Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Information

A kinetic model reveals the critical gating motifs for donor-substrate loading into *Actinobacillus pleuropneumoniae N*-glycosyltransferase

Zhiqiang Hao[#], Qiang Guo[#], Wenjie Peng and Lin-Tai Da**

Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China.

*Correspondence to: darlt@sjtu.edu.cn (L.D.) or wenjiep@sjtu.edu.cn (W.P.)

[#] These authors contributed equally.

Experimental Methods

Modeling of UDP-Glc bound ApNGT^{Q469A} and wild-type ApNGT complexes

UDP-Glc-ApNGT^{Q469A} or UDP-Glc-ApNGT complex was modeled as previously described.[1] The final complex model was then subjected to energy minimization by the steepest decent and conjugate gradient methods to relieve local steric clashes using the Amber18 package.[2, 3] The protein was described using the Amber ff14SB force field.[4] To obtain the Amber parameters of UDP-Glc, the electrostatic potential of the ligand was firstly calculated using the HF/6-31G* method by the Gaussian 09 package, and then the restrained electrostatic potential (RESP) charges of UDP-Glc were determined using the antechamber program.[5, 6] Other parameters of UDP-Glc were adopted from GAFF2.[7] The complex structure was solvated in a cubic box filled with TIP3P water molecules, and 21 Na^+ ions were added to neutralize the system. The final energy-minimized complex was used for structural analyses.

Obtaining initial ligand binding pathways by steered molecular dynamics (SMD) simulations

Based on the above energy-minimized UDP-Glc-ApNGT^{Q469A} complex, we performed SMD simulations to pull the ligand out of the binding pocket. The SMD simulations were carried out with the Amber18 package.[2, 3] The parameters of the protein and UDP-Glc were the same as described above. We firstly conducted a 100-ps heating MD simulation, followed by a 500-ps equilibration MD simulation by restraining all heavy atoms of the solutes. The last equilibrated MD snapshot was then applied for the final SMD simulations.

We performed a total of six SMD simulations using varied regions of ApNGT^{Q469A} as the reference groups, namely, the residues M510−V512, N513−P516, F517−N519, N521−I524, H542−I545 and D546−G548. For each SMD simulation, the pulling direction was determined as the vector from the given reference group pointing to UDP-Glc (directions 1−6 in Fig. S1). The pulling force was imposed on UDP-Glc with a force constant of 0.2 kcal/mol/Å^2 and a pulling distance of 40 Å. The temperature was controlled by the Langevin thermostat at 310 K, and the SHAKE algorithm was used to constrain the bonds involving hydrogen atoms.[8, 9] Non-bonded interactions were cut off at 12 Å and long-range electrostatic interactions were treated with the particle mesh Ewald (PME) method.[10] The simulation time for each SMD simulation was 20 ns. Thus, we finally obtained six ligand releasing pathways, which presumably corresponded to the reverse process of ligand binding pathways.

Generating simulation dataset for Markov state model (MSM) construction

Based on the chosen SMD trajectory, we then performed extensive unbiased MD simulations to sample the conformational space of UDP-Glc along its binding pathway. To choose the representative conformations from the SMD trajectory, we firstly projected the high-dimensional simulation dataset onto low-dimensional space using the time-structure independent component analysis (tICA) implemented in MSMBuilder 3.8.0, which is an effective dimensionality-reduction tool capable of capturing the slowest dynamics for conformational changes of biomolecules.[11, 12] In specific, the SMD conformations were featurized by binary transforms of ligand-protein contacts (only heavy atoms were taken into account) with a default distance

cutoff of 8 Å. The protein residues within 10 Å from the ligand in the energy-minimized UDP-Glc-ApNGT^{Q469A} complex were treated as the binding-pocket residues. If any of these residues had a distance less than 8 Å from the ligand, it was then considered as a contact-residue of the ligand.

Next, the above chosen features were used as the input dataset for the following tICA, thereby the high-dimensional simulation dataset was projected onto the top five tICs. We further grouped the projected low-dimensional dataset into 100 clusters using the *k*-centers algorithm. To eliminate the bias from the SMD simulation, we firstly carried out a 10-ns unbiased MD simulation for each cluster-center conformation. Then, 100 structures were randomly selected from the 100 last snapshots, and each was subjected to a 100-ns unbiased MD simulation for the final MSM construction. To ensure that enough transitions could be observed, additional 40 conformations in which UDP-Glc was located between unbound and bound states were chosen and each was subjected to a 100-ns unbiased MD simulation. Finally, we collected a total of 140 MD trajectories of 100 ns for MSM construction, with aggregated simulation time of 14 μs.

Setup of unbiased MD simulations

Unbiased MD simulations were performed with the GROMACS 2019.4 software package.[13] The parameters of the protein and UDP-Glc were the same as described above. Each complex was solvated in a triclinic box filled with 27,598 TIP3P water molecules, and 21 Na^+ ions were added to neutralize the system. The final system contained a total of 92,717 atoms. Long-range electrostatic interactions were treated with the PME method.[10] Van der Waals and short-range electrostatic interactions were

cut off at 12 Å. All chemical bonds were constrained using the LINCS algorithm.[14] The solvated system was energy-minimized with the steepest decent method, followed by a 100-ps NVT equilibration MD simulation and a 500-ps NPT equilibration MD simulation by restraining all heavy atoms of the solutes. Finally, the production MD simulation was conducted under NPT conditions at 310 K and 1 bar using the velocity rescaling thermostat and Berendsen barostat, respectively.[15, 16]

MSM construction

MSM construction proceeds from the discretization of the conformational space of the system of interest and the description of the dynamics of system as a sequence of transitions between all discrete states.[17-19] An optimally discretized MSM exhibits converging timescales with high probability of transition among kinetically similar states and lower probability between kinetically separated states.[17-19] From this model, the pathways and kinetic rates between distinct conformations can be calculated.[17-19] In this study, we constructed a Bayesian MSM by PyEMMA 2.5.7 following the recommended procedure.[20] As described above, we employed the contact map between ligand and protein as the structural features for dimensionality reduction using tICA. The final dataset was projected onto the top two tICs and then grouped into 600 microstates using the *k*-means algorithm. We validated the MSM by plotting the implied-timescale curves that converged well (Fig. S3a) and performing the Chapman-Kolmogorov test (Fig. S4). We also analyzed the slowest processes by inspecting the values of the top 2–6 eigenvectors projected onto the top two tICs (Fig. S5). To obtain a humanly interpretable model of the system, we further lumped the microstates into

six macrostates using the Perron-cluster cluster analysis (PCCA++) algorithm.[21] Transition path theory (TPT) was applied to investigate dominant transition paths of the ligand binding process.[22, 23]

The choice of the number of macrostates is still an open question.[24, 25] In spectral-based methods, like PCCA++ used in this study, the number of coarse-grained macrostates has often been chosen based on the existence of a gap in the eigenvalue spectrum of the transition probability matrix (Fig. S3b).[24, 25] However, the choice of the number of macrostates is generally very subjective due to the continuum of eigenvalues.[25] We firstly plotted the free energy landscape by mapping all MD conformations onto the top two tICs, which clearly indicates six distinct low-energy basins (Fig. S3c). From a biological viewpoint, this six-metastable-state model could well describe the ligand binding process at a desirable resolution for understanding the molecular mechanism underlying the enzyme–substrate recognition (Fig. S3d). In specific, S1 and S6 are the completely unbound and bound states, respectively; while other four states, S2–S5, are key intermediate states during the donor-substrate loading. Particularly, the S2→S4 transition is the rate-limiting step, which therefore can guide us to pinpoint the critical gating motifs responsible for regulating the loading dynamics. In comparison, we also constructed a three-metastable-state model (Fig. S3e). This relatively low-resolution model fails to isolate the completely unbound and bound states that are biologically significant (namely S1 and S6 in Fig. S3d). Therefore, we conclude that the six-metastable-state model is a reasonable model to elucidate the donor-substrate loading dynamics.

Calculations of mean first-passage time (MFPT) and equilibrium populations

We used the transition probability matrix of the constructed 600-microstate MSM to generate Monte Carlo (MC) simulation trajectories.[26] In specific, a random number between 0 and 1 was chosen at every step to determine which microstate the system would jump to in the next step according to the transition probabilities. We set the time step as 15 ns and generated a total of 300 independent 10.5-ms MC trajectories, each of which was long enough to ensure that all microstates reached equilibrium. Thus, the MFPT and equilibrium populations could be readily calculated, with the mean values averaged over the 300 MC trajectories, and the corresponding standard deviations were then obtained.

Unbiased MD simulation setup for UDP-Glc bound ApNGT^{Q469A} and wild-type ApNGT complexes

The energy-minimized UDP-Glc bound ApNGT^{Q469A} and wild-type ApNGT complexes were used as the starting structures for unbiased MD simulations. The parameters of the protein and UDP-Glc were the same as described above. Each system was immersed in a TIP3P water box with a minimum distance of 10 Å between the solutes and the edge of the box, and Na^+ ions were added to neutralize the system. MD simulations were conducted with the GROMACS 2019.4 software package.[13] After energy minimization, a 100-ps NVT equilibration MD simulation and a 500-ps NPT equilibration MD simulation were carried out in succession. Finally, three parallel 200-ns unbiased MD simulations were performed. All the parameters used for the simulations were the same as described above. For each 200-ns MD trajectory, the last

100-ns simulation dataset was used for final analyses.

Analyses of MD simulations

Structural analyses, including calculations of distance, root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), radius of gyration (Rg), hydrogen bond (HB), and contact frequency were conducted by the internal tools in GROMACS.[13, 27] Substrate pocket volumes were calculated by POVME 3.0 with default settings in the ligand-defined inclusion region mode.[28]

HB was determined with default parameters, i.e., the acceptor–donor distance was less than 3.5 Å and the hydrogen–donor–acceptor angle was less than 30°. Detailed definition of each HB in Fig. 3a is as follows: the HB between the S496 side chain and the carbonyl oxygen atom of uracil base (denoted as H1); the S496 main chain and the amine group of uracil base (H2); the S496 main chain and the carbonyl oxygen atom of uracil base (H3); the D525 side chain and the ribose ring (H4); the T438 side chain and the $P_2O_5^{2-}$ moiety (H5); the N471 side chain and the glucose moiety (H6); the H272 side chain and the $P_2O_5^{2-}$ moiety (H7); the H277 side chain and the $P_2O_5^{2-}$ moiety (H8); the S278 side chain and the $P_2O_5^{2-}$ moiety (H9); the K441 side chain and the $P_2O_5^{2-}$ moiety (H10); the N521 main chain and the $P_2O_5^{2-}$ moiety (H11); the N521 side chain and the $P_2O_5^{2-}$ moiety (H12); the S278 side chain and the glucose moiety (H13); the T520 side chain and the glucose moiety (H14).

Calculation of configurational entropy

Schlitter's formula is used for the configurational entropy calculation,[29] which derives an upper limit S_{Schl} to the true entropy S_{true}:

$$S_{true} < S_{Schl} = \frac{k}{2} \ln \det \left[\mathbf{1} + \frac{kTe^2}{\hbar^2}\, \mathbf{M\sigma}\right]$$

where k is Boltzmann's constant, T is the temperature, e is Euler's number, $\hbar$ is Planck's constant divided by 2π, **M** is the mass matrix with the masses of the atoms on the diagonal and all off-diagonal elements equal to zero, **1** is the unit matrix, and **σ** is the covariance matrix of atom-positional fluctuations.

We used the internal tools in GROMACS to analyze configurational entropy.[13] Firstly, *gmx covar* was applied to obtain the covariance matrix **σ**. Then, we employed *gmx anaeig* to calculate the configurational entropy at 310 K. For each metastable state, the mean value was averaged over all the microstates belonging to it, and the corresponding standard deviation was calculated.

Multiple sequence alignment

Multiple sequence alignment of ApNGT (ABN74719.1), AaNGT (ALC78880.1), BtNGT (WP_015433617.1), KkNGT (CRZ19328.1), MhNGT (AJE07135.1), HdNGT (AAP96624.1), HiNGT (ADO96126.1), EcNGT (CBL93373.1), YeNGT (CAL13257.1) and YpNGT (CAL21840.1) was conducted with MEGA-X and then analyzed by ESPript 3.0 (https://espript.ibcp.fr/ESPript/ESPript/).[30, 31]

Chemicals

All chemical reagents were purchased from commercial suppliers, and used without further purification unless otherwise stated. UDP-Glc was purchased from Sigma (St. Louis, MO, USA). Peptides were synthesized by GL Biochem (Shanghai, China). NahK, AGX1 and PPA were expressed as reported.[32-34] Analytical TLC was visualized under UV light (254 nm and 365 nm), or staining the plates with a solution

of H_2SO_4 in MeOH (10%, v/v). Analytical RP-HPLC was performed using a Shimadzu LC-2010A system with a PDA detector. A reverse phase C18 column (250 × 4.6 mm I.D., 5 μm, GL Sciences Inc., Tokyo, Japan) was used with linear gradients according to the polarity of the peptides used at a flow rate of 1 mL/min (mobile phase A: 5% ACN, 95% H_2O, 0.1% TFA; mobile phase B: 95% ACN, 5% H_2O, 0.1% TFA). ESI-MS analysis was performed on LTQ XL™ Linear Ion Trap Mass Spectrometer (Thermo Fisher Scientific). ^{1}H NMR were recorded at 25 ºC with Bruker AVANCE III (400 MHz) instruments (Bruker, Germany). NMR data were processed with Mnova software.

Site-directed mutagenesis

The mutants were generated by PCR using KOD-plus DNA polymerase (Toyobo, Osaka) with the pET-24b-ApNGT^{Q469A} template as described previously[1]. The PCR primers were synthesized (Sangon Biotech, Shanghai) and listed in Table S1. All mutations were confirmed by DNA sequencing.

The plasmids containing ApNGT mutants were introduced into *E. coli* BL21(DE3) and overexpressed in the same condition that was used previously.[1] The purity of ApNGT mutants was verified by SDS-PAGE and the concentration of each enzyme was measured by the BCA Protein Assay Kit (Sangon Biotech, Shanghai) with bovine serum albumin (BSA) as a protein standard.

Determination of glycosylation activity for ApNGT mutants

The glycosylation activity of ApNGT mutants against UDP-Glc was measured at 37 °C for 30 min in a reaction mixture containing Tris buffer (10 μL, 50 mM, pH 8.0), UDP-Glc (10 mM), peptide GGNWTT or GGNWST (1 mM), and the enzyme (1.25

μM). The reaction was then quenched by adding HCl (10 μL, 0.1 M) and analyzed by analytic RP-HPLC (Shimadzu LC-2010A, Kyoto, Japan) with a PDA detector. The peptide/glycopeptide was detected by UV absorbance at 220 nm. The reaction yield was determined by the change of the peak area.

The glycosylation activity of the enzyme against UDP-GlcN was measured at 37 °C for 12 h in a reaction mixture containing Tris buffer (10 μL, 50 mM, pH 8.0), UDP-GlcN (20 mM), peptide GGNWTT (1 mM), and the enzyme (25 μM). The reaction was then quenched by adding HCl (10 μL, 0.1 M) and analyzed by RP-HPLC. The peptide/glycopeptide was detected by UV absorbance at 220 nm.

Synthesis of UDP-GlcN

UDP-GlcN was synthesis following the published procedures.[35] In brief, glucosamine HCl (510 mg) and Na_2CO_3 (500 mg) were dissolved in dry MeOH (5 mL). Ethyl trifluoroacetate (0.67 mL) was then added. After being stirred overnight at room temperature, the residue was purified by flash column chromatography (EA : MeOH = 8:1, by volume) to afford GlcNHTFA (620 mg, 95%).

GlcNHTFA (100 mg, 1.0 eq.), ATP (1.5 eq.), and UTP (1.5 eq.) were dissolved in Tris-HCl buffer (18 mL, 100 mM, pH 8.0) containing $MgCl_2$ (20 mM). After the addition of appropriate amount of NahK, AGX1 and PPA, the reaction was carried out by incubating the solution at 37 °C for 8 h. The reaction was monitored by TLC (n-Butanol : NH_4OH : H_2O = 4:3:1, by volume). Once finished, the reaction was stopped by adding the same volume of ice-cold ethanol. The mixture was than purified by a BioGel P-2 gel filtration column to obtain UDP-GlcNTFA (220 mg, 92%).

UDP-GlcNHTFA (220 mg) was dissolved in 10 mL NaOH (aq., pH 10.0). The reaction was stirred at room temperature and monitored by TLC (n-Butanol : NH_4OH : H_2O = 4:3:1, by volume). Once UDP-GlcNHTFA was cleanly hydrolyzed, the reaction mixture was adjusted to pH 7.0 and purified by a BioGel P-2 gel filtration column to afford UDP-GlcN (180 mg, 96%).

^1H NMR (400 MHz, Deuterium Oxide) of UDP-GlcN: δ = 7.95 (d, J = 8.1 Hz, 1H), 5.99 – 5.95 (m, 2H), 5.80 (dd, J = 6.8, 3.4 Hz, 1H), 4.40 – 4.34 (m, 2H), 4.32 – 4.21 (m, 3H), 3.96 – 3.79 (m, 4H), 3.55 (t, J = 9.6 Hz, 1H), 3.29 (dt, J = 10.5, 3.1 Hz, 1H); ESI-MS: *m/z* calcd for $C_{15}H_{25}N_3O_{16}P_2$ (M-H) 564.07, found 564.26.

Supporting Tables

Table S1. Primers for site-directed mutagenesis.

Primer	Sequence (5’ to 3’)
F39A-F	TAGCAATgcaGGTGGCATTCATGAAATCGAGT
F39A-R	TGCCACCtgcATTGCTATCCAGCTGGCTCAGA
H272A-F	CTGCTGGAAgcaTTTCATAGCGCGCACTCTATCTA
H272A-R	TGAAAtgcTTCCAGCAGAACGACCATAACCGG
H274A-F	ACATTTTgcaAGCGCGCACTCTATCTACCGTA
H274A-R	GCGCGCTtgcAAAATGTTCCAGCAGAACGACC
H277A-F	TAGCGCGgcaTCTATCTACCGTACTCACTCTACCAGC
H277A-R	AGATAGAtgcCGCGCTATGAAAATGTTCCAGC
M349A-F	CATCGGTgcaGATATGACCACCATCTTCGCGT
M349-R	TCATATCtgcACCGATGCTCGGCATGTAGAAA
N471A-F	TCCgcaGGTATCACTCACCCTTACGTTGAACG
N471A-R	TGAGTGATACCtgcGGACGCGCCCAGTGCGAA
H495A-F	gcaTCTCCGTACCACCAGTACCTGCGTATTCT
H495A-R	TGGTGGTACGGAGAtgcCGGGTGCGCGGTAGCGCT
S496A-F	gcaCCGTACCACCAGTACCTGCGTATTCTGCA
S496A-R	TACTGGTGGTACGGtgcGTGCGGGTGCGCGGTAGC
P497A-F	CACTCTgcaTACCACCAGTACCTGCGTATTCTG
P497A-R	TGGTGGTAtgcAGAGTGCGGGTGCGCGGTAGC
Y498A-F	TCTCCGgcaCACCAGTACCTGCGTATTCTGCA
Y498A-R	TACTGGTGtgcCGGAGAGTGCGGGTGCGCGGT

Supporting Figures

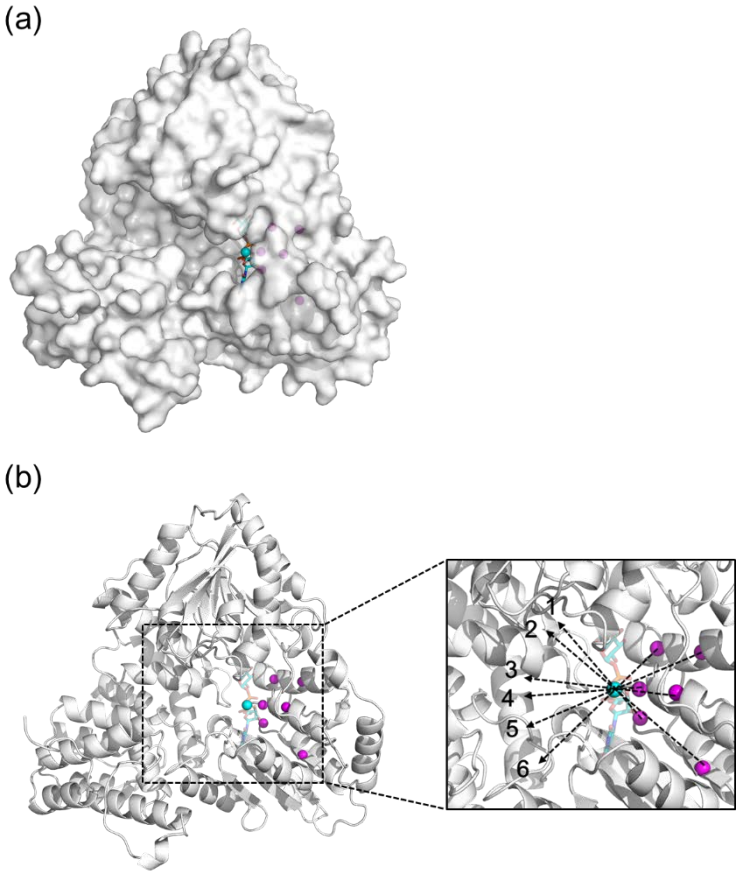


Fig. S1 (a) Surface representation of the modeled UDP-Glc-ApNGT^{Q469A} complex. (b) Six pulling directions (directions 1−6) used in SMD simulations. Accordingly, six reference groups (magenta spheres) are defined, including the center coordinates of N513−P516 C_α atoms, M510−V512 C_α atoms, N521−I524 C_α atoms, F517−N519 C_α atoms, D546−G548 C_α atoms and H542−I545 C_α atoms, respectively. Each pulling direction derives from each reference group and points to UDP-Glc (cyan sphere).

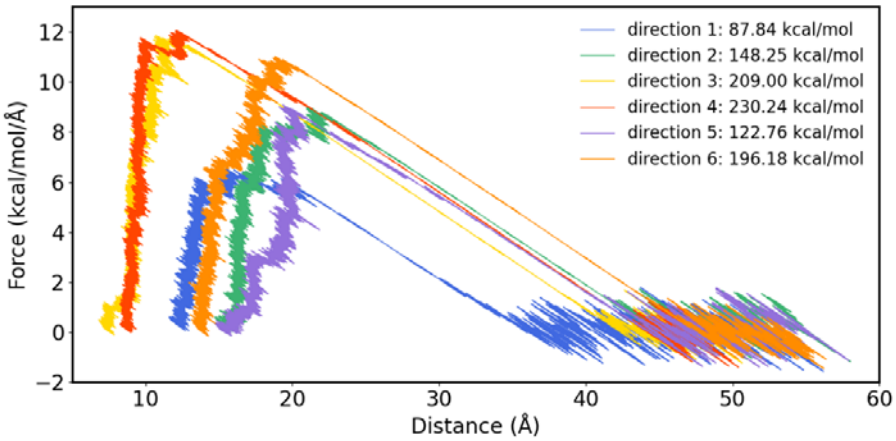


Fig. S2 Exerted pulling force against distance during SMD simulations. The energy along each pulling direction is calculated.

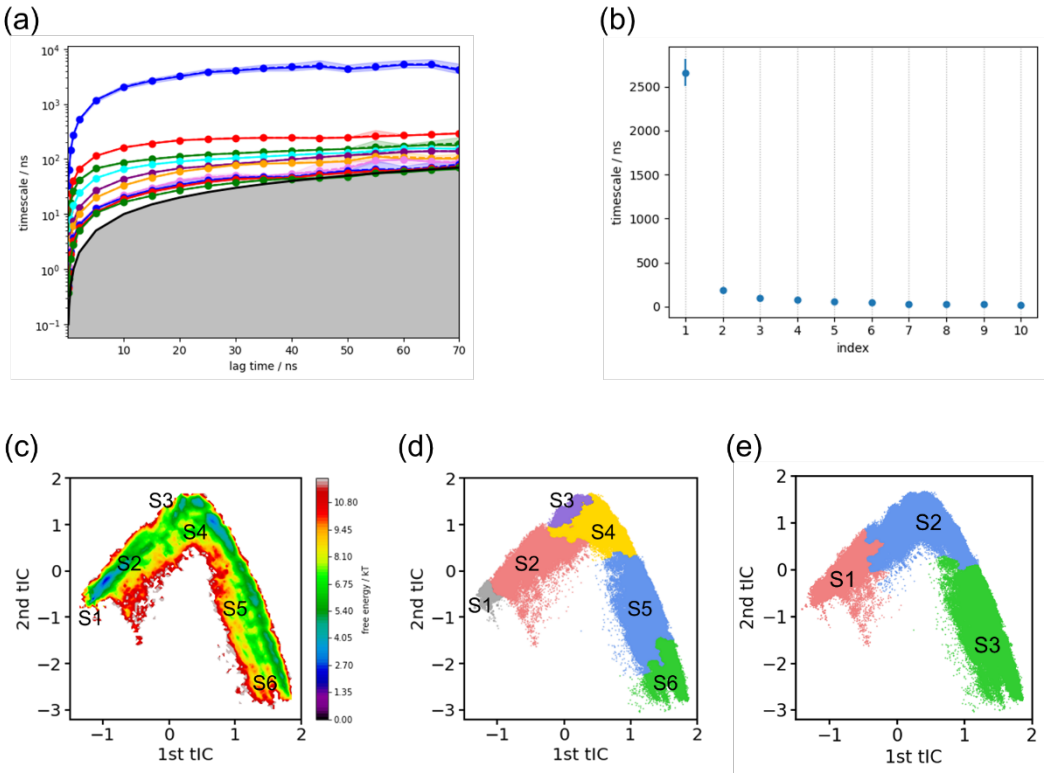


Fig. S3 (a) Implied timescales as a function of the MSM lag time τ indicate a Markovian lag time of 15 ns, and this lag time is chosen for subsequent MSM estimation. Shown as shaded regions are 95% confidence intervals. (b) Spectrum of implied timescales at τ = 15 ns. (c) Free energy projection of all MD conformations onto the top two tICs. (d) Scatter plot for the six-metastable-state model. (e) Scatter plot for the three-metastable-state model.

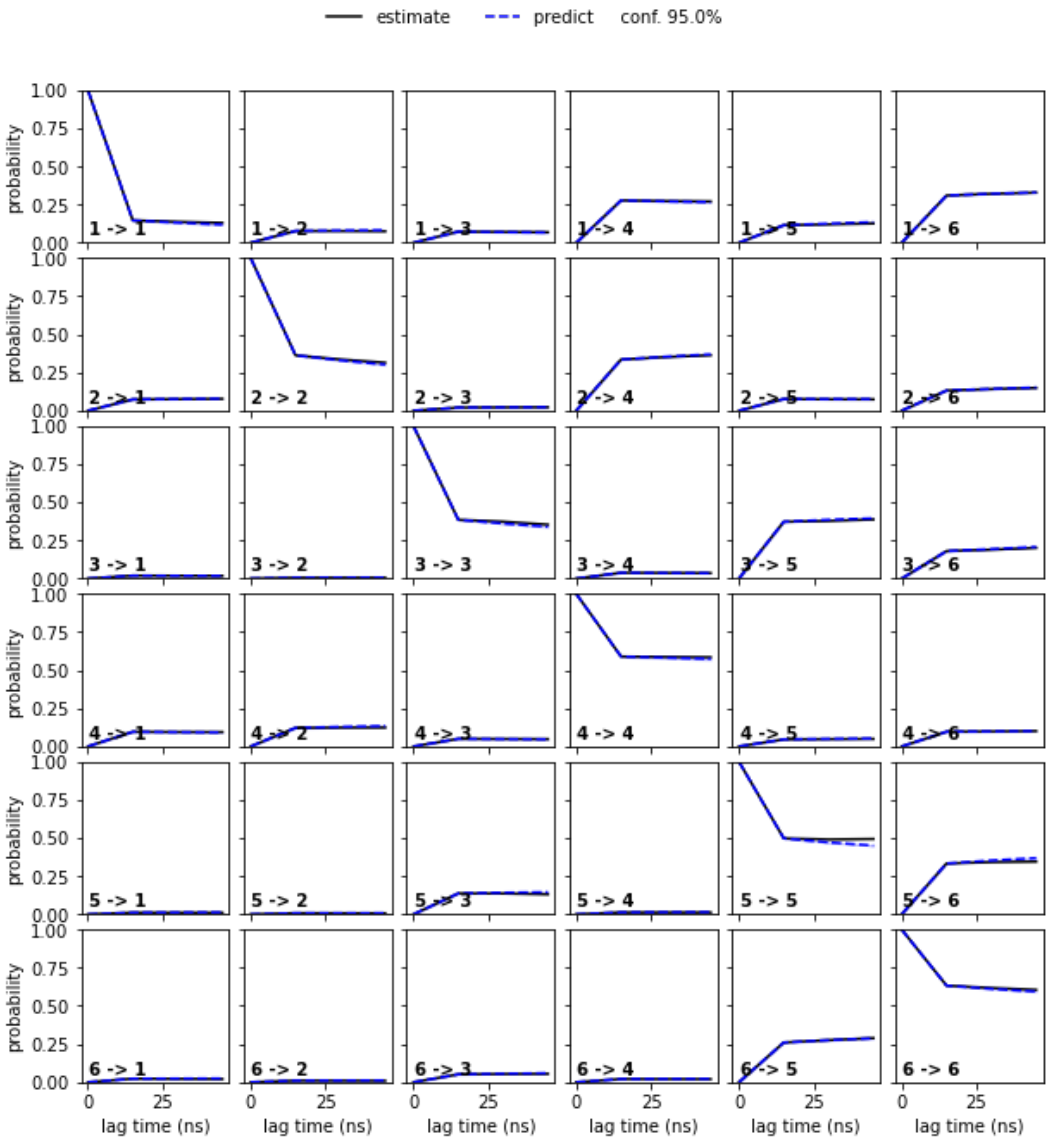


Fig. S4 Chapman-Kolmogorov test for the six-metastable-state MSM, conducted using a lag time of 15 ns.

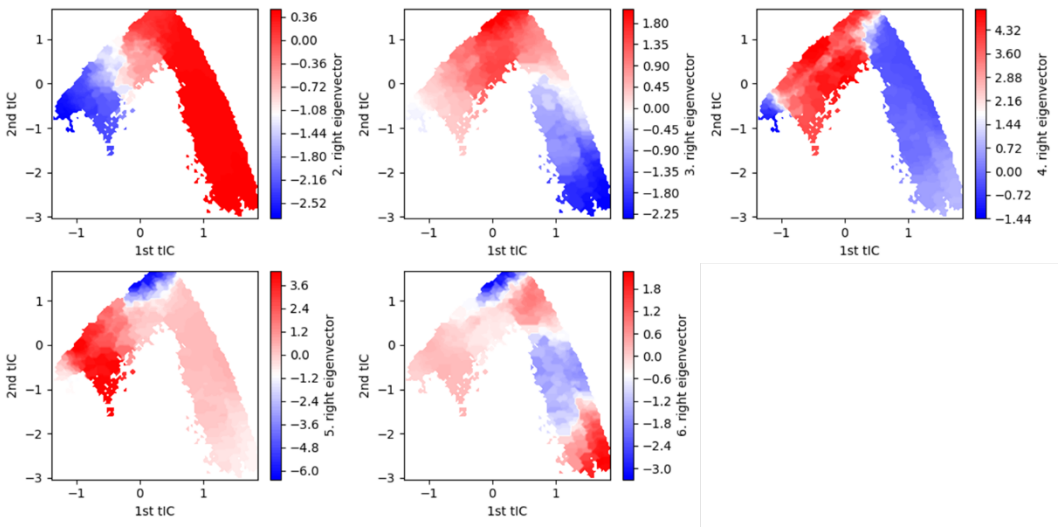


Fig. S5 Analysis of the slowest processes by inspecting the values of the top 2–6 eigenvectors projected onto the top two tICs.

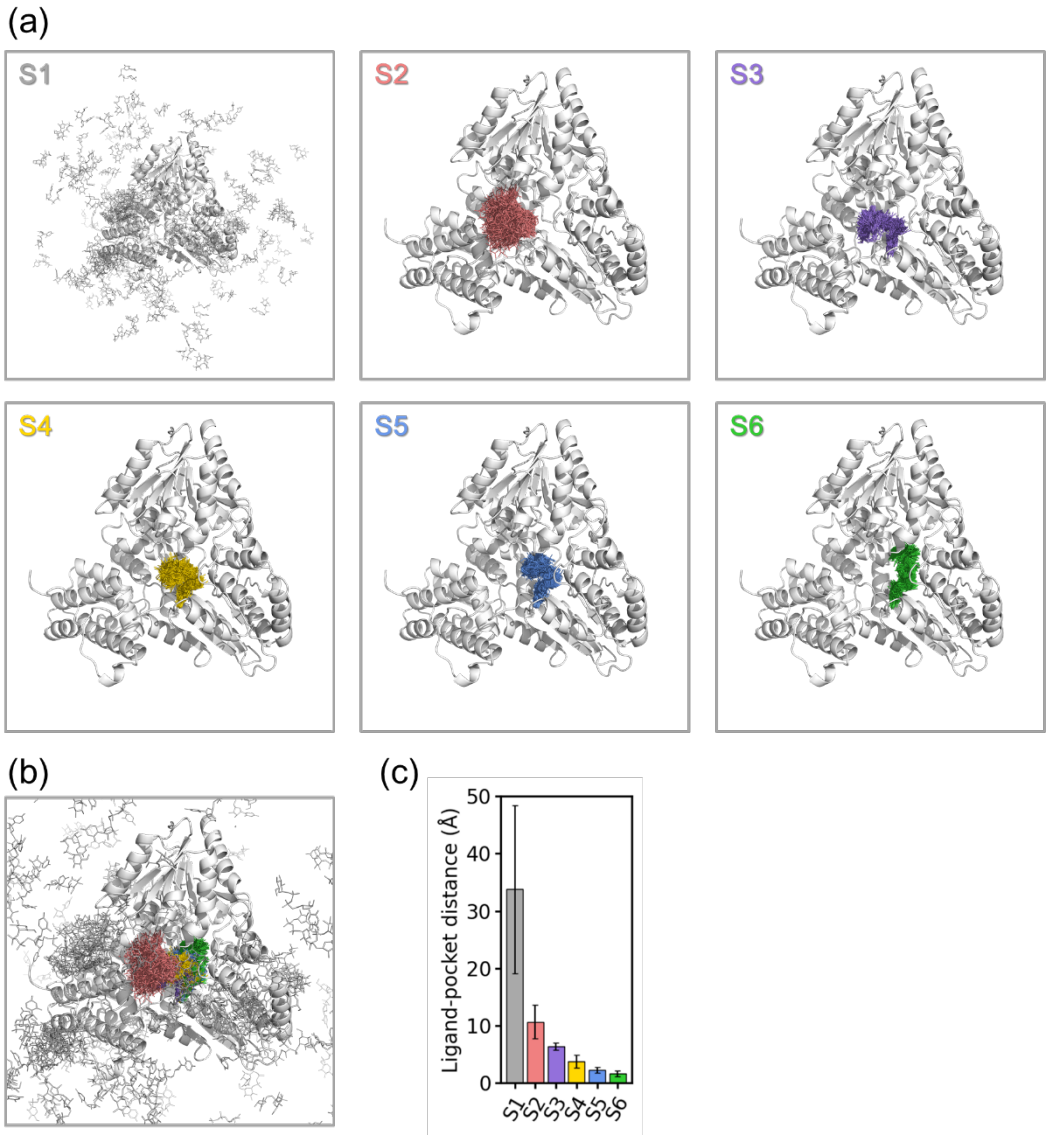


Fig. S6 Mapping 200 randomly chosen UDP-Glc conformations from the most populated microstate of each metastable state onto the donor-substrate loading pathway in ApNGT^{Q469A} with a separate (a) or overall (b) view. (c) The center of mass distance between the substrate binding pocket (the UDP-Glc bound in the energy-minimized ApNGT^{Q469A} binary complex) and the UDP-Glc conformation from each MD snapshot for each metastable state.

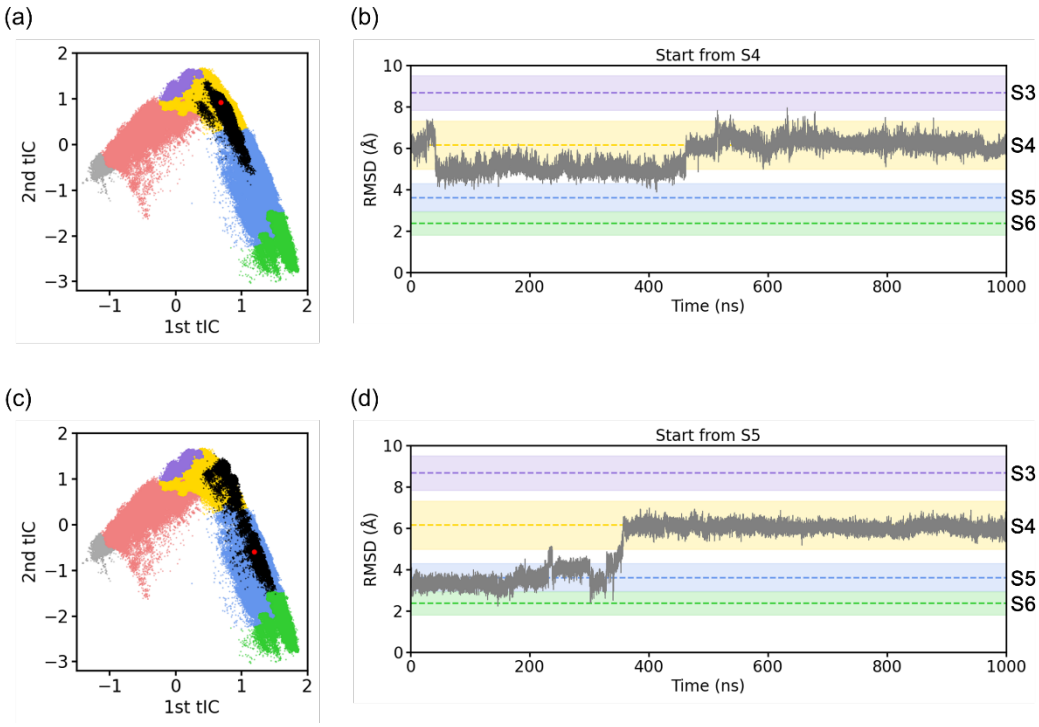


Fig. S7 (a) Projection plot of the 1-μs MD trajectory initiated from S4 onto the top two tICs (indicated as black dots). The conformations of S1–S6 are also colored as the background. The red dot indicates the starting conformation of the MD simulation. (b) RMSD of UDP-Glc along the simulation time for the 1-μs MD trajectory starting from S4. (c–d) Similar analyses with (a–b) but for the simulation starting from S5.

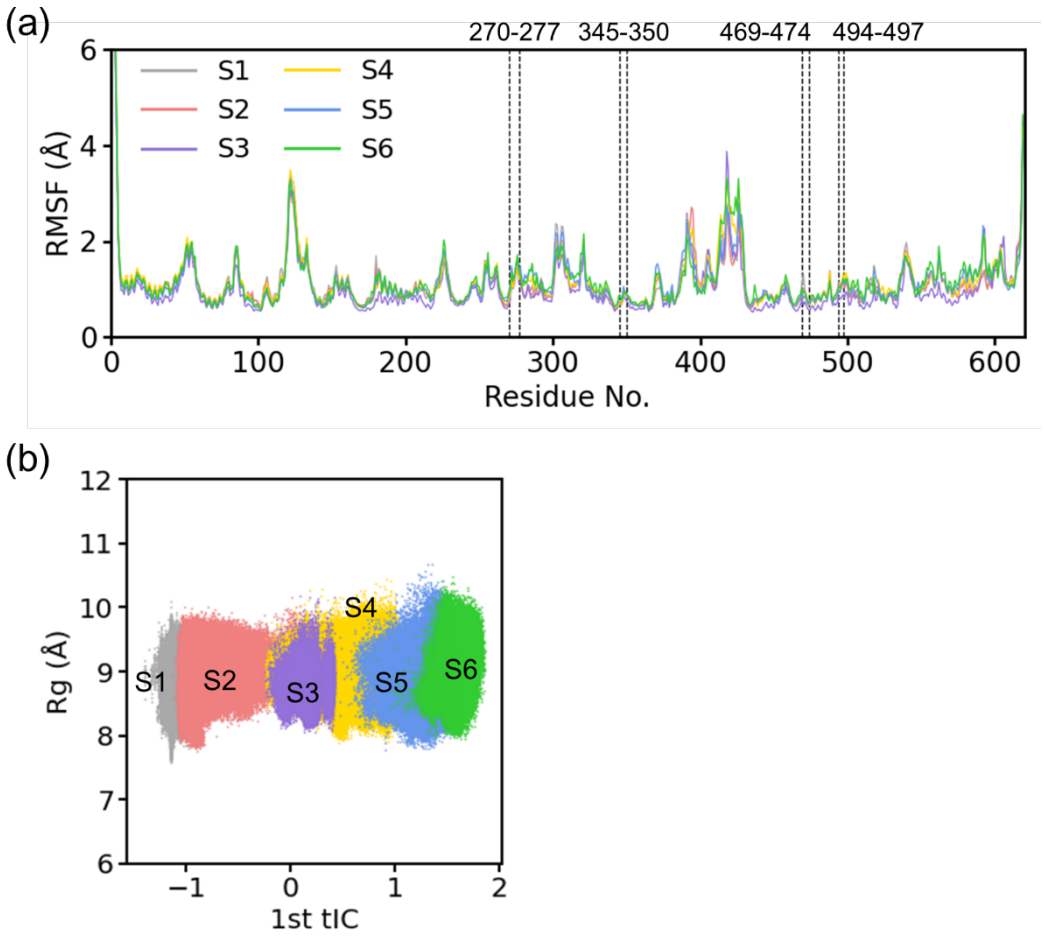
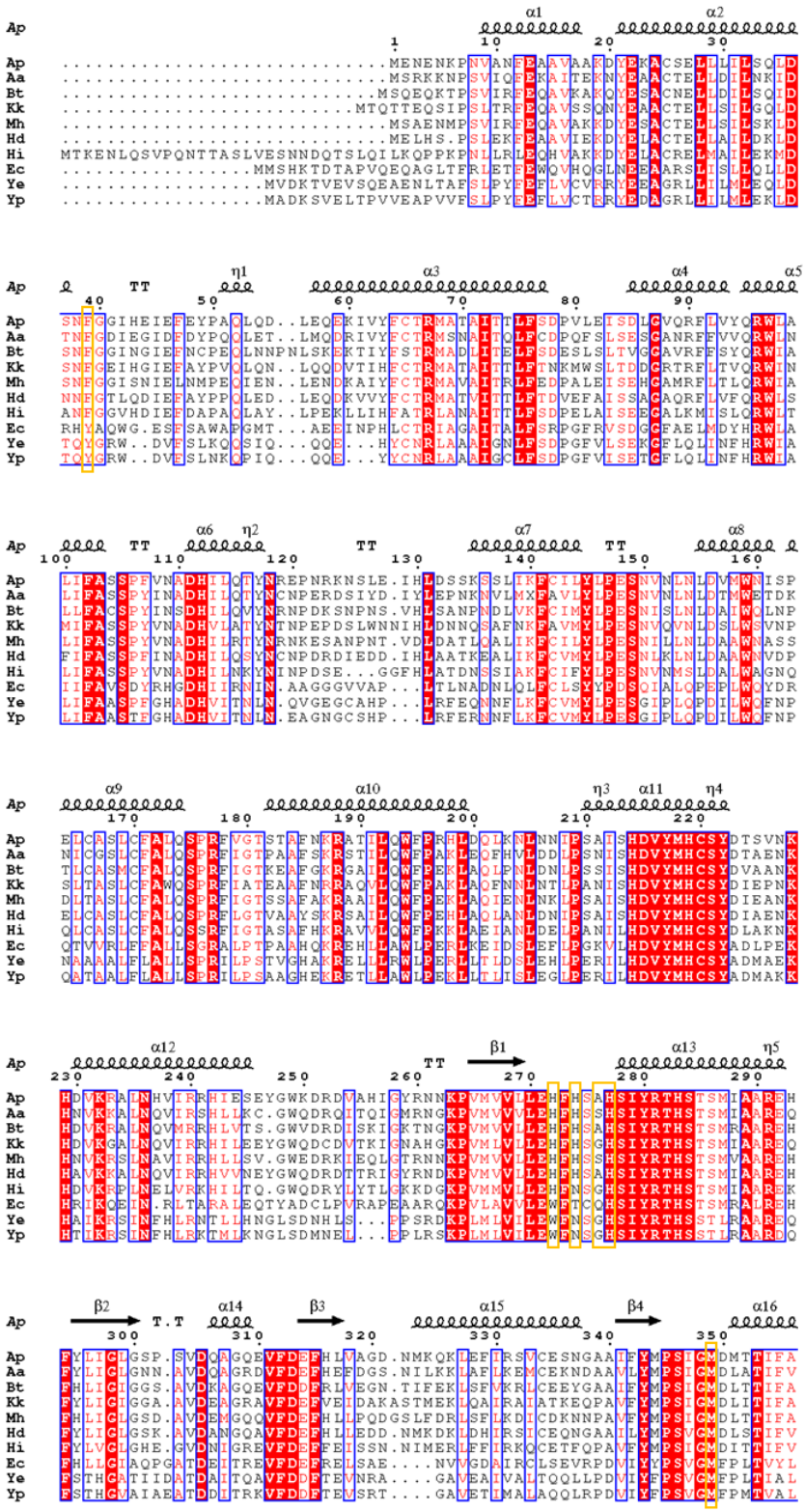


Fig. S8 (a) Plots of RMSF for each metastable state. (b) Projection of each MD conformation onto the first tIC and the Rg of the residues mentioned in Fig. 5a.



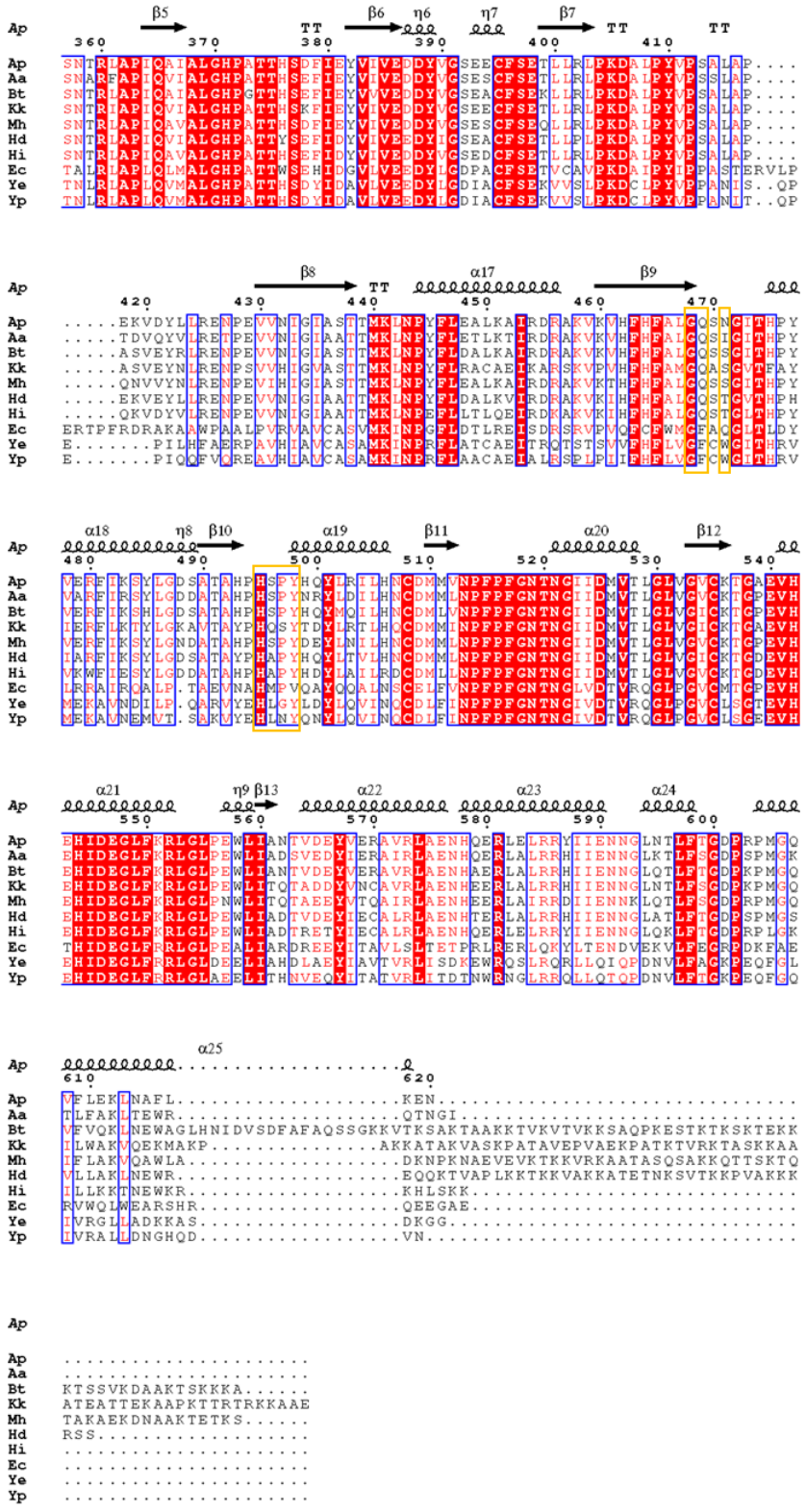


Fig. S9 Multiple sequence alignment for active NGTs. The gating residues are highlighted with yellow frames.

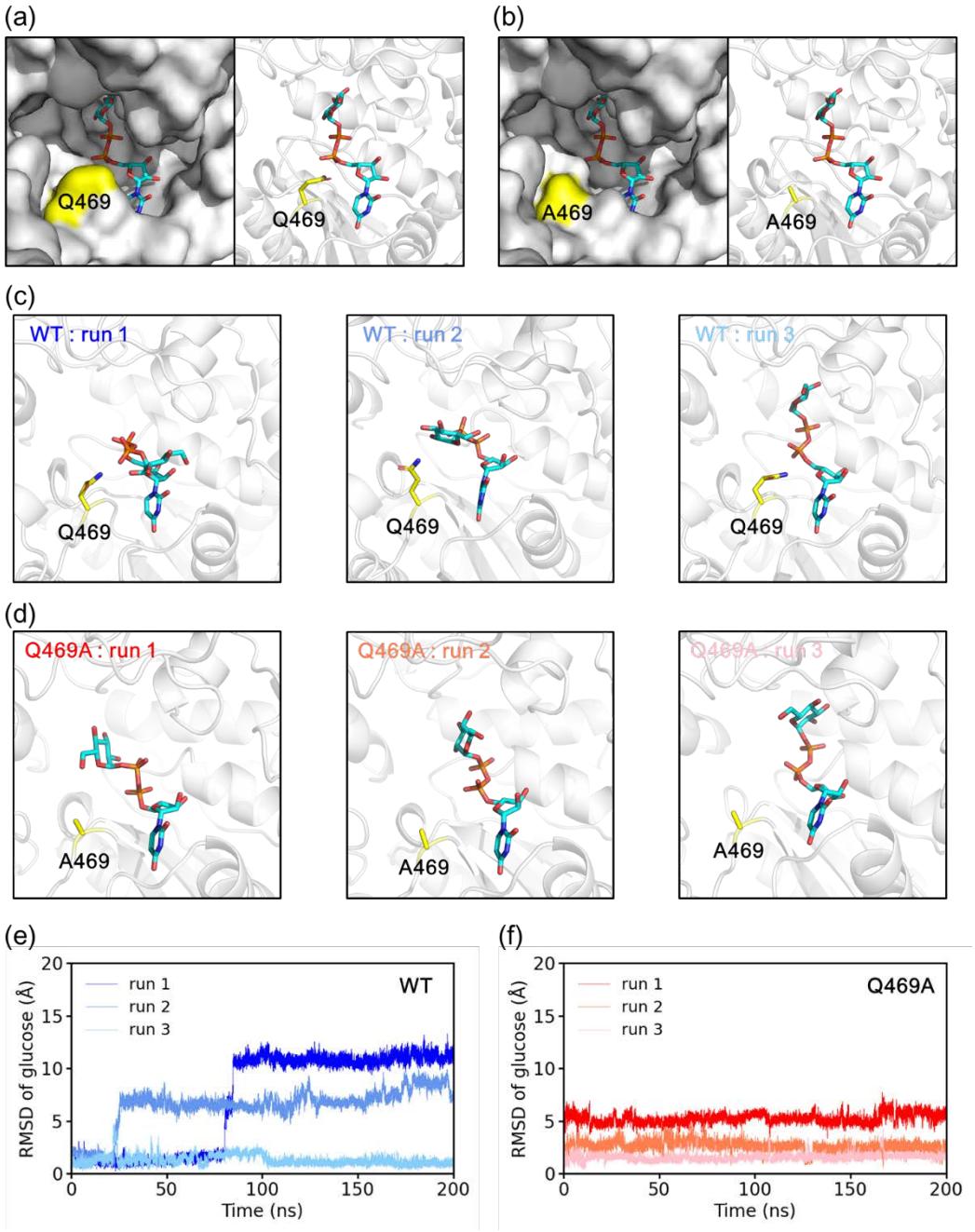


Fig. S10 Comparison of wild-type ApNGT (a) and ApNGT^{Q469A} (b) complexes. Left and right panels show surface and cartoon representations, respectively. Last snapshots of 200-ns MD simulations for wild-type ApNGT (c) and ApNGT^{Q469A} (d) complexes. RMSD plots of the glucose moiety for wild-type ApNGT (e) and ApNGT^{Q469A} (f) complexes, respectively.

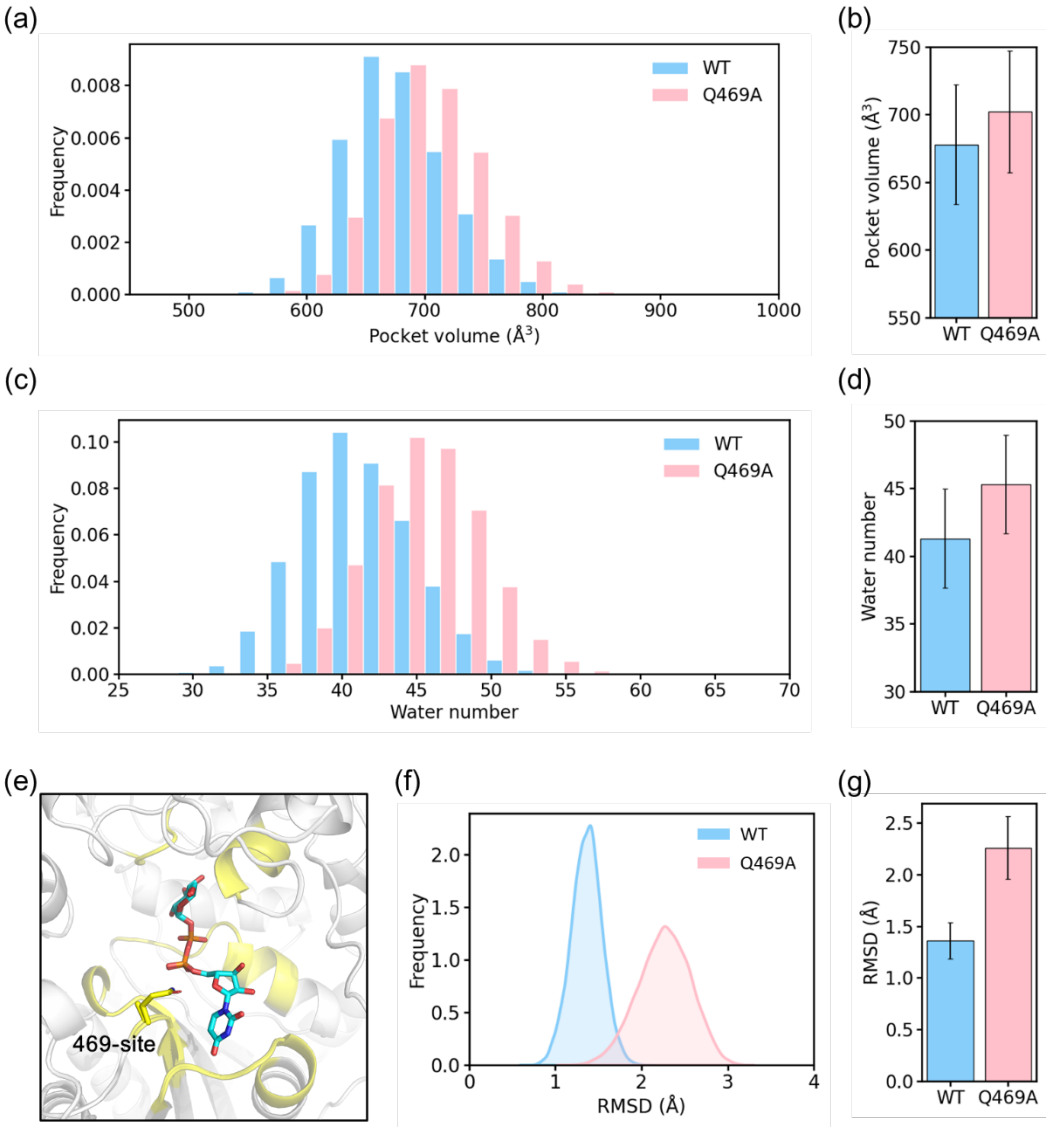


Fig. S11 The Q469A mutation results in structural changes of the active site. (a) Distributions of pocket volumes calculated by POVME 3.0. (b) Average pocket volumes. (c) Distributions of water molecules within 10 Å from the center of mass of UDP-Glc. (d) Average water numbers around the active site. (e) Active-site residues for RMSD calculation in (f) and (g), including the residues 278−282, 369−372, 437−441, 466−469, 494−501, and 517−525 (highlighted in yellow). (f) Distributions of RMSD values of the active-site residues. (g) Average RMSD values of the active-site residues.

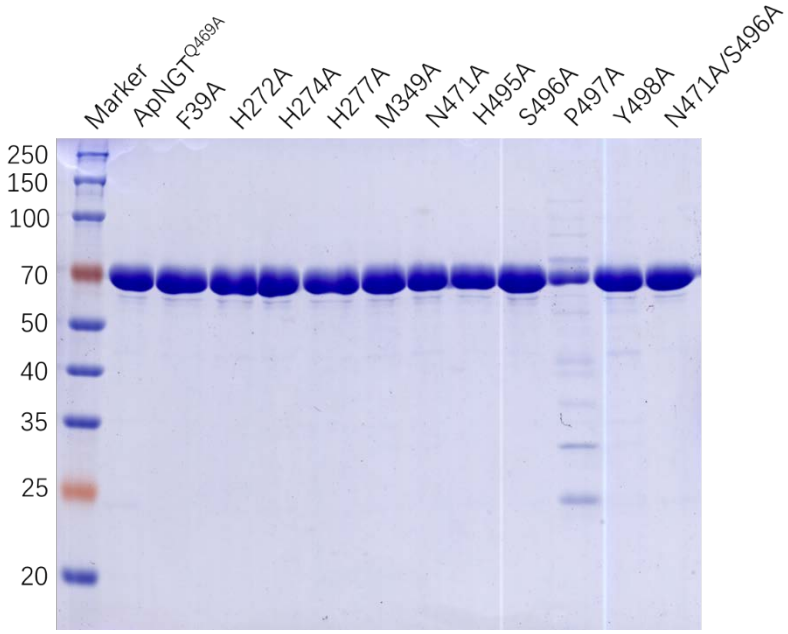


Fig. S12 SDS-PAGE analysis of purified ApNGT^{Q469A} and its mutants.

References

1. Hao, Z.; Guo, Q.; Feng, Y.; Zhang, Z.; Li, T.; Tian, Z.; Zheng, J.; Da, L.-T.; Peng, W. Investigation of the catalytic mechanism of a soluble N-glycosyltransferase allows synthesis of N-glycans at noncanonical sequons. *JACS Au* **2023,** *3* (8), 2144-2155.

2. Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2013,** *3* (2), 198-210.

3. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005,** *26* (16), 1668-1688.

4. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015,** *11* (8), 3696-3713.

5. Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph.* **2006,** *25* (2), 247-260.

6. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.;

Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Gaussian Inc., Wallingford CT: 2009.

7. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25* (9), 1157-1174.

8. Wu, X.; Brooks, B. R.; Vanden-Eijnden, E. Self-guided Langevin dynamics via generalized Langevin equation. *J. Comput. Chem.* **2016,** *37* (6), 595-601.

9. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977,** *23* (3), 327-341.

10. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995,** *103* (19), 8577-8593.

11. Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013,** *139* (1), 015102.

12. Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: statistical models for biomolecular dynamics. *Biophys. J.* **2017,** *112* (1), 10-15.

13. Van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **2005,** *26* (16), 1701-1718.

14. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997,** *18* (12), 1463-1472.

15. Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007,** *126* (1), 014101.

16. Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. v.; Dinola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984,** *81* (8), 3684-3690.

17. Husic, B. E.; Pande, V. S. Markov state models: from an art to a science. *J. Am. Chem. Soc.* **2018,** *140* (7), 2386-2396.

18. Wang, X.; Unarta, I. C.; Cheung, P. P.-H.; Huang, X. Elucidating molecular mechanisms of functional conformational changes of proteins via Markov state models. *Curr. Opin. Struct. Biol.* **2021,** *67*, 69-77.

19. Konovalov, K. A.; Unarta, I. C.; Cao, S.; Goonetilleke, E. C.; Huang, X. Markov state models to study the functional dynamics of proteins in the wake of machine learning. *JACS Au* **2021,** *1* (9), 1330-1341.

20. Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **2015,** *11* (11), 5525-5542.

21. Röblitz, S.; Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv. Data Anal. Classif.* **2013,** *7* (2), 147-179.

22. E, W.; Vanden-Eijnden, E. Towards a theory of transition paths. *J. Stat. Phys.* **2006,** *123* (3), 503-523.

23. Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106* (45), 19011-19016.

24. Wang, W.; Cao, S.; Zhu, L.; Huang, X. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2018,** *8* (1), e1343.

25. Bowman, G. R.; Meng, L.; Huang, X. Quantitative comparison of alternative methods for coarse-graining biological networks. *J. Chem. Phys.* **2013,** *139* (12), 121905.

26. Silva, D.-A.; Weiss, D. R.; Pardo Avila, F.; Da, L.-T.; Levitt, M.; Wang, D.; Huang, X. Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (21), 7665-7670.

27. Blau, C.; Grubmuller, H. g_contacts: fast contact search in bio-molecular ensemble data. *Comput. Phys. Commun.* **2013,** *184* (12), 2856-2859.

28. Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E. POVME 3.0: software for mapping binding pocket flexibility. *J. Chem. Theory Comput.* **2017,** *13* (9), 4584-4592.

29. Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **1993,** *215* (6), 617-621.

30. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018,** *35* (6), 1547-1549.

31. Robert, X.; Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **2014,** *42* (W1), W320-W324.

32. Li, Y.; Yu, H.; Chen, Y.; Lau, K.; Cai, L.; Cao, H.; Tiwari, V. K.; Qu, J.; Thon, V.; Wang, P. G.; Chen, X. Substrate promiscuity of N-acetylhexosamine 1-kinases. *Molecules* **2011,** *16* (8), 6396-6407.

33. Guan, W.; Cai, L.; Wang, P. G. Highly efficient synthesis of UDP-GalNAc/GlcNAc analogues with promiscuous recombinant human UDP-GalNAc pyrophosphorylase AGX1. *Chem.-Eur. J.* **2010,** *16* (45), 13343-13345.

34. Zhou, X.; Chandarajoti, K.; Pham, T. Q.; Liu, R.; Liu, J. Expression of heparan sulfate sulfotransferases in Kluyveromyces lactis and preparation of 3′-phosphoadenosine-5′-phosphosulfate. *Glycobiology* **2011,** *21* (6), 771-780.

35. Chen, Y.; Thon, V.; Li, Y.; Yu, H.; Ding, L.; Lau, K.; Qu, J.; Hie, L.; Chen, X. One-pot three-enzyme synthesis of UDP-GlcNAc derivatives. *Chem. Commun.* **2011,** *47* (38), 10815-10817.