Electronic Supplementary Information (ESI) for:

Mechanism of Oligosaccharide Synthesis via a Mutant GH29 Fucosidase

Tucker Burgin[a] and Heather B. Mayes[a‡]

Transition Path Sampling Analysis Methods

Here we will detail the methodology for analyzing the data produced during aimless shooting in order to obtain and evaluate the reaction coordinate, as well as to produce the energy profile along it.

Likelihood maximization.

We used the inertial likelihood maximization algorithm of Peters.[S1] This is a method for obtaining a model reaction coordinate (RC) in the form of a linear combination of configurational variables (that is, variables based only on atomic coordinates). The inertial implementation of the algorithm is demonstrably superior to older versions in that it optimizes for collective variables (CVs) whose value *and* rate of change are predictive of commitment to products or reactants (rather than only taking into account the values), and as such produces a model with less error due to recrossing of the separatrix. This is implemented by including an additional CV signifying the rate of change of each configurational CV during the first optimization step to select the most important CVs to include in the RC, and then performing another optimization step on only those configurational CVs chosen during the previous step to produce the final RC. An RC consisting of few CVs is important for both computational tractability and intuitive interpretation, so we limited our RC to three terms (plus a constant) and required that each additional term up to that maximum increase the Bayesian Information Criterion score of the model by at least 10.[S2]

Committor analysis.

After an RC has been produced by likelihood maximization, it must be validated by committor analysis.[S3] This is a procedure wherein a large number of shooting points with RC values close to the transition state value (that is, along the separatrix) are tested several times in order to approximate their relative likelihood of committing to the reactant state (A) *versus* the product state (B), measured as the ratio $p_B = N_B/(N_A+N_B)$, where N_A is the number of simulations for a given shooting point that commit to the A basin, and similarly for N_B. A successful committor analysis result is one that produces a histogram of p_B values centered on 0.5 and as narrow as possible, although in practice sampling error may be significant. We performed our committor analysis on 143 shooting points with RC values within 0.1 of the transition state value, with 10 trials per point in order to obtain a reasonable approximation of the underlying committor distribution.[S4]

Equilibrium path sampling.

Equilibrium path sampling (EPS) is a method for obtaining the energy landscape along a given RC without applying a bias to the Hamiltonian.[S5] It is analogous to aimless shooting, except acceptance is based on membership of any of the frames of a trajectory within a given window of RC values rather than on commitment to reactant and product basins. Our EPS simulations were performed across 30 windows of width 1.0 with overlap of 0.2 with adjacent windows. Simulations consisted of 10 beads separated by 5 1-fs simulation steps each. The potential of mean force (PMF) was obtained by dividing each window into 4 bins and evaluating the free energy profile in each window as:

$$G(x) = -k_B T ln(P(x)) + C, \quad (1)$$

where $k_B T$ is the Boltzmann-weighted temperature, $P(x)$ is the probability of observing an RC value in the bin indicated by x, and the constant C was chosen to align the overlapping regions of adjacent windows.

Custom Molecular Mechanics Force Fields

As described in the main text, the force field for the substrate molecules was constructed by manually modifying the Generalized Amber Force Field (GAFF). First, all of the GLYCAM06 parameters appropriate for fucose and xylose were included.[S6] The additional parameters that we added are shown in Table S1, along with their sources. The atom types tabulated therein correspond to the substrate structures as follows: the azide nitrogens are ni, ne, and nd, respectively, with ni bonded to the fucose; the fucose ring carbons are c3, and the hydrogens bonded to them are h1; the sugar ring oxygens as well as the oxygen that articulates the xylose to the nitrophenyl group are all os; the xylose ring carbons are CT; and the nitrophenyl ring carbons are ca.

The resulting molecular mechanics (MM) force field was accepted only after comparison between the MM minimized structure and the same structure minimized using the DFTB quantum mechanics (QM) model. The minimizations were allowed to run until the gradient in energy between steps converged to

[a] *University of Michigan Department of Chemical Engineering, 3074 H.H. Dow, 2300 Hayward Street, Ann Arbor, MI USA.*
‡ *E-mail: hbmayes@umich.edu*

0.1 kcal/mol-Å. The structure comparisons are shown in Figure S1.

Transition State Hypothesis Simulations

Simulations to build the 80 initial transition state hypotheses to seed aimless shooting for the α-1,4 reaction were performed as follows. We built a unique set of simulation files for each combination of the four values for each of the four bond lengths described in the main text, excluding any combination with more than one "extreme" value (that is, either the largest or smallest allowable value for a given bond length). Restraints were applied to pull the bond lengths towards the desired values using restraint weights of 80 kcal/mol-Å^2, or 160 kcal/mol-Å^2 for the bond between the acceptor O4 and its hydrogen atom, to minimize oscillations associated with the motion of the very light hydrogen atom. The simulations were run in Amber 16[S7] using the DFTB QM/MM model with the QM region set to contain both substrate molecules, the side chains of every protein residue in the first "shell" of residues around the active site, the entirety of the G224 residue, and the first shell of water molecules near the entrance to the active site cleft, as visualized in Figure S2. This QM region was chosen to minimize any errors associated with the QM/MM transition region (by keeping it far away from any of the reactive atoms), and was the same mask used throughout the QM/MM simulations in this work. There was no observed exchange of water molecules in and out of the QM region, likely owing to the short timescale of the simulations compared to the timescale of water exchange. The simulation settings were the same as those in the QM/MM equilibration described in the main text, and each ran for 100 1-fs steps.

We did not enforce a requirement that the targeted bond lengths were reached in the structures resulting from these simulations, as the goal was not these exact lengths but a variety of structures to test if they could seed pathways connecting reactants to products. To successfully start an aimless shooting search for an ensemble of transition state structures, all that is needed is one or more structures with the potential to proceed to both reactants and products when supplied with randomly chosen (Boltzmann-distributed) momenta in one simulation, and opposite momenta in another. We made the *a priori* assumption that the reaction barrier was much less than 80 kcal/mol, such that the bond stretching restraints would be able to pull the substrates toward the transition state (wherever it may lie). The reasonable aimless shooting acceptance ratios (average 15.91% in those threads that were ever accepted) that we achieved serve as an *a posteriori* validation of the acceptability of our transition state guessing procedure. Specifically, the threads that were accepted at least once, the average acceptance ratio was 15.91%, the smallest was 6.25%, and the largest (with at least 5 moves) was 31.03%. These values are a measure of the efficiency of the simulations, and do not impact the final results.

Collective Variables Included in Likelihood Maximization

Likelihood maximization provides an unbiased means of harvesting a suitable reaction coordinate (RC) from collective variables (CVs) observed during the aimless shooting simulations. Only those CVs that are explicitly included by the researcher are candidates for inclusion in the RC. In order to obtain the best possible RC for a given rare event it is necessary to include *every* CV that might reasonably contribute to prediction of commitment to the products or reactants. To that end, we included 54 CVs in our likelihood maximization. These are listed in Tables S2 and S3. These tables refer to the α-1,4 reaction; in the α-1,3 reaction, the same CV and RC definitions were used, but with the 4NX O4 and H4O atoms replaced with O3 and H3O, respectively.

Full Energy Profiles

The reaction energy profiles shown in the main text are made smooth by averaging the horizontal and vertical positions of the overlapping points between adjacent windows. This is the cause of the somewhat oscillatory nature of the error bars on that figure: there is more sampling available in the overlapping regions, and as such, lower error. For completeness, we also include the raw energy profiles here, in Figure S3.

Equilibrium Constant from Cobucci-Ponzano *et al.*

The equilibrium constant provided in the main text for the reactions performed by Cobucci-Ponzano *et al.*[S8] were calculated based on the reaction conditions described in that paper, as well as with the additional information that the total concentration of transferred fucose (donor) at equilibrium was 3 mM (based on personal communication with the authors). The calculation was performed as follows:

$$K_{eq,\alpha 1,4} = \frac{[\alpha 1,4\ product]_{eq}[free\ azide]_{eq}}{[acceptor]_{eq}[donor]_{eq}} \tag{2}$$

where:

$$[\alpha 1,4\ product] = f_{\alpha 1,4}\eta([donor]_0 - [donor]_{eq}) \tag{3}$$

$$[free\ azide] = [donor]_0 - [donor]_{eq} \tag{4}$$

$$[acceptor]_{eq} = [acceptor]_0 - \eta([donor]_0 - [donor]_{eq}) \tag{5}$$

where $f_{\alpha 1,4}$ represents the fraction of the product forming an α-1,4 bond (55%), and η represents the specificity of the reaction for transferring the fucose to the acceptor molecule rather than to water (91%). The values of the initial concentrations were $[donor]_0$ = 10 mM and $[acceptor]_0$ = 100 mM, and the reaction was performed at 70°C.[S8] An analogous calculation was performed for the α-1,3 reaction.

Notes and references

S1 B. Peters, *Chem. Phys. Lett.*, 2012, **554**, 248–253.

S2 G. Schwarz, *The Annals of Statistics*, 1978, **6**, 461–464.

S3 G. T. Beckham and B. Peters, *Computational Modeling in Lig-*

nocellulosic Biofuel Production, American Chemical Society, Washington, D.C., 2010, ch. 13, pp. 299–332.

S4 B. Peters, *J. Chem. Phys.*, 2006, **125**, 241101.

S5 B. Peters, N. E. R. Zimmermann, G. T. Beckham, J. W. Tester and B. L. Trout, *J. Am. Chem. Soc.*, 2008, **130**, 17342–17350.

S6 K. N. Kirschner, A. B. Yongye, S. M. Tschampel, C. R. Daniels, B. L. Foley and R. J. J. Woods, *J. Comp. Chem.*, 2008, **29**, 622–655.

S7 D. A. Case, R. M. Betz, D. S. Cerutti, T. E. Cheatham, III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao and P. A. Kollman, *Amber 16*, University of California, San Francisco, 2016.

S8 B. Cobucci-Ponzano, F. Conte, E. Bedini, M. M. Corsaro, M. Parrilli, G. Sulzenbacher, A. Lipski, F. Dal Piaz, L. Lepore, M. Rossi and M. Moracci, *Chem. Biol.*, 2009, **16**, 1097–1108.

S9 A. T. P. Carvalho, P. A. Fernandes and M. J. Ramos, *Int. J. Quantum Chem.*, 2007, **107**, 292–298.

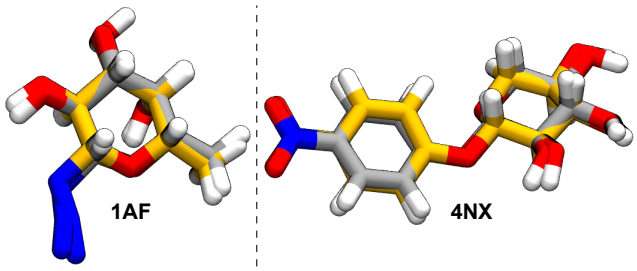

Fig. S1 Comparisons between the reactant substrate molecules minimized using the DFTB QM model (gold) and the custom MM force field using the parameters shown in Table S1 (silver). The structures are fitted atop one another to minimize the RMSD between like atoms, not including hydrogens.

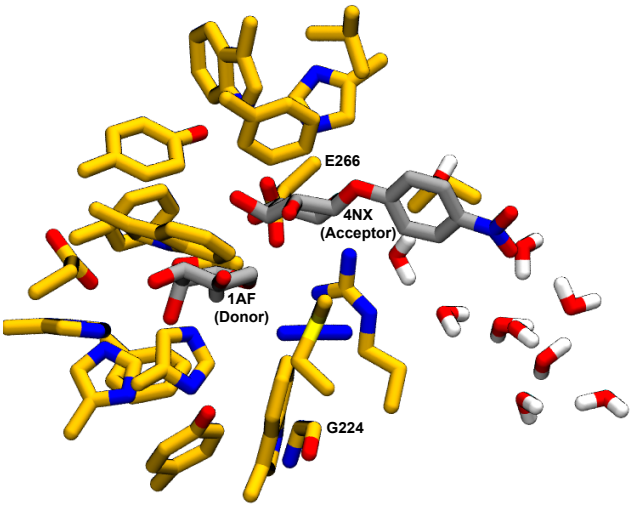

Fig. S2 Visualization of the QM region used for the QM/MM simulations. As shown, the full shell of residues and water molecules around the substrates was included. Notably, there are no water molecules in the active site, though the first layer of water molecules bordering the active site cleft was included for completeness. Hydrogens on protein residues and on the substrates were included in the QM region, but are omitted here for clarity. This snapshot shows a candidate transition state structure, but the same QM region was used for the reactant and product states. The substrates and two key residues are labeled.

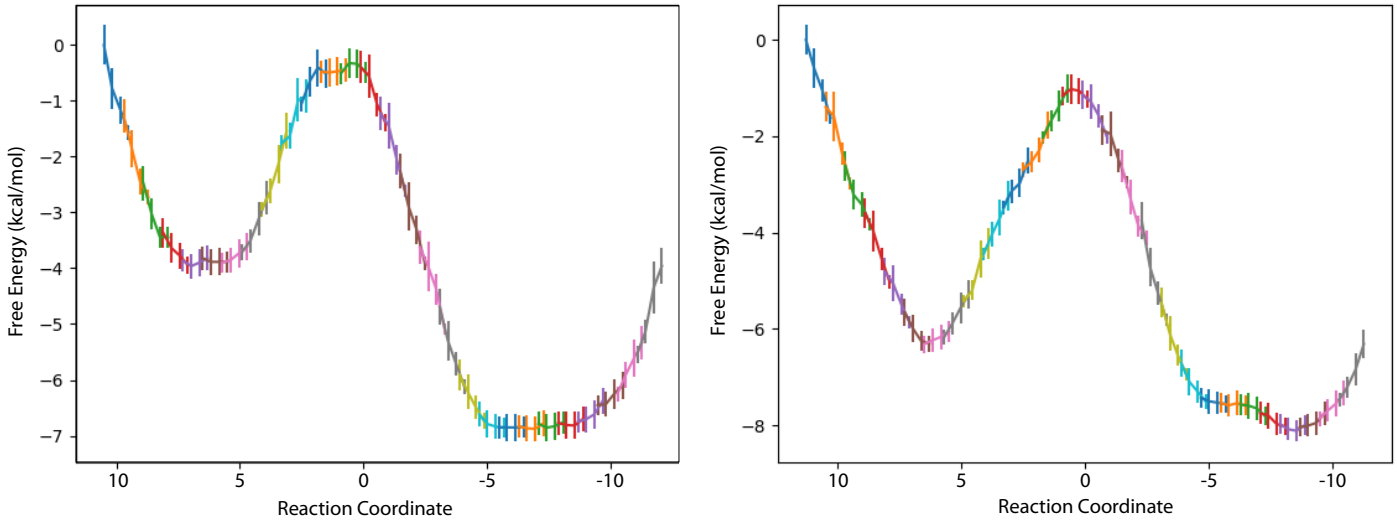

Fig. S3 Raw PMFs for the α-1,4 (left) and α-1,3 (right) reactions. Though these plots represent the same data as shown in the main paper, here they are shifted along the energy axes and mirrored along the reaction coordinate axes. Each successive color represents the energy profile recovered from a single equilibrium path sampling window.

Table S1 Parameters combined with those from GAFF and GLYCAM06 to build the custom force field. Units for equilibrium values are Å for bond distances and degrees for angles and dihedrals. Units for weights are kcal/mol-Å^2, kcal/mol-rad^2, and kcal/mol for bonds, angles, and dihedrals, respectively. Further details are available at the citations.

Parameter	Weight	Equi-librium Value	Source
c3-ni	277.5	1.49	S9
ni-nd	710.0	1.34	S9
nd-ne	1312.0	1.14	S9
c3-c3-ni	74.8	113.36	S9
h1-c3-ni	68.3	108.87	S9
c3-ni-nd	64.0	115.60	S9
ni-nd-ne	42.4	173.54	S9
ni-c3-os	70.04	111.230	Analogy to n2-c3-os from GAFF
c3-ni-nd-ne	0.25	180.00	S9
c3-c3-ni-nd	11.11	0.00	S9
h1-c3-ni-nd	11.11	0.00	S9
os-c3-ni-nd	11.11	0.00	Analogy to c3-c3-ni-nd
ca-ca-os-CT	1.410	198.800	Calculated to recreate Gaussian dihedral scan

Table S2 Complete list of CVs included in likelihood maximization. CVs with entries in only the first two columns are distances; those with three entries are angles; and those with four are dihedrals. CVs 9 and 21 are special cases: the former is the distance between the average positions of atoms 4272 and 4273, and 4272 and 4075, respectively; while the latter is the difference between the two indicated distances. Atom indices correspond to those in Table S3. The CVs that were selected by inertial likelihood maximization to appear in the final RC are marked with an asterisk (*).

CV Name	Mask 1	Mask 2		Mask 3	Mask 4
CV_1	4272	7175			
CV_2	7175	7174			
CV_3*	7174	7185			
CV_4*	7185	7186			
CV_5	7172	7174			
CV_6	7186	3584			
CV_7	7191	3584			
CV_8	7186	3582			
CV_9	4272,4273	4072,4075			
CV_{10}	7186	2704			
CV_{11}	4071	7191			
CV_{12}	7172	7185			
CV_{13}	958	4273			
CV_{14}	2044	7208			
CV_{15}	7194	496			
CV_{16}	7204	987			
CV_{17}	7186	7188			
CV_{18}	7191	7188			
CV_{19}	7186	7191			
CV_{20}	7174	4273			
CV_{21}*	4273	7175	–	7175	7174
CV_{22}	7200	7185		7186	
CV_{23}	7200	7185		7174	
CV_{24}	4265	4268		4271	
CV_{25}	4272	7173		7184	
CV_{26}	7185	7174		7172	
CV_{27}	7174	7175		4273	
CV_{28}	7186	7188		7191	
CV_{29}	7188	7186		7165	
CV_{30}	7197	7195		7196	7185
CV_{31}	7197	7195		7193	7190
CV_{32}	7195	7196		7185	7187
CV_{33}	7195	7193		7190	7187
CV_{34}	7196	7185		7187	7190
CV_{35}	7193	7190		7187	7185
CV_{36}	7196	7185		7186	7188
CV_{37}	7185	7186		7188	7191
CV_{38}	7196	7185		7174	7172
CV_{39}	7196	7185		7174	7175
CV_{40}	7185	7174		7172	7169
CV_{41}	7174	7172		7169	7168
CV_{42}	7174	7172		7176	7180
CV_{43}	7172	7169		7168	7166
CV_{44}	7172	7176		7180	7166
CV_{45}	7176	7180		7166	7165
CV_{46}	7169	7168		7166	7165
CV_{47}	7168	7166		7165	7154
CV_{48}	7166	7165		7154	7163
CV_{49}	7174	7175		4273	4271
CV_{50}	7175	4273		4271	4268
CV_{51}	4273	4271		4268	4265
CV_{52}	4271	4268		4265	4263
CV_{53}	4268	4265		4263	4274
CV_{54}	4265	4263		4274	4275

Table S3 Definitions of atom indices corresponding to those in Table S2. Atom names correspond to those in the standard residue definitions in the Amber force field[S7] for protein atoms, and standard notation for sugars. 4NX CG and CD2 refer to the nitrophenyl carbon bonded to the oxygen and another bonded to that one, respectively.

Atom Index	Identity
4272	E266 OE1
4273	E266 OE2
7175	4NX H4O
7174	4NX O4
7185	1AF C1
7186	1AF ni
7188	1AF nd
7191	1AF ne
7172	4NX C4
7200	1AF H1
3584	G244 O
3582	G224 H1
4071	R254 CZ
4072	R254 NH1
4075	R254 NH2
2704	Y171 OH
958	Y64 HH
2044	H129 NE2
7208	1AF HO2
7194	1AF O4
496	H34 HE2
7204	1AF HO3
987	E66 OE2
4265	E266 CB
4268	E266 CG
4271	E266 CD
7173	4NX H4
7184	4NX ON2
7165	4NX O1
7197	1AF C6
7195	1AF C5
7196	1AF O5
7193	1AF C4
7190	1AF C3
7187	1AF C2
7169	4NX C5
7168	4NX O5
7176	4NX C3
7180	4NX C2
7166	4NX C1
7154	4NX CG
7163	4NX CD2
4263	E266 CA
4274	E266 C
4275	E266 O