scissile bonds and the P1 sites. (A) The distributions of the SASA at the scissile bonds between residues B25 and B26 and between residues B26 and B27. (B) The distributions of the SASA at the P1 sites, residues B24 and B25.

4 Metric Correlations

We examined the correlations between the 8 measures involved in the first 3 metrics for the proteolytic stability, which resulted in 28 combinations. (The 8 measures include scissile bond SASA of B25-B26 and B26-B27, residue SASA of B24 and B25, and β-sheet propensity of residues B22 to B25.) In the correlation plots in Figure S3 to Figure S5, we annotated the Pearson correlation coefficients with its uncertainty estimated from bootstrapping where the bootstrap samples of both metric variables were drawn from values based on different wild-type models. We use Pearson correlation coefficients here by assuming a linear relationship between any two variables of interest. As a result, correlations between any two SASA measures (Figure S3) tend to be stronger compared to the ones that involve at least one β-sheet propensity measure (Figure S4 and Figure S5). This is expected because SASA measures are generally more predictive for the proteolytic stability than the β-sheet propensity measures. Correlating two measures that are highly associated with proteolytic stability naturally leads to a stronger correlation. The only exception is the correlation between

β-sheet propensity of B22 and B23, which both are not predictive for the proteolytic stability but highly correlated with each other. This is probably because these two residues are adjacent to each other and contribute to similar secondary structures.

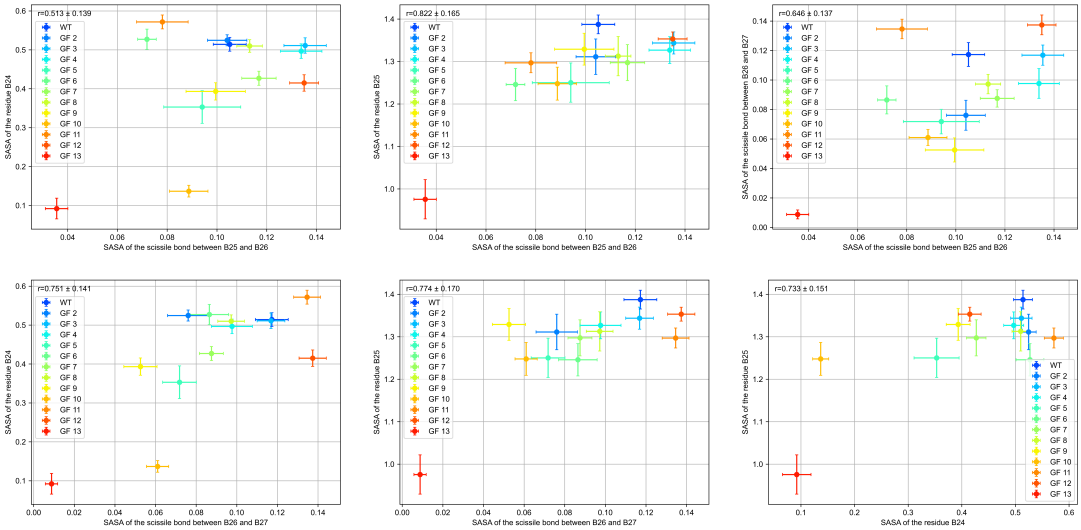

Figure S3. The correlation plots between any two SASA measures in Metric 1 and Metric 2 for the proteolytic stability. The Pearson correlation coefficients and their uncertainties determined from bootstrapping are annotated.

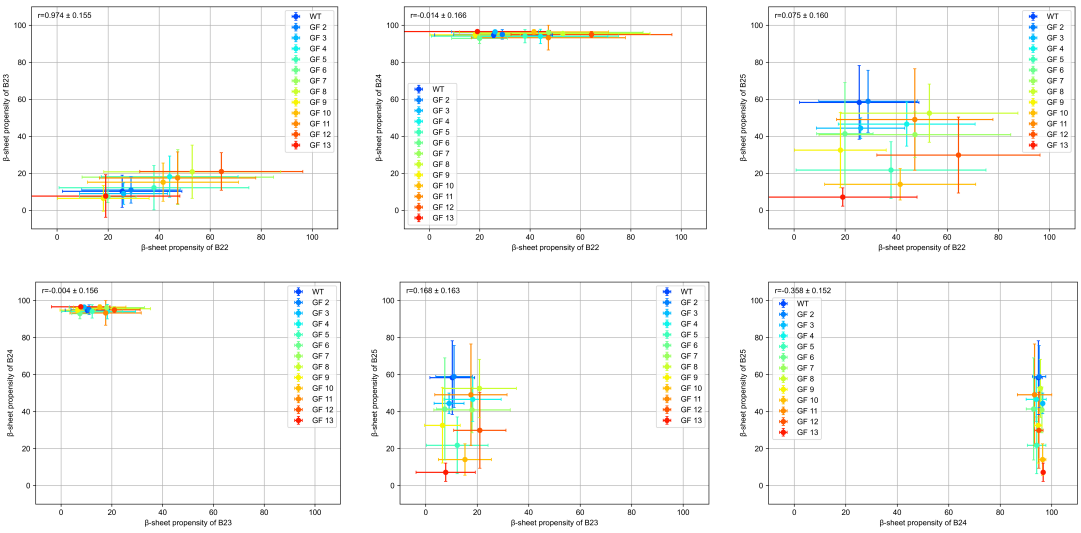

Figure S4. The correlation plots between any two β-sheet propensity measures in Metric 3 for the proteolytic stability. The Pearson correlation coefficients and their uncertainties determined from bootstrapping are annotated.

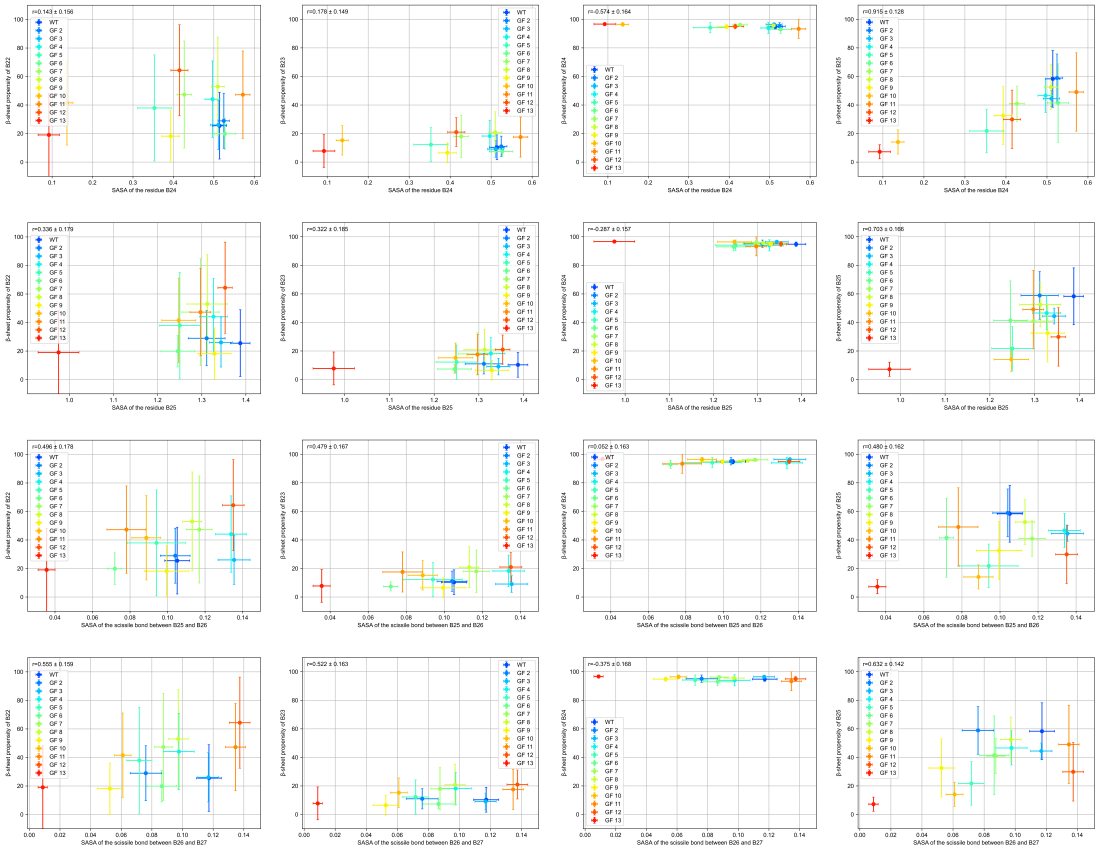

Figure S5. The correlation plots between any SASA measure in Metric 1 or Metric 2 and any β-sheet propensity measure in Metric 3 for the proteolytic stability. The Pearson correlation coefficients and their uncertainties determined from bootstrapping are annotated.

5 Additional Tables

	4EYD	4EY1	3I3Z	4EY9	2MVC
Total charges	-1	-1	-1	-1	-1
pH value	~~8.0~~ 6.9-7.8	~~8.0~~ 6.9-7.8	~~7.9~~ 6.6-6.9	~~6.9~~ 7.7-7.9	~~7.3~~ 6.5-7.3
HisB5	HIP (+1)	HIP (+1)	HIP (+1)	HIP (+1)	HIE (+0)
HisB10	HIE (+0)	HIE (+0)	HIE (+0)	HID (+0)	HIP (+1)

Table S3. ~~The comparison~~ Comparison of histidine protonation states of wild-type structures with the range of input pH values to H++ consistent with this protonation. Note that the pH ranges corresponding to the fixed total charges of -1 could vary between different models, since pH values from H++ are only approximations based on a single input structure, instead of the entire ensemble.

	GF 2	GF 3	GF 4	GF 5	GF 6	GF 7	GF 8	GF 9	GF 10	GF 11	GF 12	GF 13
4EYD	NA	NA	3.45	63.56	69.11	NA	NA	45.79	121.70	21.65	140.40	24.77
4EY1	NA	NA	6.89	371.83	71.04	NA	19.37	205.24	159.14	58.17	360.45	637.20
4EY9	NA	NA	18.53	274.66	97.34	NA	NA	12.55	496.98	0.39	89.13	57.77
3I3Z	NA	NA	93.49	62.86	199.65	NA	NA	45.27	62.80	42.78	97.81	281.52
2MVC	3.98	NA	21.31	54.78	105.28	NA	NA	114.61	59.62	58.75	66.00	81.57

Table S4. Dimer occlusion autocorrelation lag times for each of the listed models. All numbers are listed in units of nanoseconds.

#	4EYD-based GFs	4EY1-based GFs	4EY9-based GFs	3I3Z-based GFs	2MVC-based GFs
1	N/A	N/A	N/A	N/A	N/A
2	None	ThrA8(OG1)-GalNAc[1](O2N):10%	None	None	GlnA5(NE2)-GalNAc[1](O6): 14% GlnA15(NE2)-GalNAc[1](O6): 13%
3	None	None	None	None	GlnA15(NE2)-52(O2N):11%
4	None	None	None	None	None
5	None	PheB24(N)-GalNAc[1](O3): 48% ThrB30(N)-GalNAc[1](O2N): 16%	PheB24(N)-GalNAc[1](O3): 11%	ThrB27(N)-GalNAc[1](O3):15% Glu4(N)-GalNAc[1](O2N): 14%	PheB24(N)-GalNAc[1](O3): 10%
6	ThrB27(N)-GalNAc[1](O3):26%	ThrB27(N)-GalNAc(1)(O3):26%	None	None	CysA7(N)-GalNAc[1](O2N): 18% ThrB27(N)-GalNAc(1)(O3): 12%
7	None	None	None	None	GlnA15(NE2)-Man[1](O4): 18%
8	GlnA5(NE2)-Man[2](O2): 12%	GlnA5(NE2)-Man[2](O2): 11%	None	GlnA5(NE2)-Man[1](O5): 12%	None
9	PheB24(N)-Man[1](O3): 22% TyrB16(OH)-Man[1](O4): 13%	None	PheB24(N)-Man[1](O3): 27% TyrB16(OH)-Man[1](O4): 19%	None	None
10	PheB24(N)-Man[1](O3): 52% ThrB30(N)–Man[2](O2): 35% ThrB27(N)–Man[2](O6): 23%	ThrB27(N)-Man[2](O6): 46% PheB24(N)-Man[1](O3): 30% TyrB16(OH)-Man[1](O4): 15%	ThrB27(N)-Man[2](O6): 45% PheB24(N)-Man[1](O3): 28% TyrB16(OH)-Man[1](O4): 11%	ThrB27(N)-Man[2](O6): 47% PheB24(N)-Man[1](O3): 15%	ThrB27(N)-Man[2](O6): 47% PheB24(N)-Man[1](O3): 37% TyrB16(OH)-Man[1](O4): 14%
11	None	None	None	None	None
12	ThrB30(N)-Man[2](O6): 31% GlyB23(N)-Man[1](O3): 10%	TyrA19(OH)-Man[1](O4): 51% ThrB27(N)-Man[1](O3): 43% ThrB30(N)-Man[2](O6): 19%	ThrB30(N)-Man[2](O6): 24% ValA3(N)-Man[2](O2): 15%	ThrB30(N)–Man[2](O6): 38%	ThrB30(N)-Man[2](O6): 49% GlyB8(N)-Man[2](O4): 10%
13	ThrB27(N)-Man[2](O6): 61% PheB24(N)-Man[1](O3): 49% TyrB16(OH)-Man[1](O4): 16%	ThrB27(N)-Man[2](O6): 31% ThrB27(N)-Man[3](O6): 14% PheB24(N)-Man[1](O3): 12% TyrB16(OH)-Man[1](O4): 11%	ThrB27(N)-Man[2](O6): 27% PheB24(N)-Man[1](O3): 17% ThrB27(N)-Man[3](O6): 15% TyrB16(OH)-Man[1](O4): 10%	ThrB27(N)-Man[2](O6): 43% PheB24(N)-Man[1](O3): 25% TyrB16(OH)-Man[1](O4): 12%	ThrB27(N)-Man[2](O6): 44% PheB24(N)-Man[1](O3): 22%

Table S5. The glycan-involved hydrogen bonds and their existence percentages of each glycoform.

Atom type	Role	Description
N	Donor	An sp^2 nitrogen in amide group
NE2	Donor	An epsilon nitrogen.
OH	Donor	An alcohol oxygen in Tyr
OG1	Donor	An alcohol oxygen in Thr
O2	Acceptor	The oxygen atom connected to the second carbon atom of the sugar
O3	Acceptor	The oxygen atom connected to the third carbon atom of the sugar
O4	Acceptor	The oxygen atom connected to the fourth carbon atom of the sugar
O5	Acceptor	The oxygen atom connected to the fifth carbon atom of the sugar
O6	Acceptor	The oxygen atom connected to the sixth carbon atom of the sugar
O2N	Acceptor	The oxygen atom of the N-acetyl group

Table S6. The atom types involved in the glycan-involved hydrogen bonds.

	least occlusion								*most occlusion*			
4EYD	2	3	7	8	4	11	12	6	**9**	**10**	5	**13**
4EY1	2	3	7	8	4	11	6	12	**9**	**10**	5	**13**
4EY9	2	3	7	8	11	4	6	12	5	**9**	**10**	**13**
3I3Z	2	3	7	8	4	6	11	12	**9**	5	**13**	**10**
2MVC	3	7	8	2	4	11	6	12	**9**	5	**10**	**13**
	low batch				*medium batch*				*high batch*			

Table S7. Glycoforms ordered from most to least proportion occlusion, based on proportion of simulation with measured occlusion.

6 ~~Supplemental~~ Additional Figures

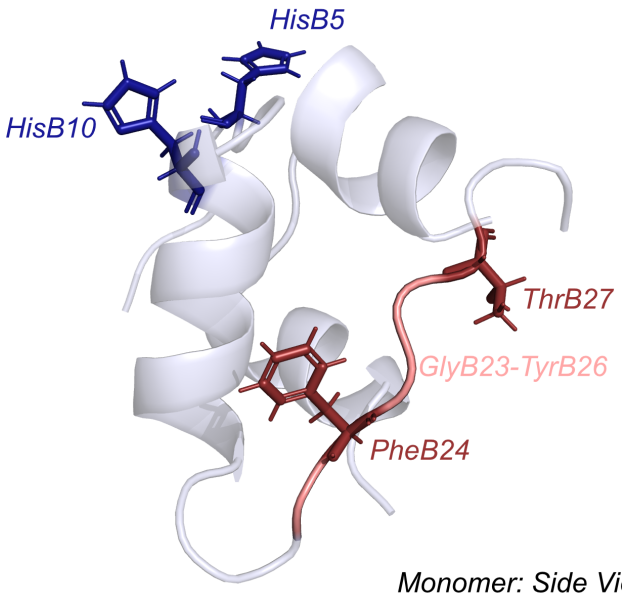

Figure S6. The histidine residues (HisB5, and HisB10, colored in blue) are shown with residues PheB24 and ThrB27 (colored red) which are important in enhancing proteolytic stability and residues ~~LeuB17-ArgB22 (colored green) and~~ GlyB23-TyrB26 (colored salmon), important in determining dimerization potential.

~~The pairwise RMSD of calculated from the 250ps-spaced MD trajectory of each wild-type model, including 4EYD (A), 4EY9 (B), 4EY1 (C), 3I3Z (D), and 2MVC (E). As shown in the figure, the major transitions occur around 500–1500 ns, we therefore concluded that at least 2000 ns was required to sample the configurational ensemble of insulin.~~

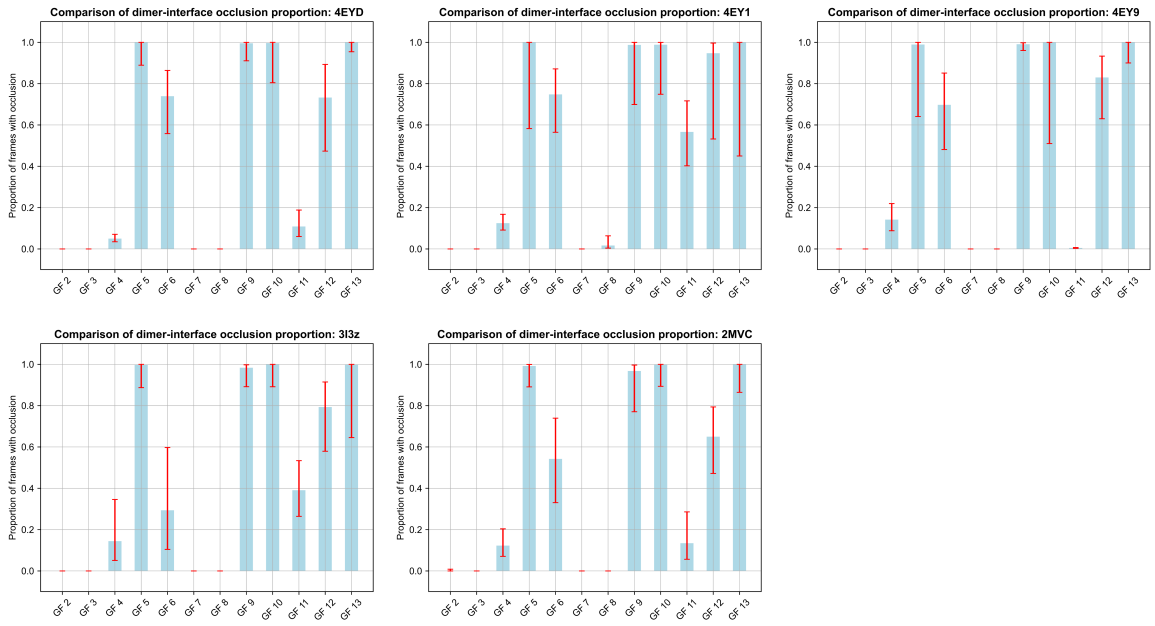

Figure S7. Proportion of frames with glycan-dimer occlusion for each glycoform. Red bars represent the asymmetric 95% Wilson score confidence interval.

References

[1] **Busto-Moner L**, Feng CJ, Antoszewski A, Tokmakoff A, Dinner AR. Structural Ensemble of the Insulin Monomer. Biochemistry. 2021; 60(42):3125–3136.

[2] **Hansmann UH**. Parallel tempering algorithm for conformational studies of biological molecules. Chemical Physics Letters. 1997; 281(1-3):140–150.

[3] **Earl DJ**, Deem MW. Parallel tempering: Theory, applications, and new perspectives. Physical Chemistry Chemical Physics. 2005; 7(23):3910–3916.

[4] **Nuske F**, Keller BG, Pérez-Hernández G, Mey AS, Noé F. Variational approach to molecular kinetics. Journal of chemical theory and computation. 2014; 10(4):1739–1752.

[5] **Lorpaiboon C**, Thiede EH, Webber RJ, Weare J, Dinner AR. Integrated variational approach to conformational dynamics: A robust strategy for identifying eigenfunctions of dynamical operators. The Journal of Physical Chemistry B. 2020; 124(42):9354–9364.

[6] **Prinz JH**, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. Markov models of molecular kinetics: Generation and validation. The Journal of chemical physics. 2011; 134(17):174105.

[7] **Bowman GR**, Pande VS, Noé F. An introduction to Markov state models and their application to long timescale molecular simulation, vol. 797. Springer Science & Business Media; 2013.

[8] **Schütte C**, Fischer A, Huisinga W, Deuflhard P. A direct approach to conformational dynamics based on hybrid Monte Carlo. Journal of Computational Physics. 1999; 151(1):146–168.

[9] **Guo Z**, Kraka E, Cremer D. Description of local and global shape properties of protein helices. Journal of molecular modeling. 2013; 19(7):2901–2911.

[10] **Tozzini V**. Minimalist models for proteins: a comparative analysis. Quarterly reviews of biophysics. 2010; 43(3):333–371.

[11] **Killick R**, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association. 2012; 107(500):1590–1598.

[12] **Davis RA**, Lii KS, Politis DN. Remarks on some nonparametric estimates of a density function. In: *Selected Works of Murray Rosenblatt* Springer; 2011.p. 95–100.

[13] **Parzen E**. On estimation of a probability density function and mode. The annals of mathematical statistics. 1962; 33(3):1065–1076.

[14] **Scott DW**. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons; 2015.