Transition Pathway Perspectives in Molecular Simulations of Enzymes

by Tucker Burgin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemical Engineering)
University of Michigan
2021

Doctoral Committee:

Assistant Professor Heather Mayes, Chair
Professor Sharon Glotzer
Assistant Professor Nicole Koropatkin
Professor Ronald Larson

Tucker Burgin

tburgin@umich.edu

ORCID iD: 0000-0002-3714-3537

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Molecular simulations of enzymes can provide a wealth of knowledge to explain and characterize these uniquely complex and beautiful molecular machines. However, the vast majority of the most interesting properties of enzymes depend on what are called “rare events” – statistically rare molecular transformations such as reactions – that cannot be observed using entirely unbiased simulations. Conversely, injection of too much bias into a simulation can mask the real mechanics of the system and lead to incorrect results. In this annotated compilation of manuscripts – one describing the rationalization of a known mechanism; the next a discovery of an unknown mechanism; and the last describing a novel software tool for automated enhanced sampling – we explore enzymatic mechanisms through simulations, relying on the minimum possible bias while capturing as complete a transition pathway perspective as possible. We also discuss the appropriate role of researcher “intuition” or general chemical knowledge in studying such complex mechanisms as are present in enzymes. Special attention is given to glycoactive enzymes, as well as to strategies for generalizing the lessons learned from studying especially unusual mechanisms or events.

CHAPTER I

Introduction

1.1 Enzymes and Intricacy

In the most general terms, “life” as it exists on Earth can be defined as the manipulation of the timescales of chemical reactions in order to produce a system that replicates itself faster than entropy disassembles it. This understanding of life in terms of the Second Law of Thermodynamics – and specifically, the theory of non-equilibrium thermodynamics pioneered by Boltzmann and Schrödinger, before eventually becoming the subject of a 1977 Nobel Prize awarded to Ilya Prigogine – is reflected in the centrality of free energy to the study of microbiological processes.[1] Enzymes, as key functional units of that timescale manipulation, can be understood as among the most fundamental units of biology.

Enzymes are extremely powerful tools, but the same property that gives them their strength is also the greatest barrier to harnessing them for other uses: exquisite chemical specificity. Intimately understanding the functioning of enzymes on a molecular level will be key to bending them more effectively towards technological needs, including but by no means limited to the manufacture of drugs and industrial chemicals,[2] the refinement of raw biomass into renewable fuels,[3] and the cleanup of environmental pollutants.[4]

Although the field of enzymology has existed at least since the 19th century,[5] the study of the molecular mechanisms of enzymes is a much younger field, owing to the lack of

means to study biology at the molecular level until the mid-20th century.[6] Koshland's famous 1958 "induced fit" proposal, whereby binding of a substrate to an enzyme induces structural changes that are in turn necessary for the enzyme to function, represents perhaps the earliest theory to gain traction by proposing an intimate and subtle relationship between enzymes and substrates.[7] This proposal arises naturally out of observations of the macroscopic behavior of enzymes; Koshland used as the basis for his reasoning the flexibility displayed by the ribosome in the formation of peptide bonds. However, viewed with an eye towards mechanisms on the molecular level, it quickly becomes clear that the structure-function relationships underlying enzymes operating under the induced fit paradigm must be exquisite indeed. Somehow, into the amino acid sequence of each enzyme must be encoded detailed physical "instructions" for the manipulation of every atom at every stage of the desired reaction – not to mention even more fundamental parameters such as the backbone fold, appropriate thermostability, and adjustments for pH conditions.

1.2 Dynamics and Transition Pathways

Owing to this extraordinary intricacy, traditional experimental tools are insufficient for answering questions regarding the detailed functioning of enzymes on the molecular scale. Although experimental kinetics studies provide some options for gleaning mechanistic information about enzymatic reactions without involving molecular structural data (and are indispensable for experimentally testing mechanistic hypotheses), more detailed discoveries are usually predicated on the analysis of three-dimensional enzyme models, such as those obtained by X-ray crystallography. However, even three-dimensional molecular models leave out what Koshland identified as the key element: dynamics. Studying the functioning of enzymes at this level motivates the field of molecular simulation.

Simulation is the best available tool if one is interested in probing the behavior of enzymatic *processes*, as opposed to merely individual slices in time. Because (as follows naturally from Koshland's insight) all of the most interesting features of enzymes are dynamic, taking a simulations perspective to enzymology is essential to obtaining a complete understanding. Simulations allow researchers to study the minutia of molecular motions over the course of a given process of interest.

Of course, not every possible dimension of motion is crucial to every enzymatic process. Indeed, the vast majority of the degrees of freedom in any given simulation are relatively unimportant in describing any given transformation. One of the most essential tasks for a molecular simulations researcher, then, is to somehow identify the key degrees of freedom, and to extract them from the surrounding noise during analysis. To this end, one of the key analytical concepts in molecular simulations is high-dimensional "phase space" within which every possible configuration of the system is represented by a single point. Molecular processes can be understood as pathways (or more exactly, ensembles of pathways) through phase space that connect different states (or, again, ensembles of states) of interest. Succinct descriptions of these pathways, and of the free energy profiles associated with them, are among the most valuable sorts of information that can be extracted from simulations.

1.3 Themes in This Work

The body of this work consists of three separate manuscripts prepared over the course of my doctoral training. Each draws from a set of themes that reflect the key concepts of this work as a whole:

(1) Explanations in terms of pathways

Rather than restricting analysis to static structures or snapshots, we will always seek to understand processes as fundamentally dynamic, and, when appropriate, as averages over ensembles. This is the fundamental strength of simulations, and pathway arguments have explanatory power that individual snapshots (even snapshots of simulations) lack.

(2) Reckoning with intricacy and human intuition

One major question in the context of simulations in general, but especially simulations of enzymes, is: what is the proper role of human intuition in guiding study? On the one hand, totally unguided simulation is generally highly inefficient for studying specific processes, and certain processes that are very slow compared to the timescale of a simulation are in fact impossible to study without some form of enhanced (biased) sampling. On the other hand, injecting too much "intuitive" bias into a simulation study may end up disguising the real underlying mechanics of the system; that is, it runs the risk of confirmation bias. A key theme in these works is the careful application of the *minimum* viable bias such that useful, interpretable results are produced without cutting out the underlying intricacy.

(3) Study of the unusual informs understanding of the usual

Here, we define "unusual" in terms of comparison to similar enzymes (unusual across an enzyme family), in terms of comparison to more typical behavior (unusual within a single enzyme), or even temporally (rare events in a time). In each case, however, we will see how study of what *stands out* is often the most useful tool for reaching a broader understanding.

CHAPTER II

Advantages of a Distant Cellulase Catalytic Base

2.1 Chapter Introduction

The first manuscript herein is also the most straightforward in concept: an unusual mechanism in need of an explanation. Although most cellulose hydrolyzing enzymes (or "cellulases") work by dividing a water molecule directly between the substrate and a catalytic base residue, the cellulase *Trichoderma reesei* Cel6A has no obvious basic residue in place to do so. Previous work by Mayes *et al.*[8] established that the mechanism instead involves a "wire" water molecule adjacent to the hydrolyzed water molecule, with the extra proton "conducted" through the water wire to the base. There is no obvious explanation for why this mechanism might've evolved, and the requirement that the wire water be coordinated in the transition state raises the free energy of the reaction compared to more traditional mechanisms. This raises the question: why does Cel6A go through the trouble?

The strategy in this work was based on noticing that nature had done something unusual, and trying to "undo" it by applying a mutation that one might expect to be more natural: a longer base residue that can reach the hydrolyzed water directly. We then set out to glean the underlying evolutionary "motivation" by investigating the reasons that the intuitively superior mutant doesn't work as well in practice. It also features a comparison to a different

enzyme that *does* follow the expected strategy. In this sense, we come at the analysis from two complementary perspectives: (1) looking from where the enzyme really is towards the expected model, and (2) looking from the expectation (that is, a similar enzyme that does meet our expectations about how such an enzyme should function) towards reality.

What results is a convincing rationalization of an unusual design. Mutating Cel6A reveals (by disrupting) an intricate set of finely tuned machinery for performing not only the reaction itself, but also the pre-reaction step of translocating the bound substrate into the active site, and the post-reaction step of shuttling a proton from the catalytic base to the catalytic acid in order to "reset" the enzyme for the next catalytic cycle. Furthermore, comparison to a more standard cellulase highlights the unexpected advantage that the distant catalytic base presents for releasing the reaction product from the active site.

Taken in the broader context of this thesis, this work serves as an introduction to all three of the key themes that will be recurring. First, **a pathway perspective**: notice while reading that the motivating "problem" of the unusual water wire mechanism only appears to be strange within the narrow perspective of the reaction step itself. Instead, explanatory power is found by broadening the context to include enzymatic steps both before and after the reaction along the full transition pathway, from before the substrate has even entered the reaction site, to the moment of its release. Second, **the limited value of intuition**: we highlight that while chemical intuition has a rightful place in guiding research into enzymatic mechanisms, it can also be misleading, as in this case, where the intuitive mutation is in fact deleterious. Finally, **the unusual informs the usual**: by comparing Cel6A's unusual mechanism to a more typical cellulase's, a more nuanced understanding of what each was "designed" to optimize is apparent. All of these themes will be formalized and developed further in future chapters.

Advantages of a Distant Cellulase Catalytic Base

Tucker Burgin, Jerry Ståhlberg, and Heather Mayes (2018). Reproduced with permission from the American Society for Biochemistry & Molecular Biology from the *Journal of Biological Chemistry* **293**(13): 4680–4687.

2.2 Abstract

The inverting glycoside hydrolase *Trichoderma reesei* (*Hypocrea jecorina*) Cel6A is a promising candidate for protein engineering for more economical production of biofuels. Until recently, its catalytic mechanism had been uncertain: the best candidate residue to serve as a catalytic base, Asp-175, is further from the glycosidic cleavage site than in other glycoside hydrolase enzymes. Recent unbiased transition path sampling simulations revealed the hydrolytic mechanism for this more distant base, employing a water wire; however, it is not clear why the enzyme employs a more distant catalytic base, a highly-conserved feature among homologs across different kingdoms. In this work, we describe molecular dynamics simulations designed to uncover how a base with a longer side chain, as in a D175E mutant, affects procession and active site alignment in the Michaelis complex. We show that the hydrogen bond network is tuned to the shorter aspartate side chain, and that a longer glutamate side chain inhibits procession as well as being less likely to adopt a catalytically productive conformation. Furthermore, we draw comparisons between the active site in *Tr*Cel6A and another inverting cellobiohydrolase to deduce the contribution of the wire water to the overall enzyme function, revealing that the more distant catalytic base enhances product release. Our results can inform efforts in the study and design of enzymes by demonstrating how counterintuitive sacrifices in chemical reactivity can have worthwhile benefits for other steps in the catalytic cycle.

2.3 Background

In order to tap into the deep reservoir of renewable energy represented by fuels derived from plant matter, an economical means of converting lignocellulose is required.[9] Decomposition of the primary component, cellulose, is catalyzed by glycoside hydrolase (GH) enzymes, which are found ubiquitously in nature;[10] therefore, improved catalytic efficiency of cellulose decomposition enzymes would help biomass to compete with non-renewable carbon sources. This motivates molecular-level studies into GH enzymatic mechanisms, as such understanding has previously proven invaluable in efforts to engineer variants with increased activities.[11–13]

A particularly important GH enzyme is *Trichoderma reesei* Cel6A (*Tr*Cel6A), which plays a key synergistic role in industrial enzyme cocktails for cellulose digestion. This enzyme is a cellobiohydrolase of GH family 6, which cleave β-1,4 glycosidic bonds processively along cellulose chains, from the non-reducing towards the reducing end, to release the glucose dimer cellobiose as the main product.[14] This processive mode of action is believed to be key to their efficiency on highly crystalline cellulose. Glycoside hydrolase family 6 (GH6) enzymes exhibit a range of activity on a continuum between cellobiohydrolase processive activity and endoglucanase activity, characterized by cleavage of internal bonds.[15–18] Endoglucanases generally exhibit higher catalytic rate constants, only show appreciable activity on less ordered regions of cellulose, and produce a broader range of products.[13] A crystal structure of *Tr*Cel6A was first solved in 1990,[19] but only recently has its molecular-level mechanism begun to be established. GH6 enzymes function via an inverting mechanism wherein the stereochemistry at the carbon of the cleaved β-1,4 bond is changed from equatorial to axial. The classical inverting mechanism (Figure 2.1, left) by which such reactions typically take place requires a catalytic acid-base pair (a proton

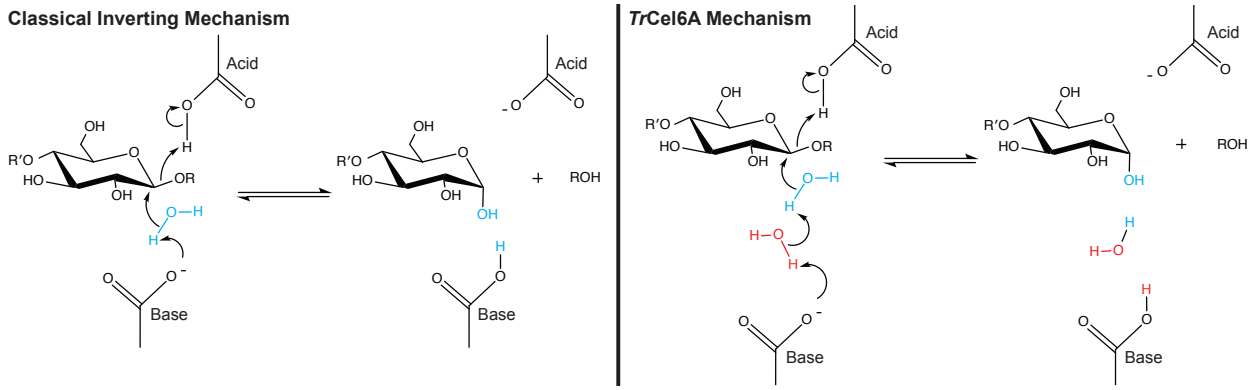

Figure 2.1: ***Tr*Cel6A vs. classical mechanism**

A comparison of the "classical" inverting cellulase hydrolase mechanism proposed by Koshland[20] (left) to that used by *Tr*Cel6A (right). The overall reaction chemistry is the same in each case, but *Tr*Cel6A requires the presence of the additional "wire water" in the active site due to the increased distance between the attacking water molecule and the catalytic base. The wire water does not change the overall chemistry. Atoms and bonds belonging initially to the nucleophilic water and the wire water are depicted in cyan and red, respectively.

donor and acceptor), the latter of which activates a nucleophilic water molecule during hydrolysis.[20] The identity of the catalytic acid in *Tr*Cel6A was identified experimentally as Asp-221; more recently, computational study confirmed the identity of the corresponding base as Asp-175.[8,21]

An interesting aspect of the mechanism revealed is that the catalytic step wherein Asp-175 accepts the excess proton requires an additional water molecule compared to the canonical mechanism, positioned between the basic carboxylate group at the end of Asp-175 and the nucleophilic (or "attacking") water, as shown in Figure 2.1 at right. This "wire" or "bridge" water molecule momentarily forms a hydronium (H_3O^+) ion during hydrolysis before offloading its excess proton to Asp-175.[8] The mechanism therefore requires the stabilization of two water molecules in the active site as opposed to only the one participating in the overall chemistry, raising the question as to why the active site of *Tr*Cel6A includes the additional water. Previous studies indicate that there is an additional energetic barrier on the order of 5 kcal/mol associated with each additional water wire in a Grotthus mechanism.[22] If the additional water molecule is only needed

to conduct a proton over the required distance, a mutant base with a longer side chain could obviate the need for the water wire. This suggests investigation of a *Tr*Cel6A D175E mutant, since glutamate (E) is identical in structure to aspartate (D) but for an additional CH_2 group in its side chain that could project its carboxylate group further into the active site. While no studies appear in the literature on a *Tr*Cel6A D175E mutant, mutation of the homologous residue in the GH6 enzymes *Thermobifida fusca* Cel6A (*Tf*Cel6A, at that time known as *Thermomonospora fusca* Endocellulase E2) (D79E) and *Cellulomonas fimi* Cel6A (formerly CenA) (D216E) has been shown to result in a decrease in activity relative to the wild type by approximately three orders of magnitude.[23,24]

In order to bridge the gap between atomistic enzymatic detail and human chemical intuition, we constructed molecular models of both wild-type and D175E *Tr*Cel6A and investigated the influence of the mutation on two non-reactive steps in the catalytic cycle: (1) procession of the substrate into the active site; and (2) the transition to the reaction-competent active site conformation following procession. We found in both cases that the longer Glu-175 residue in the mutant was a hindrance to the catalytic cycle, highlighting the importance and intricacy of the remarkable network of hydrogen bonding interactions that stabilizes the active site of the wild-type enzyme.

Finally, by means of comparison of the active site to that of a related endo-processive cellulase that functions via the classical inverting mechanism (*Tf*Cel9A), we offer an explanation for the *Tr*Cel6A mechanism in terms of reduced association between the product and enzyme. We propose a benefit in cellulases to activation of the nucleophilic water via a wire water by taking into account aspects of the enzymatic cycle outside the reaction itself, broadening the context for rationalizing enzymatic features in carbohydrate-active enzymes.

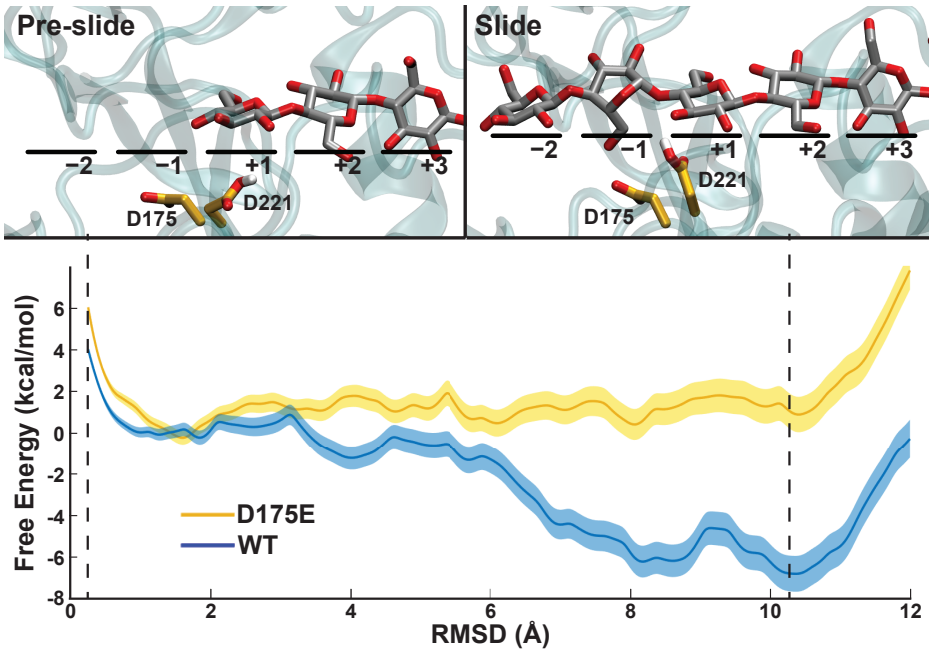

Figure 2.2: ***Tr*Cel6A substrate procession energy**

PMFs for substrate procession. Snapshots of substrate positions from the simulations with the wild-type enzyme are shown at RMSD values of 0.25 and 10.25 Å, respectively. The substrate binding sites -2 to $+3$ are labeled, along with two catalytic residues in the wild type. The RMSD compares the positions of the leading two glucose rings relative to the "pre-slide" structure (left). The dashed lines indicate the RMSD values of the "pre-slide" and "slide" conformations.

2.4 Results

2.4.1 D175E mutant: procession of the substrate into the active site

As a non-reducing-end cellobiohydrolase, each enzymatic cycle in *Tr*Cel6A requires the leading ring of the substrate chain advancing from the $+1$ site to the -2 site,[13] described as moving from the "pre-slide" to "slide" positions as shown in Figure 2.2 (at top). One hypothesis to explain the lower activity of the *Tr*Cel6A D175E homologue is that the bulkier side chain could protrude into the active site groove and hinder procession. To study this, we performed umbrella sampling simulations of procession using a CV that tracks the relative positions of the enzyme and substrate (simulation details available in the supporting material). The resulting potential of mean force (PMF) plots for both enzyme types are shown in Figure 2.2 (bottom), with the zero point on the free energy axis set at around 2 Å, where we did not expect the identity of the residue at position 175 to have a

large effect. We note that the CV used for sampling did not capture all key features that change during the transition from the pre-slide to slide positions. Specifically, as discussed in our previous study of *Tr*Cel6A wild-type procession, another key order parameter is the puckering of the second-to-leading glycosyl ring as it enters the -1 binding site, near 9 Å on the x-axis of the PMF in Figure 2.2.[8] Additionally, we found that the serine loop moves from a more-open to less-open position during procession, appearing coincident to the procession at approximately 6 Å on the x-axis of the PMF. The multidimensional CV required to properly sample all these (and potentially more) key feature changing during procession would be computationally prohibitive, and thus the PMF shown for the simplified CV (RMSD only) is most appropriately analyzed in terms of comparative qualitative, not quantitative, differences between the WT and D175E mutant (further discussion of this point is included in the supporting text). Specifically, we note that the PMF for the wild-type enzyme has a pronounced energy well that stabilizes the productive structure just after 10 Å, providing a driving force for spontaneous procession. In contrast, the D175E mutant retains a fairly flat energy profile. This effect can be at least in part attributed to steric clashes with several residues near the catalytic base, as shown in Figure 2.3. As shown, Asp-175 in the wild-type hydrogen-bonds with Arg-174 and Asn-182. This result suggests that one advantage of the shorter catalytic base is to enhance procession by widening the gap between the wall of the active site tunnel and the substrate.

2.4.2 D175E mutant: reaction-competent active site conformation

During hydrolysis, the catalytic acid loses a proton while the base obtains an excess proton. Before the next reaction, these residues likely exchange a proton when the acid (Asp-221) bends away from the reaction site towards the base (residue 175), as shown for the wild-type enzyme in Figure 2.4. After re-protonating, the acid must rotate along the

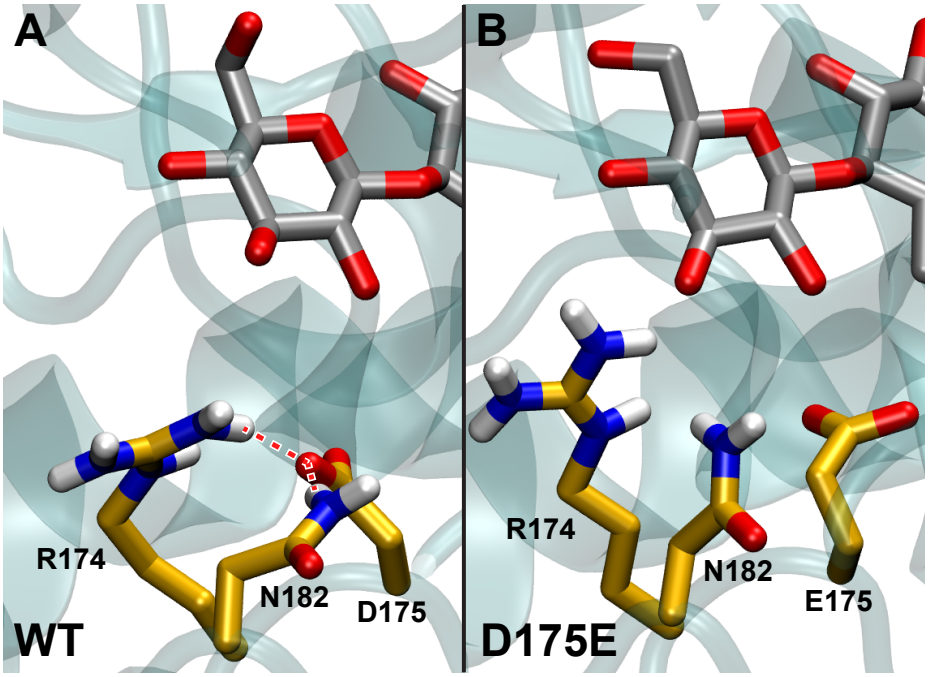

Figure 2.3: ***Tr*Cel6A D175E barriers to procession**

Snapshot of the *Tr*Cel6A wild type and D175E mutant mid-procession, corresponding to an RMSD value of 6.5 Å from Figure 2.2. While the wild-type base hydrogen bonds with nearby Arg-174 and Asn-182 to keep all three residues tucked away from the processing substrate, the added length of Glu-175 disfavors this binding and leaves the residues to clash with the leading ring.

dihedral angle indicated in the figure in order to position its proton toward the glycosidic oxygen and realign itself for the next catalytic event.[13,21]

The energy landscape in Figure 2.4 shows the barriers for the wild type and D175E mutant for this transition, with the acid and base residues further apart at the larger dihedral angles. The PMFs are similar, with the difference in barrier heights no greater than 1 kcal/mol. However, as shown in Figure 2.5, the wild-type and mutant conformations corresponding to the right-hand side energy wells in Figure 2.4 have significantly different hydrogen bonding networks. In the wild type, the hydrogen bond between the acid and base residues is broken, and the acid instead hydrogen-bonds with the glycosidic oxygen, in a reactive conformation for hydrolysis.[8] In the mutant, the base often remains hydrogen-bonded to the acid, rather than with the active-site "bridge" water, preventing formation of the hydrogen-bonding network that aligns the active site waters for hydrolysis. When

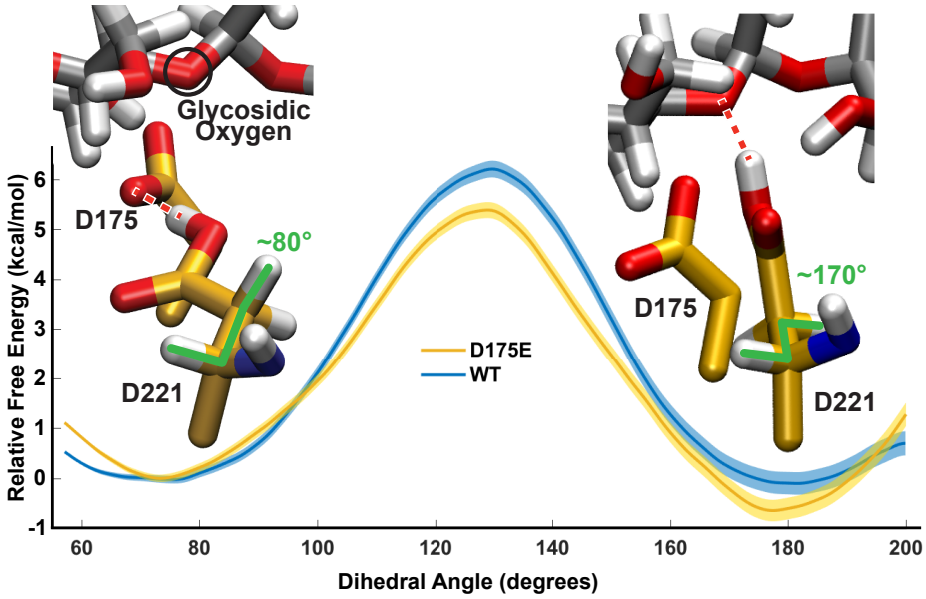

Figure 2.4: ***Tr*Cel6A proton transfer dihedral**

PMFs from umbrella sampling for dihedral rotation around the Asp-221 Cα-Cβ bond (dihedral angle indicated in green in the snapshots shown from the wild-type enzyme simulations), representing the transition from the conformation for catalytic base/acid proton transfer (at around 80°) toward the Asp-221 position for glycosidic cleavage (near 170°).

unbiased simulations were run with the acid initially at a dihedral angle of 175°, the productive hydrogen bond between the Asp-221 proton and the glycosidic oxygen was never observed over the 1-ns trajectory in the D175E mutant, compared to roughly 5% of the frames in the wild type.

In a separate simulation, the acid-base hydrogen bond in the conformation shown in Figure 2.5(B) (between Asp-221 and Glu-175) remained stable during 972 ps of simulation, at no point bonding instead with the attacking water. Efforts to obtain such a conformation indicated that it does not occupy a local energy minimum. As shown in Figure 2.6, the energy barrier associated with this active conformation was quantified using umbrella sampling with restraints applied to the dihedral angle connecting the β- and γ-carbons. This parameterization approximates the motion that the residue undergoes during unbiased simulation initiated near the high-energy region around 40°. As shown, the free energy trough in this region is shallow and readily degenerates to the inactive state at $-50°$, and

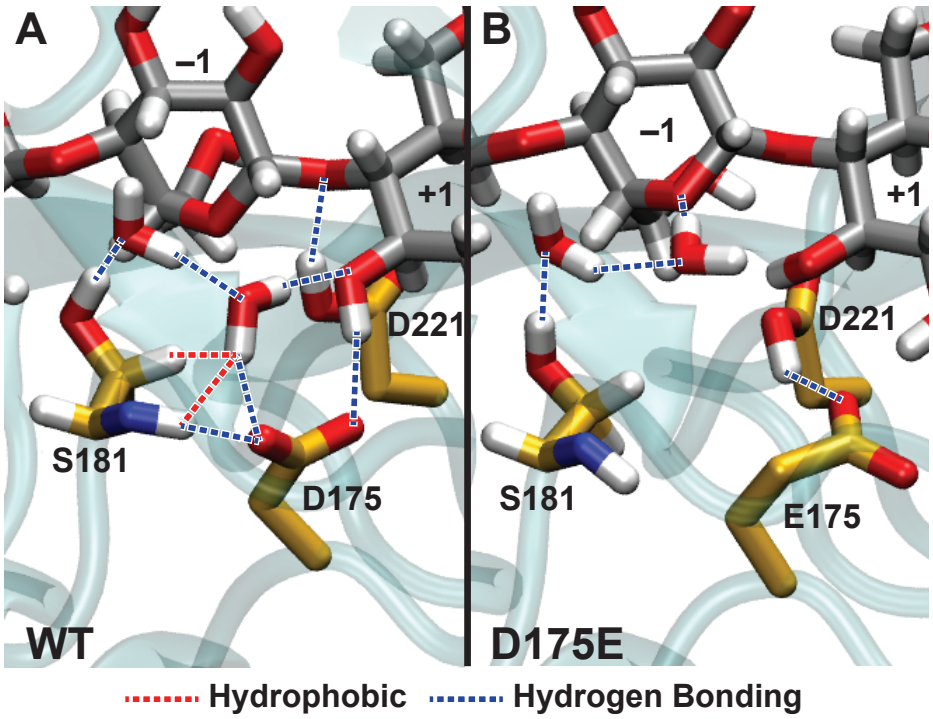

Figure 2.5: ***Tr*Cel6A mutated vs. WT active site**

Diagram of the active site of *Tr*Cel6A for the wild type and D175E mutant. An exquisite network of hydrogen bonding interactions stabilizes the wire water between the attacking water (hydrogen bonding with Ser-181 in both cases) and the wild-type base, Asp-175. The longer side chain of the mutant Glu-175 residue compels it out of the active site, chasing after the void left by the acid Asp-221 after the rotation described in Figure 2.4 and resulting in an inactive conformation.

an even lower-energy state is available at -170° (leftmost energy well in Figure 2.6, left). A total free energy activation barrier of roughly 7.3 kcal/mol separates the active state from the low-energy state at -170°, posing significant hindrance to hydrolysis.

The simulations used to construct the PMF in Figure 2.6 began from a structure with Asp-221 initially in a high-dihedral angle conformation from Figure 2.4, which in the wild type aligns the Asp-221 carboxylic proton with the glycosidic oxygen. However, in the mutant these atoms were not consistently aligned, indicating that 7.3 kcal/mol is an underestimation of the barrier to reactive alignment. Additionally, even when the attacking water is aligned by Glu-175, its lone pair is less-favorably oriented for nucleophilic attack as compared to the wild type. Considering these factors and assuming that the lack of a water wire reduces the hydrolysis barrier by 5.0 kcal/mol,[22] the net effect of the mutation

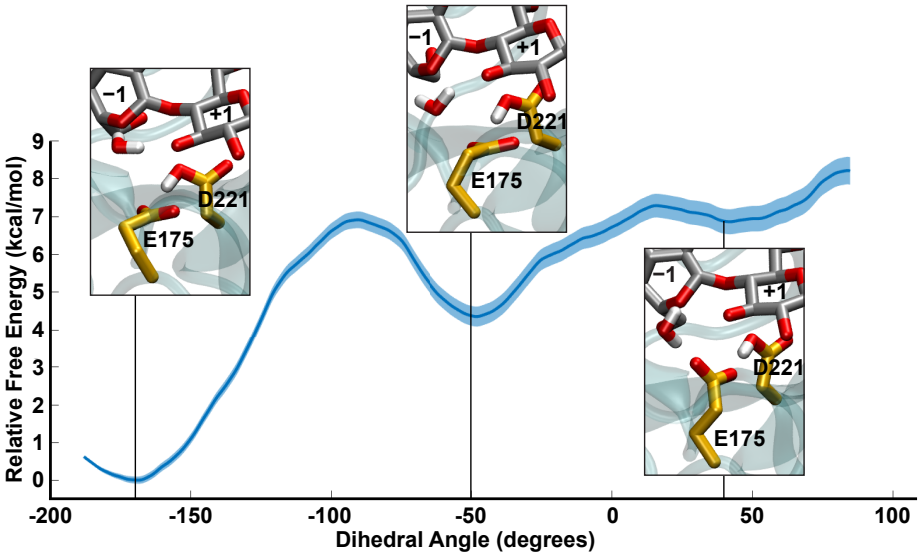

Figure 2.6: ***Tr*Cel6A catalytic acid backbone dihedral**

PMF for rotation of dihedral around the bond between the β- and γ-carbons in the mutant Glu-175 residue. Three distinct minima are observed at $-170°$, $-50°$, and $40°$. The minimum highest in free energy corresponds to a potentially hydrolytically active conformation, with Glu-175 hydrogen bonding with the nucleophilic water as shown in the corresponding snapshot.

would be to increase the barrier by a conservative (low) estimate of at least 2.3 kcal/mol. In sum, we find that although Glu-175 may adopt a conformation where it could accept a proton directly from the nucleophilic water (and thus may act as catalytic base in a classical inverting mechanism), activity in this mutant is lower than in the wild type because (a) the "active" conformation of Glu-175 is disfavored, (b) the lone pair of the attacking water is less-favorably oriented for nucleophilic attack at the anomeric carbon, and (c) the catalytic acid Asp-221 is less able to dissociate from Glu-175 (as compared to Asp-175) to enable protonation of the glycosidic bond.

The finding that a glutamate at position 175 disrupts the hydrogen bonding network that catalyzes hydrolysis indicates that active site of *Tr*Cel6A is tuned for the shorter, highly-conserved aspartate.[13] Several other GH6 enzymes have been crystallized, including *Tf*Cel6B (PDB ID: 4B4F[25]), *Chaetomium thermophilum* Cel6A (PDB ID: 4A05[26]), *Humicola insolens* Cel6A (PDB ID: 1BVW[27]), and *Tf*Cel6A (PDB ID: 2BOD[28]). The distances between

the glycosidic oxygens and catalytic bases, as well as the presence of two active site waters in structures apparently primed for hydrolysis, indicate that members of this family generally perform hydrolysis via a Grotthuss mechanism to a more distant catalytic base. Our simulations clearly indicate why a longer catalytic base decreases activity; however, the question remains as to why the active site is not tuned for a longer side chain, or why the aspartate is not positioned closer to the cleavage site.

2.4.3 Active site homology

To better understand why the *Tr*Cel6A active site is tuned to require a wire water, we compared its active site to that of another inverting glycoside hydrolase, *Tf*Cel9A, a processive endocellulase that is believed to employ the classical mechanism based on crystallography and site-directed mutagenesis studies.[29,30] We created a *Tf*Cel9A model based on a product-state crystal structure (PDB ID: 4TF4[31]) to compare with a product-state structure produced for *Tr*Cel6A in the course of our previous work,[8] as shown in Figure 2.7. Panels A and B highlight differences in the hydrogen-bonding network in the product states. In both cases, hydrogen bonding interactions that helped stabilize the nucleophilic water in the appropriate position for catalysis pre-reaction become hydrogen bonds to the product monomer in the -1 position post-reaction. Specifically, the product hydrogen bonds with Ser-181 and the "bridge" water (in turn hydrogen-bound to Asp-175) in *Tr*Cel6A, and to Asp-55 as well as directly to Asp-58 in *Tf*Cel9A. *Tr*Cel6A Asp-175 and *Tf*Cel9A Asp-58 are the base residues while *Tr*Cel6A Ser-181 and *Tf*Cel9A Asp-58 serve to stabilize the nucleophilic water in the reactant state.

Viewed from the reducing end of the substrate, an interesting geometric distinction between the two active sites becomes apparent as depicted in panels C and D. Although *Tf*Cel9A employs a glutamate residue instead of an aspartate as its catalytic acid, the

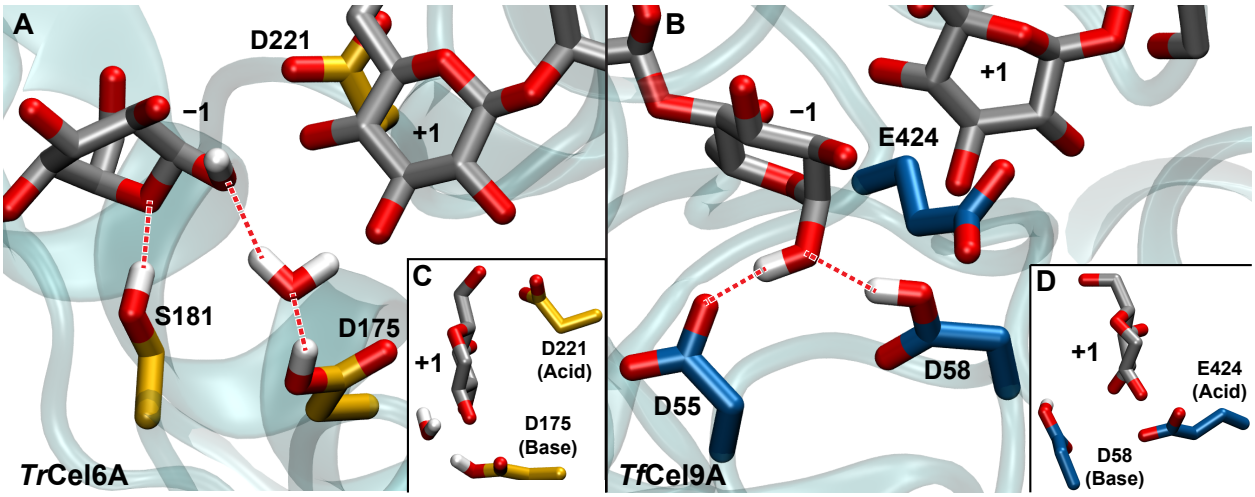

Figure 2.7: ***Tr*Cel6A comparison to homologous active site**

Comparison of the product-state active sites of *Tr*Cel6A (A and C) and *Tf*Cel9A (B and D). In A and B, the dashed lines highlight specific hydrogen bonds further discussed in the text. Panels C and D show the same conformation from a different orientation, to highlight the differences in relative positions of the product and acid/base residues.

acid/base pairs in these enzymes are in similar relative positions to one another (with 5.3 Å of separation between the carbon atoms in their carboxylate head groups in the relaxed product state), and the relative positions of the acid and base to the axis of the substrate in each enzyme are the same (making about a 90° angle in each case). However, the whole of the catalytic machinery of *Tr*Cel6A is rotated roughly 45° about the substrate. Because of the oblong cross-sectional shape of the substrate, the *Tr*Cel6A base is further from the -1 anomeric carbon than in *Tf*Cel9A (6.0 versus 4.1 Å in the product state, respectively), requiring the addition of the wire water molecule to connect the base with the attacking water. While this discussion focuses on the product state, for which a crystal structure of *Tf*Cel9A is available, the same conclusions should hold in the transition state structures based on the positions of the acid residue α-carbons.

One advantageous function of the water wire could be to destabilize the newly formed product in the -1 position after hydrolysis. We quantified this effect by measuring the contribution to the overall free energy of binding from the base residue in each enzyme. To

focus on the effect of the mediating water wire alone, not obscured by other differences in the enzymes, we calculated the total change in binding free energy associated with mutating the bases to alanine, as detailed in the supporting text. The base-to-alanine mutations remove all of the polar interactions with the catalytic bases, so the difference between the total binding energy in the mutant and wild type can be interpreted as the contribution only from the bases. As expected, the *Tf*Cel9A base (which binds directly to the substrate) has a larger contribution to the binding energy (~7.8 kcal/mol) compared to that of the base in *Tr*Cel6A (~2.1 kcal/mol). This result indicates that the GH6 enzyme greatly enhances product release by using the wire water as a buffer between the base and product.

Interestingly, the hydrogen-bonding network in the *Tr*Cel6A active site stabilizes the ^2S_O pucker of the -1 ring for the product conformation. In *Tf*Cel9A this ring is relaxed to the low-energy $^4C_1$ chair, and its C1 hydroxyl group hydrogen bonds directly with the catalytic base Asp-58. This conformation is further stabilized by hydrogen-bonding between the C1 α-hydroxyl and the nucleophilic water-stabilizing Asp-55 in *Tf*Cel9A, while the homologous residue in *Tr*Cel6A, Ser-181, stabilizes a puckered product because the latter pulls down the -1 ring oxygen. This pucker is known to promote reactivity in *Tr*Cel6A and occurs spontaneously as the second-to-leading glycosyl ring enters the -1 binding site.[8] QM calculations of monosaccharides have shown that the energetic cost for puckering an α-glucose (as in the product) in the ^2S_O orientation is approximately 4 kcal/mol greater than puckering a β-glucose (as in the reactant) in the ^2S_O orientation.[32] Holding the product in a puckered state may further promote product release by stabilizing an unfavorable conformation while bound.

In our aforementioned simulation of the *Tr*Cel6A base mutated to alanine, the -1 sugar became free to transition between skew puckers ^2S_O and $^1S_3$. Because the product is in

the α-glucose configuration, this transition stabilizes the bound product by 3 kcal/mol.[32] This result points to an additional role of the base (indirectly, through the wire water) in stabilizing the unfavorable ${}^{2}S_{O}$ pucker in the product state and promoting product release.

2.5 Discussion

The catalytic mechanism of *Tr*Cel6A involves a counterintuitive wire water molecule that is not strictly necessary to the overall chemistry of the enzyme's reaction. Motivated by a desire to rationalize this exception to the classical mechanism for an inverting glycoside hydrolase, we disrupted the active site with a D175E mutation that we hypothesized would supplant the wire water. We found that the active site is stabilized by a series of interconnected hydrogen bonds for which the wild-type base Asp-175 is perfectly suited, whereas the mutant Glu-175 was less favorably aligned for either substrate procession and hydrolysis. During procession, only the wild-type base promoted hydrogen bonding that kept Arg-174 and Asn-182 out of the tunnel, contributing to an energetically favorable, spontaneous forward motion for the wild-type enzyme, which was not manifest in the mutant. In aligning for hydrolysis, hydrogen bonding in the mutant active site favors positioning the mutant carboxylate group even further from the nucleophilic water than in the wild type, which, compared to it, both makes accepting a proton more difficult and prevents it from helping align the nucleophilic water for attack. Calculations of low-energy conformations of the system in the hydrolysis-ready position with the -2 and -1 positions occupied by the cellulose chain indicate that the overall reaction barrier in the D175E mutant would be increased by at minimum 2.3 kcal/mol, which is approximately consistent with the activity reductions observed for the *Tr*Cel6A D175E-homologous mutants *Tf*Cel6A D79E and *Cf*Cel6A D216E.[23,24]

To further understand the role of the more distant base in *Tr*Cel6A, we compared its

active site to that of an inverting cellulase, *Tf*Cel9A, which does not require a water wire to shuttle a proton to the base during catalysis. While the active site residues align similarly in the two enzymes, the substrate is rotated approximately 45° about its axis, creating the additional distance occupied by the "bridge" water in *Tr*Cel6A. We propose that the wire water buffers the post-reaction hydrogen bonding between the protonated base and the cellobiose product, easing product release. Consistent with this theory, we calculated an approximately 6 kcal/mol reduction in product binding energy due to differences in catalytic base alignment strategy.

Our results advance cellobiohydrolase enzyme engineering efforts by broadening the focus of the role of active site residues beyond the hydrolysis step. Family GH6 enzymes (and the conservation of their mechanism across different branches of life) expand our understanding of how enzymes can make small sacrifices in reactivity to enhance other aspects of the larger catalytic cycle.

2.6 Experimental Procedures

2.6.1 Molecular dynamics simulations

The approach of our investigation using molecular dynamics (MD) simulations is briefly described here, with further detail in the supplemental material. All models were based on crystal structure deposited in the Protein Data Bank (PDB).[33] The initial structure for the procession study was constructed by combining features from two crystal structures. First, the crystal structure PDB ID: 1QK2[34] (wild-type *Tr*Cel6A with a non-hydrolysable cellotetraose) was stripped of its non-protein components. Then, the cellohexaose substrate from PDB ID: 4AVO[35] was manually aligned with its leading non-reducing-end ring in the $+1$ position as in our previous work.[8] For the active site conformation studies, we started with the PDB ID: 1QJW[34] crystal structure (*Tr*Cel6A Y169F mutant complexed

with cellotetraose) reverted to the wild type, and again replaced its substrate with that of PDB ID: 4AVO,[35] this time in the active position with the leading ring in the -2 site. The crystal structures in PDB format were converted into topology and coordinate files using the CHARMM[36] package and the CHARMM36 forcefield.[37–40] The models were solvated in a periodic box using the TIP3 water model[41] and converted into Amber format using the CHAMBER program in ParmEd.[42]

All simulations were performed using the Amber14 package.[43] Unless otherwise stated, the SHAKE algorithm[44,45] was used to constrain the lengths of all hydrogen bonds and the cutoff distance for non-bonded interactions was set to at least 8.0 Å. First, structures were minimized without SHAKE over 2500 steps (1250 steepest-descent method followed by 1250 conjugate gradient method). The systems were then heated in a periodic box in the NVT ensemble, using the Andersen thermostat[46] to take the temperature from 100 K to 300 K over 10,000 2-fs time steps followed by 1000 steps of constant-temperature simulation. Velocities were randomized every 1000 steps. MD production simulations were performed under the same conditions as the heating simulations at a constant temperature of 300 K.

Simulations of *Tr*Cel6A included positional restraints on the α-carbons of residues Ser-106, Ala-150, Asp-200, Asn-247, Ala-280, Ile-330, and Gln-437, as in our previous work.[8] These atoms have been shown to have a low root mean square fluctuation,[47] and restraining them prevents bulk motion of the protein.

Models were visualized using VMD.[48] The PMF plots were produced using the umbrella integration method introduced by Kästner and Thiel[49,50] and implemented by Stroet and Deplazes.[51]

CHAPTER III

Mechanism of Oligosaccharide Synthesis via a Mutant GH29 Fucosidase

3.1 Chapter Introduction

In contrast to the previous paper, which sought to rationalize a known mechanism, this work centers on an enzyme whose reaction mechanism was not known in advance. The enzyme in question this time is a mutant of the fucose hydrolyzing enzyme *Thermotoga maritima* α-L-fucosidase, or *Tm*Afc. Unlike most mutations, which either have no identifiable effect or simply modulate the efficiency of the enzyme's native reaction, the mutation in question – D224G – has the highly unusual property of changing the dominant reaction altogether. *Tm*Afc D224G is a member of a class of enzymes called "glycosynthases": engineered glycoside hydrolases that have been mutated to disable their native hydrolytic activity in favor of new glycosyltransferase activity involving a native leaving group on the substrate.

Although many aspects of glycosynthases are poorly understood in general, *Tm*Afc D224G is unusual even among glycosynthases. Whereas most functioning glycosynthases that have been identified to date use a flourine atom as the reaction's leaving group, the reactions best catalyzed by *Tm*Afc D224G involve a departing azide (N_3) moiety instead. Furthermore, fucosynthases such as this one are rare; most glycosynthases discovered to

date are specific for other types of sugars. As there is significant bioengineering interest in developing tools for specific oligosaccharide synthesis, characterizing and understanding the full scope of glycosynthase reactions is a necessary precursor to developing more (and more efficient) enzymes for this purpose.

The central contribution of this work is a full, unbiased transition path sampling study into the mutant reaction mechanism. As in the previous chapter, all three recurring themes play a role here: **(1)** a full view of the transition pathway plays a central role; **(2)** the reaction mechanism is obtained with the minimum possible human bias by specifying in advance nothing except definitions of the reactants and products; and **(3)** the key result of the work is a study into exactly how something so unusual as a complete change in the dominant reaction scheme of the enzyme can arise by repurposing its existing machinery. The minimum bias workflow used to obtain the results in this paper is highly generalizable, and automating it will be the topic of the following chapter.

This version contains corrections not reflected in the published manuscript, regarding the pathway free energy method used to obtain the reaction energy profile. The original work used a method called equilibrium path sampling (EPS), which is a highly general (but highly inefficient) method of measuring energy along an arbitrary molecular pathway. Although our analysis at the time led us to believe that the EPS data was properly converged around the equilibrium distribution, later analysis after publication revealed that key processes on a much longer timescale are in fact dominant. The version that appears here replaces the EPS results with updated results using umbrella sampling, which is dramatically more efficient, but had not been developed for use with complex reaction coordinates such as the one that appears herein until after the original publication. The result is much stronger and in better agreement with available experimental data than what appears in the published version.

Mechanism of Oligosaccharide Synthesis via a Mutant GH29 Fucosidase

Tucker Burgin and Heather Mayes (2019). Reproduced with permission from The Royal Society of Chemistry from *Reaction Chemistry & Engineering* **4**: 402–409. Contains unpublished corrections.

3.2 Abstract

Techniques for synthesis of bespoke oligosaccharides currently lag behind those for other biopolymers such as polypeptides and polynucleotides, in part because of the paucity of satisfactory enzymatic tools to perform the synthetic reactions. One promising avenue of development for this problem is glycoside hydrolase enzymes with mutated nucleophile residues (called glycosynthases), which retain some elements of their native specificity and work with cheaply available substrates. However, the mechanistic underpinnings of this class of enzymes are not yet well-understood, and what few atomistic studies have been conducted have found different reaction pathways. In this paper, we describe the first unbiased computational study of the mechanism of a GH29 glycosynthase enzyme, *Thermotoga maritima* α-L-fucosidase (*Tm*Afc) D224G. We find a single-step endothermic reaction step with an oxocarbenium-like transition state, demonstrating how stabilization of this transition state structure (which is common to many retaining glycoside hydrolases) can be repurposed in mutant enzymes to perform synthesis instead of hydrolysis. Our results are consistent with experimental observations and help both to clarify the mechanism of the existing single-mutant and to provide directions for further engineering of this and other glycosynthases.

3.3 Introduction

Oligosaccharides have long been known to play a wide variety of important roles in biology,[52] from structural support to signaling cascades and mediation of cell-cell interactions — as one 1993 review paper put it: "all of the theories are correct."[53] Futher understanding of the properties and roles of particular oligosaccharides requires synthesis of homogeneous samples in sufficient quantities for research studies. Unfortunately, techniques for synthesizing oligosaccharides have lagged significantly behind those for other biological polymers, owing in part to the complexities of regio- and stereochemistry.[54,55] Though significant progress has been made since the 1990s, the variety of methods that have been developed are narrow in their applicability, usually taking place over many successive steps with a loss of conversion at each step, and require meticulous control over the reaction conditions to minimize competing off-pathway reactions.[56–58]

The natural alternative to arduous organic synthesis routes is the use of enzymes to catalyze highly specific glycosynthetic reactions. Enzymes remove the need for careful protection and deprotection steps or the tuning of highly sensitive reaction parameters. Unfortunately, the enzymes evolved in nature to perform these reactions, glycotransferases, are largely not amenable to biotechnological applications because of their low stability outside the cell and reliance upon expensive nucleotide-sugar substrates.[59] Although strategies to circumvent these issues are in development, such as the recycling of reacted nucleotides or directed evolution of the enzymes to accept more readily available precursors, this has proven to be a non-trivial problem.[60,61]

Other enzymatic alternatives are mutant glycosidases, dubbed "glycosynthases."[55,62] To modify a two-step retaining bond cleavage mechanism, the nucleophilic residue responsible for forming the stable intermediate is mutated into a non-reactive residue in order to

preclude the forward reaction, leaving the catalytic base (Glu-266) intact and able to aid synthesis of a glycosidic bond between suitable glycosyl donor and acceptor molecules. Such enzymes provide a powerful framework for building oligosaccharides from readily available substrates (employing much simpler leaving groups than nucleotides, such as azide groups or fluorine atoms).[63] Glycoside hydrolases are more stable and soluable than glycosyltransferases and thus more amenable to *in vitro* and industral conditions. However, because they are evolved for a different reaction pathway, they lack the high specificity and efficiency characteristic of most wild-type enzymes. Furthermore, much of the work to date on the discovery of new glycosynthases has been a series of shots in the dark: it is not well-understood which nucleophile mutations applied to which glycoside hydrolases will produce an active glycosynthase or why, and the most successful work for obtaining new or improved glycosynthases has relied on random, semi-random, or otherwise exploratory approaches.[64–66]

These shortcomings motivate attempts to rationally engineer glycosynthases to produce a given oligosaccharide with high specificity and efficiency. This endeavor will require a clear understanding of the reaction mechanism at the atomic level. However, although glycosynthases have been present in the literature for over 20 years,[62] investigations into the atomistic underpinnings of these mutants' reactions have been scant. One 2013 study by Zhang *et al.* described a metadynamics study on the *Humicola insolens* Cel7B cellulase E197S mutant in its glycosylation reaction between α-lactosyl fluoride and the flavonoid luteolin.[67] In that same year, Wang *et al.* performed partitioned-rational function optimization calculations to find energy minima and maxima in the cellulose synthase reaction of rice BGlu1 β-glucosidase mutants E386G, -S and -A.[68] Although both papers help to clarify the mechanisms of their respective reactions, taken together they underscore how individual studies do not capture the complete picture of how this class of

Figure 3.1: ***Tm*Afc reaction schematic**

The *Tm*Afc D224G reaction of interest in this work. This reaction was experimentally studied by Cobucci-Ponzano *et al.* in 2009.[66] As reported there, this reaction produces α-1,4 and α-1,3 products in nearly equal amounts, and has an overall specificity of 91% for transfer of the donor (4NX) to the shown acceptor (1AF) versus water.

enzyme functions. For example, although the transition state search algorithm employed by Wang *et al.* identified only a single step in the BGlu1 mutant reaction mechanisms, Zhang *et al.* describe a three-step reaction with stable intermediates in *Hi*Cel7B E197S. Furthermore, although in Zhang *et al.* the mutant serine participates in the reaction by stabilizing the leaving fluorine, the corresponding interaction in BGlu1 E386S was shown to be less favorable compared to the lack of interaction of E386G. Because of these stark disagreements, it is clear that further study is required in order to gain further understanding of these promising enzymes.

Herein we present a validated reaction mechanism for the transglycosylation reaction between 1-azido-β-L-fucose (1AF) and 4-nitrophenyl β-D-xylopyranoside (4NX) catalyzed by *Thermotoga maritima* α-L-fucosidase (*Tm*Afc) D224G, as diagrammed in Figure 3.1.[66] Significantly, we present the first study of the complete reaction pathway of a glycosynthase performed strictly using methods that do not bias the Hamiltonian by introducing non-physical energy terms into the model. Because the only decision made in our study that

biases the discovered reaction pathway is the choice of reactant and product definitions (as discussed in the Computational Methods section), this result represents the most rigorous computational study of a glycosynthase reaction to date.

*Tm*Afc D224G has been shown to retain the overall protein fold of the wild-type enzyme, indicating that its glycosynthetic activity arises directly as a result of the change to the active site.[66] This enzyme differs from those in previous computational studies (and most studies of glycosynthases in general) in that it uses α-fucosyl donors (compared to the α-lactosyl and α-glucosyl donors of Zhang *et al.* and Wang *et al.*, respectively) with azide leaving groups instead of fluorine atoms. Despite the underrepresentation of fucosylsynthases in the literature, fucosyl oligosaccharides are of particular interest in biomedical applications, including as anti-cancer and anti-inflammatory drugs.[69] They are also major constituents in human milk oligosaccharides present in natural breast milk and implicated in healthy gut microbiota development, but typically not included in infant formula.[70–72]

The results of our molecular models are in good agreement with the experimental observations in Cobucci-Ponzano *et al.* with respect to the relative abundances of the two isomeric products shown in Figure 3.1.[66] They reveal a one-step, endothermic reaction mechanism, wherein the dissociation of the leaving azide from the electron donor occurs in concert with the transfer of the hydrogen from the acceptor to the catalytic residue, as well as with the bond formation between the donor and acceptor, through an oxocarbenium-like transition state. Our results explain the experimental observation that, unlike in the D224G mutant, fucosidase activity in the *Tm*D224S mutant cannot be rescued by the addition of free azide,[66] and also provide clues for rational engineering of this and similar enzymes in the future.

3.4 Computational Methods

3.4.1 Model building

The *Tm*Afc model was based on the crystal structure PDB ID: 2ZXD,[73] chosen over the two earlier-published crystal structures (PDB IDs: 1ODU[74] and 2WSP,[66] respectively) because they both contain incomplete loops near the active site, and because the resolution of 2ZXD is considerably better than the next most recent (2.15 Å vs. 2.65 Å). Although the 2ZXD structure is complexed with an inhibitor molecule, the protein backbone overlays very closely with that of 1ODU and 2WSP (complexed with fucose and α-L-Fucose-(1-2)-β-L-Fucosyl-Azide, respectively), indicating that complexing with the inhibitor does not result in any major conformational changes.

The substrate models had to be custom-built for this study, as parameters for neither 1AF nor 4NX were available. In both cases, parameterization began with the Generalized Amber Force Field (GAFF),[75] to which appropriate GLYCAM06 parameters[76] and other custom parameters were added as needed to obtain qualitatively reasonable agreement between minimized structures obtained using the custom force field and quantum mechanical SCC-DFTB calculations.[77,78] For 1AF, the additional parameters required were those for the azide group and its connection to the sugar, and were taken from Carvalho *et al.*[79] and Weiner *et al.*,[80] respectively. For 4NX, one parameter was calculated directly using Gaussian 16 Rev. A.03,[81] using the B3LYP quantum mechanics model[82–85] with the 6-311+G(d,p) basis set, which has been well validated for systems similar to this one.[86–89] Comparisons between the parameterized molecular mechanics models and the quantum mechanical models, as well as documentation of all the parameters added to GAFF to build the custom force fields, are available in Appendex B.

To insert the substrates into the active site of the enzyme, first the donor 1AF was overlaid atop the inhibitor present in the crystal structure using the RMSD Visualizer tool in

VMD,[48] taking advantage of the six-membered ring structure shared between the two. The acceptor 4NX was inserted manually into the open cleft nearby the donor in such a way as to place its sugar's O4 close to the donor's C1 atom (in anticipation of the bond that forms between them). This structure was solvated in a box of TIP3P water molecules[41] such that there was everywhere at least 10 Å between the solute and the edge of the box, and one sodium counterion was added to neutralize the charge of the overall system. The model was minimized with Amber 16[90] over 2500 steps and then heated from 100 K to 300 K over 10,000 steps (followed by 1000 additional steps at 300 K) using the Andersen thermostat[46] to randomize velocities every 1000 steps in an NVT ensemble. Finally, the structure was equilibrated with molecular mechanics (MM) for 10,000 steps with isotropic pressure scaling turned on (NPT ensemble) and velocity randomization every 100 steps. A step size of 2 fs and a cutoff distance of 8.0 Å were used throughout, and the SHAKE algorithm[44] was applied to restrict the covalent bond lengths of hydrogen atoms during heating and equilibration.

3.4.2 Transition path sampling

Our transition path sampling (TPS) methodology is divided into several steps: transition state hypothesizing, aimless shooting,[91] likelihood maximization,[92] committor analysis,[93] and umbrella sampling.[94] Taken together, they represent a method of sampling the transition state ensemble without biasing the Hamiltonian, obtaining a reaction coordinate from that sample, verifying the transition state described by the resulting reaction coordinate, and then measuring the free energy surface along that reaction coordinate.

In the remainder of this section, we will detail our methodology for building the models and performing the aimless shooting step, which is responsible for producing the data that the following steps (likelihood maximization, committor analysis, and umbrella sampling)

were used to analyze. The methodologies for those analysis steps can be found in Appendex B.

Transition state hypothesizing

Aimless shooting requires at least one (and preferably more) initial structure(s) close to the separatrix (the surface in phase space along which any trajectory with randomly selected velocities for all atoms will have an equal chance of collapsing to the product state or to the reactant state.) Because we don't have an *a priori* definition of the separatrix, hypothesized transition states are created by changing the distances between the atoms involved in either formation or cleavage of bonds during the reaction of interest, to a range of distances between those observed in the reactants and products. In the case of the reaction at hand, these bond lengths (and in brackets the corresponding distances tested in Å) were those between: (1) the 4NX O4 hydrogen and closest oxygen of Glu-266 (the catalytic base) [1.1, 1.2, 1.3, 1.4]; (2) that same hydrogen and the 4NX O4 itself [1.1, 1.2, 1.3, 1.4]; (3) the 4NX O4 and the 1AF C1 atoms [2.1, 2.2, 2.3, 2.4]; and (4) the 1AF C1 atom and the primary nitrogen of the azide group [2.5, 2.6, 2.7, 2.8]. Structures with each combination of the given bond distances were built using restraints in combined quantum mechanics and molecular mechanics (QM/MM) simulations with SCC-DFTB.[77,78] This semi-empirical quantum mechanical model was selected for its good compromise between speed and accuracy.[95,96] Further details of these simulations are available in Appendex B. Combinations with more than one extreme value (defined as either the largest or smallest allowable value for a given bond length) were omitted to reduce computational expense, resulting in a total of 80 starting conformations.

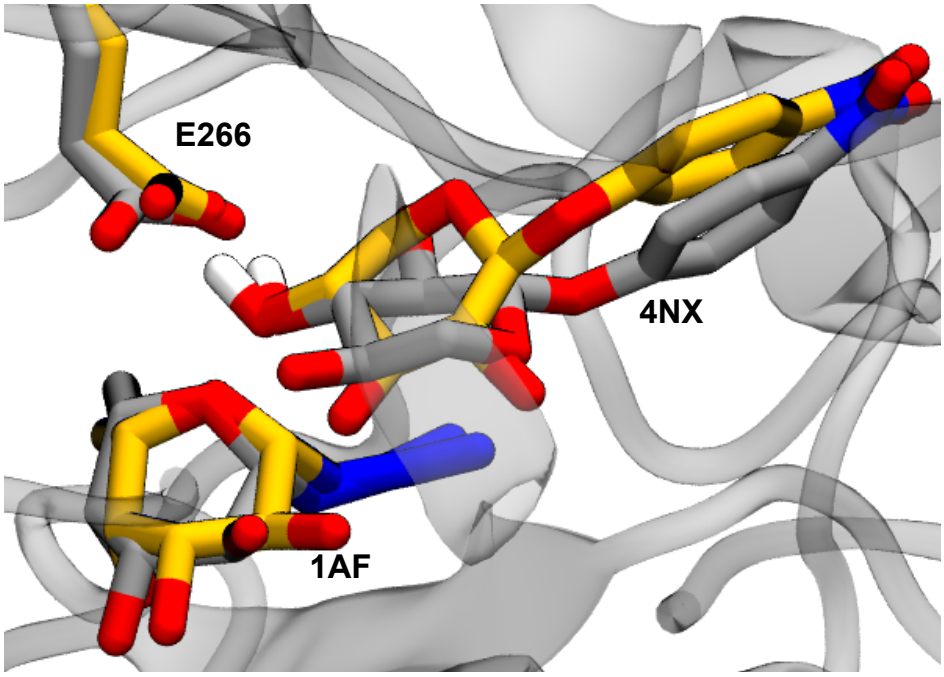

Figure 3.2: ***Tm*Afc comparison of reactant states**

Snapshots from the ensemble of reactant state structures for the α-1,4 (gold) and α-1,3 (silver) reactions. The key features – namely, the positioning of the acceptor's active oxygen and hydrogen relative to the catalytic residue and the donor – are conserved with the acceptor rotated 180° in place, motivating the hypothesis that the reaction mechanisms (and reaction coordinates) are homologous. The protein structures are fitted onto one another, though for clarity only that of the α-1,3 product is shown here. Non-reactive hydrogens are also omitted for clarity.

Aimless shooting

Each of the 80 initial conformations were used to seed two "threads" of aimless shooting using the flexible length shooting algorithm of Mullen *et al.*[91] and the SCC-DFTB quantum mechanical model,[77,78] and threads were canceled if they were rejected 10 times in a row to prevent excessive sampling of regions far from the transition state. This procedure was followed to yield 2305 unbiased shooting moves, of which 2069 committed to either the reactant or product basin from the "forward" trajectory. Of those, 275 showed the "backward" trajectory committed to the opposite basin than from the "forward" trajectory, and thus were "accepted" as points along the ensemble of pathways connecting the reactant and product basins. New shooting points were generated after an accepted shooting move by randomly choosing (with equal probability) the configuration from the first 50 1-fs frames of either

the forward or reverse trajectory, also chosen randomly with equal probability. The basin definitions were: for the products, the 4NX O4 and the 1AF C1 atoms closer than 1.60 Å and the 1AF C1 atom and the primary nitrogen of the azide group further than 2.75 Å; and for the reactants, the former distance further than 2.75 Å and the latter closer than 2.00 Å. These conservative basin definitions were chosen to avoid errors due to potential recrossing. The full set of collective variables that were included in each observation are listed in Appendex B.

3.4.3 Free energy of reaction

We calculated the overall free energy of each reaction using Gaussian 16, Rev. B.01[97] using the B3LYP quantum mechanics model[82–85] with the 6-311+G(d,p) basis set.[86,87,98] These calculations were performed on the donor, acceptor, azide, and α-1,3 and α-1,4 product structures solvated in implicit water using the polarizable continuum model.[99–101] The overall reaction energy was calculated from the Gibbs free energies of the constituent molecules as:

$$\Delta G_{rxn} = G_{product} + G_{azide} - G_{donor} - G_{acceptor}. \tag{3.1}$$

3.5 Results and Discussion

3.5.1 Results

Analysis of how 4NX binds in the active site revealed two clear modes, shown in Figure 3.2. In the mode shown in gold in the figure, the C4 of the 4NX fucosyl group is closest to the donor C1, which we hypothesized leads to the α-1,4 product, while the mode shown in silver has the C3 of the 4NX fucosyl group closest to the donor C1, corresponding to a α-1,3 product. Binding energy measurements for the α-1,3 and -1,4 reactant states

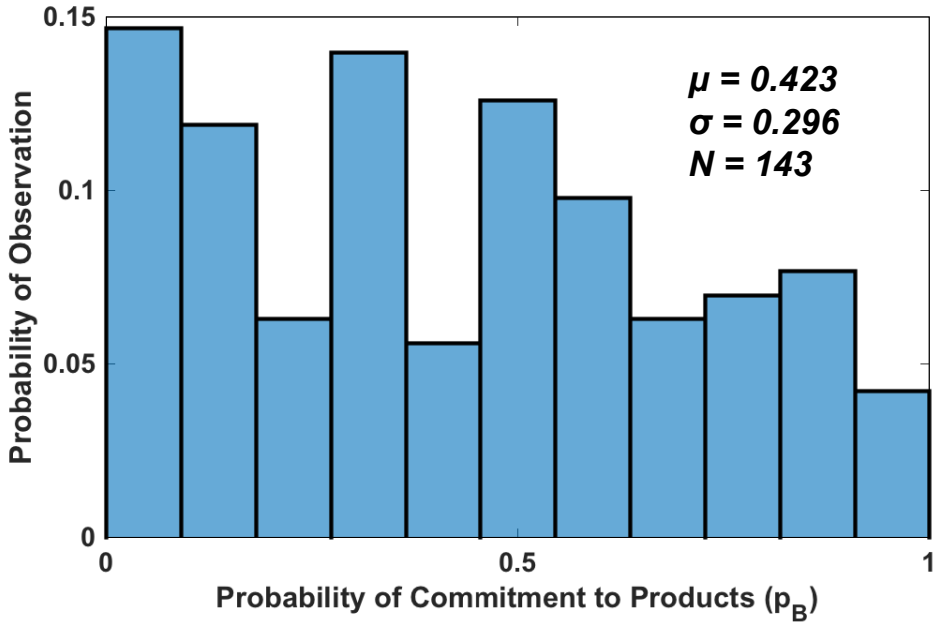

Figure 3.3: ***Tm*Afc committor distribution**

Committor distribution for the α-1,4 reaction, using the reaction coordinate definition in Equation 3.2. Each of the 143 shooting points had an RC value with absolute value less than 0.1 and was simulated 10 times in order to obtain a p_B value. Although the committor distribution is not of the ideal, sharply peaked shape, its average is appropriate and its standard deviation is reasonable, suggesting a decent fit between the reaction coordinate and the underlying committor surface.

showed no significant differences in the protein's affinity for binding either mode. For this reason, and based on the highly similar reactant state binding modes hypothesized to account for the two products, we studied only the reaction coordinate for the formation of the α-1,4 product. We propose that the reaction coordinate for the α-1,3 would be analogous to that for the α-1,4 reaction, with the identities of the donor O4 and its hydrogen changed to the O3 and its hydrogen in the definitions of the relevant CVs.

Likelihood maximization was performed for the α-1,4 reaction on a set of 54 candidate CVs. The top three CVs whose values and rates of change were most predictive of commitment to the product and reactant basins were: the distance between the acceptor O4 and the donor C1 (CV_3); the distance between the donor C1 and the primary azide nitrogen (CV_4); and the difference of the distances between the transferred hydrogen and the glycosidic and Glu-266 carboxyl oxygens, respectively (CV_{21}). See Appendex B for a

complete list of CVs. The reaction coordinate constructed only from the configurational parts of these CVs was:

$$RC = -1.35 - 3.66\ \text{Å}^{-1}\ CV_3 + 3.83\ \text{Å}^{-1}\ CV_4 - 1.69\ \text{Å}^{-1}\ CV_{21} \tag{3.2}$$

where $RC = 0$ represents the transition state and the RC is dimensionless. Because all three of these CVs represent a different bond breaking and/or forming, their importance in describing the progress of this reaction is unsurprising. Validation of the reaction coordinate was performed using committor analysis, and the results are shown in Figure 3.3.

The energy profile along this RC was obtained *via* umbrella sampling and is shown in Figure 3.4. The free energy profile for the α-1,4 reaction is in very close agreement with the experimental activation and overall free energy energy reported by Agrawal *et al.*,[102] which strongly suggests that this is the rate-determining step.

In addition to the energy of activation and ΔG value for the α-1,4 for the reaction step, we calculated the overall reaction ΔG using Gaussian[97] as described in the Computational Methods section. The overall reaction energy (the difference in energy between the free product and free reactant states) was slightly exothermic at -1.27 kcal/mol for α-1,4 and -1.24 kcal/mol for α-1,3. This difference in overall ΔG values are consistent with the ratios of products reported by Cobucci-Ponzano *et al.*[66] of 55:45 (α-1,4 : α-1,3), giving a ratio of 51:49.

The experimentally observed equilibrium constants (K_{eq}) were very low (6.6E-3 and 5.4E-3 for the α-1,4 and -1,3 reactions, respectively; personal communication, see ESI for details†). Based on the conditions reported in that study, if thermodynamics of the chemical reaction step dominated, a K_{eq} approximately two orders of magnitude higher would be expected. It is possible that the reactant binding and/or product unbinding steps have significant barriers and limited the rate of conversion. Studying these steps would

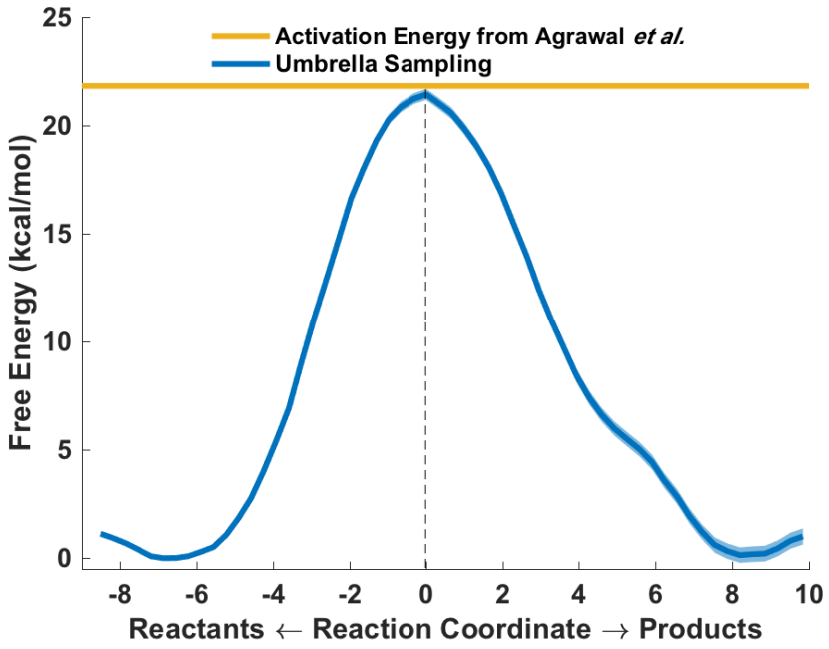

Figure 3.4: ***Tm*Afc reaction energy profile**

Energy profile for the α-1,4 reaction obtained via umbrella sampling along the unitless reaction coordinate (RC). The negative RC values represent the side of the separatrix including the reactant basin while positive values represent the side including the product basin. The orange line represents the experimentally observed activation energy from Agrawal *et al.*,[102] and is in exceptionally close agreement with our results.

thus be of interest for future work.

3.5.2 Discussion

Snapshots corresponding to the reactants, products, and transition state of the α-1,4 reaction are shown in Figure 3.5. The reaction was observed to proceed via an oxocarbenium-like transition state; unsurprising, given that this is the same transition state structure typical of wild-type glycoside hydrolases,[103] although in this case a nitrogen atom is substituted for the more typical oxygen atom. The reaction mechanism takes place in a concerted manner, with none of the relevant bond lengths changing significantly earlier than the others in either the forward or reverse directions, and a single energy barrier is observed.

One potential explanation for the poor reaction efficiency observed for this enzyme is

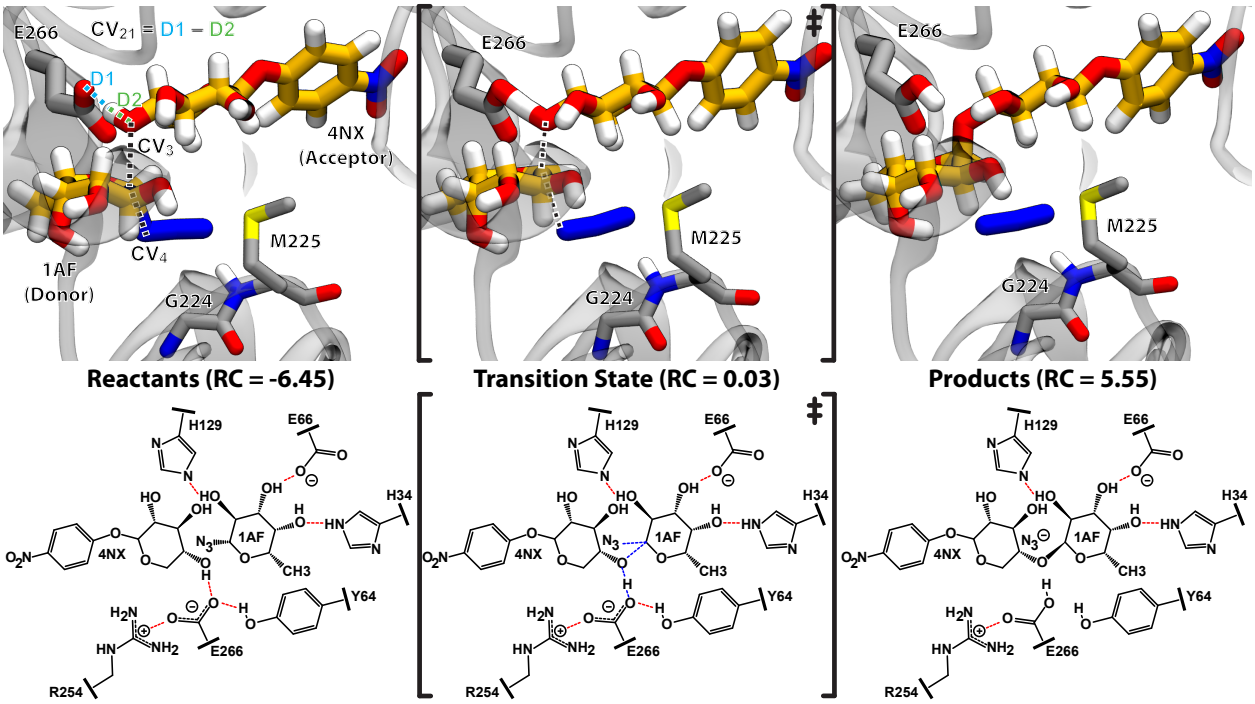

Figure 3.5: ***Tm*Afc full reaction diagram**

Snapshots and schematics of a representative transition path for the α-1,4 reaction. For this figure, a transition pathway from late in the aimless shooting procedure and with a relatively high acceptance ratio preceding it was chosen to ensure maximum decorrelation from the initial configuration. Red dashes in the schematic representations indicate hydrogen bonding interactions. Dashed lines in the reactant state snapshot (at top-left) indicate the CVs constituting the reaction coordinate obtained with likelihood maximization. Dashed lines in the transition state snapshot (top-center) help delineate the oxocarbenium-like transition state structure, where the catalytic hydrogen is caught between the two oxygen atoms and the resulting partial charge on the donor O4 is compensated by an elongated bond between the anomeric carbon and the azide group. These intermediate bonds are represented with blue dashes in the corresponding schematic. Finally, in the product state (at right) the azide is completely dissociated from the fucose and the new glycosidic bond is formed as the donor O4 bonds fully with the catalytic Glu-266 residue. Residues Met-225 and Gly-224 are also shown in the snapshots to illustrate the molecular context around the azide group, whereas various hydrogen bonding residues are shown in the schematics to depict the stabilization of the donor and the catalytic residue.

that there are no nearby residues or water molecules to stabilize the departure of the azide group. Instead, the product state azide ion is left unbound in the active site cleft, presumably until it is able to diffuse into the bulk solvent (although this was not observed during our simulations, which had limited sampling time and a limited QM region that could prevent such an observation). In this light, it is not surprising that the rescue of fucosidase activity with the addition of free azide was not observed in the *Tm*Afc D224S mutant,[66] as there is already limited space for the azide ion with a glycine at position 224; the relatively

bulky serine side chain would restrict the azide group's access to the substrate. Based on this reasoning, a possible target for engineering this enzyme is via mutation of Met-225. This bulky, hydrophobic side chain sits in prime position for interacting with the azides, so its mutation to a positively-charged side chain could potentially stabilize the departure of the leaving group.

3.6 Conclusions

We investigated the mechanism of oligosaccharide synthesis in the reaction between 1-azido-β-L-Fucose (1AF) and 4-nitrophenyl β-D-xylopyranoside (4NX) via the glycosynthase *T. maritima* α-L-fucosidase (*Tm*Afc) D224G. We propose that the reaction proceeds in a single endothermic step via a oxocarbenium-like transition state, wherein the role of the mutant residue Gly-224 is solely to provide room for the leaving azide group. Experimental results indicated that the α-1,3 product is produced in a slightly lower quantity compared to the α-1,4 product, and that fucosidase activity in the D224S mutant could not be rescued with the addition of free azide. Our results explain both of these observations, and provide new information for use in designing and engineering *Tm*Afc and other glycoside hydrolases for improved glycosynthetic activity.

CHAPTER IV

ATESA: an Automated Aimless Shooting Workflow

4.1 Chapter Introduction

As we have seen in the previous chapter, transition path sampling (TPS) is an extremely useful, flexible, and powerful approach for analyzing rare events in molecular simulations without specifying in advance how they should proceed. This latter advantage – that the mechanism of the rare event is discovered rather than specified – sets TPS apart from most other rare event sampling methods, making it among the most rigorous methods available for identifying the precise molecular mechanisms of rare events. Put another way, TPS injects the minimum possible bias into the system while still managing to study rare events that otherwise would never be observable in statistically meaningful quantities with a totally unbiased simulation.

Aimless shooting was the TPS method used to produce the key results in the previous chapter. As alluded to in the chapter introduction, the workflow in that chapter was in no way specific to the molecular system in question, and despite its apparent complexity, should in theory be highly automatable. Wanting to do so – both for the personal benefit of spending less time painstakingly curating and analyzing thousands of simulations, and for the community benefit of making it more available to other researchers interested in using TPS – was the original motivation behind what would eventually become ATESA:

Automated Transition Ensemble Sampling and Analysis.

In no small part thanks to guidance from software scientists at the Molecular Sciences Software Institute, ATESA is now a mature Python package featuring continuous test integration, support for installation and dependency management via the popular Python Package Index “pip” installer, and extensibility to different simulations engines, cluster computing batch systems, simulations methodologies, and more written into the structure of the code. The final package will not only save time and effort for researchers already familiar with TPS, but also at make TPS available at all to researchers not specialized in advanced sampling methods. It requires only the most basic competency in setting up molecular simulations from its users, and is supported by thorough documentation and tutorials to guide non-specialists from the contemplation stage (that is, deciding whether TPS is even the appropriate tool for their application) all the way to the analysis of final results. Furthermore, unlike other tools such as OpenPathSampling, ATESA is designed as a standalone software tool – not a Python API – and as such, does not require users to read or write any Python code.

ATESA represents the natural conclusion of the research trajectory of the preceding chapters: a tool to facilitate the analysis of rare events (that is, the temporally “unusual” steps in molecular processes that generally define the macroscopic behavior of a chemical system) from a minimum bias transition pathway perspective, incorporating as much of the knowledge obtained in that research as possible in such a way as to make it available to others without requiring them to undergo the same extent of specialized training. Ideally, tools such as ATESA should expand the scientific toolkit of molecular researchers who are not themselves specialists in enhanced sampling, thereby maximizing the impact that research in enhanced sampling can have on the field.

ATESA: an Automated Aimless Shooting Workflow

Tucker Burgin, Samuel Ellis, and Heather Mayes (2021). In preparation for publication.

4.2 Abstract

Transition path sampling methods are powerful tools for studying the dynamics of rare events in molecular simulations. However, these methods are generally restricted to experts with the knowledge and resources to properly set up and analyze the often hundreds of thousands of simulations that constitute a complete study. ATESA is a new open-source software program written in Python that automates a full transition path sampling workflow based on the aimless shooting algorithm, streamlining the process and reducing the barrier to use for researchers new to this approach. This introduction to ATESA includes a demonstration of a complete transition path sampling process flow for an example reaction, including finding an initial transition state, sampling with aimless shooting, building a reaction coordinate with inertial likelihood maximization, verifying that coordinate with committor analysis, and measuring the reaction energy profile with umbrella sampling. We also describe our implementation of a termination criterion for aimless shooting based on the Godambe information calculated during model building with likelihood maximization, as well as a novel approach to constraining simulations to the desired rare event pathway during umbrella sampling.

4.3 Introduction

Molecular simulations are a powerful method of probing the complex chemical and physical behavior of molecules at the atomic scale. The wide variety of tools, tutorials, and resources now readily available simplify the incorporation of molecular simulations into studies. However, many of the most interesting topics of study in molecular simulations

involve so-called rare events — transformations that involve at least one transient, high-energy state — such as chemical reactions, large conformational shifts (like protein folding), and crystal nucleation. Because by definition rare events occur infrequently, they cannot be readily observed by brute-forced, unbiased simulations, and certainly not with enough frequency to measure key quantities such as rate constants in a statistically meaningful way.[104]

The school of molecular simulations research dedicated to the study of rare events is called enhanced sampling, or rare event sampling. These terms encompass a huge variety of simulations techniques that aim to simulate rare events at a rate much higher than would be expected from simple simulations alone, and thereby allow for statistically relevant observations about those events.[105]

For a given enhanced sampling workflow, many of the steps are the same regardless of the specific nature of the rare event under investigation. When these steps are automated in an easy-to-use workflow, these studies require less researcher time to complete and become more reproducible, and it becomes easier for researchers new to the field to more quickly produce results. To this end, several software packages have been created to facilitate or automate rare event sampling simulations.

One prominent example of such a package is PLUMED, which simplifies the workflow for a number of enhanced sampling methods, most prominently umbrella sampling, replica exchange, and metadynamics, and has been cited in more than one thousand studies.[106] Other examples include OpenPathSampling,[107] PyRETIS,[108] WESTPA,[109] SSAGES,[110] and Colvars,[111] in addition to numerous tools built directly into molecular simulations packages like Amber[112] or CHARMM.[36]

Despite the success of these tools, there remain relatively few options for automating transition path sampling techniques, which are a subset of enhanced sampling where

the focus is on analyzing ensembles of pathways connecting stable states.[113] One major advantage of transition path sampling compared to more popular methods, like metadynamics and umbrella sampling, is that the dimensions of phase space along which the enhanced sampling is performed (called "collective variables," or CVs) do not need to be specified in advance; instead, the key dimensions are *discovered* over the course of the sampling.[114] This is of critical importance, as sampling along incorrect CVs can yield misleading results.[115] Furthermore, the best choice of CVs is not always obvious, especially in highly complex systems such as enzymes, but also even in ostensibly simple rare events, such as the dissolution of crystalline salts in water.[116] Given this major advantage, we propose that the relative unpopularity of transition path sampling compared to other methods can be at least partially ascribed to the relative lack of accessible tools in this space.

If there are few tools for automating transition path sampling, there are still fewer for aimless shooting in particular, in spite of evidence that aimless shooting is an especially efficient and accurate transition path sampling method.[91] To our knowledge, only two programs have been published for this purpose, and neither has been the subject of a journal publication: one unnamed program from Baron Peters' group at the University of Illinois at Urbana-Champaign (referred to by Beckham and Peters[93]), and another from Sharon Glotzer's group at the University of Michigan named LibTPS (github.com/askeys/libtps). The program from Peters also features a script for performing likelihood maximization to obtain a reaction coordinate from the aimless shooting data, but neither program facilities setup of the appropriate initial structure for aimless shooting, verification of putative reaction coordinates, or measurement of energy profiles.

Aimless Transition Ensemble Sampling and Analysis (ATESA) is a new, open-source Python program that addresses this outstanding need. It implements a particularly pow-

erful rare event sampling method called flexible length aimless shooting,[91] as well as an accompanying suite of setup, verification, and analysis tools. The full workflow of a complete enhanced sampling study is automated by ATESA, beginning with a molecular model representing either stable state separated by the rare event in question, and ending with a verified reaction coordinate that describes the event and a free energy profile along that coordinate — all handled by the software and without any need of specialized programming or enhanced sampling expertise on the part of the user. At time of publication, ATESA supports only the Amber molecular simulations package,[112] but is written to be extensible to other software. We also introduce a statistically meaningful criterion for aimless shooting termination, and a novel method for leveraging aimless shooting data to improve free energy profile analysis with umbrella sampling. This paper is not intended to serve as documentation of ATESA (documentation can be found at https://atesa.readthedocs.io/en/latest/), but instead provides an overview of the approach.

4.4 When Should We Use Aimless Shooting?

Aimless shooting (like all transition path sampling methods) is predicated on the understanding that rare events can be described as transitions between two stable (that is, relatively long-lived) states separated by a short-lived higher energy state, called the "transition state." More specifically, the transition state can be thought of as the lowest-energy state along the transition "separatrix," which is the subset of state space where unbiased, aimless simulations are equally as likely to proceed towards one stable state as towards the other. If free energy is thought of as analogous to elevation on a map, then the transition state represents the highest-elevation point along a "mountain pass" — the easiest path to traverse in the journey from one state to another.[114]

Because even very small molecular models can contain huge numbers of degrees of

freedom, many of which have only a very small influence on the energy of the transition state, it is more accurate to think in terms of an "ensemble" of transition states, weighted by their relative likelihoods (that is, their relative free energies.) And for a given ensemble of transition states, there must also be an ensemble of transition *paths* connecting the stable states through them. The extent to which this ensemble can be accurately observed is the primary parameter that distinguishes a strong transition path sampling study from a poor one. This is the origin of the term "transition ensemble" in the acronym "ATESA."

Of course, aimless shooting is not always the best tool for the job. Efficiency aside, the primary advantage of aimless shooting compared to other path sampling tools is that it is completely unbiased; the only thing you need to get started is a way to distinguish one state from the other, and a guess about what lies in between (which ATESA can help you find). A more general review of rare event sampling methods with a focus on transition path sampling is available from Dellago and Bolhuis.[117] If you're unsure whether aimless shooting is the best option for your application, consider the following:

- Aimless shooting is designed to focus sampling around transition states, rather than at the stable states that they connect. If you are more interested in comparing properties of stable states than in understanding how one transitions to another, aimless shooting is not the right tool.
- Aimless shooting is best used to discover or describe a mechanism when one is unknown or only hypothesized. If you have a known reaction coordinate or collective variable that describes the transformation already in mind and just want to characterize it, you may look to a pathway free energy method like umbrella sampling or equilibrium path sampling (both implemented in ATESA), or a path sampling method that makes use of a known collective variable like transition interface sampling. Aimless shooting can also be a suitable alternative when other methods produce results that have

hysteresis (in this context meaning different results going in one direction along a pathway compared to the other direction) or that fail mechanistic hypothesis tests.[118]

- Like all transition path sampling methods, aimless shooting is at its best where the energy barrier is high (and therefore transitions are rare). If your transition occurs quickly enough to reasonably observe many times over the course of an unbiased molecular dynamics or quantum mechanics simulation, aimless shooting may be overkill.
- Aimless shooting is best applied when you are interested in efficiently arriving at an accurate description of the transition state of a potentially complex rare event without specifying a mechanism *a priori*.

4.5 Usage and an Example Study

ATESA consists of one main script, *atesa.py*, and a few auxiliary scripts. The main script can be called through the command line along with a user-defined configuration file that defines the behavior of the program. The complete workflow of an ATESA study is depicted as a flowchart in Figure 4.1. Thorough documentation of each job type and auxiliary script is available online at https://atesa.readthedocs.io/en/latest/, but here we will simply provide brief descriptions of key configuration options as they become relevant, to help illustrate how ATESA is used. This section can help serve as an introduction to the particular transition path sampling workflow automated by ATESA.

The use of a configuration file is intended to combine the flexibility and consistency of workflows that are written as computer scripts directly, like in the case of OpenPathSampling[107] or OpenMM,[119] with the accessibility of selecting options without requiring any programming ability. The basic usage of the configuration file is to simply set each desired option based on the documentation using an *option = value* syntax. However, ATESA

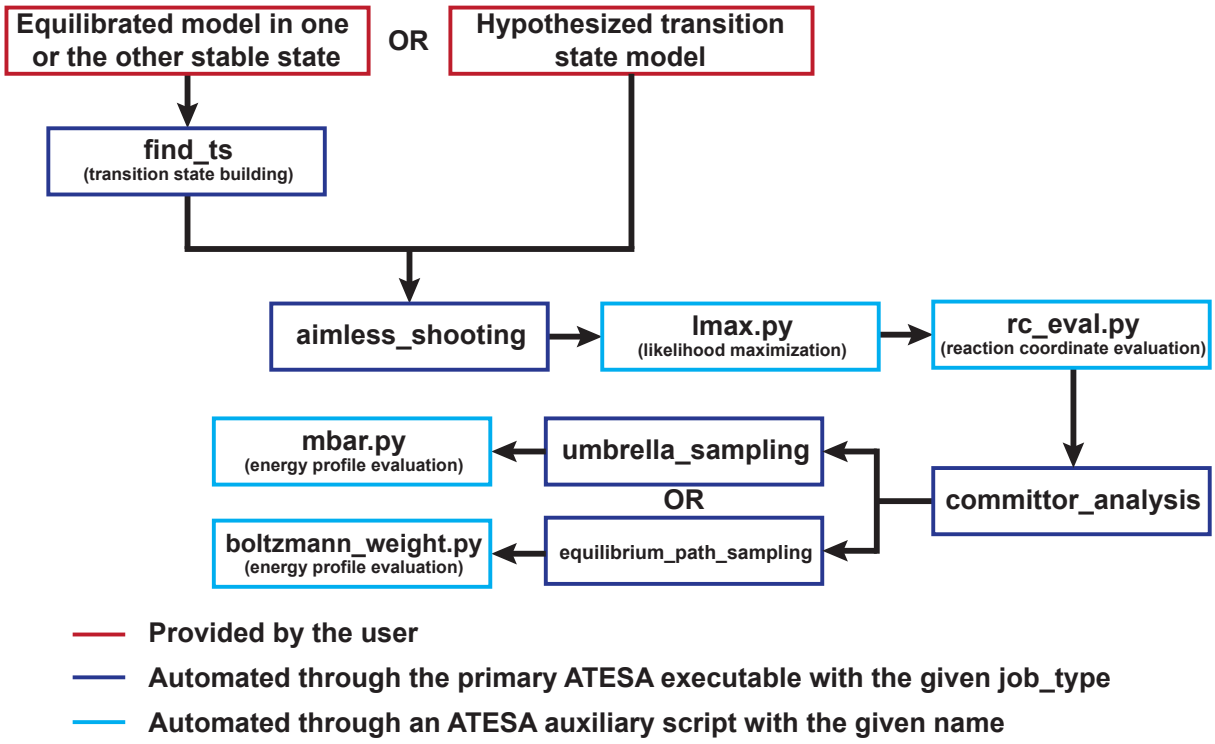

Figure 4.1: **ATESA workflow**

The standard workflow for a study with ATESA. Dark blue boxes indicate steps performed by specifying the given job_type in the configuration file provided to *atesa.py*, whereas light blue boxes indicate steps performed with other ATESA scripts. After providing the initial model, the entirety of the remaining workflow is automated through ATESA.

executes the configuration file as if it were Python code, which means that advanced users can write arbitrarily complex scripts to determine how each option should be set.

The only prerequisite to using ATESA is that the user has a working molecular model of the system that they want to study, including an equilibrated coordinate file that occupies one of the two stable states of interest, or a transition state model obtained through some other means if preferred. If appropriate, such as if the rare event of interest is a chemical reaction, the system must be set up to perform quantum mechanics (QM) or combined quantum mechanics/molecular mechanics (QM/MM) simulations, and ideally should be equilibrated at the desired level of quantum mechanical theory. An introduction to QM/MM simulations is available from Groenhof,[120] and a tutorial on setting up QM/MM simulations

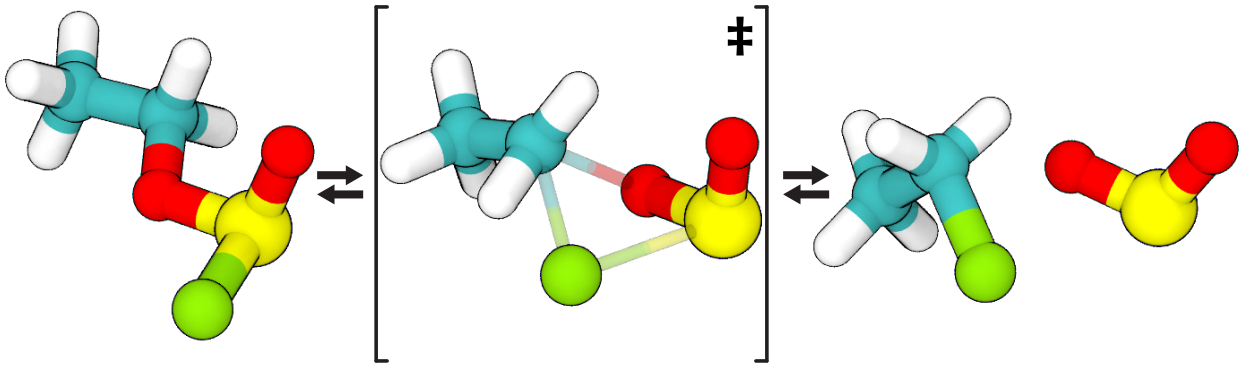

Figure 4.2: **ATESA example reaction pathway**

Reaction pathway for the example reaction, S_Ni decomposition of ethyl chlorosulfite in the gas phase. Schreiner *et al.*[121] demonstrated that this "frontside" attack (where the chlorine bonds to the same side of the carbon as the oxygen departs from) is energetically favorable compared to attacking from the opposite side. Teal: carbon; white: hydrogen; red: oxygen; yellow: sulfur; green: chlorine.

in Amber[112] is available on their website at:

https://ambermd.org/tutorials/advanced/tutorial2/index.htm.

In this section, we will go through each step in a complete aimless shooting transition path sampling study in order, and provide a step-by-step example of each step's application to the study of a simple example reaction: the S_Ni decomposition of ethyl chlorosulfite in the gas phase, which has previously been studied by Schreiner *et al.*[121] The best reaction pathway found therein is depicted in Figure 4.2. Further simulation details are provided in the supplementary information. Although a small and simple reaction was chosen here for demonstration purposes, the same workflow has been successfully applied to simulations of much larger systems, including entire enzymes.

4.5.1 Building initial transition states

The first step in preparing a molecular model for aimless shooting is obtaining a putative transition state structure. This structure does not have to be representative of the "true" transition state; the only requirement is that random ("aimless") trajectories beginning from those initial coordinates have a reasonably high probability of proceeding towards either

stable state (typically in the range of 10-30%).

ATESA automates one potential approach to building initial transition states for aimless shooting by gently forcing the system to cross the separatrix with steadily increasing energetic restraints. To use this feature, the user provides a configuration file with *jobtype = find_ts* and defines the two stable states that the simulation should connect. Then, the user provides an initial coordinate file that occupies one of those stable states, and the software automatically applies appropriate restraints to push the model into the other stable state (which must necessarily bring it through the separatrix). After this biased simulation, ATESA identifies likely transition state candidates from the resultant trajectory and tests them with a small amount of aimless shooting (by default, 10 steps) to verify that they are suitable transition states. These structures provide the initial coordinates for aimless shooting in the next step.

It is best to define the stable basins in a mutually exclusive way to ensure that any given configuration either occupies one or neither state, but not both. Figure 4.3 shows the stable basin definitions used to find the initial transition state for the example ethyl chlorosulfite reaction, as well as the initial coordinates (constructed using Open Babel[122]) and a resulting putative transition state. Although no path was specified by these basin definitions (that is, they specify the end points, not the means of getting from one to the other), the gentle restraints resulted in an initial transition state structure in close agreement with the optimized structure for this reaction presented by Schreiner *et al.*[121]

4.5.2 Aimless shooting

The key insight that drives aimless shooting is that unbiased sampling can be focused around the transition state (even though it is a relatively high-energy state by definition) by chaining together short simulations starting from a member of the transition state

```
commit_fwd = ([3,1,1],[5,5,2],[2.2,2.0,3.0],['gt','lt','gt'])
commit_bwd = ([3,1,1],[5,5,2],[1.5,3.5,2.4],['lt','gt','lt'])
```

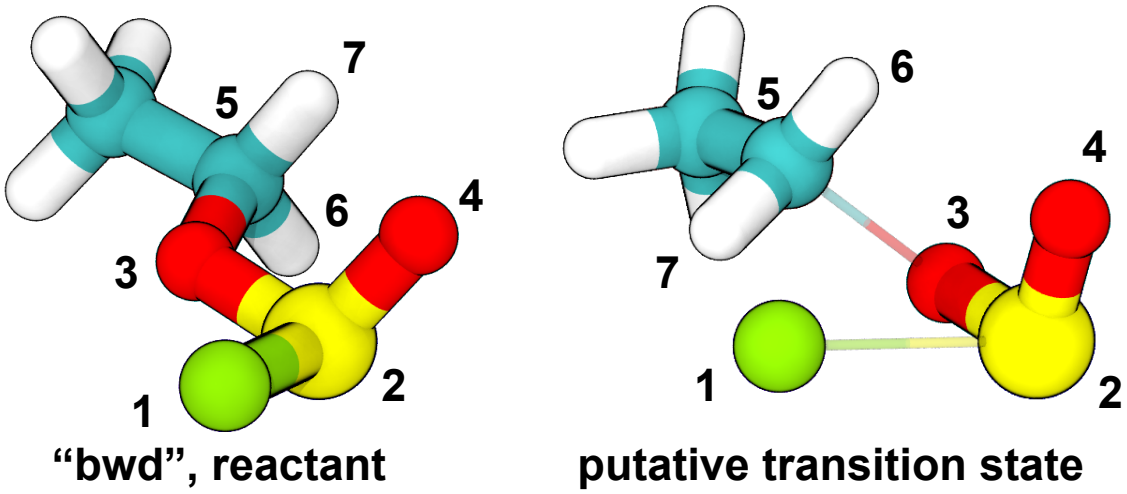

Figure 4.3: **ATESA example transition state**

Definitions of stable states and initial and final structures from the example *find_ts* job. The stable state definitions are read by inner index; for example, the first element of the definition of the “bwd” state is read as “the distance between atom 3 and atom 5 is less than (*‘lt’*) 1.5 Å”. Based on these definitions, the initial coordinates (at left) occupy the “bwd” state, and restraints are automatically constructed to build a transition state (at right) that has roughly equal probabilities of relaxing to either state. The narrow, transparent bonds in the transition state structure show the original topology of the model, for comparison.

ensemble, as described in the publications that introduced this approach[123,124] and the recent extension of the method, flexible length aimless shooting.[91] Briefly, starting from a putative member of the transition state ensemble (called a “shooting point”), an unbiased simulation with random atomic velocities is initiated, at the same time as another simulation starting from the same initial coordinates but with exactly opposite velocities. If one of these simulations proceeds towards one stable state, and the other towards the other stable state, then the shooting move is “accepted,” and a new shooting point is picked from an early time step of one of the previous simulations to continue the procedure. These trajectories can be quite short (on the order of femtoseconds to picoseconds depending on the simulation parameters) as the system quickly relaxes to a lower energy conformation, rendering it feasible to sample thousands of transition pathways in a short time. As new shooting points are selected, the system relaxes along all dimensions orthogonal to the

underlying reaction coordinate.

ATESA automates aimless shooting using an arbitrary number of independent "threads," each of which represents its own unique chain of shooting points. Multiple threads can start from the same initial coordinates, and they will rapidly diverge (depending on the steepness of the local energy landscape) due to the pseudo-random velocity initialization and choices of starting points from previous trajectories. For the example ethyl chlorosulfite reaction, we initialized 50 unique threads from the putative transition state produced by the *find_ts* job, using the default settings to automatically identify the CV measurements to make for each shooting move in preparation for the next step, as well as for the termination criterion (see the Information Error Termination Criterion section).

4.5.3 Likelihood maximization and reaction coordinate evaluation

Likelihood maximization[123,124] is a robust numerical method of selecting a model from a set of observations about a system that is maximally predictive of a given outcome. In the case of aimless shooting, the outcomes are the energy basins at the endpoints of simulations starting from a shooting point, and observations are measurements of collective variables characterizing that shooting point, such as distances, angles, dihedrals, or any other descriptor such as contact angle or number of hydrogen bonds. Likelihood maximization can be used to discern the combination of parameters that is maximally predictive of the outcome.

Likelihood maximization produces a model reaction coordinate (though the event need not be a chemical reaction), a unitless scalar value describing the progress of the rare event from one stable state to another through the transition state, customarily located at zero. The accuracy of the coordinate is limited by whether the tested set of collective variables includes appropriate system descriptors and how many variable combinations

are tested, which are typically modeled as linear combinations. Although the reaction coordinate produced by likelihood maximization will always be an oversimplification of the true dynamics, it can be an extremely valuable tool in describing the key features of the system as it proceeds through the rare event.

Other model selection schemes such as the Bayesian[125] or Akaike[126] information criteria may also be reasonable choices, although these are not currently included in ATESA. ATESA implements an improved "inertial" version of likelihood maximization that takes into account both the values and rates of change of collective variables.[92] For the example reaction, we performed inertial likelihood maximization over hundreds of automatically chosen CVs using ATESA's built-in "two_line_test" option to produce a highly predictive model without specifying the number of terms in the reaction coordinate *a priori*. This algorithm iteratively adds further collective variables onto the model until the marginal rate of improvement from each additional dimension compared to the average rate of improvement from the addition of earlier dimensions falls below some threshold (in this case, 0.5), resulting in the reaction coordinate shown in Figure 4.4. In this case, ATESA has produced a non-trivial reaction coordinate: one term describes the relative proximity of the reactive carbon to either of the other atoms it may bond to, while the two angle terms describe the orientation of the partially charged face of the carbon atom, using the sulfur atom as a reference point.

After obtaining the reaction coordinate, the auxiliary script *rc_eval.py* can be used to measure the value of that reaction coordinate at each shooting point. This facilitates the selection of an appropriate set of initial coordinates for the next step, committor analysis.

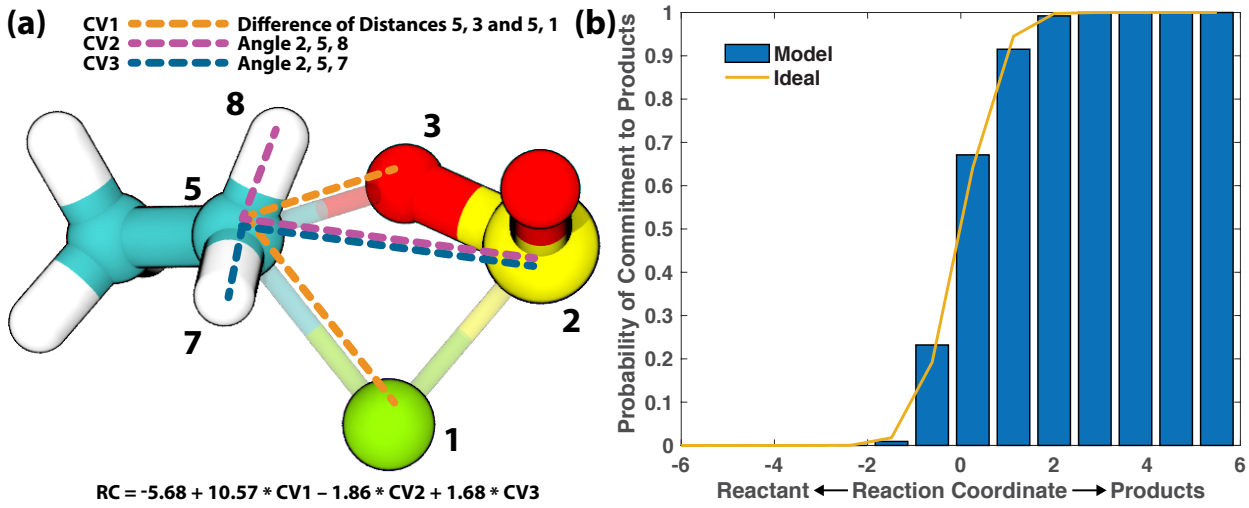

Figure 4.4: **ATESA example likelihood maximization**

Likelihood maximization results for the example reaction. (a) Visualization of the reaction coordinate model chosen automatically using ATESA's "two_line_test" option. See documentation at atesa.readthedocs.io/ for details. In this case, a three-dimensional reaction coordinate model was chosen including three CVs as shown. (b) Comparison between the modeled and ideal probabilities of commitment to the "products" state for the selected reaction coordinate.).

4.5.4 Committor analysis

In transition path sampling, one of the best tools for validating a reaction coordinate is committor analysis.[118] This method is implemented by first collecting a diverse set of initial coordinates close to the predicted transition state (hundreds of maximum likelihood reaction coordinate values near zero (say, to within 0.05% of the distance from the nearest stable state to zero)) and initiating multiple simulations from each point, recording the energy basin to which each trajectory led. If the model reaction coordinate is a good description of the "actual" reaction coordinate, then the observed probability that any given simulation will proceed towards one stable state should be roughly equal to the probability of proceeding towards the other stable state. Poor reaction coordinates will instead produce simulations that are heavily biased towards one or the other stable state.

Figure 4.5 shows the committor analysis distribution for the example model reaction coordinate. The distribution is peaked towards the middle and lower on either extreme, as

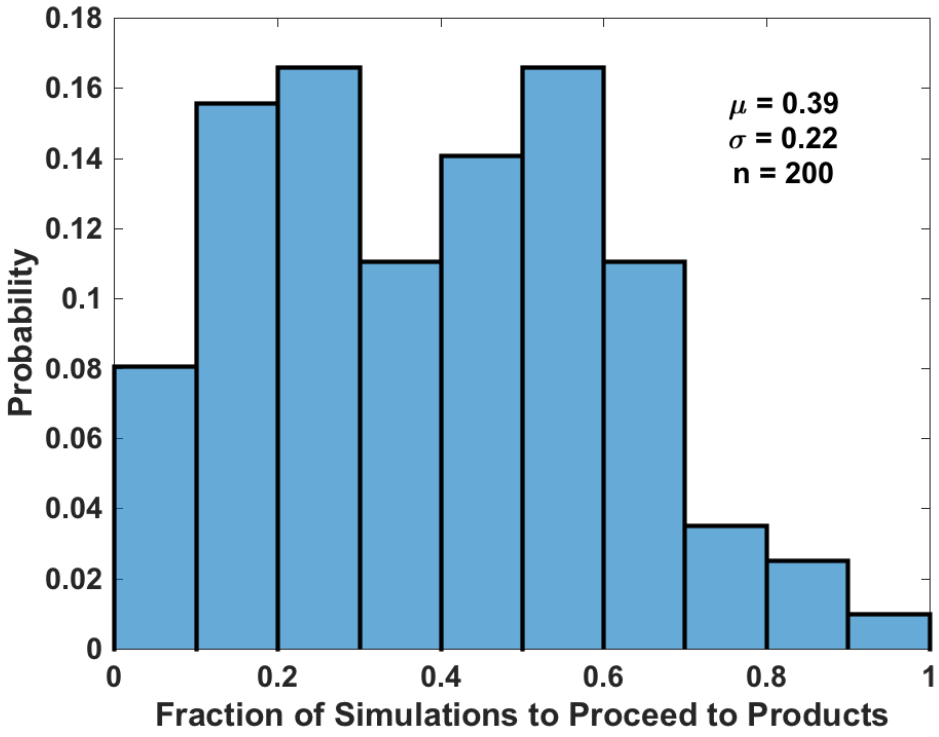

Figure 4.5: **ATESA example committor analysis**

Committor analysis result for the example reaction model. This histogram conveys the ratio of the number of simulations that proceeded to the products ("fwd") state over the total number of simulations. This plot represents n = 200 sets of initial coordinates with 20 simulations each. μ is the mean value and σ is the standard deviation of the distribution.

expected.

4.5.5 Free energy analysis

Once a reaction coordinate has been validated with committor analysis, the final step is to evaluate the free energy profile along it. The free energy profile provides the predicted activation energy for the rare event, and these energies can be used to calculate rate coefficients, *e.g.* by using transition state theory, and the equilibrium distribution of stable states according to the Boltzmann distribution. ATESA automates two methods for calculating the free energy profile: umbrella sampling and equilibrium path sampling.

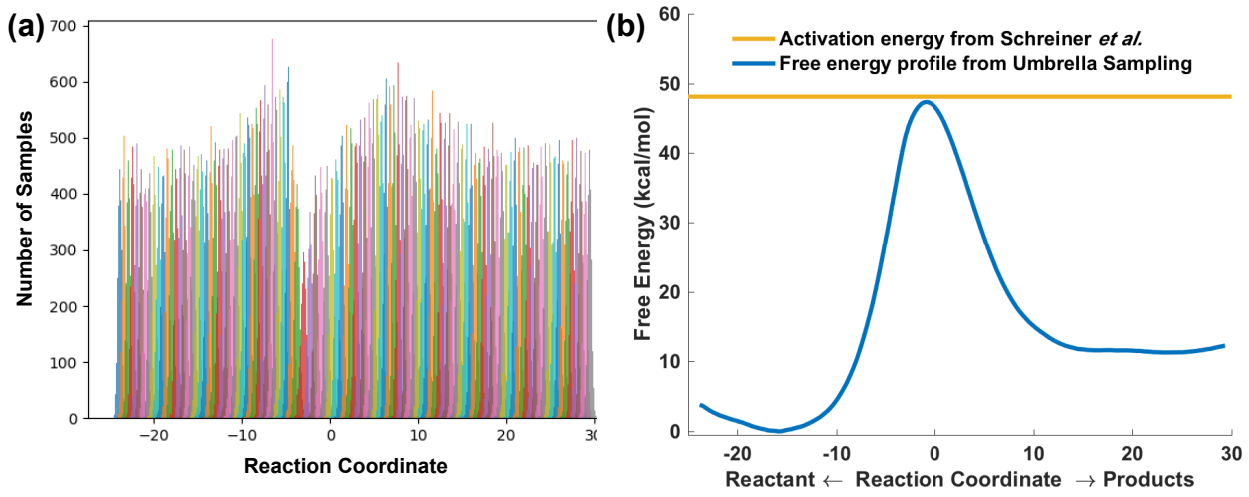

Figure 4.6: **ATESA example umbrella sampling**

Results from analysis of the umbrella sampling data using ATESA's mbar.py. **(a)** The histograms of samples from each window plotted together show that there are no gaps in the data, which is a prerequisite for a continuous free energy profile. **(b)** The free energy profile produced using MBAR. The activation energy is in very close agreement with calculations from Schreiner *et al.*[121] The error on this calculation is smaller than the width of the line itself.

Umbrella sampling

One of the most efficient methods of pathway free energy analysis is umbrella sampling.[127] Harmonic restraints are applied to an array of initial structures representing several discrete states along the reaction coordinate, and the contribution of those restraints to the resulting distribution of samples can be removed *post hoc* in order to produce the underlying free energy profile. Applying these restraints is trivial for very simple coordinates, such as one-dimensional distance or torsion coordinates, but such functionality is not always build-in or available in plugins such as PLUMED for more complex restraints for collective variables that are included in the reaction coordinate. Fortunately, a development version of the Amber simulation package has been created which supports restraints along linear combinations of distances, angles, dihedrals, and/or differences of distances.[128] These are the types of collective variables used by ATESA in creating putative reaction coordinates, and thus this version of Amber was used to apply umbrella sampling along the aforementioned reaction coordinate (Figure 4.4) using 109 windows of width 0.5 and

restraint weight 5 kcal/mol. The sampling histograms and resulting free energy profile obtained with *mbar.py* (a wrapper script that relies on Shirts' PyMBAR package[129,130]) are shown in Figure 4.6. This method also produces estimates of the uncertainty, but in this case, they are too small to make out on the plot.

Pathway-restrained umbrella sampling

Although committor analysis can be used to confirm that the chosen reaction coordinate contains all of the key CVs to describe the *transition state* ensemble, one of the weaknesses of this analysis is that there is no guarantee that the appropriate set of CVs to describe the transition pathway ensemble remains the same along the full path from one stable state to the other. This can pose a problem when attempting to apply umbrella sampling along the reaction coordinate: if there exist any dimensions along which the transition pathway ought to be restrained for some portion of it, but those dimensions ought *not* to be restrained at the transition state (where the reaction coordinate was defined), then relaxation along those dimensions will result in misjudging the shape of the free energy profile along those portions during umbrella sampling.

Fortunately, sampling data from aimless shooting can be leveraged to address this issue. To the extent that aimless shooting explored the ensemble of transition pathways, the regions of state space represented among its accepted trajectories describe the boundaries of the transition pathway in *every* dimension (not just those that contribute to the reaction coordinate). Our approach, which we call "pathway-restrained" umbrella sampling, is to apply additional restraints during umbrella sampling simulations to every dimension that was recorded during aimless shooting. The restraints have flat (zero) weight in the range of values observed during frames of accepted aimless shooting trajectories with reaction coordinate values closest to the umbrella sampling window in

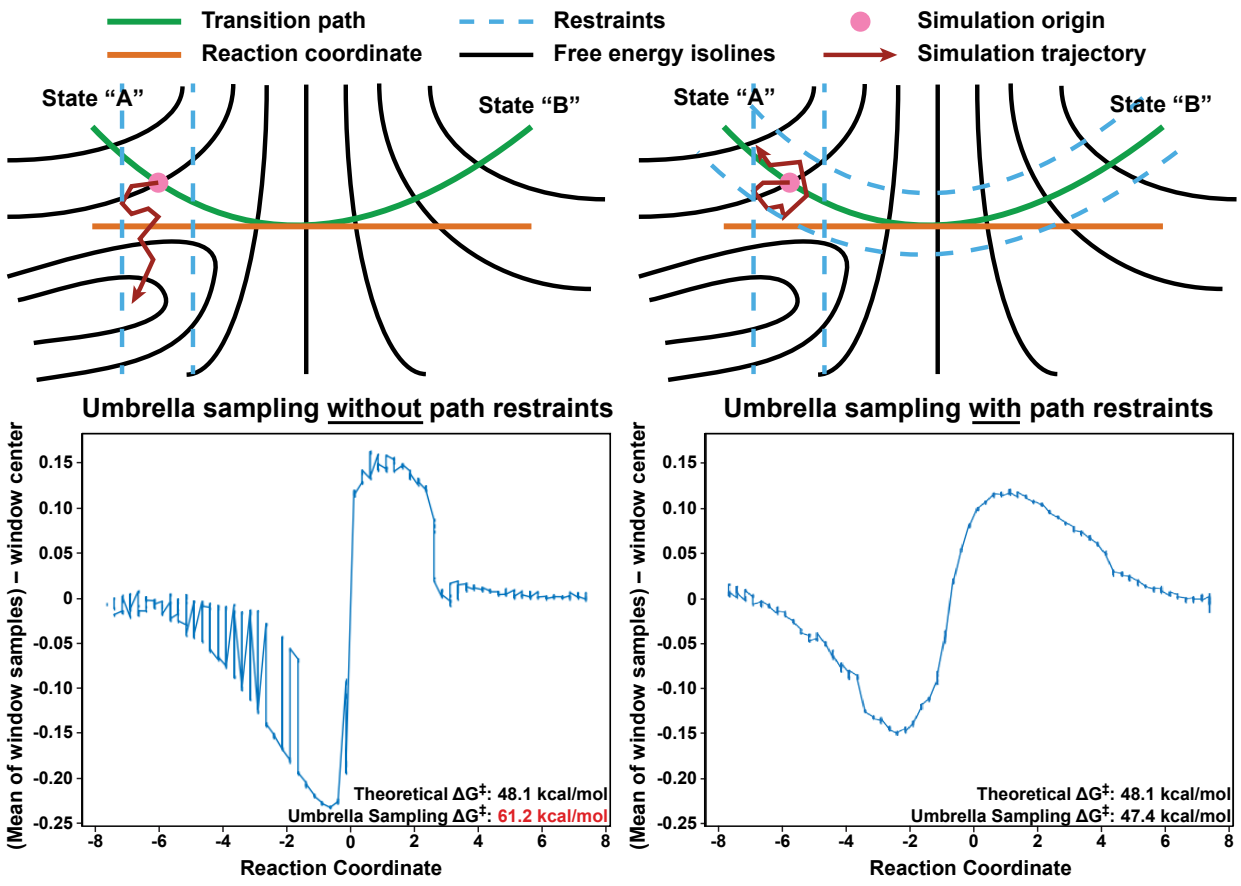

Figure 4.7: **ATESA pathway restrained umbrella sampling**

(Top) Schema depicting a simple free energy surface on which umbrella sampling is being performed with and without pathway restraints derived from aimless shooting data. Likelihood maximization optimizes the reaction coordinate only at the separatrix, where the orange and green lines intersect. The lack of pathway restraints in the scheme at left leaves the depicted simulation trajectory free to relax into the off-pathway free energy basin, which would cause errors in measuring the free energy along the pathway. At right, pathway restraints are added to prevent this. Note that in practice umbrella sampling restraints are harmonic, not rigid walls as depicted here for clarity. **(Bottom)** The difference between the mean reaction coordinate value of the samples collected in each umbrella sampling window and the reaction coordinate value of the corresponding harmonic restraint for that window, plotted against the reaction coordinate value, with and without pathway restraints. These plots represent real data collected with a reaction coordinate for the ethyl chlorosulfite system that produced a strong committor analysis result (albeit not the same reaction coordinate as shown in Figure 4.4). Each point represents a single umbrella sampling simulation (five at each reaction coordinate value). Unsmoothness in the plot without pathway restraints (at left) arises from sampling of two or more energetically distinct states with similar reaction coordinate values, caused by off-pathway local minima such as shown in the scheme above. Theoretical activation energy is from Schreiner *et al.*[121]

question, and steeply increasing weight outside that range. As a result, each umbrella sampling simulation is only able to explore the same regions of state space that were already explored along the corresponding point along the transition pathway during aimless shooting. Figure 4.7 helps visualize the sort of transition pathway and free energy surface where this could be necessary, as well as the impact this has on umbrella sampling data. As illustrated, umbrella sampling restraints are applied along the reaction coordinate, which may not match the shape of the true reaction pathway far from the transition state. Although this does not pose a problem for analyzing the free energy along the transition path when there are no accessible off-pathway free energy minima in the vicinity (like near State "B" in the schema), if there *are* off-pathway minima accessible by relaxation along dimensions orthogonal to the reaction coordinate (like near State "A"), then traditional umbrella sampling simulations can fall into them in error. Application of pathway restraints, which can be handled automatically by ATESA's "us_pathway_restraints_file" option, prevents this.

The bottom half of Figure 4.7 depicts real data collected during umbrella sampling simulations with and without pathway restraints for a specific reaction coordinate describing the ethyl chlorosulfite reaction. Although this reaction coordinate is not the same one shown in Figure 4.4 (this coordinate was chosen specifically to illustrate pathway restraints), it was a strong model with a comparable log likelihood score, and with a similarly strong committor analysis result, indicating good agreement with the shape of the transition pathway ensemble at the transition state. However, off-pathway umbrella sampling along dimensions crucial to describing the shape of the transition path elsewhere resulted in serious error in measuring the activation energy ($\sim$27% error). Application of pathway restraints successfully prevented this error, resulting in an extremely strong estimate of the activation energy ($\sim$1.5% error).

Usage of pathway-restrained sampling does have the potential to impart an additional source of error in umbrella sampling data and should only be used when necessary due to the presence of significant off-pathway sampling in otherwise-unrestrained umbrella sampling simulations. Some off-pathway regions of state space may still exist within the confines of the restraints, especially when the excursion is not along one of the restrained dimensions. Furthermore, although this approach can help remedy errant relaxation along dimensions that should be restrained, it cannot help when a dimension *is* restrained at the transition state and should not be elsewhere.

Equilibrium path sampling

Equilibrium path sampling is also included in ATESA, as it is a method that can be used with *any* reaction coordinate, including ones that would be difficult to impose as umbrella sampling restraints. Like in umbrella sampling, equilibrium path sampling uses "windows" along the reaction coordinate, but instead of imposing restraints on the simulations, equilibrium path sampling chains together short simulations (on the order of femtoseconds) to allow sampling of higher-energy conformations along the reaction coordinate, similar to the aimless shooting algorithm.[93] While this difference obviates the need for a molecular dynamics package that can impose constraints on the chosen collective variables, equilibrium path sampling explores the degrees of freedom orthogonal to the reaction coordinate very slowly compared to umbrella sampling, which can make it difficult to reach equilibration and obtain an accurate free energy profile. Equilibrium path sampling was not used in the example study.

4.6 The Information Error Termination Criterion

As researchers continue to increase molecular dynamics system sizes enabled by increasing access to computational resources, the challenge of adequately exploring the

relevant state space becomes more difficult, as does identifying when sufficient sampling has been achieved. To address the question of when a sufficient number of shooting points have been collected during aimless shooting, ATESA introduces a new method based on an assessment of the error in the likelihood maximization procedure. ATESA measures the mean value of the parametric standard errors from the Godambe information matrix[131] (a generalization of the related Fisher information matrix with more lax assumptions regarding the distribution of samples) for a given reaction coordinate model as a function of the amount of data collected, and by default uses a threshold on this parameter as its termination criterion during aimless shooting. This "information error" is a property of likelihood maximization derived from the first and second derivatives of the log likelihood function evaluated at the model optimization solution (the "maximum likelihood estimator"). It is a measure of the "information" about the optimization parameters that is stored in the dataset, as described by the sensitivity of the optimization result to changes in the values of the parameters. As the information error decreases, the confidence that the optimized model is the "best" possible one for the given sets of data/observations and included CVs increases. A more formal treatment of the theory is available in the Supplementary Information.

The information error in the model can be interpreted as a metric of the statistical convergence of sampling *within the explored regions of state space* and for a given reaction coordinate. It is specific to a given reaction coordinate because, insofar as each CV is independent, the precision with which the aimless shooting samples collected represent the underlying distribution for that CV is also independent. This is, a CV with a narrow and/or rapidly explored distribution of values within the transition state ensemble requires less sampling to achieve a given precision, compared to one with a broader and/or more slowly explored distribution of values.

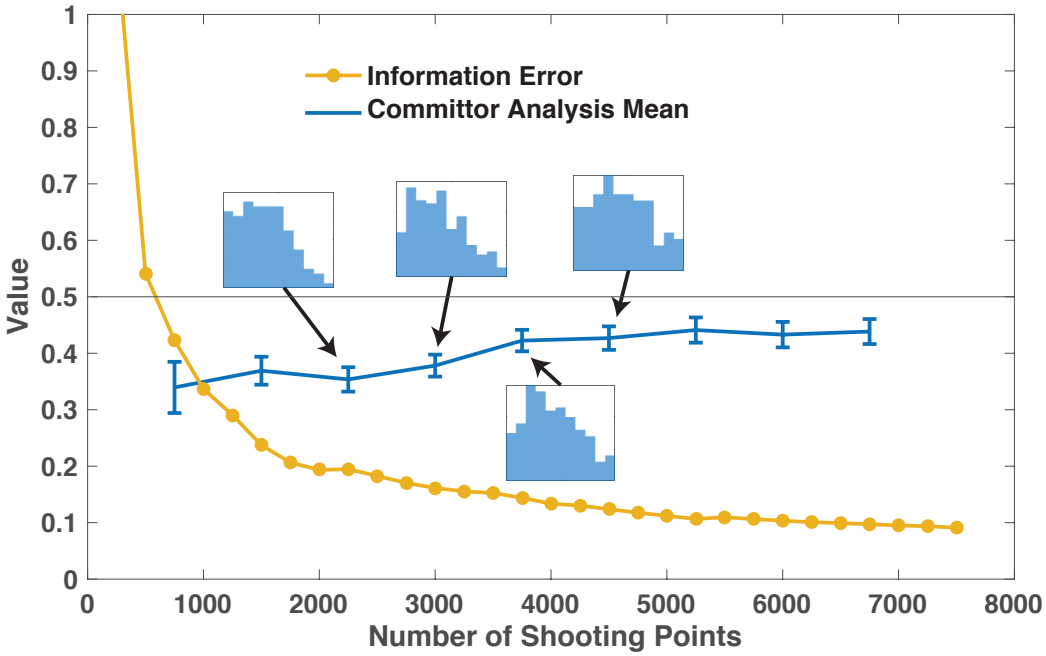

Figure 4.8: **ATESA information error**

Relationship between information error and committor analysis result as aimless shooting sample size increases for an engineered enzyme reaction from our previous work.[132] Each point on the blue line represents the mean value (with standard error of the mean) of the distribution from a separate committor analysis result for each maximum-likelihood three-CV model obtained with the corresponding number of aimless shooting samples (on the horizontal axis), with four of the distributions plotted as histograms in insets to show their shape. The yellow points indicate the information error evaluated for the maximum likelihood three-CV model at each point. When the default information error tolerance threshold of 0.1 is reached after 6250 samples, the committor analysis result at that point is roughly beyond improvement for the given set of observations and sampled regions of state space.

Put another way, the information error indicates the extent to which further sampling from the same distribution is likely to change the maximum likelihood estimate for the chosen set of CVs making up the reaction coordinate. Since information error is derived from likelihood maximization, no additional molecular dynamics simulations are needed, allowing the test to be be efficiently and repeatedly applied during the course of aimless shooting.

Information error should not supplant committor analysis in verifying a reaction coordinate, but should augment it. While the information error measures the *precision* of the maximum likelihood estimate, a mechanistic hypothesis test such as committor analysis is required to assess the *accuracy* of the reaction coordinate as a description of the

transition state ensemble. That is, information error determines whether additional aimless shooting simulation is needed in order to optimize the reaction coordinate to within the desired precision, while committor analysis determines whether the reaction coordinate is composed of an appropriate set of CVs.

A comparison between the information error and the committor analysis result for a three-CV reaction coordinate as more aimless shooting data is gathered for an example reaction is shown in Figure 4.8. As the information error decreases, the committor analysis result improves towards an average close to 0.5 (the ideal value) with a decreasing standard error. That the committor analysis mean levels out at a value other than exactly 0.5 is a function of the likelihood maximization procedure used to build the models at each point (specifically, of leaving out less-important dimensions in order to keep the reaction coordinate model simple).

4.7 Conclusion

Once a scientific tool has been established as important and useful, it becomes important to make it as widely available as possible. ATESA was created to enable more researchers to efficiently use transition path sampling, a powerful approach to determine complex reaction and other transition pathways. Furthermore, automation of nearly every step in the transition path sampling workflow and the introduction of a statistically meaningful sampling termination criterion facilitates consistency between different studies, even when conducted by different researchers. This open-source, Python-based software is publicly available at github.com/team-mayes/atesa, with documentation provided at atesa.readthedocs.io. ATESA has been written with extensibility in mind, and researchers are encouraged to consult the GitHub repository to find information about the latest updates to the software and to report any suggestions, feature requests, or bugs.

CHAPTER V

Conclusion

The previous three chapters represent the first-author manuscripts completed over the course of my doctoral training. Two have been published,[132,133] while the last is currently in preparation. The first was an explanation of a known mechanism; the second a discovery of an unknown mechanism; and the third, a generalization and expansion of the tools used to complete the second. Each is an escalation in terms of ambition and generality, especially with regards to the three themes that I highlighted: **(1)** analysis from a transition pathway perspective; **(2)** a careful modulation of the use of human intuition or bias; and **(3)** the use of the unusual or rare to help explain or understand the usual or common.

5.1 Further Works

In addition to the three preceding manuscripts, I have contributed to two other published projects during this time, which I will briefly describe here.

Click-chemistry enabled directed evolution of glycosynthases for bespoke glycans synthesis.[102]

A related project to the work presented here in Chapter III, this work by our experimental collaborators at Rutgers University (led by Assistant Professor Shishir Chundawat) describes a directed evolution study performed on the same enzyme, the glycosynthase

*Tm*Afc. Having discovered a quadruple mutant with marginally improved glycosynthetic activity compared to the single mutant whose mechanism we discovered in that paper, our collaborators asked us to build a model to explain their findings. Our contribution to this paper includes a free energy profile obtained by reanalyzing the reaction pathway within the context of the quadruple mutant, which matches with their experimental results exceptionally well, as well as a simple study of the changes to the flexibility of the enzyme's backbone caused by the mutations that helps explain the observed improvements.

Biochemical and genetic analysis identify CSLD3 as a β-1,4-glucan synthase that functions during plant cell wall synthesis.[134]

This work involved the study of a series of simulations on family-2 transglycosidase (GT2) enzymes responsible for cell wall synthesis. An open question in the field of plant cell biology is the substrate specificity of the key transglycosidase CSLD3, with conflicting evidence present in the literature for a potential role in both glucan and mannan synthesis pathways. Here, by analogy to three similar transglycosidases (two known glucan synthases and one known mannan synthase), we presented structural arguments supporting our collaborator's hypothesis that CSLD3 is in fact a glucan synthase. Our arguments were largely predicated on a trio of aromatic residues that we identified as being responsible for coordinating mannose substrates in the mannan synthase. Though these aromatic residues are highly conserved across other members of that family, they are totally absent in all of the observed glucan synthases; nevertheless, they were key to informing our understanding of how GT2 enzymes coordinate their substrates in general, exemplifying how unusual features can shed light on more typical ones.

5.2 Generalization and Future Directions

Taken as a whole, the work presented in this thesis contributes to the study of enzymatic processes in the specific (that is, within the contexts of the particular enzymes being studied), but also more broadly in terms of methods and tools. The techniques described here are applicable to solving a huge range of scientific problems within the context of rare event sampling in molecular simulations, and with regard to enzymes especially. However, the direct engineering applicability of the work in these manuscripts is mostly constrained to their applicability as guides for experimental engineering efforts.

To address this, future directions for my research include a software project currently in development, entitled *in silico* Enzyme Engineering, or "*is*EE". Architecturally similar to ATESA, this program automates the full-ensemble analysis of the binding free energy between an enzyme and a transition state (not an analogue) discovered by ATESA, as well as strategic mutations to the enzyme in order to automate the discovery of promising candidates for verification by experiment. Early results on the *Tm*Afc D224G model are highly promising.

Even further down the line, ATESA and *is*EE should prove useful tools for training machine learning algorithms aimed at engineering enzymes, with an eye towards training artificial intelligence to "compose" desired enzymes. Because the extraordinary intricacy of enzymes that bestows upon them their greatest strengths is also responsible for making them more-or-less impenetrable to human intuition, machine learning will certainly prove essential to delivering on the promises of the field of enzymology. Computational tools such as those described here – that incorporate the themes laid out in this work – will be essential tools in developing artificial intelligences that can accomplish these tasks, by putting them directly in dialogue with key simulations results for unsupervised training.

APPENDICES

APPENDIX A

Supplementary Information for: Advantages of a Distant Cellulase Catalytic Base

A.1 Simulation Details

A.1.1 Procession study

Our procession simulations were performed using umbrella sampling simulations with Amber's Targeted MD function. The ring atoms of the leading two β-glucose residues of the substrate (C1 through C5 plus the ring oxygen atoms) were targeted with a force constant of 10 kcal/mol-Å^2 (Amber calculates Targeted MD restraint energy as $V = kN(x - x_0)^2$ where k is the force constant and N is the number of atoms in the restraint mask, here 12) to reach a given RMSD value between 0.25 and 12 Å in increments of 0.25 Å relative to the initial position (where the leading ring occupied the -1 site, as it does just after product release), resulting in 48 "windows" for sampling. Steric effects from the enzyme's active site groove constrain this targeted motion to approximate translation along the axis of the groove. In order to ensure that the build steps did not encounter forces so high as to potentially disrupt covalent bonds in the active site, the simulations were performed in blocks of 1 Å with the last step of each block serving as the initial structure for the next block (while the reference structure for targeted MD remained the 0 Å structure). Furthermore,

to minimize the possibility of the targeted motion forcing the enzyme into a conformation outside its natural phase space, the entire protein structure was restrained with a force of 200 kcal/mol-Å^2 while building the windows and the simulation timestep was reduced to 0.5 fs. The first half of the procession (from 0.25 to 5.75 Å) was initiated using the "pre-slide" structure as described in our previous work, while the second half (6 to 12 Å) began with the "slide" structure (see Figure 2.2 in the main text).

After each of the windows was built, the general protein restraint was removed and Targeted MD on the substrate remained turned on with a force constant of 5 kcal/mol-Å^2 to act as the controlled transformation on the energy landscape that defines umbrella sampling. For each of the windows in both the wild-type and mutant protein, at least 500 ps of sampling MD was produced, of which at least the first 150 ps were discarded to allow the protein structure to relax around the new substrate position. Conversion of these data into the PMFs shown in the main text was performed using umbrella integration at 300 K over 2000 bins, with the integration error reference at the left-most side of the plot.[49,50] While umbrella integration does not strictly require histogram overlap (in contrast to WHAM; here, it affects the calculated error), we plotted histograms of sampling in each bin to verify that all regions of the CV were sampled (Figure A.1).

As mentioned above, the PMFs presented were constructed from windows that used two separate initial structures: "pre-slide" for the windows with RMSD values less than 6 Å, and "slide" for the others. The key difference between these structures, apart from relative substrate position, are the conformations of the serine (active-site) loop. We examined our simulations for potential hysteresis due to this deviation. Adding a bias only along the RMSD used for the umbrella sampling did not lead to freely sampling this loop site motion. As shown in Figure A.2, two conformations that both correspond to an RMSD of 6 Å each still reflected the loop position of the initial structure. In the "pre-slide" structure and Figure

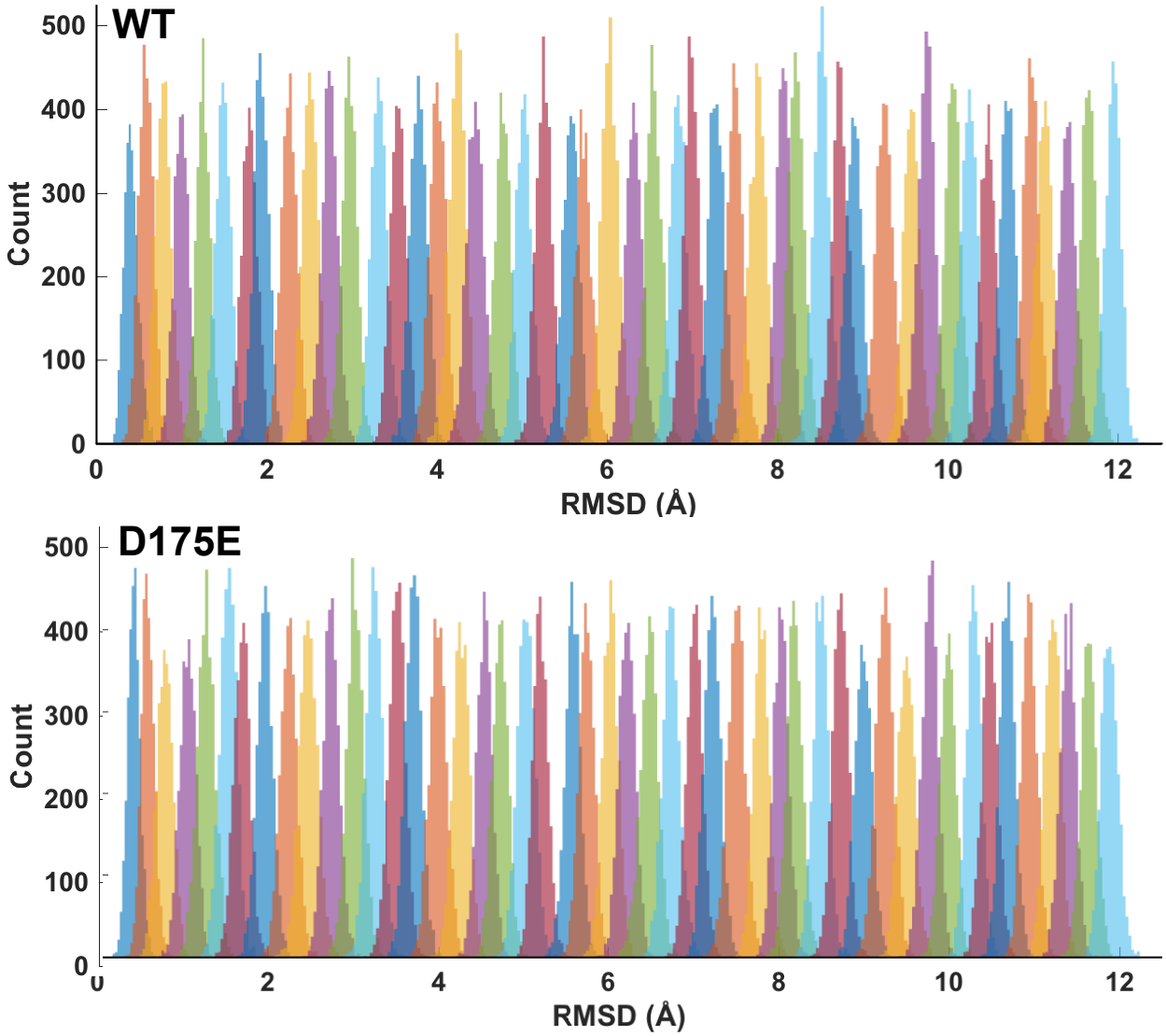

Figure A.1: ***Tr*Cel6A umbrella sampling histograms: procession**

Histograms of processivity sampling data for each of the 48 sampling windows, for both enzyme types, used in creating the PMFs shown in Figure 2.2 in the main text.

A.2A, Ser-181 is hydrogen bonded to Asp-175 and Asn-182 is projected into the active site tunnel, whereas in the “slide” structure and Figure A.2B, Asn-182 is hydrogen bonded to Asp-175 instead (see also the main text Figure 2.3).

The present approach to creating the wild-type PMF and that in our previous work[8] differs only for the windows with centers with RMSD less than 6 Å: in the previous work, all windows were created from initial “slide” conformations, in contrast to the current work

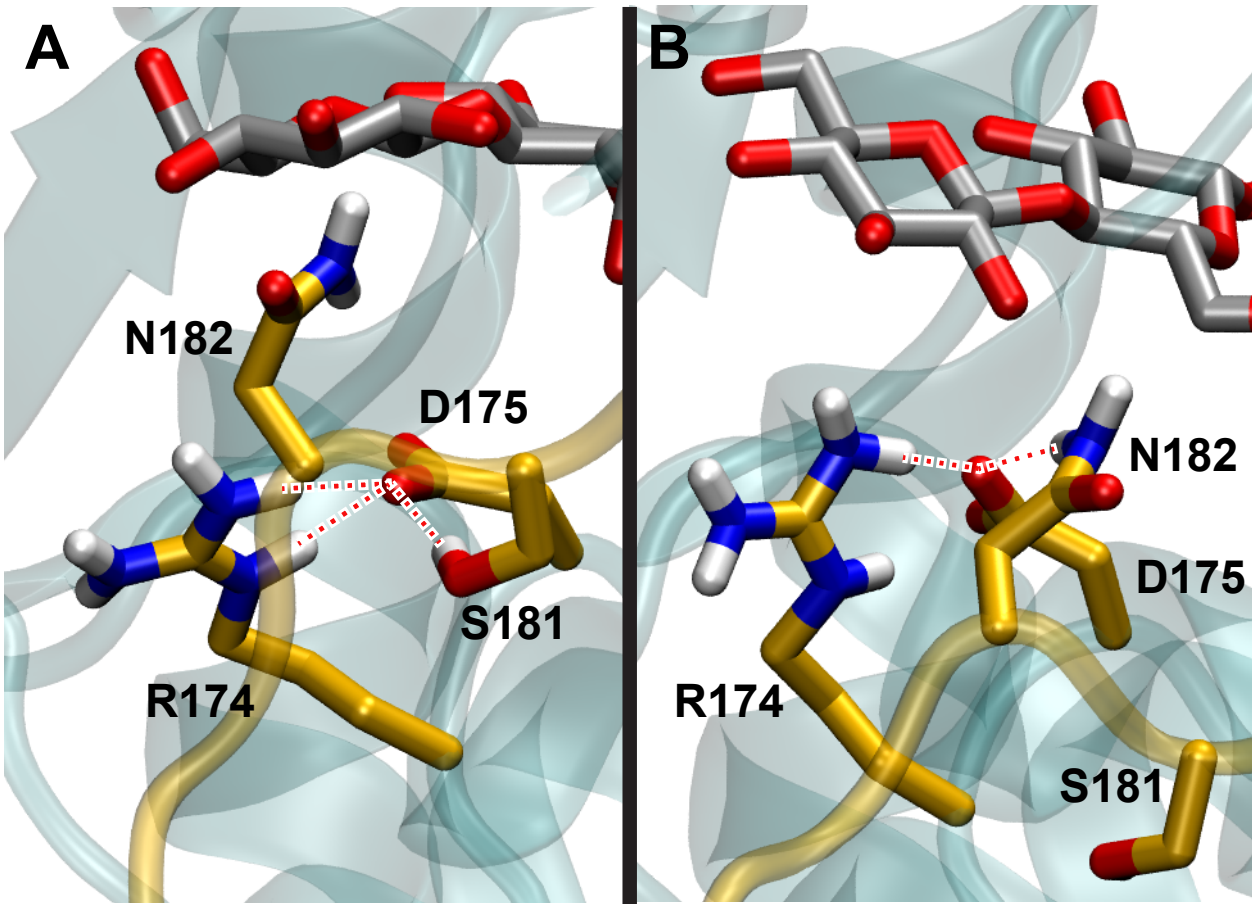

Figure A.2: ***Tr*Cel6A slide vs. pre-slide structures**

Snapshots of two conformations of wild-type enzyme which both have an RMSD of approximately 6 Å on the CV used to create the PMF in Figure 2.2 of the main text. (A) The serine (active-site) loop (backbone shown as yellow cartoon) in the position from the "pre-slide" conformation. (B) The serine loop in the "slide" conformation, which is closer to the hydrolytically active position.

using the "pre-slide" conformation to seed the windows with centers less than 6 Å. Despite this difference, the PMFs are similiar (within expected differences due to finite sampling). This indicates that the PMF is insensitive to the loop orientation at RMSD values less than 6 Å. This region of the PMF, closer to and including the pre-slide conformation, has the -2 binding site unoccupied. However, at higher values (greater than 6 Å) interactions between the substrate and residues on the serine loop (including Asn-182) were unfavorable when sampled using structures produced entirely from initial configurations in the "pre-slide" conformation (data not shown). The window centered at 6 Å was the first window where

the substrate approaches close enough to the relevant residues for the energy to be affected, and thus was chosen as the dividing point. At the timescales accessible to our simulations, we were not able to sample this loop repositioning. This deficiency, as well as the previously found deficiency in sampling ring puckering of the second-to-leading glycosyl ring as it enters the -1 binding site, indicates that at least two features of the processive mechanism are not properly captured by the CV that we chose here for computational efficiency. Thus, barrier heights for processivity are likely underestimated, and absolute values from the PMF are not expected to be accurate. However, in this study we are focused on the comparison of the wild-type PMF to that from the D175E mutant. Since the catalytic base residue is not on the serine loop and is not involved in puckering, the difference should not affect the portion of the potential energy surface not sampled. It is likely that we properly sampled the portions of the PMF that are affected by this difference, which allows us to discuss the qualitative differences in procession for these enzymes.

A.1.2 Active site conformation study

The Asp-221 and Glu-175 umbrella sampling simulations corresponding to Figures 4 and 6 in the main text, respectively, were performed using windows with dihedral angle center values from 55 to 200 degrees for the Asp-221 study and -190 to 85 degrees for the Glu-175 study, in steps of 5 degrees. The Glu-175 dihedral was applied to the angle defined by the sequence of atoms: HB2, CB, CG, HG1. In each window, the restraint was harmonic with weight 250 kcal/mol-rad^2. Separate "build" simulations to produce the starting structure for each window were performed before sampling using 10,000 1-fs steps, while sampling was performed over 500,000 2-fs steps. The raw data from these simulations is shown in Figure A.3 and Figure A.4, respectively. Conversion of these data into the PMFs shown in the main text was performed using umbrella integration at 300 K

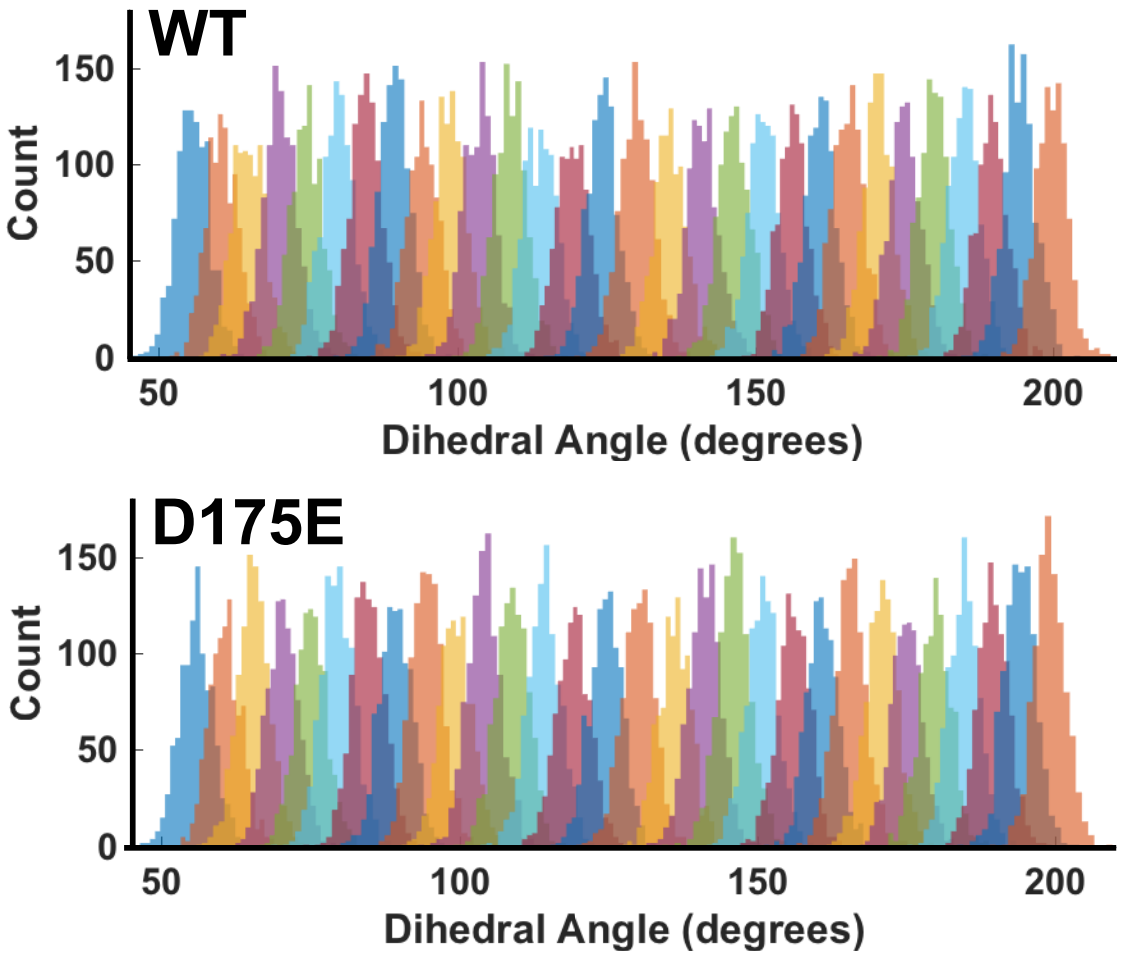

Figure A.3: ***Tr*Cel6A umbrella sampling histograms: Asp-221**

Histograms of Asp-221 umbrella sampling data for each of the sampling windows for both enzyme types. These are the raw data constituting the results of our Asp-221 dihedral rotation study, Figure 2.4 in the main text. These data were fed into the umbrella integration method.

over 2000 bins, with the integration error reference at the left-most side of the plot.[49,50]

A.1.3 Homology study

Simulation-ready structures were built as described above from the *Tf*Cel9A crystal structure described in the main text, and from the product structure obtained in our previous work.[8] In addition, copies of each enzyme with the catalytic base (*Tr*Cel6A Asp-175 or *Tf*Cel9A D58) mutated to alanine (*via* manual edits to the .pdb structure files) were prepared. 300-ps trajectories of both wild-type and both mutant systems were produced

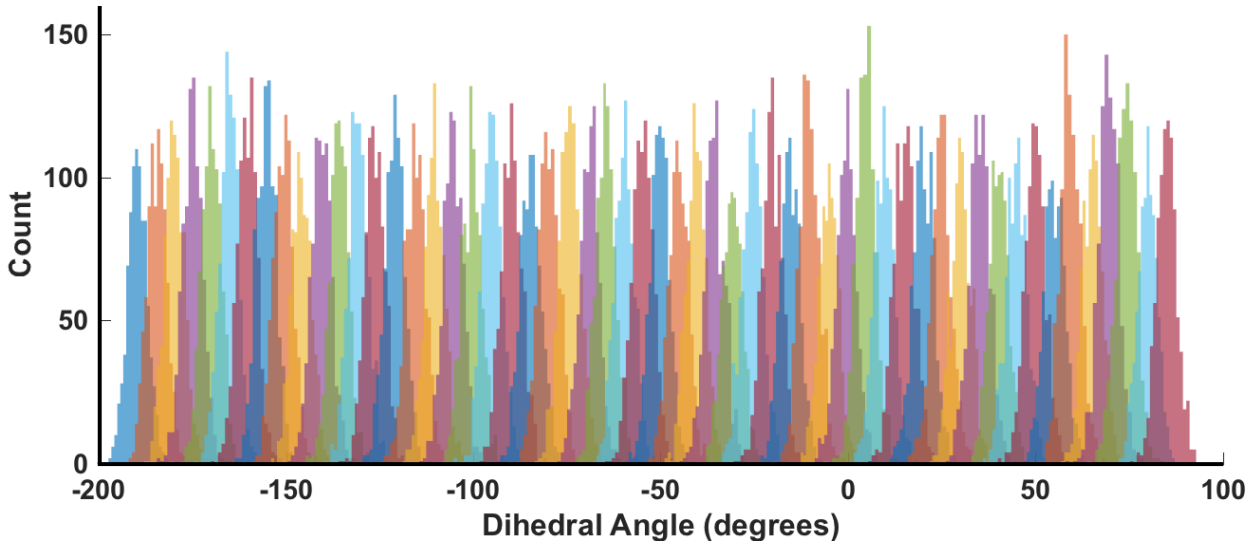

Figure A.4: ***Tr*Cel6A umbrella sampling histograms: Glu-175**

Histograms of Glu-175 umbrella sampling data for each of the sampling windows for the mutant enzyme. These are the raw data constituting the results of our Glu-175 dihedral rotation study, Figure 2.6 in the main text. These data were fed into the umbrella integration method.

for analysis.

The free energy of ligand binding for each system was approximated using Amber's MMPBSA.py program to perform a generalized Born implicit solvation analysis of the system.[135–137] This program requires separate topology files for the ligand and protein, which were produced using the "strip" command in another Amber program, cpptraj.[138] Because in the product state the ligand has been broken into two molecules and because we are uninterested in the interaction potential between the base residue and the portion of the substrate not subject to hydrolytic attack during catalysis, this section of the ligand was included in the "protein" mask for the purposes of generalized Born analysis, leaving the measured effects wholly attributable to the binding energy of the substrate in product binding site (-2 and -1 binding sites).

Although the term "free energy" is used here to refer to the energy terms measured in this analysis, no measurement was made of the entropic contribution to the free energy

that is associated with ligand binding; *i.e.*, no attempt was made to calculate the entropy of the unbound states of the ligands. Therefore, the reported energy value is not a complete description of the relative free energies of binding, but the error attributable to this effect is expected to be quite small because the bound and unbound states in each case are not expected to be significantly different in terms of conformational dynamics. See Srinivasan *et al.* for a more detailed discussion of this type of error.[139]

APPENDIX B

Supplementary Information for: Mechanism of Oligosaccharide Synthesis via a Mutant GH29 Fucosidase

B.1 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.

B.1.1 Likelihood maximization.

We used the inertial likelihood maximization algorithm of Peters.[92] 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.[125]

B.1.2 Committor analysis.

After an RC has been produced by likelihood maximization, it must be validated by committor analysis.[93] 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.[140]

B.2 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.[76] The additional parameters that we added are shown in Table B.1, 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,

Table B.1: ***Tm*Afc substrate force field parameters**

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	Equilibrium Value	Source
c3-ni	277.5	1.49	[79]
ni-nd	710.0	1.34	[79]
nd-ne	1312.0	1.14	[79]
c3-c3-ni	74.8	113.36	[79]
h1-c3-ni	68.3	108.87	[79]
c3-ni-nd	64.0	115.60	[79]
ni-nd-ne	42.4	173.54	[79]
ni-c3-os	70.04	111.230	Analogy to n2-c3-os from GAFF
c3-ni-nd-ne	0.25	180.00	[79]
c3-c3-ni-nd	11.11	0.00	[79]
h1-c3-ni-nd	11.11	0.00	[79]
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

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 0.1 kcal/mol-Å. The structure comparisons are shown in Figure B.1.

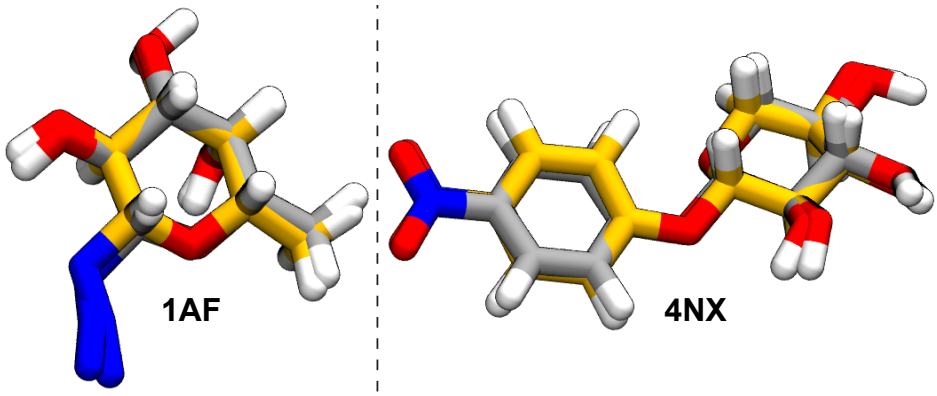

Figure B.1: ***Tm*Afc substrate: QM vs MM**

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 B.1 (silver). The structures are fitted atop one another to minimize the RMSD between like atoms, not including hydrogens.

B.3 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[90] 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 B.2. 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

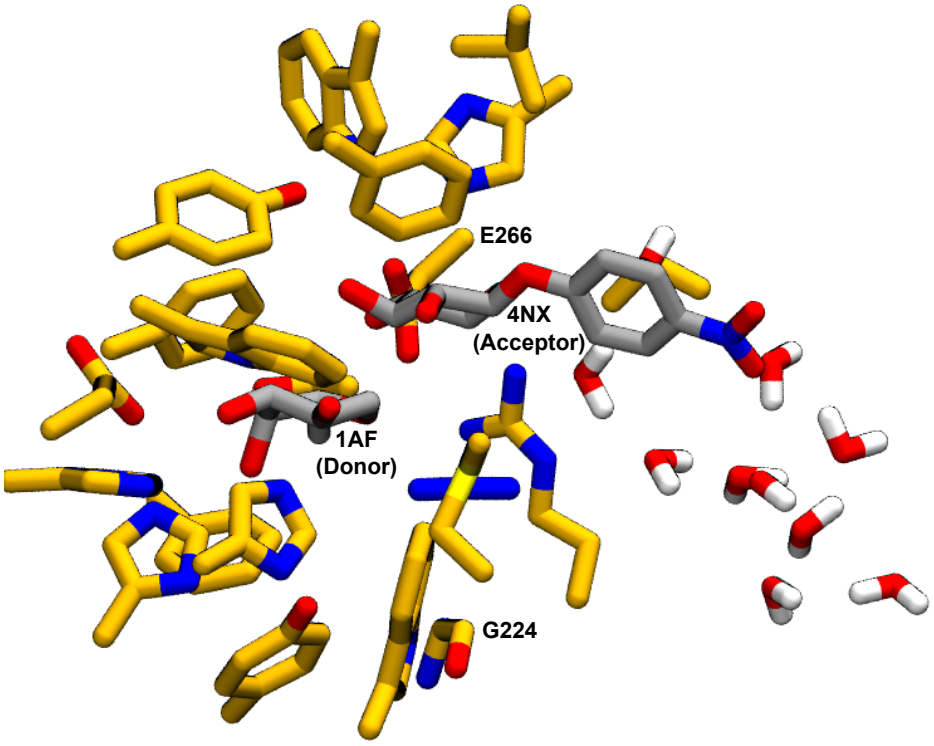

Figure B.2: ***Tm*Afc QM region**

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.

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, among 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.

B.4 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 B.2 and B.3. 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.

B.5 Umbrella Sampling

Umbrella sampling is a means of measuring the free energy along a given coordinate using a series of restrained simulations where the restraint is defined along that coordinate. The distribution of samples collected in each window can be deconvolved from the

Table B.2: ***Tm*Afc likelihood maximization dimensions**

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

Table B.3: ***Tm*Afc CV atom identities**

Definitions of atom indices corresponding to those in Table B.2. Atom names correspond to those in the standard residue definitions in the Amber force field[90] 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

applied restraint in order to measure the underlying free energy profile using a number of algorithms. Specifically, we used the Multistate Bennett Acceptance Ratio, or "MBAR", implemented in Python as pymbar.[129] Our umbrella sampling simulations were performed using four redundant simulations in each window, with windows spaced in steps of 0.5 from -9 to 10 along the unitless reaction coordinate. The restraint weight was 20 kcal/mol, the step size was 0.5 fs, and a DFTB electronic temperature of 100 K was applied to smooth out inconsistencies associated with non-convergent QM calculations. The data was equilibrated and decorrelated automatically using the "pymbar.timeseries.detectEquilibration" method.[130] The equilibrated and decorrelated samples used to calculate the free energy profile in Figure 3.4 in the main text are shown graphically in Figure B.3.

B.6 Equilibrium Constant from Cobucci-Ponzano *et al.*

The equilibrium constant provided in the main text for the reactions performed by Cobucci-Ponzano *et al.*[66] 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:

(B.1) $$K_{eq,\alpha1,4} = \frac{[\alpha1,4\ product]_{eq}[free\ azide]_{eq}}{[acceptor]_{eq}[donor]_{eq}}$$

where:

(B.2) $$[\alpha1,4\ product] = f_{\alpha1,4}\eta([donor]_0 - [donor]_{eq})$$

(B.3) $$[free\ azide] = [donor]_0 - [donor]_{eq}$$

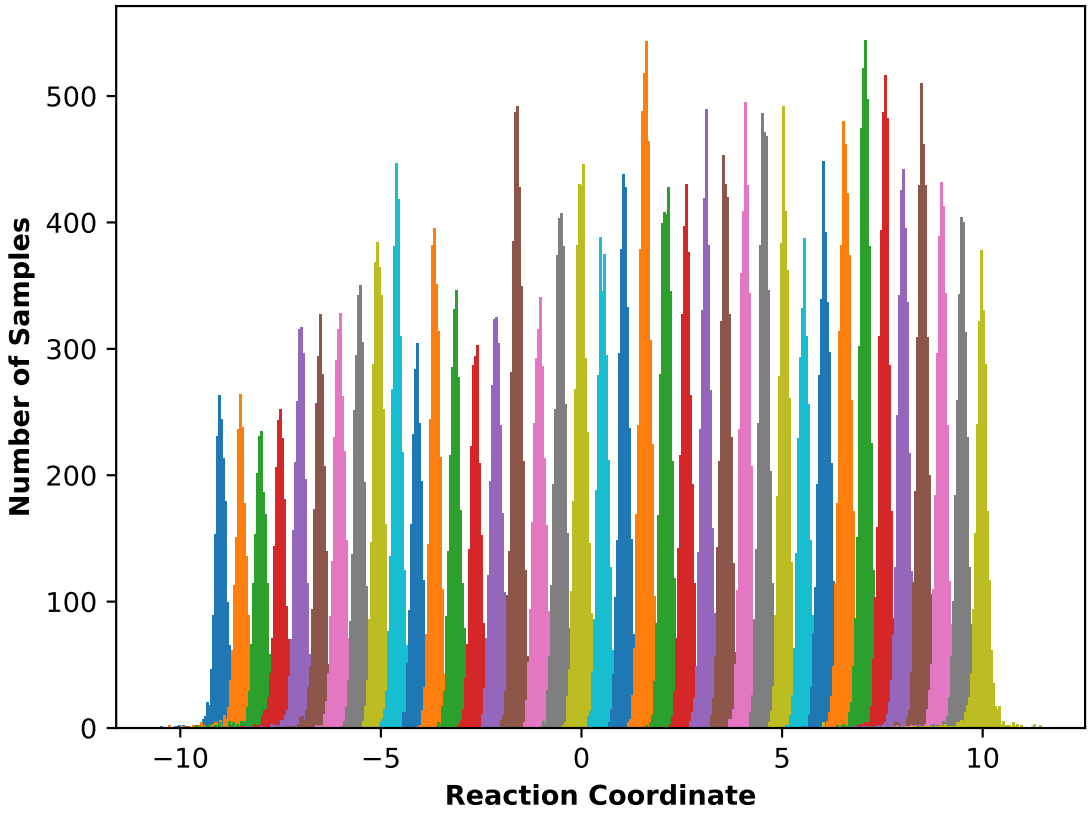

Figure B.3: ***Tm*Afc umbrella sampling histograms**

Histograms of umbrella sampling data after decorrelation and equilibration. The overlap of each histogram with its neighbors demonstrates that there are no unsampled regions of the free energy profile.

(B.4) $$[acceptor]_{eq} = [acceptor]_0 - \eta([donor]_0 - [donor]_{eq})$$

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.[66] An analogous calculation was performed for the α-1,3 reaction.

APPENDIX C

Supplementary Information for: ATESA: an Automated Aimless Shooting Workflow

C.1 Example Study Simulation Details

Here we detail the settings used in the ethyl chlorosulfite decomposition simulations. All simulations were performed using Amber 18,[112] except for umbrella sampling, which was based on the custom version of Amber 12 described in the main text. The simulation timestep in all cases was 1 femtosecond. The simulations were performed in non-periodic vacuum at a temperature of 300 Kelvin using the Andersen thermostat with a randomization frequency of 100 steps.[141] The entire molecule except for the methyl group was treated using the semi-empirical PM3 quantum mechanical ("QM") model[142] during the *find_ts*, *aimless_shooting*, and *committor_analysis* simulations, and the QM method was changed to RM1[143] for *umbrella_sampling* in order to obtain a higher degree of energetic accuracy. The methyl group was treated using the generalized Amber force field ("gaff").[75] This partial QM treatment was chosen to preclude an off-pathway reaction involving a bond between the methyl group and the leaving group. Within the non-QM region, SHAKE was used to fix the carbon-hydrogen bond lengths.[45]

C.2 A More Formal Treatment of Information Error

In the main text, we introduce the concept of Godambe information as a convergence criterion for aimless shooting. Here, we provide an overview of the Godambe information method and describes its application to aimless shooting.

The Godambe information is a more general form of the more familiar Fisher information, which is a property of maximum likelihood estimators (MLEs).[144] Under certain assumptions about the distribution of the underlying data, the inverse of the Fisher Information Matrix for a given MLE provides an estimate of the covariance matrix, and can therefore be used to evaluate the parameter error.

The Fisher Information Matrix can be evaluated empirically during likelihood maximization as the Hessian of the likelihood function evaluated at the calculated MLE:

$$I_{ij} = \frac{\partial^2 L}{\partial\theta_i \partial\theta_j}|_{\theta_{MLE}} \tag{C.1}$$

where $\theta = \{\theta_1, \theta_2, ...\}$ is the vector of model parameters being optimized, L is the log likelihood function, and θ_{MLE} is the vector of model parameter values that maximizes L.

For this relationship to hold true requires that the data being used for likelihood maximization be, among other assumptions, independently sampled from the underlying distribution. However, data obtained through aimless shooting is *not* independently sampled: each sample is, by definition, chosen to be very nearby to a previous sample, as a means of ensuring that it remains nearby the rare event separatrix. For this reason, an MLE built using aimless shooting data is called "misspecified." Contrary to the name, misspecified MLEs can still produce valid models, particularly when we are more interested in obtaining a *useful* model than a mathematically ideal one; however, it requires that we employ a more generalized version of the information error:

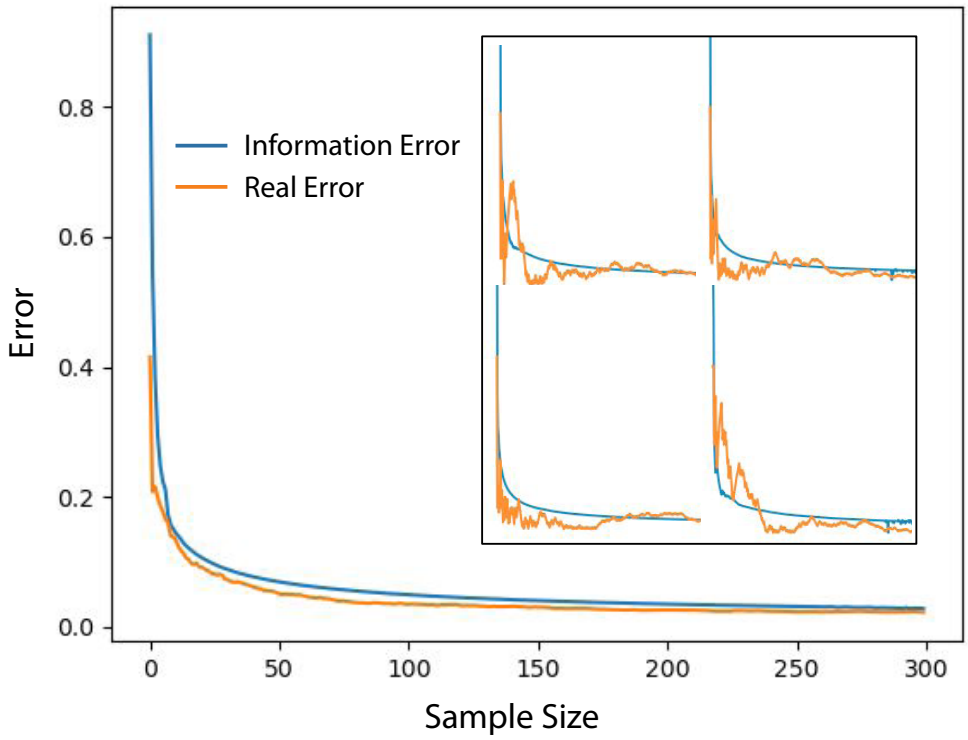

Figure C.1: **ATESA information error coin flip model**

Relationship between real error (orange) and Fisher information error (blue) for the coin-flipping "experiment". Inset are four examples of individual trials of 300 coin flips each, while the larger plot is the mean across 100 such trials.

$$G(\theta) = H(\theta)J(\theta)^{-1}H(\theta) \tag{C.2}$$

where $H(\theta)$ is the Hessian of θ, $J(\theta)$ is the Jacobian, and $G(\theta)$ is the Godambe information matrix (sometimes also called the "Sandwich Estimator" due to the structure of its relationship to the Hessian and Jacobian).[145]

In general, misspecified models retain the standard $n^{1/2}$ asymptotic behavior for measures of parameter error. However, in practice and in particular for smaller numbers of samples, the asymptotic behavior is not always well-behaved, owing to phenomena such as sample autocorrelation. For this reason, our proposed implementation of Godambe information as a termination criterion in aimless shooting is to evaluate the mean value of the parameter errors for the best available model every so often during sampling and to

place a threshold on this value, without making any assumptions about the shape of the relationship between the mean parametric error (the “information error”) and the number of samples.

As an example to graphically demonstrate the applicability of information error for models obtained through MLE, we present a simple “experiment” involving the flipping of a weighted coin (Figure C.1). For simplicity, the samples here are indeed independently distributed. A one-dimensional model of the coin’s weight is maximized via log-likelihood maximization based on successive trials. The weight of the coin is 0.5564 (chosen such that the exact correct value cannot be arrived at with a small number of samples by coincidence) and the initial guess for the model is 0.1. As expected, the information error (square root of the first (and only) term in the Fisher information matrix) is an excellent estimator of the actual difference between the model estimate and the real weight.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] E. D. Schneider and J. J. Kay. Life as a Manifestation of the Second Law of Thermodynamics. *Math. Comput. Model.*, 19(6-8):25–48, 1994.

[2] A. S. Wells, G. L. Finch, P. C. Michels, and J. W. Wong. Use of enzymes in the manufacture of active pharmaceutical ingredients - A science and safety-based approach to ensure patient safety and drug quality. *Org. Process. Res. Dev.*, 16(12):1986–1993, 2012.

[3] D. Klein-Marcuschamer, P. Oleskowicz-Popiel, B. A. Simmons, and H. W. Blanch. The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnol. Bioeng.*, 109(4):1083–1087, 2012.

[4] L. Gianfreda, M. A. Rao, R. Scelza, and M. de la Luz Mora. *Role of Enzymes in Environment Cleanup/Remediation*. Elsevier Inc., 2016.

[5] C. M. Heckmann and F. Paradisi. Looking Back: A Short History of the Discovery of Enzymes and How They Became Powerful Chemical Tools. *ChemCatChem*, 12:6082–6102, 2020.

[6] E. Kellenberger. The evolution of molecular biology. *EMBO Reports*, 5(6):546–549, 2004.

[7] D. E. Koshland, Jr. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.*, 44(2):98–104, 1958.

[8] H. B. Mayes, B. C. Knott, M. F. Crowley, L. J. Broadbelt, J. Ståhlberg, and G. T. Beckham. Who's on base? Revealing the catalytic mechanism of inverting Family 6 glycoside hydrolase. *Chem. Sci.*, 7:5955–5968, 2016.

[9] M. J. Biddy, R. Davis, D. Humbird, L. Tao, N. Dowe, M. T. Guarnieri, J. G. Linger, E. M. Karp, D. Salvachúa, D. R. Vardon, and G. T. Beckham. The techno-economic basis for coproduct manufacturing to enable hydrocarbon fuel production from lignocellulosic biomass. *ACS Sustain. Chem. Eng.*, 4:3196–3211, 2016.

[10] S. M. Cragg, G. T. Beckham, N. C. Bruce, T. D. H. Bugg, D. L. Distel, P. Dupree, A. G. Etxabe, B. S. Goodell, J. Jellison, J. E. McGeehan, S. J. McQueen-Mason, K. Schnorr, P. H. Walton, J. E. M. Watts, and M. Zimmer. Lignocellulose degradation mechanisms across the Tree of Life. *Curr. Opin. Chem. Biol.*, 29:108–119, 2015.

[11] F. García-Guevara, M. Avelar, M. Ayala, and L. Segovia. Computational tools applied to enzyme design – a review. *Biocatal.*, 1:109–117, 2015.

[12] H. Renata, Z. J. Wang, and F. H. Arnold. Expanding the enzyme universe: Accessing non-natural reactions by mechanism-guided directed evolution. *Angew. Chem., Int. Ed.*, 54:3351–3367, 2015.

[13] C. M. Payne, B. C. Knott, H. B. Mayes, H. Hansson, M. E. Himmel, M. Sandgren, J. Ståhlberg, and G. T. Beckham. Fungal Cellulases. *Chem. Rev.*, 115:1308–1448, 2015.

[14] C. Boisset, C. Fraschini, M. Schülein, B. Henrissat, and H. Chanzy. Imaging the enzymatic digestion of bacterial cellulose ribbons reveals the endo character of the cellobiohydrolase Cel6A from *Humicola insolens* and its mode of synergy with cellobiohydrolase Cel7A. *Appl. Environ. Microbiol.*, 66(4):1444–1452, 2000.

[15] J. Ståhlberg, G. Johansson, and G. Pettersson. Trichoderma reesei has no true exo-cellulase: all intact and truncated cellulases produce new reducing end groups on cellulose. *Biochimica et Biophys. Acta*, 1157(1):107–113, 1993.

[16] B. Mertz, R. S. Kuczenski, R. T. Larsen, A. D. Hill, and P. J. Reilly. Phylogenetic analysis of family 6 glycoside hydrolases. *Biopolym.*, 79(4):197–206, 2005.

[17] G. T. Beckham, J. Ståhlberg, B. C. Knott, M. E. Himmel, M. F. Crowley, M. Sandgren, M. Sørlie, and C. M. Payne. Towards a molecular-level theory of carbohydrate processivity in glycoside hydrolases. *Curr. Opin. Biotechnol.*, 27:96–106, 2014.

[18] A. Nakamura, T. Tasaki, D. Ishiwata, M. Yamamoto, Y. Okuni, A. Visootsat, M. Maximilien, H. Noji, T. Uchiyama, M. Samejima, K. Igarashi, and R. Iino. Single-molecule imaging analysis of binding, processive movement, and dissociation of cellobiohydrolase *Trichoderma reesei* Cel6A and its domains on crystalline cellulose. *J. Biol. Chem.*, 291:22404–22413, 2016.

[19] J. Rouvinen, T. Bergfors, T. Teeri, J. K. C. Knowles, and T. A. Jones. Three-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*. *Sci.*, 249:380–386, 1990.

[20] D. E. Koshland, Jr. Stereochemistry and the mechanism of enzymatic reactions. *Biol. Rev.*, 28:416–436, 1953.

[21] A. Koivula, L. Ruohonen, G. Wohlfahrt, T. Reinikainen, T. T. Teeri, K. Piens, M. Claeyssens, M. Weber, A. Vasella, D. Becker, M. L. Sinnott, J.-y. Zou, G. J. Kleywegt, M. Szardenings, J. Ståhlberg, and T. A. Jones. The active site of cellobiohydrolase Cel6A from Trichoderma reesei: The roles of aspartic acids D221 and D175. *J. Am. Chem. Soc.*, 124(34):10015–10024, 2002.

[22] D. Riccardi, P. König, H. Guo, and Q. Cui. Proton transfer in carbonic anhydrase is controlled by electrostatics rather than the orientation of the acceptor. *Biochem.*, 47(8):2369–2378, 2008.

[23] D. E. Wolfgang and D. B. Wilson. Mechanistic studies of active site mutants of Thermomonospora fusca endocellulase E2. *Biochem.*, 38(30):9746–9751, 1999.

[24] H. G. Damude, S. G. Withers, D. G. Kilbum, R. C. Miller, and R. A. J. Warren. Site-Directed Mutation of the Putative Catalytic Residues of Endoglucanase CenA from Cellulomonas fimi. *Biochem.*, 34(7):2220–2224, 1995.

[25] M. Sandgren, M. Wu, S. Karkehabadi, C. Mitchinson, B. R. Kelemen, E. A. Larenas, J. Ståhlberg, and H. Hansson. The structure of a bacterial cellobiohydrolase: The catalytic core of the thermobifida fusca family GH6 cellobiohydrolase Cel6B. *J. Mol. Biol.*, 425(3):622–635, 2013.

[26] A. J. Thompson, T. Heu, T. Shaghasi, R. Benyamino, A. Jones, E. P. Friis, K. S. Wilson, and G. J. Davies. Structure of the catalytic core module of the Chaetomium thermophilum family GH6 cellobiohydrolase Cel6A. *Acta Crystallogr. Sect. D: Biol. Crystallogr.*, 68(8):875–882, 2012.

[27] A. Varrot, S. Hastrup, M. Schülein, and G. J. Davies. Crystal structure of the catalytic core domain of the family 6 cellobiohydrolase II, Cel6A, from Humicola insolens, at 1.92 Å resolution. *The Biochem. J.*, 337:297–304, 1999.

[28] A. M. Larsson, T. Bergfors, E. Dultz, D. C. Irwin, A. Roos, H. Driguez, D. B. Wilson, and T. A. Jones. Crystal structure of Thermobifida fusca endoglucanase Cel6A in complex with substrate and inhibitor: the role of tyrosine Y73 in substrate ring distortion. *Biochem.*, 44(39):12915–12922, 2005.

[29] W. Zhou, D. C. Irwin, J. Escovar-Kousen, and D. B. Wilson. Kinetic studies of Thermobifida fusca Cel9A active site mutant enzymes. *Biochem.*, 43(30):9655–9663, 2004.

[30] Y. Li, D. C. Irwin, and D. B. Wilson. Processivity, Substrate Binding, and Mechanism of Cellulose Hydrolysis by Thermobifida fusca Cel9A. *Appl. Environ. Microbiol.*, 73(10):3165–3172, 2007.

[31] J. Sakon, D. C. Irwin, D. B. Wilson, and A. P. Karplus. Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca. *Nat. Struct. Biol.*, 4(10):810–818, 1997.

[32] H. B. Mayes, L. J. Broadbelt, and G. T. Beckham. How sugars pucker: Electronic structure calculations map the kinetic landscape of fivebiologically paramount monosaccharides and their implications for enzymatic catalysis. *J. Am. Chem. Soc.*, 136(3):1008–1022, 2014.

[33] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[34] J.-y. Zou, G. J. Kleywegt, J. Ståhlberg, H. Driguez, W. Nerinckx, M. Claeyssens, A. Koivula, T. T. Teeri, and T. A. Jones. Crystallographic evidence for substrate ring distortion and protein conformational changes during catalysis in cellobiohydrolase Cel6A from Trichoderma reesei. *Struct.*, 7:1035–1045, 1999.

[35] M. Wu, L. Bu, T. V. Vuong, D. B. Wilson, M. F. Crowley, M. Sandgren, J. Ståhlberg, G. T. Beckham, and H. Hansson. Loop motions important to product expulsion in the Thermobifida fusca glycoside hydrolase family 6 cellobiohydrolase from structural and computational studies. *J. Biol. Chem.*, 288(46):33107–33117, 2013.

[36] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[37] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and J. MacKerell, Alexander D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and sidechain χ1 and χ2 dihedral angles. *J. Chem. Theory Comput.*, 8(9):3257–3273, 2012.

[38] J. MacKerell, Alexander D., M. Feig, and C. L. Brooks. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.*, 126(3):698–9, 2004.

[39] J. MacKerell, Alexander D., D. Bashford, M. Bellott, J. Dunbrack, R. L., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. I. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, J. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102(97):3586–3616, 1998.

[40] D. Beglov and B. Roux. Finite Representation of an Infinite Bulk System: Solvent Boundary Potential for Computer Simulations. *J. Chem. Phys.*, 100(12):9050–9063, 1994.

[41] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[42] M. F. Crowley, M. J. Williamson, and R. C. Walker. CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software. *Int. J. Quantum Chem.*, 109:3767–3772, 2009.

[43] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, Walker, J. Wang, R. M. Wolf, X. Wu, and P. A. Kollman. *Amber 14*. University of California, San Francisco, 2014.

[44] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[45] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 1992.

[46] T. A. Andrea, W. C. Swope, and H. C. Andersen. The role of long ranged forces in determining the structure and properties of liquid water. *J. Chem. Phys.*, 79(9):4576–4584, 1983.

[47] L. Bu, M. F. Crowley, M. E. Himmel, and G. T. Beckham. Computational investigation of the pH dependence of loop flexibility and catalytic function in glycoside hydrolases. *J. Biol. Chem.*, 288(17):12175–12186, 2013.

[48] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *J. Mol. Graph.*, 14:33–38, 1996.

[49] J. Kästner and W. Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "umbrella integration". *J. Chem. Phys.*, 123(14):144104–144108, 2005.

[50] J. Kästner and W. Thiel. Analysis of the statistical error in umbrella sampling simulations by umbrella integration. *J. Chem. Phys.*, 124(23):234106–234112, 2006.

[51] M. Stroet and E. Deplazes. Python implementation of the umbrella integration method for potential of mean force (pmf) calculations. https://doi.org/10.5281/zenodo.164996.

[52] A. Varki, R. D. Cummings, J. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, and P. H. Seeberger. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 3rd edition, 2017.

[53] A. Varki. Biological roles of oligosaccharides: All of the theories are correct. *Glycobiol.*, 3(2):97–130, 1993.

[54] P. H. Seeberger. Automated oligosaccharide synthesis. *Chem. Soc. Rev.*, 37(1):19–28, 2008.

[55] P. M. Danby and S. G. Withers. Advances in Enzymatic Glycoside Synthesis. *ACS Chem. Biol.*, 11(7):1784–1794, 2016.

[56] L.-X. Wang and B. G. Davis. Realizing the promise of chemical glycobiology. *Chem. Sci.*, 4(9):3381–3394, sep 2013.

[57] D. Lloyd and C. S. Bennett. An Improved Approach to the Direct Construction of 2-Deoxy-β-Linked Sugars: Applications to Oligosaccharide Synthesis. *Chem. - Eur. J.*, 24:7610–7614, 2018.

[58] M. Panza, S. G. Pistorio, K. J. Stine, and A. V. Demchenko. Automated Chemical Oligosaccharide Synthesis: Novel Approach to Traditional Challenges. *Chem. Rev.*, 118:8105–8150, 2018.

[59] Z. Armstrong and S. G. Withers. Synthesis of glycans and glycopolymers through engineered enzymes. *Biopolym.*, 99(10):666–674, oct 2013.

[60] S. M. Hancock, M. D. Vaughan, and S. G. Withers. Engineering of glycosidases and glycosyltransferases. *Curr. Opin. Chem. Biol.*, 10(5):509–519, 2006.

[61] K. Schmölzer, M. Lemmerer, and B. Nidetzky. Glycosyltransferase cascades made fit for chemical production: Integrated biocatalytic process for the natural polyphenol C-glucoside nothofagin. *Biotechnol. Bioeng.*, 115(3):545–556, 2018.

[62] L. F. Mackenzie, Q. Wang, R. A. J. Warren, and S. G. Withers. Glycosynthases: Mutant glycosidases for oligosaccharide synthesis. *J. Am. Chem. Soc.*, 120(22):5583–5584, 1998.

[63] B. Cobucci-Ponzano, A. Strazzulli, M. Rossi, and M. Moracci. Glycosynthases in biocatalysis. *Adv. Synth. Catal.*, 353(13):2284–2300, 2011.

[64] W. Huang, J. Giddens, S.-Q. Fan, C. Toonstra, and L.-X. Wang. Chemoenzymatic glycoengineering of intact IgG antibodies for gain of functions. *J. Am. Chem. Soc.*, 134(29):12308–12318, 2012.

[65] J.-H. Shim, H.-M. Chen, J. R. Rich, E. D. Goddard-Borger, and S. G. Withers. Directed evolution of a β-glycosidase from Agrobacterium sp. to enhance its glycosynthase activity toward C3-modified donor sugars. *Protein Eng. Des. Sel.*, 25(9):465–472, 2012.

[66] 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. β-Glycosyl Azides as Substrates for α-Glycosynthases: Preparation of Efficient α-L-Fucosynthases. *Chem. Biol.*, 16(10):1097–1108, 2009.

[67] Y. Zhang, S. Yan, and L. Yao. Mechanism of the Humicola insolens Cel7B E197S mutant catalyzed flavonoid glycosides synthesis: A QM/MM metadynamics simulation study. *Theor. Chem. Accounts*, 132(6):1–10, 2013.

[68] J. Wang, S. Pengthaisong, J. R. K. Cairns, and Y. Liu. X-ray crystallography and QM/MM investigation on the oligosaccharide synthesis mechanism of rice BGlu1 glycosynthases. *Biochimica et Biophys. Acta - Proteins Proteomics*, 1834(2):536–545, 2013.

[69] P. T. Vanhooren and E. J. Vandamme. L-Fucose: occurrence, physiological role, chemical, enzymatic and microbial synthesis. *J. Chem. Technol. & Biotechnol.*, 74(6):479–497, 1999.

[70] P. A. Prieto. Profiles of Human Milk Oligosaccharides and Production of Some Human Milk Oligosaccharides in Transgenic Animals. *Adv. Nutr.*, 3(3):456S–464S, 2012.

[71] J. T. Smilowitz, C. B. Lebrilla, D. A. Mills, J. B. German, and S. L. Freeman. Breast milk oligosaccharides: structure-function relationships in the neonate. *Annu. Rev. Nutr.*, 34:143–169, 2014.

[72] D. L. Ackerman, R. S. Doster, J.-H. Weitkamp, D. M. Arono, J. A. Gaddy, and S. D. Townsend. Human Milk Oligosaccharides Exhibit Antimicrobial and Antibiofilm Properties against Group B Streptococcus. *ACS Infect. Dis.*, 3(8):595–605, 2017.

[73] H.-J. Wu, C.-W. Ho, T.-P. Ko, S. D. Popat, C.-H. Lin, and A. H.-J. Wang. Structural basis of alpha-fucosidase inhibition by iminocyclitols with K(i) values in the micro- to picomolar range. *Angewandte Chemie - Int. Ed.*, 49(2):337–340, 2010.

[74] G. Sulzenbacher, C. Bignon, T. Nishimura, C. A. Tarling, S. G. Withers, B. Henrissat, and Y. Bourne. Crystal structure of Thermotoga maritima α-L-fucosidase: Insights into the catalytic mechanism and the molecular basis for fucosidosis. *J. Biol. Chem.*, 279(13):13119–13128, 2004.

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

[76] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, and R. J. Woods. GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates. *J. Comput. Chem.*, 29(4):622–655, 2008.

[77] R. Rüger, A. Yakovlev, P. Philipsen, S. Borini, P. Melix, A. F. Oliveira, M. Franchini, T. Soini, M. de Reus, M. G. Asl, D. McCormack, S. Patchkovskii, and T. Heine. *ADF DFTB 2017*. SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, http://www.scm.com.

[78] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B*, 58(11):7260–7268, 1998.

[79] A. T. P. Carvalho, P. A. Fernandes, and M. J. Ramos. Parameterization of AZT—A Widely Used Nucleoside Inhibitor of HIV-1 Reverse Transcriptase. *Int. J. Quantum Chem.*, 107(2):292–298, 2007.

[80] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 7(2):230–252, 1986.

[81] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, , and D. J. Fox. *Gaussian 16, Revision A.03*. Gaussian, Inc., Wallingford CT, 2016.

[82] F. J. Devlin, J. W. Finley, P. J. Stephens, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields: A comparison of local, nonlocal, and hybrid density functionals. *J. Phys. Chem.*, 99(46):16883–16902, 1995.

[83] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652, 1993.

[84] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

[85] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58(8):1200–1211, 1980.

[86] A. D. McLean and G. S. Chandler. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11-18. *J. Chem. Phys.*, 72(10):5639–5648, 1980.

[87] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.*, 72(1):650–654, 1980.

[88] M. P. Andersson and P. Uvdal. New scale factors for harmonic vibrational frequencies using the B3LYP density functional method with the triple-ζ basis Set 6-311+G(d,p). *J. Phys. Chem. A*, 109(12):2937–2941, 2005.

[89] G. I. Csonka. Proper basis set for quantum mechanical studies of potential energy surfaces of carbohydrates. *J. Mol. Struct. THEOCHEM*, 584(1):1–4, 2002.

[90] 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.

[91] R. G. Mullen, J. E. Shea, and B. Peters. Easy Transition Path Sampling Methods: Flexible-Length Aimless Shooting and Permutation Shooting. *J. Chem. Theory Comput.*, 11(6):2421–2428, 2015.

[92] B. Peters. Inertial likelihood maximization for reaction coordinates with high transmission coefficients. *Chem. Phys. Lett.*, 554:248–253, 2012.

[93] G. T. Beckham and B. Peters. New Methods To Find Accurate Reaction Coordinates by Path Sampling. In M. R. Nimlos and M. F. Crowley, editors, *Computational Modeling in Lignocellulosic Biofuel Production*, chapter 13, pages 299–332. American Chemical Society, Washington, D.C., 2010.

[94] G. M. Torrie and J. P. Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.*, 23:187–199, 1977.

[95] T. H. Choi, R. Liang, C. M. Maupin, and G. A. Voth. Application of the SCC-DFTB method to hydroxide water clusters and aqueous hydroxide solutions. *J. Phys. Chem. B*, 117(17):5165–5179, 2013.

[96] M. Elstner. The SCC-DFTB method and its application to biological systems. *Theor. Chem. Accounts*, 116(1-3):316–325, 2006.

[97] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. *Gaussian 16, Revision B.01*. Gaussian, Inc., Wallingford CT, 2016.

[98] D. A. McQuarrie and J. D. Simon. *Molecular Thermodynamics*, volume 63. Sausalito, CA: University Science Books, 1999.

[99] S. Miertuš, E. Scrocco, and J. Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys.*, 55(1):117–129, 1981.

[100] S. Miertuš and J. Tomasi. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chem. Phys.*, 65(2):239–245, 1982.

[101] J. L. Pascual-ahuir, E. Silla, and I. Tuñon. GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solventâĂŘexcluding surface. *J. Comput. Chem.*, 15(10):1127–1138, 1994.

[102] A. Agrawal, C. K. Bandi, T. Burgin, Y. Woo, H. Mayes, and S. Chundawat. Click-chemistry enabled directed evolution of glycosynthases for bespoke glycans synthesis. *BioRXiv*, pages 1–32, 2020.

[103] T. M. Gloster and G. J. Davies. Glycosidase inhibition: Assessing mimicry of the transition state. *Org. Biomol. Chem.*, 8(2):305–320, 2010.

[104] C. Hartmann, R. Banisch, M. Sarich, T. Badowski, and C. Schütte. Characterization of Rare Events in Molecular Dynamics. *Entropy*, 16(1):350–376, 2014.

[105] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao. Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.*, 151:070902, 2019.

[106] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.*, 180(10):1961–1972, 2009.

[107] D. W. H. Swenson, J.-H. Prinz, F. Noe, J. D. Chodera, and P. G. Bolhuis. OpenPathSampling: A Python Framework for Path Sampling Simulations. 1. Basics. *J. Chem. Theory Comput.*, 15:813–836, 2019.

[108] A. Lervik, E. Riccardi, and T. S. van Erp. PyRETIS: A Well-Done, Medium-Sized Python Library for Rare Events. *J. Comput. Chem.*, 38(28):2439–2451, 2017.

[109] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman, and L. T. Chong. WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis. *J. Chem. Theory Comput.*, 11(2):800–809, 2015.

[110] H. Sidky, Y. J. Colón, J. Helfferich, B. J. Sikora, C. Bezik, W. Chu, F. Giberti, A. Z. Guo, X. Jiang, J. Lequieu, J. Li, J. Moller, M. J. Quevillon, M. Rahimi, H. Ramezani-Dakhel, V. S. Rathee, D. R. Reid, E. Sevgen, V. Thapar, M. A. Webb, J. K. Whitmer, and J. J. de Pablo. SSAGES: Software Suite for Advanced General Ensemble Simulations. *J. Chem. Phys.*, 148(4), 2018.

[111] G. Fiorin, M. L. Klein, and J. Hénin. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.*, 111(22-23):3345–3362, 2013.

[112] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York, and P. A. Kollman. Amber 18. University of California, San Francisco, 2018.

[113] W. E and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.*, 61:391–420, 2010.

[114] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.*, 53(1):291–318, 2002.

[115] F. Pietrucci. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Rev. Phys.*, 2:32–45, 2017.

[116] M. N. Joswiak, M. F. Doherty, and B. Peters. Ion dissolution mechanism and kinetics at kink sites on NaCl surfaces. *Proc. Natl. Acad. Sci.*, 115(4):656–661, 2018.

[117] C. Dellago and P. G. Bolhuis. *Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events*, pages 167–233. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[118] B. Peters. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. Phys. Chem.*, 67:669–690, 2016.

[119] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, 13(7):1–17, 2017.

[120] G. Groenhof. *Introduction to QM/MM Simulations. In: Monticelli L., Salonen E. (eds) Biomolecular Simulations. Methods in Molecular Biology (Methods and Protocols)*, volume 924. Humana Press, Totowa, NJ, 2013.

[121] P. R. Schreiner, P. v. R. Schleyer, and R. K. Hill. Mechanisms of Front-Side Substitutions. The Transition States for the SNi Decomposition of Methyl and Ethyl Chlorosulfite in the Gas Phase and in Solution. *J. Org. Chem.*, 59:1849–1854, 1994.

[122] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An Open chemical toolbox. *J. Cheminformatics*, 3(10):1–14, 2011.

[123] B. Peters and B. L. Trout. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.*, 125(5), 2006.

[124] B. Peters, G. T. Beckham, and B. L. Trout. Extensions to the likelihood maximization approach for finding reaction coordinates. *J. Chem. Phys.*, 127(3), 2007.

[125] G. Schwarz. Estimating the Dimension of a Model. *The Annals Stat.*, 6(2):461–464, 1978.

[126] H. Akaike. Fitting autoregressive models for prediction. *Annals Inst. Stat. Math.*, 21:243–247, 1969.

[127] J. Kästner. Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(6):932–942, 2011.

[128] V. S. Bharadwaj, B. C. Knott, J. Ståhlberg, G. T. Beckham, M. F. Crowley, and G. W. Hart. The hydrolysis mechanism of a GH45 cellulase and its potential relation to lytic transglycosylase and expansin function. *J. Biol. Chem.*, 295(14):4477–4487, 2020.

[129] M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, 2008.

[130] J. D. Chodera. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.*, 12(4):1799–1805, 2016.

[131] V. P. Godambe. An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals Math. Stat.*, 31(4):1208–1211, 1960.

[132] T. Burgin and H. B. Mayes. Mechanism of oligosaccharide synthesis via a mutant GH29 fucosidase. *React. Chem. & Eng.*, 4:402–409, 2019.

[133] T. Burgin, J. Ståhlberg, and H. B. Mayes. Advantages of a Distant Cellulase Catalytic Base. *J. Biol. Chem.*, In Press, 2018.

[134] J. Yang, G. Bak, T. Burgin, W. J. Barnes, H. B. Mayes, M. J. Peña, B. R. Urbanowicz, and E. Nielsen. Biochemical and genetic analysis identify CSLD3 as a beta-1,4-glucan synthase that functions during plant cell wall synthesis. *Plant Cell*, 32(5):1749–1767, 2020.

[135] B. R. I. Miller, T. D. Mcgee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.*, 8:3314–3321, 2012.

[136] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.*, 246(1-2):122–129, 1995.

[137] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *The J. Phys. Chem.*, 100(51):19824–19839, 1996.

[138] D. R. Roe and T. E. I. Cheatham. PTRAJ and CPPTRAJ: software for processing and analysis of molecular synamics trajectory data. *J. Chem. Theory Comput.*, 9:3084–3095, 2013.

[139] J. Srinivasan, T. E. I. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *J. Am. Chem. Soc.*, 120(37):9401–9409, 1998.

[140] B. Peters. Using the histogram test to quantify reaction coordinate error. *J. Chem. Phys.*, 125(24):241101, 2006.

[141] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2364–2393, 1980.

[142] J. J. P. Stewart. Optimization of Parameters for Semiempirical Methods II. Applications. *J. Comput. Chem.*, 10(2), 1989.

[143] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. P. Stewart. RM1: a Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.*, 27(10):1101–1111, 2006.

[144] K. Miura. An Introduction to Maximum Likelihood Estimation and Information Geometry. *Interdiscip. Inf. Sci.*, 17(3):155–174, 2011.

[145] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Stat. Sinica*, 21(1):5–42, 2011.