SUPPLEMENTARY MATERIALS FILE 1

Systematic Functional and Computational Analysis of Sugar-Binding Residues in *Thermoanaerobacterium xylanolyticum* Family GH116 β-Glucosidase

Meng Huang[1,2], Salila Pengthaisong[1,2], Ratana Charoenwattanasatien[3], Natechanok Thinkumrob[4], Jitrayut Jitonnom[4]*, James R. Ketudat Cairns[1,2]*

[1]School of Chemistry, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

[2]Center for Biomolecular Structure, Function and Application, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

[3]Research Facility Department, Synchrotron Light Research Institute (Public Organization), Nakhon Ratchasima 30000, Thailand

[4]Unit of Excellence in Computational Molecular Science and Catalysis, and Division of Chemistry, School of Science, University of Phayao, Phayao 56000, Thailand

*Corresponding authors

Table S1-1. Proteins from GH116 family used for sequence alignment

Number	Uniprot ID	Species	Protein name
1	Q69ZF3	*Mus musculus*	GBA2 MOUSE Non-lysosomal glucosylceramidase
2	Q7KT91	*Drosophila melanogaster*	DROME Non-lysosomal glucosylceramidase
3	A0A3Q7ERA4	*Solanum lycopersicum*	SOLLC Non-lysosomal glucosylceramidase
4	A0A4X3NVV9	*Pristionchus pacificus*	PRIPA Non-lysosomal glucosylceramidase
5	A7SX93	*Nematostella vectensis*	NEMVE Predicted protein (Fragment)
6	A8XB73	*Caenorhabditis briggsae*	CAEBR Non-lysosomal glucosylceramidase
7	A9SXS9	*Physcomitrium patens*	PHYPA Non-lysosomal glucosylceramidase
8	B3RSC0	*Trichoplax adhaerens*	TRIAD Non-lysosomal glucosylceramidase
9	D3AW83	*Polysphondylium pallidum*	POLPP Uncharacterized protein
10	E0VQ36	*Pediculus humanus* subsp. corporis	PEDHC Non-lysosomal glucosylceramidase
11	F1NYQ3	*Gallus gallus*	CHICK Non-lysosomal glucosylceramidase
12	F6U9S8	*Ciona intestinalis*	CIOIN Non-lysosomal glucosylceramidase
13	H2LUF8	*Oryzias latipes*	ORYLA Non-lysosomal glucosylceramidase
14	H3AXX8	*Latimeria chalumnae*	LATCH Non-lysosomal glucosylceramidase
15	I1I0F1	*Brachypodium distachyon*	BRADI Non-lysosomal glucosylceramidase
16	K3YG42	*Setaria italica*	SETIT Non-lysosomal glucosylceramidase
17	K3ZH71	*Setaria italica*	SETIT Non-lysosomal glucosylceramidase
18	M0T848	*Musa acuminata* subsp. malaccensis	MUSAM Non-lysosomal glucosylceramidase
19	M0ZKD8	*Solanum tuberosum*	SOLTU Non-lysosomal glucosylceramidase
20	P73619	*Synechocystis* sp.	SYNY3 Sll1775 protein
21	Q17L47	*Aedes aegypti*	AEDAE AAEL001478-PA (Fragment)
22	Q1IR11	*Koribacter versatilis* (strain Ellin345)	KORVE Uncharacterized protein
23	Q54D71	*Dictyostelium discoideum*	DICDI Uncharacterized protein
24	Q6EUT3	*Caenorhabditis elegans*	CAEEL Non-lysosomal glucosylceramidase
25	Q7NHS5	*Gloeobacter violaceus*	GLOVI Gll2460 protein
26	Q7XDG7	*Oryza sativa* subsp. japonica	ORYSJ Non-lysosomal glucosylceramidase
27	Q8A3P6	*Bacteroides thetaiotaomicron*	BACTN Uncharacterized protein

Table S1-2. Proteins from GH116 family used for sequence alignment

Number	Uniprot ID	Species	Protein name
28	Q8DMC7	*Thermosynechococcus elongatus*	THEEB Tlr0193 protein
29	Q8GUI9	*Arabidopsis thaliana*	ARATH Non-lysosomal glucosylceramidase
30	Q8LPR0	*Arabidopsis thaliana*	ARATH Non-lysosomal glucosylceramidase
31	Q97VF2	*Saccharolobus solfataricus*	SACS2 DUF608 domain-containing protein
32	Q97YG8	*Saccharolobus solfataricus*	SACS2 Uncharacterized protein
33	W1NWJ7	*Amborella trichopoda*	AMBTC Non-lysosomal glucosylceramidase
34	F6BL85	*Thermoanaerobacterium xylanolyticum*	β-Glucosidase (glucosylceramidase)

Table S2. Glycone sugar binding interactions related residues and their response mutations.

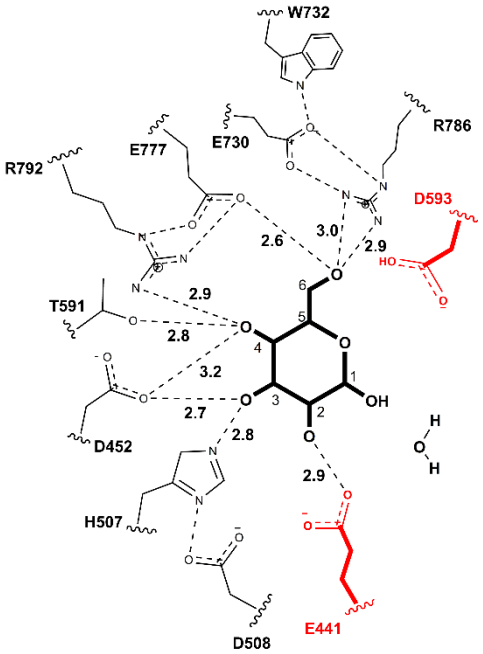
Primer name	Primer sequence
*Tx*GH116_E777Q_for	5'- CTATTGGTTTCGTACCCCG**CAG**GCCTGGACGAAAGATG -3'
*Tx*GH116_E777Q_rev	5'- CATCTTTCGTCCAGGC**CTG**CGGGGTACGAAACCAATAG -3'
*Tx*GH116_E777A_for	5'- GTTTCGTACCCCG**GCG**GCCTGGACGAAAGATG -3'
*Tx*GH116_E777A_rev	5'- CATCTTTCGTCCAGGC**CGC**CGGGGTACGAAAC -3'
*Tx*GH116_R786K_for	5'- ACGAAAGATGGCAATTAT**AAA**GCAAGTATGTATATGCGTCCGCTG -3'
*Tx*GH116_R786K_rev	5'- CGCATATACATACTTGC**TTT**ATAATTGCCATCTTTCGTCCAGGC -3'
*Tx*GH116_R786A_for	5'- CGAAAGATGGCAATTAT**GCG**GCAAGTATGTATATGCGTCCGC -3'
*Tx*GH116_R786A_rev	5'- GCATATACATACTTGC**CGC**ATAATTGCCATCTTTCGTCCAGG -3'
*Tx*GH116_W732F_for	5'- CAGGCTCAAGAAGTG**TTC**ACCGGTGTTACGTATGCAC -3'
*Tx*GH116_W732F_rev	5'- CATACGTAACACCGGT**GAA**CACTTCTTGAGCCTGAATATC -3'
*Tx*GH116_E730A_for	5'- CTGATATTCAGGCTCAA**GCG**GTGTGGACCGGTGTTAC -3'
*Tx*GH116_E730A_rev	5'- GTAACACCGGTCCACAC**CGC**TTGAGCCTGAATATCAG -3'
*Tx*GH116_E730Q_for	5'- CTGATATTCAGGCTCAA**CAG**GTGTGGACCGGTGTTAC -3'
*Tx*GH116_E730Q_rev	5'- GTAACACCGGTCCACAC**CTG**TTGAGCCTGAATATCAG -3'
*Tx*GH116_R792A_for	5'- GCAAGTATGTATATG**GCC**CCGCTGAGTATTTGGAGTATGG -3'
*Tx*GH116_R792A_rev	5'- CTCCAAATACTCAGCGG**GGC**CATATACATACTTGCGC -3'
*Tx*GH116_R792K_for	5'- CGCAAGTATGTATATG**AAA**CCGCTGAGTATTTGGAGTATGGAAG -3'
*Tx*GH116_R792K_rev	5'- CCAAATACTCAGCGG**TTT**CATATACATACTTGCGCGATAATTGC -3'
*Tx*GH116_T591A_for	5'- CATCCCGGACCAG**GCG**TACGATACGTGGTCAATG -3'
*Tx*GH116_T591A_rev	5'- GACCACGTATCGTA**CGC**CTGGTCCGGGATGCCTTC -3'
*Tx*GH116_D452A_for	5'- ATTACGAAACCCTG**GCG**GTTCGTTTTTATGGCAGCTTC -3'
*Tx*GH116_D452A_rev	5'- CTGCCATAAAAACGAAC**CGC**CAGGGTTTCGTAATAATTG -3'
*Tx*GH116_D452N_for	5'- ATTACGAAACCCTG**AAC**GTTCGTTTTTATGGCAGCTTCC -3'
*Tx*GH116_D452N_rev	5'- GCTGCCATAAAAACGAAC**GTT**CAGGGTTTCGTAATAATTG -3'
*Tx*GH116_H507A_for	5'- CAGGGTATGATTCCG**GCT**GATCTGGGCTCATCGTACG -3'
*Tx*GH116_H507A_rev	5'- GATGAGCCCAGATC**AGC**CGGAATCATACCCTGGAC -3'
*Tx*GH116_H507Q_for	5'- CAGGGTATGATTCCG**CAG**GATCTGGGCTCATCGTACG -3'
*Tx*GH116_H507Q_rev	5'- GATGAGCCCAGATC**CTG**CGGAATCATACCCTGGAC -3'
*Tx*GH116_H507E_for	5'- CAGGGTATGATTCCG**GAG**GATCTGGGCTCATCGTACG -3'
*Tx*GH116_H507E_rev	5'- GATGAGCCCAGATC**CTC**CGGAATCATACCCTGGAC -3'

All of the mutated genes will be sequenced (Macrogen) to confirm that only the desired mutations were inserted.

Table S3. Data collection and refinement statistics.

Dataset	**E730A glucose**	**E730Q glucose**	**R786A glucose**	**R786K glucose**	**R786K**
PDB code	7W2S	7W2T	7W2V	7W2W	7W2X
Data collection					
Space group	$P2_12_12$	$P2_12_12_1$	$P2_12_12$	$P2_12_12$	$P2_12_12$
Cell dimensions					
a, *b*, *c* (Å)	*a* = 177.4	*a* = 80.4	*a* = 177.4	*a* = 177.3	*a* = 176.8
	b = 54.7	*b* = 124.5	*b* = 54.2	*b* = 54.5	*b* = 54.3
	c = 83.2	*c* = 174.3	*c* = 83.1	*c* = 83.1	*c* = 83.0
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Resolution range (Å)	30–1.85	50– 2.15	30–1.80	50–1.79	50–1.80
	(1.92–1.85)	(2.23–2.15)	(1.86–1.80)	(1.85–1.79)	(1.86–1.80)
Completeness (%)	99.8 (99.5)	99.3 (96.6)	98.3 (90.4)	98.5 (86.7)	98.4 (86.9)
Redundancy	6.7 (5.6)	6.3 (4.0)	5.6 (2.6)	7.2 (5.8)	7.1 (5.7)
I/σ (I)	16.5 (1.9)	11.7 (2.2)	17.4 (2.1)	16.3 (1.9)	17.3 (2.0)
$R_{(merge)}$ (%)	11.3 (72.2)	18.0 (67.0)	10.1 (48.5)	12.1 (81.0)	11.1 (77.2)
CC1/2	(0.765)	(0.814)	(0.635)	(0.708)	(0.692)
Refinement					
Resolution range (Å)	30–1.85	50– 2.15	30–1.80	50–1.79	50–1.80
No. reflections	66151	77195	70183	70343	69631
R_{factor} (%)	15.5	15.3	15.3	15.0	14.7
R_{free} (%)	19.3	20.3	18.2	18.6	18.1
No. atom					
Protein	6273	6262/6170	6241	6299	6321
Carbohydrate	12	12/12	12	12/12	-
Hetero	63	44	53	82	72
Water	437	617	429	463	523
Mean B-factor (Å^2)					
Protein	22.1	28.5/36.8	23.0	21.7	21.4
Carbohydrate	17.3	26.1/33.7	23.3	10.7/21.8	-
Hetero	49.3	49.9	39.7	40.9	37.5
Water	32.9	36.3	31.8	32.0	33.1
R.m.s. deviations					
Bond lengths (Å)	0.014	0.013	0.013	0.012	0.011
Bond angle (°)	1.57	1.58	1.53	1.53	1.50
Ramachandran favored (%)	95.8	94.5	95.8	96.6	96.2
Ramachandran outliers (%)	0.13	0.13	0	0	0

Values in parentheses are for the outer resolution shell.



Residue	Interactions
E777	H-bond to C6OH, along with R786, interacts with R792
R786	H-bonds to C6OH, Glc2 C2OH, Y445
W732	CH-pi interactions with C6, H-bond to E730
E730	2 H-bonds to R786 guanidinyl group
R792	H-bonds to C4OH, E777
T591	H-bonds to C4OH
D452	H-bonds to C3OH, C4OH
H507	H-bonds to C3OH

Figure S1. WT *Tx*GH116 glycone binding related residues and their interactions with glucose ligand.

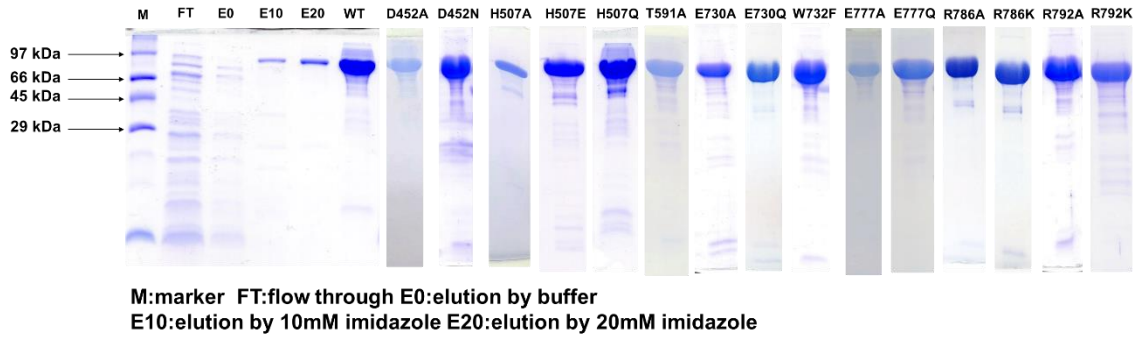
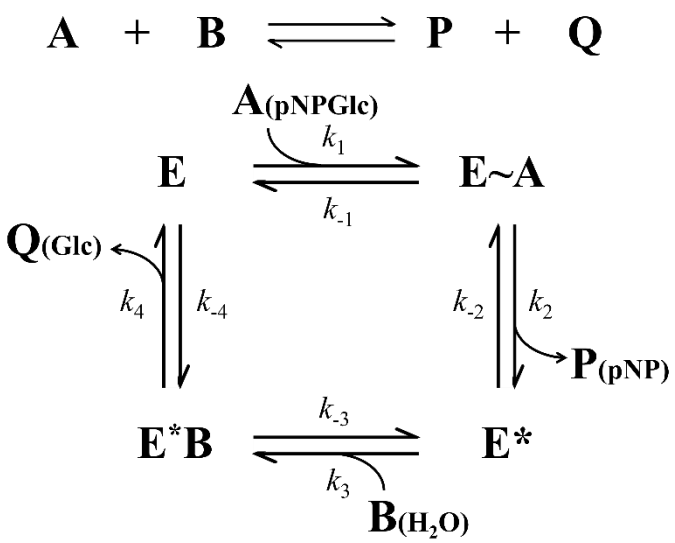


Figure S2. SDS gel after protein expression and 1st IMAC purification of all mutants. The protein concentrations are reported as an average of the biological triplicates after being measured with the spectrophotometer at an absorbance of 280 nm and converted to molar using the enzyme coefficient [1]. All protein samples were running in 12% SDS-PAGE. The wild type has a molecular weight of 96.4 kDa. This gel indicates the presence and purity of reported proteins.



$$K_M = \frac{k_4(k_{-1} + k_2)}{k_1(k_2 + k_4)} \qquad k_{cat} = k_2k_4 / (k_2 + k_4)$$

Figure S3. Schematic view of Pi**ng-Pong Bi Bi reaction of *Tx*GH116 with *p*NPGlc substrate.**

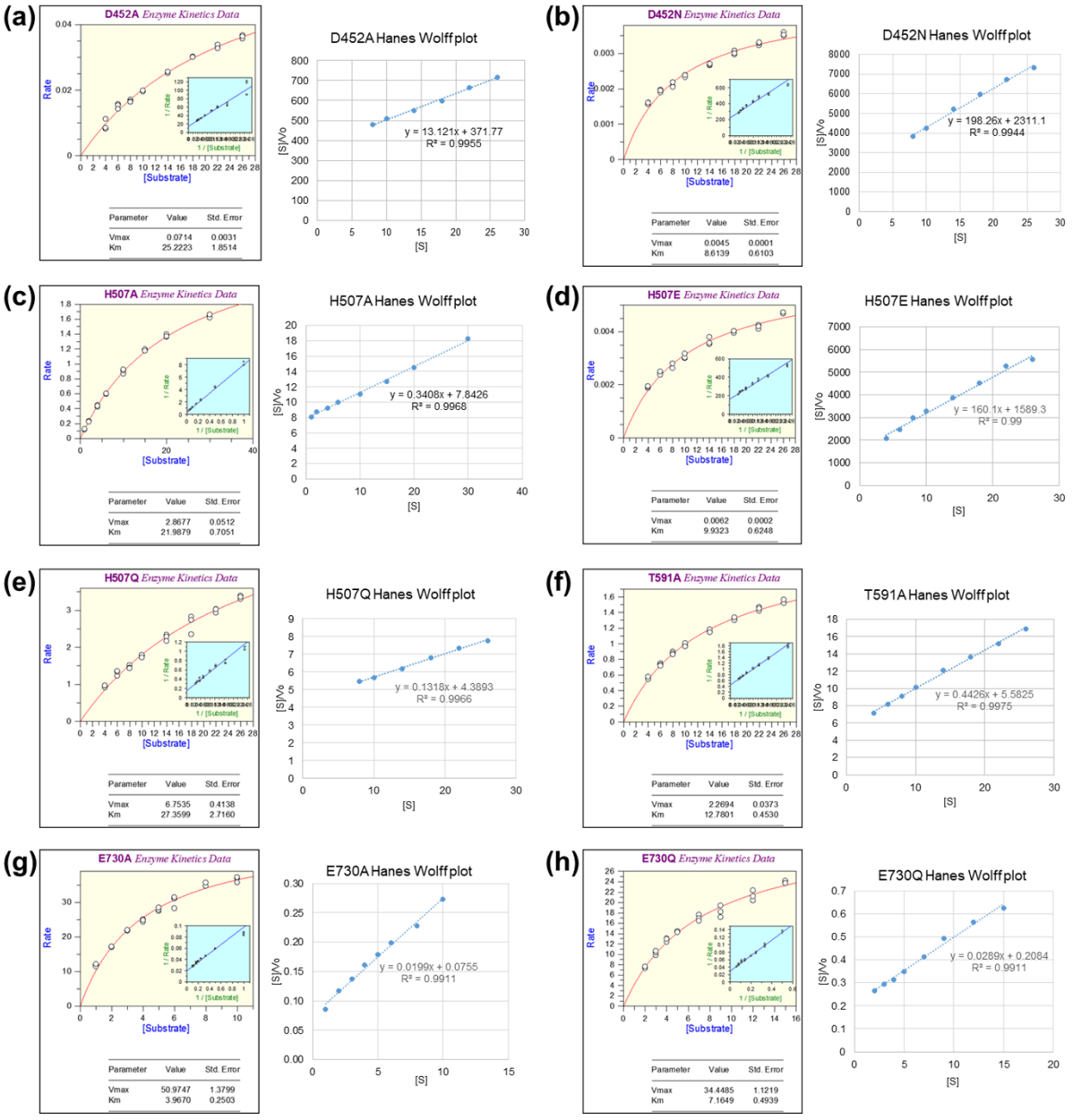


Figure S4-1. ***p*NPGlc kinetics of glycone sugar binding related residues mutants. The kinetic parameters, including k_{cat}, K_M, and k_{cat}/K_M, of purified *Tx*GH116 mutants from *E. coli* were calculated with linear (Hanes-Wolff [S]/v_0 vs. [S]) plots and checked by nonlinear regression of Michaelis-Menten plots with the Grafit 5.0 computer program (Erithacus Software, Horley, UK).**

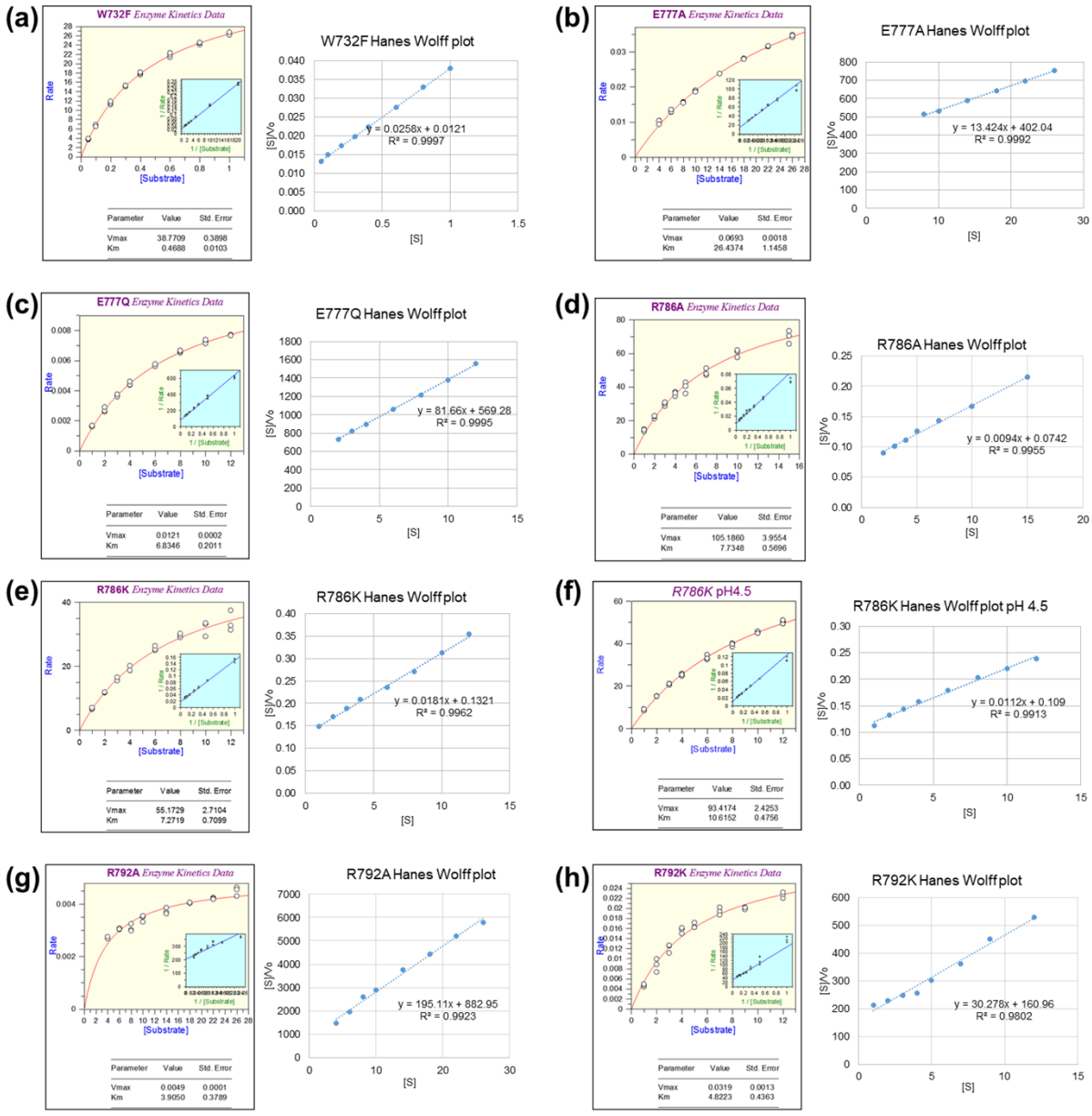


Figure S4-2. ***p*NPGlc kinetics of glycone sugar binding related residues mutants. The kinetic parameters, including k_{cat}, K_M, and k_{cat}/K_M, of purified *Tx*GH116 mutants from *E. coli* were calculated with linear (Hanes-Wolff [S]/v_0 vs. [S]) plots and checked by nonlinear regression of Michaelis-Menten plots with the Grafit 5.0 computer program (Erithacus Software, Horley, UK).**

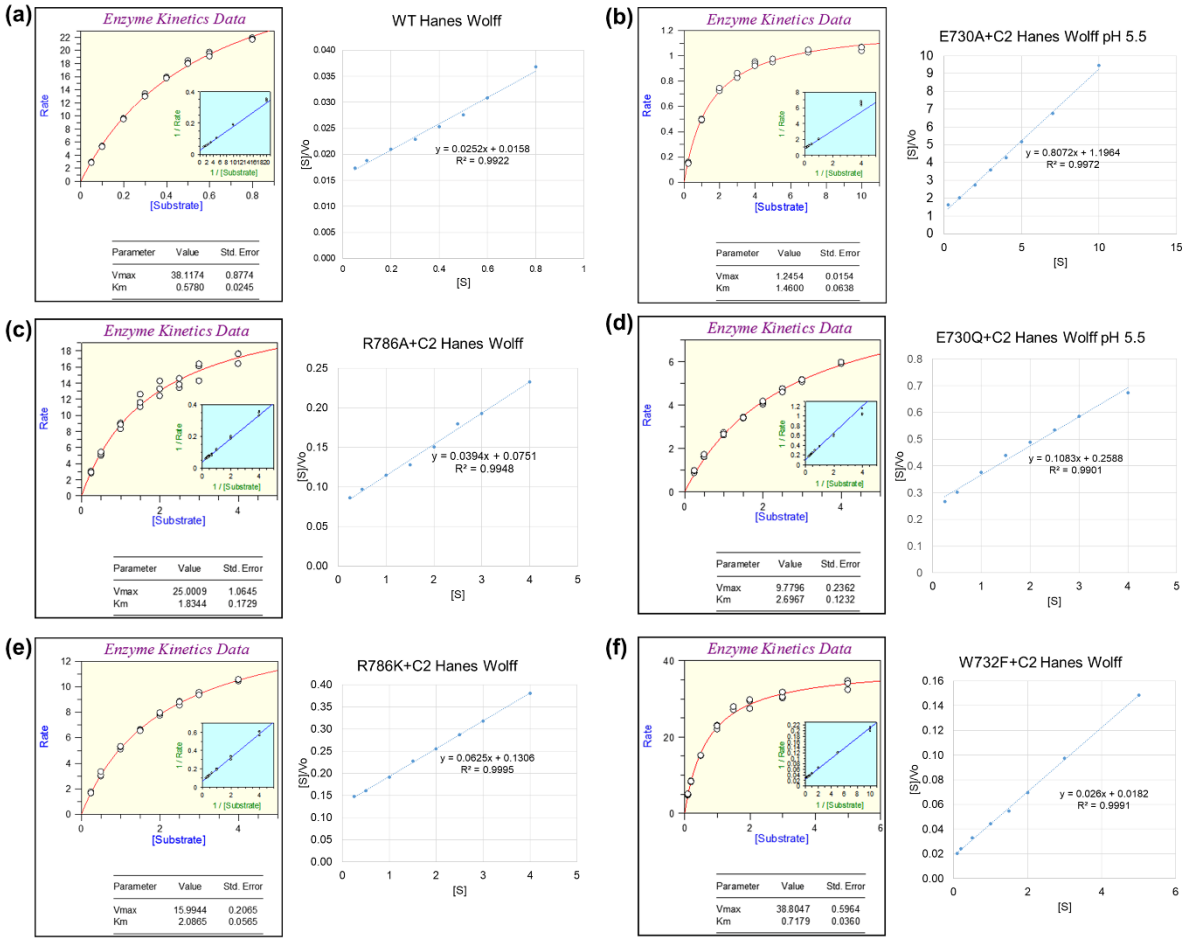


Figure S5. Cellobiose kinetics of *Tx*GH116 variants with mutations of glycone sugar binding related residues. Kinetics values were determined by dividing the amount of glucose released by 2, since two glucose molecules are released per glycosidic bond hydrolyzed.

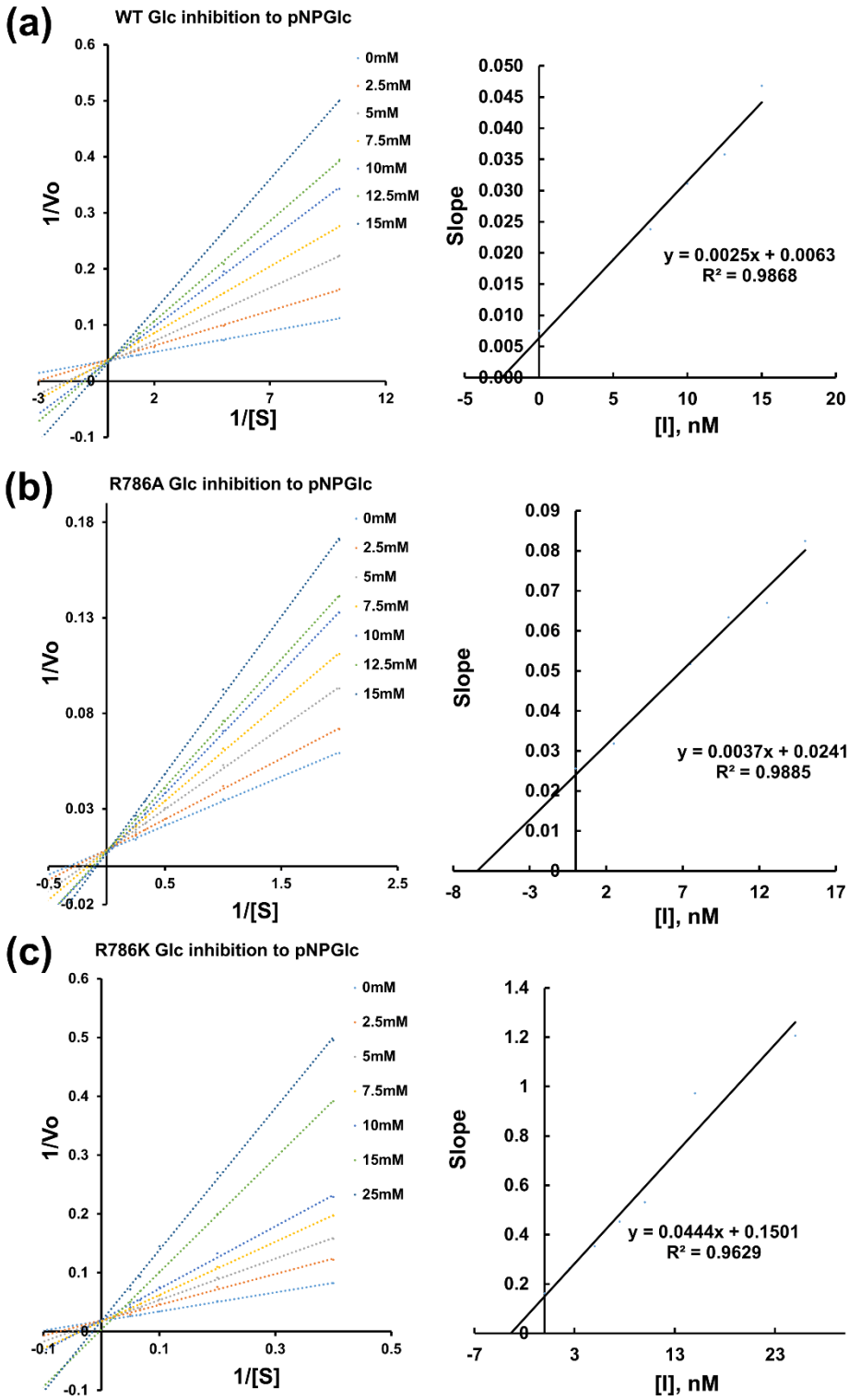


Figure S6. Competitive inhibition of glucose to *Tx*GH116 and its mutants. *Tx*GH116 (a), R786A (b), R786K (c) were pre-incubated with different concentrations of glucose in 50 mM sodium acetate, pH 5.5, at 37 °C for 10 min and then assayed for release of *p*NP from 0.1, 0.2, 0.5, 0.75, and 1 mM *p*NPGlc at 60 °C for 20 min. (Left) 1/v versus 1/[S] plot in the presence of different concentrations of an inhibitor. (Right) Replot of slope from reciprocal plot versus [I].

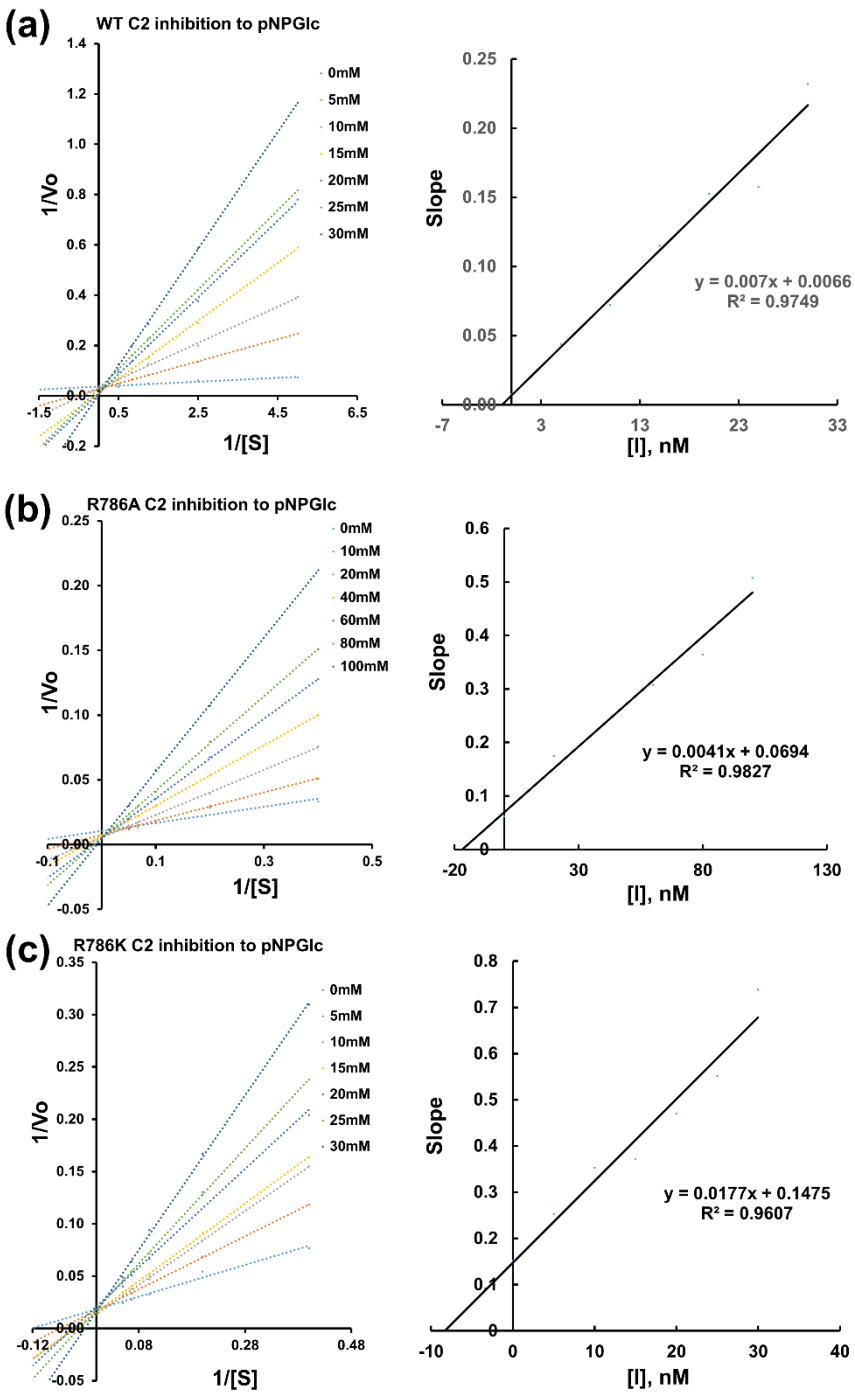


Figure S7. Competitive inhibition of cellobiose to *Tx*GH116 and its mutants. *Tx*GH116 (a), R786A (b), R786K (c) were assayed for release of *p*NP from 0.2, 0.4, 0.8, 1.2, and 2 mM *p*NPGlc at 60 °C for 10 min (WT); 2.5, 5.0, 10.0, 15.0 and 20.0 mM *p*NPGlc at 60 °C for 15 min (R786A and R786K); (Left) 1/v versus 1/[S] plot in the presence of different concentrations of an inhibitor. (Right) Replot of slope from reciprocal plot versus [I], for which the x-intercept is the negative value of the competitive inhibition constant, K_{ic}.

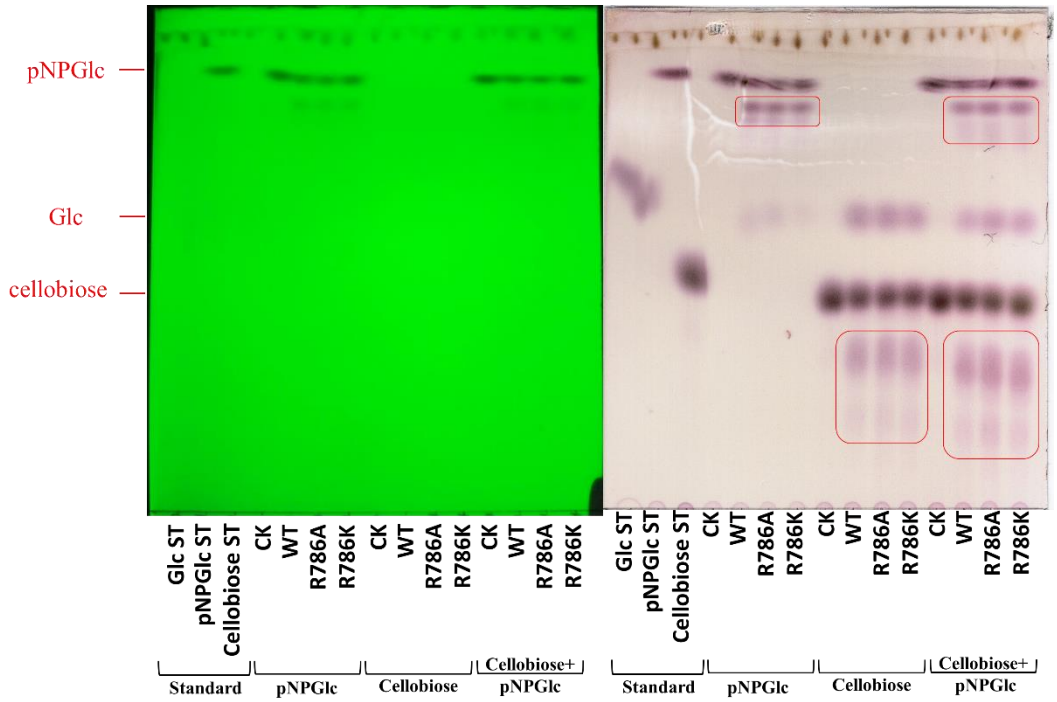


Figure S8. Tranglycosylation analysis between WT and R786 mutation variants on hydrolysis of *p*NPGlc and cellobiose substrate. WT and R786 mutants show similar level transglycosylation products when reaction have single substrate, but R786 mutants have more oligosaccharide products (red boxes) when both *p*NPGlc and cellobiose substrates are included in the reaction. The reaction contained 10 mM substrate in 50 mM NaOAc buffer, pH 5.5, incubated at 60 °C for 18 hours. Silica gel TLC was developed in EtOAc:Acetic acid:H_2O (11:5:4 v/v/v) for two rounds. UV lamp light was used to visualize nitrophenol molecular componds on the plates (left). 12% sulfuric acid in ethanol was sprayed on TLC plate to see carbohydrates (right).

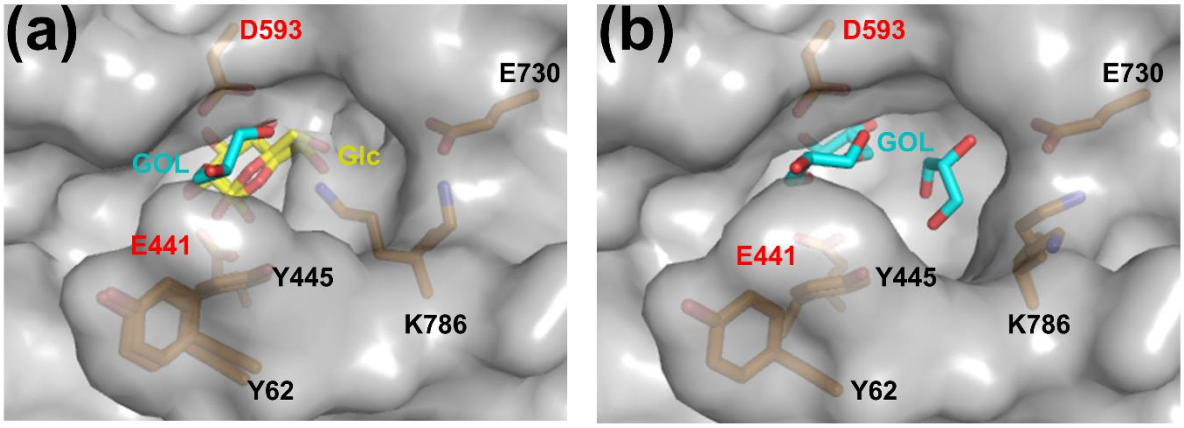


Figure S9. Slot like active site entrance comparison between *Tx*GH116 R786K with (a) and without (b) glucose ligand. R786K with two conformations show narrow entrance space (a), while native R785K have an enlarged entrance space (b) with an extra glycerol molecule in the place of the original arginine side chain. The glucose (Glc) ligand is shown as yellow sticks, glycerol (GOL) molecules are shown as cyan sticks, and catalytic residues are labeled in red.

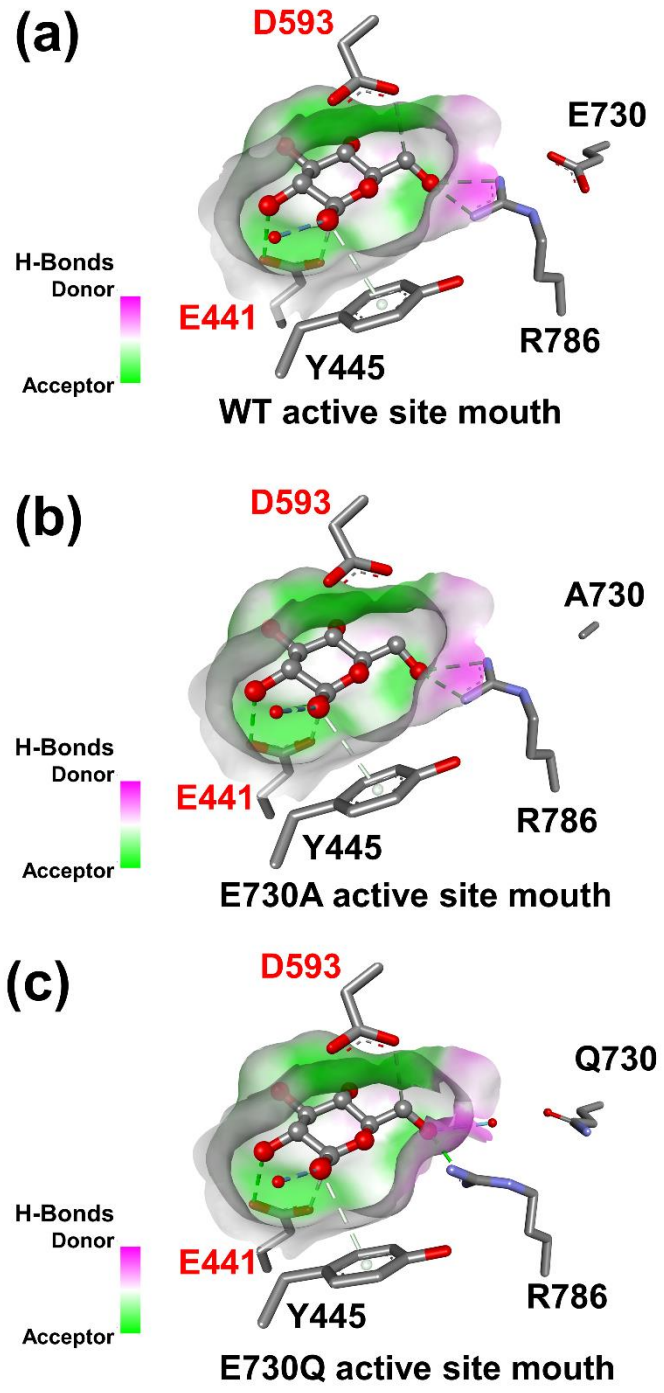


Figure S10. E730A and E730Q active site interactions. Little difference is seen compared with WT, but the R786 residue shows more flexibility because of the elimination of the hydrogen bond from E730 to the R786 guanidinyl group. Catalytic residues are labeled in red, and glycone sugar-binding residues and residue 730 in black. The surface coloration represents the hydrogen bond donor/acceptor potential of the underlying residues, with donors colored in green and acceptors in purple. This structures were generated with Discovery Studio Visualizer™ Software, Version 3.5.

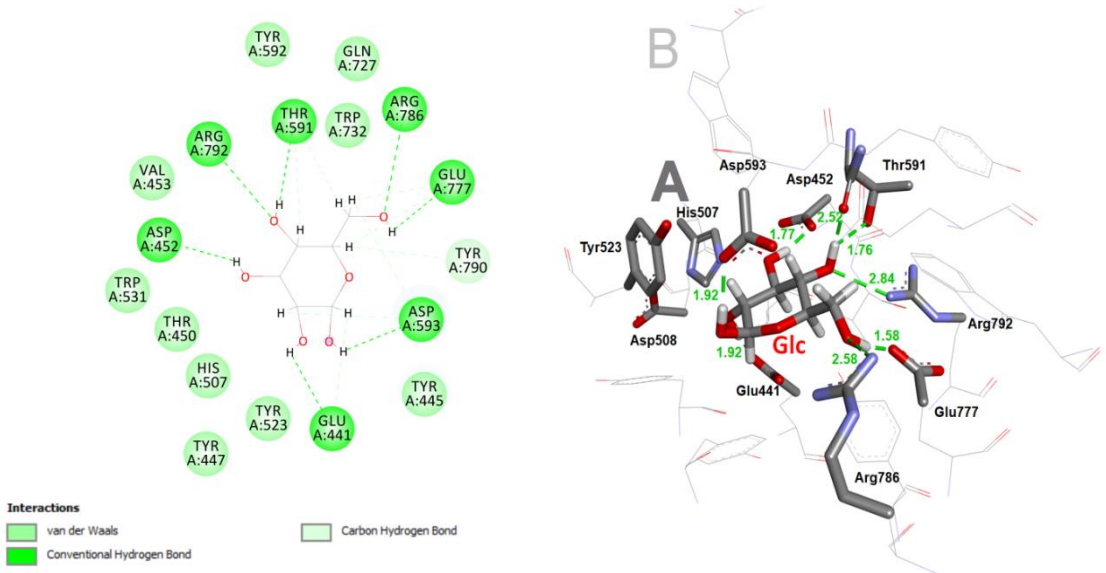


Figure S11. 2D protein-ligand interaction map (left) and 3D (optimized) structure (right) of WT *Tx*GH116-glucose (Glc) complex obtained at ONIOM2 (B3LYP/6-31G(d,p):PM3) level. Regions A (shown in *tube model*) and B (shown in *wire model*) used in the ONIOM calculations. Green dash lines indicate H-bond distances (in Å) between glucose, Glc, and *Tx*GH116 active site. Important interactions are also indicated, which mainly involves the hydrogen bond and van der Waals interactions.

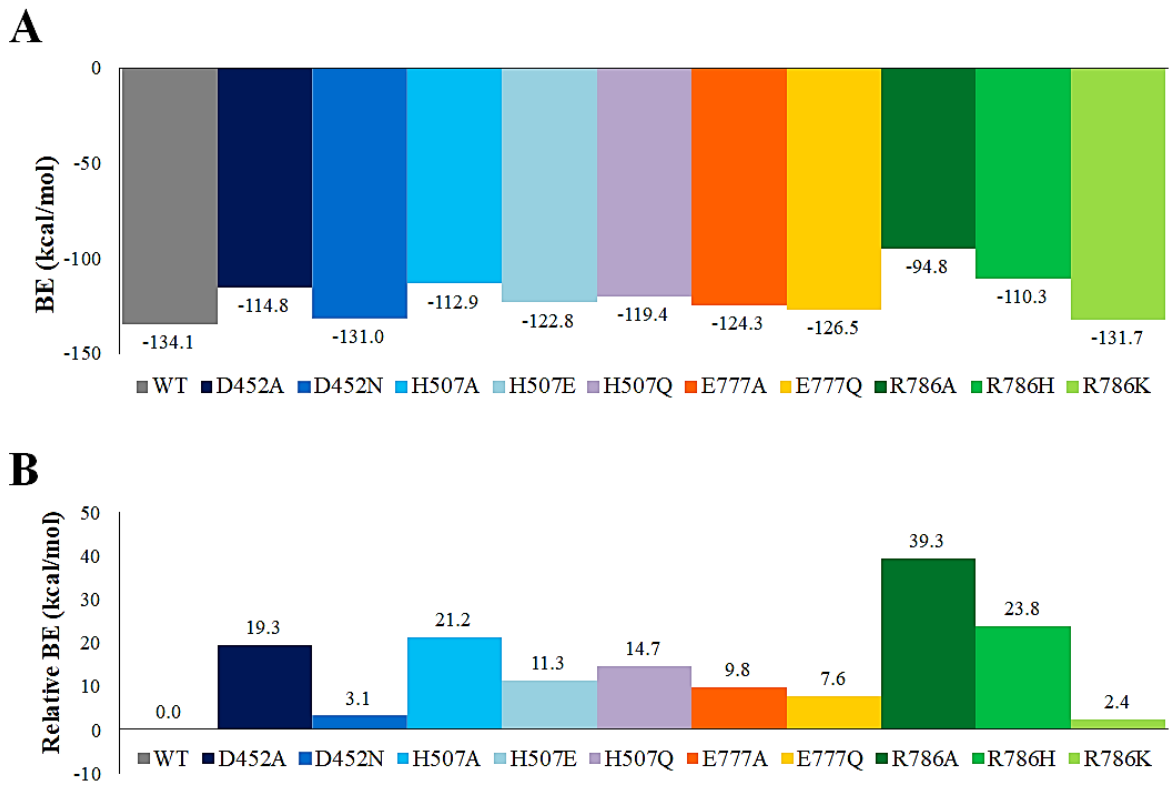


Figure S12. Binding interaction energies (*BE*, kcal/mol) of glucose for WT and mutant (D452A, D452N, H507A, H507Q, H507E, E777A, E777Q, R786A, R786H, R786K) models (a). The corresponding binding energies relative to the BE of WT were also included (b). Positive and negative values of the relative BE (Δ*BE*) for particular mutant models indicate the unfavorable and favorable binding of the glucose ligand in the *Tx*GH116 pocket, respectively.

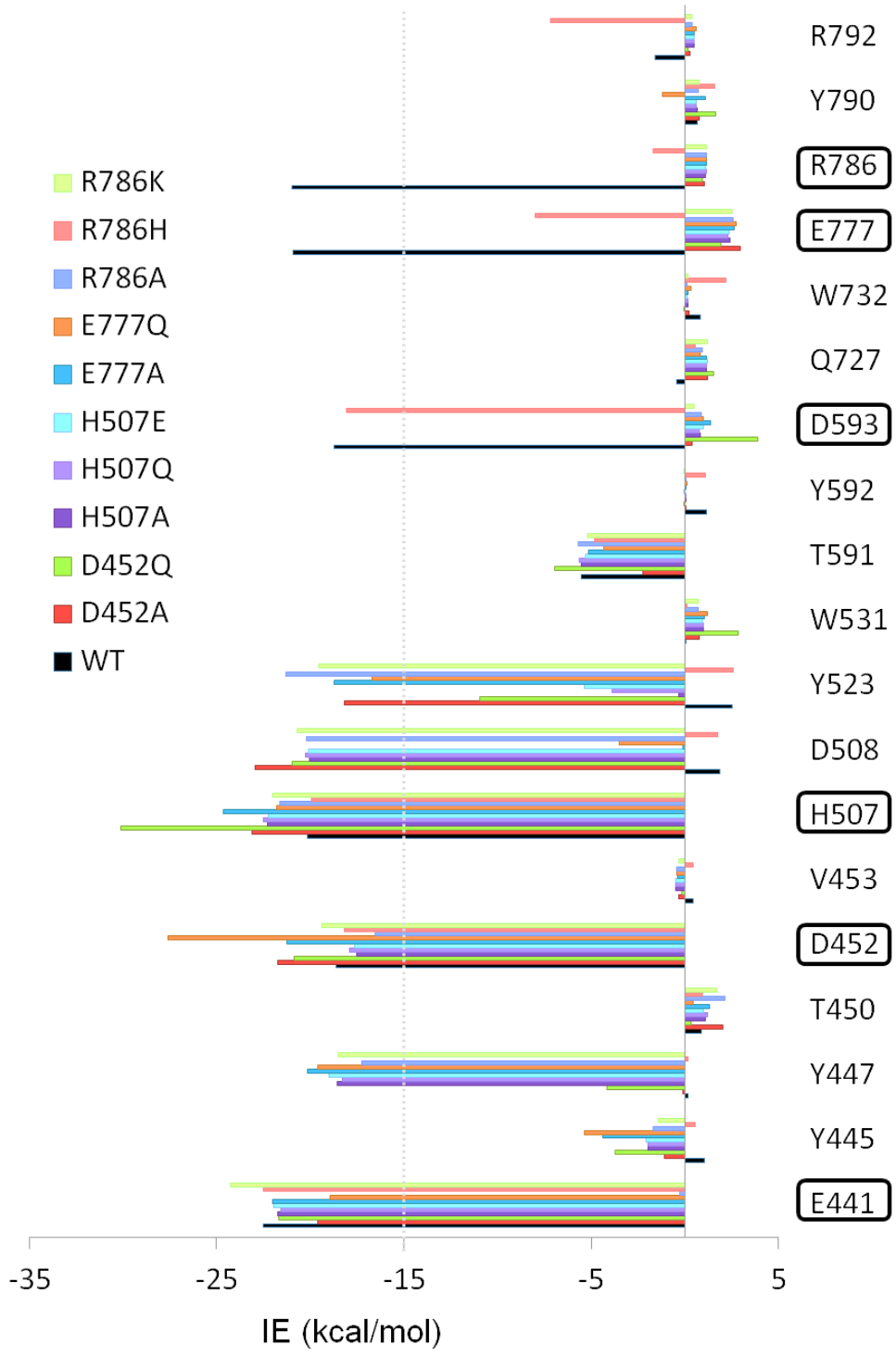


Figure S13. Particular interaction energy (*IE*, kcal/mol) between the glucose and individual residues for wildtype (WT) and mutant *Tx*GH116 models (D452A, D452N, H507A, H507Q, H507E, E777A, E777Q, R786A, R786H, R786K). Six residues (E441, D452, H507, D593, E777, R786) were identified as strong electrostatic contributors for the *Tx*GH116-glucose binding, with $|IE| > 15$ kcal/mol (shown in rectangular line).

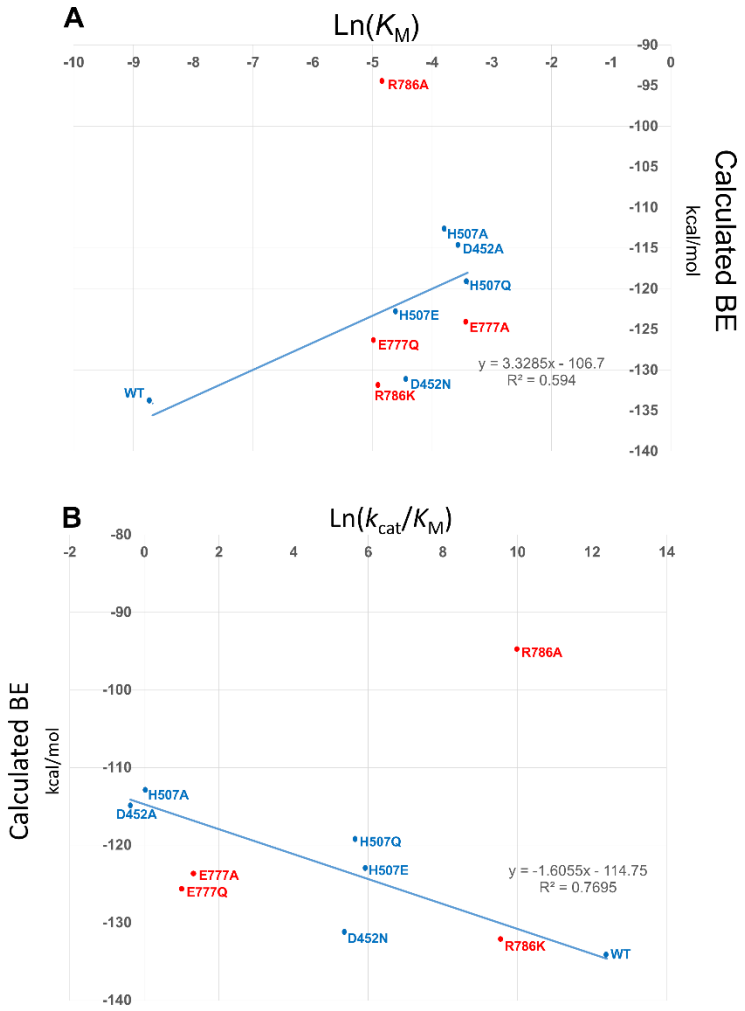


Figure S14. Correlation of calculated binding energy (BE) with kinetic constants. A. Correlation with of BE with ln(K_M), which should correlate if the binding energies are accurate and K_M is correlated with the dissociation constant, K_D, since $\Delta G = -RT\ln(K_A) = RT\ln(K_D)$, $K_D = 1/K_A$. B. Correlation of BE with ln(k_{cat}/K_M), which should correlate if the binding energies are accurate and the binding is proportional to the binding of the energy of binding of the transition state of the first committed step in the reaction $\Delta G = -RT\ln(K_A) \propto -RT\ln(k_{cat}/K_M)$. The lines represent the linear regression of the points in blue, representing the wild type and mutations at residues not binding to the C6OH. The mutations of residues binding to the C6OH are shown in red. Inclusion of the mutations of C6OH-binding residues, particularly R786A, decrease the correlation to an insignificant level.

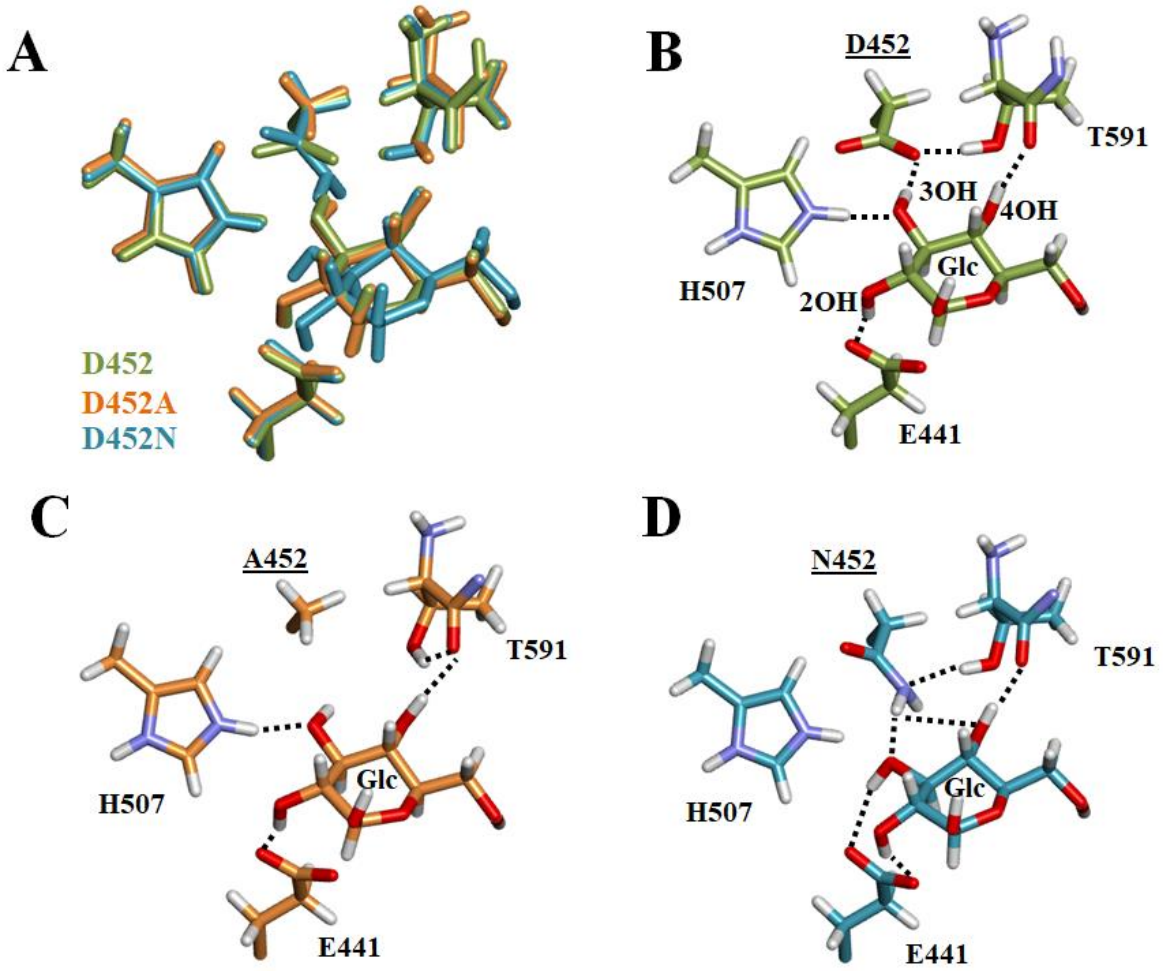


Figure S15. Active site comparison of ONIOM2 models of wildtype, D452A, and D452N. Structural overlap of the *Tx*GH116-glucose complex models (obtained at the ONIOM2 [B3LYP/6-31G(d,p):PM3] level) between the WT (green) and the mutants, D452A (orange) and D452N (cyan) (a) and their hydrogen bond (H-bond) interactions involved; (b) WT, (c) D452A, (d) D452N. For clarity, only the residues making hydrogen bonds at C2OH, C3OH and C4OH of glucose are shown. The mutated residue 452 is also indicated with underlined labels.

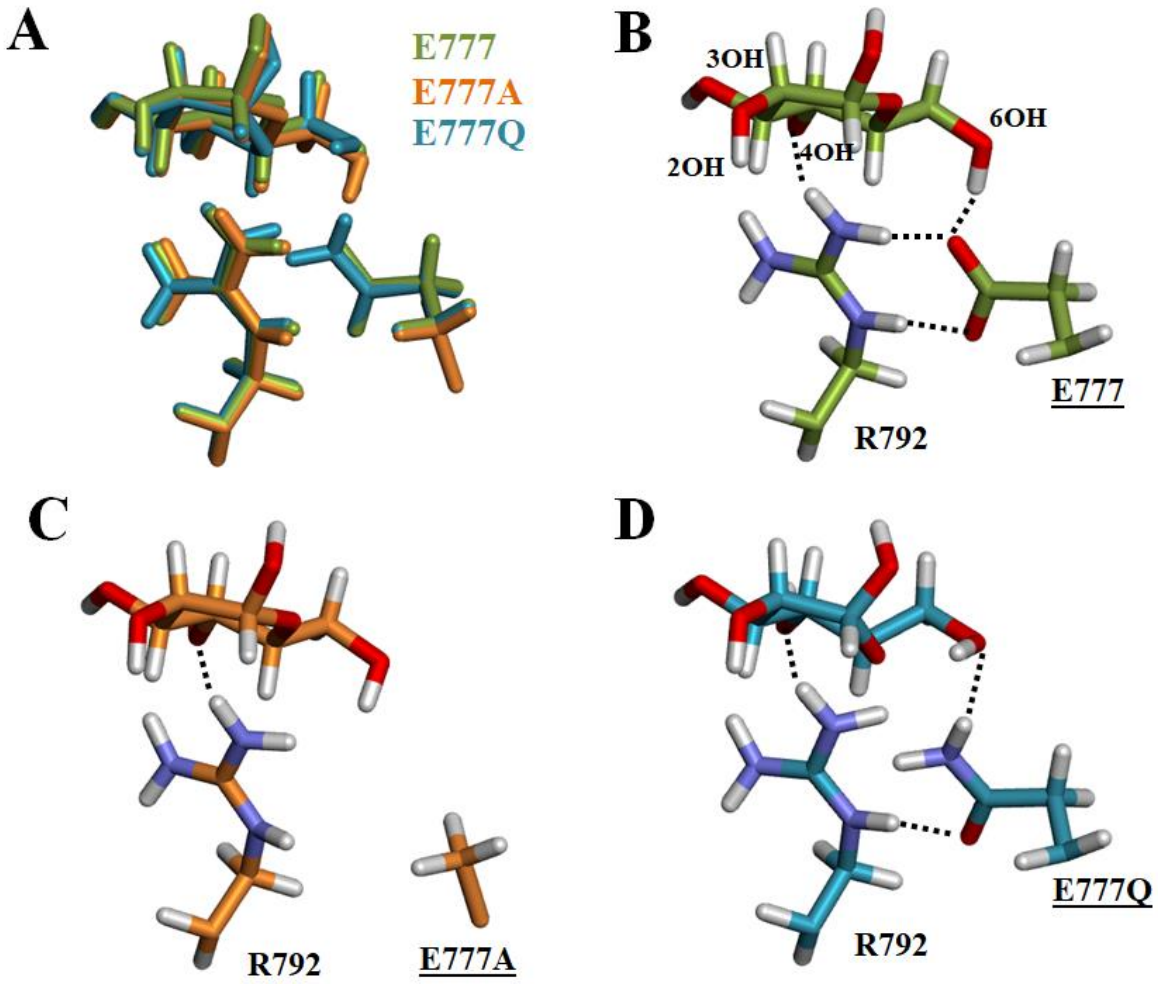


Figure S16. Active site comparison of ONIOM2 models of wildtype, E777A, and E777N. Structural overlap of the *Tx*GH116-glucose complex models (obtained at the ONIOM2 [B3LYP/6-31G(d,p):PM3] level) between the WT (green) and the mutants, E777A (orange) and E777Q (cyan) (a) and their hydrogen bond (H-bond) interactions involved; (b) WT, (c) E777A, (d) E777Q. For clarity, only the residues R792 and E777 (E777Q/E777A) were shown. The mutated residue 777 is also indicated with underlined labels.

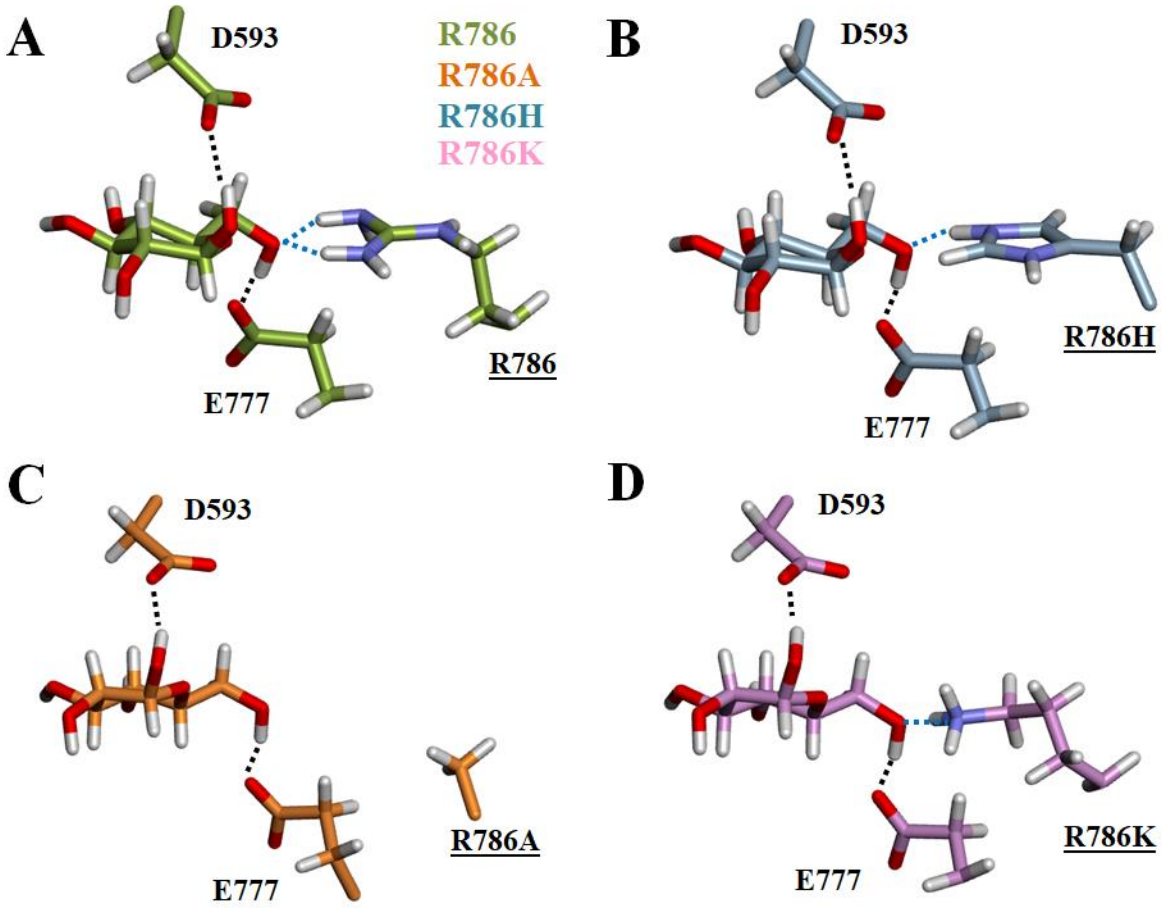


Figure S17. Active site comparison of ONIOM2 models of wildtype, R786H, R786A, and R786K. H-bond interactions in the *Tx*GH116-glucose complex models (obtained at the ONIOM2 [B3LYP/6-31G(d,p):PM3] level) for (a) WT (green), (b) R786H (light blue), (c) E786A (orange) and (d) R786K (pink). For clarity, only the residues D593, E777 and R786 (R786Q/R786A) were shown. The mutated residue 786 is also indicated with underlined labels.

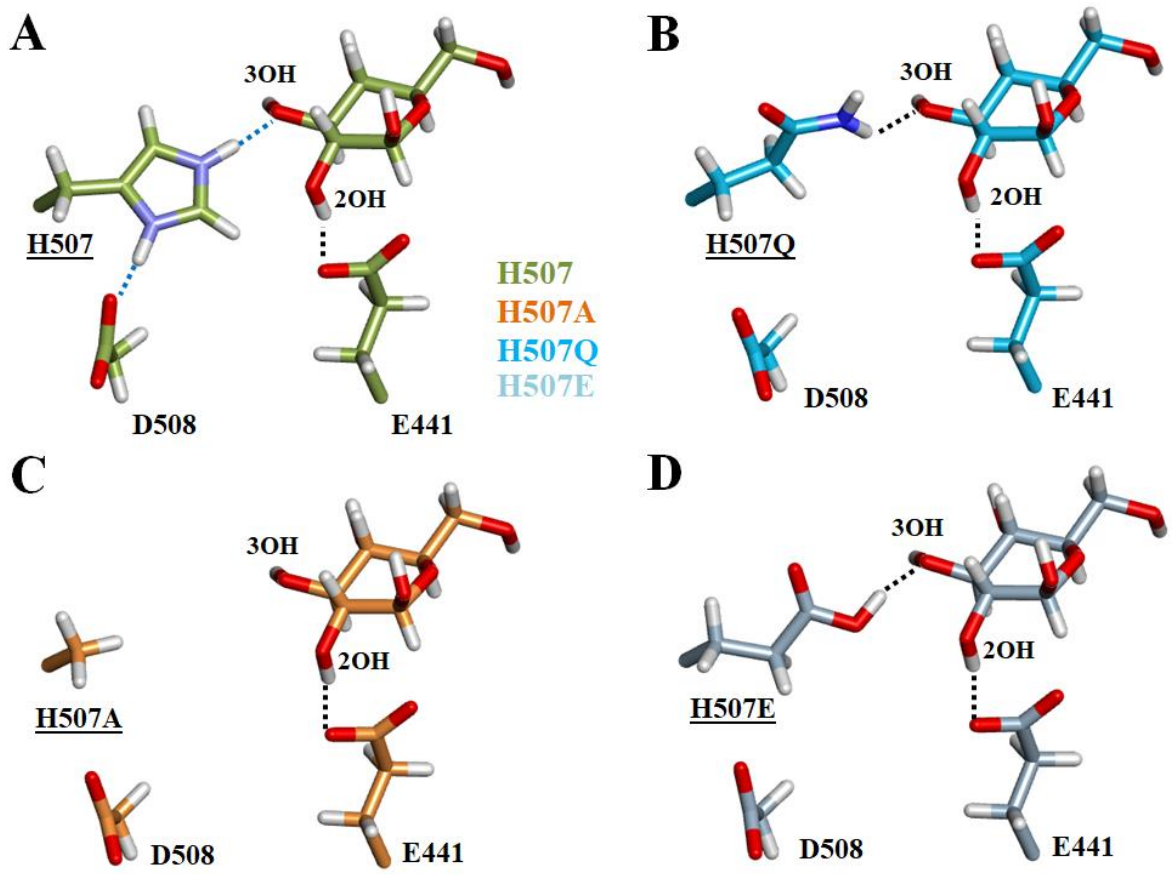


Figure S18. Active site comparison of ONIOM2 models of wildtype, H507Q, H507A and H507E. H-bond interactions in the *Tx*GH116-glucose complex models (obtained at the ONIOM2 [B3LYP/6-31G(d,p):PM3] level) for (a) WT (green), (b) H507Q (cyan), (c) H507A (orange) and (d) H507E (light blue). For clarity, only the residues E441, D508 and H507 (H507Q/H507A/H507E) were shown. The mutated residue 507 is also indicated with underlined labels.

Reference

1. Charoenwattanasatien, R.; Pengthaisong, S.; Breen, I.; Mutoh, R.; Sansenya, S.; Hua, Y.; Tankrathok, A.; Wu, L.; Songsiriritthigul, C.; Tanaka, H.; et al. Bacterial beta-Glucosidase Reveals the Structural and Functional Basis of Genetic Defects in Human Glucocerebrosidase 2 (GBA2). *ACS Chem Biol* **2016**, *11*, 1891-1900.