This electronic thesis or dissertation has been downloaded from Explore Bristol Research, http://research-information.bristol.ac.uk

Author:
Johns, Sam T

Title:
Towards simulation-led engineering of a natural Diels-Alderase

General rights
Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy
Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

Towards simulation-led engineering of a natural Diels-Alderase

Samuel Tredget Johns

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

School of Biochemistry

April 2021

Word count: Sixty thousand

Abstract

The Diels-Alder (DA) reaction is a fundamental reaction in organic synthesis, and the discovery of enzymes that can catalyse this reaction (or DAases) has thus long been a desirable goal. AbyU is a DAase which performs the cyclisation step in the formation of abyssomicin C, a spirotetronate antibiotic. There is therefore interest in engineering this enzyme to use it as a platform for more general spirotetronate synthesis. In this thesis, the origin of catalysis in AbyU is firstly studied via multiscale computational simulation. It is determined that the role of the active site is chiefly to provide complementary binding for the reactive conformation of the substrate. The potential for expanding the substrate scope of AbyU is then explored by predicting its activity (and that of several point mutants) with a panel of alternative substrates. A workflow of efficient computational protocols including docking, molecular dynamics (MD) and combined quantum mechanics / molecular mechanics (QM/MM) reaction simulation is therefore developed for screening the different combinations (utilising the same protocols which were used to study the native reaction). Due to the simplicity of the enzyme, it is found to accept most alternative substrates, which are all predicted to be fairly active with the enzyme. In addition, the enzyme was rationally redesigned for specific substrates, and redesigned variants were evaluated using the same *in silico* screening protocols. This demonstrates a promising approach to simulation-led engineering of this enzyme, where good scope for optimisation was suggested for one particular substrate. Overall, the *in silico* screening protocols developed in this thesis provide a valuable example of computationally efficient simulations that can be used to generate activity predictions for β-barrel spirotetronate forming enzymes such as AbyU, in a reasonably high throughput manor. The remaining aspects of enzymatic catalysis are then lastly studied. The origin of stereoselectivity is probed, where it is found that short MD simulations of the Michaelis complex alone are sufficient to indicate this. Finally, the behaviour of the catalytically important capping loop is studied and its potential for optimisation explored. A particular enhanced sampling technique shows utility for this, enabling the potentially rate limiting loop opening (for product release) step to be captured on practical simulation timescales.

Acknowledgements

I would firstly like to thank my supervisor, Dr. Marc van der Kamp, who has been of tremendous support throughout the entirety of my PhD. He has always been incredibly patient and helpful, and made it a lot easier to get started in this field. I also appreciate the considerable feedback and guidance he has provided during the writing of this thesis, his input has been invaluable at each step along the way. I would also like to thank the various members of my group, past and present, who helped make our office a vibrant, fun and inspiring place to work towards the goal of a PhD. Next, I would like to thank the friends I met on my degree programme, you have been a constant source of moral support and commiseration, always having been up for a trip to the pub to find some consolation for our various research struggles. A special mention also goes to all my other friends, who have been understanding of my complete radio silence while writing up. Lastly, I would like to thank my parents and my brother, who have provided continuous moral support and encouragement throughout the years. In particular during this last phase, they have helped to keep me going in the face of many months of prolonged writing.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: DATE: 04/04/2021

Table of Contents

List of Figures

Chapter 1. Introduction

1.1 Diels-Alder importance and the existence of Diels-Alderases

The Diels-Alder (DA) reaction is a [4+2] cycloaddition which takes place between an alkene (termed the dienophile) and a conjugated 1,3 diene resulting in a substituted cyclohexene ring (Figure 1.1). This is known as a pericyclic reaction meaning that, rather than by a stepwise process, bond breaking/formation happens in a single step via the concerted transfer of 6 π electrons through a single cyclic transition state (TS) (see section 3.1.1 for detailed background).

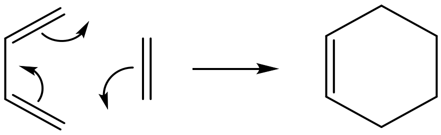


Figure 1.1: The Diels-Alder Reaction between butadiene and ethylene to form cyclohexene.

Due to its ability to yield complex molecules with remarkable regio- and stereo-selectivity this reaction has long been a powerful tool in the repertoire of synthetic chemists, both in the creation of designer molecules as well as for retrosynthesis of natural products with important biological activities (1). In terms of natural products, cyclohexene formation is used by nature to generate the molecular scaffold for a diverse array of compounds. Additionally, as surveys have identified >400 compounds (2) which show the characteristics of being synthesised via a biosynthetic DA reaction (defined stereochemistry about the cyclohexene moiety (3)) this had led to speculation of the existence of natural DAases which either enhance or control cyclohexene formation. Subsequent searching of the biosynthetic clusters for these has now turned up many putative examples of such enzymes. However, it is not immediately clear whether these are true DAases i.e. whether they utilise a pericyclic reaction mechanism, since cyclohexene formation can occur through multiple routes e.g. stepwise via diradical intermediates. In addition, as initially discovered DAases were multifunctional this complicated studies of their mechanism. However, evidence has since mounted to support the existence of several bonafide DAases and patterns have emerged which give insight into how they operate. Understanding how these enzymes work is the key to harnessing their power and designing novel biocatalysts for DA reactions. This is of great interest as, while the DA reaction has been a powerful tool in synthetic organic chemistry, biocatalysis offers great potential as a greener alternative for industrial synthesis and in cases when chemicals are unable to be synthesised conventionally, while providing exquisite control over product formation (4). In addition, multiple enzyme can be combined into single cascade reactions which mimic entire biosynthetic pathways thereby providing a highly simplified and more efficient way of synthesising complex molecules from simple precursors (4).

1.2 History of Diels-Alderases

1.2.1 Bifunctional decalin forming

The decalin scaffold represents one of the most frequently occurring potential [4+2] adducts found in nature. Polyketide Synthases (PKSs), and in particular bacterial multimodular/fungal iterative PKSs, generate various skipped polyene configurations (2), including a diene and alkene separated by 4 sp^3 carbons which forms a decalin ring upon intramolecular cycloaddition of the diene to the alkene. It is thus unsurprising that two of the first naturally occurring putative DAases discovered were Solanapyrone (**1**) synthase (Sol5) (5) and Lovastatin (**2**) nonaketide synthase (LovB) (6) that are responsible for decalin formation in the fungal polyketides **1** and **2** respectively (Figure 1.2) (their activity being detected by stereocontrol rather than catalysis (3)).

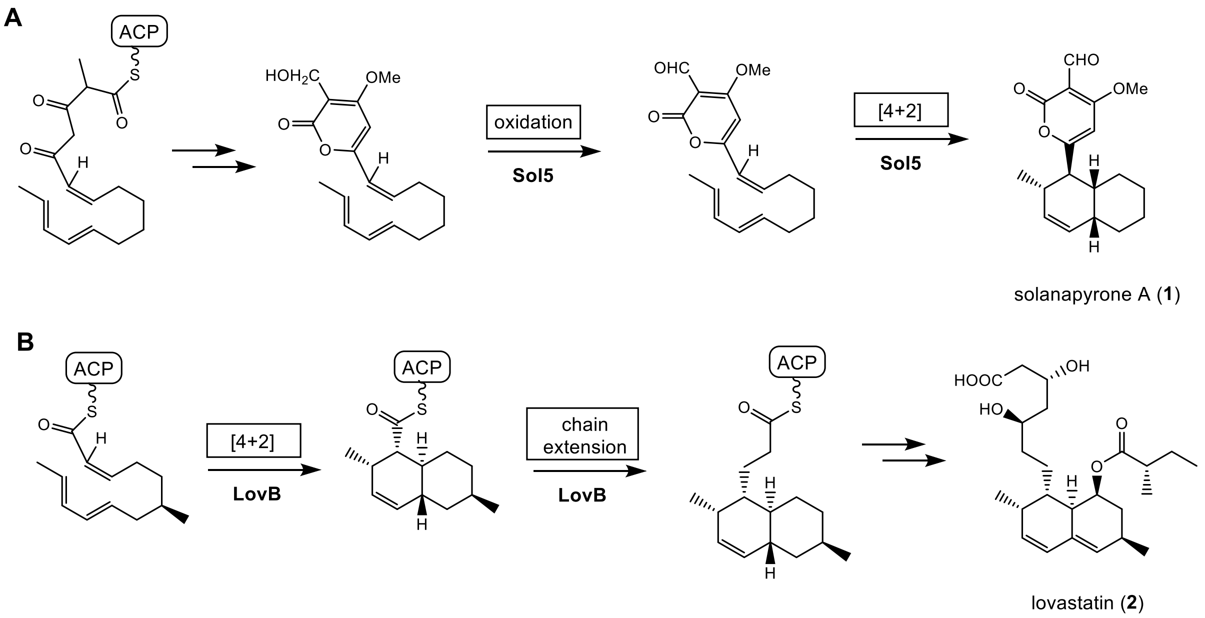


Figure 1.2 (Adapted from (2)): Decalin formation performed by multifunctional DAases Sol5 (A) and LovB (B).

However, in addition to decalin formation these enzymes also catalyse one or more of the previous biosynthetic steps in their respective pathways. Sol5 also catalyses the oxidation step which occurs prior to decalin formation, after release of the substrate from the highly reducing iterative PKS (HR-IPKS). Meanwhile LovB is itself a HR-IPKS, and is thus involved in all biosynthetic steps in the pathway including the DA step (recruiting the separate protein LovC which provides the ER domain required for certain reaction steps). These two enzymes have little in common, and since DA activity is only one part of their overall function (particularly in the case of LovB), this greatly complicates any investigation into the biocatalysis of DA reactions which is based on them. In addition, as they will have added complexity and likely sacrifice some DA activity to enable their other activities, they are inappropriate models for the design of enzymes whose sole function is to catalyse a DA reaction.

1.2.2 Monofunctional

1.2.2.1 Decalin forming

A type of monofunctional decalin forming enzyme has been discovered in the pathways for fungal polyketides fusarisetin A (**3**) and Sch210972 (**4**), which contain tetramic acid moieties (Figure 1.3). Fsa2 was found to found to be solely responsible for decalin formation in the production of **3** due to the *endo*-selectivity found in the presence of the enzyme (7). Sol5, Fsa2 and the other monofunctional putative DAases described here work after release from the PKS machinery. Fsa2 was a previously unrecognised protein, showing no homology with existing proteins and thus having no previously known function. However, the functional homologue CghA was also found to perform the analogous role, catalysing the *endo*-selective cycloaddition in the biosynthetic pathway for **4** (8). The similarity in product and homology of these two enzymes suggest they have a common ancestor and evolved alongside the substrate producer enzymes to accommodate divergence in their substrate. Further homologues whose cycloaddition activity is not yet confirmed have also been found in the biosynthetic gene clusters for other fungal polyketides which are potential [4+2] adducts containing tetramic/tetronic acid moieties, e.g. those for pyrrolocin (9) and chytochalasin (10). (In the case of chytochalasin, a decalin ring is not formed but instead a larger carbocycle since the reacting alkene is provided by the tetronic acid moiety itself.) This suggests further examples of specialisation of this emergent enzyme type.

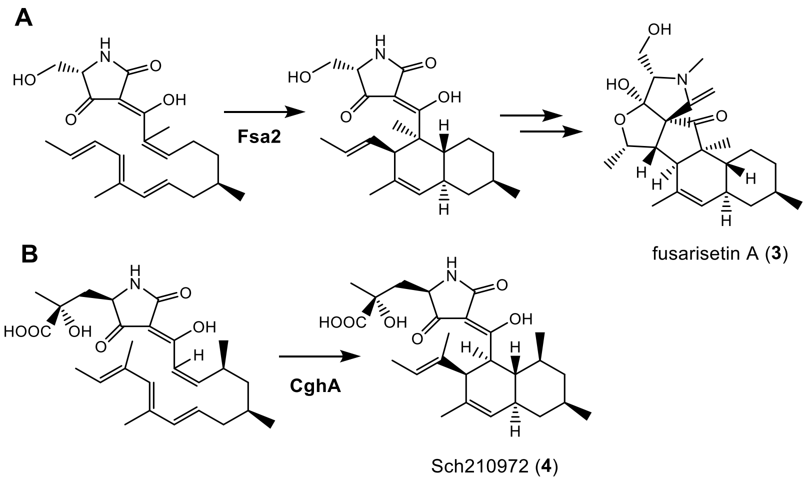
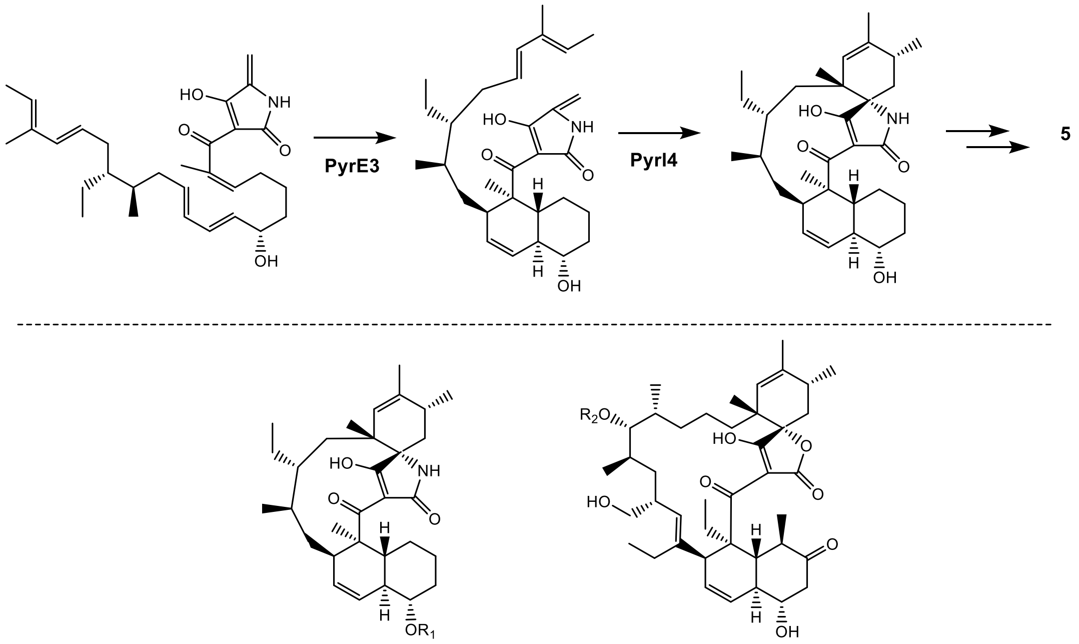


Figure 1.3: Decalin formation in the pathways for fungal polyketides fusarisetin A (A) and Sch210972 (B).

Another related group of monofunctional decalin forming enzymes has been discovered in bacterial PKS pathways which produce spirotetramates/spirotetronates. PyrE3 is a representative of this group which was found to exclusively catalyse formation of the decalin ring in the production of pyrroindomycin A (**5**) (Figure 1.4), following cleavage from the PKS (11). Its ancestral function was that of a FAD dependent oxidase, showing that this enzyme has thus been repurposed and given a new activity. Interestingly, as FAD is still essential for decalin formation (11, 12), this cofactor has simply been incorporated into its new function. However, rather than its redox potential being important for catalysis, it now instead serves an essential structural role within the active site (12). Homologues of PyrE3 were then also found in the pathways for other bacterial decalin containing spirotetramates/spirotetronates (e.g. VstK in the versipelostatin (**6**) pathway (13) and ChlE3 in the chlorothricin pathway (14)), which are presumed to perform the same function in those pathways. While both groups share the decalin ring and tetronate/tetramate moiety, unlike the compounds produced by FSa2/CghA, the family of compounds represented in Figure 1.4 are also spirolinked into larger macrocycles via a substituted cyclohexene, and for pyrroindomycin A and versispelostatin this is known to be performed by yet another type of monofunctional DAase (discussed in the next section). This spirolinking occurs after decalin formation, where another 1,3-diene provided by the linear precursor reacts through a further [4+2] cycloaddition with the exocyclic methylene of the tetramate/tetronate ring (which, unlike for the representatives of the decalin containing family shown in Figure 1.3, is present as they contain an enolate form of the tetramate/tetronate moiety).



pyrroindomycin A (**5**)

versipelostatin (**6**)

Figure 1.4: The biosynthetic pathway for decalin containing spirotetramate pyrroindomycin A (**5**) contains two enzymatic DA steps performed by PyrE3/Pyrl4. Homologues of PyrE3/Pyrl4 are also found in the pathway for decalin containing spirotetronate versipelostatin (**6**).

For the monofunctional enzymes described thus far, while structural information has in some cases been obtained, the mechanism of enzymatic cycloaddition has not yet been studied in detail. It is thus not known for sure that decalin formation proceeds via a true DA reaction in enzyme. Nevertheless, the phylogenetic origins of the decalin forming enzymes provide useful context for the later discussion of how DAase activity may arise and what the requirements are for it. For the non-decalin forming enzymes discussed in the next section, structural information has been obtained and more detailed mechanistic insights have been gained which provide evidence for true DA catalysis.

1.2.2.2 Non decalin forming

1.2.2.2.1 SpnF

In 2011 SpnF, from bacteria *S. spinosa*, was determined to be solely responsible for formation of the cyclohexene in the as-indocene core of the polyketide insecticide spinosyn A (**7**) (Figure 1.5) making it technically the first discovered monofunctional putative DAase (15). As well as selectively yielding the *endo*-product the enzyme was found to provide a 500-fold rate enhancement for the cycloaddition, making this an intriguing subject due to its potential mechanism for DA catalysis as well as just providing stereocontrol.

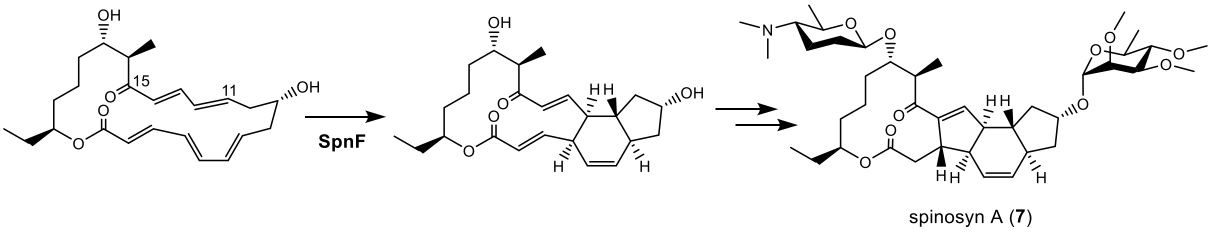


Figure 1.5: Role of SpnF in the biosynthetic pathway for spinosyn A (**7**).

SpnF is a homologue of a SAM-dependent methoxytransferase, showing an example of another co-factor associated enzyme being repurposed for a DA reaction. Again, the co-factor most likely now just acts as part of the overall enzyme structure since its original role is no longer relevant. In the crystal structure of SpnF (16), SAH was the cofactor bound and extensive interactions with the enzyme further suggest its likely role as a primarily structural element (although in vivo this would most likely be provided by SAM co-factor). Consensus docking of the substrate/product also gave some initial insight into the mechanism of catalysis. In the most probable docking mode, it was found that residue T196 was in a position to withdraw electron density from the dienophile via the C15 carbonyl and C11-C15 π system (Figure 1.5). As this is a way to lower the energy barrier for a normal demand DA (see section 3.1.1) it thus suggested this may be a true DAase. It was also thought that the enzyme may stabilise the reactive conformation of the substrate to enhance the reaction rate. As the rest of the substrate is fairly rigid due to its cyclical structure and the extended conjugated π systems, this mostly just involves stabilising the *s-cis* conformation of the reacting diene which is required for the DA reaction to occur (and is typically the less stable conformation – see section 3.1.1). However, some uncertainty as to whether this reaction proceeded via a DA route came from a QM study of the reaction in gas phase (17). An ambimodal TS was found for the cycloaddition which could either lead directly to the [4+2] adduct, or to a [6+4] intermediate which would then rapidly be converted into the [4+2] adduct via a Cope rearrangement (dynamics simulations showed that the [6+4] route was slightly preferred by a 2.5:1 ratio in gas phase). In the enzyme, an ambimodal TS was then also found (18), but reactive trajectories through this TS more often proceeded directly to the [4+2] adduct (in a ratio of 11:1). Although the direct DA route was therefore found to be favoured by modification of the post TS energy surface, the fact that this reaction exhibits such a complex ambimodal pathway may make it seem to be a less relevant case study for general DA catalysis. Nevertheless, the way in which catalysis and lowering of the barrier for the ambimodal TS may be achieved - namely electron-withdrawing interactions through H-bonding residues (as well as stabilising of the reactive conformation, which is relevant regardless of the electronic details of the cycloaddition mechanism) – should also be applicable to a simple un-bifurcated DA reaction pathway. This enzyme therefore provides useful context for the study of DA catalysis.

1.2.2.2.2 Spirotetronate/tetramate forming

As discussed in the previous section, the biosynthetic pathways for decalin containing spirotetronates/tetramates pyrroindomicin A (**5**) and versipelostatin (**6**) were found to contain a second type of DAase (Pyrl4 and VstJ respectively), which carries out the spiro-linking of the molecule post decalin formation (Figure 1.4) (homologues of this also exist for other members of this family of products e.g. chlorothricin, which presumably carry out the same function but have yet to be functionally characterised). The structure of Pyrl4 has been obtained both in the apo state and for the product complex (19) which has provided considerable mechanistic insights. Structurally, the enzyme is made up of a β-barrel scaffold, where the core of the barrel forms the active site cavity (Figure 1.6A). An N-terminal region then forms a capping motif which acts to trap the substrate (see below). Looking at the product complex, hydrophobic regions of the active site can be observed to interact with the hydrophobic spirotetronate scaffold, while polar groups interact and hydrogen bond with the polar groups of the product (Figure 1.6B).

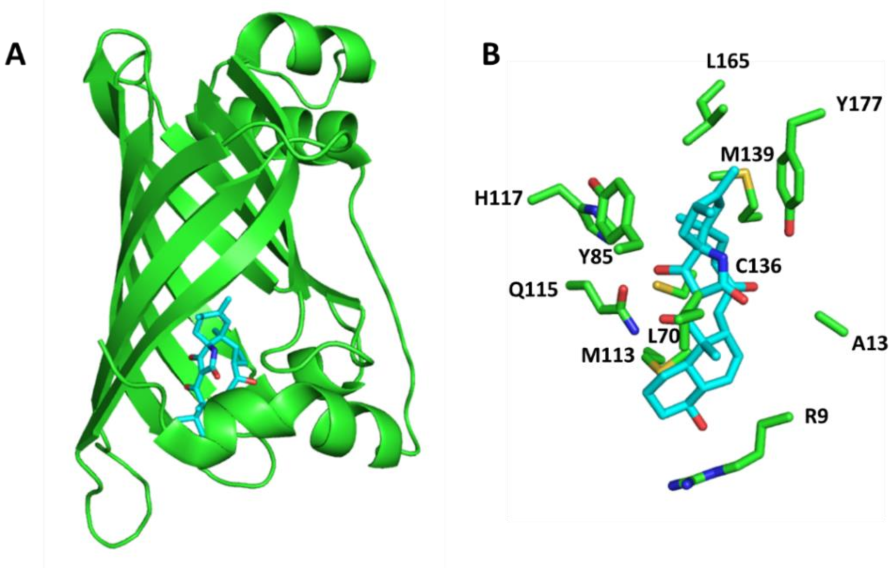
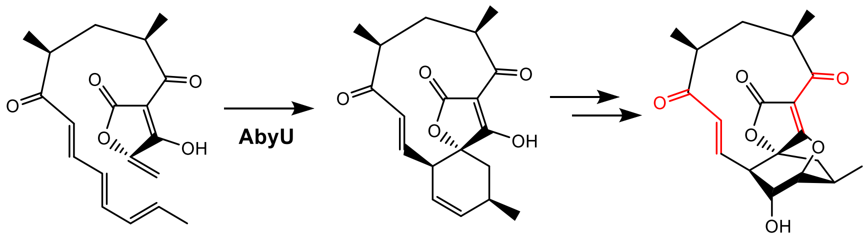


Figure 1.6: A) Crystal structure of Pyrl4 with product bound (from (19) - PDB ID: 5BU3). B) Close up view of active site of product complex crystal structure (key residues highlighted).

Mutagenesis studies then gave insight into the contribution of active site residues towards activity (19). Q115 was found to be a key residue, likely important in activating the dienophile (via the electron-withdrawing interaction with the substrate that is indicated in the crystal structure), as mutating it reduced activity by around 60%. A recent computational study (20) of Pyrl4 further investigated the catalytic contribution of this residue, given that the electron-withdrawing interaction was also observed during MD simulations of the Transition state complex. In addition to this, the existence of catalytic triad, in which H117 and Y85 serve to increase the activating effect of Q115 by making it more acidic, was proposed and tested. This was because a Hydrogen bonding network that would appear to achieve this was found to form between these three residues during the aforementioned MD. To assess the catalytic influence of Q115 alone, as well as when part of the hypothesised triad, theozyme models were constructed containing the ligand and sidechains of the respective residues. Co-ordinates for these models were extracted from the cluster representing the most stable state found during the MD of the transition state complex (in which the aforementioned network is formed), replacing the alpha carbons for the amino acid sidechains with methyl groups. Activation free energies were then determined for both model systems via QM optimisation and frequency calculations for the reactant and transition states. On its own, Q115 was found to lower the activation energy by 0.9 kcal/mole compared to that for the uncatalyzed reaction, thus indicating that it can indeed act to catalyse the reaction by activating the dienophile. When the nearby residues H117 and Y85 were also included in the modelling, the activation energy was then lowered by a further 2.4 kcal/mole. This suggests that these residues do indeed serve to further enhance the catalytic effect of Q115, and together form an effective catalytic triad. For the remaining active site residues, much smaller effects were found in the mutagenesis study showing that none of these are, on their own, essentially catalytic. However, making multiple mutations of active site residues led to a cumulative effect on activity indicating that other residues act together synergistically to constrain the substrate in the reactive conformation for the observed *exo* product. Unlike for SpnF, where the substrate begins in a cyclic state, the spirolinking reaction performed by Pyrl4 (and homologues) involves cyclising a linear precursor and so requires first bringing the reacting diene and alkene together. Therefore, as well as the reactive conformation of the diene likely being unfavourable, there is likely an entropic cost to the substrate adopting a reactive conformation due to the constraining of bonds (which would otherwise be free to rotate) required when bringing the diene and dienophile together (see section 3.3.3). Therefore, since this is also indicated by the mutagenesis results, some catalysis likely comes from the enzyme providing a complementary binding site where residues synergistically constrain the substrate in the entropically unfavourable reactive conformation (thereby functioning as a so-called 'entropy trap' (21)). The follow-up computational study of Pyrl4 supports this hypothesis since the reactive conformation of the substrate was found to be 6.8 kcal/mole less stable compared to a relaxed conformation. This factor also has interesting implications for the homologous reaction catalysed by VstJ (as well as the reaction catalysed by AbyU – see later). As, in the case of VstJ, the linear substrate is very long (and thus flexible – see Figure 1.4) it has been suggested that the likelihood of the substrate adopting a reactive conformation may be the main reason that this substrate is intrinsically unreactive (13). Therefore, the entropy trap effect likely contributes a more significant rate enhancement for VstJ than for Pyrl4. The final key aspect of the Pyrl4 catalytic mechanism is the capping motif formed by the N-terminal region. This capping motif was found to be essential for catalysis as mutating the 10-aa α-helical region completely abolished activity (19). The crystal structures indicated that this region acts in a lid-like manner to trap the substrate within the active site in the high energy reactive conformation as, while it is found in a disordered state in the apo structure, with the product bound it adopts a structured α-helical form and closes around the active site (Figure 1.6A). It was also speculated, based on different NMR signals of the loop region in presence of substrate or product, that the increased rigidity of the product then induces the subsequent weakening of the lid-barrel interactions in order to enable product release (19). Interestingly, while the β-barrel is found to be quite conserved, for homologs of Pyrl4 (e.g. ChlL) the N-terminal region varies considerably (19), indicating that the ideal design is highly dependent on the substrate in question. Indeed, in the case of VstJ (13), no equivalent N-terminal region exists indicating that very different mechanisms have emerged for trapping substrates within this class of [4+2] cyclases. Another homologue of Pyrl4, called AbyU (discussed in the next section), also lacks this N-terminal region and a capping function is instead performed by a 10-aa loop region between two neighbouring strands.

1.2.2.2.2.1 AbyU

AbyU is a homologue of Pyrl4 which performs the analogous spirolinking in the production of the bacterial spirotetronate Abyssomicin C (**8**) (see Figure 1.7. Note: unlike the pyrroindomycin A/versipelostatin family of products, Abyssomicin C does not contain a decalin ring and so no equivalent of the decalin forming enzyme PyrE3/VstK exists in its pathway). Abyssomicin C was found to be a potent inhibitor of bacterial pABA biosynthesis (22) (part of the folate metabolic pathway), and is effective against Mycobacterium tuberculosis (23) and strains of multi-drug resistant Staphylococcus aureus (22). This inhibition is achieved by irreversible binding of Abyssomicin C to a cysteine residue of the PabB subunit of amino-deoxychorismate synthase (ADCS), thus covalently inhibiting the enzyme and preventing pABA formation (24). Binding of Abyssomicin C to ADCS occurs via a double Michael addition, where the two Michael acceptors involved are the two α, β unsaturated ketones of Abyssomicin C highlighted in red in Figure 1.7. Due to the antibiotic activity of Abyssomicin C, there is interest in engineering AbyU and using it as a platform for the synthesis of related spirotetronates which may be useful as novel antibiotics. As it is also a co-factor free enzyme that works independently from the PKS machinery (like its homologues), it makes an ideal candidate for engineering.



Abyssomicin C (**8**)

Figure 1.7: AbyU catalysed DA reaction in biosynthetic pathway for polyketide antibiotic Abyssomicin C (**8**). Michael acceptors which are involved in the mechanism of action of Abyssomicin C are highlighted in red.

The structure of AbyU was recently solved (with buffer molecule HEPES bound in the active site - PDB ID: 5DYV (25)) which has enabled subsequent mechanistic insight to be gained. Like Pyrl4, the core of the β-barrel forms the active site cavity, where entry to the active site is at the 'bottom' of the barrel i.e. opposite the N-terminal of the leading strand (Figure 1.8). However, rather than the N-terminal region found in Pyrl4, active site access is instead gated by a capping loop region located between the first two strands. Having obtained the structure, this prompted a mechanistic study wherein the uncatalyzed/catalysed reactions were investigated via docking, QM/MM reaction simulations and QM (gas phase) calculations (25). A concerted [4+2] mechanism was found both in the enzyme and via gas phase QM calculation, indicating that AbyU is indeed a true DAase. Different binding poses were tested, and the likely reactive pose was indicated by the obtained activation free energies (Figure 1.8). Other than a key H-bond between Tyr76 and the tetronate carbonyl of the substrate, enzyme-substrate interactions are largely hydrophobic, and binding is thus primarily driven by steric complementarity. It was also suggested that the main catalytic role of the enzyme may be to provide a complementary binding site to stabilise the reactive conformation of the substrate (as is also indicated to play a role in Pyrl4). However, further investigation is warranted in order to determine the involvement of the active site in the reaction. The capping loop was indicated to behave as a substrate trapping mechanism, since MD simulations showed it to be more flexible - opening easily - for the apo form, and locked tightly shut with product bound. Further investigation is also required for this aspect of the mechanism, in particular to determine whether this may be leading to product inhibition in this enzyme.

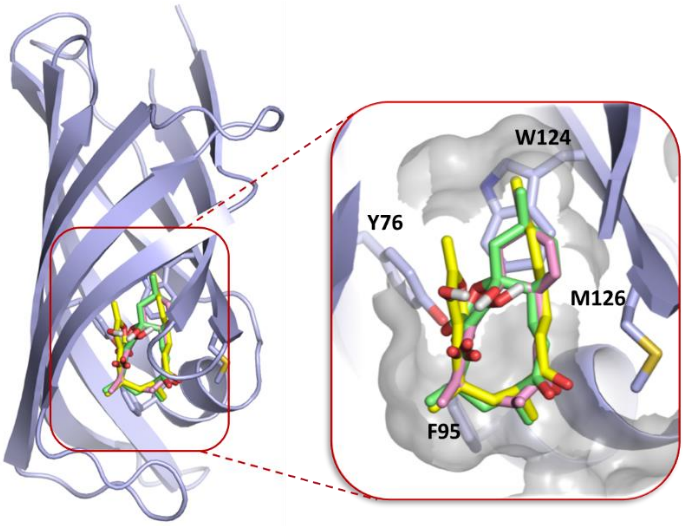


Figure 1.8: Crystal structure of AbyU with docked substrate (yellow) and the two possible atropisomers of the product (see section 3.1.2) in the reactive binding mode suggested by previous work (25). Key residues which line the active site and were treated flexibly in the docking are highlighted.

1.3 Insights from Nature and strategies for DAase design

Reviewing the various natural DAases which have so far been discovered, there appears to be no such thing as a general DAase class (or a specific architecture that is conducive to DAase activity), and it may therefore be the case that practically any enzyme can become a viable DAase. This is evident from the structural diversity found in the different types of putative/confirmed monofunctional DAases e.g. the PyrE3 type enzymes which utilise the Rossman fold for a FAD dependent oxidase, the methyltransferase fold combined with a SAM/SAH cofactor for SpnF and, finally, the β-barrel architecture of Pyrl4 and homologues (see also IdmH – another intramolecular DAase which uses a different architecture again in the α+β barrel (26)). This likely comes from the fact that catalysis can generally be achieved by simply constraining the substrate in the necessary reactive conformation – whether this involves just the enforcement of the correct reactive conformation/orientation of the diene for stereoselective cycloaddition (as in the case of SpnF), or also the trapping of a linear substrate in a cyclic conformation (as in the case of the decalin forming/spirolinking enzymes). The approach taken by Nature thus appears to be to take a somewhat arbitrary protein scaffold and then add complementary binding for the substrate and reaction in question. Reactivity may then be further enhanced by activation of the diene and/or dienophile by the appropriate electrostatic interactions (as was demonstrated in both SpnF and Pyrl4). Conceptually, this is reminiscent of *de novo* protein design, where a desired catalytic activity/active site design is incorporated into a generic starting scaffold (27). Indeed, this approach has previously been successfully applied in the computational design of an artificial DAase for a particular bimolecular DA reaction (28). To design this enzyme, the active site of a β-propellor scaffold was first modified to bind the diene and the dienophile in the correct reactive orientation, and activation of the reactants was achieved by the addition of appropriately positioned hydrogen bond donors/acceptors. While the initial design showed modest activity, further design and refinement by directed evolution then greatly improved activity, partly by further sculpting of the active site (29). Another available strategy for DA engineering, now that natural DAases have been discovered, is to simply modify the active site of an existing DAase so it may accept a wider range of DA substrates. This is particularly relevant in the case of AbyU, where modified DA adducts are of interest. Beyond the general scaffold/active site design, a final feature which is employed by Natural DAases to enhance catalysis is some means of trapping the substrate once bound, in order to tightly lock it in the high-energy reactive conformation. This is demonstrated by the different trapping motifs found for the β-barrel enzymes Pyrl4/AbyU (the recently investigated IdmH provides another example of a capping loop motif utilised by a [4+2] cyclase). The importance of this was also shown by the *de no*vo designed DAase previously mentioned, where a notable improvement in catalysis was achieved by the integration of a computationally designed lid-like element to close the active site off from bulk solvent (30). However, as subsequent opening of the active site is required to enable product release, these elements must also be finely tuned, or perhaps incorporate a way to sense the subtle differences between substrate and product (as may be the case for Pyrl4), to trigger product release. Product inhibition tends to be an issue for catalytic antibody DAases, which were technically the earliest known examples of DAases (31, 32). This is because these enzymes are typically designed to bind TS analogues, which are likely closely resembling the reaction product in DA reactions. Attempts have thus been made to retroactively incorporate product release mechanisms into catalytic antibody DAases (33). However, future attempts to improve catalytic efficiency in this way will benefit greatly from gaining more mechanistic insights into how this is accomplished in the natural enzymes.

1.4 Thesis outline

In this thesis, computational modelling is applied to provide deeper insight into the function of AbyU and explore its potential for engineering, in particular for accepting alternative substrates. A variety of modelling methods have been applied to study different aspects of the reaction, protein dynamics and protein-ligand interactions, and these methods are described in Chapter 2. In Chapter 3, the intrinsic DA reaction of the native AbyU substrate is characterised. In addition, benchmarking is performed to determine an appropriate semi-empirical method for efficient QM/MM modelling of the enzymatic reaction. The potential energy profile for the reaction is obtained using an appropriate level of density functional theory (DFT). This both demonstrates the intrinsic activation energy for the reaction and is used as a benchmark to assess the performance of various semi-empirical QM methods. Further insight into the intrinsic reactivity comes from QM calculations which provide an idea of the energetics involved in the substrate adopting the reactive conformation.

In Chapter 4, the reaction in context of the enzyme and the catalytic role of the enzyme active site is investigated with pure QM, QM/MM and MM/GBSA binding affinity calculations. In addition to being of general interest for DA biocatalysis, this informs the investigation of alternative substrates in Chapter 5 and provides a modelling strategy which can be used to predict enzyme activity when alternative substrates are bound. Firstly, the possible catalytic role of Trp124 is investigated. Then, a suitable reaction co-ordinate for QM/MM umbrella sampling of the enzymatic reaction is determined. Next, to determine whether there may be multiple productive binding modes that contribute to overall turnover, reactivity and substrate binding affinity are predicted for the different binding modes via QM/MM umbrella sampling and MM/GBSA calculations. The protocols used for this are designed to be efficient, such that they can then be used for quick screening of substrate-variant combinations to explore substrate scope in Chapter 5. Finally, to determine if the enzyme lowers the reaction free energy barrier (or simply provides stabilisation of the folded reactive conformation), QM/MM umbrella sampling is performed for the reaction in solution and the resulting free energy profile compared to those found in enzyme.

In Chapter 5, the ability of WT AbyU and several point mutants to turn over alternative substrates is explored using an efficient activity screening workflow based on the protocols used in Chapter 4. The most productive binding mode identified for the native reaction is used to narrow down the poses for testing the non-native substrates, thereby assuming that this still provides the dominant contribution towards activity. Redesign of the active site using the computational protein design tool Rosetta was then also tested, to optimise productive binding for particular substrates. The resulting redesigned complexes were tested with the same protocols as used for the combinatorial screening.

Up to Chapter 5, only the observed (or assumed, in the case of alternative substrates) stereoisomeric product(s) have been considered. In Chapter 6, the origin of stereoselectivity in AbyU is investigated, in relation to the homologous AbmU (which yields the opposite stereochemistry at a particular position in its closely related substrate). Since stereo-control can be an important design criterion, it is of interest how these opposite stereoselectivities are achieved. The Michaelis complexes for the two alternate stereoproducts in question are therefore studied in each enzyme. After generating an ensemble of structures for each complex using an efficient MD protocol, both binding affinity and estimated reactivity (by measuring the key reactive distances) are investigated to help explain the observed stereochemical outcomes. Comparing the results for the two enzymes provides some key lessons as to how stereoselective control can be achieved/how these enzymes can be manipulated to produce a different stereoisomer. Furthermore, the protocols used provide an example for stereoselectivity prediction which may be useful in the future design of spirotetronate forming enzymes with desired stereoselectivities.

In Chapter 7, the capping loop of AbyU is investigated to see if it may be contributing to product inhibition and to subsequently test its potential for optimisation. Since product release is found to be rate limiting, it is possible that this may be due (in part) to a slow rate of loop opening. Therefore, the behaviour of the loop is probed at pressure using long MD simulations to determine if the rate of loop opening may be responsible for the observed drop in activity at high pressure. The potential for tuning loop opening through mutation is then explored. Loop mutants with experimentally known activities are first tested computationally to better understand how loop behaviour contributes to activity, as well as how extreme a mutation is permissible. Based on these results, mutants which subtly alter the stability of the closed conformation are proposed and tested *in silico*. Enhanced sampling (Gaussian Accelerated MD) is used to determine how these mutations affect the ease of loop opening for the product complex and highlight potentially useful mutants. As well as being of interest for harnessing AbyU itself, the understanding gained in this Chapter will be useful for inspiring design of future DAases that incorporate capping loop type trapping mechanisms.

Chapter 2. Theoretical methods

2.1 System description

2.1.1 Molecular mechanics

An efficient (but approximate) way to describe a molecular system at the atomic level is using molecular mechanics (MM). This description is therefore necessary when dealing with large systems such as biomolecules. With this treatment electrons are ignored, and the system is represented in terms of the location of atomic centres only. The interactions between atoms resulting from the electronics are then described empirically as a function of the nuclear locations. The system is therefore treated classically, where the various interactions are considered as forces acting on the atomic masses according to Newtonian mechanics. Each interaction is modelled by an empirical potential function, which takes a different form for each type of interaction. All of the potentials defined for the system together (which include the appropriate parameters defined for each interaction) then make up what is known as the force field, which thus describes the total potential energy of the system. A simple form for this force-field which includes all the essential interactions for an MM representation shown by equation 1.

$$E_{ff} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\varphi(1 + \cos(n\varphi - \delta))$$

$$+\sum_{i}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon r_{i,j}}\right) \tag{1}$$

The first three terms describe the bonded interactions, where there are separate terms for bond stretching, bending and torsion. The bond stretching and bending terms are represented by harmonic potentials which act to restore bond lengths and angles (r and θ respectively) back to their equilibrium values (r_0 and θ_0 respectively). Bond stretch/bending forces are therefore modelled by simple linear and angular springs obeying Hooke's law (with force constants k_b and k_θ respectively). The third term represents the torsional energy which acts to restore the dihedral angle φ (Figure 2.1) between 4 sequentially bonded atoms to the ideal (e.g. staggered as shown in Figure 2.1). This is described by the periodic function shown, where *n* defines the periodicity, which is the number of minima that occur as φ rotates through 360° (3 for dihedral shown in Figure 2.1 as there are 3 staggered conformations within a 360° rotation), and δ is the phase factor which defines the values of the dihedral at which the minima occur. k_φ is then equal to the half the energy barrier for the dihedral to cross between these minima (since the height of the peaks created by the periodic function inside the brackets in the torsional term is 2).

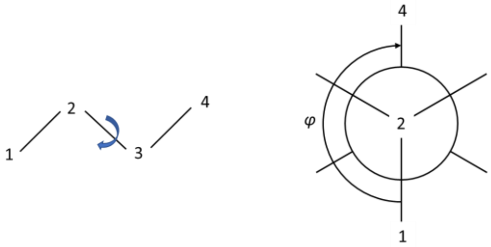


Figure 2.1: Rotation about bond between atoms 2-3 determines dihedral angle ϕ between atoms 1-2-3-4.

The final two terms in equation 1 represent the non-bonded interactions, which are those between atoms that are not directly connected. The first term is known as the 6-12 Lennard Jones potential. This describes both Van der Waals attraction between atoms (since the term is attractive when the distance between atom centres ($r_{i,j}$) is greater than the ideal) and repulsion due to overlap between the electron clouds when the atoms get too close (term becomes exponentially repulsive when $r_{i,j}$ is lower than the ideal distance). The parameters $\varepsilon_{i,j}$ and $\sigma_{i,j}$ are the depth of the attractive energy well (from an infinite distance between the atoms) and the collision diameter for a given atom pair respectively. These are derived from the respective values defined for the elements comprising the atom pair via mixing rules. The final term describes electrostatic interactions, i.e. the repulsion or attraction of atoms due to the charges which are developed on them as a result of their bonding arrangements. This is described by a simple Coulombic potential, where q_i and q_j are the partial charges for the atoms in each pair and ε is the effective dielectric constant (set to 1 when solvent is treated explicitly). Note: although equation 1 shows the non-bonded interactions calculated for every single atom pair, directly connected (i.e. 1-2 bonded atoms, as well as 1-3 typically) are excluded from this (also, a cut-off may be applied so that interactions between atoms over a set distance apart are ignored - see section 2.2.3).

Having established the form of the empirical interaction potentials, parameters must then be found which adequately describe the interactions for every context that will be encountered. The general strategy is to first assign atom types which describe the chemical properties of the atoms e.g. chemical element, hybridisation. Then, parameters are developed which best capture the interactions for all the different possible combinations of these atom types and chemical contexts they are found in. This is somewhat easier for proteins since they are always formed from the same amino acid building blocks and thus all the possible combinations of atom types and chemical contexts (in terms of neighbouring atoms within a residue) is known. Standard protein force fields have thus been developed which describe well the interactions for each atom type/combination in each residue context. Determining what atom types/parameters to apply to the set of atoms/interactions in an arbitrary protein then simply requires specifying a standard atom name for each atom and which residue it is part of (i.e. in PDB format). In terms of development, the bonded parameters are typically optimised to reproduce relevant small molecules geometries/conformational energies from QM/experiment. Similarly, partial charges can be found by fitting to electrostatic potentials of relevant compounds found from QM calculations. The Lennard Jones parameters, being the final unknown, can then be optimised so that the overall system description then reproduces well experimental observables.

As small molecules are more random than proteins, parameterisation is not quite as straightforward. There are however general force-fields which have been developed that include a set of atom types and parameters that cover the majority of conceivable small molecules within a certain class. The General Amber Force Field in one such example which is specialised towards organic molecules (and is used for parameterisation of the substrates/products in this thesis). For commonly occurring combinations of atom types, parameters that fully describe the corresponding interactions are typically available (e.g. dihedral parameters for four sequentially bonded sp^3 carbons). However, generic parameters are also typically included (e.g. torsional parameters for any generic combination of atoms with two middle sp^3 carbon atoms) to ensure that at least some parameter can be assigned to a particular interaction if complete parameters are not available. Therefore, as generic parameters may not accurately describe a given interaction, the caveat with these force fields is that accurate parameterisation may require manually finding or developing certain parameters. Assigning the correct atom types/parameters for a given molecule can also be performed in an automated fashion, despite there being no standard notation for organic molecules for example. Provided the chemical identity of the atoms is provided, connectivity can be algorithmically determined from the geometry of the input structure and the appropriate atom types and parameters assigned. Within the Amber package (34), this is performed by the programs antechamber and parmchk2 (these are run by the Enlighten PREP protocol (35) which was used in this thesis for small molecule parameterisation – see section 4.2.2.2). However, since there is no real way to establish standard values for partial charges, these are instead calculated with an *ad hoc* QM calculation on the small molecule when preparing the system (as part of the antechamber procedure). For the parameterisation performed in this thesis the bcc option was specified to antechamber meaning that the AM1-BCC charge model (36) is used for calculating point charges. This method takes point charges derived from a QM calculation using semi-empirical QM method AM1 and applies empirical Bond charge corrections (BCC) which are parameterised to reproduce the charges obtained from high level HF/6-31G* electrostatic potential fitting. This method therefore enables the generation of high-quality charges with the speed of a semi-empirical method for the QM calculation.

2.1.2 Quantum mechanics

The most fundamental way to describe a molecular system is to use quantum mechanics (QM). Since with QM theory the nuclear and electronic nature of atoms and their interactions are modelled explicitly, this no longer relies upon the coarse empirical description of electronic phenomena that MM provides. The advantage of this is that there is no need to worry about idealised MM parameters always providing an accurate description, and chemical reactions can be modelled. However, this approach is much more computationally expensive than MM so a large system would generally never be treated using QM (other than modelling the individual components of the system with QM in order to generate parameters for an MM model). QM is thus typically only applied for modelling small molecules. However, even for small molecules, when using the most accurate first principles (or *ab initio*) QM methods (which make very few approximations and no reduction of system complexity through parameterisation/fitting), only simple optimisation or energy calculations, rather than actual dynamics simulations, are typically feasible.

A QM system can generally be accurately represented by the time-independent Schrödinger equation (Equation 2). The energy of the system, E, can be obtained by first finding the wavefunction, ψ, since the Hamiltonian operator $\boldsymbol{H}$, when applied to the wavefunction, returns the energy of the system as its eigenvalue.

$$\boldsymbol{H}\psi = E\psi \tag{2}$$

Wavefunction based QM methods thus seek to solve this equation by finding the wavefunction of the system. A fundamental approximation that is first made to simplify the Hamiltonian is called the Born-Oppenheimer approximation (37). This assumes that nuclei can be treated as being fixed points that the electrons travel around, and thus that the equation can be solved for just the electrons (eliminating terms from the Hamiltonian). This is valid as nuclei are much heavier than the electrons, and the speed of the electrons means they will respond almost instantaneously to changes in the nuclei position. However, for all but the very simplest one electron system, the exact wavefunction still cannot be found analytically.

An approach to make solving this equation possible for practical molecules is that of molecular orbital theory. In this approach, it is firstly assumed that the overall electronic wavefunction for the system can instead be described by a set of individual wavefunctions, which represent the molecular orbitals containing the individual electrons. This is known as the molecular orbital approximation (MO). The molecular orbitals are then further described as a linear combination of atomic orbitals (LCAO). These two approximations together (MO-LCAO) provide the fundamental theoretical framework for molecular orbital theory. To describe the atomic orbitals, a set of gaussian functions is typically used which, altogether, form the basis set. This basis set must be sufficiently detailed to describe the molecule and state in question. For example, it must allow sufficient flexibility in the resulting spatial arrangements of electrons to accurately describe the system. Thus, orbitals for valence shell electrons are typically described using a more sophisticated set of gaussian functions. Further, for describing non-ground state electronics, polarisation functions can be used in the basis set (for example the p-type functions for hydrogens and d-type functions for second row elements). These functions provide a higher angular momentum which is important for describing electrons that are farther from the nucleus. Having obtained the atomic orbital description, the problem then becomes finding the correct linear combination of atomic orbitals which make up the individual molecular orbitals and thus the overall wavefunction. However, since the separate molecular orbitals will influence each other, the atomic orbital coefficients cannot be directly calculated, and must instead be found iteratively. Starting from an initial guess for the molecular orbitals, this is used to calculate a new set of orbitals, thus obtaining a better molecular orbital description. This process is repeated until a set of molecular orbitals that are consistent with their surroundings, or so-called self-consistent field (SCF), is obtained. Practically, this involves recalculating the molecular orbitals until the energy of the system converges to within a certain tolerance. As is, this procedure then forms the basis of the Hartree-Fock (HF) theory and method (38). However, the issue with this is that no allowance is yet made for the correlation of the spatial position of electrons due to their repulsion. HF thus yields an energy which is always higher than that for the exact solution of the Schrödinger equation (this systematic error is known as the correlation energy). Post-HF methods (e.g. CCSD which is based on coupled-cluster theory (39) and MP2 which is based on perturbation theory (40)) thus seek to correct for this by extending the HF method to account for correlation. However, these are far more costly, which is where DFT methods show a considerable advantage.

An alternative to wavefunction theory based (or *ab initio*) QM methods is to use density functional theory (DFT). With this approach, the system is described in terms of its electron density rather than the wavefunction. In terms of most of the energy contributions, it is simply the formalism of the DFT approach which differs from the wavefunction based methods. However, the way in which electron exchange and correlation are accounted for (via the so-called exchange-correlation functional) is fundamentally different, as well as being what distinguishes the different DFT functionals from each other. In terms of the correlation term, there is no exact way in which this can be calculated, and many different density functionals have thus been proposed for this. The exchange term can technically be calculated the same way as in HF, i.e. exact exchange. However, this typically does not behave well when combined with the DFT approach to correlation, and so many purpose-built exchange functionals have also been developed. The exchange functional is then either used as is to calculate the exchange contribution (as in the pure DFT functionals), or instead combined with a certain amount of HF exchange (as in the hybrid DFT functionals). The value of DFT methods is that they provide an accounting for correlation, and thus strongly improve on HF, while being effectively as cheap. The issue with these methods is that, unlike the post-HF methods, they cannot be trivially improved upon to improve accuracy (by increasing the basis set for example). Good performance relies upon an appropriate functional being used for the system at hand. Fortunately, many different DFT functionals have been developed, and certain methods have been found to give generally good performance for particular types of systems. As shown in section 3.1.4, the hybrid functional M06-2X (41) has been found to give accurate results when studying DA reactions, and this method has therefore been used for most of the QM calculations in this thesis.

While DFT provides good results with a relatively low computational cost (compared to equivalently accurate *ab initio* methods), it is still too costly for simulating dynamics efficiently, as this requires a large number of energy and force calculations to simulate useful timescales. Further approximations and a reduction of the complexity of the system through fitting can therefore be made, resulting in much faster semi-empirical methods. One group of these methods derive from HF theory and make two key approximations which greatly reduce the computational effort in finding the energy of the system. Firstly, integrals involving overlap between the atomic orbitals of separate atoms are neglected/approximated. Two different approaches to this are thus termed the ‘neglect of diatomic differential overlap’ (NDDO) and ‘modified neglect of diatomic overlap’ (MNDO). Examples of some popular MNDO based methods that were used for testing in this thesis are AM1 (42) and PM6 (43). The second key simplification made by these methods is that only valence electrons are treated explicitly (using a minimal basis set). Inner shell electrons are then described by empirical functions, parameterised by fitting to experimental data. These approximations give these semi-empirical methods a considerable speed advantage over the HF method from which they derive, where a 100-1000 fold reduction in computation time can be expected for a typical HF implementation (44). Further, under ideal circumstances they may actually give more accurate results, since experimental fitting of the introduced parameters can provide somewhat of an empirical correction for the neglected electron correlation in HF. The standard caveat for empirical methods is of course that the general parameters available for them may not always give accurate results, and reparameterization for the specific type of molecule or reaction under investigation may be necessary to ensure this. As for the HF based methods, a semi-empirical treatment has also been applied to DFT. One popular semi-empirical DFT-based method is known as density functional tight binding (DFTB) (45), and this makes similar approximations to those for the wavefunction based methods (resulting in a similar 100-1000 fold speedup in computation time (44)). An extension to this method, called self-consistent charge density functional tight-binding (SCC-DFTB) (46), provides charge redistribution to introduce self-consistency and make this method more applicable to organic molecules containing atoms with different electronegativities. Generally, DFTB methods (and in particular DFTB3) have been shown to perform well for a wide variety of organic molecules and reactions, often reproducing high level *ab initio* results (47). In this work the SCC-DFTB method (also known as DFTB2) was used for QM/MM reaction simulations, while benchmarking was also performed using AM1 and PM6 for comparison.

2.1.3 Combined quantum mechanics/molecular mechanics

Using the combined quantum mechanics/molecular mechanics (QM/MM) approach (48), the molecular system is divided into both a QM and MM region. This means that only the part of the system where it is strictly necessary (i.e. that changes chemically) need be treated with computationally expensive QM, while the rest of the system can be described with the much cheaper MM representation. By combining the advantages of both levels of theory, this makes it possible to model chemical reactions happening in the presence of a large (mostly MM) environment, for example in an enzyme active site surrounded by solvent (an approach pioneered in 1976 by Warshel and Levitt (49)). Generally, when modelling enzyme catalysed reactions, the QM region will therefore consist of just the reacting small molecule and any nearby residues for which a QM representation may be important.

The potential function describing a QM/MM system can be represented by the following equation (this equation shows an additive QM/MM scheme, where the term accounting for the interactions between the QM and MM regions, $E_{QM/MM}$, is explicitly determined):

$$E_{total} = E_{QM} + E_{MM} + E_{QM/MM} \quad (3)$$

E_{MM} represents the energy of the MM region, and since this contains purely MM interactions it is thus calculated solely using the MM forcefield established for the system. Likewise, E_{QM} is the energy of the QM subsystem which is evaluated using the chosen QM method. $E_{QM/MM}$ is then the interaction energy between the QM and MM regions, which can be approached in different ways. In a mechanical embedding scheme, interactions between the QM and MM regions are handled by treating the QM region as if it were also MM. Partial charges and Lennard Jones parameters are assigned to the atoms in the QM region (potentially updating these throughout a simulation as the electrostatics of the QM region change) and interaction energies are calculated according to the standard MM potentials for nonbonded interactions. This approach therefore only captures the classical influence of the environment, i.e. the steric and Couloumbic forces that are imposed on the QM region by the MM region. However, in reality, the electrostatic field of the protein will also influence the position of electrons in the QM region, thereby polarising the electron density (Note: strictly it is the electronic distribution of the environment which does this but since this is not modelled with MM this cannot be represented). A much better approach is therefore that of electrostatic embedding. In this case, steric interactions are still modelled using MM (and LJ parameters therefore still need to be supplied for the QM region). However, the electrostatic field created by the MM region is then also included within the QM calculation. This is done by including a series of one electron operators in the QM Hamiltonian to represent the point charges for each MM atom. This means that electrostatic interactions due to the charge densities in the two regions will then be handled quantum mechanically, but more crucially, that the charge density of the MM region will polarise the electron density of the QM region. This is obviously a much more realistic description in a situation where the environment creates any kind of electrostatic field around the QM region (effectively always). Further, including the effect of environmental electrostatics can be especially important for modelling reactions in condensed phases, given that the resulting polarisation may be (at least partly) what is driving catalysis.

What has not yet been considered is how to handle the situation when a covalent bond crosses the boundary between the QM and MM region. If the QM region consists of just a non-covalently bound ligand, then the only interactions between the QM and MM regions will be nonbonded interactions which are handled according to the different approaches described above. However, if for example a sidechain of a protein residue were to also be included in the QM region, this will then involve a QM atom being covalently bonded to an MM atom. Different methods have been developed to deal with this, however the most straightforward and widely used method is to include a ‘dummy’ link atom (usually hydrogen) which bridges the two regions, and satisfies the empty valence of the connecting QM atom (50). With this method, care must be taken when deciding where to place the boundary, in order to avoid the QM region being over-polarised by surrounding MM atoms. It is thus best placed between a carbon-carbon single bond and far from highly charged MM atoms. It has been shown that, provided care is taken with the charge distribution of MM atoms around the interface, this method can give comparable accuracy to more sophisticated methods (51).

Finally, with respect to the method used in the QM calculation, while any level of theory can technically be applied, when simulating dynamics (for example to sample reactions and obtain free energies) it is most practical to use a semi-empirical method. However, even using these methods, practical MD simulation timescales for QM/MM simulations are still limited to something on the order of 10's of ns. As this is still too short to meaningfully sample reactions, enhanced sampling techniques are therefore typically used to obtain sufficient sampling when studying an enzyme catalysed reaction (see section 2.3.1).

2.2 Energy minimisation and dynamics simulations

2.2.1 Minimisation

Given an initial set of atomic co-ordinates which define the structure of a molecular system, minimisation provides a way to optimise those positions by systematically lowering the energy of the system (as determined by the system description that is applied - as covered in the previous section). The way in which the energy varies with the atomic positions is known as the potential energy surface (PES). Wherever the system starts on this PES, minimisation will thus decrease the energy until a minimum is found. This is done by moving the system down the local energy gradient, which amounts to simultaneously moving each atom in the direction of the net force that acts upon it. This is repeated until a minimum is reached, at which point no net force then acts on the system. A minimum on the PES thus represents a stable conformation, which the system will be pulled towards (from the starting point) as a result of all the potentials acting on it. However, for a large biomolecule the PES is highly complex and there are therefore many such minima. Since minimisation can only move downhill on the PES it will only find the minimum energy conformation which is apparent from the starting point (or local minimum), and not necessarily the overall minimum energy state (or global minimum). In terms of how minimisation algorithms work, as mentioned these lower the energy of the system by moving down the local energy gradient. The simplest method, known as *steepest descent* (SD), simply calculates the gradient or first derivative of the energy, then moves the system by some step size in this direction. This is best for initial minimisation where the system is far from the minimum and the gradient is steep, such that the system moves quickly towards the minimum energy. However, this method will not typically converge well on the minimum as it will tend to oscillate around it before getting very close. Thus, a method called *conjugate gradient* (CG) is then typically used for final minimisation as this can get closer to the true minimum by also considering the gradient at the previous minimisation step. A combination of SD and CG was generally used in the way described for the minimisation steps performed in this thesis.

As mentioned, minimisation algorithms on their own can only optimise the system towards the local minimum. This is therefore a good way to resolve issues with a starting structure which is not optimal, by finding its lowest energy representation. This will ensure that any subsequent MD simulation will not blow up due to large initial forces, or be unphysically catapulted into some other state which is less relevant. However, it could be that the local minimum found from the starting structure is not very stable, and in reality the system will not spend a significant amount of time in this state. In this case we may therefore want to start MD simulations from a more stable nearby state. This is particularly important if running limited sampling (to ensure sampling is obtained for the state that the system can be expected to reside in more of the time), as is the case for the approximate simulations run in this thesis using the Enlighten protocols (see section 4.2.2.2). In this case, a more global optimisation procedure could first be performed where the system can explore the nearby PES. This is the purpose of the simulated annealing stage of the struct protocol within Enlighten. After an initial minimisation of the starting structure, struct performs short MD simulation at high temperature before cooling down (also known as simulated annealing). This provides the kinetic energy for the system to quickly climb and explore the nearby energy landscape, then either reaching a more stable state or confirming the stability of the starting conformation. The system is then reoptimized to the closest minimum, which then more likely represents a state that will be well populated at the target temperature.

2.2.2 Molecular dynamics

Having obtained a relaxed structure via minimisation, it is then of interest to sample the typical conformations that will be visited by the system in a given physical regime (i.e. temperature and pressure). This is essentially a statistical mechanics problem, where one would like to sample states which reflect the correct statistical ensemble. One way to do this is using molecular dynamics (MD), which involves physically simulating the motion of the atoms according to Newtons second law, $F = ma$ (i.e. given their mass and the forces acting on them due to all the potentials surrounding them), thus exploring the different possible conformations. This is valid since nuclei are heavy enough to be modelled classically (and in the case of QM atoms, the Born-Oppenheimer approximation means that only motion of the nuclei themselves need be considered, where electrons can be presumed to rearrange instantly around them). For a molecular system, this equation can be written as follows:

$$-\frac{dV}{dr} = m\frac{d^2r}{dt^2} \tag{4}$$

Where r is the vector of atomic co-ordinates, and V the potential energy of the system (as described previously). This equation can be numerically integrated using the *Verlet* algorithm (52), thereby propagating the positions and velocities forward in time, with a discrete timestep. By doing this, a trajectory of the conformations visited over time is thus obtained. Given sufficient sampling, this should then approximate the correct ensemble distribution, since this will be determined by the average time spent in each conformation. The smaller the timestep used, the more accurate the simulation becomes, but the longer it will take to simulate a given length of time (and obtain sufficient sampling for states of interest). Since, in large biomolecules, it can take a long sampling time to reach relevant conformations (and given that calculating each successive conformation requires a large computational effort due to the number of atoms/interactions), it is of interest to use the maximum timestep possible which still yields accurate results[1]. The absolute maximum timestep that can be used for a simulation to remain stable is determined by the fastest process in the system. In a molecular system this will be the stretching vibration of bonds involving hydrogen atoms. However, since these do not greatly affect the results, hydrogen bonds can be constrained using the SHAKE algorithm so that a larger timestep may then be used. When using SHAKE, a timestep of 2 fs is typically used, while a 1 fs timestep is required without it (thus doubling the required computation time). This setup was therefore used for the MD simulations in this thesis.

If a molecular system as so far described were simulated i.e. using equation 4, this would be a simulation of an isolated system which simply maintains its starting energy. This could be described as the microcanonical or NVE ensemble, where the number of particles, volume and energy are constant (assuming also some boundary conditions are present to keep volume constant – see next section). However, a more relevant ensemble may be that which will be found in reality, where the energy in the system will change as there is energy transfer with the environment. In this case, the other macroscopic observables (i.e. temperature and pressure) will instead be constant as they are controlled by the surrounding environment. The conditions of interest must therefore be mimicked in simulation by controlling either temperature alone, or both temperature and pressure (simulation in the NVT or NPT ensemble, respectively). If simulations are run in the NVT ensemble, then the energy of the system (and pressure) will fluctuate as energy transfer occurs to maintain a constant temperature. One way to achieve this is via a so-called thermostat method, such as the weak-coupling algorithm (53) illustrated by equations 5 below. In this approach the system is coupled to a “heat bath”, which provides gradual energy transfer with the system in order to maintain the desired temperature (rate of heating is controlled by τ, the time constant for coupling to the heat bath). The kinetic energy of the system is adjusted at each timestep by scaling the velocities as indicated in equation 5, to achieve the desired rate of change in temperature. Another method for controlling

[1] As well as this, multiple simulations - with different starting conformations or random initial velocities - will generally be run to obtain better sampling of the ensemble distribution for a fixed length of sampling time.

temperature is to use Langevin dynamics. This introduces a frictional component and random collisions into the equations of motion, thus providing the necessary kinetic energy[2]. If pressure is also to be controlled, thus running in the NPT ensemble (which corresponds to the typical experimental setup), then the volume of the system is also varied as necessary in order to maintain a constant pressure. Similarly to that used for temperature control, a barostat method (53) can be used for this where the system is then also coupled to a "pressure bath". All production MD simulations in this thesis (for full periodic boundary simulations – see next section) were run in the NPT ensemble, while the NVT ensemble was used for certain equilibration steps.

$$\frac{dT_{actual}}{dt} = \frac{1}{\tau}(T_{desired} - T_{actual})$$

$$scaling\ vector = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{desired}}{T_{actual}} - 1\right)} \qquad (5)$$

2.2.3 Boundary conditions

When running a molecular simulation with explicit solvent, the molecule of interest will be surrounded by a small finite region of solvent. This will typically only be large enough to surround the solute and provide a buffer region around it of a comparable size to the solute itself (as much more will lead to a prohibitively large number of atoms for efficient simulation). Without anything to stop this, solvent molecules will gradually drift away at the boundary, evaporating into the surrounding vacuum (i.e. the simulation will eventually become a gas phase simulation instead of that for a proper aqueous environment). The simplest way to address this is to fix the positions of atoms at the surface, or atoms than are more than a certain distance from the region of interest. Solvent and solute will thus stay confined within the region defined by the frozen atoms. This could technically be described as an NVT representation as the volume of the system will then be fixed. However, this is a very unrealistic situation as the system will be simulated as if it exists within a very small finite container which is restricting its motion. In reality, the solute will exist within a bulk aqueous environment, where it is generally free to move around without coming into contact with anything other than solvent. A way to replicate the bulk environment with a finite number of explicit solvent molecules is to instead use periodic boundary conditions (PBC). With this setup, the solute is contained within a polyhedral (generally cubic) box of solvent. The system is then duplicated in all directions, creating a repeating array of periodic images. If a molecule of solvent were to leave the central box, thus moving into a

[2] Langevin dynamics can help ensure a more even spread of temperature across the system – it is particularly useful for implicit solvent simulations where there are no collisions with solvent molecules to aid this.

neighbouring box, its image from the opposite neighbour then also moves into the central box. In this way it is as if the surrounding solvent environment is never-ending and the solute may thus move around freely within it. However, since the periodic images are all copies, interactions and motion only need to be calculated for the atoms making up a single periodic image. Of course, there will now be infinite non-bonded interactions to calculate, since there are infinite copies of each atom in the system (thus making these interactions impossible to compute on a per-atom basis). A cut-off distance is therefore used for certain non-bonded interactions, such that atoms pairs that are more than a set distance apart are ignored (typically 8 Å is used for this). This is suitable for VDW interactions as these are effectively negligible beyond a certain distance. However, electrostatic interactions can be much longer range and so it is not generally valid to apply such a cut-off. Instead, the problem of calculating interactions for an infinite set of periodic images can be solved using the Particle-Mesh-Ewald method (54), which is then used to calculate the full electrostatics based on Ewald summation. A final issue to mention for PBC simulations is that the box of solvent must be made large enough to avoid certain periodic artefacts, such as a macromolecule interfering with its periodic image.

PBC were used for MD simulations in this thesis studying loop motion, since using fixed boundary conditions would interfere with loop motion as the protein meets the boundary (and it is thus important that the enzyme be realistically modelled as freely moving in a bulk environment). However, QM/MM reaction simulations and MM MD simulations of substrate complexes (for calculating binding affinity) were run using the Enlighten protocols, which employ fixed boundary conditions (see section 4.2.2.2). Using these protocols, a solvent sphere was first added which was centred on the enzyme active site, and made just large enough to surround it by a few layers of water molecules. The region which is allowed to move in subsequent MD is then just the atoms within this solvent sphere. In this case, rather than being allowed to move in a small container of solvent, the overall protein molecule was then also fixed in space, since part of it extended outside the solvent sphere. By using this setup, far fewer solvent molecules were required (compared to PBC), thus greatly reducing the total number of atoms and making the simulations very computationally efficient. The assumptions made to justify this approach were 1) that overall protein motion within a bulk solvent environment was not essential for studying the reactivity and interaction energy for the protein-ligand complex, and 2) that only the solvent environment near the active site would strongly affect these properties.

2.3 Enhanced sampling

2.3.1 Umbrella sampling

As mentioned, reactions typically occur on longer timescales than can feasibly be simulated in QM/MM MD simulations (even when using semi-empirical QM methods). Determining the average i.e. macroscopic rate would also require observing many reactive trajectories, which is therefore not possible using simple unbiased MD. What can instead be done is to use an enhanced sampling method such as umbrella sampling to determine the free energy profile for the reaction. With umbrella sampling, a biasing potential is introduced that allows conformational sampling to be performed at incremental points along the reaction path. From this a free energy barrier for the reaction can be determined which can then be related to the macroscopic rate. What is first required is therefore a reaction co-ordinate (r) which describes how the system changes during the reaction, and can thus be used to restrain it to a certain point in terms of reaction progress. A series of harmonic biasing potentials (see Equation 6) are then specified which act to restrain the system near a desired value of the reaction co-ordinate (r_0).

$$U(r) = k_U(r - r_0)^2 \tag{6}$$

The specified restraints are centred around incremental values of the reaction co-ordinate, going from reactant to product, and applied in separate umbrella sampling simulation windows. Each of these windows consists of a separate QM/MM MD simulation, in which the system will thus be biased to sample around the reaction co-ordinate value corresponding to the window. In this way sampling can be obtained across the relevant range of the reaction co-ordinate. To ensure that good sampling is obtained along this entire range, the force constant (k_U) used for the biasing potential should be strong enough that the system is held near to the desired reaction co-ordinate value, but also result in sufficient overlap between the distributions of the sampled reaction co-ordinate values for adjacent windows. Assuming sufficient sampling, the unbiased probability distribution along the reaction co-ordinate (or the probability that the system will be found at a given value of the reaction co-ordinate, without the bias present) can then be accurately found via reweighting. This can in turn be related to the free energy profile or ‘potential of mean force’ (PMF) along the reaction co-ordinate, where the free energy barrier for the reaction is then given by the free energy change in going from the reactant minimum to the transition state.

In terms of reweighting, different possible approaches exist such as Free energy perturbation (FEP) (55), weighted distribution function (W-DF) (56) and the weighted histogram analysis method (WHAM) (57). WHAM has the advantage that it combines results from all the separate umbrella sampling simulation windows. In this way it benefits from the convergence properties of using more umbrella sampling windows, which then offsets the error introduced by the added uncertainty in matching more adjacent windows (58). This is computationally advantageous as using more windows to cover a given range means that less total simulation time is required to sample across the windows (58). For the work in this thesis, the Grossfield implementation of the WHAM method (59) was used for reweighting to remove the bias from the simulation results and obtain the free energy profile along the reaction co-ordinate.

2.3.2 Gaussian Accelerated MD

Gaussian Accelerated MD (GaMD) (60) is an approach to enhance sampling of configurational space in MD simulations of biomolecules. Large conformational changes for biomolecules can occur on timescales which are not easily accessible using conventional MD (cMD), due to the inherent energy barriers for transitions between the different stable states. This therefore makes it infeasible to obtain sufficient sampling to quantify the stability of different states/rate of transitions between them. GaMD works by adding a boost potential to flatten the PES and reduce energy barriers, thus accelerating transitions between states. Similarly to umbrella sampling, once sufficient sampling has been obtained from the modified PES, reweighting can be performed to account for the bias and infer the unbiased probability (and thus free energy) for observing the different states of interest. Key with this approach is that no specific reaction co-ordinate need be specified in advance to describe the process under investigation, or even what the target state looks like for example. GaMD simply enables efficient exploring of all conformational space, where the states of interest can then be analysed once they have been sampled. Of course, if one is interested in the energy barrier for a transition in order to give information on the expected rate, then an appropriate reaction co-ordinate may subsequently be needed to characterise the pathway (and obtain the relevant free energy profile along it from the obtained sampling). However, the key advantage with GaMD is that this information is not needed *a priori* in order to observe the transition.

The GaMD boost potential is a harmonic term which gets added to the potential energy for the system, $V(r)$, when it is below a threshold energy, E. This gives the modified system potential as follows:

$$V^{*}(r) = V(r) + \frac{1}{2}k\big(E - V(r)\big)^{2} \ , \quad V(r) < E \qquad (7)$$

Some prior MD must first be run to sample the potential energy surface and determine what the boost parameters should be. Given some maximum and minimum potential energies sampled for the system (V_{max} and V_{min} respectively); for the boost potential to flatten the energy surface (i.e. lower energy differences while maintaining their order, and thus the overall shape of the PES) for energies within the sampled range, the threshold energy must be set as follows:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k} \tag{8}$$

The force constant k must therefore also satisfy the following:

$$k \leq \frac{1}{V_{max} - V_{min}} \tag{9}$$

k determines the magnitude of the applied boost potential and how much the energy surface is flattened. A higher k will therefore provide greater barrier lowering and increased acceleration. As mentioned, the above upper limit for k is determined so that the boost potential preserves the shape of the energy surface while lowering energy differences. However, k will also affect the standard deviation of the boost potential, $\sigma_{\Delta V}$, as follows (when using cumulant expansion to the 2nd order):

$$\sigma_{\Delta V} = k(E - V_{av})\sigma_V \leq \sigma_0 \tag{10}$$

where V_{av} and σ_V are the average and standard deviation of the system potential energy respectively. Since a narrow distribution of $\sigma_{\Delta V}$ is required for accurate reweighting, an upper limit, σ_0, is set for this value (e.g. 10k_bT which is the default). This thus imposes another limit for k in order to enable accurate reweighting (when using cumulant expansion to the second order), given also the system energy statistics and threshold energy used. When using the lower energy threshold (V_{max}), an upper limit on k is found from equation 10 and so the lower of the two limits will be applied to give the maximum allowable acceleration. When using the upper limit for the threshold energy ($V_{min} + \frac{1}{k}$) a lower limit is obtained for k from equation 10. Since, in this case, a lower value of k gives a higher threshold energy, a value of k up to the upper limit can be used to give the greatest force constant, or a value of k down to the lower limit can instead be used to the highest threshold energy. The current GaMD implementation has different defaults for this depending on the value calculated for the lower limit of k.

The boost parameters E and k are initially calculated based on sampling for the system from cMD simulation. However, the boost potential is then adaptively added to the system (as the boost parameters are dynamically updated) during several preparatory accelerated MD phases i.e. the final boost parameters will be determined after first exploring the PES with some acceleration. This means that the boost parameters used for the production simulations will be appropriate for smoothing the overall system energy surface, thus providing more global acceleration while still leading to a narrow enough distribution of the boost potential $\sigma_{\Delta V}$ to allow for accurate reweighting. Finally, in the current GaMD implementation, the boost potential can be applied to either just the total potential, the dihedral potential, or to both the total and dihedral potentials (dual boost setting). Using the dual boost setting tends to result in better acceleration, and this was therefore used for the work in this thesis.

2.4 Molecular Mechanics / Poisson Boltzmann (or Generalized Born) Surface Area (MM/PB(GB)SA) Calculations

MM/PB(GB)SA (61) is a method for finding free energy differences between two states of a system (for example the bound and unbound state of a ligand with its receptor) which was used in this thesis for predicting substrate binding affinities. Unlike other methods for finding relative free energies such as Free Energy Perturbation or Thermodynamic Integration, it considers only the end states of the binding process. This makes it far less computationally demanding, since intermediate states are not required to be simulated. Furthermore, as implicit solvent is used to capture the effect of solvation on the binding energy, this reduces the noise (compared to explicit solvent) since the solvent effect is averaged out[3].

In the MM/PB(GB)SA method, the binding free energy is obtained by calculating solvation free energies for representative conformations of the component species in the bound and unbound state. The solvation free energy represents the free energy required for transferring a particular conformation of a solute from gas phase into solvent. The overall binding free energy in solvent ($\Delta G_{bind,solv}$) is thus calculated via an equivalent thermodynamic cycle, splitting the calculation into the binding free energy in gas phase ($\Delta G_{bind,gas}$) and the solvation free energies for the individual components ($\Delta G_{solv,x}$):

$$\Delta G_{bind,solv} = \Delta G_{bind,gas} + \Delta G_{solv,complex} - \left(\Delta G_{solv,ligand} + \Delta G_{solv,receptor}\right) \qquad (11)$$

The solvation free energies are further split into their polar and nonpolar components, where the implicit solvent model is used to calculate the polar contribution and the nonpolar contribution is estimated using the LCPO method (62). The Poisson-Boltzmann (PB) method is the most rigorous implicit solvent formulation and thus the most accurate way to calculate polar solvation free energies.

[3] If explicit solvent were used in the direct calculation of energy differences between bound and unbound states, fluctuations in the solvent-solvent interactions would dominate over the binding energy and it would thus take a very long time to obtain converged results.

However, calculating the solvation free energy for a given solute geometry using this method requires solving second order partial differential equations, which therefore makes it computationally intensive. The Generalised Born (GB) model thus provides a computationally efficient approximation to solving the PB equations. The approximation involves calculating effective Born radii for solute atoms which describe the degree of their burial within the solute and thus the level of screening of electrostatic interactions for a given atom pair by the surrounding solvent (63).

In order to determine the gas phase binding free energy, this is split up into its energy and entropy contributions:

$$\Delta G_{bind,gas} = \Delta E_{bind,gas} - T\Delta S_{bind,gas} \tag{12}$$

The binding energy $\Delta E_{bind,gas}$ is the difference in gas phase energy between the complex and that of the isolated ligand and receptor. This is calculated with the MM forcefield for some representative conformations of each species (if obtaining these from a single trajectory (see below) this will simply be the interaction energy between the ligand and receptor in the complex). The contribution due to the entropy change upon binding ($T\Delta S_{bind,gas}$) can then be calculated by estimating the entropy of each species. The translational and rotational entropies can be estimated with standard statistical mechanical formulae, and the vibrational entropies calculated by either normal mode analysis or the quasi-harmonic approximation. Normal mode calculations are costly however, and quasi-harmonic calculations require a large ensemble of conformations (thus requiring extensive sampling for each state). If comparing the relative binding free energy of similar ligands with the same receptor, the entropic contribution is thus often neglected as it may be assumed that the entropy change upon binding will be similar and so will effectively cancel out.

Other than the entropy component of the gas phase binding energy, each of the contributions to the binding free energy are calculated by averaging them for an ensemble of representative structures for the bound and unbound states. The simplest and cheapest way to obtain these is via the so-called single trajectory approach. In this case, a single (typically explicit solvent) MD simulation is run for the complex, and the gas phase structures for the individual components are extracted from that. This therefore assumes that similar conformations will be explored by the ligand and receptor, whether free in solution or in the bound complex. If this were not the case, then separate simulations can also be run to generate representative structures for the both the free ligand and receptor.

Chapter 3. The Diels-Alder reaction involved in abyssomicin C formation

3.1 Introduction

The aim of this chapter is to characterise the intrinsic Diels-Alder (DA) reaction catalysed by AbyU in the Abyssomicin C pathway (see section 1.2.2.2.2.1). In the first instance, this will provide some general insight into spirotetronate synthesis. More directly relevant to this work, an understanding of the baseline reaction is necessary to understand how the reaction is affected by the presence of the enzyme. The intrinsic reaction was primarily studied via gas phase DFT calculations, with implicit solvent included for a particular aspect to approximate the effect of the environment. In addition, benchmarking was performed to assess the suitability of different semi-empirical QM methods for describing the intrinsic reaction to help validate their use in the modelling of the enzymatic reaction in Chapters 4 and 5. Before describing the details of the investigation in this Chapter, the background theory for the DA reaction and the principles which govern it are reviewed as this provides a foundation for understanding the AbyU reaction.

3.1.1 Diels-Alder reaction background & theory

3.1.1.1 Overview

The Diels-Alder reaction takes place between a conjugated i.e. 1,3 diene (which must be in the reactive s-*cis* conformation) and an alkene group. The simplest example of this reaction is between butadiene and ethylene to yield cyclohexene (Figure 3.1). During the reaction, the three formal π bonds in the reactants are converted into two new σ bonds and a new π bond (Figure 3.1). Upon reaction, the 3 pairs of electrons involved in the 3 π bonds rearrange simultaneously; one pair moves to form a new π bond in the centre of the diene and the other two move to form new σ bonds between the two reactants thus joining them into a single cyclohexene ring. The reaction is thus also known as a [4+2] cycloaddition as it involves the 4 π electrons in the diene and the 2 in the dienophile, and adds the two reacting groups together to form a single cyclic product.

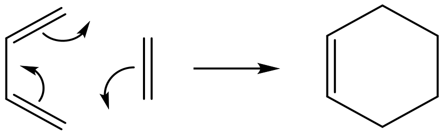


Figure 3.1: The Diels-Alder Reaction between butadiene and ethylene to form cyclohexene.

To facilitate the electron rearrangement, as well as the diene being in the s-*cis* conformation, the end π orbitals in the two groups need to overlap. The planes in which the two reactants sit must thus also be roughly parallel as they approach each-other. Also, although the reaction is concerted, with all bond breaking and making happening in a single step, it need not be synchronous i.e. one bond can start to form before the other (64). In the conventional or 'normal demand' DA (which is the relevant case for the AbyU substrate), the reaction is initiated by electron transfer from the diene HOMO to the dienophile LUMO. This type of reaction is favoured by the diene being electron rich and the alkene group electron deficient, and the greater the disparity in electron distribution, the more prone they are to react. Therefore, making the diene more electron rich by adding an electron donating group (EDG) or making the dienophile more electron poor by adding an electron withdrawing group (EWG) will enhance the reaction rate. This rate enhancement occurs because the EDG raises the energy of the diene HOMO, while the EWG lowers the energy of the dienophile LUMO (therefore lowering the energy gap between the Frontier Molecular Orbitals and the activation energy required to initiate the electron transfer) (65). Reactants with these groups can be considered to be more 'activated' (tending to also produce more asynchronous bond formation – see section 3.1.4).

Finally, it is worth mentioning that, although the concerted s-*cis* mechanism described is by far the most favourable way for this reaction to proceed, for butadiene and ethylene at least the same outcome could be achieved with the diene starting in the s-*trans* conformation. In this case the reaction still occurs in essentially the same way, but has a free energy barrier approximately 20 kcal/mole higher and results in a higher energy conformer of the product (66). Also, for both the s-*cis* and s-*trans* conformations there exist stepwise routes (where the new bonds forms in separate steps and the reaction thus goes through biradical intermediate species), however, as the overall activation free energies are considerably higher for these, they are unlikely to occur (66).

3.1.1.2 Diene conformation

As the diene is conjugated, it adopts one of two possible planar conformations: s-*cis* or s-*trans* (Figure 3.2). These are analogous to the *cis* and *trans* conformations for alkenes, where *s* indicates they are separated by a rotation about a formal σ bond (with partial double bond character) rather than a true π bond as for alkenes.

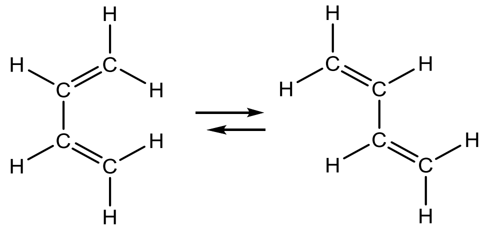


Figure 3.2: s-*cis* (left) and s-*trans* (right) conformations of butadiene and interconversion between them. Steric hinderance caused by end carbons being on the same side of the middle C-C bond makes s-*cis* the less stable conformation.

The diene must be in the *s-cis* conformation to make the DA reaction energetically feasible. Therefore, the more favourable this conformation, the greater the proportion of reactant that will be prone to react through a low energy route and thus the faster the reaction will be (Figure 3.3). However, due to steric hinderance in s-*cis* (Figure 3.2, Figure 3.3) this is usually not the preferred conformation. In butadiene, for example, s-*trans* is preferred by a modest energy difference of about 3-4 kcal/mole (66). In addition, as nonplanar conformations are higher in energy (as conjugation is broken) there is an associated energetic barrier in switching between s-*cis*/s-*trans*, and so this barrier must also be low enough to allow the diene to reach the s-*cis* conformation. The barrier for conversion is 7 kcal/mole for butadiene (66) which is low enough to allow for fast conversion (e.g. at room temperature).

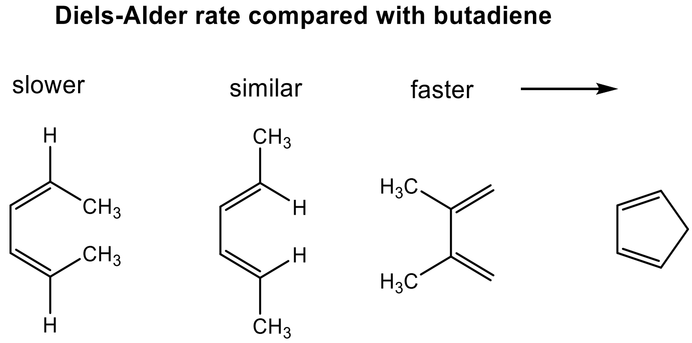


Figure 3.3: Effect of the favourability of the s-*cis* conformation on the resulting DA reaction rate. Going from left to right dienes have increasingly more favourable s-*cis* conformations.

3.1.1.3 Stereoselectivity

In the simple case of ethylene and butadiene only a single product is formed (cyclohexene). However, if there is more than one type of substituent on any of the 4 carbons between which the new σ bonds form, then the reaction can result in different stereoisomers.

The Diels-Alder reaction typically provides good regio- and stereoselectivity, and the favoured product can often be predicted by consideration of intramolecular interactions which result from the various substituents. For example, if the dienophile has an EWG attached then stereochemical outcome can be predicted, since stabilising secondary interactions will be made between the EWG and the π orbitals of the diene in the TS corresponding to one particular diastereomer product (Figure 3.4). These interactions will only be made when the EWG is in the *endo* position and so, since the way in which the reactants face each other cannot change over the course of the reaction, the diastereomer product which forms will also have this group in the *endo* position. However, although the *endo* product is favoured kinetically, it is typically not thermodynamically favoured as there is more steric hinderance by having a bulky EWG in the *endo* position (this does of course depend on the nature of the other substituents and, if they are equally bulky for example, this could mean the *endo* product is also thermodynamically preferred). Therefore, the *endo* product will generally be produced under kinetic rather than thermodynamic control.

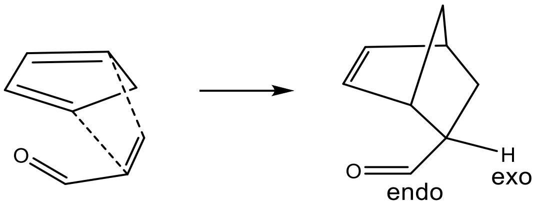
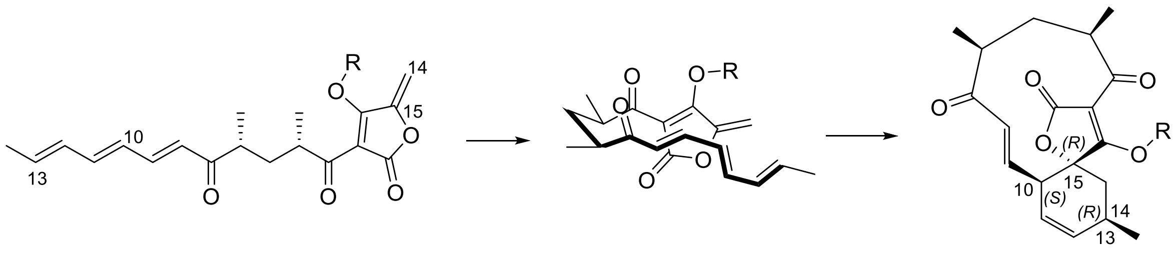


Figure 3.4: When the dienophile has an EWG attached, stabilising interactions in the TS favour the product with the EWG in the endo position.

3.1.2 Details of the AbyU reaction

The reaction catalysed by AbyU is intramolecular which means that the diene and dienophile are on the same substrate molecule. The substrate **9** (Figure 3.5) consists of a tetronate ring, attached through a polyketide chain to a triene moiety. The exocylic methylene of the tetronate forms the dienophile for the reaction (C14-C15), while the end diene of the triene (between C10 and C13) forms the other reacting group. Thus, what happens in the reaction is that the substrate first folds into a reactive conformation to bring the diene and dienophile together, such that the new bonds can then form between them (one forming between C10 and C15 and the other between C13 and C14). As in the general case, the diene must be in the s-*cis* conformation and the two groups must approach each other in parallel planes to make the reaction energetically feasible. During the resulting cycloaddition, the new single bonds which form between the two groups joins the two ends of the substrate through the resulting cyclohexene ring, forming the cyclised spirotetronate product.



R=H: **9**
R=Me: **10**

R=H: **9P**
R=Me: **10P**

Figure 3.5: Reaction of the AbyU substrate (both the native substrate **9** and the methoxy analogue substrate **10** which was used in experiments). Showing formation of the 'P' atropisomer of the product with the chirality which is formed in both the enzymatic and spontaneous reactions.

In terms of regioselectivity, only one structural isomer is possible. However, multiple products are possible due to the new chiral centres which form in the product (at C10, C13 and C15). The resulting product will depend on the relative orientations of the substituent groups in the reactive conformation, as this will dictate the arrangement of these groups at the new chiral centres. In this sense, there are 4 different reactive conformations which correspond to 4 possible reaction products (see below). In addition to this, when forming the reactive conformation corresponding to a particular chiral product, the polyketide chain can adopt two distinct conformations (characterised by whether the carbonyl groups in the chain point in the same (as in Figure 3.5) or opposite directions). Since, upon cyclisation, the chain cannot easily interconvert between these conformations, this is therefore a further source of isomerism, known as atropisomerism. (In order to be considered actual atropisomers, rather than just conformers, the barrier to this interconversion must be about 23.3 kcal/mole such that they are fully resolvable for >=1000 seconds at room temperature (67). For each possible chiral product there are therefore also two possible atropisomers which can form. Note: to distinguish between the two atropisomer products in this thesis these are named by appending a 'P' or 'A' to the compound number for the substrate from which they are formed, indicating that the carbonyl groups point in either the same or opposite directions, respectively (see Figure 3.5 for example).

The four possible product outcomes in terms of chirality are shown in Figure 3.6 (where the two possible atropisomers are shown for each). While the enzymatic Diels-Alder product itself has not yet been formally identified, an X-ray structure exists for the end product of the pathway Abyssomicin C (Figure 3.7). By comparing the stereochemistry at each chiral centre, the DA adduct which goes on to form Abyssomicin C can be seen to be **A** in Figure 3.6, which has stereochemistry *S*, *R*, *R* on C10, C13, C15 respectively (Note: only the atropisomer of Abyssomicin C corresponding to the product in Figure 3.5 is shown - but both are known to form in the biosynthetic reaction). A second possible product is the diastereomer of **A** (**B** in Figure 3.6) which results from the substituent groups being arranged oppositely on the two chiral centres originating from the diene (C10, C13), leading to a stereochemistry of *R*, *S*, *R* on C10, C13, C15 respectively. Which of these two products will form is determined by which face of the diene gets presented towards the dienophile in the reactive conformation (while keeping the rest of the molecule in the same orientation). In terms of the expected stereoselectivity, **A** could be described as the *endo* product (see section 3.1.1.3) since the only EWG is that formed by the methoxy group (or hydroxyl in the case of the native substrate) attached to the tetronate ring (which is in the *endo* position for **A**). This consideration would therefore support this also being the favoured product in absence of the enzyme, which is exactly what has been found by previous experimental and computational efforts (see section 3.1.3).

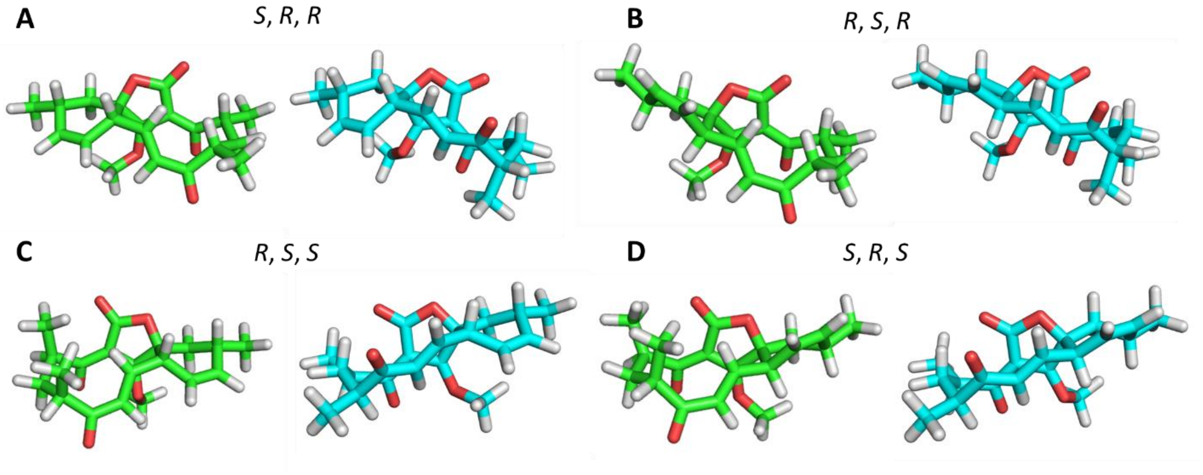


Figure 3.6: 4 possible product outcomes A-D (in terms of chirality) resulting from the reaction of **10** (both atropisomers are shown). Chirality shown in panel **A** is that formed in the biosynthetic/spontaneous reactions.

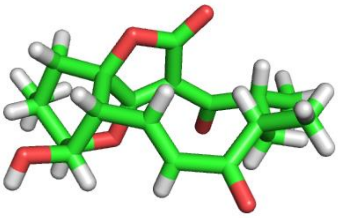


Figure 3.7: One particular atropisomer of Abyssomicin C. Isolated from actinomycete Verrucosispora strain AB 18-032 via HPLC and structure determined by X-ray crystallisation and confirmed by NMR (68).

The two remaining possible products (**C** and **D**) are the equivalents of **A** and **B** with opposite stereochemistry at the remaining chiral centre C15 (and are thus their enantiomeric counterparts). These have stereochemical configurations of *R*, *S*, *S* and *S*, *R*, *S* (for **C** and **D** respectively). What dictates whether these are formed is if the polyketide chain folds around to meet the opposite face of the tetronate ring than it does for **A** and **B**. This is clearer when viewing the products side by side and looking approximately into the plane of the macrocycle (Figure 3.8). From this viewpoint it can be seen that, in terms of the spirotetronate skeleton, the two enantiomeric counterparts (**A** and **C** or **B** and **D**) are mirror images of each other. However, it should be pointed out that, because of the methyl groups attached to the chiral carbons in the polyketide chain, symmetry is broken and so the products are not actually mirror images. As such, the energetics involved in the formation of **C** and **D** will not be the same as for **A** and **B**. However, the same structural distinction separates **C** and **D** as does **A** and **B**, thus **C** (the enantiomeric counterpart to **A**) can be called the *endo* product and **D** the *exo* product.

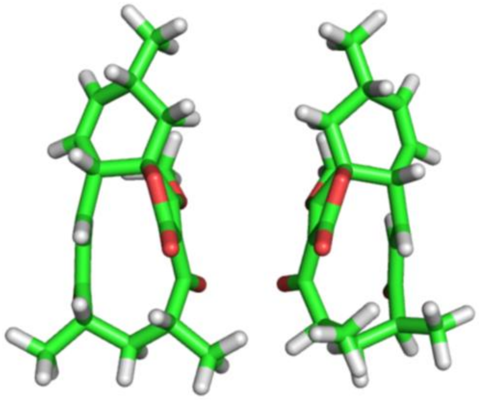


Figure 3.8: Side-by-side view of stereoisomer S,R,R shown in Figure 3.6A (left) and its enantiomeric counterpart R,S,S shown in Figure 3.6C (right).

3.1.3 Existing knowledge and aims of this study

All experimental evidence for the reaction reported in this section is for the methoxy analogue substrate **10**. Firstly, the reaction has been shown to occur spontaneously (25, 69), albeit much more slowly than the enzyme catalysed reaction (25). A non-enzyme catalysed rate of 0.0014 min^{-1} was found while a k_{cat} of 564 min^{-1} was found for the enzymatic reaction (representing a 4x10^{4} min^{-1} rate enhancement). The spontaneous reaction has also been found to be highly stereoselective, yielding the same products (Figure 3.6A) which are formed in the biosynthetic reaction (69). This has been rationalised previously by comparing TS energies for the various possible reaction products at the Hartree Fock (HF) level, demonstrating a lower energy for the preferred product (69). Mechanistically, the reaction appears to take place via a normal demand route, since making the dienophile of the substrate less electron deficient by replacing the carbonyl group conjugated to the tetronate ring with a hydroxyl group (resulting in a mixture of epimeric alcohols) led to no observable cyclisation under the same reaction conditions as for **10** (25). Gas phase transition states optimised for the two atropisomer products of **9** at the M06-2X/6-31G(d,p) level (25) indicated an asynchronous mechanism, where bond formation is more advanced for the C13-C14 bond (~2.0 Å) than the C10-C15 bond (~2.70 Å).

To understand how this contributes to the intrinsic reactivity (as well as for benchmarking semi-empirical QM methods), the activation energy for the reaction is now determined in gas phase. In the previous study (25) an approximate gas phase energy profile was calculated for the reaction by doing optimised scans along the two bond distances starting from the product (optimising at the B97D/6-31G(d,p) level and then recalculating energies at the M06-2X/6-31G(d,p) level). To obtain the proper reaction path and energy profile (in contrast to the previous approximate scans), IRC calculations are now performed. In addition, M06-2X has also been used as the optimisation method here as well as the method for energy calculation (although B97D has also been used for comparison). This is a hybrid DFT functional which has emerged as being useful (for both optimisations and energy calculations) for studying DA reactions (see section 3.1.4 below). For benchmarking, reaction profiles were obtained using the semi-empirical methods DFTB2, PM6 and AM1. Next, to gain further insight into reactivity, the thermodynamics and kinetics involved in the diene adopting the s-*cis* conformation have been investigated. As seen previously, the reacting diene needs to be in the s-*cis* conformation for the reaction to take place, and as this is typically not the favoured conformation due to steric hinderance, a DA reaction may be enhanced by encouraging the formation of this conformation. The torsional potential energy profile for the reacting diene was therefore obtained by scanning along the relevant dihedral angle. The final factor investigated is the likelihood of the substrate being in a folded up reactive conformation as opposed to a more extended one, by comparing their stability (in terms of free energy). As was discussed in Chapter 1 (see section 1.2.2.2.2), entropy may lead to this conformation being energetically unfavourable. Constraining the substrate in this conformation could thus be the primary source of catalysis in AbyU.

3.1.4 DFT functionals for DA reactions

While B3LYP (70) is a popular and widely used DFT functional which has been used in the past for studying DA reactions, it has been shown to consistently overestimate activation energies (71) while underestimating reaction exothermicities (71, 72), which is further exacerbated by a larger basis set size (71, 72). Although other DFT functionals are found to - in some cases considerably - overestimate reaction exothermicities, energies are instead improved by moving to a larger basis set (72). In (72), to identify the source of this error and its relationship with basis set size, the enthalpy for a π > σ transformation (2 of which are involved in the DA reaction) was calculated using different DFT functionals at the different basis set sizes. The change in calculated transformation energy with basis was found to be similar for all methods (~3-4 kcal/mole) and accounted for the systematic change to reaction energies found from increasing the basis set size. Since this is a systematic error, and given that B3LYP already typically underestimates reaction exothermicities (by 8 kcal/mole on average for reactions in (72)) with a smaller basis set, this makes it a generally inappropriate method for studying DA reactions. Of the other DFT functionals, M062-X performed the best since it on average overestimates reaction exothermicities by just a few kcal/mole at the lower basis set size, and thus typically has only a very small error using a larger basis set. M06-2X has thus gone on to be used for many studies of DA reactions, although B3LYP has still been relied upon for geometry optimisations in many cases (72, 73). However, more recent studies by Brinck *et al.* suggested that B3LYP may not be the best method for geometries either, as it was found that it tends to produce overly asynchronous TS's compared to M06-2X and MP2 for strongly activated systems (74, 75). A recent benchmarking study (76) by the same authors therefore set out to determine, and to some extent rationalise, the applicability of various DFT functionals for studying DA reactions. TS geometries (plus activation energies) were obtained for a range of DA reactions with variously activated reactants using a selection of pure and hybrid DFT functionals (where different combinations of the different correlation and exchange components were also tested). In terms of activation energies, M06-2X (along with several of the other DFT functionals) was again found to be a good method, giving predictions within 1.5 kcal/mole of the CCSD result, while B3LYP gave the well-known overestimate. Looking at geometries, all methods generally correctly predicted an increase in TS asynchronicity with greater degree of activation. Asynchronicity was defined as the value of the longer incipient bond distance d_β minus the shorter distance d_α, and for the benchmark CCSD results this value ranged from ~0.5-1 Å across the systems (it is primarily the longer incipient bond distance d_β which varies across the systems, while d_α was consistently ~2 Å). All the DFT functionals tested generally correctly predict the value for d_α, so where there was a divergence from the CCSD results, this tended to be in the prediction of the longer bond d_β. Including some amount of HF exchange was found to be important at low activations since 'pure' functionals considerably overestimate asynchronicities for these systems (for more activated reactants, correlation becomes more important and pure methods become more reliable). B97D was somewhat of an exception to this as it fared better for the less activated systems (but still tends to overestimate asynchronicity by ~>0.15 Å). Of the hybrid methods, other than B3LYP, most perform well across reactants and it is difficult to claim a best method since no method performed best for all reactions. However, M06-2X performs very well for the less activated systems (with asynchronicities from CCSD of ~0.5 Å and 0.6 Å) and well overall, with a low mean deviation in incipient distance of 0.034 Å compared to the CCSD geometries (the second best functional was ωB97X with average deviation of 0.046 Å). A general conclusion made by the authors was that, for predicting the correct DA TS geometry, both a balanced combination of HF exchange along with a well-behaved correlation functional is required. For the present work, M06-2X was chosen for geometry optimisation (as well as energy calculation) since it was the most widely applicable DFT functional, while also having the edge for less activated systems (as noted in (76), this may be due to the 45% of HF exchange included in its exchange functional).

3.2 Methods

3.2.1 Gas phase IRC calculations

IRC calculations for the reaction products **9A** and **9P** were performed at the M06-2X/6-31G(d,p) and B97D/6-31G(d,p) levels as well as the semi-empirical levels DFTB2, AM1 and PM6. For the DFTB2 calculations, the Gaussian implementation using tabulated (rather than analytical) matrix elements was used (with Slater-Koster files from the mio-1.1 parameter set (46)).

For M06-2X, the IRC calculations were started from the transition states optimised in the previous study (25). The reaction path was followed in both directions, using the default step size of 0.1 Bohr, for as many steps as possible. The end points of the path were then optimised to reach the substrate and product minima and obtain the complete reaction path for determining activation and reaction energies.

To start IRC calculations for **9A** and **9P** at the B97D and semi-empirical levels, the corresponding transition states were first obtained. The QST3 method was used for TS optimisation which requires a ‘product like’ and ‘reactant like’ structure as well as a starting guess for the transition state. For the ‘product like’ structure, the products optimised at the M06-2X level were used. For the ‘reactant like’ structure the reactant end point of the relevant M06-2X IRC path was used. For the TS guess the corresponding TS optimised at the M06-2X level was used. Transition states were confirmed via frequency calculations as having 1 imaginary frequency. The reaction paths were then followed in both directions using each method. For B97D, the default step size of 0.1 Bohr was used, and the path was followed for 30 steps in each direction. For the semi-empirical methods, a step size of 0.05 Bohr was used and the paths were followed for as many steps as possible in each direction. For B97D, DFTB and PM6, the end points of the path were then optimised to reach the substrate and product minima and obtain the complete reaction path. This was not done for PM6 since, for both products, an unintended proton transfer (from the hydroxyl oxygen on the tetronate ring to the first carbonyl oxygen in the chain) occurred on the path towards the substrate therefore leading to an irrelevant species being formed.

3.2.2 Diene dihedral scan

An optimised scan of the reactive diene dihedral for native substrate **9** was performed to obtain the relative energies of the s-*cis* and s-*trans* conformations and the energy barrier for conversion between them. This was done with the substrate in the folded up reactive conformation set up to produce **9P** and with the sidechain of the active site Tryptophan positioned relative to the substrate as it would be in the proposed reactive binding mode. This was done so that, as well as indicating the intrinsic *cis*/*trans* energetics, the calculation could give some idea of the barrier for entering the reactive conformation should the substrate bind the enzyme with the diene in the s-*trans* conformation. To get the starting point for the scan, the energy minimum was first obtained for the s-*cis* conformation. To do this, the 'substrate like' structure which was used for the M06-2X QST3 transition state optimisation for **9P** combined with the Tryptophan sidechain (see section 4.2.1) was optimised at the B97D/6-31G(d,p) level. An optimised scan of the full range of the dihedral was then performed, scanning in 10 degree increments. Energies of the structures along the path were then also recalculated at the M06-2X/6-31G(d,p) level.

3.2.3 Gibbs free energy calculations of folded reactive and linear conformations of substrate

To calculate approximate Gibbs Free Energies of the reactive and extended conformations, energy and frequency calculations were performed on DFT optimised structures for the methoxy analogue substrate **10** (this was done at the B97D/6-31G(d,p) level). To obtain an input structure for the folded reactive conformation for optimisation, the structure for **10P** was used as a starting point. This is the structure displayed in Figure 3.6A (atropisomer shown in green) which was built and minimised in Chem3D using the MM2 forcefield (77). To generate the reactive conformation from this, the relevant bonds were broken/made and the structure briefly minimised within Chem3D to obtain a reasonable starting structure for QM optimisation. To enable a stable minimum to be found with the substrate in a reactive conformation, the two reactive C-C distances were first frozen to their starting values and the rest of the structure optimised around this. The restraints were then removed, and a free optimisation was performed, followed by the frequency calculation. For the extended conformation, a representative conformation was first built and optimised in chem3D (structure was optimised with the dienes in the s-*trans* conformation). The QM optimisation and frequency calculations were then performed for the resulting conformation. The reported Gibbs free energies for the two conformations were those taken directly from the Gaussian thermochemical output, where the vibrational entropy contribution is determined via the rigid-rotor-harmonic-oscillator approximation. While frequency calculations were only performed in gas phase, the energies of the optimised structures were also recalculated with implicit PCM solvent (78) to get a more realistic estimate of the relevant enthalpy difference between the two conformations.

3.3 Results and discussion

3.3.1 IRC reaction paths for the Diels-Alder reaction of the AbyU substrate

IRC calculations were performed at the M06-2X/6-31G(d,p) level to obtain the full reaction path for the DA reaction of the AbyU substrate (to form the expected enzymatic product). This reaction was already known to be asynchronous: previous TS optimization at the M06-2X/6-31G(d,p) level indicated that bond formation was more advanced for the C13-C14 bond. The newly optimised entire path at this level (Figure 3.9) shows that bond formation is also fairly staggered and almost stepwise in appearance since the C13-C14 bond is effectively completely formed in the initial part of the downhill path toward the product, while the second bond is mostly formed in the remainder of the downhill path (resulting in two distinct regimes on the path with a strong inflection point between them). Nevertheless, the process is still a concerted DA reaction since both bonds form on the downhill path from a single TS, with the shape of the energy profile reflecting the distinct stages of bond formation (Figure 3.10). B97D predicts an overly asynchronous TS by about 0.1 Å (asynchronicity=~0.8 Å compared to ~0.7 Å for M06-2X) with the error mostly being in the larger incipient distance. This is in line with the prior benchmarking study (76) (see section 3.1.4) where, for the most comparable system (in terms of asynchronicity), a similar overestimate was found, potentially reinforcing this as a general issue with B97D for systems of this type. The overall reaction path is qualitatively similar for both methods, with an offset between the paths in the d(C10-C15) direction corresponding to the different predicted asyncronicities. The predicted activation energy with M06-2X is around 16 kcal/mole and differs by about 1 kcal/mole for the two atropisomers (this may just be because of the reaction co-ordinate values at which the reactant minima are found, which may be due to the particular starting structures used and not reflect a difference in reactivity in practice). B97D underestimates the activation energy by around 7 kcal/mole compared to M06-2X, again similar to that found in the prior benchmarking study (~5 kcal/mole). Taking the M06-2X result to be the most accurate, this can be compared to the activation enthalpy for the DA reaction of butadiene and ethylene which has been found to be about 24 kcal/mole both experimentally and computationally via CCSD (79) (while M06-2X has been shown to underestimate CCSD by 2 kcal/mole (80)). These results therefore show that, in terms of potential energy, cyclisation should be significantly easier for **9** once in the reactive conformation. Since the DA reaction between butadiene and ethylene is essentially a baseline reaction where no electron withdrawing/donating groups are activating the reacting groups, some activation of these groups is therefore presumably occurring to enhance the reactivity of **9** (as also indicated by the asynchronous TS). In this case, the methoxy group attached to the tetronate ring may be acting as the electron-withdrawing group to activate the dienophile, while conjugation within the triene system may also be providing increased electron density around the reacting diene.

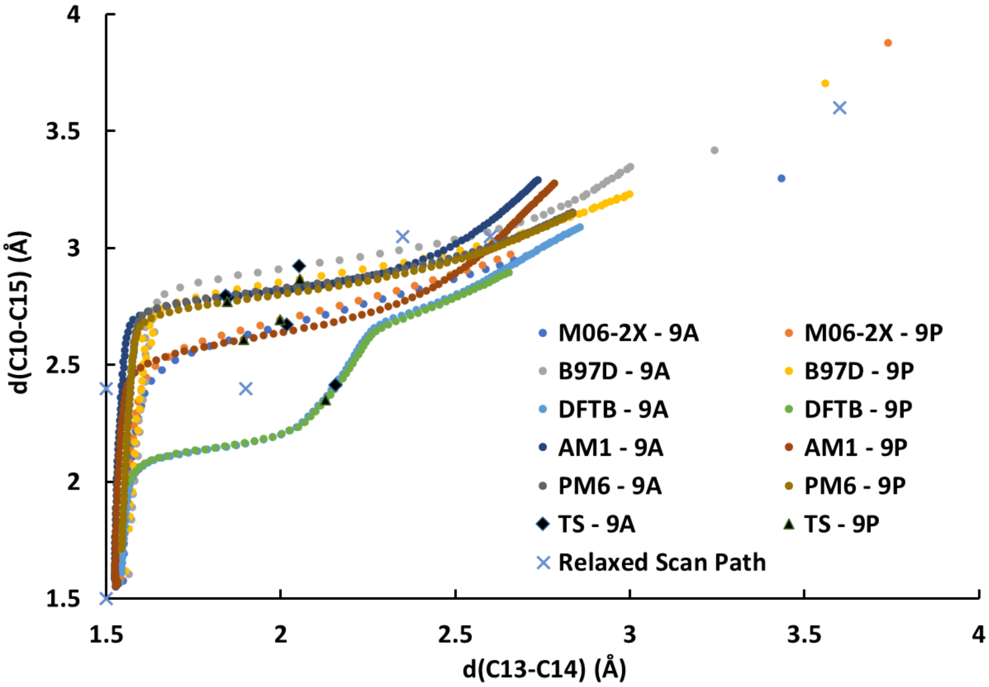


Figure 3.9: IRC reaction paths and optimised end points for the reactions to form **9A** and **9P** at the different QM levels. Reactant minima are only shown for the DFT methods as for AM1 this was not obtained, and because the other semi-empirical methods do not correctly predict a folded substrate minimum. Key points along the approximate scan path from the previous study (25) are indicated.

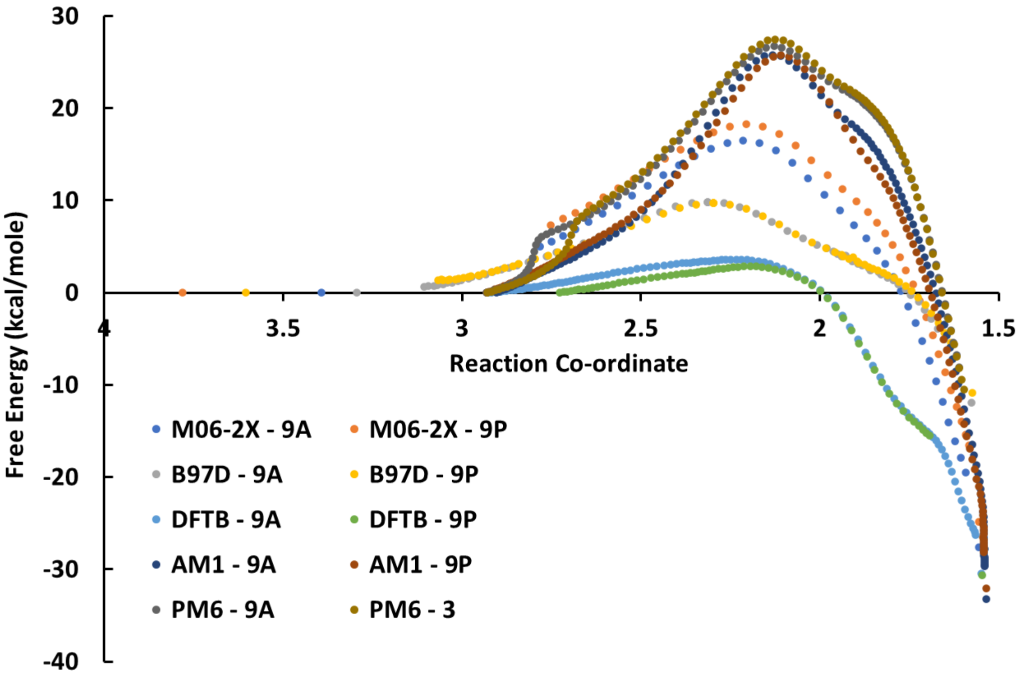


Figure 3.10: Potential energy profiles for the reactions forming **9A** and **9P** from IRC/optimisation of IRC endpoints at the different QM levels. Reactant minima are only shown for the DFT methods as for AM1 this was not performed, and the other semi-empirical methods do not correctly predict a proper reactive substrate minimum. Profile is projected along the optimised reaction co-ordinate from Chapter 4 and normalised to the reactant minima for the DFT methods and the end of the IRC path toward the substrate for the semi-empirical methods.

The IRC path found for M06-2X is fairly similar to the approximate path followed in the relaxed scans in the previous study (Figure 3.9 and see Figure 3.11). Therefore, despite the geometries being optimised with B97D, the energy profile/activation energies found previously for M06-2X (Figure 3.11) are quite similar to those now found for the improved IRC path. Similarly, since the optimal reaction paths shown by IRC are not too dissimilar between the two methods, the B97D energies from the approximate scans are also approximately in line with the IRC results. However, as there is still an appreciable difference between the IRC paths and the path followed in the relaxed scans this seems to indicate that the energies are not too sensitive to the exact geometry, thereby indicating a reasonably flat reaction energy surface. Flat energy surfaces appear to be a general feature of DA reactions (81, 82), which may thus also be the case for this reaction.

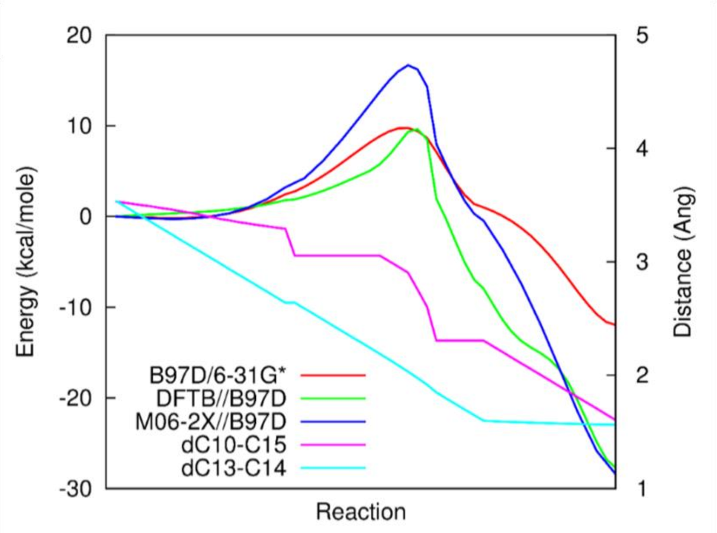


Figure 3.11 (From (25)): Approximate potential energy profiles for the reaction of **9**, from B97D/6-31G* relaxed scans starting from the product, where either dC10-C15 or dC13-C14 is increased (corresponding distances along the scan path are superimposed). Single-point energy calculations are shown for M06-2X/6-31G* and (SCC-)DFTB.

While the energy profiles for the DFT methods are similar to the previous study, the results for the semi-empirical method SCC-DFTB are more significantly different. For the B97D optimised scan path, the activation energy found with DFTB single point energies was equal to that for B97D (~10 kcal/mole). However, along its optimal path from IRC, the DFTB energy profile is much shallower, resulting in <3 kcal/mole barrier when calculated from the end of the IRC path (as a reactive folded minimum was not found by optimising IRC end points this cannot be used for reference). Although this path is qualitatively quite different to the DFT paths, it is not so different in terms of the actual location of the TS. This shows that the energy is more sensitive to the exact path, therefore suggesting a more curved energy surface for SCC-DFTB. This then explains why the energies were significantly higher along the scan path as this would then represent a more extreme diversion from the optimal path in energetic terms. These results therefore show that, when using DFTB for geometry calculation (e.g. when simulating the reaction in later QM/MM), the inherent activation energy will be considerably more underestimated than previously thought. This is not an issue in itself as this may be correctable, and if this is the case then DFTB could even just be used to discern relative activation energies (i.e. for different binding modes/substrates), at least qualitatively. In any case, it is difficult to assign an appropriate correction factor; the true DFTB activation energy from the reactant minimum in gas phase could not be used for this since a proper reactive folded minimum is not predicted (and so using activation energy from this point (~10 kcal/mole) to determine the underestimate will not reflect what the underestimate would be from a proper folded reactive conformation as will be seen in the enzyme, where this is enforced by surrounding residues). The difference between M06-2X and DFTB barriers taken from a common point in the IRC paths (e.g. reaction co-ordinate value of 2.7) indicates that the underestimate would be 8.7 kcal/mole (for atropisomer **9A**). Again, despite this difference, relative energies (for different substrates or enzyme-substrate conformations) may still be correctly predicted, providing the underestimate is consistent.

In terms of the reaction path, PM6 and AM1 provide a more qualitatively accurate description compared to DFTB (Figure 3.9). In terms of energies, the activation energy is strongly overestimated by these methods (Figure 3.10). Similarly to DFTB, AM1 does not correctly predict a properly folded reactant minimum and so is no better in this regard (end points were not optimised for PM6 due to proton transfer occurring at end of IRC – this coincides with a steep drop in energy in the IRC profiles). Further, the fact that AM1 finds different paths for the two atropisomers is a concern, as the DFT results suggest that path should be similar for both. A final point in favour of DFTB vs. AM1/PM6 is that it performs better in terms of predicting the overall TS geometry than the other two methods. The all-atom RMSD of the DFTB TSs (compared to M06-2X geometries) was 0.506 Å for **9P** and 0.468 Å for the **9A**. For AM1/PM6 this was 0.792/0.559 Å respectively for **9P** and 0.814/0.749 Å respectively for **9A**. On balance, DFTB seems like the best choice of semi-empirical method and was therefore used in Chapter 4 for enzymatic reaction modelling using umbrella sampling.

3.3.2 Conformations of the diene in the AbyU substrate

The potential energy profile for the diene dihedral of **9** is shown in Figure 3.12. As can be seen a converged profile was obtained and the final structure from the complete scan was identical to the starting structure. Both the s-*cis* and s-*trans* stationary points are approximately planar (see also Figure 3.13), however the s-*cis* conformation is slightly less planar than s-*trans* (C10-C11-C12-C13 dihedral angles are 2.2° and 179.6° for s-*cis* and s-*trans* respectively) due to the greater steric hinderance between C10 and C13 in the s-*cis* conformation which penalises a truly planar conformation. This is different to butadiene where, instead of just a single s-*cis* minimum, there is a planar s-*cis* stationary point which is a TS between two symmetrical off-planar minima (66), which the system will thus reside in when in the s-*cis* conformation. The s-*cis* minima for butadiene are significantly less planar than the single s-*cis* minimum for **9** and thus more strongly relieve the associated steric hinderance. The reason for this could be partly due to increased conjugation stabilisation in **9** from the overall triene system, compensating for the increased steric hinderance from being closer to planar. Further, due to the lack of symmetry in **9**, unlike for butadiene, the dihedral energy profile is asymmetrical. This is probably exacerbated by the fact that the molecule is in the folded reactive conformation rather than a more extended one, meaning that there are significantly different intramolecular interactions between 0-180° than between -180-0°.

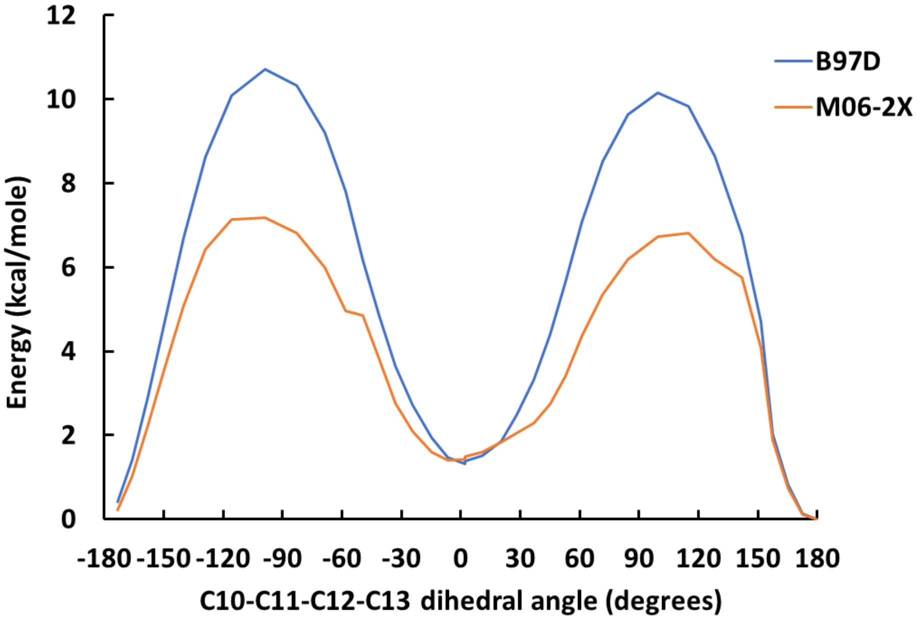


Figure 3.12: Potential energy profile for the reactive diene dihedral of **9** with M06-2X and B97D. Energies are obtained from a B97D/6-31G(d,p) optimised scan of the relevant dihedral, with single-point energies also calculated using M06-2X/6-31G(d,p). Energies are shown relative to the s-*trans* minimum.

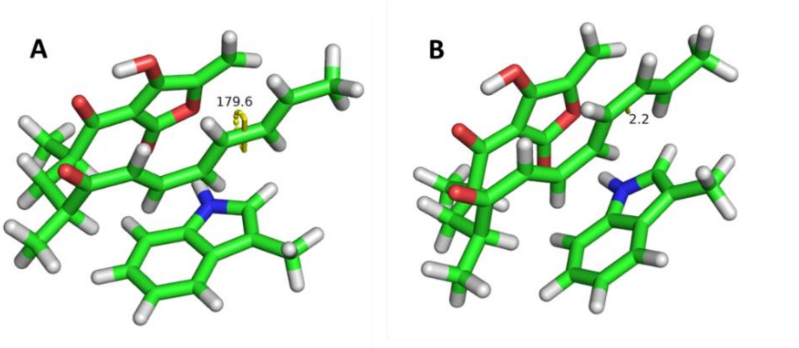


Figure 3.13: s-*cis* (A) and s-*trans* (B) B97D minimum energy conformations for **9** from the optimised scan of the diene dihedral.

For both **9** and butadiene, s-*trans* is the lower energy conformation, which is expected due to the greatly relieved steric hinderance. For butadiene with M06-2X, the s-*cis* minima are found to be 3 kcal/mole less stable than s-*trans* (66), while for **9** (with both B97D and M06-2X) this difference is only about 1.4 kcal/mole. This shows that the equilibrium will lie further towards the reactive s-*cis* conformation for **9** than is the case for butadiene; i.e. a greater proportion of substrate can be expected to have the diene in the reactive s-*cis* conformation than for butadiene and thus (from the diene point of view), be primed to react. Comparing again the results obtained with M06-2X, the lowest barrier between s-*trans* and s-*cis* is slightly higher for **9** than for butadiene (6.8 kcal/mole vs. ~6 kcal/mole), but still comparable, showing that equilibrium could be achieved on a similar timescale. Given the lower barrier found for the reaction step, these results show that the intrinsic rate of reaction for **9** would thus be expected to be at least as fast as for butadiene with ethylene. However, to complete the picture of reactivity, it must also be determined what proportion of the substrate is likely to be folded up in the reactive conformation.

3.3.3 Stability of the folded reactive conformation of the AbyU substrate

The difference in Gibbs Free Energy (which has been broken down into the enthalpic and entropic contributions) between the extended and folded conformations is shown in Figure 3.14 (the conformations from which this was calculated are shown in Figure 3.15). Looking at the gas phase results, enthalpically the folded conformation is considerably more favourable. This is presumably due to increased packing in the folded state leading to increased favourable Van der Waals interactions. The benefit of this will be overestimated by gas phase calculations, as in solvent environment there will also be favourable packing interactions in the extended state by surrounding water molecules. Including implicit solvent in the energy calculation therefore compensates for this overestimation (by approximating this favourable packing via a dispersion interaction term based on the solvent accessible surface area of the molecule) and reduces the enthalpic favourability of the folded form.

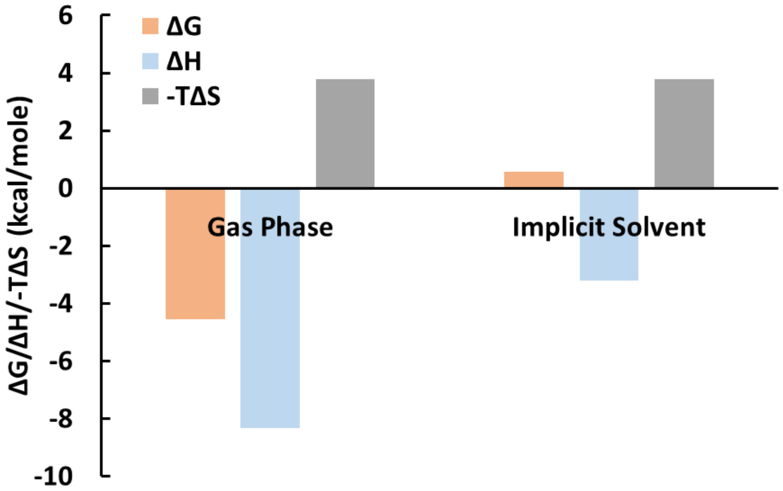


Figure 3.14: Thermodynamic stability of the folded reactive conformation of **10**. Difference in Gibbs free energy (ΔG) reported as the free energy for the folded state minus that for the extended form (thus a negative value would indicate a favourable folded form). Contributions to the Gibbs free energy difference ΔH and -TΔS are reported likewise. In gas phase G/H was calculated for the two optimised conformations via energy/frequency calculations at the B97D/6-31G(d,p) level. For implicit solvent ΔH was calculated via single point energy calculations on the gas phase optimised conformations. Frequency calculations were not performed in implicit solvent so the ΔG value shown is that found by combining ΔH with the value calculated for -TΔS in gas phase.

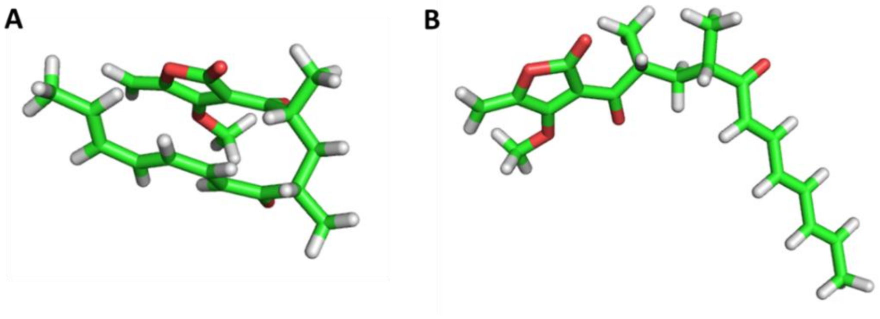


Figure 3.15: Folded (A) and extended conformations (B) from which Gibbs free energies/enthalpies were calculated to determine stability of the folded form. Conformations optimised at the B97D/6-31G(d,p) level.

Examining the entropic component (only calculated in gas phase) shows that there is a penalty involved in the substrate adopting the folded conformation which significantly reduces the overall stability (and when combined with the enthalpy difference calculated in implicit solvent gives a lower Gibbs Free Energy for the extended conformation). Cyclisation reactions are known to incur an entropic cost due to the restriction of internal rotations which were previously free to move in the acyclic substrate (83, 84). Torsional entropy loss due to the fixing of a single C-C bond has been estimated previously to be ~4.5 entropic units (eu) by comparing data for a series of analogous cyclisation reactions (84). In another study (83), using data for ~60 different cyclisation reactions it was found that, for n<7 (where n is the number of skeletal single bonds making up the cyclic product), entropy loss due to cyclisation was approximately independent of reaction specifics, and increases in proportion to n (where a proportionality constant of 4 was found - roughly consistent with the estimated ~4.5 eu torsional entropy for one C-C bond). This therefore indicates that a typical cyclisation reaction leads to near complete loss of entropy from all single bonds which are part of the cycle. In addition to the cyclic product, this concept of restricting rotations applies when forming the cyclic TS (or indeed a TS-like folded reactive conformation) and thus similar values can be observed for activation entropies (84). Therefore, given that acyclic substrate **10** is reasonably long (with 6 unrestricted single bonds linking the diene and dienophile), it is understandable that we see a significant decrease in favourability of the folded conformation due to entropy (which therefore greatly lowers the likelihood of the substrate adopting the folded conformation).

It should be noted however, that the calculation of the entropic component is fairly approximate. Firstly, the simple harmonic oscillator approximation is used to estimate vibrational entropies, which is itself an approximation. Additionally, as frequency calculations were only performed for a single minimum energy conformation, conformational entropy is thus unaccounted for, and the obtained entropy difference likely underestimates the true entropy loss for folding. Accounting for conformational entropy would require running frequency calculations for multiple minimum energy conformations. As noted in (85), this makes entropy estimation for flexible molecules challenging, due to the existence of many accessible low energy conformers. Finally, as frequency calculations were performed in gas phase, solvent is not taken into account here in the calculation of entropies. However, solvent rearrangement between the two states is likely to also involve additional entropic effects. Given the aforementioned issues, it may well be the reported entropic component is considerably underestimated, and entropy may serve to make the folded conformation significantly more unstable than these calculations indicate.

Overall, these results indicate that, while the folded conformation is enthalpically favourable, in a realistic environment this will not compensate for the likely entropic penalty. Given the reaction was found to be favourable in terms of the other aspects explored in this Chapter, these results therefore give insight into the relatively poor reactivity found in experiment. The reactive conformation for the Pylr4 substrate (which undergoes a similar spiro-linking DA reaction to form spirotetramate pyrroindomycin A – see Figure 1.4), was similarly found to be unfavourable when compared to a relaxed extended conformation. In this case, the reactive conformation was found to be 6.8 kcal/mole higher in free energy (as determined in a CPCM modelled ether environment using M06-2X and applying a quasi-harmonic correction), which may explain why the reaction does not occur spontaneously. This is presumably also due in large part to entropy, given that a similar number of bonds are restricted in the formation of the macrocycle. It has been suggested that both Pyrl4 and

AbyU may be functioning as 'entropy traps' (21) by binding and thus stabilising the entropically unfavourable reactive conformations of their respective substrates, and the evidence described here indicates that this is likely the case. Although, in the case of Pyrl4, other catalytic effects have also been found (see section 1.2.2.2.2), whether these may exist for AbyU (or whether stabilisation of the reactive conformation is the primary function of the enzyme) is explored in the next Chapter.

3.4 Conclusions

The work in this Chapter has further developed the knowledge of the intrinsic DA reaction of the AbyU substrate to form the enzymatic product, laying the groundwork for understanding the role of the enzyme itself. The activation energy for the reaction has been determined and the full reaction path elucidated. This has demonstrated a reasonably low barrier which is therefore likely not the limiting factor in the intrinsic activity of the substrate (but whether the enzyme also acts to further enhance this remains to be seen). Likewise, the reactive conformation of the diene is not found to be especially unfavourable. Instead, the modest enzyme-free reactivity found experimentally likely comes from the instability of the folded reactive conformation, where entropy significantly lowers the likelihood of the substrate adopting this conformation. The entropic consideration is also considerably oversimplified when considering the addition of solvent and conformational flexibility. The role of entropy in the enhancement of the uncatalyzed reaction rate may thus yet be an even greater factor than these calculations suggest. This provides compelling evidence for the enzyme providing a significant rate enhancement by trapping the substrate in this conformation, where the reaction may then occur with a low barrier thanks to the intrinsic reactivity of the substrate. Whether or not the enzyme also acts to further enhance reactivity by lowering the barrier to the reaction once substrate is productively bound, is explored in the next chapter. Finally, semi-empirical benchmarking has demonstrated that no semi-empirical method performs particularly well in capturing the energy in terms of the kinetics of the reaction. DFTB2 in particular, which was the method chosen for later reaction simulations in the enzyme, is found to strongly underestimate the barrier. However, this method may still be able to correctly discern relative reaction barriers, and thus be sufficient for determining the relevant reactive binding mode of the enzyme and performing reactivity screening of alternative substrates and mutants.

Chapter 4. Modelling the enzymatic reaction

4.1 Introduction

Having studied aspects of the intrinsic reaction (Chapter 3), the study presented in this chapter aims to develop mechanistic understanding of the enzyme through computational simulation. In addition to providing valuable insight for the design of Diels-Alderases in general, this will inform the investigation of the reaction of alternative substrates and mutants presented in Chapter 6. In particular, the modelling strategy developed here for the native reaction will then also be applied in that chapter. The previous computational study on this enzyme (25) provides the starting point for this work. In this study, four distinct candidate binding modes were identified for the reaction via molecular docking of the substrate and reaction products. Comparing the consistency between substrate and product docking provided an initial indication of the likely mode in which the substrate is turned over (Figure 4.1). Due to the nature of the substrate, binding is largely driven by steric complementarity between the enzyme and substrate. However, in addition to this a specific hydrogen bonding interaction is formed with the Tyr76 in this mode, which may therefore contribute to substrate specificity. To indicate the reactivity in each mode, activation free energy barriers were obtained by performing QM/MM umbrella sampling runs starting from the product state, using the semi-empirical QM method SCC-DFTB. This gave more evidence that the identified mode was the relevant one for the reaction as, out of the two poses for which sufficient umbrella sampling runs resulted in valid reaction paths, it gave the lowest average reaction barrier and clear substrate minima in each of the reaction profiles obtained. In addition, after accounting for the approximate correction factor due to the semi-empirical QM method, the resulting reaction barrier would suggest good reactivity for this mode. However, since for the two remaining modes sufficient results could not be obtained, some uncertainty remains whether the identified mode is really the mode through which the reaction occurs, or indeed if there are multiple such modes. Since recent experimental evidence suggests the possibility of multiple productive binding modes, the present study now more thoroughly investigates the different binding modes, after demonstrating an improved protocol for reactivity prediction (see section 4.2.2.3). The source of the different reactivities is also investigated, which provides further insight into the role of the enzyme.

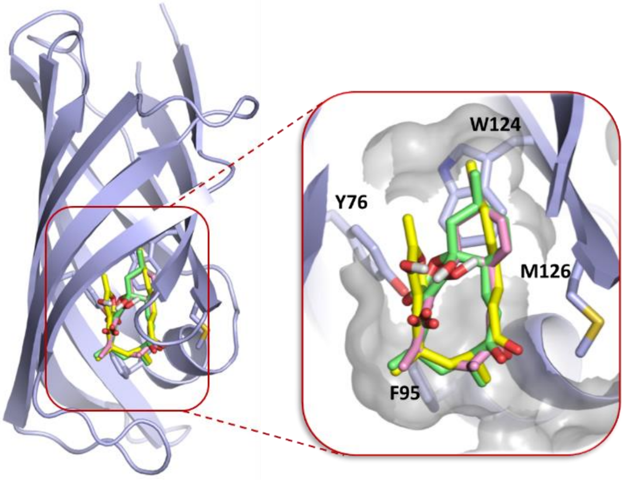


Figure 4.1: Crystal structure of AbyU with natural substrate **9** (yellow) and possible reaction products (**9A** in pink and **9P** in green) docked in the reactive binding mode suggested by previous study. Key residues which line the active site and were treated flexibly in the docking are highlighted.

In terms of mechanism, the reaction simulations in the previous study showed that AbyU is indeed a true Diels-Alderase as the reaction was found to proceed through a concerted (but asynchronous) path in the enzyme. The main function of the enzyme was thought to be to provide a preorganised binding site to capture the substrate in a reactive conformation. This was suggested by the similarity between the transition state (TS) in gas phase (found with M06-2X) and the approximate TS found in the enzyme (as found by averaging structures from the transition state window of the umbrella sampling). It could therefore be speculated that, once the substrate binds, the reaction then proceeds with an energy barrier which is dictated largely by the intrinsic reactivity of the substrate. This now seems more plausible as in Chapter 3 it was shown that the substrate is intrinsically already reasonably reactive. Furthermore, the stability found for the reactive folded conformation of the substrate in solution was such that capture of the substrate alone could potentially be responsible for the rate enhancement. However, some additional rate enhancement may also come from the enzyme further promoting reactivity once substrate is bound. To investigate this the free energy barrier for the enzymatic reaction was compared to that found in solvent, since solvent may itself be providing some catalysis of the DA step.

To study both the enzymatic and solvent reactions, semi-empirical QM/MM umbrella sampling has been applied as this is an effective way to obtain activation free energies while incorporating the influence of environment (48). First of all, how the enzyme is represented was considered to ensure that the main catalytic effects are captured and will be reflected in the obtained activation free energy. If the enzyme is serving to simply bind a reactive substrate conformation then MM should be sufficient as it is likely just the overall steric and electrostatic interactions, rather than detailed electronic interactions, which are important for this. A discussion of other roles the enzyme may be playing which, although not explicitly tested for here, may be adequately captured with an MM representation is in section below titled ‘MM environment modelling for DA reactions’. In terms of specific residues which may have a catalytic effect that requires explicit electron treatment of the sidechain to capture, the active site tryptophan (Trp124) came to mind. Considering that, in the presumed reactive binding mode, its aromatic side-chain is nearby the reacting diene of the substrate which has a conjugated π system (Figure 4.1), the specific electronic distribution of the Trp124 could be important (for this reason, this side-chain was included in the QM region in the previous study (while the rest of the enzyme was treated with MM)). To test whether QM modelling of this sidechain is in fact necessary, calculations were performed in gas phase at the DFT level to show if the Tryptophan, on its own, provides any lowering of the reaction barrier. If catalysis were to be found, QM treatment may be important for this residue (although if catalysis stems from dispersion/Van der Waals interactions this could, in principle, be captured via MM).

Having investigated the role of Trp124 and validated the choice of QM region, QM/MM umbrella sampling was then used to obtain free energy profiles and activation free energies for the different binding modes (to confirm the reactive mode and determine the likely reactivity in the other modes). An efficient protocol has been used for this since, by first demonstrating its applicability for the native reaction, it can then be used as part of a high throughput workflow for activity screening in Chapter 6. In addition to reaction barriers, the Michaelis complex has now been studied for the different binding modes via MD simulation (again using an efficient protocol). From this, binding affinity is calculated to indicate which pose is preferred in binding to further confirm the relevant mode as well as the contribution of the other binding modes. In addition, the distance between the reacting groups in the Michaelis complex has been studied to determine if closer (average) positioning of the two groups may explain the estimated higher activity for certain modes. Finally, to determine if there are any other interactions made by the enzyme which serve to catalyse the reaction step (i.e. more so than those provided by solvent) the umbrella sampling protocol developed has been used to obtain the reaction free energy barrier in solution for comparison.

4.1.1 MM environment modelling for DA reactions

As discussed in Chapter 3, activation energy for the Diels-Alder reaction can be lowered by attaching an Electron Donating Group to the diene and/or an Electron Withdrawing Group to the dienophile as this lowers the HOMO-LUMO energy gap. Likewise, Electron donating/withdrawing interactions could be being made by the environment which polarise the substrate and thereby achieve the same thing. This is suggested to be behind the catalysis in IdmH (26) where, using essentially the same protocol to the previous study with AbyU, it was found that in the binding mode with the lowest free energy barrier the enzyme makes an interaction which may withdraw electron density from the dienophile. This interaction comes in the form of a hydrogen bond between either a threonine or serine residue in the active site and the carbonyl group of the substrate which is conjugated to the dienophile. It is possible that Tyr76 in the active site of AbyU may be performing a similar role since, in what is currently thought to be the most likely binding mode, it forms a hydrogen bond with the carbonyl of the substrate which is also conjugated to the dienophile through the aromatic tetronate ring. This role can of course also be performed by solvent as well as the enzyme and so must be captured for both environments to determine their relative catalytic effects. Either way, since using an electrostatic embedding scheme the MM region can polarise the QM region (see section 2.1.3), this type of catalysis may therefore be adequately captured by representing the environment with MM. Indeed, in the IdmH study the QM region consists of just the ligand so this may be demonstrating the ability of a hybrid QM/MM approach to capture this type of effect. More generally, catalysis can be a result of the environment interacting more favourably with the Transition state than with the Reactant State and thereby lowering the reaction barrier compared to the intrinsic reaction. A computational study (86) found that this type of textbook catalysis may be occurring in the promiscuous lipase CALB, which was shown to enhance reactivity for a particular intermolecular Diels-Alder reaction (as exhibited by QM/MM simulations which showed a lower free energy barrier for the reaction in enzyme than in solution). Since a build-up of negative charge occurs on the carbonyl of the dienophile during the reaction, a pocket formed by hydrogen bonding residues (which stabilise the oxyanion in CALB's original activity) is in a position to stabilise that charge and thus lower the reaction barrier. During the simulations, the residues making up the oxyanion hole were shown to be well set up to do so and a by-residue decomposition of the QM/MM interaction energy (considering that only the substrate was in the QM region) demonstrated that they did indeed interact more strongly with the transition state than Reactant State. The reason that this electrostatic charge stabilisation is well captured with an MM treatment is essentially because the point charges of the MM region can interact favourably with the electron density of the QM region. Although, in this case, the charge stabilisation was not thought to be what gave the enhanced reactivity compared to the reaction in solution (due to the fact that solvent molecules were equally well oriented to stabilise the negative charge during the simulations) this nevertheless demonstrates the utility of MM for modelling important catalytic interactions.

4.2 Methods

4.2.1 QM modelling of reaction with Trp124 sidechain

To test whether inclusion of the Trp124 in the QM region may be important to capture enzyme catalysis, a simple model was constructed consisting of just the substrate and sidechain of the Tryptophan residue (positioned approximately where it would be for the suspected reactive pose). Then, the reaction profile was obtained with the Tryptophan sidechain present at the M06-2X/6-31G(d,p) level and compared to the previous profile obtained for the reaction in gas phase. Using this simple reduced model, compared to including the rest of the enzyme for example, provides a clear test of the catalytic importance of the Tryptophan as it shows whether, under ideal circumstances (i.e. optimising for just the Tryptophan/substrate energies), it is at all capable of catalysing the reaction, and therefore gives more confidence in ruling it out if not. In addition, the small number of atoms involved makes using high level QM practical.

To obtain the reaction profile with the Tryptophan sidechain at the M06-2X/6-31G(d,p) level, a transition state optimisation was performed followed by an IRC calculation. This was attempted for both atropisomer products, however, results are reported for **9A** only since a transition state was not able to be obtained for **9P**. No restraints were applied during the transition state optimisation and subsequent IRC/optimisations, and thus all atoms were allowed to move freely. To obtain the transition state, the QST3 method was used which requires a starting guess for the transition state geometry as well as structures on either side of the transition state (i.e. reactant and product sides). To obtain the starting guess for the transition state the co-ordinates for the gas phase optimised transition state were firstly aligned to the approximate transition state in the enzyme from adiabatic mapping (see section 4.2.4). The co-ordinates of the atoms making up the Tryptophan sidechain in the in-enzyme structure (with a hydrogen added on the C_β using Gaussview to replace the missing C_α atom) were then combined with the aligned gas phase transition state and this structure used as input to the QST3 method. The same procedure was applied to obtain the input co-ordinates for the product and reactant structures, where the starting ligand co-ordinates used were those input in the previous gas phase transition state optimisation. A frequency calculation was then performed to check if the optimised structure was a true transition state with 1 imaginary frequency. However, as well as the relevant imaginary frequency corresponding to the bond formation, a second much smaller imaginary frequency was found. The motion corresponding to this second frequency involves mostly rotation of the indole group. In order to obtain a true transition state, it is possible that this frequency could be eliminated by perturbing the structure along the vector corresponding to the frequency and re-optimising. However, since the second frequency was much smaller and likely negligible, it was decided to proceed with the existing optimised structure. The IRC calculation was then performed starting from the transition state and following the reaction path in both directions for 20 steps using the default step size of 0.1 Bohr. However, for the path towards the substrate only 18 steps were completed at which point the calculation failed to converge. The end points of the path were then optimised to reach the substrate and product minima and obtain the complete reaction path for determining activation and reaction energies. The energy profile is then reported along the optimised reaction co-ordinate used for umbrella sampling.

4.2.2 QM/MM Umbrella sampling of the enzymatic reaction

Umbrella sampling of the enzymatic reaction was performed for each of the 4 binding modes identified in the previous study, starting from the poses of the products obtained via docking. Rather than the natural products **9A** and **9P** which were tested in the previous study (25), the methoxy analogues **10A** and **10P** (Figure 4.2) were instead used here, where the analogous poses were therefore tested.

4.2.2.1 Docking

Starting conformations (in PDB format) of the products were generated in Chem3D, optimised using the built-in MM2 minimiser. The coordinates for the enzyme were taken from the crystal structure of AbyU, removing the buffer molecule HEPES bound in the active site (PDB ID: 5DYV, chain A).

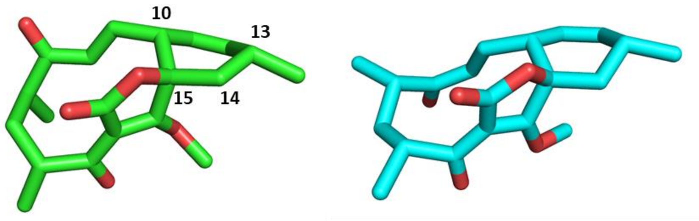


Figure 4.2: Structures of the two possible methoxy analogue reaction products **10A** (green) and **10P** (blue) input to the docking.

AutoDockTools 1.5.6 was used to create the input files for each protein/product docking by assigning AutoDock atom types to the relevant atoms as well as defining rotatable bonds. Polar hydrogens are required in the input structures of both the ligand and protein to determine Hydrogen bond donor/acceptors. Therefore, as the protein crystal structure does not contain hydrogens, polar hydrogens were firstly added in AutoDockTools (utilising open babel) where all were in their standard protonation states. During docking, the side-chains of residues Tyr76, Phe95, Trp124 and Met126 (which line the cavity) were treated flexibly; all formally single bonds were defined as rotatable. The standard active torsions detected by AutoDockTools were assigned (all formally single bonds not part of a cycle are treated as rotatable) for the products.

Docking was then performed with AutoDock Vina (87) using a search grid of size 21.4 x 16.5 x 17.6 Å centred on the active site, and an exhaustiveness of 16. 4 poses (corresponding to binding modes A-D in Figure 4.8) were selected for **10A** and 3 poses for **10P** (corresponding to modes A, C, D – a binding mode corresponding to B was not found) for further simulation to predict the reactivity and binding affinity of the different modes. These binding modes are analogous to those found from the previous study, except for the fact that this work is now using the methoxy analogue products since these were used in the kinetics experiments.

As the poses output by Vina consist of just the co-ordinates of the ligand and flexible sidechains, these were then combined with the rest of the protein to get the complete structure of the complex for each pose. The hydrogens which were added to the protein during the docking setup were left out when making these structures as the protein is then protonated during the simulation setup.

4.2.2.2 Generating starting structures for umbrella sampling

Repeat umbrella sampling runs were performed starting from MM MD snapshots of the product complex. To obtain the starting snapshots, short, efficient MD runs were performed for each pose using the Enlighten protocols (www.github.com/marcvanderkamp/enlighten) (35) PREP, STRUCT and DYNAM. These are an automated set of protocols for running simple and computationally inexpensive optimisation and MD for a protein-ligand system. The protocols employ fixed boundary conditions where the system is described with just a small sphere of solvent around the active site and everything outside this region is fixed during simulation. Since this results in a system with far fewer atoms than using a full periodic box setup, this means the simulations require much less computational time. This setup is carried through to the umbrella sampling simulations and following MD of the substrate complex, making these also efficient simulations. The basic assumption here is that only protein motion in the vicinity of the substrate is relevant in either the sampling of conformations of the bound product/substrate or along the reaction pathway (see also section 2.2.3).

The PREP protocol is used to prepare the structures of the product complex from docking for simulation. PREP takes the combined protein and docked product structure; it then adds hydrogens (to the protein only) via the AmberTools (34) programme reduce. This assigned the only histidine, His88, as doubly protonated. Asp43 was protonated manually since the predicted pK_a from PropKa 3.1 (88, 89) was 9.7. Then, in addition to the crystallographically determined water molecules, a 20 Å solvent sphere is added around the active site (with the AmberTools program tleap), missing heavy atoms (and their hydrogens) are added, and finally the topology and co-ordinate files are generated which describe the resulting system for simulation. The protein is parameterised by the ff14SB force field (90), and TIP3P is used as the water model. AM1-BCC partial charges and Generalised Amber Force Field (91) parameters were generated for each atropisomer using Antechamber.

Structural optimisation and Molecular Dynamics were then performed using the Enlighten protocols STRUCT and DYNAM, using sander from AmberTools16, with the entire structure treated with MM. The STRUCT protocol performs brief simulated annealing and minimisation to optimise the structure. DYNAM then takes the output from STRUCT and performs brief heating to 300K, followed by MD simulation. Throughout, all atoms outside of the 20 Å solvent sphere are kept fixed. Rather than taking starting snapshots from a single MD trajectory, multiple independent simulations were performed from which to start the repeat umbrella sampling simulations. This was done to enable as diverse as possible starting structures to be generated using a short total simulation time, thus making this procedure part of an efficient overall protocol for reactivity prediction. 10 repeat DYNAM runs were therefore performed to obtain 10 short (10ps) MD trajectories (2 fs timestep), the final snapshots of which were then used to start multiple umbrella sampling runs.

4.2.2.3 Umbrella sampling

The reaction involves the concerted formation of two new carbon-carbon bonds, one between C10 and C15 and the other between C13 and C14 (see Figure 4.2). As simulations are started from the product the reaction is sampled in reverse, forcing these two bonds to break. To do this, distances between the respective carbons must be gradually increased. A single reaction co-ordinate which can describe this is the distance between the centres of mass of the two bond forming pairs from the diene/dienophile respectively. This was chosen in the previous study to allow freedom in the bond distances and therefore find the natural (asynchronous) minimum energy reaction path. However, due to the nature of the underlying energy surface, this has been shown here to be unreliable in ensuring sufficient sampling of the correct minimum (free) energy path (see results section 4.3.2). Since a short sampling procedure has been used for efficiency, it is important to choose an effective reaction co-ordinate which enables good sampling of the reaction path, to minimise error in the predicted barrier (this will then allow accurate enough barriers to be obtained to reliably discern between reactivities in different binding modes). Another DA reaction co-ordinate which has been used before for umbrella sampling is a symmetric combination of the two bond distances (86). However, this was also found to lead to poor sampling and inaccurate barriers when compared to those found along the minimum energy path in full 2D free energy surfaces. Therefore, to ensure good sampling, a weighted combination of the bond distances (see below) has been used here as the 1D reaction co-ordinate, where different weights have been trialled to find the optimum.

$$RC = k_1 d(C10 - C15) + \ k_2 d(C13 - C14)$$

$$0 < \ k_1, k_2 < \ 1, \qquad k_1 + k_2 \ = \ 1$$

Since the weights determine the relationship that the two distances must satisfy for a given value of the reaction co-ordinate, along with the underlying energy surface they will determine the path taken in the reaction simulations. Therefore, to find the optimal reaction coordinate which leads to sampling along the correct minimum free energy path, umbrella sampling was performed for different weight combinations. This was done for **10P** in mode A. To determine the reaction coordinate that lead to the best sampling along the minimum free energy path, the underlying potential energy surface for the reaction was also obtained (Figure 4.4, Figure 4.6b; see computational details in the section “2D Potential Energy Surface” below). 10 repeat umbrella sampling runs were performed for each weight combination, starting from the MD snapshots of the product complex. Umbrella sampling was done at the SCC-DFTB/ff14SB level with the QM region consisting of just the ligand and using 26 windows from reaction coordinate values 1.3 Å to 3.8 Å (i.e. steps of 0.1 Å). A restraint of 200 kcal mol^{-1}Å^{-2} and 2 ps of simulation was used for each umbrella sampling window. The simulations were run with the Amber16 program sander using a 1 fs timestep (with remaining simulation parameters as for the DYNAM MD stage) and keeping all atoms outside the 20 Å solvent sphere fixed. Reaction coordinate values were recorded every 1 fs and used as input for the Weighted Histogram Analysis Method (WHAM) (57, 59) to obtain the potential of mean force (free energy profile) along the reaction coordinate. In addition to this, the bond distances themselves were recorded every fs to assess sampling quality. Comparing the distances sampled using the 1D co-ordinate to the 2D potential energy surface, the optimal weight combination was found to be k_1=0.3, k_2=0.7 as this ensured that the system sampled the correct (i.e. minimum energy) reaction path through the transition state (Figure 4.6b).

The umbrella sampling protocol described above was then run for the remaining product poses using the optimised reaction co-ordinate, and in each case similar sampling was obtained. For each pose, individual free energy profiles were obtained with WHAM for each of the repeat umbrella sampling runs. The barrier was then calculated for each profile as the difference between the transition state and substrate minimum and the average (and standard deviation) taken from the 10 repeats. Overall free energy profiles were also obtained by running WHAM with the sampling from all 10 runs combined.

4.2.3 Molecular dynamics of substrate complexes

Molecular dynamics was run for the reactive substrate complexes corresponding to the different binding modes of the products. This was done to generate an ensemble of structures of the Michaelis complex for each mode to determine approximate binding affinities using MM/GBSA (61). In addition, the average distance between the bond forming carbon atoms was calculated for each mode from the same ensemble of structures.

To run the MD, multiple starting structures were first obtained for each mode by taking them from the final window of the umbrella sampling runs for the corresponding product poses. For each starting structure, after a brief QM/MM optimisation, the Enlighten protocols PREP and STRUCT were run as before and DYNAM was then used to run 100ps of MD. As structures were saved every 1 ps, each of the 10 starting structures provided 100 snapshots and this therefore gave an overall ensemble of 1000 structures for each pose from which to calculate binding affinity and reactive distances. To calculate binding affinity the MM/GBSA single trajectory approach was used since this requires only structures of the complex. The average and standard deviation of the binding energy over all the snapshots are then reported for each pose. Cpptraj (92) was used to calculate the average distance between bond forming carbons for each snapshot from which overall and per run averages were then calculated. The standard deviation was then calculated over the averages for the separate runs to indicate error.

Representative structures (as shown in Figure 4.8) were generated for each pose by using cpptraj to cluster all 1000 structures based on the all-atom RMSD of the substrate, without fitting (using the default hierarchical agglomerative clustering algorithm) into a single cluster and taking the cluster centroid. (Prior alignment of the trajectory was not necessary since the simulation setup keeps the overall protein position fixed in space.)

4.2.4 2D potential energy surface

To obtain the potential energy surface for the enzymatic reaction at the SCC-DFTB/ff14SB level, a series of energy minimizations is performed, ensuring no jumps to different potential energy surfaces occur (here referred to as 'adiabatic mapping'). This was performed for product **10P** starting from binding mode A from docking.

To prepare the system, the relevant docked complex structure output after the reduce stage of the PREP protocol was taken and the remaining preparation steps were performed as before, with the exception that the system was instead solvated with a periodic box of TIP3P (distance/closeness - 11 Å / 0.75 respectively) and neutralised with 4 Na+ counterions (this is to circumvent issues with the minimization algorithms in Amber16 when using a solvation sphere only). Then, to obtain a reasonable starting structure for adiabatic mapping, the system was briefly equilibrated with the following procedure, treating the whole structure with MM (all steps performed with the Amber16 program pmemd.MPI, using periodic boundary conditions, a cut-off for direct-space non-bonded interactions of 8 Å and Particle-mesh Ewald summation for electrostatic interactions outside the cut-off):

1) optimize solvent and solute hydrogen positions (minimization for 300 steps with restraints on solute heavy atoms, force constant: 100 kcal mol^{-1}Å^{-2});
2) heating/equilibration of solvent (initial random velocities assigned at 50K followed by 50 ps NPT heating to 300K with a 2 fs timestep using SHAKE restraints, using Langevin dynamics with collision frequency 2 and the Berendsen barostat with a pressure relaxation time of 1 ps. All solute atoms were restrained with force constant 25 kcal mol^{-1}Å^{-2});
3) 300 steps of minimization and quick heating to 298 K (20 ps in NVT ensemble with 2 fs timestep, using Langevin dynamics with collision frequency 1, SHAKE and initial random velocities assigned at 25K);
4) 300 ps equilibration in the NPT ensemble (Langevin dynamics with 2 fs timestep, SHAKE and collision frequency of 1 and using the Berendsen barostat with a pressure relaxation time of 1 ps).

Following this, the product and the sidechain of Trp124 were treated with SCC-DFTB. The structure at the end of the 300 ps equilibration was first minimised for 500 steps using steepest descent/conjugate gradient and then further minimised with the LBFGS method with a convergence criterion of 0.002 mol^{-1} Å^{-1} for energy gradients (all calculations performed using sander in AmberTools16).

To get the full energy surface, an initial scan was first performed along a diagonal path, with both bond distances restrained to the same value. To do this, optimisations were performed (again using the LBFGS method with a convergence criterion of 0.002 kcal mol^{-1}Å^{-1}) harmonically restraining the C10-C15 and C13-C14 bond distances (with 2500 kcal mol^{-1}Å^{-2} potential) to increasingly higher values. The first structure was optimised with distances restrained to 1.3 Å and then the following structures were optimised in 0.1 Å steps, using the previous optimised structure as the starting structure, up to a distance of 3.8 Å. This path was scanned back and forth until energies converged to a smooth profile. During the optimisations, only residues that had an atom within 5 Å of the ligand were allowed to move. All other residues were restrained to the starting co-ordinates by a 50 kcal mol^{-1}Å^{-2} harmonic potential. This was to ensure no discontinuities occur between progressive energies due to structural changes unrelated to the reaction. Separate scans were then performed along the C10-C15 distance starting from each optimised structure from the previous diagonal scan, and keeping the C13-C14 distance restrained to the same value that it was previously. Again, structures were optimised in 0.1 Å steps, starting from the initial structure and either increasing the C10-C15 distance to 3.8 Å, or decreasing it to 1.3 Å, to cover the entire surface. In this case only a single scan was performed in either direction as this already resulted in a smooth energy surface (except for part of one edge, which is unimportant due to its high relative energies).

4.2.5 Solvent only umbrella sampling

The free energy barrier for **10P** in solution was obtained by umbrella sampling using the reaction co-ordinate optimised for the enzymatic reaction. However, to prevent unphysical conditions being imposed on the reacting species, a periodic box setup was used for both the umbrella sampling procedure and the procedure for generating starting structures.

The system was initially prepared by solvating the product structure previously built with a periodic box of TIP3P (distance/closeness – 20 Å / 0.75 respectively). The topology and co-ordinate files were then generated using the parameters already generated for the product. The system was equilibrated as follows, treating the whole structure with MM (all steps performed with the Amber16 program pmemd.MPI, using periodic boundary conditions, a cut-off for direct-space non-bonded interactions of 8 Å and Particle-mesh Ewald summation for electrostatic interactions outside the cut-off):

1) optimize solvent and solute hydrogen positions (minimization for 300 steps with restraints on solute heavy atoms, force constant: 100 kcal mol^{-1}Å^{-2});
2) heating/equilibration of solvent (initial random velocities assigned at 50K followed by 50 ps NPT heating to 300K with a 2 fs timestep using SHAKE restraints, using Langevin dynamics with collision frequency 2 and the Berendsen barostat with a pressure relaxation time of 1 ps. All solute atoms were restrained with force constant 25 kcal mol^{-1}Å^{-2});
3) 300 steps of minimization and quick heating to 298 K (20 ps in NVT ensemble with 2 fs timestep, using Langevin dynamics with collision frequency 1, SHAKE and initial random velocities assigned at 25K);
4) 600 ps equilibration in the NPT ensemble (Langevin dynamics with 2 fs timestep, SHAKE and collision frequency of 1 and using the Berendsen barostat with a pressure relaxation time of 1 ps).

10 snapshots were then taken at regular intervals from the final 200ps of the equilibration stage to start umbrella sampling simulations. The umbrella sampling procedure was identical to that for the enzymatic reaction with the exception that simulations are now run in the NPT ensemble (using the same simulation parameters as the equilibration stage).

4.3 Results and discussion

4.3.1 Establishing the catalytic role of Trp124

The reaction barrier for the substrate in the presence of the Trp124 sidechain was obtained via IRC calculation. This was done to both establish the potential catalytic influence of Trp124, and to determine whether it should be treated with QM in subsequent enzymatic reaction modelling. Comparing the resulting reaction profile with that found for the reaction found in gas phase (see Chapter 3) indicates that very little (0.9 kcal/mole) stabilisation is provided by the Tryptophan sidechain (Figure 4.3). Furthermore, the reaction path is not affected by the presence of the Tryptophan in any meaningful way, which further indicates that it does not participate strongly in the reaction step. Considering that these results also reflect the ideal situation in which the geometry has been optimised purely for the substrate and Tryptophan energies it therefore seems unlikely that any appreciable catalysis would be occurring in the whole enzyme where many other interactions will prevent this ideal geometry from being found. In addition, it may that this stabilisation comes from Van der Waals/dispersion interactions which could be adequately captured by an MM representation. Therefore, it appears that the Tryptophan may be excluded from the QM region without losing essential detail from the reaction modelling.

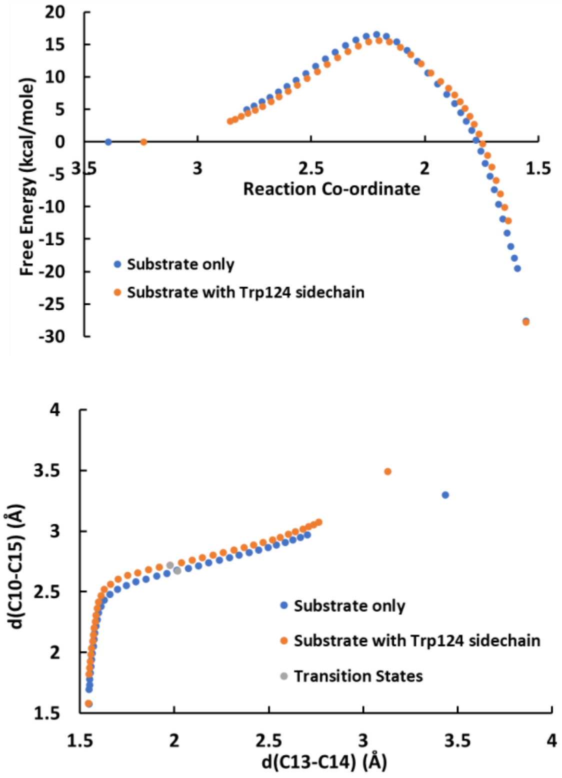


Figure 4.3: Reaction profiles (Top) and paths (Bottom) for the reaction in gas phase (with and without the sidechain of the Trp124 present). Profiles/paths obtained from IRC calculations and subsequent optimisations of IRC endpoints to obtain reactant/product minima. Calculations performed at the M06-2X/6-31G(d,p) level.

Profiles have been projected onto the optimised reaction co-ordinate used for umbrella sampling and energies are shown relative to the substrate minima.

4.3.2 Reaction co-ordinate optimisation

Using the distance between centres of mass as the reaction co-ordinate with the current umbrella sampling protocol, the system cannot be reliably steered through the minimum energy path and true transition state (Figure 4.4), generally resulting in a discontinuous path around the TS. This can be rationalised by considering the shape of the underlying energy surface and how this is explored by the individual reaction simulations. (Interestingly, the 2D Free Energy Surface found for CALB (86) exhibits a similar shape and shows a very similar discontinuity in the sampling with the 1D reaction co-ordinate). The energy surface for the reaction contains a large well in the product basin where just the C10-C15 bond distance is increasing. It can be seen that, although in all runs the system initially follows the minimum energy path, in certain runs, rather than passing through the transition state, the system moves up through this well before eventually crossing into the reactant basin at a point higher up the transition surface. It then re-joins the correct path but at a reaction co-ordinate far beyond that corresponding to the TS (therefore capturing only part of the downhill free energy change from transition state to Reactant State which makes up barrier).

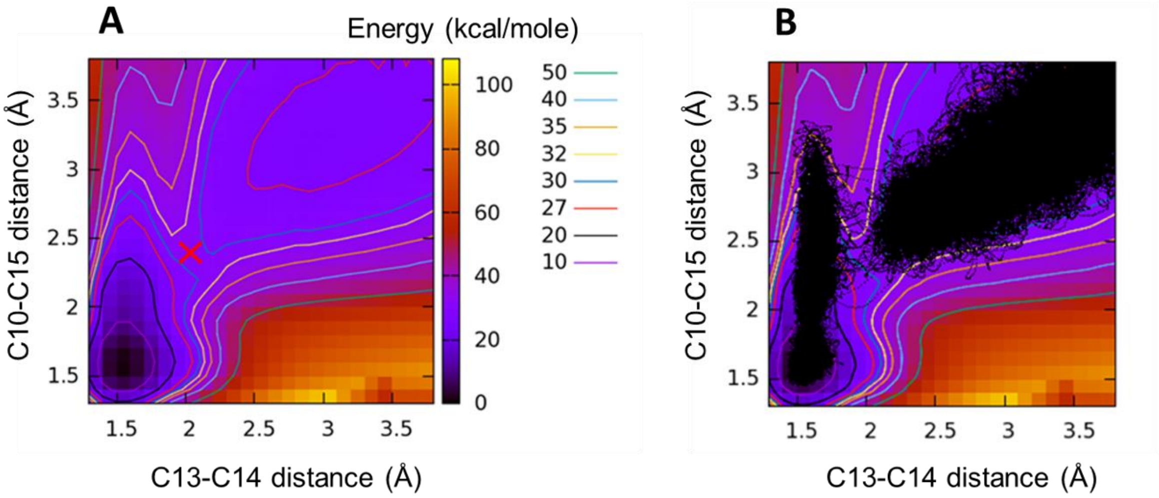


Figure 4.4: A) 2D potential energy surface for the reaction of **10P** in binding mode A (found at the SCC-DFTB/ff14SB level). Approximate transition state location is indicated by a cross. B) 2D potential energy surface with distances sampled during the 1D sampling superimposed (black dots). 1D sampling results shown for the 10 repeat umbrella sampling runs using the centre of mass distance reaction co-ordinate.

The reason this happens is, as the reaction co-ordinate is increased during umbrella sampling, since only the centres of mass need move apart, this can be satisfied by either moving up through the well or by taking the correct path through the transition state. Energetically speaking, when a value of the reaction co-ordinate is reached near that corresponding to the transition state, there is a fork in the energy surface where there are then two low energy paths available with increasing reaction co-ordinate; the correct one through the transition state, and another moving up into the well in the product basin. Even though the path through the transition state is ultimately the lower energy path to the reactant, as the alternative path is initially a lower energy route with increasing reaction co-ordinate, the system can get temporarily trapped here. This is also exacerbated by the umbrella sampling protocol; since the simulations are performed sequentially and the sampling in each window is short, this means that the system is not given enough time to escape the well before the reaction co-ordinate is increased and the system forced further along this path. By the time the system finally escapes into the reactant basin, which may not be until the energy surface starts to flatten, the reaction co-ordinate is then much more advanced and so the system re-joins the correct path at some point downhill from the TS.

The problem here is essentially that the centre of mass restraint allows too much freedom in how the bond distances can change. If a weighted average of the bond distances is instead used as reaction co-ordinate the system is constrained such that the bond distances are made to satisfy a linear relationship for any given value of the reaction co-ordinate. It can then be ensured, by adjusting the weights, that as the reaction co-ordinate approaches the value corresponding to the transition state, the desired path is the only low energy route available. Different weight combinations were therefore tested to find the optimum reaction co-ordinate which provides the best sampling of the reaction path. To do this the weights were systematically modified, and four different combinations tested which cover the range of possibilities (see Figure 4.5). The two best combinations, which give the most complete sampling of the correct reaction path through the transition state, are those with a lower k_1 and higher k_2 (Figure 4.5A and Figure 4.5B). As k_1 is increased beyond 0.4 and k_2 is decreased below 0.6 (combinations C and D), the sampling then becomes progressively worse. This is because the vertical energy well in the product basin then becomes the significantly lower energy route with increasing reaction co-ordinate. As a result of this the system crosses into the reactant basin progressively later for these combinations, such that the downhill path from the transition state is not fully sampled. It was therefore decided to use the weight combination lying between that for A and B, with k_1=0.3 and k_2=0.7, as the final optimised reaction co-ordinate.

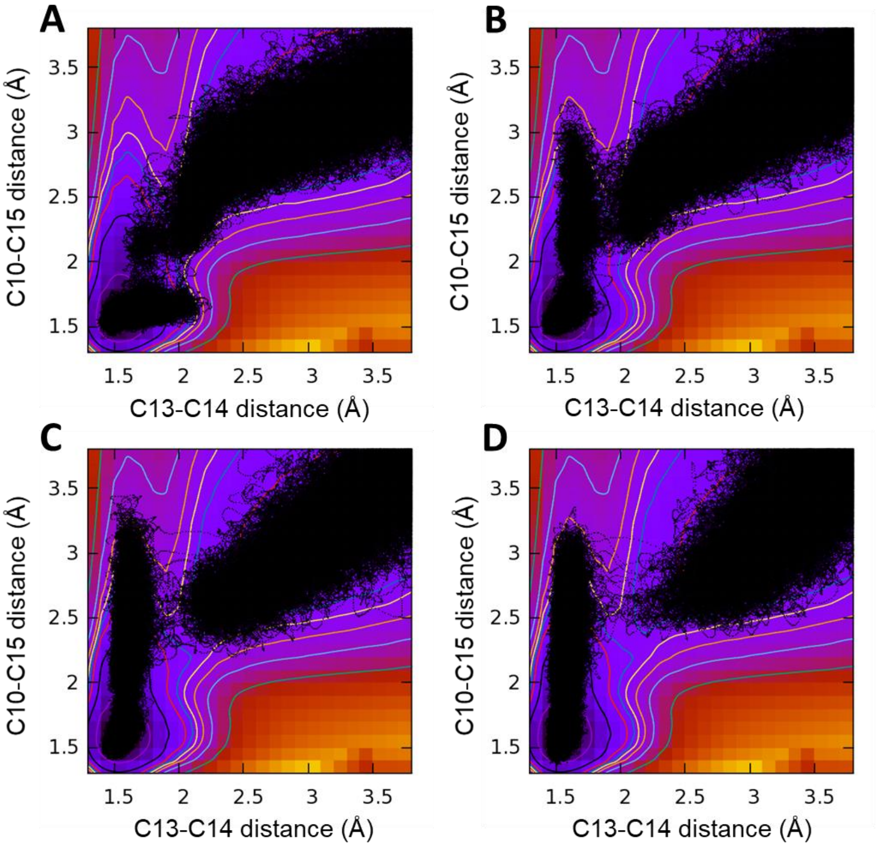


Figure 4.5: Sampling of the reaction surface for **10P** in binding mode A obtained from 1D umbrella sampling using a linear combination reaction co-ordinate with 4 different weight combinations (A-D). 1D sampling results for 10 repeat runs with each reaction co-ordinate are shown superimposed on the 2D potential energy surface. Weight combinations tested were as follows: A – k_1=0.2, k_2=0.8, B – k_1=0.4, k_2=0.6, C – k_1=0.6, k_2=0.4, D – k_1=0.8, k_2=0.2.

Using this optimised reaction co-ordinate, the system is consistently steered through the transition state and therefore always samples the entire downhill path towards the reactant as desired (Figure 4.6). Finally, although the potential energy surface and optimised reaction co-ordinate were obtained for the reaction in binding mode A, similar sampling was obtained using this reaction co-ordinate for the remaining binding modes. This indicates that the overall shape of the energy surface, at least in the region around the TS, is a function of the substrate itself and is not strongly affected by the enzyme. This can also be seen by comparing to the reaction pathway found in gas phase via IRC (section 3.3.1), which corresponds to the minimum energy pathway now found in the enzyme. Since this energy surface appears to apply to the intrinsic/gas phase reaction, this also explains why the reaction energy barrier obtained with DFTB was found to be sensitive to the path followed. The strongly curved nature of the energy surface found using this method penalises any divergence from the ideal minimum energy path (resulting in the much higher barrier found for the approximate path followed in the previous study on the enzyme – see section 3.3.1).

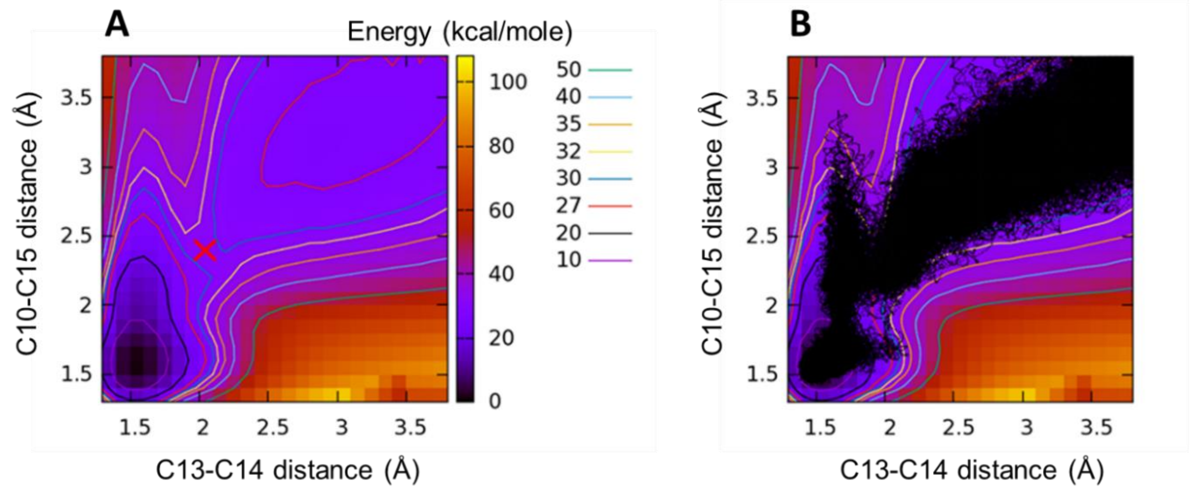


Figure 4.6: A) 2D potential energy surface for the reaction of **10P** in binding mode A (found at the SCC-DFTB/ff14SB level). Approximate transition state location is indicated by a cross. B) 2D potential energy surface with distances sampled during the 1D sampling superimposed (black dots). 1D sampling results shown for the 10 repeat umbrella sampling runs using the optimised reaction co-ordinate.

4.3.3 Productivity of different binding modes

To help establish the relevant binding mode, reactivity and binding affinity was studied for the 4 identified binding modes[4] via QM/MM umbrella sampling simulations and MM/GBSA binding affinity calculations. Free energy profiles obtained from individual umbrella sampling runs for the different binding modes are shown in Figure 4.7, and the average barriers derived from these depicted in Figure 4.8 (alongside the overall binding affinity predictions). For binding modes A and C, clear substrate minima are observed in all profiles. For the remaining binding modes, clear minima are generally always observed except for a couple of runs, which thus indicates that in these cases the profiles may not be giving a true estimate of the reaction barrier.

[4] Note: results are only reported for one of the atropisomers in each of the binding modes A-D. The trends in reaction barriers were the same for both atropisomers. For modes A and C, the substrate always ended up in a reactive substrate conformation corresponding to **10P** at the end of umbrella sampling, regardless of which product it came from; reaction barriers and binding affinities are therefore reported for simulations starting from **10P**. For mode D, although distinct substrate conformations were obtained for the two atropisomers, the average barrier and binding energy obtained was the same (within 0.2 kcal/mole); so therefore only the result obtained when starting from **10P** are reported (as it would be redundant to show both in terms of investigating substrate reactivity in the different binding modes). Finally, for mode B, results are only shown for **10A**, since this docking mode was not obtained with the standard docking procedure for **10P**.)

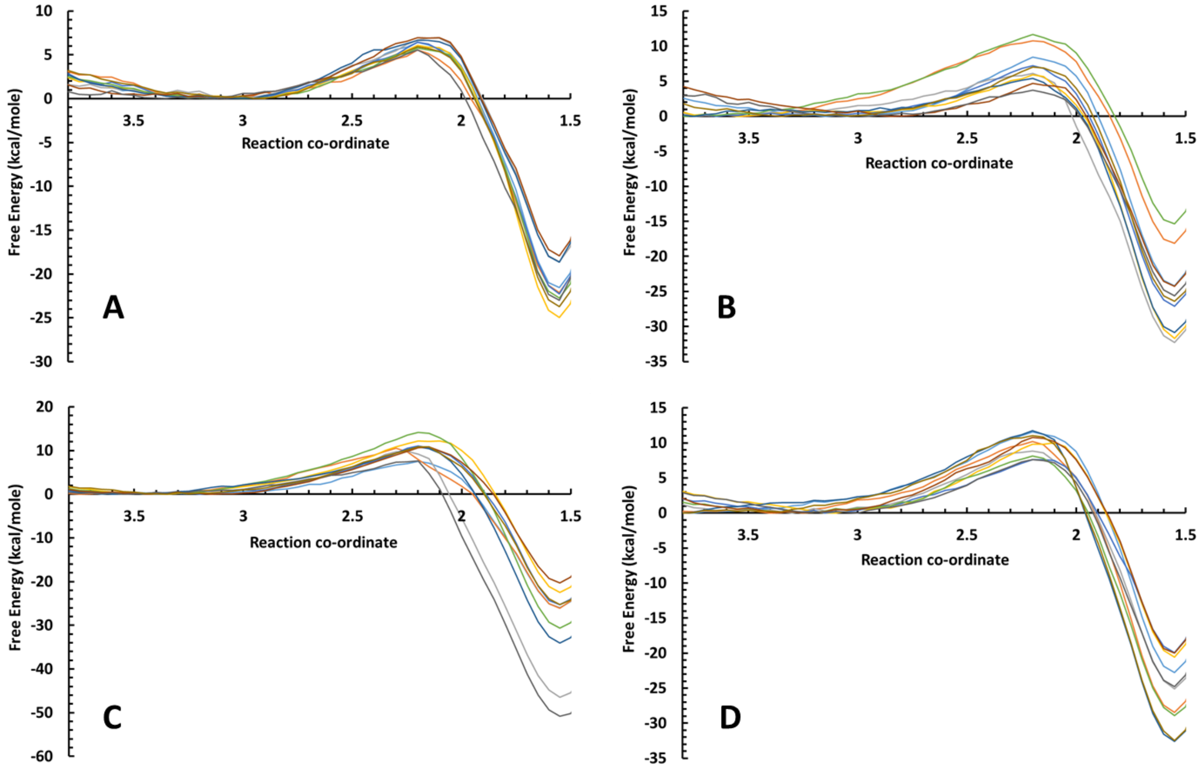


Figure 4.7: Individual free energy profiles (not including the SCC-DFTB correction factor) obtained for the 10 repeat umbrella sampling simulations in each of the 4 binding modes A-D. Profiles obtained by reweighting with WHAM (not including the SCC-DFTB correction factor). Energies shown relative to the calculated reactant minima.

Comparing the overall results (Figure 4.8), the average free energy barriers now give stronger evidence that the reactive binding mode suggested by the previous study (mode A) is indeed the most reactive. While binding mode B has an average barrier predicted to be within that of mode A, there is a much larger uncertainty in this, partly due to a couple of higher barrier runs (Figure 4.7). Also, an additional free energy penalty is expected for this mode because longer distances between the diene and dienophile were observed in MD simulations of the substrate complex (see the representative substrate structure in Figure 4.8 and average C-C distance in Figure 4.10), which is not accounted for in the reaction simulation protocol. As mentioned, a lack of sampling of the true substrate minimum for this mode is also suggested by the free energy profiles for certain simulations. Overall, there is therefore a clear indication from the reactivity predictions that mode A is indeed the most active binding mode. Since mode A also has the highest predicted binding affinity, it is therefore indicated to be the mode which contributes most strongly to the activity of the enzyme.

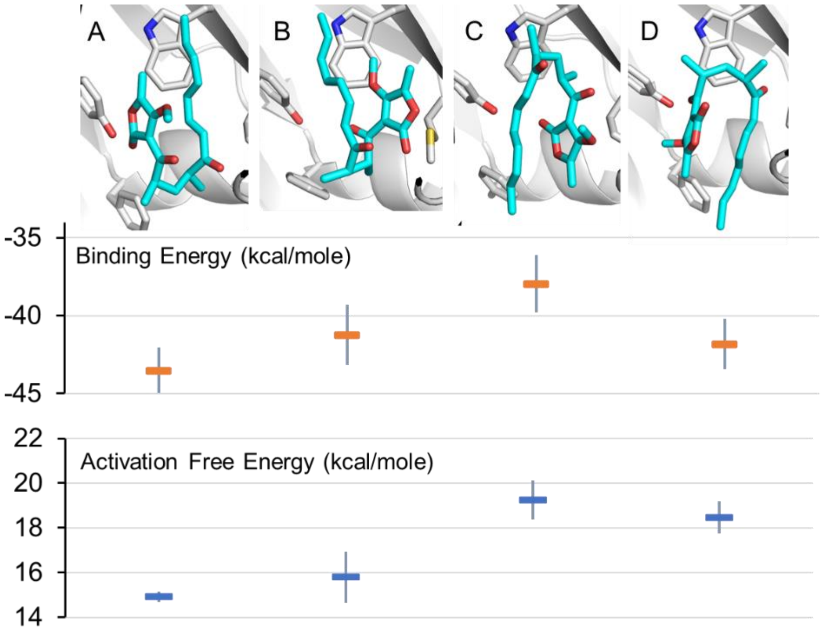


Figure 4.8: Top – representative structures from molecular dynamics (MD) simulations of the substrate complex for each of the 4 binding modes obtained from docking. Middle – Average binding energy for Michaelis complex of each mode (as calculated using the MM/GBSA approach using 1000 snapshots of each mode taken from 10 independent MD runs). Bottom – Predicted Diels-Alder activation free energies for each mode (obtained from 10 independent QM/MM umbrella sampling molecular dynamics simulations using SCC-DFTB/ff14SB, with a 8.7 kcal/mole correction for each to M06-2X/6-31+(d,p) level) – see section 3.3.1. Standard deviations are indicated for binding and activation free energies. See Methods section for details.

However, as multiple distinct rates of turnover were detected in pre-steady state kinetics experiments, it was of interest whether the substrate may also be binding and being turned over in alternative binding modes, with different associated activities. As well as the true barrier likely being higher for mode B, binding modes C and D both have predicted barriers that are notably higher than that for mode A and would therefore be expected to result in an observably lower, but still measurable, rate of turnover. Therefore, providing that there is a comparable level of binding in these modes as for the reactive mode, this supports the hypothesis that the different rates measured in the kinetics experiments are indeed the result of multiple productive binding modes. Alternatively, it may be that binding in one of the less reactive modes may lead to a 'dead-end', at which point unbinding and rebinding in a productive mode is required for the reaction to proceed (where the overall rate for this process will show up as the lower rate seen in experiment). Approximate binding energy calculations (using the MM/GBSA method) indicate that, while substrate binding mode A is thermodynamically preferred, modes B and D have comparable binding affinity. In addition, there could be an additional binding penalty for mode A due to the short distances seen in the Michaelis complex (Figure 4.10), as this suggests a potentially high energy conformation of the substrate when bound in this mode (which would not be reflected in the binding energy as, using the single trajectory approach, the conformations of the individual species are taken from the conformations of the complex to calculate energy of the unbound state). This suggests that substrate binding could occur mostly in a combination of modes A, B and D, and that there may indeed be multiple productive binding modes.

To help understand the significance of the calculated binding energies and the factors which may be influencing binding, the individual contributions to the overall binding energy can be considered (Figure 4.9). The Van der Waals interaction energy provides the dominant contribution to the gas phase binding energies, which makes sense given the mostly hydrophobic nature of the substrate and binding site. Moreover, this term is approximately the same for all modes which indicates that the protein can sterically accommodate the ligand well in various orientations. In terms of electrostatics, mode A clearly has the strongest interactions, which makes sense given the key H-bond interaction between the tetronate carbonyl and Tyr76, which is ideally made in this mode (this was also one of the main reasons this mode was thought to be the likely binding mode in the previous study). However, reasonably strong electrostatics are also seen for other modes, in particular B and D, which help them to yield similar overall affinities. Overall though, it is the hydrophobic binding and lack of specific interactions which is the main reason the substrate is predicted to bind with similar affinities in the different modes. Looking at the effect of solvent, in all cases the electrostatics of solvation are predicted to strongly favour the unbound state over the complex (nonpolar solvation energies however are much smaller and do not vary between the modes). This is thus indicative of greater electrostatic interactions with solvent being possible when the ligand and protein are free than when they are bound in the complex. In addition, polar solvation energies can be seen to disfavour the bound state more for binding mode A, which then brings the overall binding energy more in line with the other modes (counteracting its stronger electrostatic interaction energy).

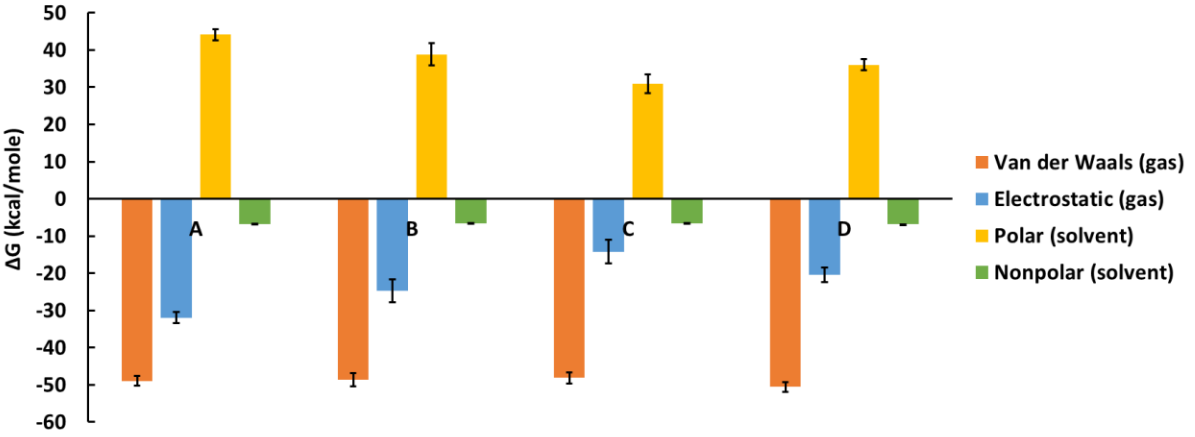


Figure 4.9: Decomposition of MM/GBSA binding energy contributions for modes A-D. Van der Waals (gas) and Electrostatic (gas) are the differences in the Van der Waals and Electrostatic energies between the bound/unbound state as calculated with the MM forcefield. As the single trajectory approach is used these

simply represent the interaction energy between protein and ligand due to the respective interactions, and together make up the gas phase binding energy. Polar (solvent) and Nonpolar (solvent) are the polar and nonpolar solvent contributions to the binding energy which arise from the difference in the solvation energies for the bound and unbound state. Polar solvation energies are calculated with the GB implicit solvent model while nonpolar solvation energies are determined via the LCPO method.

However, the implicit solvent model used to calculate polar solvation energies is approximate, and the real effect of solvation is much more difficult to predict. This is particularly true when a buried hydrophobic environment like the AbyU cavity is concerned, as it is difficult to know how solvated this site will be in reality. Implicit solvent may therefore be overestimating the electrostatics interactions that would be made with the unbound receptor and biasing these results. Finally, it should be mentioned that these calculations do not consider entropy change of binding for the ligand/receptor. One might speculate that this could be similar for each binding mode, as is the standard assumption for binding of small molecules. However, it may well be that greater conformational flexibility of the substrate and protein in different binding modes means that this is not the case. To gain more confidence in these results, gas phase entropies could therefore be calculated for the bound and unbound states by running either normal mode or quasi-harmonic entropy calculations as described in Chapter 2. An alternative to MM/PB(GB)SA is to obtain binding free energies via free energy perturbation (FEP) or thermodynamic integration (TI). Examples of this approach are alchemical free energy methods (93) for calculating relative binding free energies of different ligands and Waterswap (94) for determining the absolute binding free energy of a single ligand. In principle these methods should be very accurate since they are based on statistical mechanics. However, they are computationally expensive as they also require sampling of artificial intermediate states of the binding process. In practise their accuracy is also system dependent, and accuracy has been found to be better (95) or worse (96) than equivalent binding free energies obtained via MM/PBSA (i.e. where entropy calculations are also incorporated in order to obtain true free energies). To recap, the current MM/GBSA protocol gives some indication that a similar binding affinity may be expected for several modes (and thus be leading to the multiple rates mentioned), and an idea of why this may be the case is suggested by the individual contributions. However, due to limitations such as the exclusion of entropy and the uncertainty with the handling of solvation effects by the MM/GBSA method, this should only be taken as an initial indication at present.

To explain the different reactivities, as was suggested it could be that Tyr76 withdraws electron density from the dienophile in mode A and thus lowers activation energy. A similar hydrogen bond exists for the most reactive binding mode found for IdmH (as predicted by QM/MM umbrella sampling) which also catalyses an intramolecular DA reaction (26). However, a separate factor which was investigated to explain the differences was the distance between reacting groups in the Michaelis complex. This was suggested to be the reason for the reaction barrier for a particular bimolecular DA reaction being lower in the enzyme CALB than in solvent (86), as shorter distances between the bond forming carbon atoms were found for the reactant state in enzyme (when comparing stationary points from IRC calculations). While bringing the reactants closer together should provide a geometric advantage, shorter C-C distances were also shown to correlate with a lower HOMO-LUMO energy (86) so also lowers the electronic part of the activation energy. Comparing the activation energies for the different poses with the average C-C distances seen during simulations of the Michaelis complex (Figure 4.8, Figure 4.10), there is a clear correlation (ignoring mode B since the activation energy cannot reflect the large distances seen during MD due to the umbrella sampling protocol). This is also reflected in the overall free energy profiles (Figure 4.12) in terms of the positions of the reactant minima, where for the most reactive mode A it occurs closer to the TS at a lower reaction co-ordinate value. So, proximity of the reacting groups in the Michaelis complex may be the primary reason for the different reaction barriers. Although, for mode A there may be some specific interactions which provide a small amount of additional catalysis given the significantly lower barrier e.g. the possible electron withdrawing effect of Tyr76, as well as the potential 0.9 kcal/mole catalysis provided by Trp124 (which may be captured in these QM/MM barriers if this is indeed a diffusive effect – see section 4.3.1).

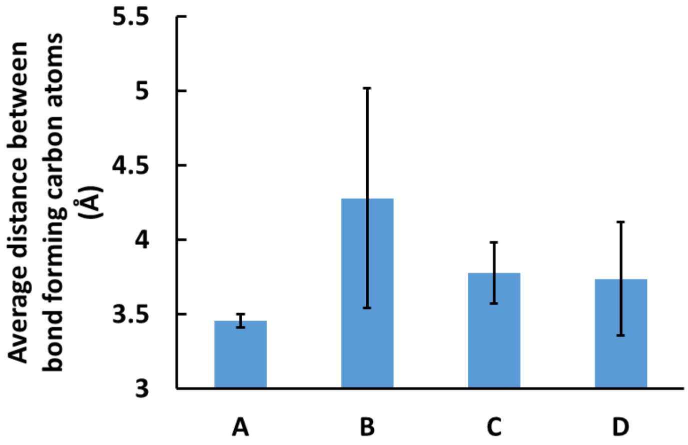


Figure 4.10: Average reactive distances for the Michaelis complexes of the different binding modes. Average is calculated from 1000 snapshots which came from the 10 independent MD simulations of the substrate complex. Error bars show the standard deviation in the average distances found for the individual runs.

Overall, these results suggest that, in line with the previous hypothesis, the primary function of the enzyme is to provide a complementary binding site for the reactive conformation of the substrate. From this it follows that all modes capable of accommodating a reactive conformation of the substrate will provide some baseline level of reactivity (where, as suggested by these results, the exact reactivity of each binding mode is largely determined by exactly how reactive a conformation it supports). As a further consequence of this, given that in this case binding is largely hydrophobic (and thus non-specific), it follows that there can be multiple binding modes in which the substrate is turned over. This is an interesting result, as it demonstrates a possible complex mechanism that may be generally applicable to other similar enzymes. Although for IdmH a single reactive binding mode was emphasised (and rationalised by the existence of an electron-withdrawing H-bond), there is actually a similar range in reactivities found for the different binding modes tested as for AbyU. This suggests a similar situation to AbyU may be the case, where little specific catalysis is performed such that there is then not a single reactive binding mode. These results therefore suggest a possible generalisation which can be made, which is that, in Diels-Alderases with hydrophobic binding sites and where the enzyme mostly functions to bind a reactive substrate conformation, a complex reaction mechanism may be found where reactivity is ultimately the combined result of various possible binding modes.

4.3.4 Confirming catalytic role of enzyme

Comparing the reaction free energy barrier in solution to that for the enzymatic reaction gives further insight into where the rate enhancement originates from i.e. whether this is chiefly due to capturing of the reactive conformation or if the reaction step itself is also enhanced. A lower reaction barrier was found in the enzyme than in solution (Figure 4.11) indicating that the enzyme does provide some catalysis of the reaction step itself. Looking at the overall reaction profiles (Figure 4.12) the observed reactant minimum is significantly closer to the TS in the enzyme, so again this indicates that the enzyme may be doing little more than trapping the substrate in a conformation nearer the TS than it would be in solution to achieve this (i.e. this does not contradict the interpretation of the binding mode results that the enzyme may have only a slight effect in terms of specific catalytic interactions).

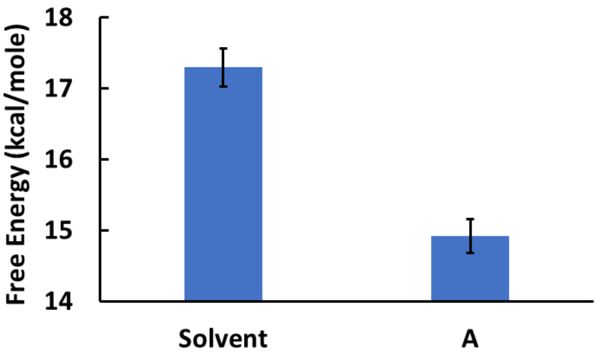


Figure 4.11: Average and standard deviation in the Diels-Alder activation free energies predicted for reaction in solution and in the reactive binding mode (obtained from 10 independent QM/MM umbrella sampling molecular dynamics simulations using SCC-DFTB/ff14SB or SCC-DFTB/TIP3P, with a 8.7 kcal/mole correction for each to M06-2X/6-31+(d,p) level) – see section 3.3.1.

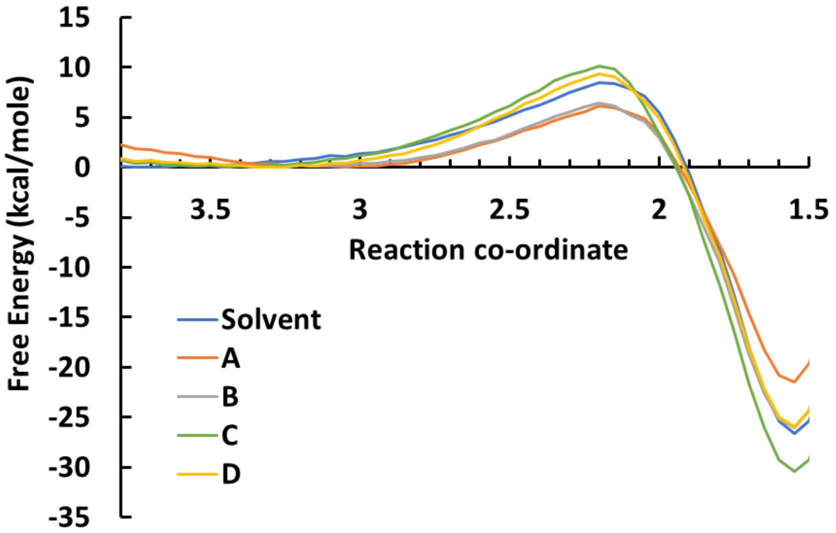


Figure 4.12: Overall free energy profiles (not including the SCC-DFTB correction factor) for the reaction in solution and in binding modes A, B, C and D. Profiles obtained by running WHAM on the combined sampling from the 10 repeat QM/MM umbrella sampling simulations for each environment. Energies are shown relative to the reactant minima.

However, a potentially confounding issue with this is that the semi-empirical method used in the reaction simulations does not correctly predict a stable folded minimum for the substrate in gas phase. This casts doubt on the reactant state being at such a high value of the reaction co-ordinate in solution, since the incorrect tendency of the substrate to open in gas phase could be contributing to this. However, the fact that the reaction profile does flatten out in solution and give a proper minimum shows that this opening issue may be mitigated by solvent and thus that the prior interpretation (i.e. that lowering of the reaction barrier is due to the closer proximity of the reacting groups in the reactant state in enzyme) is valid. Further, the reactant minimum in solution does not occur at a higher reaction co-ordinate than that found from IRC in the gas phase with the more accurate M06-2X level (see Figure 3.10). This thus gives further confidence that the observed difference in the reactant minima is the true result of the different environments and not just due to the erroneous behaviour of DFTB. However, as future work it would be recommended to obtain reactant state minima in the enzyme and in solution from e.g. QM/MM IRC calculations with M06-2X for the QM region (as was also done for CALB (86)), in order to verify that the enzyme does indeed enforce closer docking of the reactive groups. Finally, comparing the reaction profile in solution to those for the binding modes with the highest barriers (Figure 4.12) provides some potential insight into the catalytic effect of solvent. The barrier in solution is slightly lower than those for modes C and D despite the reactant minimum occurring at a slightly later reaction co-ordinate. This could suggest that, while little if any may be provided by the enzyme in these modes, solvent does in fact provide some TS stabilisation (which thus makes up for it having a less reactive substrate minimum). This would make sense given that solvent is more mobile, giving it the potential to rearrange to provide TS stabilisation.

4.4 Conclusions

In this chapter, the enzymatic AbyU reaction has been studied via both DFT calculations and QM/MM simulation, giving insight into the role of the enzyme active site in catalysis. The role of the Trp124 was investigated first, which indicated only modest catalysis (and was therefore excluded from the QM region in subsequent QM/MM calculations). QM/MM umbrella sampling was then used to obtain reaction free energies for both the presumed reactive binding mode as well as the other possible binding modes. An efficient protocol was specifically developed for this, which still provided good precision in the predicted reaction free energies (in order that this can then also used for the high-throughput reactivity screening of alternative substrates in Chapter 5). The alternative binding modes were found to be less reactive, while having a comparable predicted binding affinity to the reactive mode; this suggests that a significant proportion of substrate may bind in a suboptimal conformation and be responsible for the attenuated rates found in pre-steady state kinetics. Reactivity in the different modes appears to be mostly the result of proximity of the reacting groups in the Michaelis complex, while for the most reactive mode some slight additional catalysis may come in the form of other subtle catalytic interactions (e.g. from Tyr76 and Trp124). Finally, the origin of the enzymatic rate enhancement was further probed by comparing to the reaction barrier found in solution. While it was determined in Chapter 3 that capture of the folded reactive conformation of the substrate alone may provide a reasonable rate enhancement, it appears the enzyme then encourages the reaction further by bringing the diene and dienophile into even closer proximity than they would be in solution. Interestingly, some slight TS stabilisation may be provided by solvent (given the lower barrier compared to less reactive poses), however, the enzyme ultimately provides the more favourable environment given the highly reactive Michaelis complex for the reactive binding mode (along with the other possible catalytic interactions previously mentioned).

Chapter 5. Towards simulation-led engineering of AbyU substrate scope

5.1 Introduction

Due to the antibiotic potential of spirotetronates and the practicality of biosynthesis, it is a desirable goal to find enzymes capable of producing a diverse set of spirotetronate products (see Chapter 1). Therefore, in this chapter the potential for AbyU and engineered variants to accept alternative substrates has been explored computationally. To do this, an *in silico* set of protocols have been used to predict substrate binding affinity and reactivity to determine if the enzyme is likely to effectively turn over a particular substrate. Rather than rationally redesigning the enzyme for a specific alternative substrate, the first approach taken was to perform high throughput combinatorial screening of a panel of existing enzyme variants with various alternative substrates in order to identify promising combinations, whose activity could then be verified experimentally. The native substrate was also run through this screening for reference. The enzyme variants and alternative substrates tested were those which had already been produced or were being attempted to be produced in a parallel experimental effort to test the substrate scope of the enzyme. The idea was therefore, rather than testing all possible combinations experimentally in the first instance (which is far more time consuming and expensive), to use a quick and efficient activity prediction protocol to narrow down the array of possible substrate-mutant combinations to only those showing potential catalytic activity.

The alternative substrates which have been studied are shown in Figure 5.1 alongside the methoxy analogue **10** of the native substrate. These were chosen either for their chemical importance or simply to test the scope of the enzyme. They all share the same overall structure in terms of the tetronate ring linked to a triene group and can undergo the analogous cyclisation reaction as the native substrate where the end diene of the triene reacts with the exocyclic methylene of the tetronate via a concerted DA. Substrates **11**-**16** share the same skeletal structure as the native substrate and differ only in the substituent groups attached to the backbone. **11**-**14** differ in terms of the number or position/orientation of the methyl substituents along the polyketide linker, while **15** and **16** have different substituents replacing the methyl which is attached to the end carbon of the triene group in the native substrate. Substrate **15** has an alkyne group which is of interest due to possible click chemistry which it enables, while **16** adds a bulky phenyl group. Substrates **18** and **19** have the same number of atoms making up the backbone of the molecule as **10** but have different groups substituted within the polyketide linker (and unlike **10** have no methyl substituents attached to the polyketide part of this linker). For **18**, the carbon preceding the second carbonyl carbon in the chain is replaced by a NH_2 group thus forming an amide. **19** is then the same as **18** except that the second carbonyl group is also replaced by a sulfonyl group, forming a sulfonamide. These two compounds were chosen as they modify the ability of one of the Michael acceptors of Abyssomicin C (which the second carbonyl group forms part of) that confers antibiotic activity (see section 1.2.2.2.2.1). Finally, in **17** the polyketide linker has been made longer by one atom by substituting an additional CH_2 group within it (and also has no methyl substituents attached to the polyketide chain). As with the native substrate, the alternative substrates can all form two atropisomers. For **11**-**18**, similarly to the natural, these can be characterised by whether the carbonyl oxygens point in the same or opposite directions in the product. For **19** the second carbonyl group is not present but the two atropisomers can be similarly characterised by whether the two sulfonyl oxygens point in the same or opposite direction to the first carbonyl oxygen in the chain.

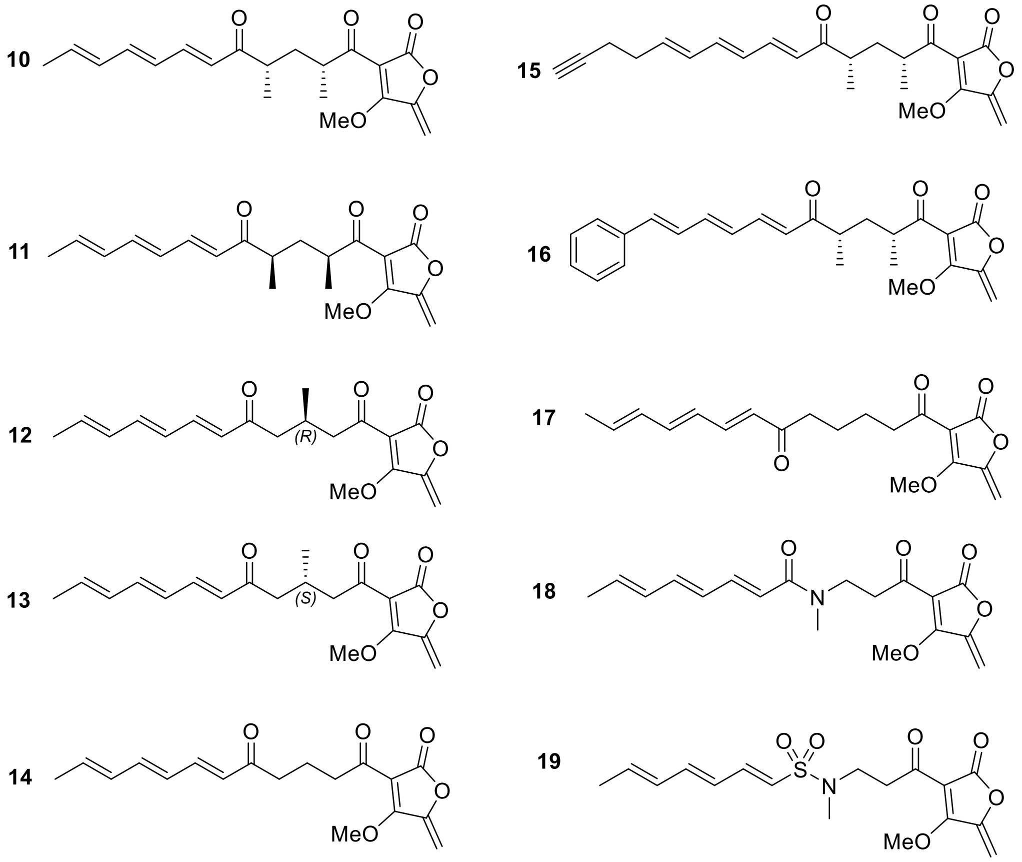


Figure 5.1: Substrates tested in the combinatorial screening.

The enzyme variants tested included the WT and 4 active site point mutants. These were W124F, W124A, Y106F, Y76F (Figure 5.2). These were all successfully expressed and soluble[5], indicating that the enzyme was able to tolerate these mutations. However, as no structures had been solved when

[5] L. Maschio, K. Tiwari & P.R. Race, personal communication

the simulations were being performed, starting structures were obtained from the WT crystal structure, making the relevant mutations *in silico*. When experimental mutant structures are not available, for example in protein design applications, various mutational protocols exist for generating mutant structures starting from a WT structure and modelling in the modified sidechain. It is important that the mutated sidechain orientation is correctly predicted as this can strongly impact the perceived effect of a given mutation (97). While the accuracy of sidechain conformations predicted by these protocols can be confirmed by running MD, for protein design applications where many mutations are to be evaluated it is desirable to not have to first search conformational space to be sure the most relevant rotamer has been found. Programs like Rosettafixbb (98) and Scrwl4 (99), which use backbone dependent rotamer libraries to suggest orientations (which are then scored to determine the best rotamer), have been found to predict well the dominant conformations seen in MD simulations (100). However rather than using one of these programs, since few mutations were involved in this study, these were more simply made in PyMOL using the mutagenesis wizard. This suggests rotamers from the PDB in either a backbone dependent or independent manner (ordered according to their frequency of occurrence). The accepted rotamer is then decided by the user by visual inspection (where PyMOL provides a visual indication of favourable/unfavourable contacts).

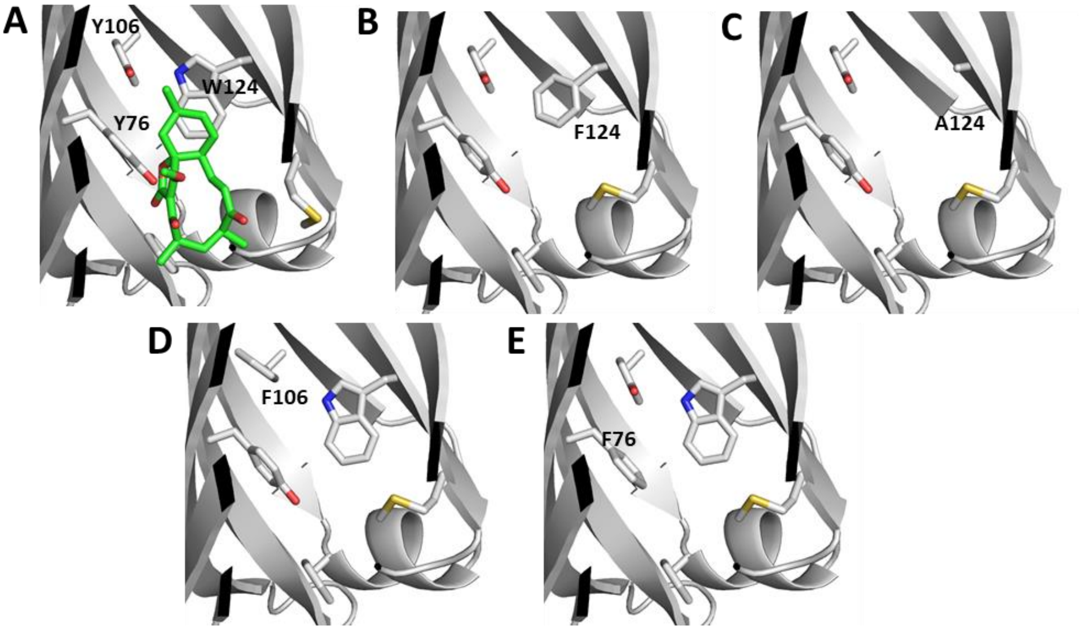


Figure 5.2: A) Docked pose of **10P** with WT enzyme in reactive binding mode identified in Chapter 4. Key residues are highlighted with sticks and positions at which screened points mutations were made are indicated. B-E) Starting structures generated for mutants W124F (B), W124A (C), Y106F (D) and Y76F (E). Mutant structures were built from WT crystal structure making mutation manually in PyMOL and selecting rotamer shown.

5.1.1 Outline of screening approach

Based on the native reaction presented in Chapter 4, assumptions were made to allow the creation of a standard automated workflow for catalytic activity prediction. Firstly, it was assumed that the overall binding characteristics/product outcome are not changed by the modifications to the substrate/enzyme. Therefore, in terms of chirality, only production of the analogous reaction product to that for the native reaction was considered for the alternative substrates (although formation of both atropisomers was tested). Further, substrate reactivity and binding affinity were predicted based on the most similar binding mode to the most reactive binding mode identified for the native substrate in Chapter 4. These simplifying assumptions can be justified by the fact that only single point mutations are made which effectively maintain the overall shape of the active site and either just subtly change interactions in the binding site (e.g. Y to F mutations) or create more space which may help accommodate changes to the substrate given similar binding (e.g. W124A creates space at top of cavity which may accommodate the phenyl group of **16**). The same overall procedure for reaction barrier/binding affinity prediction as used in Chapter 4 (for assessing the reactivity and binding affinity for the different poses of the native substrate) was then also used here. Therefore, to predict the catalytic potential of a particular mutant/substrate combination, each atropisomer of the analogous reaction product to that for the native reaction was first docked into the enzyme active site. The poses most closely resembling the reactive mode for the native substrate (Figure 5.2) were then selected and used as the starting point for Umbrella Sampling simulations followed by corresponding simulations of the Michaelis complex. Although poses were selected manually, to facilitate a truly automated workflow this could also have been done via calculation of a similarity metric (e.g. RMSD of the common ligand atoms). The described workflow is depicted graphically in Figure 5.3, where the output of each protocol can be seen to feed into an overall prediction of catalytic activity (although if for example it was found at the docking stage that the substrate did not fit into the enzyme in a pose analogous to that for the native reaction it could perhaps be concluded from this alone that binding is suboptimal and the enzyme-substrate combination could be rejected based on this alone (see Results)). The simulation protocols themselves were the same as those used in Chapter 4 for studying the native reaction. These protocols were developed with this in mind as they are designed to be computationally efficient to enable quick screening of substrate-mutant combinations. Also, having demonstrated their applicability for studying the native reaction already somewhat validates the of use of these protocols in a standardised workflow for alternative substrates. For example, as the overall systems are quite similar (as should therefore be the conformational sampling requirements), the achieved precision in the predicted values can be expected to be similar for alternative substrates/mutants. Therefore, as the umbrella sampling protocol used for reactivity prediction was shown to be able to discriminate reactivities to within a few kcal/mole for the native reaction (and thus demonstrate significantly different reactivities between binding modes) this should hopefully make it sufficient for effectively discriminating between more or less active substrate/mutant combinations. Finally, Figure 5.3 shows possible next steps that can be taken given the findings of this study (see Results).

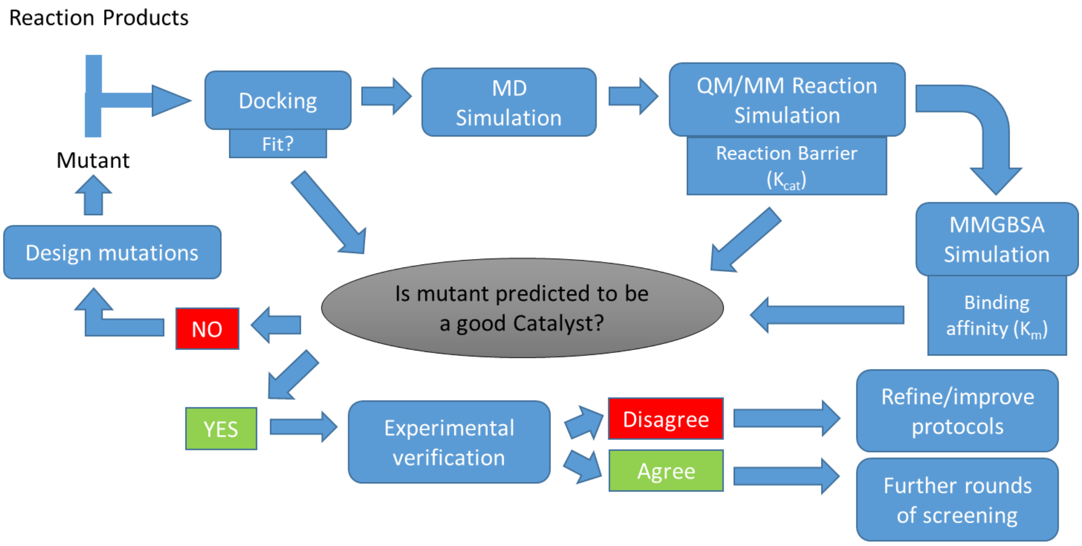

Reaction Products
Mutant
Docking
Fit?
MD Simulation
QM/MM Reaction Simulation
Reaction Barrier (K_{cat})
MMGBSA Simulation
Binding affinity (K_m)
Design mutations
Is mutant predicted to be a good Catalyst?
NO
YES
Experimental verification
Disagree
Agree
Refine/improve protocols
Further rounds of screening

Figure 5.3: Flowchart illustrating proposed iterative workflow for catalytic activity prediction/design. Output of each computational protocol is used as input for the next, and results generated feed into an overall prediction of catalytic activity. Information gleaned from this can be used to guide further mutation. Comparison with experimental results is used to refine prediction protocols.

5.1.2 Rational redesign

In addition to the screening of specific point mutants, a second strategy employed to increase substrate scope was to rationally redesign the binding site to optimise it for specific alternative substrates. This was done with the Rosetta Enzyme Design package (27), which enables automated design of a binding site by optimising ligand placement, along with the identity and orientation of selected nearby residues (optionally incorporating specific interactions via constraints). As a stochastic optimisation procedure is used to find overall low energy structures (which also optimise protein-ligand interactions) this is typically used to generate a diverse set of different designs which are then narrowed down and tested to see if they perform well in practise. This tool has been used both as part of a *de novo* design process by designing a novel binding site into a generic scaffold (27) as well as to redesign existing binding sites to accommodate modified substrates (101, 102). For example, the aspartase AspB, which catalyses the deamination of aspartate into ammonia and fumarate, was redesigned to be used for the reverse hydroamination reaction with α,β-unsaturated carboxylic acids other than fumarate (101). To do this, the residues which form the α-carboxylate binding pocket for aspartate were allowed to change to accommodate the respective groups in the alternative substrates, while at the same time maintaining the interactions which are relevant for all substrates (those made by the amine and β-carboxylate binding pockets and the catalytic serine which abstracts a proton from the β carbon in the deamination direction). This design process was successful in finding variants with the desired activities (also showing 99% stereoselectivity towards the product they were optimised for in some cases). In a similar fashion, Rosetta was used to redesign limonene epoxide hydrolase (LEH) to instead catalyse the epoxide hydrolysis of cyclopentene oxide (102). Again, the active site was redesigned to accommodate the changes to the substrate while preserving the interactions which facilitate epoxidation. As, in this case, the goal was to obtain a pair of complementary stereo-selective enzymes which produced either the RR or SS diol, two separate design runs were performed, where each optimised the binding site for the catalytic configuration corresponding to the relevant product. These variants were then screened computationally to identify the most stereoselective variants before proceeding to the lab. As the initial designs generated by Rosetta were not found to be stereo-selective mutations designed to inhibit production of the unwanted product were manually specified in subsequent design runs. This case therefore highlights the important fact that while Rosetta optimises the active site for a given binding mode, it has no way to guarantee by itself that unwanted binding modes will not also be present. Therefore, this reinforces that, in general, many designs should be generated to increase the probability that one showing the desired behaviour can be found.

A similar strategy to that described for AspB and LEH has been employed here for AbyU. The binding site was redesigned for substrates **14** and **19** to accommodate the changes to the substrate while retaining the native binding characteristics where relevant (for the most reactive binding mode identified in Chapter 4). Although, since there are not thought to be any specific interactions essential for catalysis (catalysis likely coming mainly from general shape complimentary), the entire active site could have potentially been redesigned for the novel substrates (i.e. without stipulating any constraints), it was decided to maintain some of the native binding interactions in order to diverge less from a binding mode known to provide good substrate affinity/reactivity. Specifically, since these likely work together to enforce close docking between the diene and dienophile, the Tyr76 and Trp124 were retained and enforced in the orientations they adopt for the native binding mode. Furthermore, the native H-bond interaction between the Tyr76 and tetronate carbonyl of the substrate was enforced which both provides a favourable electrostatic interaction and, along with the other constraints, helps enforce the same overall orientation of the substrate within the active site as for the native reaction. It was envisioned that, since the chosen alternative substrates differ from the native substrate in the polyketide chain region which links the diene/dienophile (which is positioned at the ‘bottom’ of the active site in the chosen binding mode), there may be potential for Rosetta to optimise the active site by modifying the residues lining the bottom of the active site cavity. For **14**, since methyl substituents are removed compared to the native substrate, it was anticipated that packing may be increased by introducing bulkier residues to fill the gaps. For **19**, it was speculated that electrostatic interactions may be introduced by Rosetta to stabilise the additional polar oxygen present as a result of the sulfonyl group. The few best designs obtained from Rosetta (according to Rosetta’s scoring function) were evaluated computationally using the same screening workflow as used for the point mutants.

5.2 Methods

5.2.1 Combinatorial screening

5.2.1.1 Structure preparation and docking

The two atropisomers of the analogous reaction product which forms in the native reaction were built for each of the alternative substrates. These structures (in PDB format) were generated in Chem3D, optimised using the built-in MM2 minimiser. The coordinates for the WT enzyme were taken from the crystal structure of AbyU, removing the buffer molecule HEPES bound in the active site (PDB ID: 5DYV, chain A). The mutant structures were built from the WT structure using the mutagenesis wizard in PyMOL. In all cases, backbone dependent rotamers were used and the number one ranked rotamer (i.e. most frequently occurring) was chosen as this also had the least clashes with surrounding residues. The same docking setup and procedure as used in Chapter 4 was used here, except for the Y106f mutant where mutated residue 106 was also made flexible. The poses most similar to the product poses for the most reactive binding mode identified in Chapter 4 were then selected for simulation. In cases where an analogous pose could not initially be found, the docking was rerun (since it is stochastic in nature different poses can be identified in separate runs). If an analogous pose was still not found, then the combination was not screened. The exceptions to this were for the products of **15** and **16**, where docking indicating that the native binding mode was no longer supported by the enzyme (see docking results section). Therefore, for these substrates, poses were tested which differed significantly different from the native reactive binding mode. To evaluate the similarity of the selected product poses to the native binding mode, RMSD of the poses were calculated (relative to the respective natural product pose) for select heavy atoms that are part of the overall spirotetronate scaffold. The atoms chosen for comparison are those that come from the tetronate ring/attached ketone and the triene of the substrate (Figure 5.4) since these are common and adopt the same conformation in all products.

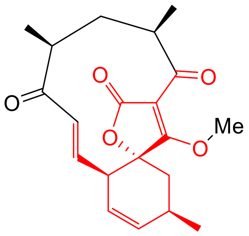


Figure 5.4: Region of molecule used to compare similarity between different product poses. Atoms selected for RMSD calculation are shown in red.

5.2.1.2 Reaction barrier and binding affinity prediction protocols

The docked product complexes were prepared using the same setup procedure, and input to the same reactivity/binding affinity prediction workflow as was used for evaluating the different binding modes for the native reaction in Chapter 4 (see Figure 5.3 and Chapter 4 methods section).

5.2.2 Rational redesign

5.2.2.1 Rosetta Enzyme Design

The binding site was optimised around the reaction products rather than the reactive conformation of the substrate as this is more well defined and simplified the design procedure. While ideally the binding site would be optimised with the substrate, it was thought that the changes required to optimise for product binding may be similar to those required for the substrate in the reactive conformation, since the region where binding will be less optimal is around the polyketide linker of the substrate (which does not vary much structurally between bound substrate and product). The atropisomer referred to in this thesis as the 'A' product was chosen for this. The starting structures used in the design were the docked product complexes chosen for simulation in the combinatorial screening (i.e. the poses found which were the analogues of the most reactive native binding mode - shown in Figure 5.5) after protonation by the PREP protocol. The Rosetta parameter file was generated for the ligand and the dummy partial charges replaced with those generated via the PREP protocol. Typically, an ensemble of relevant low energy ligand conformers is also generated to which the active site redesign is targeted. Since the rigid product was used, however, this was not deemed necessary here. Before running the design algorithm, the structures from docking were pre-optimised by relaxing with the Rosetta score function. This aids design as it ensures that Rosetta will suggest structural changes which result in an actual improvement to the WT, not just because they relieve the issues in a poor starting structure (103). Residues allowed to mutate during the design were manually specified with a resfile. These are indicated in Figure 5.5 and were essentially all those lining the active site cavity with the exception of Trp124 and Tyr76. For **14**, Arg130 was mistakenly included as this does not point into the cavity and separates the cavity from solvent. In addition, it makes a salt bridging interaction with Glu30 which is likely important for protein stability. This residue was thus excluded for the subsequent redesign for **19**, and an additional valine was also included which was mistakenly omitted for **14**. All designable residues were allowed to mutate to any other residue, with the exception of the two Phenylalanine residues at the bottom of the cavity which were constrained to remain as any of the canonical nonpolar amino acids so as not to disrupt the hydrophobic interactions at the capping loop interface (see Chapter 7). In order to maintain the native binding mode, the orientations of the Trp124 and Ty76 were fixed during the design and a restraint was placed to enforce the H-bond between the Tyr76 and tetronate carbonyl of the product. Specifically, the distance between the donor H atom and the acceptor O atom was restrained to the ideal H-bond distance of 2 Å (with force constant of 100 Rosetta Energy Units (REU)/ Å^2). All other non-designable residues were repacked (i.e. their sidechain orientations allowed to change) if they had their C$_\alpha$ atom within 6 Å of any ligand atom. In addition, the backbone was allowed some flexibility during the minimisation cycles of the design process by including the bb_min option. Rosetta searches for design structures with a lower overall energy score than the input structure, thereby optimising the entire complex. However, the contribution of ligand-protein interactions to the energy score can be increased (by increasing their weighting via the lig_packer_weight parameter) to prioritise optimisation of the protein-ligand interface. The lig_packer weight parameter was therefore set to be 1.8 to improve binding energy. 100 designs were generated in the design run for **14** and 150 for **19**. The resulting designs were filtered down to only those showing both the lowest overall and ligand interface energy scores, as well as having negligible constraint energies (to ensure the ligand was not being artificially held in the designed binding mode). For both ligands, a cutoff of -11 REU was used for the ligand interface score (best scoring designs in terms of interface energy had scores of -13.04 REU and -12.31 REU for **14** and **19** respectively), while for the total energy cutoffs of -220 REU and -232.5 REU were used for **14** and **19** respectively (lowest total energy scores were -223.91 REU and -234.63 REU for **14** and **19** respectively). The filtering procedure resulted in 4 design structures remaining for **14** and 5 structures for **19**, but in both cases only 3 of these had unique sequences. For the designs that had the same sequence the structure with the lowest total score was chosen to test the variant as this should be the most representative.

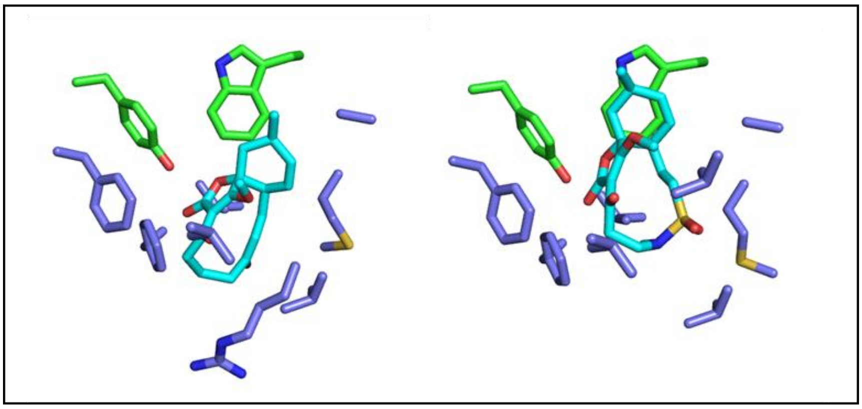


Figure 5.5: Poses from docking of ‘A’ products of **14** (left) and **19** (right) with the WT enzyme which were input to Rosetta for redesign of the active site. Residues which were set as designable are those shown in purple.

5.2.2.2 Evaluating redesigned variants

The redesigned product complexes were input to the same setup and activity prediction workflow as used for the docked complexes in the combinatorial screening. Although it is mainly the binding affinity of the substrate which is of interest (as this is the aspect of the reaction which the design process is hoped to have optimised), the results of the umbrella sampling (which is mainly used here to generate the substrate complex from the product complex for testing) are also reported. However, it should be noted that these were obtained using the original centre-of-mass based reaction co-ordinate, so may be less reliable (see section 4.3.2). Finally, to evaluate the success of the optimisation procedure itself (since the design process is technically optimising the binding site for the reaction product), the binding affinity was also calculated for the product using the same binding energy calculation protocol as for the substrate. An ensemble of 1000 snapshots were therefore generated for the redesigned product complex by running 10 independent 100 ps MD runs with DYNAM (starting from the output of the STRUCT protocol for the complex which was run when generating the umbrella sampling starting structures). This then gave the same number of snapshots as for the substrate complex which were input to MM/GBSA to calculate the binding affinity. This was done for both the WT product complex tested in the combinatorial screening as well as the redesigned complexes for comparison.

5.3 Results and discussion

5.3.1 Combinatorial screening

5.3.1.1 Docking

For most cases other than the few which are indicated, docked poses of the products could be found which were approximately analogous to the reactive binding mode for the native substrate (Table 5.1). Except for the case of the W124A mutant, the poses indicated as being analogous in Table 5.1

generally have an RMSD of less than 1 Å compared to the binding mode for the equivalent product of the methoxy analogue substrate **10** (see Methods for how this was calculated). In terms of the ligand orientation, the pose was considered as being analogous if the diene/dienophile end of the product was at the 'top' of the cavity (i.e. running alongside the Trp124 as in Figure 5.2) and the tetronate ring of the substrate was on the 'left' side of the cavity such that, for the variants with a Tyrosine in position 76 (all except Y76F), the tetronate carbonyl may be feasibly able to form the H-bond interaction with the Tyrosine. In terms of sidechain orientations, selected poses all had the analogous orientation of residue 76 and, where relevant (i.e. for all except W124A), residue 124. Therefore, selected poses that are indicated as being different to Figure 5.2 are so because of ligand orientation, specifically this was because the product was rotated along its long axis so that the tetronate carbonyl is no longer in range of residue 76.

Table 5.1: Docking ranks (R) and RMSDs of the product poses selected for screening of the alternative substrates with the WT enzyme and point mutants (docking scores are reported in Figure 5.9 and figures showing all selected poses are shown in the Appendix).

Substrate Ref. (Figure 5.1)	Atropisomer	WT		W124F		W124A		Y76F		Y106F	
		R	RMSD (Å)	R	RMSD (Å)	R	RMSD (Å)	R	RMSD (Å)	R	RMSD (Å)
10	A	3	0.000	6	0.297	14	1.198	3	0.089	3	0.212
	P	1	0.000	1	0.084	6	0.245	1	0.335	2	0.021
11	A	-[a]	-	-[a]	-	-[a]	-	8	1.109	8	0.879
	P	1	0.895	2	0.991	11	1.326	1	0.928	1	0.854
12	A	1	0.842	5	0.690	5	1.181	1	0.755	1	0.841
	P	4	0.215	6	0.275	11	2.213	7	1.543	7	0.242
13	A	3	0.598	7	0.499	9	1.170	4	0.636	4	0.619
	P	1	0.300	2	0.361	11	3.114	3	0.324	1	0.317
14	A	2	0.338	5	1.418	8	1.348	1	1.211	2	0.414
	P	3	0.439	4	0.446	10	1.291	4	0.319	3	0.272
15	A	1	0.269	1	0.166	9	2.450	1	0.150	1	0.201
	P	1[b]	2.149	1[b]	2.135	5[b]	2.046	1[b]	2.170	1[b]	2.132
16	A	2[b]	4.495	1[b]	3.387	11	2.551	1[b]	3.441	1[b]	3.396
	P	3[b]	2.444	2[b]	2.441	3[b]	2.734	1[b]	3.512	2[b]	3.483
17	A	3	0.238	6	0.349	10	3.212	5	0.301	7	0.250
	P	3	0.360	3	0.470	11	2.279	3	0.454	2	0.408
18	A	7	0.573	9	0.519	10	2.346	7	0.570	-[a]	-
	P	1	0.861	3	0.800	6	1.472	5	1.550	1	0.831
19	A	-[a]	-	8	0.694	-[a]	-	6	0.723	12	0.745
	P	5	1.512	8	1.499	7	2.334	6	1.521	8	1.521

[a] ***Pose analogous to binding mode shown in Figure 5.2 (see description in text) could not be found and did not screen combination (see text)***

[b] ***Consistently could not find analogous pose for product/s indicating binding mode not possible – thus a different type of pose was tested (see text)***

Finally, it should be noted that for the W124A mutant, the poses selected as being analogous are generally quite different to those for the other variants (where the RMSD relative to the native pose is almost always greater than 1 Å) due to the void introduced at the top of the cavity. Since the cyclohexene ring end of the product is then no longer supported by a sidechain at position 124, it instead occupies the vacancy left by the Tryptophan (see example in Figure 5.6). As shown in the following section, this mutation almost universally impairs both reactivity and binding of the substrate.

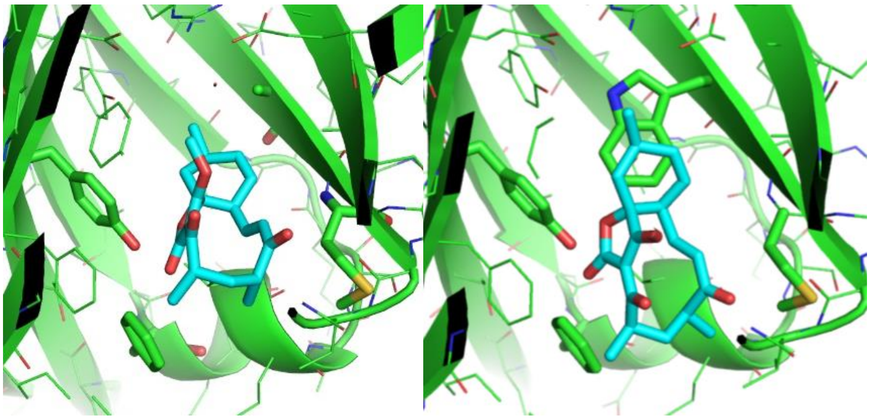


Figure 5.6: Left - Example of pose selected as being analogous to the native binding mode for W124A mutant. Diene end of molecule occupies void left by residue 124. Pose shown is that for the 'P' atropisomer of **11**. Right - Native binding mode for 'P' atropisomer shown for comparison.

In isolated cases where an analogous pose was not found (see footnote *a* in Table 5.1), it was not clear whether this precluded the binding mode being possible (or whether the individual docking runs had simply not been able to find the pose). However, as docking was not rerun a third time these combinations were skipped. In cases where analogous poses were consistently not found (footnote *b* in Table 5.1), this indicates that the binding mode is no longer supported. Firstly, for substrate **16**, analogous poses were essentially never found from docking: the only poses found that had the diene at the 'top' of the cavity then had the tetronate ring at the 'front' or 'back' instead of at the 'left' side (see example in Figure 5.7). This is due to the benzene ring which forces the main skeleton to sit lower in the cavity when it is oriented with the diene end at the 'top' and means that there is simply not enough room for this substrate to fit in the active site in an analogous pose. This could therefore already be an indication of suboptimal binding for this substrate. To test this substrate the top ranking alternative poses (with diene at the 'top' of the cavity) were instead selected for simulation.

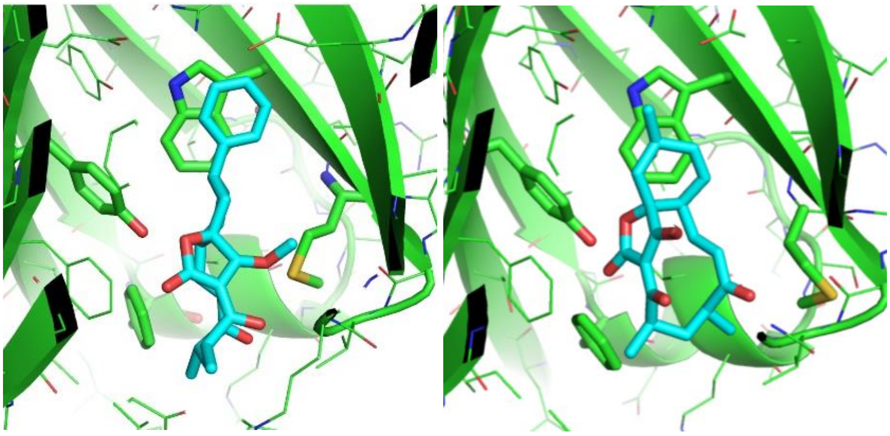


Figure 5.7: Left - Example of a pose selected for screening of substrate **16** (where analogous poses could generally not be found). Showing pose for **16P** with WT enzyme. Right - Native binding mode for 'P' atropisomer shown for comparison.

Secondly, although generally it is the case that analogous poses were found for both atropisomers, this was not true for **15**, where poses were only found for the 'A' product. This may be due to the slight difference between the native binding modes for the two atropisomers and the presence of the bulky alkyne group. As can be seen in Figure 5.8, the alkyne group forces the diene end of the molecule into a position in which only the native binding mode for the 'A' product is supported. This is an interesting result, as docking suggests a significant structural difference in how the substrate needs to bind to form the two atropisomers and could mean selectivity towards the 'A' product if the native-like binding mode is still favoured for this substrate. Again, the alternative pose for the 'A' product represented in Figure 5.8 was selected for screening. Finally, it should be mentioned that, although the chosen poses are not always the most highly ranked in terms of the 'binding affinity' docking score, scores of alternative poses are typically similar and docking scores are known not to be very reliable (e.g. due to the absence of conformational sampling). To determine if other poses are likely to compete with the chosen poses would require running extensive MD and binding affinity studies. Nevertheless, the docking scores for the selected poses are shown in Figure 5.9 for reference and later comparison to the binding affinities calculated with MM/GBSA. Although these scores are a very approximate measure of binding affinity, they can however already give a strong indication that W124A is a detrimental mutation due to the consistent reduction in affinities this mutation leads to.

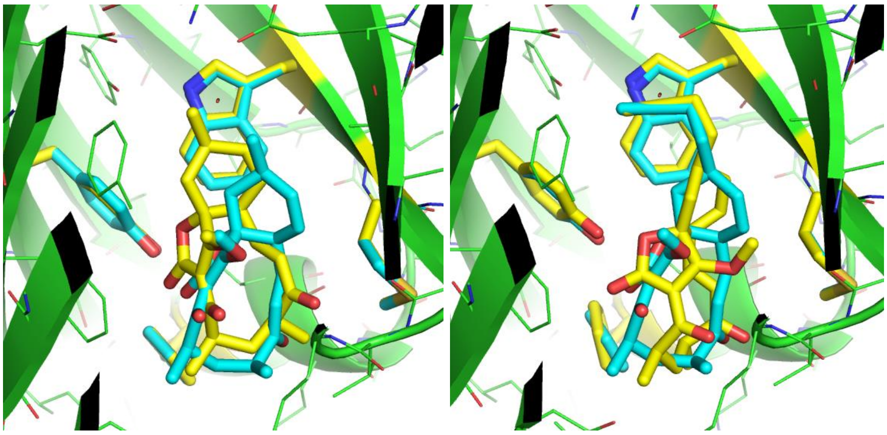


Figure 5.8: Comparison between docking poses of the two atropisomers for the native reactive binding mode (left) and the most similar poses selected for screening of **15** with the WT enzyme (right). 'A' products are coloured in cyan and 'P' products in yellow.

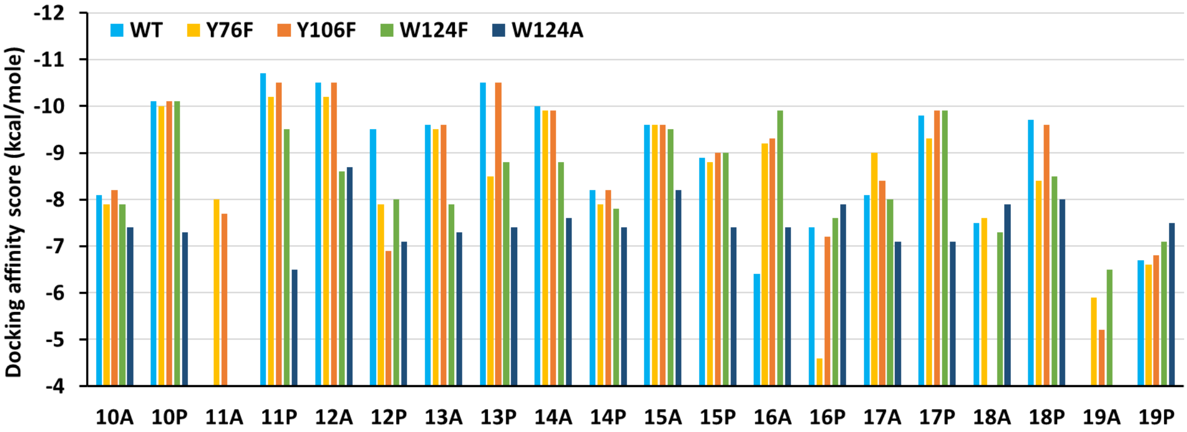


Figure 5.9: Docking scores of the product poses selected for screening of substrates **10**-**19** with the WT enzyme and point mutants.

5.3.1.2 Reaction barrier and binding affinity predictions

Results of the reaction barrier and binding affinity prediction protocols for all substrate/enzyme variant combinations are shown in Figure 5.10. The different substrates are generally predicted to be fairly reactive with multiple enzyme variants as, at least for one atropisomer, barriers can be found within 2-3 kcal/mole of that for the native reaction. Given that these results are, in most cases, for a broadly analogous pose to the native binding mode, this may indicate (assuming similar intrinsic reactivities) that the reactivity conferred by the native binding mode is more or less retained for the alternative substrates (potentially with assistance from a given mutation in the cases where mutant barriers are lower). This may have been anticipated given that the reactivity of this mode seems to stem from it supporting the substrate in a reactive 'near attack conformation' with short C-C distances, i.e. there are no specific catalytic interactions that a change of substrate may disrupt. However, of the substrate/mutant combinations that fall within this 2-3 kcal/mole range, a strong prediction cannot be made as to which should be the most active due to the errors involved. When looking for mutations which may improve activity a similar problem is encountered, in that differences are generally too small to be deemed significant given the error. Only in cases like product **16A** where the barrier is very high to begin with, is there an indication that mutation may improve activity. Although for other substrates analogous poses were generally found (indicating the active site was already fairly optimal), this was not the case for **16**. This may explain why it would be easier to find enzyme variants with predicted improvement in reactivity. Despite the general lack of significant differences between mutants, there is one clear trend that emerges: W124A increases reaction barriers compared to the WT enzyme for the poses analogous to the native binding mode (ignoring **15P** and **16A**/**P**). Comparing this with the differences between average reactive distances during simulations of the substrate complex (Figure 5.11) shows that this is the likely culprit in many cases, and that Trp124 is thus important in forming a reactive conformation for the 'native like' binding mode. Looking at the binding affinity results, a similar issue is encountered as for the reactivity screening, where there are broadly similar values for most substrates/variants (indicating that the binding strength in the native binding mode is not significantly disrupted by changes to the substrate) but the differences are too small to make any strong conclusions. Although for substrates **15**, **16** and **19** higher affinities can be found, relative binding affinities for such different substrates may be unreliable due to the approximate nature of the MM/GBSA method. It is thus safer to only compare binding affinities for the same or very similar substrates. In terms of the effect of mutation, W124A is again predicted to negatively impact the substrate binding affinity and thereby the overall catalytic activity.

Comparing the predicted substrate affinities with the previously calculated docking scores (Figure 5.9) shows that, other than the fact that W124A has an almost universally negative impact on the affinity, there is little agreement between the two. This highlights the limitation of using docking scores for affinity estimates, mostly because they are evaluated for a single structure and therefore do not consider any conformational sampling of the system. As an example, the docking scores predict the binding of the A product of **10** to be significantly stronger than the P product. Meanwhile, this stark difference is not reflected in the MMGSBA energies. This is likely because the static structures identified by docking, wherein the pose for the A product exhibits stronger interactions, are not very representative. In reality, the system will not simply remain in these idealised poses but will visit various conformations due to thermal fluctuations. Therefore, as seen here the real-world benefit of the improved interactions seen in a single static structure may be dramatically overestimated. However, the fact that the docking does correctly predict the effect of W124A is noteworthy, as it shows that even a crude docking score evaluated for a single structure can identify a poor mutation if it involves a large-scale and systematic change to the binding.

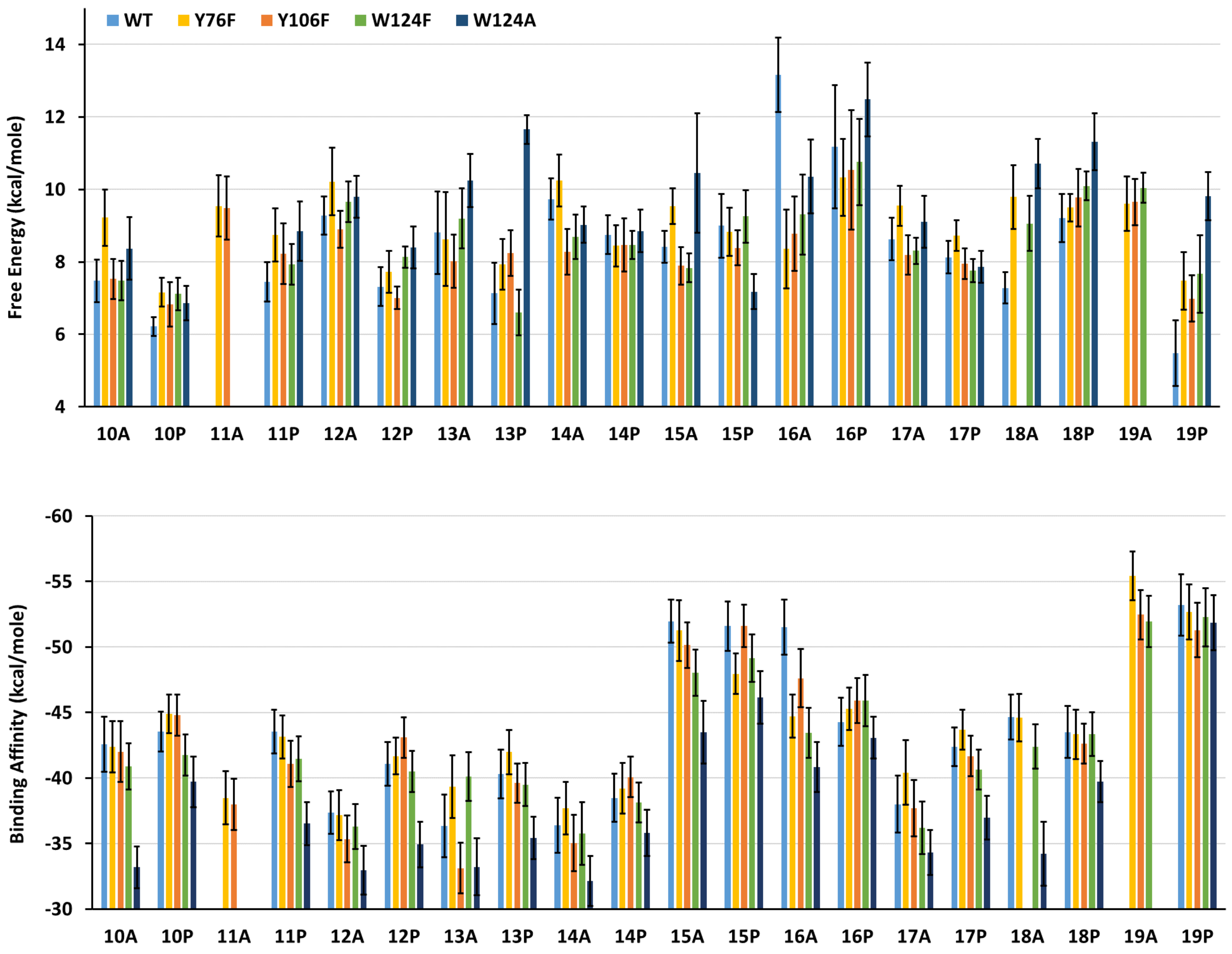


Figure 5.10: Combinatorial screening results for the two atropisomer products of substrates **10**-**19** with the WT enzyme and point mutants. Top - Predicted Diels-Alder activation free energies (obtained from 10 independent QM/MM umbrella sampling molecular dynamics simulations using SCC-DFTB/ff14SB). Bottom - Average binding energy of the Michaelis complex (as calculated

using the MM/GBSA approach using 1000 snapshots of the complex taken from 10 independent MD runs). Standard deviations are indicated for binding and activation free energies. See Chapter 4 Methods section for details.

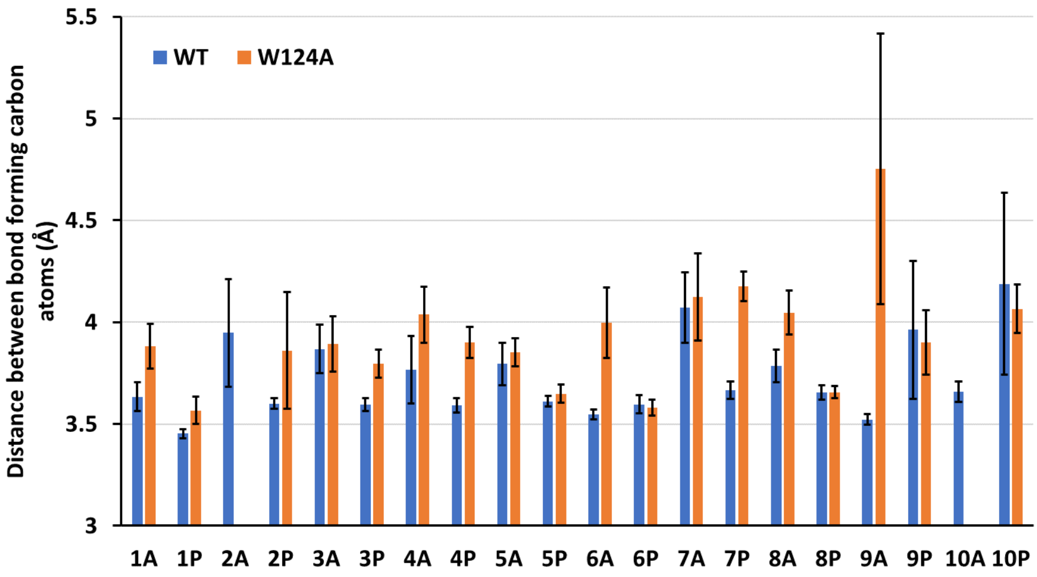


Figure 5.11: Difference between the average reactive distances seen for the Michaelis complexes of the products with the WT enzyme and W124A mutant. Averages were calculated from the 1000 snapshots which came from the 10 independent MD simulations of the complex. Error bars show the standard deviation in the averages for the individual runs.

The lack of significant differences in predictions is compounded by the lack of quantitative experimental results available to compare the predictions to. However, experiments have found that both the native substrate **10** and alternative substrates **14** and **17** are all turned over to a similar degree by the WT enzyme and tested point mutants (all except W124A) as indicated by the height of product peaks after a fixed length incubation time. Although using peak heights to discern turnover is approximate (and provides no kinetic information to determine k_{cat}/K_m), these results are consistent with the indications from simulation. Without proper experimental results for comparison and the fact that differences are too small to be significant on their own (other than the few exceptions mentioned) all that can be taken from these predictions is that most substrates will likely be accepted by most variants and turned over at some nominal rate. Therefore, the results do not narrow down the possible combinations to test experimentally (as was intended), apart from excluding the W124A mutant. The main reason that difficulty has been found in narrowing down active combinations is that, given the similarity between substrates and the nonspecific nature of the enzyme, more sensitive protocols are likely required to discern the resulting small differences in reactivity/affinities. A more effective initial use of these approximate protocols may thus have been to look at more drastically different substrates where, as was found for **16**, binding is less optimal and therefore mutations which improve activity may be more feasibly identified. However, the results obtained here could in future be correlated with detailed kinetics results from experiment to see how well they predict the relevant experimental observable and refine them accordingly (see Figure 5.3). For example, if it were found that predicted activation energies agree well with trends in experimentally obtained k_{cat} values, this would indicate that efforts would be best focussed on improving the binding affinity protocol to obtain greater predictive capability. Having gained more confidence in their predictive power, the protocols could then be used in subsequent screening efforts. For example, as depicted in Figure 5.3 they can be used as part of an iterative design process, where additional mutations are suggested to further optimise the enzyme (e.g. where a substrate-mutant combination is promising in terms of its reactivity but not binding affinity, further mutations could be proposed and tested in further rounds of screening that specifically target binding).

5.3.2 Rosetta redesign

5.3.2.1 Design results

The 3 designs which were selected for testing for the two products are shown in Figure 5.12 with the input WT pose for reference. The mutations made for **14** have not obviously increased packing around the bottom of the active site to compensate for the methyl groups lost from the substrate. However, it may have been that, by subtle changes to substrate binding orientation, fairly optimal packing was already achieved for this binding mode with the WT enzyme. This would imply that there is not much room for improvement with this substrate. This may be further indicated by the fact that, although 100 designs were generated, only 24 of these had unique sequences. Interestingly, all of the generated designs had 2 long flexible residues in the binding site (ignoring the less relevant residue Arg125 accidentally included in the design region) which were generally methionines and occasionally lysines. This is also consistent with there not being much room for improvement as there are many ways these residues can be fit into the active site, which therefore allows optimisation by achieving a highly complementary binding site. The issue with this is that the binding of the ligand is then not well defined given that these are such mobile residues and so this may not actually translate to stronger binding. Overall, the design suggestions do not seem particularly promising based on these observations, however one potentially useful mutation which was discovered is the introduction of a Serine in DE_96. This provides an H-bond interaction which will only exist for the 'A' atropisomer and therefore offers potential for engineering selectivity towards this product.

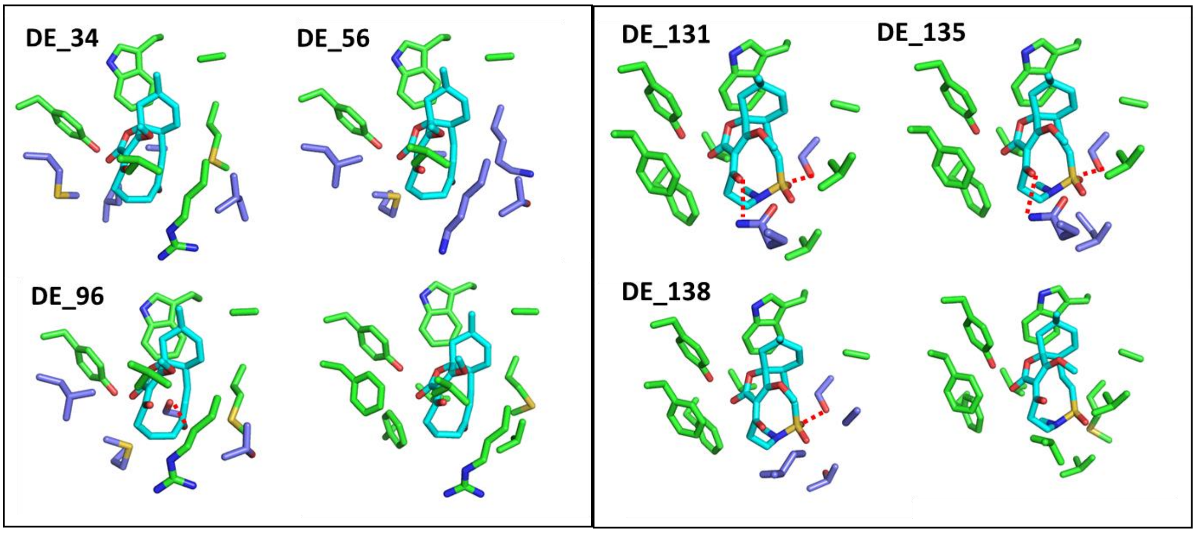


Figure 5.12: Binding sites of the redesigned variants for substrates **14** (left panel) and **19** (right panel) which were selected for testing alongside the input WT complex (bottom right in each panel). Designable residues, along with the key Tyr and Phe residues which were maintained, are displayed. Mutations made by Rosetta are coloured in purple (new H-bonds introduced are shown in red).

For **19**, the design process was able to introduce a new H-bond with the sulfonyl group as anticipated. Note: although not pictured in Figure 5.12, there was an existing H-bond to the sulfonyl group in the WT which is provided by a backbone amino group. This is still present in the redesigned active sites, but an additional H-bond to the second sulfonyl oxygen is now also provided by mutating the methionine to a serine. Additionally, for designs DE_131 and DE_135 a Glutamine has been introduced which H-bonds with the first carbonyl oxygen in the chain. The fact that 2 new H-bonds have been introduced indicates that, contrary to for **14**, there may well be good optimisation potential for this substrate. Perhaps a further indication of this is the fact that the optimisation procedure was able to discover more unique design sequences for this substrate (60 designs out of the total 150 had unique sequences).

5.3.2.2 Screening of designed variants

Reactivity predictions were obtained for the redesigned variants since this was part of the overall screening workflow (Figure 5.13). It should be noted that these were found using the original centre of mass reaction co-ordinate which is less reliable (see section 4.3.2). The predicted reaction barriers are approximately the same as for the WT, although no significant differences are found either way. This is not surprising since the Enzyme Design application itself just optimises for protein-ligand interaction energy, as well as overall complex energy. In order to ensure good activity, it is generally up to the user to specify the necessary catalytic interactions. Or, if the active site were optimised around a transition state then this may also be a way to specifically stabilise that state and thus improve activity. In this case, as activity for the WT enzyme is mostly the result of enforcement of a

Michaelis complex with short reactive distances, it is possible that optimising for the product could have improved activity since the two are structurally similar. However, if this is occurring, the difference is clearly marginal and thus too small to be deemed significant with these protocols.

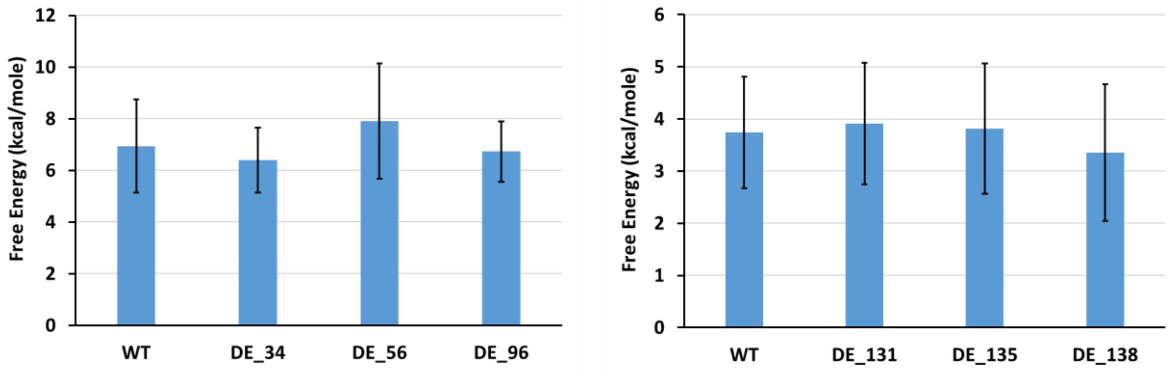


Figure 5.13: Predicted reaction barriers for substrates **14** (left) and **19** (right) with WT enzyme and redesigned variants. Obtained from 10 independent QM/MM umbrella sampling molecular dynamics simulations using SCC-DFTB/ff14SB with the original centre of mass reaction co-ordinate. Standard deviations are indicated - see Chapter 4 Methods section for details.

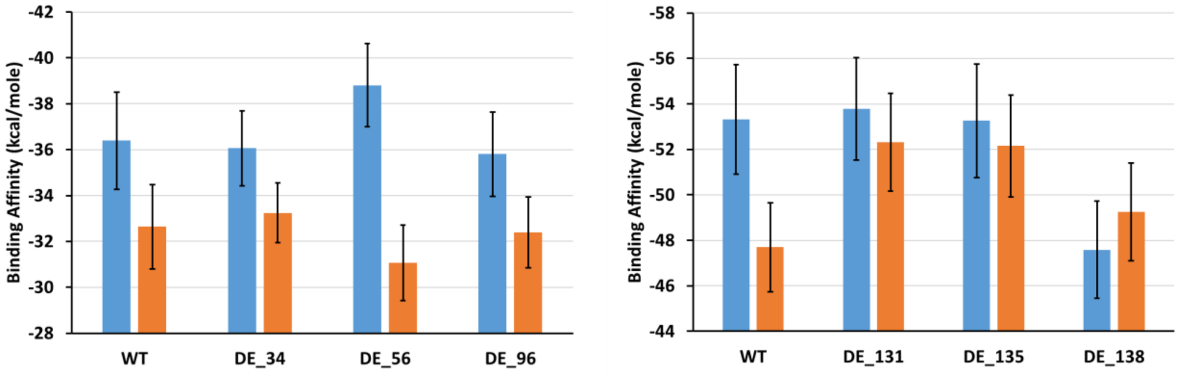


Figure 5.14: Average binding energies for the substrate (blue) and product complexes (orange) for substrates **14** (left) and **19** (right) with the WT enzyme and redesigned variants. Calculated with the MM/GBSA approach using 1000 snapshots of the relevant complex taken from 10 independent MD runs. Standard deviations are indicated - see Chapter 4 Methods section for details.

Unfortunately, the predicted substrate binding affinity has also not been significantly increased for either substrate as was hoped (Figure 5.14). For **14**, the product binding affinity, which the design process specifically targeted, has also not been significantly increased. Although in theory designed structures should have been generated which interact more strongly with the product, as shown this does not necessarily translate to improvement for a more realistic measure of binding affinity which this then incorporates conformational sampling (rather than just the interaction energy for a single optimal structure), influence of solvent etc. This will be especially true if the optimisation gains were small to begin with, or if they rely on a very specific conformation of highly flexible residues to achieve high complementarity, as is likely the case for this substrate. Furthermore, in terms of real-world binding affinity the situation may be even worse than this, since fixing of the highly flexible sidechains introduced by the designs in the bound state may mean binding incurs a further entropic cost (which will not be reflected in the MM/GBSA energies which do not, as-is, account for solute entropy changes). Overall then, it seems there is simply too little room for improvement with this substrate to make the confident prediction of an optimised structure possible. For **19**, although substrate affinity has not increased, the optimisation process itself has been more successful as, for the 2 designs which introduced 2 new H-bonds (DE_131 and DE_135) there has been a fairly significant increase to the predicted binding affinity for the product. Interestingly, although DE_138 was scored higher by Rosetta in terms of interaction energy than DE_135, the additional H-bond is clearly far more important when it comes to a more realistic measure of binding affinity. This therefore stresses the importance of basing selection on features like this as well as general chemical intuition rather than just relying on Rosetta energy scores. These types of qualitative criteria were also used effectively for filtering down the Rosetta redesigned AspB variants for experimental testing (101). Although the Rosetta optimisation performed well for the design ligand, as substrate affinity has not increased for **19** the overall design goal has not been achieved. Indeed, as product release may be rate limiting for this enzyme, increasing product affinity alone may be detrimental to activity. Overall, this shows that the approach of optimising for the product in order to improve substrate binding affinity (and reactivity) is probably too naïve and simplistic, and future efforts should therefore focus on redesigning specifically for the substrate (or transition state). Finally, despite the lack of success, these results still highlight the advantage of using computationally efficient *in silico* screening for designed variants, as poor designs can already be ruled out by this before proceeding to more expensive and time-consuming experimental testing. This is similar to the strategy employed in (102) which found that running many short MD simulations (thus providing efficient sampling) for a set of Rosetta designed structures allowed highly stereoselective variants to be discovered with minimal experimental testing.

5.4 Conclusions

In this chapter, two different approaches have been taken to explore the potential of AbyU to be used as a starting point for general spirotetronate synthesis. In the first approach, combinatorial screening of point mutants and alternative substrates has been performed using an efficient computational activity screening workflow to narrow down combinations by their predicted activity. Apart from excluding the W124A mutant (as likely being detrimental to catalysis), the results of the initial screening were mostly inconclusive. This likely stemmed from the fact that alternative substrates did not depart strongly enough from the native substrate such that, given the approximate protocols, the resulting activity differences were too small to be reliably identified. Future work should therefore focus on experimental validation of the predictions and refining of the individual protocols as necessary such that more confidence is gained in them. The improved workflow can then be used as part of an iterative design process, selecting specific substrate/enzyme variant combinations to be taken forward for successive rounds of mutation and *in silico* screening. It is worth considering how the protocols may be improved to this end, by resolving some of their current limitations. Improving predictivity for the substrate and enzyme variants currently considered (which mostly involve subtle changes to the native reaction) will likely require improving their accuracy by putting less focus on high efficiency and throughput, thereby using more sophisticated protocols and improving confidence in subsequent predictions. Firstly, increased sampling could be performed for both the reactivity and binding affinity protocols to ensure convergence has been achieved in the predictions. In terms of binding affinity calculations, improvements could be gained by for example adding an entropy estimate into the calculation. The reactivity prediction protocol could also be improved by obtaining energy barrier corrections for individual substrates with higher level DFT calculations. In this way there would be more confidence in the different intrinsic reactivities of the various substrates being properly reflected, and the approximate semi-empirical method would only be relied on to resolve changes to the reactivity based on the environment. Another current weakness is the lack of consideration of other (potentially unproductive) binding modes and the assumption of analogy in the most reactive mode to that of the native reaction. As seen in chapter 4, with this enzyme there can be competitive binding with other modes, which may be less productive and thus inhibit activity. How much of an effect this has on the activity may also be different for different substrates/enzyme variants, thereby confounding predictions. Further, the assumption of an analogous reactive mode to the natural reaction may not always be correct. This is particularly so when less trivial changes to the substrate are made like with **16** for example, where it was found that a closely analogous mode is no longer possible. In this case there is therefore far less certainty in selecting a pose based on similarity to the native reaction. For these reasons, an improved activity prediction protocol should most likely incorporate some consideration of alternative binding modes.

In the second approach to expand substrate scope, the biding site was rationally redesigned with the aim of improving binding for a specific substrate, the same screening protocols being used to test redesigned variants. To an extent this suffers from the same issue as the first approach, namely that binding is already fairly optimal and provides too little room for measurable improvement. For the substrate that differed most from the natural one, more promising designs were obtained, indicating that this approach may be useful for further alternative substrates. However, the results also indicate that the approach of redesigning based on a product-complex to improve productive binding of the substrate may not necessarily lead to designs with higher turnover. Despite the lack of initial success, these results do serve to highlight the potential of approximate binding affinity calculations for evaluating the outputs of computational protein design by reducing the amount of experimental testing required. However, as described above, improving accuracy of the protocols would also be advantageous to help yield more significant and reliable predictions when optimising for similar substrates (where only subtle improvements may therefore be realisable). In particular, the lack of an entropy consideration in the binding affinity protocol may be a drawback when performing complete active site redesign. As seen, the design process can strongly change the characteristics of the active site, by introducing more conformational flexibility for example. Therefore, in these cases the assumption of a similar entropy change upon binding as the baseline WT complex, may become less valid. Further, lack of consideration of competing binding modes could also be an issue here. Although, in this approach binding is being optimised for a specific binding mode, there is no guarantee that other less productive modes will not also be promoted. The existence of other binding modes which may interfere with activity could therefore also be investigated when attempting to predict activity for redesigned enzyme variants with the developed protocols.

Chapter 6. Comparison of stereoselectivity in the homologous Diels-Alderases AbyU and AbmU

6.1 Introduction

AbmU is a recently discovered homologue of AbyU (104) which performs the analogous role within the abm biosynthetic pathway from *Streptomyces koyangensis*. The end products of this pathway are Abyssomicins 2 and 4 and neoabysomicins A and B, which are all spirotetronates similar to Abyssomicin C. Like AbyU, AbmU catalyses the spirocyclisation step that converts the linear substrate consisting of the polyketide chain/triene attached to a tetronate ring into the resulting spirotetronate product. As with AbyU, it does this via a Diels-Alder reaction between the end diene and exocylic methylene of the tetronate (Figure 6.1). The substrate is very similar to the AbyU substrate, the only differences being a methyl group added at C12 and one removed from the polyketide chain. The AbmU product then goes on to form the final end products via several oxidation and reduction steps thought to be carried out by various putative tailoring enzymes that have been identified within the pathway. In particular, the neoabyssomicins are proposed to undergo a biosynthetic Baeyer-Villiger oxidation which inserts an oxygen in the polyketide chain and thereby separates them into a new subclass of abyssomicins (104).

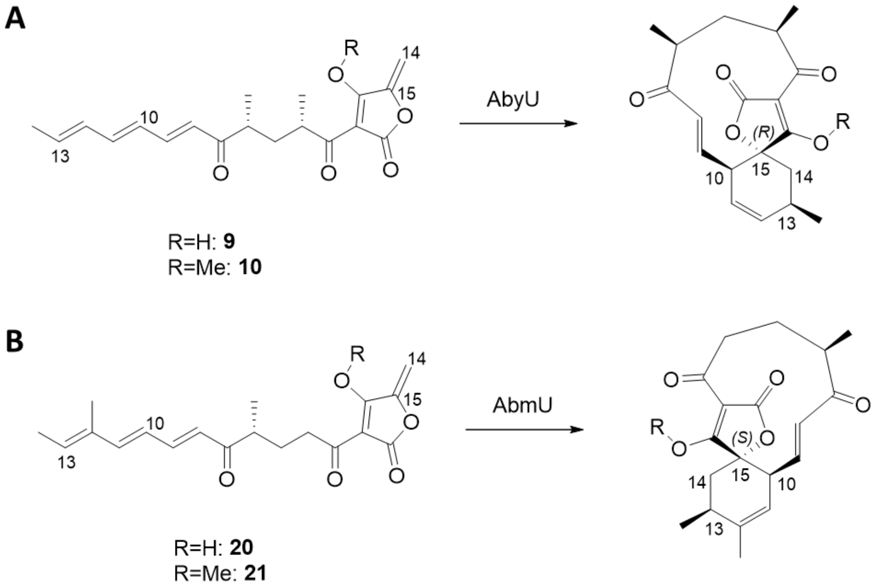


Figure 6.1: Comparison of the Diels-Alder reactions of AbyU (A) and AbmU (B). A) Natural AbyU substrate **9** is cyclised into spirotetronate product with *R* stereochemistry at C15 via the Diels-Alder reaction. B) Natural AbmU substrate **20** is cyclised into spirotetronate product with *S* stereochemistry at C15 via the Diels-Alder reaction. Methoxy analogue substrates **10** and **21** are also indicated which were used in simulations/experiments.

As well as the structural differences due to the difference in the Diels-Alder substrate and post-cyclisation tailoring steps, the products of the abm pathway bear the opposite stereochemistry to abyssomicin C at C15. The abm products are therefore defined as Type II abyssomicins in contrast to Abyssomicin C, which is a Type I abyssomicin (24). Since this stereochemistry is set during the Diels-Alder step this means that the 2 enzymatic reactions have different stereoselectivities, where AbyU yields the *R* product and AbmU yields the *S* product (Figure 6.1). Understanding how these different product outcomes are achieved is highly valuable from an industrial biosynthesis perspective. In the first place it is of interest to discover how these spirotetronate-forming enzymes function in general, since they typically produce high value compounds which can be difficult to synthesise chemically. This is particularly relevant for antibiotic resistance, where Abyssomicin C for example has shown good antibiotic potential (see section 1.2.2.2.2.1), and where there is thus a drive to develop Diels-Alderases which can turn over ever more diverse substrates. In Chapter 4 and 5 the catalytic ability of AbyU was therefore studied and its ability to turnover modified substrates explored. However, it is equally important how stereochemical control is achieved in these enzymes as only one of their possible products is typically relevant for a given application. Since AbyU/AbmU have opposite stereoselectivities, comparison of these enzymes may therefore provide valuable insight into how to design highly selective biocatalysts for industrial spirotetronate biosynthesis.

The structure of the apo AbmU enzyme has now been solved using X-ray crystallography[6] (Figure 6.2). The overall structure shows very close resemblance to AbyU; both enzymes are 8-stranded β-barrels with a central active site cavity and an active site capping loop between strands 1 and 2, as well as a salt bridge blocking the top of the cavity. One difference is that for AbmU there also exist N- and C-terminal strands/helices attached to the barrel which make up the interface between subunits in the dimer. While AbyU also exists as a dimer, the two subunits instead interface directly at the barrel surface.

[6] C. Back & P.R. Race, personal communication

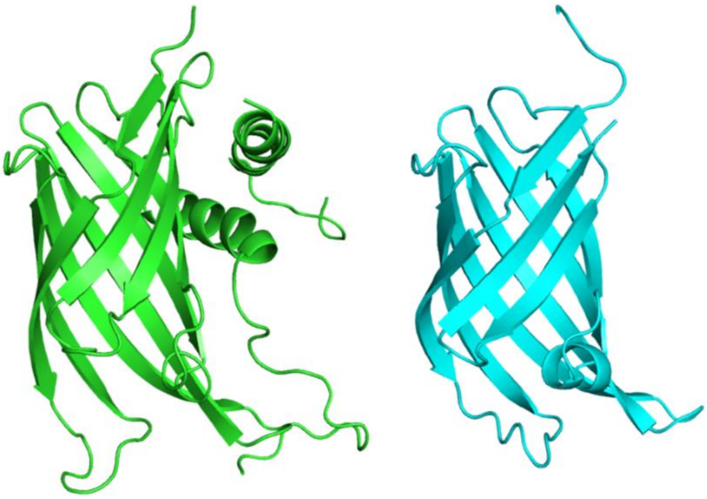


Figure 6.2: AbmU (green) and AbyU (cyan) crystal structures. Chain A of the dimer is shown for AbmU where the structure was solved everywhere except for the region between the N-terminal strand and helix.

Because of the strong structural similarity between the two enzymes, AbmU is expected to act via a similar mechanism, namely that the enzyme captures the substrate in the folded reactive conformation in order to facilitate the reaction, where the capping loop serves to mediate active site access. This suggests a possible cause of the opposite stereoselectivity of AbmU being changes in the enzyme active site which lead it to favour the binding of a pro-*S* conformation of its substrate. The element of the reactive conformation that dictates which of the two products will form is the direction in which the polyketide chain is folded, as this then determines the resulting arrangement of atoms about the C15 in the product (Figure 6.1 and see also Figure 3.6). As this is not a subtle difference it would seem that the binding site would need to be set up quite differently to accommodate this change. Overall, the binding pockets are quite similar for both enzymes (Figure 6.3); the key Tyr76 in AbyU is preserved (Tyr104 in AbmU) and two other residues have simply swapped between Tyr and Phe thereby maintaining a similar shape. However, there are still several key differences that alter the shape and electrostatics of the pocket which could therefore lead to a different reactive conformation being favoured. Studying the substrate complex alone for the different reactive conformations may therefore reveal key insights into how exactly the different stereoselectivities are achieved.

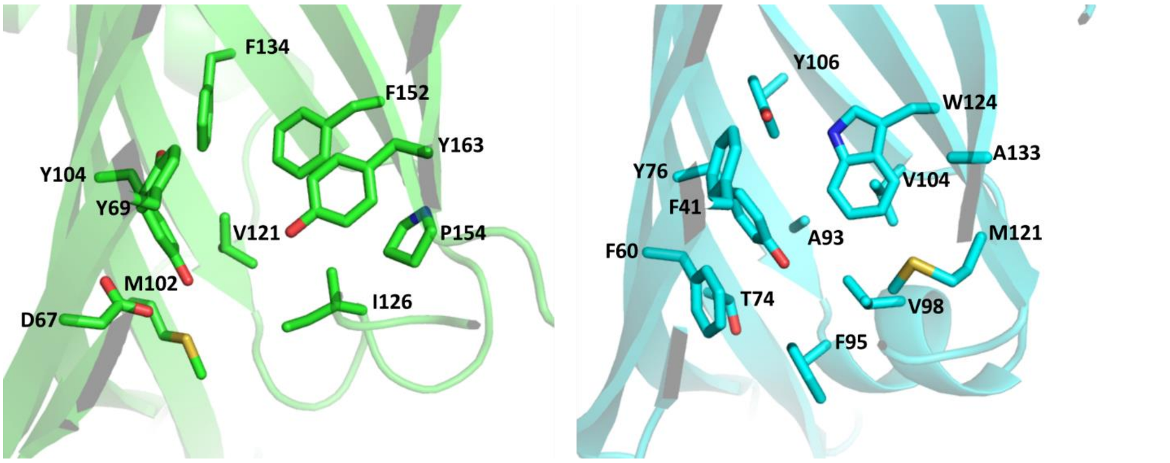


Figure 6.3: Comparison of active sites in the AbmU (green) and AbyU (cyan) crystal structures. Sidechains of the respective residues which line the active site cavity are shown as sticks.

Also of interest, since the substrates of the respective enzymes are slightly different, is whether the difference in stereoselectivity solely arises from changes to the enzyme, or if the difference in the substrate itself contributes to this. To investigate this, experiments were performed by the Race/Willis groups where they fed each enzyme with the substrate of its counterpart. However, rather than the true native substrates the methoxy analogues **10** and **21** were used where the hydroxyl group on the tetronate ring is replaced with a methoxy group (Figure 6.1). Since, strictly speaking, these are not the native substrates, they were also fed to their native enzyme as a control, but it was expected that the stereoselectivity would be the same as for the true native substrate i.e. as depicted in Figure 6.1. Indeed, for the experiments feeding the native substrates, a single product was formed in each case. For AbmU, this was identified by NOE assignment to be the *S* product as expected. For AbyU, although the product has not yet been formally identified, this is strongly assumed to be the *R* product, in line with the assignment of Abyssomicin C as a Type I abyssomicin (68) (the general *R* selectivity exhibited by AbyU is also a very strong indication that product outcome is unchanged for the methoxy analogue substrate – see below). In both cases of feeding the non-native substrate it was turned over showing firstly that each enzyme can at least accept the substrate of its counterpart. When **21** was fed to AbyU, a single product was obtained and this was identified by NOE assignment to be the *R* stereoisomer as is also formed with the native substrate. This shows that AbyU is generally selective towards this product outcome. Considering that the change from a pro-*R* to pro-*S* conformation involves a drastic change in substrate conformation, it makes sense that, even with some changes to the substrate, the active site could still be better set up to form the *R* product. When **10** was fed to AbmU, two products were formed. These are assumed to be the *R* and *S* stereoisomers, although they have not yet been formally identified. If this is the case, then it shows that AbmU selectivity is more dependent on specific interactions with its native substrate as, although it confers some change in outcome to the AbyU substrate, the original product is still formed to a large extent.

To obtain insight into the stereoselectivity in each case, substrate binding in both pro-*R* and pro-*S* reactive conformations has been studied via docking and MD simulations. In addition to probing the origin of stereoselectivity (thereby providing insight into to how this can be achieved artificially), this will demonstrate how computational simulation can be used as a tool to predict stereoselectivity and thus aid in the design of stereoselective spirotetronate variants. The first factor investigated is the reactivity in different prochiral conformations, where it is possible that the stereoselectivity is a result of the enzyme enforcing a more reactive substrate conformation when it binds the conformation that would yield the observed product. It has been shown that an enzymatic Diels-Alder reaction can be favoured by forming a reactive conformation with shorter distances between the relevant carbon atoms of the substrate between which the new bonds form (86). This is essentially because this brings the energy of the substrate minimum closer to the transition state and thus lowers the activation energy of the reaction. This effect can be both steric and electronic as shorter C-C distances were also shown to correlate with a lower HOMO-LUMO gap which contributes to the electronic part of the activation energy. More directly relevant, in Chapter 4 it was found that for the different binding poses of AbyU, reaction free energy barriers obtained from umbrella sampling correlated well with the average carbon-carbon distances seen in MD simulations of the substrate complex (ignoring the case where the distances seen in the simulations of the substrate complex were greater than those sampled in the umbrella sampling). Thus, reactive distances during MD can be seen to provide a reasonable descriptor of reactivity, and therefore offer a simpler alternative to performing actual reaction simulations. As such, substrate reactivity has been explored here as a potential cause of stereoselectivity via simulations of the substrate complex alone. In order to do this, candidate pro-*R* and pro-*S* binding modes were first obtained via molecular docking. MD simulations were then run for the selected binding modes and the average C-C distances for pro-*R* and pro-*S* poses compared to see if this could explain the observed stereoselectivities. Another factor which may be playing a role in stereoselectivity is the affinity with which the two enzymes bind pro-*R* vs pro-*S* conformations of the substrates, and so the binding affinity was also calculated for each binding mode to see if this gave any further insight. Finally, it should be noted that, for the reaction where the products have yet to be formally identified, these simulation results first serve to validate the assumed product outcome.

6.2 Methods

6.2.1 Enzyme structure preparation

For AbyU, the coordinates of the enzyme were taken from the crystal structure, removing the buffer molecule HEPES bound in the active site (PDB ID: 5DYV, chain A).

The AbmU structure was determined by X-ray crystallography by members of the Race lab. For two chains (A and B) the entire barrel structure including the capping loop region were solved, the only missing density being between the N-terminal strand/helix and barrel where a break occurs in the chain (Figure 6.2), which is unimportant for the purpose of these simulations. As only one chain is required for simulations of ligand binding, chain A was used. However, the capping loop is in an open state in the crystal structure (in both chains), such that the cavity is accessible to solvent. As the catalytically active state is presumed to be the same as for AbyU and therefore with the capping loop closed, a structure was firstly obtained with the loop in a closed conformation. Initially, this was done by using Targeted Molecular Dynamics (TMD) to steer the loop backbone structure to that of the equivalent closed loop in the crystal structure of AbyU. However, due to differences between the crystal structures, in order for the loop to adopt this conformation this meant that the strand preceding the loop also required a shift in register. This introduces a potentially unrealistic change to the residues in the active site. Therefore, a second model of AbmU with a closed loop was obtained by loop remodelling. Molecular docking and dynamics simulations were performed for both models, allowing direct comparison.

6.2.1.1 Targeted Molecular dynamics

Reference co-ordinates for a closed conformation of the loop were obtained by building a homology model of AbmU based on the AbyU crystal structure for which the loop is closed. The homology model was generated using I-TASSER (105) by inputting the AbmU sequence and accepting the default options. I-TASSER searches the PDB automatically to find suitable template proteins for the model building. The template it selected for use in the threading alignments (from which the final models are then generated) was AbyU (PDB accession code: 5dyv). The best alignment showed 27% sequence identity where the entire AbyU sequence minus two residues in the loop connecting strands 7 and 8 was used for the alignment. The model with highest confidence (in terms of C-score) was then used. This bears very close structural similarity with AbyU (Figure 6.4A), the only regions where the structures differ being the aforementioned terminal strand/helices which are not present in AbyU, and the turn between strands 7 and 8 which is longer for AbmU. While there appears to be a difference in terms of the location of the break in the first strand, and where the penultimate strand begins, these are purely visualisation artefacts due to PyMOL's definition of secondary structure, where in fact the backbones line up well in both cases. Although no residues in the capping loop region are conserved between the two proteins, the loop is still able to be built in an almost identical closed conformation as in AbyU. This provides further evidence that the loop likely functions in a similar way for both enzymes and that the loop is thus required to be closed for catalysis.

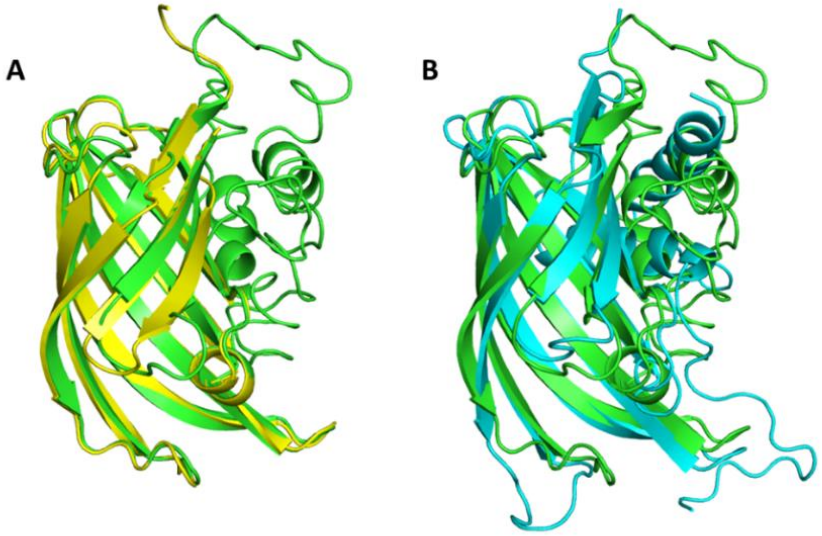


Figure 6.4: Comparison of the AbmU homology model with its template (AbyU) and its crystal structure. A) AbyU crystal structure (yellow) with AbmU homology model (shown in green) after aligning on Cα atoms. B) AbmU homology model (green) and AbmU crystal structure (blue) after aligning on Cα atoms (ignoring terminal strands).

Except for the desired change in the capping loop conformation, the homology model shows reasonably good agreement with the crystal structure (Figure 6.4B). A Cα RMSD of 2.421 was calculated for the two structures after alignment on the Cα's (minus the terminal helices). This is perhaps unsurprising given the good homology with the template and thus the confidence with which it was built. However, in addition to the capping loop itself being different the preceding strand (strand 1) is in a different conformation, and instead adopts the bulge which is seen in the AbyU structure (Figure 6.4A). While the AbmU crystal structure follows the conventional antiparallel structure in this region, in the homology model the pattern is interrupted at residue 50 before then continuing on a residue later (Figure 6.5). This means that, from that point onwards, the strand is shifted up by one residue relative to its neighbouring strand. Given this, in order for the loop conformation of the crystal structure to be driven towards what it is in the homology model, it must begin at the same point and so the part of the strand which differs between the two structures must now also be driven to the conformation that it adopts in the homology model.

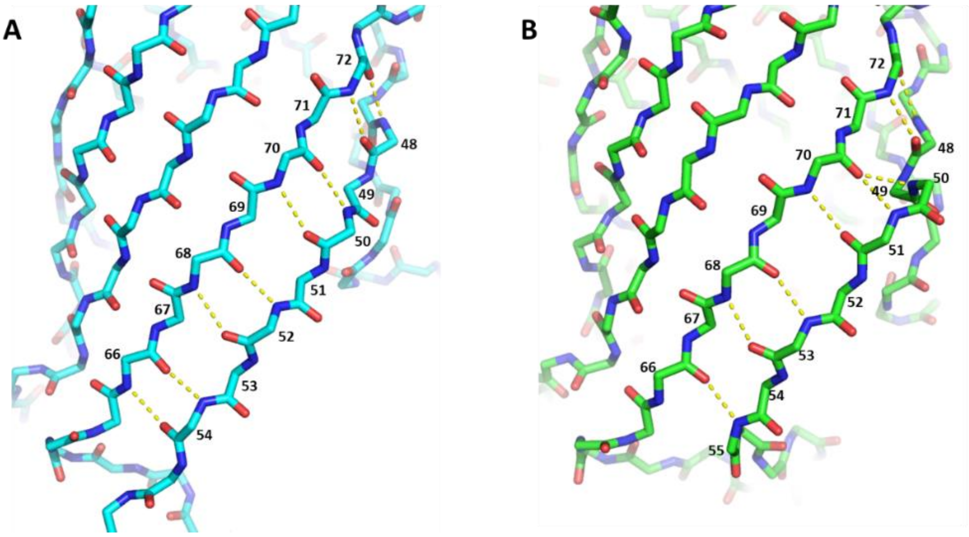


Figure 6.5: Difference in backbone structure of strand 1 between the homology model and crystal structure of AbmU. A) In the crystal structure strand 1 adopts the conventional antiparallel β structure. B) The antiparallel structure is temporarily broken in the homology model at residue 50 such that the remainder of the strand is shifted up by one residue relative to strand 2 (this is also seen in the AbyU crystal structure).

To prepare the crystal structure for Targeted MD, the terminal strands/helices that make up the dimer interface were firstly removed as these are irrelevant for simulations of the monomer. The N-terminal was capped with an Acetyl residue [−C(=O) −CH_3] in place of Arg37 and the C-terminal with a N-CH_3 residue in place of Gly174. Hydrogens were added by the AmberTools program tleap, where all titratable residues were assigned their standard protonation states. A solvent box was then added by tleap with a minimum distance between any protein atom and the edge of the box of 11 Å (and a closeness parameter of 0.75), along with 5 sodium ions to neutralise the net charge.

Prior to performing targeted MD, the structure was equilibrated with the following procedure, using the crystal structure as the reference for positional restraints throughout (simulations were performed with the Amber14 programs sander (for minimization) and pmemd.cuda using the AMBER ff14SB force field (90) for the protein and the TIP3P water model. Particle-mesh Ewald summation was used in conjunction with periodic boundary conditions, and a cut-off for direct-space non-bonded interactions of 8 Å):

1) optimisation of solvent and solute hydrogen positions (minimization for 300 steps with restraints on solute heavy atoms, force constant: 100 kcal mol^{-1}Å^{-2});
2) heating/equilibration of solvent (initial random velocities assigned at 25K followed by 50 ps NPT heating to 300K with a 2 fs timestep using SHAKE restraints, using Langevin dynamics with collision frequency 2 and the Berendsen barostat with a pressure relaxation time of 1 ps. Solute heavy atoms were restrained with force constant 25 kcal mol^{-1}Å^{-2});

3) 300 steps of minimization and quick heating to 298 K (20 ps in NVT ensemble with 2 fs timestep, using Langevin dynamics with collision frequency 1, SHAKE and initial random velocities assigned at 25K. 5 kcal mol^{-1}Å^{-2} restraint applied to solute Cα atoms);
4) 2ns equilibration in the NPT ensemble (Langevin dynamics with 2 fs timestep, SHAKE and collision frequency of 1 and using the Monte Carlo barostat with a pressure relaxation time of 1 ps. 5 kcal mol^{-1}Å^{-2} restraint applied to solute Cα atoms).

Targeted MD was then performed on the equilibrated structure to drive the co-ordinates of the target region to match the co-ordinates of the reference structure (reference was the homology model after alignment to the crystal structure – see Figure 6.4B). The target region was selected to be the Cα atoms of residues 49 – 64 (using original numbering i.e. before removing terminal regions). This region corresponds to the capping loop and relevant region of the preceding strand as previously discussed. The rest of the protein was restrained to the crystal structure, leaving some unrestrained residues between the two regions to allow flexibility since the two structures do not line up perfectly (restrained region was the Cα atoms of residues 37 – 45 and 66 - 174). During Targeted MD the the RMSD of the target region (calculated relative to the reference structure) was driven towards zero to thus converge on the reference structure. Since the restraint energy applied to the target region is determined by the difference between the actual RMSD and the target RMSD, the target RMSD was lowered gradually from its initial value (i.e. at the end of the equilibration protocol) to gently steer the system to the desired conformation. Therefore, the initial RMSD was firstly calculated to determine the starting point for the target RMSD. This was found to be 5.55 Å, and so the target RMSD was lowered from 5 Å to 0 Å in steps of 1 Å, running 300ps of targeted MD at each step to give the system time to reach the target using a reasonable force constant. The force constants used for the restraints applied to the target region and region held to the crystal structure were 25 kcal mol^{-1}Å^{-2} and 15 kcal mol^{-1}Å^{-2} respectively. Targeted MD steps were performed in the NPT ensemble (using SHAKE with a 2 fs timestep, Berendsen temperature control algorithm with the time constant for coupling to the heat bath of 10 ps and the same pressure regulation as for the final equilibration step). At the end of the targeted MD protocol the target region had effectively converged on the reference structure with an RMSD of 0.458 Å (Figure 6.6). This structure was then minimised for 300 steps, and the solvent and ion atoms were removed to prepare it for docking.

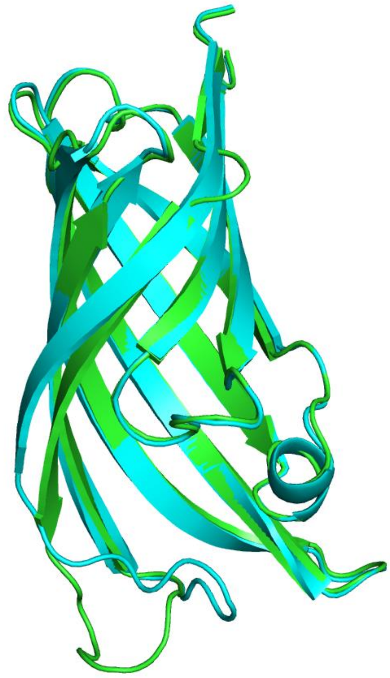


Figure 6.6: Closed-loop structure obtained via targeted MD (referred to as AbmU-TMD) superimposed alongside the starting crystal structure (with open loop) input to targeted MD. The AbmU-TMD structure is shown in blue and the crystal structure in green. The Cα's of the capping loop and partial strand region which can be seen to deviate from the crystal structure were driven to within 0.5 Å of the conformation that they adopt in the AbmU homology model based on AbyU. The remainder of the structure was restrained to the starting crystal conformation on the Cα's.

6.2.1.2 Loop remodelling

To obtain a closed loop structure without impacting the barrel, automated loop modelling and refinement was performed using Modeller (106). Although Modeller is more generally used for comparative modelling in order to generate a homology model for an entire protein, it can also be directed to just model a specific region into an existing structure. Rather than using a template structure for the modelled region it can also be used to generate an entirely *de novo* structure. This is useful for insertions where no template exists or flexible regions like loops where comparative modelling is less reliable. Modeller therefore also has dedicated scoring functions and optimisation protocols which have been adapted for loop modelling (107). In order to obtain a closed loop structure, the crystal structure was thus input to Modeller and loop modelling/optimisation was requested for the capping loop region which was defined to be from Phe55 to Gly64. As before, the terminal regions were first removed truncating residues before Glu36 and after Gly174 as these are not relevant for the modelling of the capping loop or the following simulations of the substrates in the β-barrel. To ensure that closed structures were obtained, a harmonic upper bound restraint was placed on the distance between the Cα atoms of Pro58 and His160, such that this distance was restrained to be less than 10 Å (with standard deviation parameter 0.1 Å). A dihedral restraint was also placed on the peptide bond between Pro58 and Pro59 to preserve the conformation found in the crystal structure. Modeller was then run generating 1500 loop models which were optimised using the second highest MD refinement level. The loop model with the lowest DOPE score (Figure 6.7) was then chosen for the following simulations.

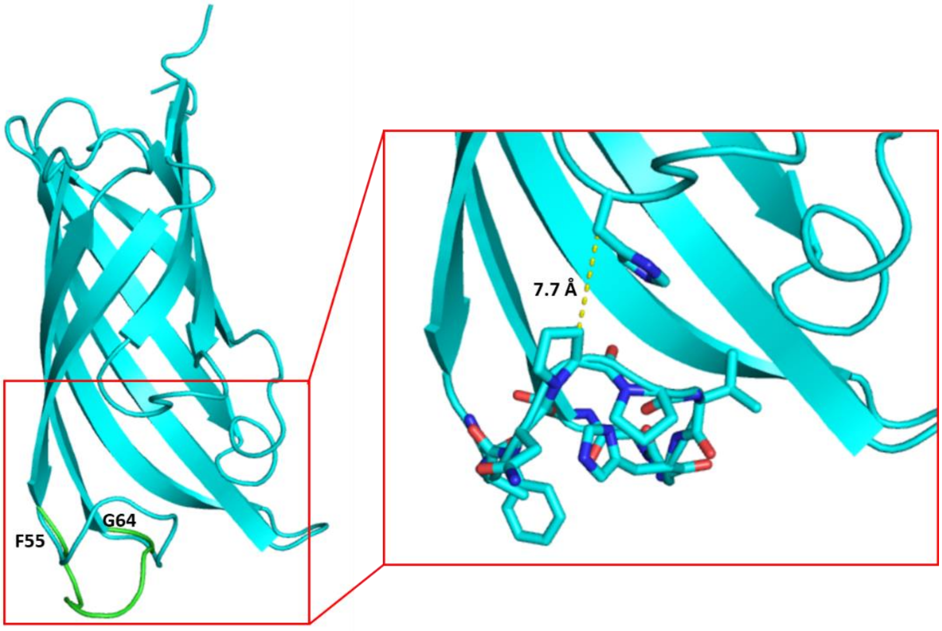


Figure 6.7: AbmU with closed loop conformation obtained via loop remodelling (referred to as AbmU-loop). Original crystal structure conformation where loop is open is shown in green. Start and end residues for remodelled loop region are labelled. Inset – closeup of the remodelled loop region, with remodelled residues and His160 which was used in modelling restraint highlighted with sticks (distance indicated was restrained to be < 10 Å).

6.2.2 Docking

For both substrates, structures were built (in PDB format) of the methoxy analogue since these were used in the experiments. For **21** the pro-*S* and pro-*R* reactive conformations were docked rigidly into the active sites. This was done to ensure that reactive conformations consistent with the products identified by NOE assignment were used for the simulations. Specifically, this was concerned with the prochirality of the additional stereo-centres which form at C10 and C13 as well as the relevant conformation of the second (non-reactive) diene, which is s-*cis* for the pro-*R* product and s-*trans* for the pro-*S* product.

For the docking of **10** with AbyU and the AbmU structure from TMD (AbmU-TMD), flexible ligand docking was performed, fixing only key bond rotations. The substrate structure was created and optimised in Chem3D. As well as the essential reacting diene being in the s-*cis* conformation, the structure was optimised with the second diene also in the s-*cis* conformation as this was assumed to be the relevant conformation for all cases based on the product obtained in the native reaction (inferred from the Abyssomicin C crystal structure (68) – see Figure 3.7). For **10** with the AbmU structure with the remodelled loop (AbmU-loop), rigid docking of the pro-*R* and pro-*S* reactive conformations was performed. This was done for a better direct comparison to the docking for **21** by enabling analogous poses to be found and tested. This was not necessary for AbyU since the reactive pro-*R* pose (known from the work in Chapter 4) was discovered in the flexible docking and sufficient pro-*S* reactive poses were found to test the different binding modes. Likewise, for the AbmU-TMD structure sufficient reactive poses were found to test both pro-*R* and pro-*S* binding modes. The reactive conformations of **10** docked into the AbmU-loop structure were built from and thus analogous to those of the reactive conformations built for **21**. However, the reactive pro-*S* pose was still not found. Therefore, a pro-*S* pose was built manually based on the equivalent pose (including flexible residues) for **21**. As well as this, since the products have not been experimentally identified in the case of feeding **10** to AbmU, a pro-*S* pose was tested for the AbmU-loop structure where the manually docked pose mentioned above was modified, changing the conformation of the second diene (and thus the pro-chirality in terms of the other formed stereo-centres) to be in line with the pose found from flexible docking of **10** with the AbmU-TMD structure.

AutoDockTools 1.5.6 was used to create the input files for each protein/substrate docking by assigning AutoDock atom types to the relevant atoms as well as defining rotatable bonds. Polar hydrogens are required in the input structures of both the ligand and protein in order to determine Hydrogen bond donor/acceptors. Therefore, as the protein crystal structures do not contain hydrogens, polar hydrogens were firstly added in AutoDockTools (utilising open babel) where all were in their standard protonation states. For AbyU, the sidechains of residues Tyr76, Phe95, Trp124 and Met126 (which line the cavity) were treated flexibly; where all formally single bonds were defined as rotatable. For the AbmU-TMD structure Tyr69, Met102, Tyr104, Phe134, Phe152, Tyr163 were made flexible. For the AbmU-loop structure, Tyr69 and Phe134 were removed from the flexible region as it was found that they essentially always had the crystal conformation in the discovered poses. However, two new residues were added to the flexible region which were Asp67 and Ile126. These were required in order to create more space in the cavity and enable the suspected reactive pro-*S* pose for **21** suggested by the AbmU-TMD results to be found. The difference in space between the two structures is mostly due to His160, which encroaches slightly on the active site in the loop-remodelled structure i.e. when it is in its crystal orientation (while for the TMD structure it ended up pointing into solution). Making His160 flexible was tested, but this led to many poses where the substrate was partially outside of the cavity and thus discouraged poses where it was properly docked within the active site.

The standard active torsions detected by AutoDockTools were assigned for **10** when docking with AbyU and the AbmU-TMD structure (where all formally single bonds not part of a cycle are treated as rotatable), except for the two dienes previously mentioned and the first bond in the chain which were all fixed in their starting conformation to ensure that correct reactive conformations were found. For the docking of **10** with the AbmU-loop structure and the docking of **21**, all bonds were fixed since the reactive conformations themselves were docked.

Docking was performed with AutoDock Vina (87) using a search grid centred on the active site of size 18 x 18 x 14 Å for AbyU and 20 x 20 x 18 Å for AbmU, with an exhaustiveness of 16. At least one binding pose was then selected for both pro-*S* and pro-*R* reactive conformations to test in the following simulations. In cases where several alternative reactive poses were found (e.g. when docking the reactive conformations themselves) the overall orientation of the molecule in the cavity was used as criterion for selection. For AbyU with its native substrate, the most reactive pro-*R* binding mode is already known from previous work on the enzyme and so this was the only pose studied. For the other cases, the first criterion used for selecting a pro-*R* pose was whether the pose was analogous to the known reactive binding mode for AbyU. The basis for this was that this is presumably still the most likely way for the substrate to form the *R* product. The next criterion, used also for selecting pro-*S* poses, was simply whether the diene was at the top of the cavity. Although it is not known *a priori* which the most relevant pro-*S* pose is, it is assumed that it is likely to have the same overall orientation within the active site as the known reactive pro-*R* pose for AbyU (as determined in Chapter 4).

As the poses output by Vina consist of just the co-ordinates of the ligand and flexible sidechains, these were then combined with the rest of the protein to get the complete structure of the complex for each pose. The hydrogens that were added to the protein during the docking setup were left out when making these structures as the protein is then protonated during the simulation setup.

6.2.3 Molecular dynamics and binding affinity

For each of the poses, the structure was prepared and then optimised using the Enlighten protocols PREP and STRUCT and then multiple repeat 100ps MD runs were performed using DYNAM (see Chapter 4 Methods for a description of these protocols). For **10** with AbyU and the AbmU-TMD structure 10 repeat DYNAM runs were performed to obtain 10 independent 100ps MD trajectories (2 fs timestep) of the substrate complex for each pose. For **21** 20 DYNAM runs were performed. For both **21** and **10** with the AbmU-loop structure 10 DYNAM runs were performed. The distances between the bond-forming carbon atoms (C13 – C14 and C10 – C15) were measured every 0.1 ps during the MD and averaged to assess the reactivity of each binding mode. The binding affinity was calculated for each pose using MM/GBSA (61) by averaging the binding energy over the ensemble of snapshots from combining all the individual MD trajectories. To ensure that data from a proper reactive conformation was used in the analysis, runs were excluded where the system left a reactive conformation for a significant portion of the trajectory. This was defined as the average distance between the bond forming carbons being greater than 6 Å for at least 100 consecutive measurements (10ps). In addition, runs were excluded where the relevant diene was in the unreactive s-*trans* conformation for more than 10 snapshots. For **10** with AbyU and the AbmU-TMD structure, this led to no runs being discarded. For **21**, the fewest number of unaffected runs for a single pose was 18 and so analysis was performed on this number of runs for all poses to ensure consistency. For **21** and **10** with the AbmU-loop structure, additional DYNAM runs were performed as necessary, where runs were discarded, to give a total of 10 runs for each pose.

6.3 Results and discussion

6.3.1 Stereoselectivity of AbyU from substrate-complex simulations

Flexible docking of the native substrate **10** led to the discovery of 12 poses. Of these, several distinct pro-*R* poses were found. Since the dominant reactive pro-*R* binding mode is already known from the work presented in Chapter 4, this was the only pose selected for simulation (Figure 6.8A). The other pro-*R* reactive poses correspond to orientations found for the product in Chapter 4. Only a single pose was found with the substrate in a pro-*S* conformation (Figure 6.8B), which was therefore selected to investigate pro-*S* reactivity/binding affinity. For the docking of the pro-*R* conformation of non-native substrate **21**, 4 poses were obtained that were broadly like those found for the native substrate; this is unsurprising given that the two substrates are quite similar. The most likely reactive pose (Figure 6.8C) was chosen for simulation by analogy to the known reactive pose for **10**. The fact that this binding mode is still accessible is already some indication that, despite the changes to the substrate, the active site is still well set up to produce the *R* product. However, as the reactive pro-*R* pose has not strictly been investigated for this substrate, simulations were also run for the pose considered to be the second most likely reactive pose, where the diene was also at the top of the cavity (Figure 6.8D). For the pro-*S* conformation of **21**, 9 poses were found. The poses selected for simulation were the analogous one to the pro-*S* pose found for **10** (Figure 6.8E) and the other pose found where the diene was at the top of the cavity (Figure 6.8F).

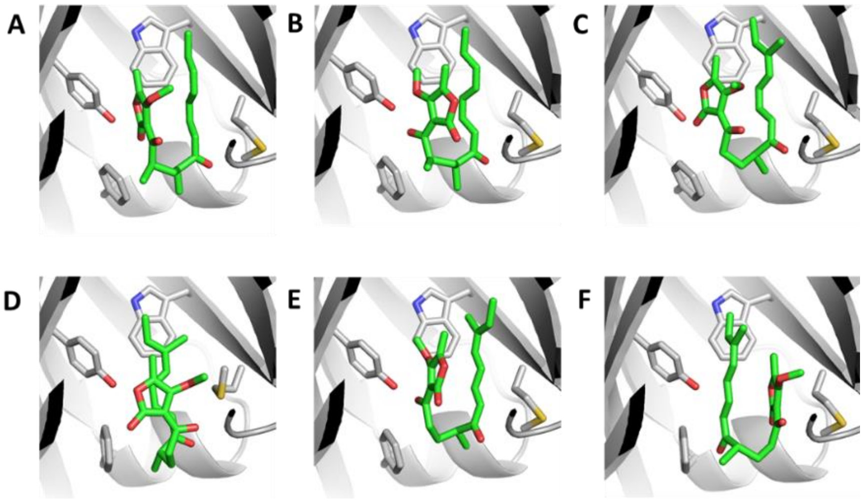


Figure 6.8: Docking poses selected for simulation of pro-*R* and pro-*S* conformation of substrates with AbyU. A) pro-*R* pose for **10**. B) pro-*S* pose for **10**. C-D) pro-*R* poses for **21**. E-F) pro-*S* poses for **21**. Residues treated as flexible in docking shown in sticks.

For AbyU with native substrate **10**, the average distance between the carbons atoms during MD was significantly shorter for the pro-*R* pose (Figure 6.9). Given that the pro-*R* pose tested was known to be the most reactive pose, this was expected to be reflected in terms of short carbon-carbon distances (as was also found in Chapter 4). Providing that the one pro-*S* pose found is also the most reactive pro-*S* mode, these results show that the distance between the bond forming carbons in the Michaelis complex could indeed be the primary factor driving stereoselectivity in the enzyme. With non-native substrate **21**, the pro-*R* pose with the shortest average carbon-carbon distance (Figure 6.8C) was analogous to the reactive pose for **10**, where the average distance found was also very similar for both substrates. This shows that the active site is still well set up to produce the *R* product for **21** through the same binding mode as for the native substrate. The difference between the average distances for the most reactive pro-*R* pose and the most reactive pro-*S* pose (Figure 6.8E) has somewhat diminished for this substrate compared to that for **10**. Therefore, according to this measure, the selectivity might be lessened. However, as there is still a clear separation, this result may explain the *R* selectivity seen in experiment. Overall, it seems that the general selectivity of AbyU towards the *R* product is conferred by the ability of it to form a highly reactive Michaelis complex when binding the pro-*R* conformation of either substrate (i.e. bringing the reacting groups into close proximity). However, for the non-native substrate, the active site is still cable of binding the pro-*S* conformation in a reasonably reactive Michaelis complex, perhaps hinting at the possibility that AbmU accomplishes the shift to an *S*-selective reaction by further enhancing the reactivity of this binding mode.

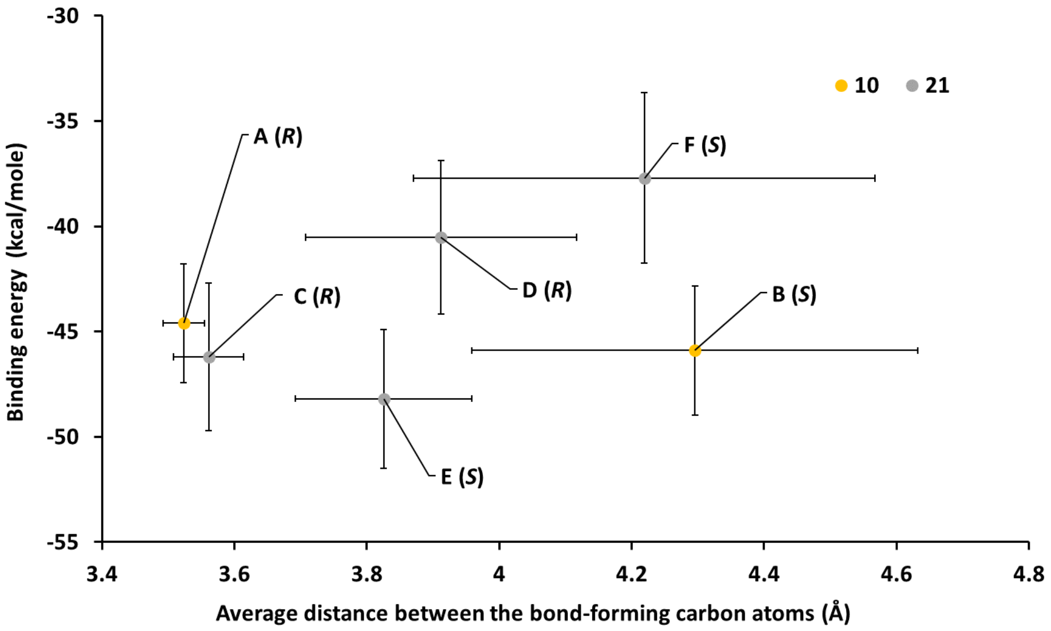


Figure 6.9: Assessment of binding affinity and reactivity of pro-*R* and pro-*S* substrate binding poses in AbyU. Poses that the data points correspond to are indicated by their equivalent labels in Figure 6.8 (pro-chirality of each pose is shown in brackets). Starting from the docking poses obtained for each prochiral conformation of each substrate, multiple independent molecular dynamics simulations were run. Average distance between the bond-forming carbons was measured from these runs to assess reactivity, with mean value and standard deviations obtained over the values for individual runs. Average and standard deviation of the binding affinity calculated (using MM/GBSA) over the snapshots from all the runs combined. Results for **21** were from 18 runs while 10 runs were used for **10**.

6.3.2 Stereoselectivity of AbmU from substrate-complex simulations (AbmU-TMD structure)

Flexible docking of the non-native substrate **10** led to finding 18 poses. Of these, only 1 was in the pro-*R* conformation and was thus selected for simulation (Figure 6.10A). This pose is quite similar to the reactive pro-*R* pose for AbyU in terms of its overall orientation within the active site (compare Figure 6.10A, Figure 6.8A), except for the fact that the tetronate carbonyl is now interacting with the Tyrosine on the other side of the cavity rather than the one that is common between the two enzymes. This could be an indication as to why the suspected product outcome does not completely shift towards the *S* product for this substrate, given that the pose is somewhat similar to the optimal native pro-*R* binding mode. 3 poses were found with the substrate in a pro-*S* conformation and the two with the diene at the top of the cavity were selected for simulation (Figure 6.10B, C), with one also resembling the pro-*S* pose found for AbyU (Figure 6.10B). For the pro-*R* conformation of the native substrate **21**, 18 poses were found. 2 poses were tested for which the diene was at the top of the cavity (Figure 6.10D, E) where one of these similarly resembled the known reactive binding mode for AbyU (Figure 6.10D). Assuming this is still reactive in this context, then there is no real hint from docking alone why the *R* product is disfavoured and the reaction becomes entirely *S*-selective. For the pro-*S* conformation of **21**, 11 poses were obtained. The two top scoring poses (Figure 6.10F, G) were chosen for simulation as these also had the diene at the top of the cavity.

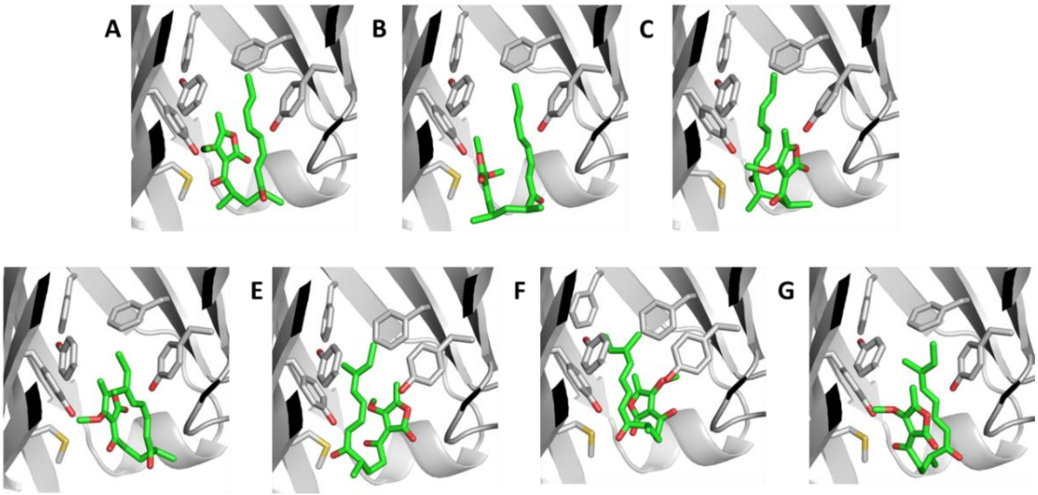


Figure 6.10: Docking poses selected for simulation of pro-*R* and pro-*S* conformation of substrates with AbmU structure from TMD. A) pro-*R* pose for **10**. B-C) pro-*S* poses for **10**. D-E) pro-*R* poses for **21**. F-G) pro-*S* poses for **21**. Residues treated as flexible in docking shown in sticks.

For AbmU with non-native substrate **10**, comparing the results for the pro-*R* pose with the most reactive pro-*S* pose (Figure 6.10B) there is no longer a clear separation between the average distances (Figure 6.11), which suggests the substrate may thus be turned over in both binding modes. This may therefore explain why two products were seen in the experiments and helps confirm that these are indeed probably the *R* and *S* products. The most reactive pro-*S* pose is analogous to the single pro-*S* pose found with AbyU and the most reactive pro-*S* pose of those tested for **21** with AbyU (Figure 6.8E). However, compared to AbyU with **10**, this pose has become more reactive along with the pro-*R* pose becoming less reactive. Therefore, these results suggest that the reason that complete *R* selectivity may have been lost and a mixture of products allowed to form is that the enzyme is simultaneously promoting the reactivity of the pro-*S* pose while no longer enforcing such a reactive conformation for the pro-*R* pose. However, the pro-*R* pose has obviously not been impacted so greatly as to lead to total *S*-selectivity as is the case for the native substrate. The results for the native substrate **21**, however, are less conclusive. It is known that the enzyme is purely *S*-selective with its native substrate but looking at the average distances alone this is not reflected (Figure 6.11). While the pro-*S* poses do have the shortest average distances, the average distance seen for the pro-*R* pose similar to the reactive binding pose for AbyU (Figure 6.10D) is only slightly higher and the error bars clearly overlap. Looking at the binding affinity results may at least provide a hint towards the answer. Although the difference between binding affinity is not so dramatic between the most reactive pro-*R* and pro-*S* modes (2.6 kcal/mole), there is a more significant difference between the reactive pro-*R* mode and the other pro-*R* pose tested (3.7 kcal/mole), which also happens to be considerably less reactive based on carbon-carbon distance. This suggests that the turnover of the pro-*R* substrate could be inhibited by the preferential binding of the substrate in a less reactive pro-*R* binding pose. This is not particularly convincing however because of the error involved and so, given also the potentially unrealistic alteration to the active site introduced by the TMD protocol, this casts strong doubt over the accuracy of the results for this structure.

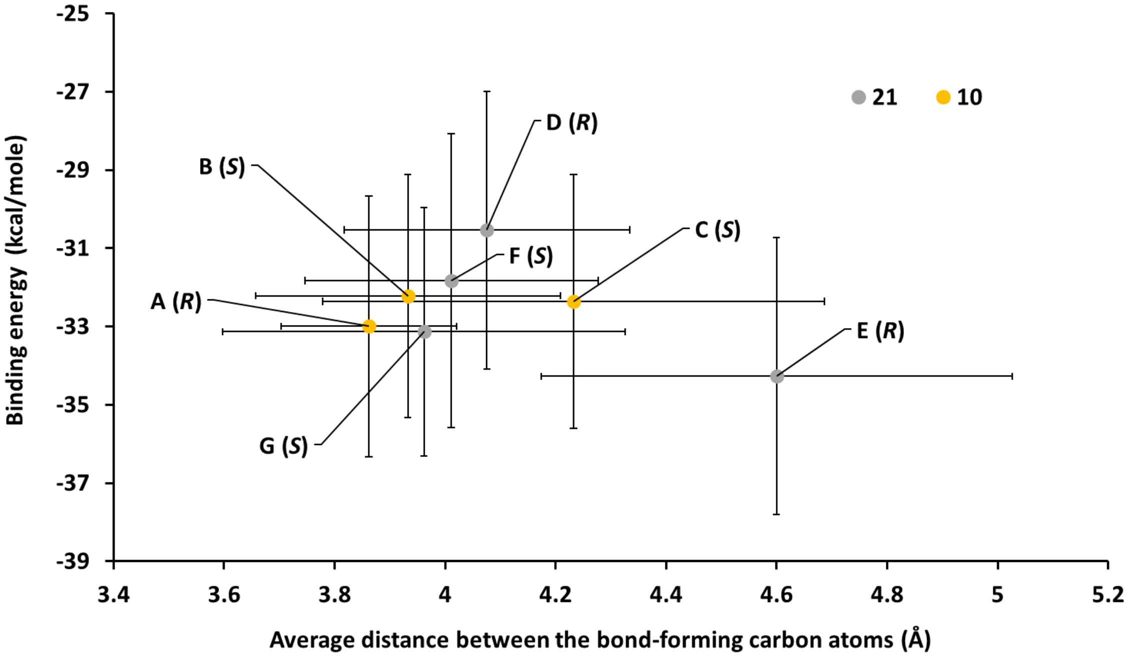


Figure 6.11: Assessment of binding affinity and reactivity of pro-*R* and pro-*S* substrate binding poses in AbmU structure from TMD. Poses that the data points correspond to are indicated by their equivalent labels in Figure 6.10 (pro-chirality of each pose is shown in brackets). Starting from the docking poses obtained for each prochiral conformation of each substrate, multiple independent molecular dynamics simulations were run. Average distance between the bond-forming carbons was measured from these runs to assess reactivity, with mean value and standard deviations obtained over the values for individual runs. Average and standard deviation of the binding affinity was calculated (using MM/GBSA) over the snapshots from all the runs combined. Results for **21** were from 18 runs while 10 runs were used for **10**.

6.3.3 Stereoselectivity of AbmU from substrate-complex simulations (AbmU-loop structure)

For the pro-*R* conformation of non-native substrate **10**, 6 poses were found from docking. The 2 poses tested both had the diene at the top of the cavity (Figure 6.12A, B). One of these (Figure 6.12B) much more closely resembles the reactive binding mode for AbyU (Figure 6.8A) than the pro-*R* pose tested for the AbmU-TMD structure (Figure 6.10A). Therefore, this still provides supports for the hypothesis that *R* product formation is retained for the non-native substrate. However, while the *S* product is thought to also be produced for this substrate, only one pose out of the 4 found for the pro-*S* conformation had the diene approximately at the top of the cavity (Figure 6.12C), and this was not the expected reactive pro-*S* mode suggested by the simulations with the AbmU-TMD structure (Figure

6.10B). Therefore, in order to explicitly test this mode, it was instead obtained via manual docking of the pro-*S* conformation (Figure 6.12D). However, during the structural optimisation protocol for this manually generated pose, the diene group rotated into an unreactive conformation where it was no longer coplanar with the dienophile and the reactive distances were correspondingly large. As this remained the case throughout the MD simulations, this pose was deemed unreactive. It was therefore decided to alter the conformation of the secondary diene for the manually docked pose to s-*cis* (Figure 6.12E) since this is the conformation it adopted in the corresponding pose found from flexible docking with the AbmU-TMD structure (Figure 6.10B). In this case, while still pro-*S* in terms of the stereocentre formed at C15, the substrate is then set up to yield a product with opposite stereochemistry at the other two stereocentres (C10, C13). This is reasonable, since it was merely assumed by docking in the analagous pro-*S* conformation as for **21** that this is the product formed, whereas it may actually be forming the product suggested by the flexible docking of **10** with the AbmU-TMD structure. This makes sense as, given the substrates differ in the substituents attached to the reacting diene, this may induce a change in its orientation when bound in this mode. For the pro-*R* conformation of native substrate **21**, 5 poses were found. The two poses with the diene at the top of the cavity were selected for simulation (Figure 6.12F, G). Interestingly, with this structure, a pose is no longer found where there is a favourable interaction between the tetronate carbonyl of the substrate and one of the tyrosines in the active site. As this is an actual structural indication of how the *R* product may be disfavoured with AbmU it suggests that this structure and docking result is giving a more realistic picture of substrate binding. For the pro-*S* conformation of **21**, 9 poses were found. The 3 poses for which the diene was at the top of the cavity were selected for simulation (Figure 6.12H, I, J). Of these, one pose (Figure 6.12J) is analogous to the reactive pro-*S* mode for AbyU (Figure 6.8E) and one of the two equally reactive pro-*S* poses tested for the AbmU-TMD structure (Figure 6.10G). An analogue of the other pose tested for the AbmU-TMD structure was also tested for this structure (Figure 6.12H).

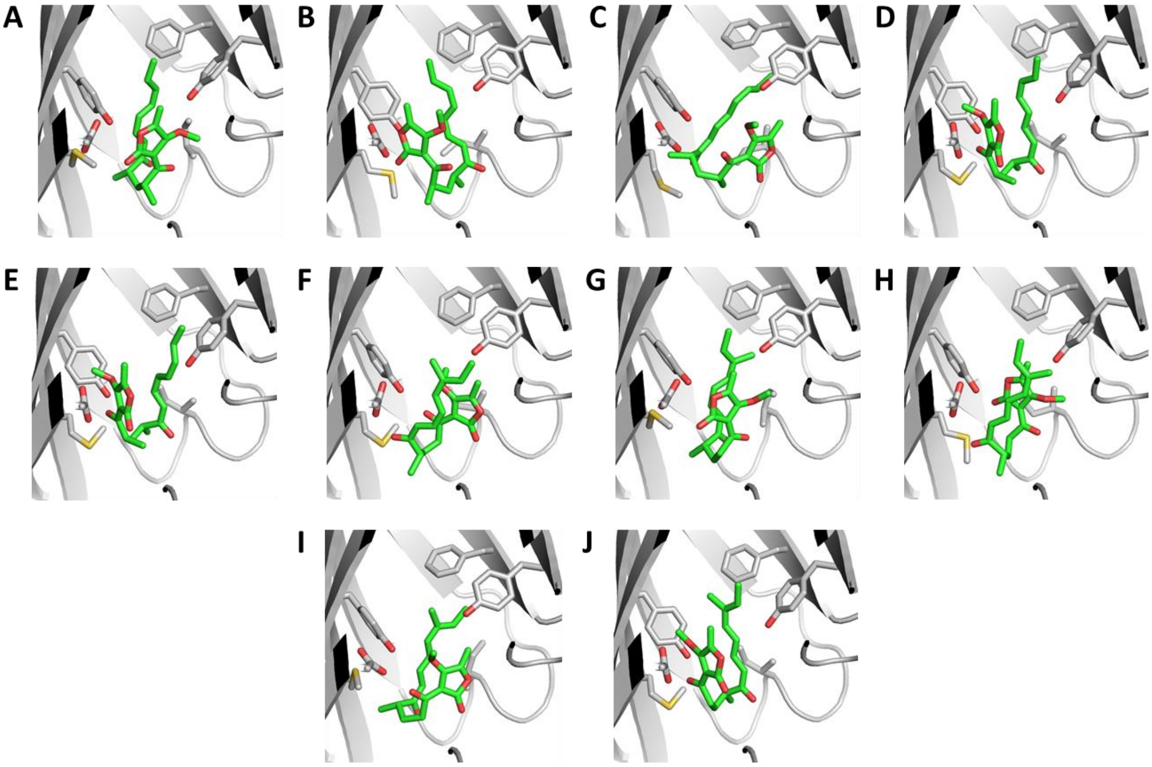


Figure 6.12: Poses selected for simulation of pro-*R* and pro-*S* conformation of substrates with AbmU structure with remodelled capping loop. A-B) pro-*R* poses for **10**. C-E) pro-*S* poses for **10**. F-G) pro-*R* poses for **21**. H-J) pro-*S* poses for **21**. Residues treated as flexible in docking shown in sticks.

The MD results for the non-native substrate **10** (Figure 6.13) present a similar picture to the previous simulations. Firstly, the 2 pro-*R* poses tested both lead to short reactive distances, so this still supports the hypothesis that one of the products seen in experiment is indeed the *R* product. Despite one pose (Figure 6.12B) seeming now more similar to the reactive pose for AbyU compared to the pro-*R* pose found for the AbmU-TMD structure, the reactivity of the pro-*R* conformation has still similarly diminished compared to the native enzyme, suggesting this is part of the reason for the suspected partial loss of *R* selectivity. The pro-*S* pose from manual docking with the alternative diene conformation (Figure 6.12E) seems far more likely to be the relevant pro-*S* mode (in terms of reactivity) than that from automated docking of the reactive conformation, as the both the average and standard deviation of the carbon-carbon distance are considerably lower. Although this mode is predicted to be somewhat less reactive than the pro-*R* modes in terms of average distances, there is sufficient overlap to support the second product seen in experiment being the *S* product. However, the results suggest that the *S* product formed for this substrate is not the exact analogue to the product formed by the reaction with **21**, but instead the product corresponding to the reactive pose found from the flexible docking with the AbmU-TMD structure (Figure 6.10B) i.e. which differs in chirality at the other two stereo-centres. Again, this makes sense given the difference in the substrate.

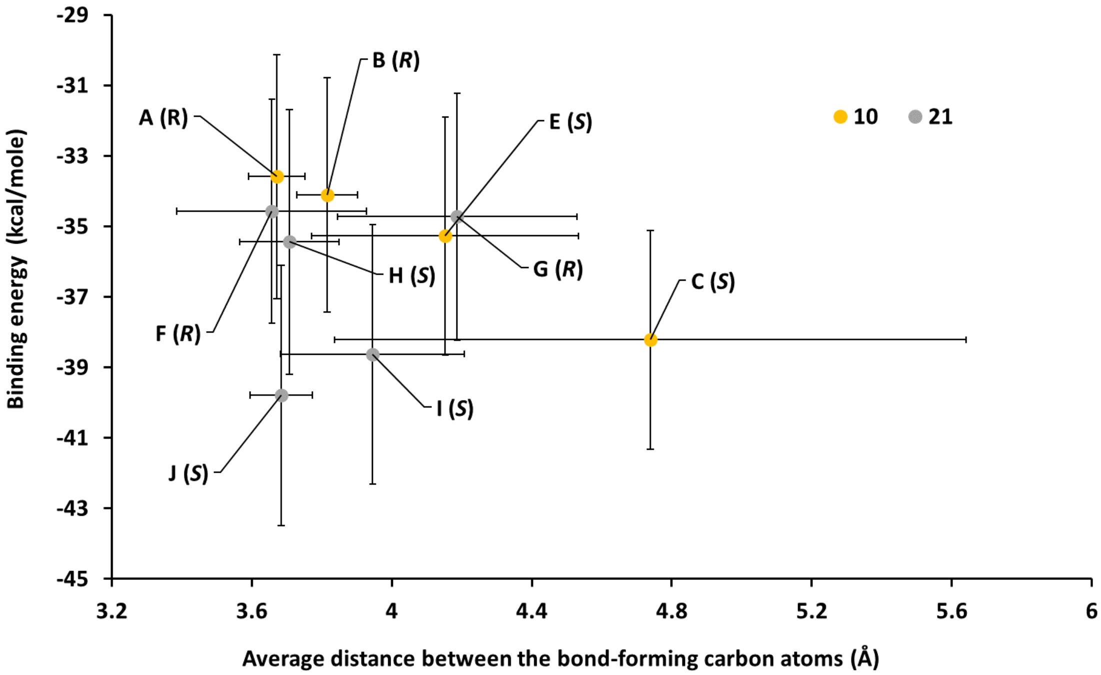


Figure 6.13: Assessment of binding affinity and reactivity of pro-*R* and pro-*S* substrate binding poses in AbmU structure with remodelled capping loop. Poses that the data points correspond to are indicated by their equivalent labels in Figure 6.12 (pro-chirality of each pose is shown in brackets). Starting from the docking poses obtained for each prochiral conformation of each substrate, multiple independent molecular dynamics simulations were run. Average distance between the bond-forming carbons was measured from these runs to assess reactivity, with mean value and standard deviations obtained over the values for individual runs. Average and standard deviation of the binding affinity was calculated (using MM/GBSA) over the snapshots from all the runs combined. Results were from 10 runs.

For native substrate **21**, comparing the binding affinities between the pro-*R* and pro-*S* modes with the shortest average distances, the pro-*R* mode is now 5.22 kcal/mole weaker in terms of binding affinity than the pro-*S* mode. This may be because the interaction between the tetronate carbonyl of the substrate and the active site Tyrosine is now missing in the pro-*R* binding modes found (such that there is now no effective equivalent for the reactive pro-*R* mode in AbyU). This provides a more convincing explanation for the total *S* selectivity of the AbmU reaction compared to that provided by the simulations with the AbmU-TMD structure. Given that the AbmU-loop structure with which these simulations were performed also completely retains the experimentally observed binding site configuration, it thus seems likely that the results for this structure are more reliable. Meanwhile, for the AbmU-TMD structure, the register shift induced in strand 1 (in order to obtain the analogous closed loop conformation as for AbyU - see section 6.2.1.1) alters the sidechains which point into the cavity for part of this strand, thus making the results for this structure much more questionable. Interestingly, the most reactive pro-*S* mode (Figure 6.12J) turns out to be the analogue to the most reactive pro-*S* mode in AbyU (Figure 6.8E), which was also one of the two reactive modes found for the AbmU-TMD structure (Figure 6.10G) (the analogue of the other pro-*S* mode tested for the AbmU-

TMD structure (Figure 6.12H), which was of comparable reactivity/affinity in that case, also has comparably short carbon-carbon distances to the most reactive pro-*S* mode (3.68 Å vs 3.71 Å) for this structure, but is now disfavoured in binding). Therefore, it seems that the most reactive pro-*S* mode for AbyU, while not reactive enough to yield any *S* product in that case, is the relevant mode for *S* production in AbmU. So, it seems that AbmU partially accomplishes a shift to an *S*-selective reaction for its native substrate by enhancing the reactivity of the pro-*S* mode beyond what it seen in AbyU. However, full *S* selectivity apparently involves also energetically disfavouring pro-*R* binding by preventing a key tyrosine H bond from forming for a pro-*R* conformation. It should be mentioned though that this explanation of stereoselectivity is less clear cut than for AbyU, for which a significant disparity was found in the reactivity of the pro-*R* and pro-*S* complexes. Since in these simulations a comparably reactive pro-*R* mode was identified, the explanation instead relies on the binding energy difference which, while large, is not statistically significant. Together with the fact that the method itself is rather approximate, this casts strong doubt on the reported average affinities as being representative and therefore able to explain the observed stereoselectivity.

6.3.4 Implications of results on engineering and computational prediction of stereoselectivity in spirotetronate forming enzymes

The results presented here provide considerable insight into stereoselectivity in the two enzymes studied and suggest ways in which stereochemistry may be manipulated in spirotetronate forming enzymes. Firstly, stereo-control in formation of the native product in AbyU appears to come from the enzyme supporting a significantly more reactive (in terms of the key carbon-carbon distances) pro-*R* than pro-*S* complex. In addition, the enzyme also conveys *R* selectivity when accepting the substrate of its counterpart enzyme AbmU, since the overall pro-*R* binding conformation is able to be maintained and is predicted to have the same level of reactivity as with the native substrate (and although the selectivity is predicted to diminish somewhat - due to the pro-*S* mode becoming more reactive - this is clearly still not enough to overcome the strong *R* preference). This general *R* preference suggests that selectivity mostly comes from the overall shape of the active site better supporting the skeleton of the substrate (which yields the spirotetronate scaffold) when arranged in a pro-*R* conformation. It would also seem that the interaction between Tyr76 and the substrate is critical to this general stereo-control as this will encourage the orientation of the substrate in the highly productive pro-*R* mode. To verify the importance of this interaction experimentally, one could mutate the Tyr to Phe to see if this led to a loss of stereoselectivity (although complicating this is that the fact that pro-*S* binding - and thus reactivity - may also be impacted). This leads to a useful lesson for engineering stereoselectivity in enzymes where binding mostly consists of hydrophobic interactions: incorporating a specific interaction is important to ensure the substrate binds in the desired reactive binding mode. Any slight divergence from this (e.g. going from binding mode shown by Figure 6.8C to that of Figure 6.8D) may impact reactivity enough to lead to a loss of stereoselectivity. In AbmU, the overall binding characteristics are similar to AbyU; a mode resembling the reactive pro-*R* mode for AbyU is still possible for the non-native substrate, and this seemingly allows a mixture of both *R* and *S* products to be formed (although this has yet to be confirmed experimentally), despite the *S* selectivity of the native reaction. In addition, formation of the *S* product appears to involve the same pro-*S* binding mode that was most reactive in AbyU (where its reactivity is enhanced to make it competitive with the pro-*R* mode). For the native substrate, the enzyme then favours the *S* product. To accomplish this, as well as pro-*S* mode reactivity being enhanced compared to AbyU, a pro-*R* binding mode which has the equivalent Tyr76 interaction as seen for AbyU is no longer able to form, leading to the pro-*S* mode being energetically favoured. As this is only apparent for the native substrate, this shows that the enzyme may leverage the molecular shape specific to the native substrate (e.g. via steric interactions) in order to achieve this, thereby making selectivity less general. Nevertheless, these findings make the case that, when engineering an enzyme to alter stereoselectivity (by promoting an existing less favoured binding mode), achieving total selectivity may require actively disrupting formation of the original product by e.g. preventing the unwanted mode (as seen in AbyU) from forming. Overall, the results for both AbyU and AbmU suggest that, in order to achieve stereoselectivity in spirotetronate forming enzymes in general, the enzyme should not allow significant binding of a substrate conformation with close reacting groups, when the substrate is set up to yield an undesired product. This makes sense, given that a large portion of catalysis will likely always come from the overall enforcement of a reactive substrate conformation. Thus, stereoselectivity will not necessarily be able to rely solely upon the fact that specific catalytic interactions are only present in the desired binding mode for example. This is echoed by a recent computational study for spiro-linking enzyme Pyrl4 (20), in which the origin of stereoselectivity was investigated. In this enzyme, there is a key catalytic electron-withdrawing interaction provided by Q115 (see section 1.2.2.2.2), and this may confer some stereo-control since it is only present in the binding mode leading to the observed *exo* product. However, the initial bond forming carbon-carbon distance sampled during MD was also found to be significantly shorter for the binding mode corresponding to the observed product. Therefore, this again suggests that an effective strategy for engineering stereoselectivity in these enzymes will typically require also ensuring that a reactive conformation with short carbon-carbon distances is not able to form for the unwanted products.

Further to the mechanistic insights gained, the results in this Chapter provide a good example of how simulation can be effectively used to give insight into stereoselectivity in β-barrel spirotetronate forming enzymes. This could be useful for probing stereoselectivity in similar newly discovered enzymes, as well as aiding future (re)design attempts. Firstly, the results demonstrate the importance of considering both the reactivity and binding affinity of relevant binding modes as either factor (or indeed both) could be responsible for stereoselectivity. In terms of the simulations themselves, the results show that, provided different modes are tested and the most relevant modes established for the different stereoisomers of interest, relatively short MD simulations of the substrate complex (starting from poses of the substrate predicted via automated docking) may be enough to correctly predict stereoselectivity (see below for exceptions to this). In terms of the requisite structural enzymatic data for running these simulations, results for the AbmU-loop structure show that, given at least an intact structure for the overall barrel (and thus also the binding interface), reliable results can potentially be obtained where a loop region making up the active site entrance is modelled by automated loop modelling tools (though this may involve further caveats in terms of the subsequent simulation procedures required as described below). Results for the AbmU-TMD structure meanwhile show that, using homology models based on experimental structures of homologous β-barrel enzymes poses a risk, as differences in the barrel architecture (such as a 'frame-shift' in a β-sheet interaction) can lead to inaccurate active site models and thereby inaccurate prediction results. As seen, even modelling in a single strand where there is a conflict can introduce enough of a departure from the true structure to give a very different (and thus misleading) picture of substrate binding. Regarding docking, it is important that the flexible region for the protein is carefully considered as this may prevent relevant modes from being discovered if not chosen appropriately. For the AbmU-loop structure, some testing of the flexible region was required to find the most relevant pro-*S* pose (see Methods). Thankfully, here there was some idea of the relevant pro-*S* mode *a priori* due to the simulations for the other structures; in other cases, more exhaustive testing may be required. In terms of reactivity/affinity predictions based on substrate complex simulations, distance between the bond forming carbon atoms during MD appears to be an effective proxy for reactivity in different modes (since this factor provides a convincing explanation for the observed stereoselectivity in AbyU), while approximate MM/GBSA binding energy calculations (based on just the conformations of the complex) can be sufficient to identify which modes will be favoured in binding. However, it should be mentioned that convincing results may not always be obtained with these procedures, as was found to be the case for AbmU. Even for the AbmU-loop structure, which likely provides the most sensible results for this enzyme, the results are far from convincing. This is due to the reliance on binding energies which exhibit significant statistical uncertainty to explain the stereo-preference, in conjunction with the questionable nature of MM/GBSA itself. Assuming that there is not another factor behind the stereo-selectivity (e.g. catalytic effects other than just the reactive distances) and binding strength is indeed the cause, it could be that the calculated binding affinities do not capture the true extent of the disparity. This could be due to the approximations of the specific MM/GBSA protocol that was applied here, such as the uncertainty in solvation and the fact that entropy has been neglected. MM/GBSA is also known to be an inherently imprecise method (108), so it may just be that further runs are required to achieve sufficient precision. However, another factor which may be at least partly responsible for the lack of a clear explanation in the results is the uncertainty in the starting structure. Firstly, it could be the case that, as well as just being closed, the specific loop conformation is important in establishing the interactions in the bound state. Therefore, if the proper closed loop conformation had not been identified by the modelling, clear results cannot be expected. Assuming that the loop conformation was reasonable or that the absolute conformation of the closed state is less important, as the starting structure input to enlighten came directly from Modeller this suggests the possibility that it may not have been well enough equilibrated prior to running the production MD. If the structure had not fully relaxed to a stable minimum, then the results would be biased by transient conformations that do not represent a valid Michaelis complex. In this case, increasing the number of runs would not help as a relevant minimum would still not be reached and sampled within the short length of individual runs. Although the Enlighten protocols do provide an equilibration step, this procedure is by no means rigorous. This therefore stresses that using efficient and short sampling procedures requires that the input structure be already well equilibrated or at least fairly close to a relevant stable conformation, such that an extensive equilibration procedure is not first required to achieve this. Bearing in mind the aforementioned caveats, the types of efficient protocols used here may prove to be generally useful for studying stereoselectivity in this class of enzymes. As well as providing insight into stereoselectivity, they could be useful in design for rapidly testing whether desired stereoselectivities have been achieved. For example, when designing stereoselectivity by means of enhanced reactivity, they can be used to ensure that a sufficiently more reactive complex is enforced for the desired prochiral conformation, and that binding is sufficiently strong for the desired binding mode (compared to other less reactive modes, as well as binding modes for the alternative product).

6.4 Conclusions

This chapter demonstrates that rapid (but careful) docking combined with efficient, multiple short MD simulations of resulting enzyme-substrate poses can give in-depth insight into the stereoselectivity of spirotetronate Diels-Alderases. These types of protocols will therefore be useful for aiding future prediction/design of stereoselectivity in such enzymes. However, detailed structural information from experiment is also required to aid such simulations; homology models may not always be sufficient. By running the aforementioned simulations for the complimentary AbyU and AbmU with their different substrates, the molecular basis for their opposite stereoselectivities was suggested, giving insight into how stereoselectivity can be achieved in future design attempts. In addition, by further testing the enzymes with each other’s substrates it was determined whether the difference in selectivity arises primarily from changes to the substrate or enzyme. General *R* selectivity was found for AbyU which appears to come from the greater reactivity of the pro-*R* binding mode; which is as effective regardless of which substrate is bound. However, the enzyme may become slightly less selective with different substrates due to pro-*S* conformations becoming more reactive (as is the case for the AbmU substrate). As well as general shape complementarity, specific interactions such as the Tyr76 interaction in AbyU are important to ensure the substrate binds in the aforementioned reactive conformation, since other conformations may be less reactive (leading to a loss of stereoselectivity). In the case of AbmU, the shift to an *S*-selective reaction for the native substrate appears to come from the fact that, along with a promotion of the reactivity of the pro-*S* binding mode, a similarly reactive pro-*R* binding mode to AbyU is now energetically disfavoured due to the disruption of the key Tyr76 interaction. This shows that, in addition to promoting the desired mode, achieving an acceptable level of stereo-control may also require explicit disfavouring of the unwanted mode. However, as this selectivity is seemingly lost with the AbyU substrate (as the native pro-*R* binding mode is largely maintained), this ability to disrupt the reactive pro-*R* binding mode is highly contingent on the specific shape of the native substrate, making selectivity less general for AbmU. Lastly, the results for AbmU and AbyU (as well as those for homologous Pyrl4), stress the importance of considering the key distances between the bond forming carbon atoms to establish the differential catalysis that will be provided by the enzyme towards the different product outcomes.

Chapter 7. AbyU capping loop: behaviour under pressure and with mutation

7.1 Introduction

The active site capping loop of AbyU (Figure 7.1) plays a key role in its catalytic cycle by gating active site access. As previous computational and experimental evidence suggest that loop motion could be rate limiting for the enzyme (see below), this has therefore been investigated computationally in this chapter. Firstly, the ability of the loop to open was studied in relation to activity of AbyU found at different hydrostatic pressures. In addition, since it forms such an important part of the catalytic cycle, the effect of mutations in the loop was investigated, both to rationalise experimental findings and to explore potential for loop optimisation.

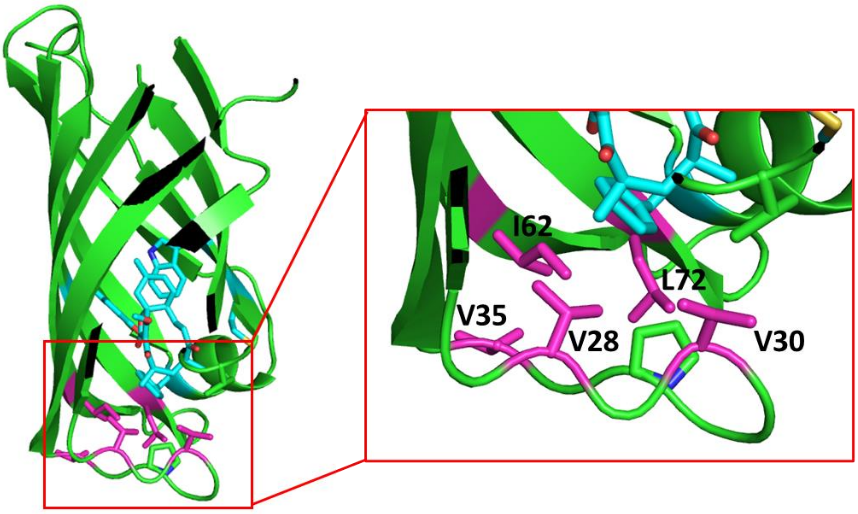


Figure 7.1: AbyU active site capping loop in closed conformation with natural product bound. Showing X-ray crystal structure of AbyU (green) and docking pose for product **10P** (cyan) in the reactive binding mode identified in Chapter 4. Capping loop connects strands 1 and 2. Key hydrophobic contacts which hold the loop closed are highlighted with sticks. Positions which were mutated are coloured pink and labelled in inset.

The closed conformation of the capping loop, where the active site is inaccessible to solvent, can be seen in the crystal structure for the apo form (Figure 7.1). Hydrophobic interactions between the loop and the residues at the bottom of the cavity are made in this state which thus stabilise this conformation. However, despite being found in the closed state in the crystal structure, MD simulations have shown that the loop is highly flexible for the apo enzyme (25), opening easily to enable substrate binding. Conversely, in simulations with the product of **9** bound (in what is now known to be the primary binding mode - see Figure 7.1) the loop remained shut during 300 ns of conventional MD. This is presumably because, as well as gating access to the active site, the loop also forms part of the hydrophobic binding cavity when in the closed conformation. Thus, when product is bound, the loop can form a maximum number of hydrophobic contacts by also interacting with the hydrophobic region of the product, thus greatly stabilising the closed conformation. As these interactions will also be present with the substrate, similar behaviour is therefore assumed for the substrate complex. This suggests that, while the loop is initially flexible in the apo state, once the active site is occupied, the loop becomes much more ordered and adopts the closed conformation in order to trap the substrate and facilitate the reaction step. However, although this was not captured on the timescale of the previous simulations, the loop must still be able to open up to release the cyclised product on a short enough timescale so as to not bottleneck the reaction (i.e. leading to product inhibition) and allow efficient turnover of the substrate. There is therefore a trade-off between substrate capture and product release and the loop flexibility/interface is likely finely tuned to provide the optimum balance given the conditions under which the enzyme has evolved to operate. In a similar enzyme Pyrl4, this trade-off may be mitigated by the active site capping mechanism being ligand inducible, such that the substrate is trapped more strongly than product, and thus avoiding product inhibition (19). However, no evidence currently exists of such a mechanism in AbyU, and the previous simulations indeed point to loop opening being a slow process when product is bound. Therefore, since pre-steady state multiple turnover kinetics experiments also indicate that product release is rate limiting, there is reason to believe that this may be due to the rate of loop opening for product release. The ability of the loop to open was therefore investigated as a possible cause of the loss of activity found at high hydrostatic pressures.

To give insight into the pressure tolerance of AbyU, it has been tested experimentally to see if it retains structure and activity at pressures up to 2000 bar. High hydrostatic pressure can denature an enzyme by forcing solvent into the hydrophobic core (109). However, high-pressure synchrotron radiation circular dichroism (HP-SRCD) showed secondary structure content does not change significantly at higher pressures indicating that the structure is not denatured at the pressures tested. Pre-steady state multiple turnover kinetics were also performed using stopped flow across the pressure range in question. From this it was found that the steady state rate for the enzyme-catalysed reaction begins to decrease significantly at pressures beyond 750 bar. As the structure appeared to remain stable, and burst phase data precluded substrate binding or cyclisation being responsible for the drop in activity (i.e. showing that the rate limiting step is still product release), it was thought that it may instead be due to the capping loop being less able to open at higher pressures, thus slowing the rate of product release. As pressure is increased higher volumes become energetically penalised; thermodynamically this leads to equilibria moving towards the state with lowest volume (E.g. Le Chatelier’s principle for gaseous reactions). Similarly, according to absolute rate theory (110), given a positive activation volume where volume is greater in the transition state, activation energy will increase at higher pressure thus disfavouring transitions which go through a state with higher volume. Therefore, if the transition state between the open and closed loop occupies a greater volume than that for the closed state this would decrease the rate of loop opening at higher pressure which could then be responsible for the observed drop in activity. To investigate this, AbyU was simulated at different pressures between 1-1000 bar to see if any difference in loop opening could be observed. Although it is technically the product complex which is of interest, since no loop opening was observed for the previous simulations within a reasonable timeframe, the apo structure was instead simulated. Although the energetics of loop opening are clearly quite different between the two, assuming the loop goes through a similar transition in either case then the activation volume and thus effect of pressure for the two states should be similar. Unlike the previous simulations using the TIP3P water model, a modified version of the TIP4P model (111) was used for these simulations to more accurately represent the environment at high pressure. Because of this, the product complex was also simulated again at 1 atm to enable direct comparison. Finally, to confirm the secondary structure content is also not appreciably affected by the increase in pressure this was analysed using the DSSP method (112). This stands for ‘dictionary of secondary structure of proteins’ which is a method to assign secondary structure information based on the atomic co-ordinates of a protein.

To give more insight into the role of the loop, several loop mutants have been created and tested experimentally. These were V28S, V30S and I62V (Figure 7.1), which were designed to disrupt the native hydrophobic contacts made by the loop in the closed conformation to varying degrees and see the overall effect this had on activity. By changing Valine to Serine, a nonpolar loop residue is replaced by a polar one. This should thus strongly disrupt the hydrophobic contacts and the stability of the closed conformation. In contrast, I62V just removes a single methyl group and may therefore only slightly weaken the interactions at the loop/barrel interface. To rationalise the effect of the mutations on activity, these loop mutants were tested *in silico* by simulating both the apo and product bound form and comparing to the WT enzyme. Based on the results for the experimentally tested mutants, four new mutants were also proposed and tested *in silico* which, in the same vein as I62V, were designed to slightly weaken hydrophobic interactions, with the aim of slightly easing loop opening for the product complex without overly compromising substrate capture. These mutants were V28A, V30A, V35A and L72V (Figure 7.1), for which the product complex alone was simulated. For the mutants which subtly disrupt the loop interactions (all except the Valine to Serine mutants), enhanced sampling was also performed using Gaussian Accelerated MD (GaMD) (60). This was done to enable significant sampling of the different loop states to be obtained for the product complex within a reasonable simulation time. Accelerated MD (aMD) in general is a simple approach for enabling slow processes like large conformational changes to be observed as, rather than specifically biasing the event of interest as in a technique like Metadynamics for example (thus requiring an appropriate reaction co-ordinate which describes the motion to be found), it accelerates dynamics of the entire system by applying a boost potential to the total potential energy and/or dihedral energy (one drawback of this however is that unwanted effects like unfolding can also occur which can then confound analysis of the behaviour of interest - see section 7.2.2.2). Using the structures sampled in the GaMD simulations, reweighting was then performed (60) to estimate free energy differences between the relevant loop states and determine whether certain mutations may make loop opening more energetically favourable.

7.2 Methods

7.2.1 Conventional MD simulations

Explicit solvent periodic box MD simulations were performed for the WT enzyme and V28S/V30S/I62V/V28A/V30A/V35A/L72V mutants. For the WT enzyme and V28S, V30S and I62V mutants both the apo-enzyme and product complex were simulated while only the product complex was simulated for the additional mutants. Simulations for the apo WT enzyme were carried out at 1, 250, 500 and 1000 bar while all others were carried out at 1 bar.

To get the initial coordinates for the apo WT enzyme chain A of the crystal structure of AbyU was taken from the PDB (ID 5DYV) and the bound HEPES molecule removed. The mutant structures were built from the WT structure using the mutagenesis wizard in PyMOL. In all cases, backbone dependent rotamers were used and the number one ranked rotamer (i.e. most frequently occurring) was chosen as this also had the least clashes with surrounding residues. For the complex structures, the ligand and flexible sidechain co-ordinates for the docked pose of **10P** corresponding to binding mode A in Chapter 4 were used (see Figure 7.1).

In order to give more separation between the protein and its periodic images (and since it was not required to capture the behaviour of interest), the short, disordered N-terminal region which precedes the first β strand in the sequence (Figure 7.1) was removed. 5 residues were therefore removed, and the N-terminal capped with an Acetyl residue [−C(=O)−CH_3] in place of Gln10. Hydrogens and missing heavy atoms were added by the AmberTools program tleap. Asp43 was treated as protonated (predicted pK_a value for the apo and complex WT structure by PropKa 3.1 (88, 89) were 9.7 and 11.1 respectively (25)) and His88 was doubly protonated (based on the hydrogen bonding pattern as determined by reduce - see section 4.2.2.2). All other titratable residues were left in their standard states. A truncated octahedral solvent box of TIP4P/2005 (111) water was then added by tleap with a minimum distance between any protein atom and the edge of the box of 15 Å (and a closeness parameter of 0.75), along with 4 sodium ions to neutralise the net charge.

MD simulations were performed with the Amber16 programs sander (for minimization) and pmemd.cuda (for MD on GPUs, using the default SPFP model) using the AMBER ff14SB force field for the protein and the TIP4P/2005 water model. Particle-mesh Ewald summation was used in conjunction with periodic boundary conditions, and a cut-off for direct-space non-bonded interactions of 8 Å. Prior to production simulation at 298 K and either 1, 250, 500 or 1000 atm, the following procedure was carried out: 1) optimize solvent and solute hydrogen positions (minimization for 300 steps with restraints on solute heavy atoms, force constant: 100 kcal mol^{-1}Å^{-2}); 2) heating/equilibration of solvent (50 ps NPT heating to 300k with 2 fs timestep, using Langevin dynamics with collision frequency 2, SHAKE and initial random velocities assigned at 50K, and using the Berendsen barostat with a pressure relaxation time of 1 ps. All solute atoms were restrained with force constant: 25 kcal mol^{-1}Å^{-2}); 3) 300 steps of minimization and quick heating to 298 K (20 ps in NVT ensemble with 2 fs timestep, using Langevin dynamics with collision frequency 1, SHAKE and initial random velocities assigned at 25K) with harmonic positional restraints on Cα atoms only (force constant: 5 kcal mol^{-1}Å^{-2}); 4) 2 ns equilibration of temperature and pressure in the NPT ensemble (Langevin dynamics with 2 fs timestep, SHAKE and collision frequency of 1 and using the Monte Carlo barostat with the default pressure relaxation time of 1 ps) with harmonic positional restraints on Cα atoms only (force constant: 5 kcal mol^{-1}Å^{-2}); 5) gradual release of positional restraints in 4 steps of 20 ps MD in the NPT ensemble (conditions as in previous step) such that in the final step the force constant is then at 1 kcal mol^{-1}Å^{-2}. The unrestrained production simulation was then performed in the NPT ensemble for 500ns, saving snapshots every 10ps (MD with SHAKE and a 2 fs timestep, using the Weak coupling temperature control algorithm with the time constant for coupling to the heat bath of 10 ps and the same pressure regulation as in the previous step). Four runs of the whole procedure above were performed to obtain 4 independent production MD trajectories for each system for analysis.

RMSD analysis of the trajectories was performed with the AmberTools program cpptraj for the 20,000 conformations sampled from 300 to 500 ns in each of the 4 repeat MD simulations. RMSD was calculated based on the positions of the Cα atoms in the loop (residues 26-36) relative to those in the crystal structure after first aligning frames to the crystal structure on the Cα atoms of the β-barrel (residue ranges 11-25, 37-45, 53-64, 71-82, 85-93, 103-111, 118-127, 131-139). For the simulations of the apo WT enzyme at the different pressures, secondary structure analysis was performed using the DSSP method implemented in cpptraj for the complete 500ns MD trajectories for each of the 4 runs.

7.2.2 Gaussian Accelerated MD

7.2.2.1 Simulation procedure

GaMD simulations were run for the product complex of the WT enzyme and I62V, V28A, V30A, V35A and L72V mutants. Simulations were performed using the Amber18 version of pmemd.cuda, with the starting structure coming from the end of the final equilibration stage of one of the previous conventional MD simulation runs. The initial GaMD setup MD run was 70.4 ns, consisting of a conventional MD phase of 6.4 ns (where potential statistics are collected to initially determine the GaMD boost parameters), followed by an equilibration stage of 64 ns (during which the boost potential is applied and the boost parameters updated). This was run in the NVT ensemble at 298K where initial random velocities were assigned (using Langevin dynamics with 2 fs timestep, SHAKE and collision frequency of 2). It was ensured that the equilibration stage of the GaMD setup run was long enough so that the potential statistics and boost parameters had settled down. The average and SD of the potential energy were calculated every 320 ps which in simulation steps (160000) was approximately four times the number of atoms in the system as recommended. Following the setup MD, 7 independent 1000 ns production GaMD simulations were spawned (applying the final boost parameters), where initial random velocities were assigned for each (using same conditions as for setup MD). It was ensured that the starting structure used for these always had the loop in the closed conformation so as not to introduce bias between mutants (although with ideal converged sampling the results should be irrespective of starting structure). For all systems except V28A (where the final structure from the end of the final equilibration stage of the previous conventional MD had to be used) this was the structure at the end of the GaMD setup MD simulation. The GaMD simulations were run with the “dual boost” setting where boost potentials are applied to both the dihedral and total potential energy terms. The default value of 6.0 kcal/mole was used for the upper limit of the boost potential SD (σ0) for both the dihedral and total potential energy terms. The threshold energy (up to which energies are boosted) was set to the upper bound as using the lower bound was found to result in insufficient acceleration.

7.2.2.2 Discarding simulations

Due to the acceleration, partial unfolding was occasionally observed in the GaMD production simulations. This manifested itself in partial dissociation of the two β strands which connect to the capping loop. Unfortunately, this dissociation occurred between the “bottom” ends of the strands i.e. where they connect to the capping loop (see Figure 7.2). As this would therefore invalidate the later geometric measurements which were used to quantify loop openness (loop RMSD and number of contacts) simulation runs which were significantly affected were discarded from further analysis. To decide whether a run should be discarded, a criterion was first found which was able to describe the observed dissociation. For this the antiparallel β strand fraction (as determined by DSSP) for the 5 pairs of residues which form the two strands in the relevant region was used (indicated in Figure 7.2). This measurement represents the fraction of residues pairs in the specified region for which the necessary backbone hydrogen bonds exist (as determined by a distance cut-off) in order for the residues to be classed as forming an antiparallel bridge by the DSSP method (112). These can be either the two complementary H bonds between the residues in the pair itself, or both of the neighbouring H bonds between the two residue pairs adjacent to the pair in question (see Figure 7.3). As this was calculated for 5 pairs of residues this fraction can therefore vary between 0 (where no β structure exists in the specified region) and 1 in steps of 0.2 (depending on how many residue pairs qualify as forming an antiparallel bridge). For example, for the crystal structure this fraction is 0.8 as the bottommost residue pair has only 1 hydrogen bond (see Figure 7.2) and is thus not defined as forming an antiparallel bridge. It was found that only in cases where this fraction was 0 (i.e. the region showed no β structure) for a continuous period of the trajectory did the large-scale dissociation previously described occur i.e. if at least one pair of residues showed β structure then the strands remained in close proximity. Runs were thus discarded if this fraction was equal to 0 for greater than 15000 consecutive frames (150ns). Although large scale dissociation can occur before this length time this threshold was reached as a compromise so as not to discard too many runs. Using this criterion 2 runs were discarded for V28A and 1 run for the WT.

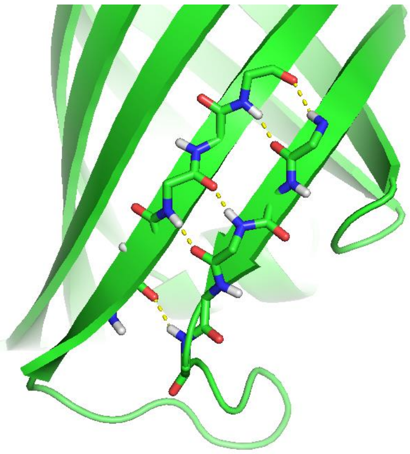


Figure 7.2: Region for which β structure was determined using DSSP to detect strand dissociation. Showing crystal structure with backbone atoms of region used in calculation highlighted with sticks representation. H-bonds which exist and form part of an antiparallel bridge according to DSSP definition indicated with dashed lines.

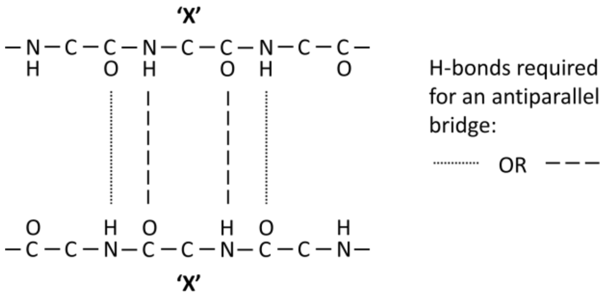


Figure 7.3: Definition of an antiparallel bridge in the DSSP method. Showing the two possible ways in which residue pair 'X' can be considered to form a bridge.

7.2.2.3 RMSD analysis and reweighting

RMSD was performed as for the conventional MD simulations on all 100,000 frames of the 1000 ns production GaMD trajectories. The PyReweighting toolkit (113) was used to perform energetic reweighting (using cumulant expansion to the first order) for the structures sampled in the production MD and obtain a free energy profile or 'potential of mean force' (PMF) across the relevant states of the loop i.e. closed, open and those in between. Instead of loop RMSD, the reaction co-ordinate used to describe the different states of the loop and obtain the PMF was the number of contacts between residues 28-34 in the capping loop and residues 26, 39, 60, 62, 65, 72, 95-6, 99, 129-30 at the bottom of the barrel (which should interface with the loop when it is in a closed conformation). This was calculated for each frame of the trajectory using cpptraj, where a contact was counted as any atom in region A being within 5 Å of any atom in region B (Figure 7.4). The PMF profiles reported are those obtained from performing the reweighting on the combined repeat 1000ns production GaMD trajectories which were used for analysis (i.e. ignoring discarded runs), where a bin size of 20 contacts was used. A leave-one-out procedure was used to obtain an approximate error on the PMF profiles: the reweighting was rerun multiple times leaving out a different run from the combined set each time. The error bars shown at each point along the profile are then the standard deviations across the profiles where a run was left out.

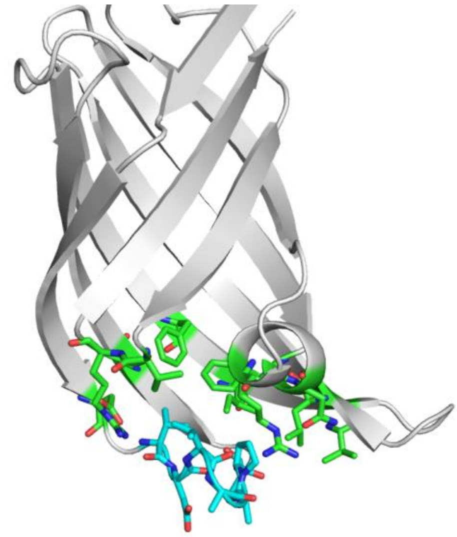


Figure 7.4: Selected regions for defining number of loop contacts. A contact was defined as any atom of any residue in the blue highlighted region of the loop being within 5 Å of any atom of any residue in the green highlighted region at the bottom of the cavity.

7.3 Results and discussion

7.3.1 WT simulations at atmospheric pressure

The results of the new simulations at 1 bar run here with the TIP4P/2005 water model show similar behaviour of the loop as for the simulations in the previous study (25) using TIP3P. In the apo state the loop is more flexible, sampling a large range of conformations as shown by the wide RMSD distribution, while for the product complex the loop remained firmly closed (close to the starting crystal conformation at an RMSD ~1 Å) for the duration of the simulations (Figure 7.5). For the apo state, there were two overall populations which conformations fell into. The first was the population sampled in 3 of the runs and corresponds a peak at an RMSD of ~3 Å where the loop still closed enough to prevent active site access (see Figure 7.6). However, in these runs the loop is not as close to the tightly closed crystal conformation as it is for the product complex, which on its own indicates some relative weakening of the hydrophobic interactions at the loop/barrel interface (Note: a low RMSD state coincident with that for the product complex is observed in the apo trajectories up until ~100ns, but this is likely just due to the simulations starting in the crystal conformation which, as only briefly held, indicates that it is not a stable conformation for the apo-enzyme). A second population which was reached in the remaining run (which also samples the closed conformation seen for the other 3 runs to an extent) shows up as a broader peak at an RMSD of ~5 Å. Since the loop is likely to be open for an RMSD >5 Å (Figure 7.6) the loop spends a significant amount of time in an open conformation during this run. This shows that this a stable conformation and thus that a significant proportion of the apo-enzyme is likely to adopt a conformation in which substrate can bind. Since this conformation was also reached within 200ns this also suggests a relatively low barrier to loop opening for the apo-enzyme.

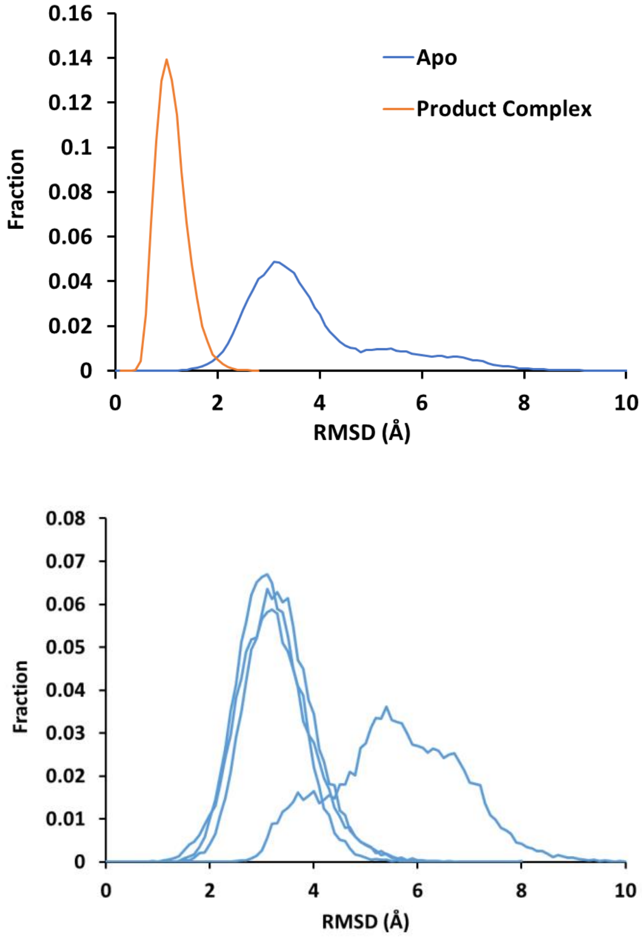


Figure 7.5: Histograms (0.1 Å bin width) showing the distribution of the capping loop Cα RMSD (residues 26-36) from the final 200ns of the MD runs for the apo and product bound enzyme. Top – Distributions from the 4 combined runs for the apo and product bound enzyme. Bottom – Separate distributions for the individual runs for the apo enzyme. RMSD is calculated relative to the closed loop conformation from the crystal structure (after aligning on the β-barrel Cα atoms) and thus represents the deviation in the loop backbone from that conformation; above ~5 Å, the loop is likely to be open.

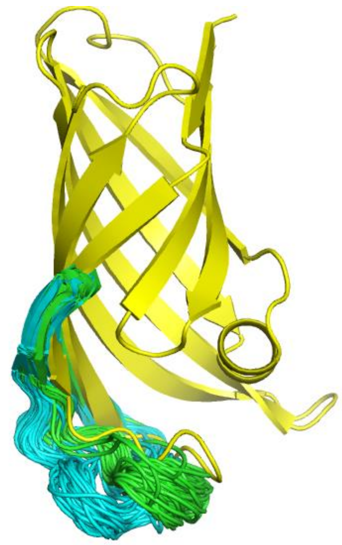


Figure 7.6: Difference between high and low RMSD structures. Showing residues 21-41 of 50 structures which have RMSD between a range of 2.5-4 Å (green) and 5-6 Å (blue). Structures are taken at regular intervals from (and so as to span) the entire set of structures which fall within the given RMSD ranges after concatenating these from the 4x500ns MD trajectories for the apo WT enzyme at 1 bar. Crystal structure shown in yellow.

7.3.2 Sensitivity of WT enzyme to pressure

The overall enzyme structure remained stable during the simulations and DSSP confirms that the secondary structure is not affected at pressures up to 1000 bar (Figure 7.7). This indicates that that the β-barrel architecture confers remarkable pressure stability to the enzyme, which is consistent with HP-SCRD studies that also show the secondary structure content to remain fairly constant up to 2000 bar (L. Maschio & P.R. Race, personal communication).

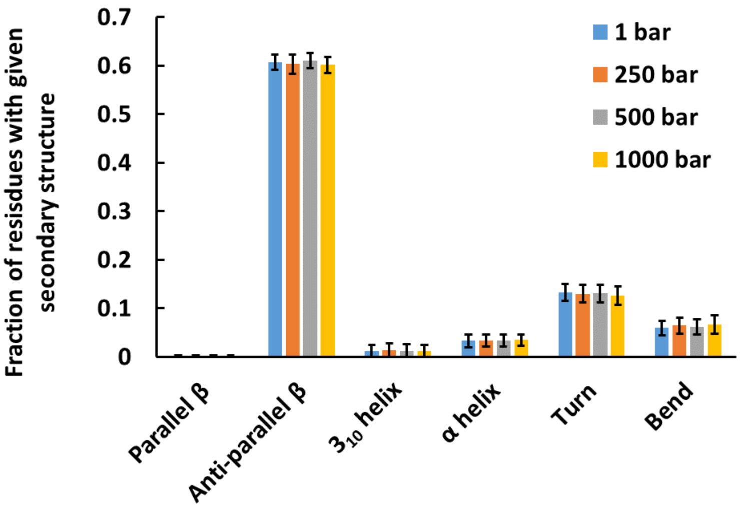


Figure 7.7: Secondary structure content of enzyme determined by DSSP at each of the 4 pressures. Averaged over all frames from the 4 repeat 500ns MD trajectories at each pressure. Error bars show standard deviation in the fraction of each type of secondary structure over the frames.

While the overall enzyme structure is unaffected, it was thought that the precise behaviour and flexibility of the capping loop (which is essential to enzyme function) may be more susceptible to high pressures and therefore responsible for the loss of activity observed at pressures above 750 bar. However, the difference in sampling of open conformations across the pressure range tested could not explain the loss in activity. At all pressures, the closed conformation described previously for the apo simulations at 1 bar was the dominant conformation, being sampled almost exclusively for 3 of the runs, where 1 run at each pressure is then seen to sample from a more open population (Figure 7.8). Therefore, there is no difference indicated by these simulations in terms of the ease of loop opening.

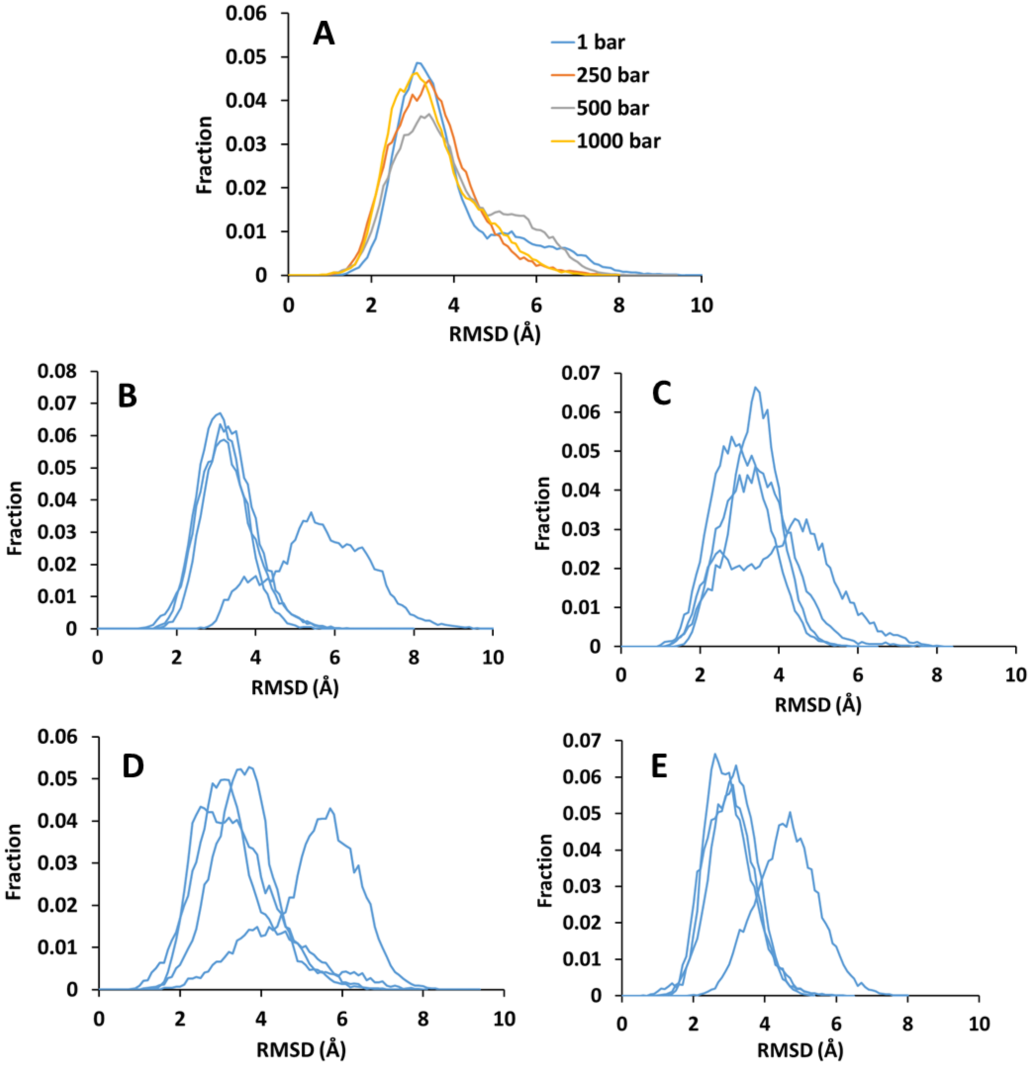


Figure 7.8: Histograms (0.1 Å bin width) showing the distribution of the capping loop Cα RMSD (residues 26-36) from the final 200ns of the MD runs at the different pressures. A) Distributions from the 4 combined runs at each of the 4 pressures. B-E) Distributions for the individual runs at pressures 1 (B), 250 (C), 500 (D) and 1000 (E) bar. RMSD is calculated relative to the closed loop conformation from the crystal structure (after aligning on the β-barrel Cα atoms) and thus represents the deviation in the loop backbone from that conformation; above ~5 Å, the loop is likely to be open.

However, since opening only occurs in a single simulation there is also not significant enough sampling to claim that loop opening is unaffected by pressure. Another possibility suggested by these results comes from the fact that the higher RMSD peak seen at 1000 bar appears at a lower RMSD than that for 1 bar - this could suggest the loop being open less of the time when sampling the more open state (thus potentially hindering product release). However, as the same can also be said for 250 bar (where activity is the same as at atmospheric conditions) this difference is clearly not statistically significant and would again require running more sampling to confirm. It is of course possible that the loop itself is not responsible for the loss of activity and that loop function is fact robust in the face of high pressure. This would also be an interesting finding in the context of enzymatic adaptation to a high-pressure environment. Many insights into environmental adaptation have come from studying homologous enzymes from organisms which have evolved in different environments (114, 115). As proteins are dynamic flexible molecules, which comprise a network of interacting structural elements such as loops, enzymes must achieve some way of retaining their flexibility, while also retaining structural integrity if they are to function under different conditions. It is thus generally found that proteins show similar flexibility and stability at their growth conditions or so-called “corresponding states”. For example, variants of dihydrofolate reductase (DHFR) found in organisms evolved to survive in high pressure and temperature environments have recently been studied (114), where it was found that they exhibit similar flexibility (as shown by their RMSF) at their respective growth conditions. For the high pressure adapted DHFR, simulations also indicated that, by replacing an Aspartate with a Glutamate (as compared to the thermophilic enzyme) and thereby making a key hydrogen bond between two loops more flexible, this may enable the enzyme to cope with the increase in pressure (and reduction in available volume that it creates) by allowing sufficient flexibility for the loop motions to remain correlated with each other while also still connecting with the other structural elements. As for other enzymes then, the capping loop of AbyU may be similarly tuned to achieve sufficient flexibility to counter the opposing force of pressure which could otherwise restrict its motion (while still allowing a sufficiently ordered closed conformation to form at milder conditions). However, the question still remains as to what the limiting factor is for AbyU activity beyond 750 bar. As mentioned, to confirm whether the loop is to blame would require running more simulations to obtain a significant indication as to whether loop opening does in fact become more difficult at higher pressures, or if indeed a lower RMSD open population may be favoured at 1000 bar for example which could hinder product release. Alternatively, GAMD simulations could be performed (as seen later for mutant study) to obtain significant sampling across the relevant loop states (and then obtaining from this an estimate for the energetics of loop opening/to see where the true open state minimum resides at different pressures). Another possibility is that product dissociation, rather than loop opening, is the rate limiting step, and responsible for the drop in activity at higher pressure. This seems plausible since unbinding of the product from the cavity will also involve an increase in volume and thus be energetically penalised by an increase in pressure. Future investigations may therefore also wish to focus on modelling this aspect.

7.3.3 Effect of mutation

7.3.3.1 Conventional MD

Loop conformations sampled in unbiased simulations for the apo and product bound states for the mutants V28S, V30S and I62V are now compared to the WT enzyme to understand their effect and how this relates to activity. Firstly, examining the behaviour of the apo enzyme for the Valine to Serine mutants; as for WT, 2 main populations are seen in the distribution of the RMSD values found for the loop (Figure 7.9). One peak occurs at an RMSD of ~3 Å where the loop is effectively closed, and another at an RMSD of >6 Å where the loop is then most likely open. Although for V28S this high RMSD population comes from 2 simulations (as opposed to the WT for which only a single simulation reached a stable high RMSD conformation), this alone is not a statistically significant indication that it is easier for the loop to reach an open conformation for this mutant. (this would of course be expected to be the case though since intuitively the mutation should have disrupted the hydrophobic interactions which serve to keep the loop closed). For V30S, a stronger indication is found that loop behaviour has been perturbed since the peak for the open conformation occurs at a significantly higher RMSD (~8 Å) than is seen in any of the previous simulations, indicating that a more open conformation is favoured for this mutant. Finally, for the I62V mutant, rather than two distinct populations, a much flatter distribution is found (which is comprised of peaks from the individual runs that are found at a range of RMSD's), indicating a wider range of stable global conformations and higher overall loop flexibility.

Looking at the product complex results for the Valine to Serine mutants (Figure 7.9), just one run maintains the closed loop conformation seen for the WT product complex, while the remaining runs sample more open states. This is markedly different behaviour than that for the WT and thus gives a strong indication that introducing a polar residue has significantly destabilised the closed conformation, thus making loop opening a more common occurrence for the product complex. This gives considerable insight into the experimental activity where it was found that, for the V28S mutant, activity is completely abolished[7]. Given this, these results suggest this mutation goes much too far in disrupting the native loop interactions as, although it may ease product release (which is rate limiting for the WT), it presumably has also had far too detrimental an impact on substrate capture/binding

[7] L. Maschio, K. Tiwari & P.R. Race, personal communication

affinity. While V30S was insoluble[7], and its activity therefore not tested, these simulations suggest it may also be too extreme as the product complex easily reaches an RMSD of at least 4 Å.

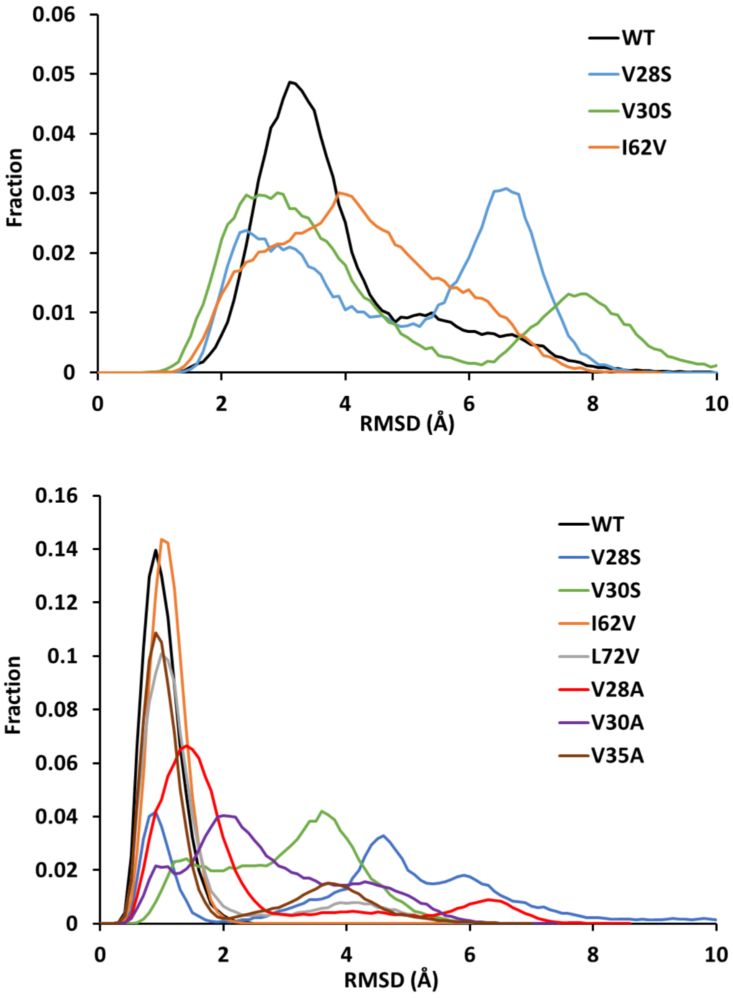


Figure 7.9: Histograms (0.1 Å bin width) showing the distribution of the capping loop Cα RMSD (residues 26-36) from the final 200ns of the 4 MD runs for the apo enzyme (Top) and product complex (Bottom) of the WT and loop mutants. RMSD is calculated relative to the closed loop conformation from the crystal structure (after aligning on the β-barrel Cα atoms) and thus represents the deviation in the loop backbone from that conformation; above ~5 Å, the loop is likely to be open.

For the I62V mutant, the results for the product complex were effectively the same as for the WT as the loop remained closed (with an RMSD of ~1 Å) for the duration of the simulations. Clearly, any potential difference between these cannot be detected on a reasonable simulation timescale using conventional MD. As preliminary kinetics also indicated that this mutant may be quite active[6] (although not whether it is significantly more active than the WT), this suggests that this mutation is subtle enough so as to not detrimentally disrupt the native behaviour. As the apo simulations also suggest increased flexibility of the loop there was thus interest in whether this mutation may in fact enhance product release and thereby activity. The additional mutants V28A/V30A/V35A/L72V have not been tested experimentally, but their product complexes were simulated to determine if, like

I62V, they may be promising in terms of subtly enhancing product release without greatly compromising substrate capture. Some appreciable sampling of higher RMSD conformations (Figure 7.9) than those seen for the WT/I62V (and in the case of V28A of a definitely open conformation) is found for these mutants. As they also appear to have had more of a subtle effect than V28S, this makes them promising candidates for further investigation to see whether they may subtly enhance loop opening.

7.3.3.2 Gaussian Accelerated MD (GaMD)

To further investigate the possible effect of mutation on loop opening, GaMD simulations were run for the mutants showing promise based on the conventional MD (all those except the V to S mutants). This was done to obtain significant sampling for the entire range of accessible loop conformations and an approximate PMF across the different loop states to determine whether, and by how much, loop opening may be eased by these mutations. In all cases, GaMD enabled significant sampling to be obtained for a broad range of loop RMSD values, including above 5 Å where the loop is likely to be significantly open (Figure 7.10).

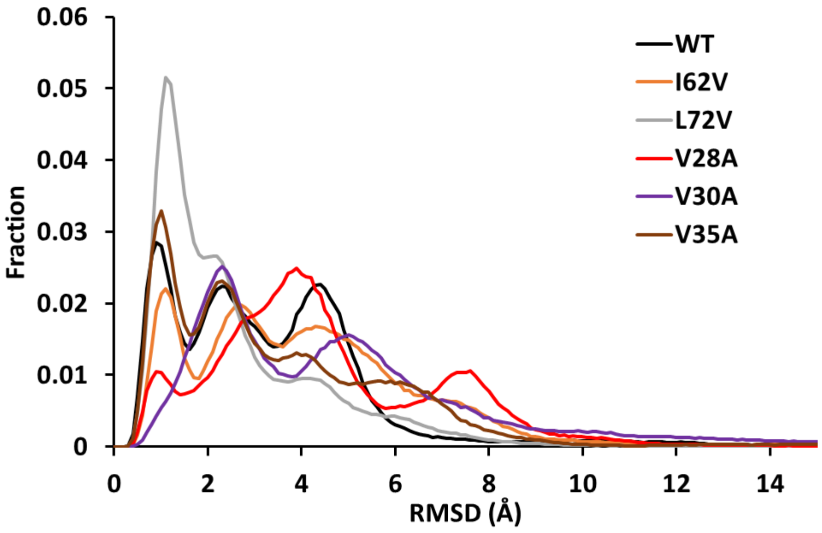


Figure 7.10: Histograms (0.1 Å bin width) showing the distribution of the capping loop Cα RMSD (residues 26-36) from the 7 1000ns GaMD runs for the product complex of the WT and loop mutants (ignoring runs which were discarded for the WT/V28A). RMSD is calculated relative to the closed loop conformation from the crystal structure (after aligning on the β-barrel Cα atoms) and thus represents the deviation in the loop backbone from that conformation; above ~5 Å, the loop is likely to be open.

The different populations found in the RMSD distributions roughly correspond to where sampling was found for the conventional MD (cMD) simulations, with additional sampling now obtained between these states and in some cases strong peaks emerging in the distributions which are only slightly apparent from the cMD results (e.g. at RMSD's of ~5 Å for V30A and ~4 Å for V28A) which likely represent additional stable global conformations. As these are biased sampling distributions, reweighting is required to recover the unbiased distribution across states (and then, assuming significant enough sampling is obtained, this gives the probability of observing a particular state which can be related to its free energy). Throughout this Chapter, RMSD has been used to describe the different loop states in terms of their openness, however it is not necessarily the best measure for this as it only shows how different a given structure is to the reference crystal structure (where the loop is closed). Therefore, for the reweighting, another metric has been used which should provide a more direct measure of loop "openness". This was chosen to be the number of contacts between residues in the capping loop and those at the bottom of the barrel which should interface with the loop when it is in a closed conformation (Figure 7.4). The sampling distributions found in terms of the number of contacts can be seen to be much flatter than the RMSD distributions (Figure 7.11); the distinct populations present in the RMSD distributions (which presumably represent different global conformations) disappear when conformations are instead grouped by their number of contacts. This highlights that RMSD is not the best measure of openness as conformations within the distinct RMSD populations do not necessarily share a similar number of contacts (and since number of contacts is a more direct measure of loop openness, this therefore shows that RMSD provides a less relevant distribution to use to discern the probability of the loop being in a more or less open state). However, an RMSD >5 Å does most likely correspond to an open conformation, so can be used to identify if loop opening has occurred (as in the previous sections). With the more extensive sampling of higher energy states along the path between an open and closed state though, a more precise measure is required to help determine the free energy change when moving between these states.

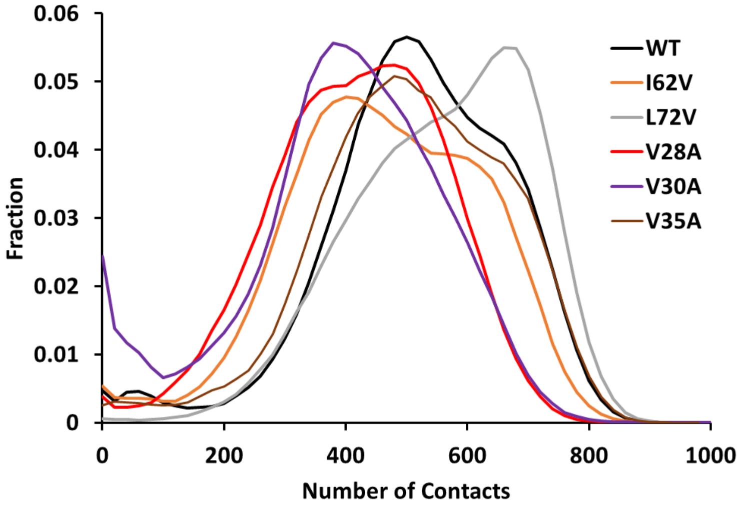


Figure 7.11: Histograms (bin width 20 contacts) showing the distribution of the number of loop contacts from the 7 1000ns GaMD runs for the product complex of the WT and loop mutants (ignoring runs which were discarded for the WT/V28A). See section 7.2.2.3 for how a loop contact was defined.

Looking at the PMF from reweighting for the WT (Figure 7.12), the minimum energy state corresponds to where there are many contacts (~700) and the loop is thus closed (see Figure 7.13). As expected, this minimum is roughly consistent with the crystal conformation for which there are 762 contacts. The energy then climbs, initially steeply and then more gradually, as the number of contacts is reduced to 200. This makes sense, given that reducing the number of favourable hydrophobic interactions should result in an energetic penalty. In qualitative terms at least, this energetic penalty (which is most severe when breaking the first few contacts) helps explains why the system never left the tightly closed crystal conformation in the cMD simulations (see later for discussion of the energies themselves which are unrealistic). The same qualitative behaviour as seen for the WT is then also seen for the mutants, excepting V28A and V30A. Around 200 contacts, the loop is likely open enough to allow product release (Figure 7.13), however for all except V28A/V30A no real minimum is found for an open state by this point. Below 200 contacts, the profile is not shown, as the errors rise significantly for most systems (along with large fluctuations in the PMF) due to insufficient sampling. However, it might be expected that as energies converge in this region, they will produce a minimum for an open state (which is the case for the WT for which better sampling was obtained in this region). This is because there should be more possible conformations as the number of contacts diminishes (and thus at least an entropic favouring of a fully open state). As no minima are found, no effective barrier to reaching a stable open state can be determined to compare between these mutants. However, the profiles can show how much energy must be overcome to reach a low contact state. It should be mentioned at this stage that, while the PMFs may provide a qualitatively accurate picture of loop energetics, the absolute energy differences are significantly underestimated. According to the PMF for the WT, for example, there is only ~2 kcal/mole free energy that must be overcome to reach a low contact state. Using TST and the Eyring equation, this would mean this should be achieved on average once every 5ps (and therefore that many thousands of opening-closing events would occur within 500 ns of cMD, which is clearly is not the case). This underestimate appears to be a typical result when obtaining PMFs from reweighted GaMD simulations. For example, in two studies which used GaMD to investigate binding events involving large conformational changes in a protein, similarly small barriers to those seen here were found in PMFs from reweighting (116, 117). This was in spite of the fact that these transitions are not observable within µs length cMD simulations. However, in the aforementioned studies, the PMFs were generally used qualitatively to give a picture of the conformational landscape and to characterise the binding pathways (e.g. by identifying intermediate states and getting an idea of the relative difficulties for the different transitions). In a similar fashion, comparing the PMFs for loop opening here may still provide accurate qualitative predictions about the relative difficulty of leaving a closed state for the various mutants. Firstly, we shall investigate the mutants for which the discovered energetics are qualitatively similar to WT. Looking at L72V/V35A, although conformations with higher RMSD than seen for the WT were reached in the cMD simulations, the PMFs indicate the closed conformation is actually more stable for these relative to a state with fewer contacts. This likely shows that the sampling obtained from the cMD simulations was not sufficient to yield a significant comparison in these cases. Similarly, for I62V (where no difference to WT was observed with cMD) the PMFs indicate that it is in fact slightly easier to reach a state with fewer contacts. The results for these three mutants therefore highlight the power of enhanced sampling in resolving the relative effect of different mutations which all subtly affect loop behaviour (and can therefore not be reliably detected through a few relatively short cMD simulations). The implication of these results in terms of loop optimisation is that L72V/V35A are likely not useful mutants since the energy profiles suggest it will be more difficult on average to depart from a closed conformation than for the WT. On the contrary, as I62V appears to slightly ease loop opening, this may therefore be a mutation that can enhance activity by aiding product release without seriously compromising substrate capture.

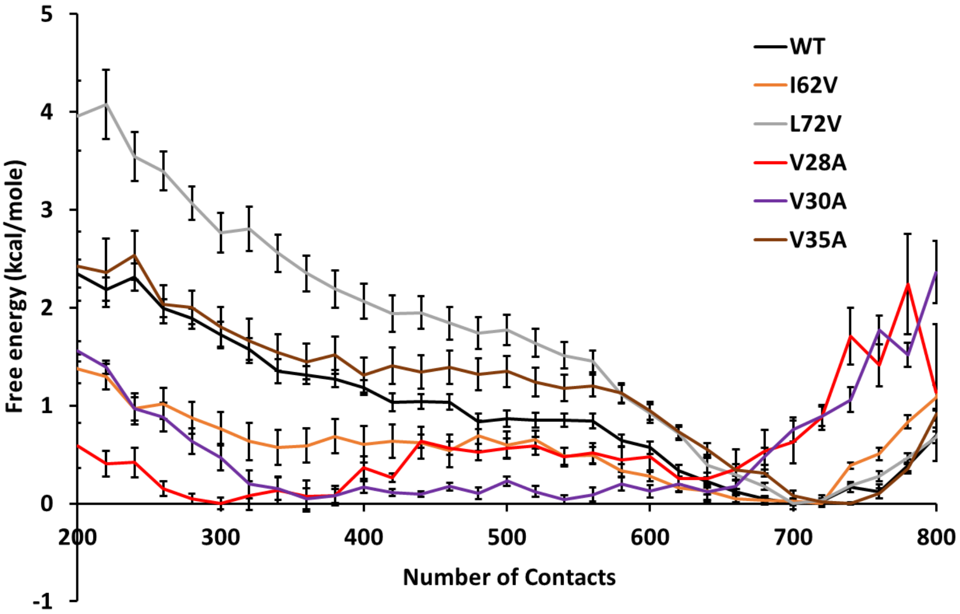


Figure 7.12: 1-dimensional PMFs in terms of the number of loop contacts for the product complexes of the WT and loop mutants. Obtained from reweighting of the distributions (bin size 20) from the 7 1000ns GaMD runs for each system (ignoring runs which were discarded for the WT/V28A). Below 200/above 800 contacts PMF’s are not shown as errors become very large for most systems due to insufficient sampling (see Figure 7.11). See section 7.2.2.3 for how error bars were calculated. Number of loop contacts in the WT crystal structure is 762. For the (in-silico) mutant crystal structures number of contacts is 706, 726, 659, 672 and 762 for I62V, L72V, V28A, V30A and V35A respectively.

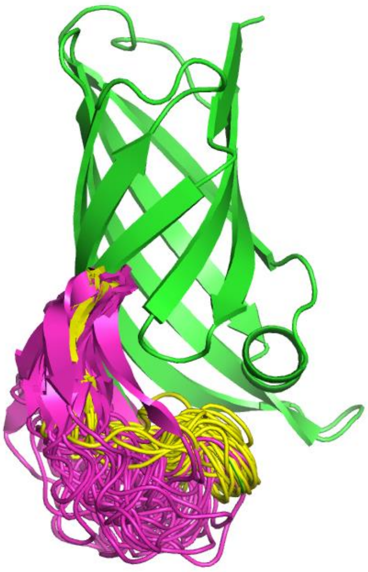


Figure 7.13: Comparison of high and low contact structures. Showing residues 21-41 of 50 structures which have between 700-800 (yellow) and 100-200 (pink) contacts. Structures are taken at regular intervals from (and so as to span) the entire set of structures which fall within the given contact ranges after concatenating these from the 7x1000ns GaMD trajectories for I62V. Crystal structure shown in green.

Unlike for the other mutants, the PMFs for V30A/V28A show qualitatively different behaviour to the WT and that these have likely too drastically destabilised the closed conformation. For V28A, both a closed and open minima are found with a low barrier between them, however the closed minimum is actually less stable than the open one (Note: although the location of the closed minimum in terms of number of contacts is different than for the WT, it is consistent with the number of contacts for this mutant when in the crystal conformation – see legend in Figure 7.12). Comparing the resulting barrier for leaving the closed conformation to the energetics of initial loop opening for WT shows this has likely made loop opening too easy, and the fact that the open state is more stable further suggests that substrate capture will be too strongly affected. Based on these results it makes sense that an open state was easily reached by running just a few cMD simulations, and in this case it could have been concluded from this alone that loop opening had been made too much easier. However, performing GaMD allowed us to quantify energetics and discount the possibility that the opening observed was a rare event that may have just happened to have been observed by running a few simulations. Finally, for V30A, a flat profile is found between 300-700 contacts i.e. no stable closed state exists. This shows that this is clearly too extreme a mutation as the loop will too easily wander into a conformation where substrate could escape thus severely impacting substrate capture. This is also already somewhat evident from the cMD results where a flat RMSD distribution was found (and no well-isolated low RMSD population). Overall, the results for these mutants suggest that loop opening is indeed a rare enough event for the WT that it was unlikely to be observed on the timescale of the cMD simulations. Therefore, any mutation which makes it observable on the same timescale when running just a few simulations is most likely too extreme. Again, the power of GaMD in this regard is that much longer effective timescales are simulated. This allows the behaviour of interest to be sufficiently sampled, and a subtle change in that behaviour to then be resolved by inferring what the energetics of it would be without the presence of the biasing potential.

However, while GaMD has proved to be a useful method here, it is worth contrasting it with approaches that have been used in similar studies. In a recent study (118) of Triosephosphate Isomerase (TIM), several different enhanced sampling techniques were used to study a catalytically important loop movement. In this case, the goal was to observe closure of the loop which is required for the enzyme to achieve a catalytically competent state. Loop motion is also thought to be partially rate limiting in this enzyme, although its high activity means that it may also be partially diffusion limited. Loop closure has been estimated based on experiment to occur on the timescale of ~100 μs with a barrier of 12 kcal/mole (119) for the substrate bound enzyme and is thus not practically observable via cMD. Both Hamiltonian replica exchange MD (HREX-MD) and bias-exchange metadynamics (BE-METAD) were therefore employed to study this. HREX-MD (120) is similar to aMD in the sense that no collective variables are required to describe the process under investigation. Using HREX-MD, the Hamiltonian is modified in a particular part of the system in order to only enhance sampling of configurational space in the so-called "hot" region which is of interest (this improves computational efficiency compared to temperature replica exchange MD (T-REMD) where the whole system is enhanced by running individual replicas at different temperatures (121)). However, loop closure was not observed after running ~300 ns for each replica. It was noted that this was unsurprising since REMD simulations have been estimated to give only an order of magnitude speedup (122), which is therefore insufficient given the ~100 μs required for closure. It is worth comparing this speedup to the enhancement found here with GaMD where, in simulations for the WT, loop opening (to an RMSD >5 Å) was consistently observed within 400 ns. Comparison of k_{cat} values indicate that loop opening may be a similarly slow process in AbyU (564 s^{-1} for AbyU vs 300-2100 s^{-1} for TIM). This may therefore be demonstrating superior acceleration for GaMD. However, as seen GaMD has the drawback of potentially sampling unwanted conformations, such as partial unfolding. The other enhanced sampling technique used in the TIM study was BE-METAD (123). In regular metadynamics (124), a biasing potential is generated during simulation time along a collective reaction co-ordinate (which describes the process of interest). As the simulation is run, the shape and height of this potential grows flattening the energy surface along this co-ordinate and allowing the system to explore the range of interest (where the free energy along the co-ordinate can then be inferred from the shape of the converged potential). In BE-METAD, a set of collective variables (CVs) are instead assigned which are all relevant to the process in question. A number of MD replicas are then run in parallel, with each replica being biased by a history dependent metadynamics potential acting along one of the CVs. This biasing potential is then exchanged between replicas during the simulations so that the replicas periodically change which CV they are being biased along. For complex processes, this typically reduces the sampling required to obtain good convergence (compared to regular metadynamics) since you explicitly sample more of the relevant energy surface. In this case, a native contacts parameter was used for the CVs which described how close to the reference closed structure a conformation is in terms of the strength of the specific contacts which are made in the closed state. Using this method, full loop closure was successfully observed for the substrate bound enzyme and convergence of the pathway achieved with 500ns sampling for each of the 7 replicas. This method could have possibly been used here in a similar way (thereby avoiding the unwanted unfolding of the protein). However, a key difference worth mentioning with the work here for AbyU is the different measure used for describing the loop state. In the TIM study, the native contact-based measure was used, since it was desired to study the process of loop closure to the specific conformation observed in the crystal structure (this is because the loop has to be in this exact conformation for the subsequent reaction to be energetically feasible). In this chapter, however, the focus was on product release, and describing simply how open or closed the loop was thus more relevant (in order to then find the relative energy penalty for leaving any conformation which prevents active site access). Regular metadynamics may thus have been a more appropriate method than BE-METAD, as using the total number of loop/barrel contacts (as used for the GaMD reweighting) as a collective variable, for example, could have been a straightforward way to sample many different possible opening pathways.

Another technique which has been used in similar studies to provide additional insight into loop conformational change is principal component analysis (PCA). Performing this analysis based on the sampled loop conformations would help to characterise the overall conformational states that the loop occupies (by extracting representative structures for these for example), as well as understanding how the loop transitions between these states. By projecting the sampling statistics along principal components, it could be determined whether the loop adopts a simple binary open or closed state (with the associated transition between these occurring along a single principal component), or, as has been found for other enzymes with catalytically important loop motions (118, 125), whether more complex dynamics are involved with transitions between other meta-stable states. As well as giving insight into the different conformational states of the loop, studying the conformational dynamics via PCA may help to better understand what is the relevant transition that will dictate the kinetics of loop opening.

7.4 Conclusions

In this chapter, the catalytically important capping loop of AbyU has been investigated in detail via both conventional and enhanced MD simulations. Conventional MD (cMD) was first run in an attempt to explain the pressure dependence of the enzyme and determine whether loop opening may be a bottleneck in product release and hampering activity at high pressure. However, using the current simulation protocol, a significant difference could not be identified in the flexibility/ease of loop opening (of the apo enzyme) at different pressures. Further simulations (e.g. using enhanced sampling techniques) are required to determine whether loop opening in the product bound state may be responsible for loss of activity. The possibility remains that functional loop movement is tolerant to high pressure, which would be an intriguing result from the perspective of enzymatic adaptation. If this is the case, product dissociation itself may be the rate limiting step that changes with increase in pressure. This aspect could thus also be studied in future work. cMD was then applied to understand the effect of mutations in the loop on behaviour, then correlating this to experimental activity to give further insight into the loop's role in catalysis. These results showed that introducing polar residues can strongly destabilise the closed loop conformation, even when the active site is occupied. This will thus severely affect the ability of the enzyme to trap substrate, which is clearly crucial since this type of mutation lead to a complete loss of activity in experiment. Loop mutants which more subtly affect loop behaviour were thus tested in an attempt to further optimise the loop (by enhancing loop opening and speeding up the rate-limiting product release step, without overly compromising substrate capture). Since loop opening was not observable for the product complex on reasonable timescales using cMD, the enhanced sampling technique GaMD was instead used. Using this technique, loop opening was made to be observable within the same simulation timescale as the previous cMD, while also enabling extensive sampling of open, closed and intermediate loop states. Energy profiles obtained from GaMD allowed meaningful differences to be discerned in the ease of loop opening for different mutants. The protocols applied here thus show promise to aid precise rational engineering of loop dynamics which take place on timescales that are not practically reachable via conventional MD simulations.

Chapter 8. Conclusions and future work

In this thesis, a broad range of modelling techniques has been applied to answer key questions about catalysis, specificity and turnover in AbyU. There has also been a focus on protocol development for enzyme engineering, with application of protocols in an activity prediction workflow in order to expand the substrate scope of the enzyme. Since modified substrates are of key interest, the main focus has been on engineering of the active site. However, some effort has also been directed towards engineering of the capping loop and thus achieving an improved overall enzyme architecture (although, as the loop makes contact with the substrate, this may also be dependent on the specific substrate of interest).

To help understand the catalytic role of the enzyme, the intrinsic DA reaction was investigated in Chapter 3 using QM calculations. IRC calculations were performed using DFT to obtain the reaction barrier in gas phase. Along with helping to understand the intrinsic reactivity and the reaction path, this enabled benchmarking to find an appropriate semi-empirical method for later QM/MM calculations. Further DFT level calculations demonstrated that the cause of the poor reactivity seen in experiment was likely due to the instability of the folded reactive conformation, suggesting that stabilisation of this conformation by the enzyme could be the primary driver of catalysis.

In Chapter 4, the enzymatic reaction was studied to confirm the origins of catalysis as well as the contribution of different binding modes to turnover. Efficient protocols using QM/MM umbrella sampling and MM/GSBA calculations were developed to determine reactivity and substrate binding affinity, respectively, in the different binding modes. This demonstrated that, while one binding mode is most reactive, non-negligible quantities of substrate could also be turned over through other modes. This appears to be due to the non-specific nature of the enzyme, where catalysis is largely the result of 'simple' complementary binding for the reactive substrate conformation, and thus that differences in reactivity come from the ability of the binding modes to support a more or less reactive conformation (in terms of proximity between reacting groups). Observing the reaction profile in solution provided further evidence that this is the case. However, there may be a further contribution to catalysis still unaccounted for in the most reactive mode, which could be coming from specific catalytic interactions with the enzyme. In particular, Tyr76 may enhance reactivity by withdrawing electron density from the dienophile. Therefore, future work could be to test for catalysis from this residue, for example by DFT level IRC calculations with this residue present, similar to those performed in a computational study of the homologous Pyrl4 (20). In addition, although they give an indication that the substrate may indeed bind in multiple binding modes (with different associated activities), the reported binding affinities still carry a large degree of uncertainty due to the shortcomings of the MM/GBSA approach applied. Mainly these the uncertainty in how solvation effects are captured, and the lack of an entropy consideration for the ligand and receptor. One obvious avenue of future work would therefore be to determine binding affinities with greater confidence by at least including an entropy correction in the calculation.

In Chapter 5, the reactivity and binding affinity prediction protocols developed in Chapter 4 were used to predict the activity of AbyU variants with modified substrates. This was done both in a combinatorial fashion (testing all substrates with a set of lab-produced point mutants) as well as with rationally designed variants intended to enhance affinity for certain substrates. In terms of the combinatorial screening, it was found that most substrate/mutant combinations are likely to have the same nominal level of activity. This may be due to the non-specific nature of the enzyme, which meant that, given the resulting lack of significant differences found using the approximate protocols (and lack of existing experimental results for validation), strong predictions could not be made in identifying more or less active combinations. The main future work can therefore concentrate on validating these predictions by comparison with experimental kinetics, and refining them if necessary. To this end, greater sampling may be required to achieve higher precision and confidence in the predictions, so that they are able to resolve more subtle differences in reactivity/binding (and therefore be of more practical use for slightly modified substrates and enzyme variants). In addition, accuracy of reaction barrier estimates may be improved by performing further benchmarking of the semi-empirical QM method, and establishing correction factors for individual substrates for example. In terms of the binding affinity protocol, entropy corrections could also be added, at least when comparing more different substrates (where the assumption of similar binding entropy may therefore no longer hold). Work can then focus on an iterative design process using an improved set of protocols, and starting with the best combinations from the initial screening (following experimental validation). Additional mutations can be proposed for these and screened using the validated/improved protocols to find variants which are an even better fit for the modified substrates, giving higher affinity, and closer docking of reactive groups. For the rational redesign strategy, the main finding was that redesigning based on the reaction product is too naive an approach to reliably achieve an improvement in productive substrate binding. Therefore, future efforts could focus on a design process which targets the reactive substrate conformation (or transition state) itself. Improvement of the screening protocols may also be useful for this engineering strategy. In particular, an entropy correction may be necessary where more drastic changes that alter the flexibility of the binding site are concerned. Finally, any redesign strategy should likely also incorporate the influence of other (potentially less reactive) binding modes into activity prediction workflows, as these may otherwise confound attempts to screen for more active enzyme/substrate combinations.

In Chapter 6, stereoselectivity was studied in the homologous AbyU/AbmU pair, and the origin of their opposite stereoselectivities ascertained. To give insight into this, both reactivity and binding affinity for binding modes corresponding to the two stereo-products in question was predicted. Given the finding of earlier chapters, reactive distances in the Michaelis complex were used to predict reactivity and so the Michaelis complexes alone were simulated. This was done using effectively the same efficient protocol used for binding affinity calculation in the earlier chapters. It was found that, in AbyU, *R* selectivity comes from the enzyme supporting a much more reactive pro-*R* than pro-*S* complex. For AbmU, the switch to *S* selectivity appears to be achieved by an energetic disfavouring of the reactive pro-*R* mode (e.g. by disruption of the key Tyr76 interaction) along with promotion of the reactivity of the existing competent pro-*S* mode. As well as providing insight into how stereoselectivity can be achieved, the protocols used in this chapter provide a good recipe for stereoselectivity prediction that can be used to aid in the design of stereoselective spirotetronate forming enzymes. In future, these protocols could be combined with those developed above for engineering substrate scope, in order to search for active variants that also demonstrate stereochemical control.

While the preceding work on the enzyme has focused on more approximate and computationally efficient simulations, in Chapter 7 more sophisticated and longer timescale periodic boundary MD simulations were used to study the large conformational change involved with the enzyme capping loop. Conventional MD (cMD) simulations were first used to determine whether the potentially (partially) rate-limiting process of loop opening for product release may be what is causing a reduction in activity at high pressure. Although the simulations support stability of the AbyU barrel at high pressure, insufficient data was gathered to make a significant comparison on differences in loop opening. Future work could thus focus on obtaining more sampling of loop opening at different pressures to enable a statistically significant comparison. This could be done either by running many more cMD simulations, or via enhanced sampling techniques. cMD was then also used to determine the effect of mutation on the loop and how this relates to activity. It was found that introducing a polar residue too greatly destabilised the closed loop conformation, thereby likely detrimentally impacting substrate capture and thus activity. More subtle mutations that slightly weaken hydrophobic loop interactions were therefore focused on, as these may potentially enhance activity by easing product release without greatly compromising substrate capture. Using Gaussian accelerated MD, mutations that subtly change the energetics of loop opening for the product complex were able to be found, highlighting I62V as the most promising mutant. Future work could therefore involve detailed kinetic characterisation for this mutant to determine whether activity has indeed been improved and whether there may still be further optimisation potential through this route, in case product release is still found to be rate limiting.

8.1 General outlook

Reflecting on the more general significance of this work in this thesis, in the first place it has demonstrated the continued and growing relevance of using a computational approach when investigating enzymatic catalysis. As seen, computation provides an invaluable tool, both for uncovering the subtleties of enzyme mechanisms and well as for engineering. The work in this thesis has demonstrated how multiple aspects of enzyme catalysis can be studied computationally, in order to develop a complete picture of enzyme function. By using different computational techniques and levels of description, this allows one to study individual aspects of an enzyme catalysed reaction at the atomistic level, accessing relevant timescales through sophisticated enhanced sampling techniques where necessary. This therefore allows one to gain molecular level insights that are unachievable by conventional experiment alone. In terms of its industrial relevance, this work has made strides in unlocking the biosynthetic potential of spirotetronate/tetramate DAases. These are well balanced and robust catalysts, which produce high value compounds and are therefore ripe for exploitation in biosynthesis. By firstly obtaining a thorough mechanistic understanding of how AbyU works this has contributed to a growing body of knowledge on how this class of DAase works in general, therefore helping to make them more harnessable and customisable. This work has already made in-roads in this pursuit and begun testing the limits and potential of AbyU for repurposing with the help of computation, thereby setting the stage for future engineering efforts. Specifically, this work has shown that the AbyU is a versatile catalyst, readily accepting substrates with slight modifications and providing reasonably good activity (for which its activity could be more precisely tuned in the future with the help or more refined protocols). The β-barrel architecture itself is found to also be a versatile platform which provides a good general starting point for more complete redesign and engineering with computational protein design (CPD) tools. In terms of predictive computational protocols (which are useful for guiding and evaluating the results of engineering), this work has shown the utility of efficient, but atomistically detailed simulations, for making reasonably precise catalytic activity predictions. These types of protocols thereby lend themselves to large scale *in silico* screening in which (for example) promising substrates can be identified, and their activity improved upon by iterative *in silico* screening rounds. As well as this, these protocols can be used for the high throughput screening of a large number of designs output by CPD procedures. While some preliminary attempts at engineering using both these approaches have been conducted on a relatively small scale, this serves mostly as demonstration and to help establish useful procedures. The procedures developed, and lessons learnt, within this work thus provide a useful example of a viable computational strategy for the engineering of β-barrel spirotetronate DAases, which can inform and empower future engineering efforts.

Chapter 9. Appendix

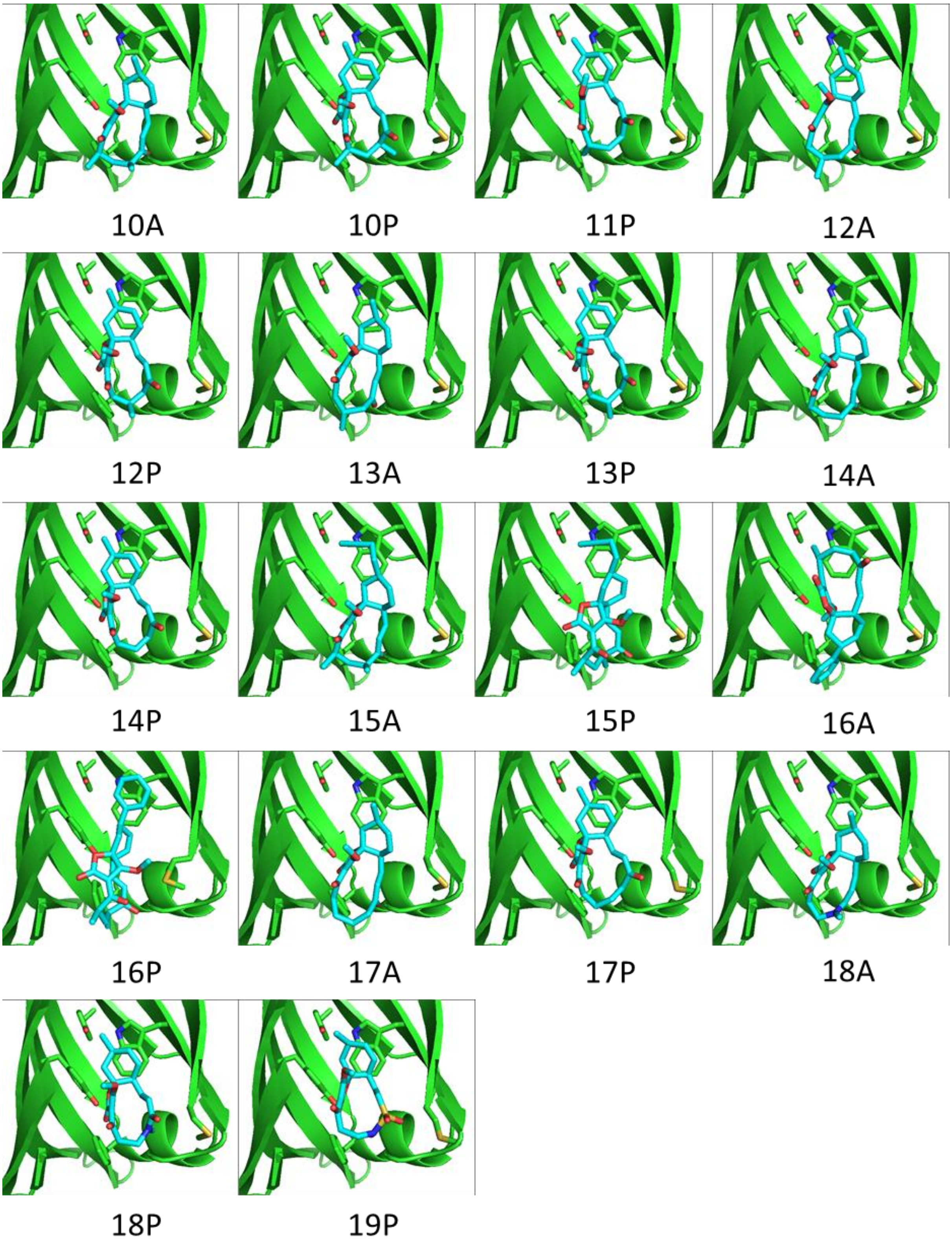
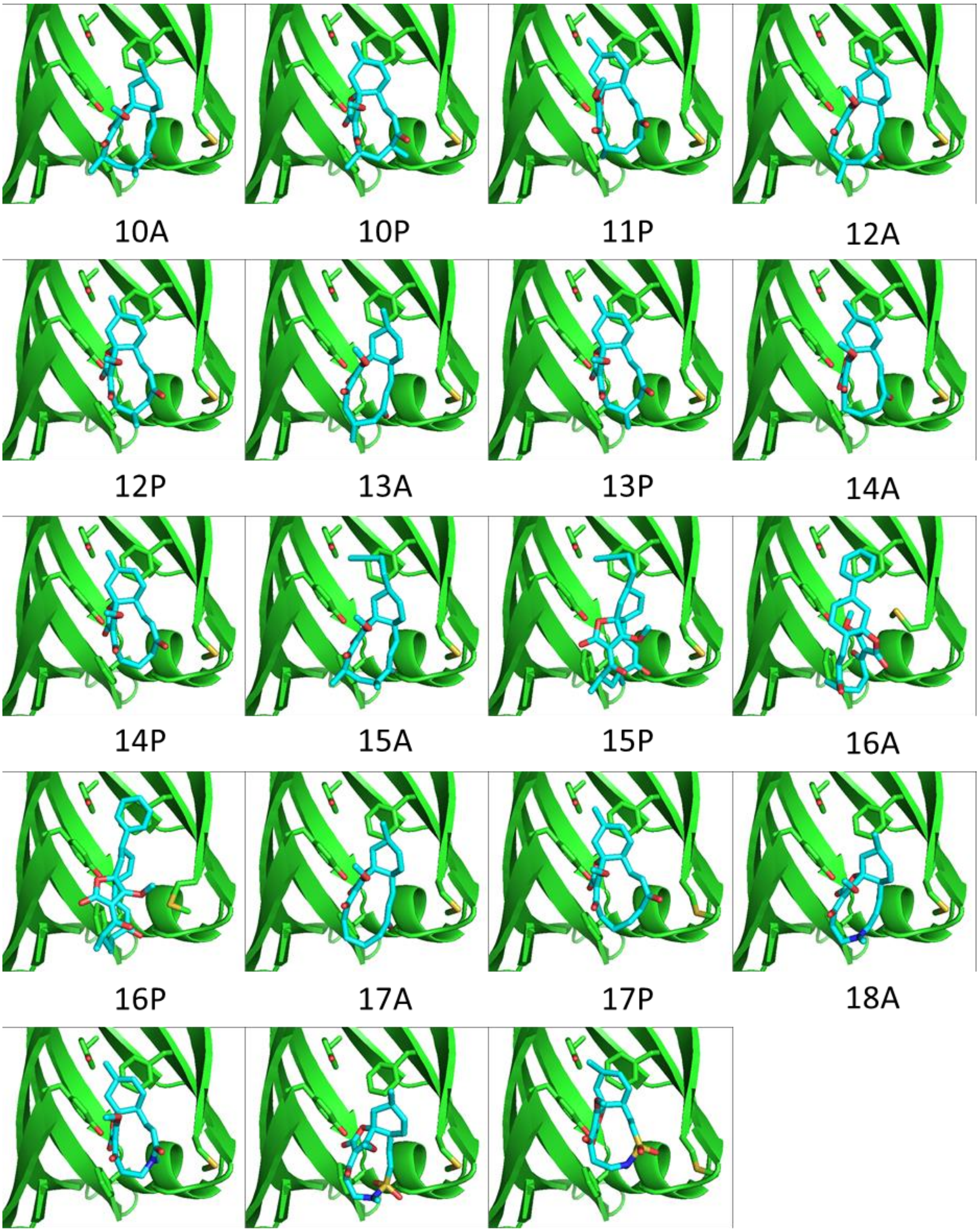


Figure 9.1: Docked poses selected for screening of the two atropisomer products of substrates **10**-**19** with the WT enzyme. Key residues treated flexibly in the docking, as well as Y106, are highlighted with sticks.



10A 10P 11P 12A

12P 13A 13P 14A

14P 15A 15P 16A

16P 17A 17P 18A

18P 19A 19P

Figure 9.2: Docked poses selected for screening of the two atropisomer products of substrates **10**-**19** with the W124F mutant. Key residues treated flexibly in the docking, as well as Y106, are highlighted with sticks.

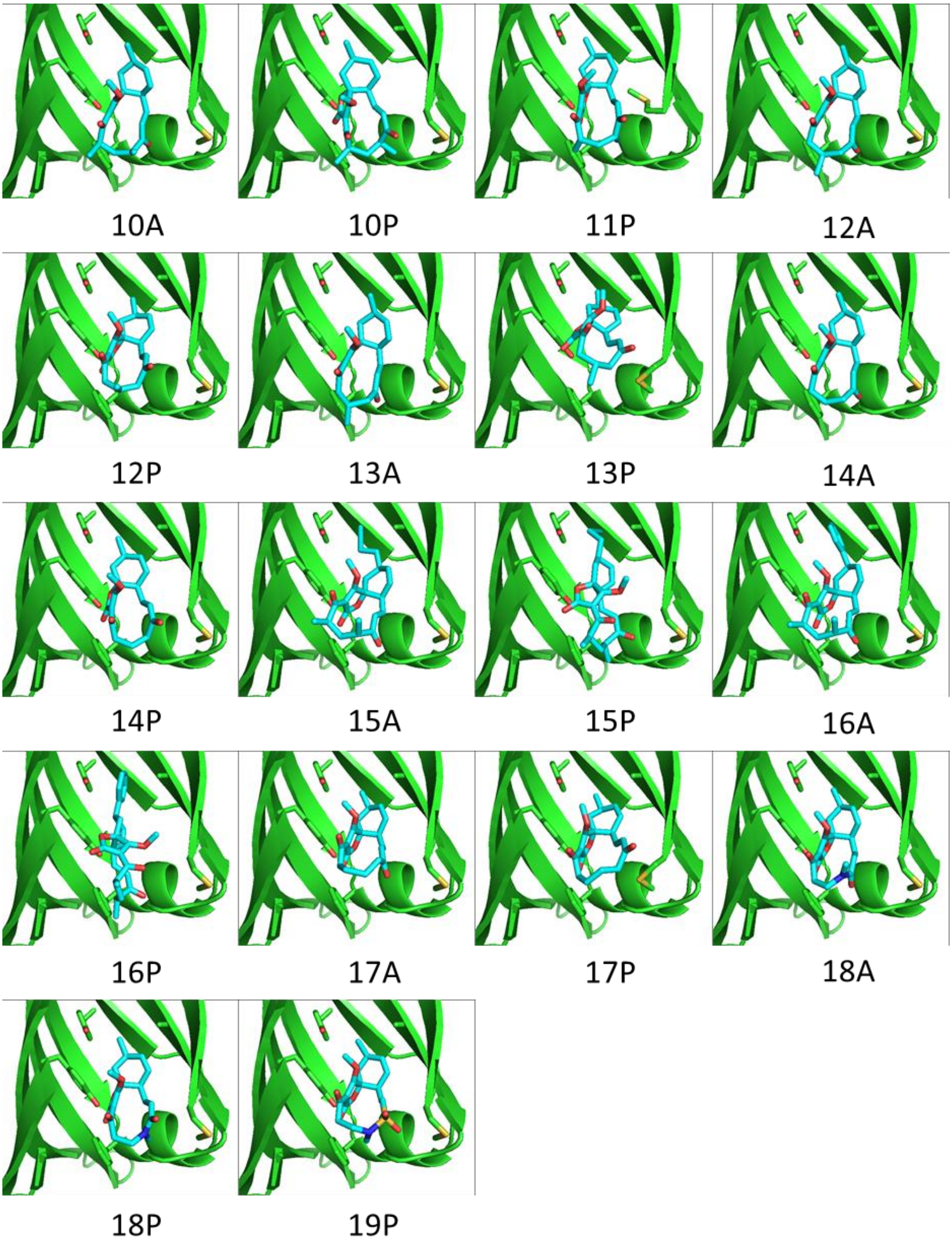


Figure 9.3: Docked poses selected for screening of the two atropisomer products of substrates **10**-**19** with the W124A mutant. Key residues treated flexibly in the docking, as well as Y106, are highlighted with sticks.

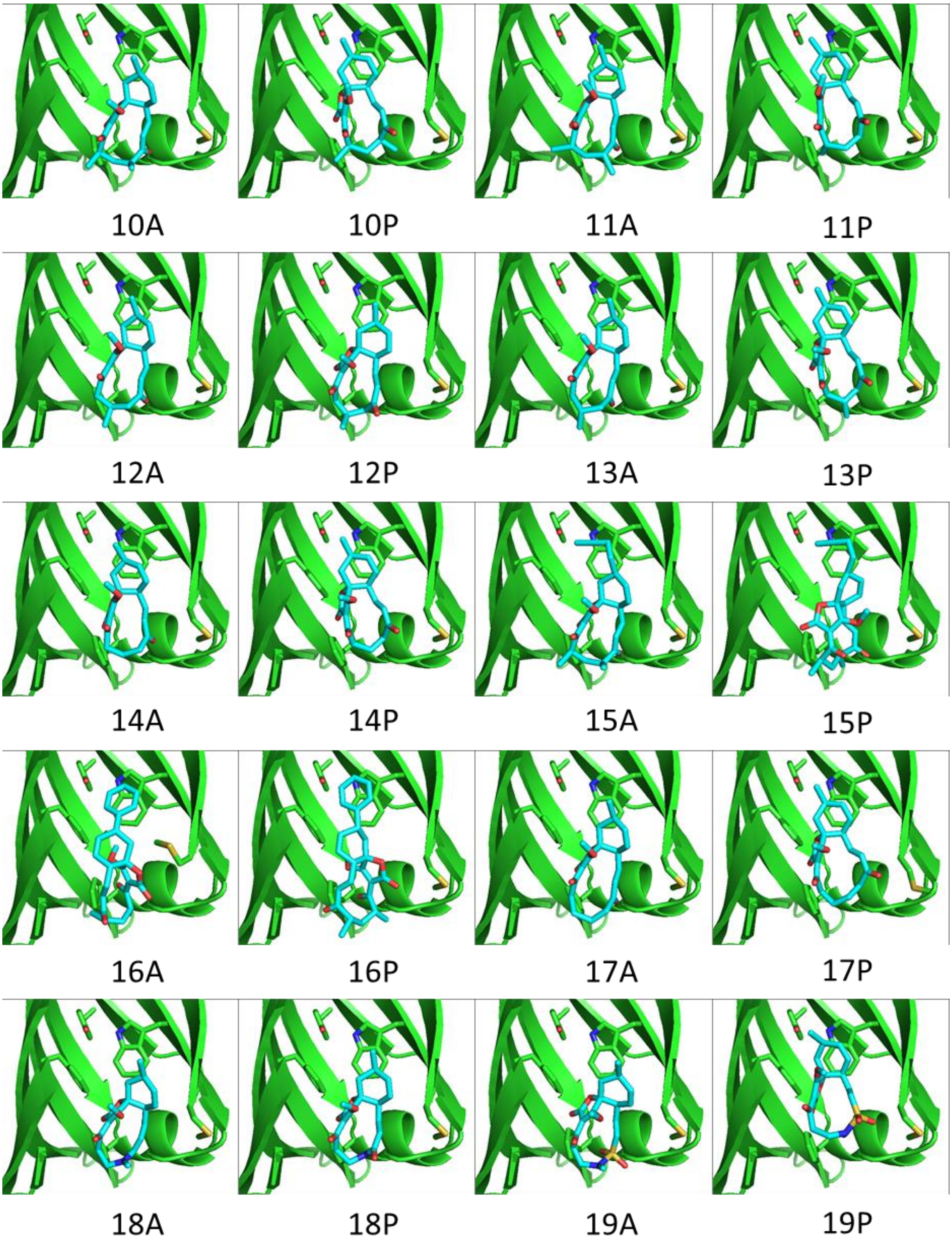


Figure 9.4: Docked poses selected for screening of the two atropisomer products of substrates **10**-**19** with the Y76F mutant. Key residues treated flexibly in the docking, as well as Y106, are highlighted with sticks.

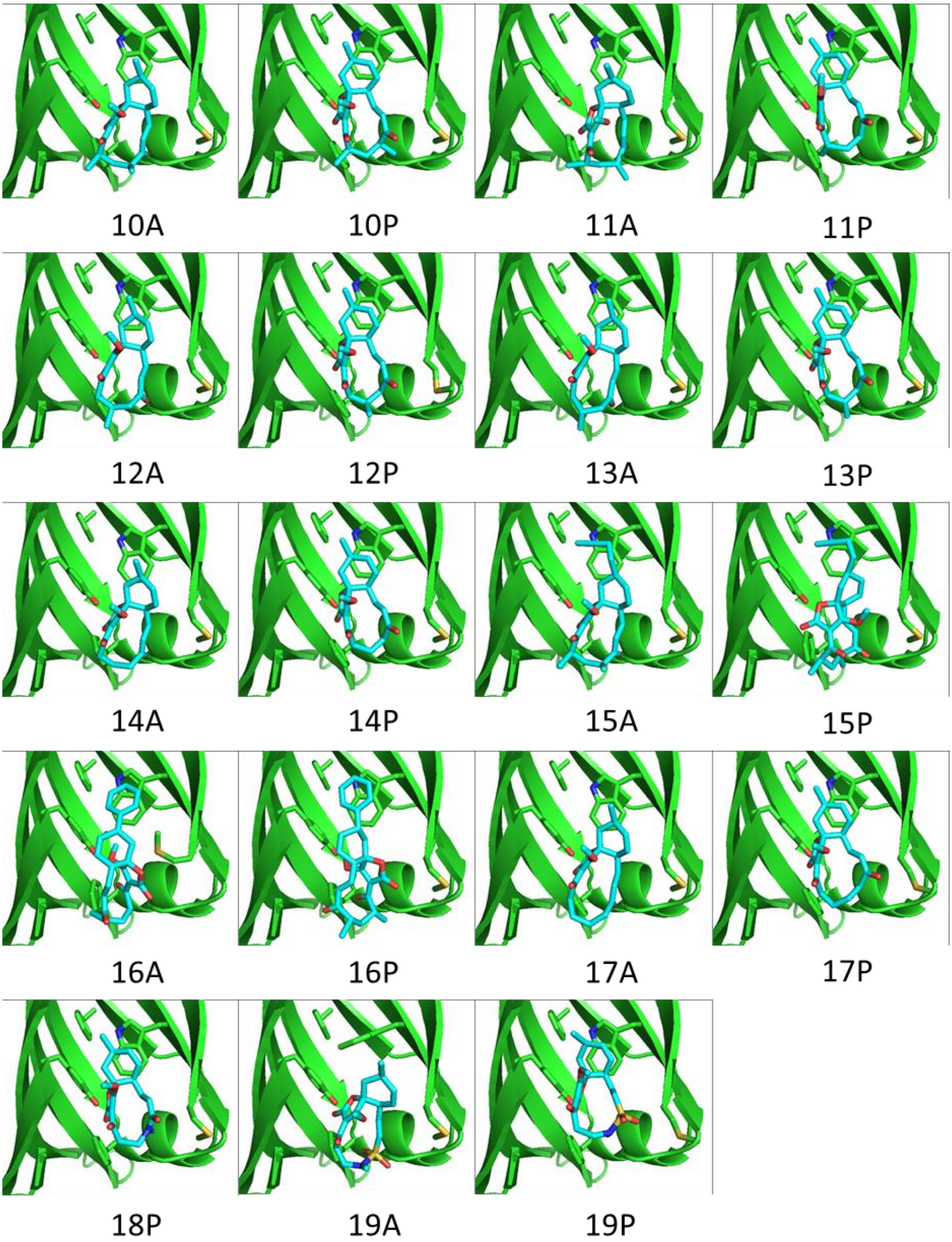


Figure 9.5: Docked poses selected for screening of the two atropisomer products of substrates **10**-**19** with the Y106F mutant. Residues treated flexibly in the docking are highlighted with sticks.

References

1. Nicolaou KC, Snyder SA, Montagnon T, Vassilikogiannakis G. The Diels-Alder reaction in total synthesis. Angewandte Chemie-International Edition. 2002;41(10):1668-98.

2. Minami A, Oikawa H. Recent advances of Diels-Alderases involved in natural product biosynthesis. J Antibiot. 2016;69(7):500-6.

3. Kim HJ, Ruszczycky MW, Liu HW. Current developments and challenges in the search for a naturally selected Diels-Alderase. Curr Opin Chem Biol. 2012;16(1-2):124-31.

4. Schmidt-Dannert C, Lopez-Gallego F. A roadmap for biocatalysis - functional and spatial orchestration of enzyme cascades. Microb Biotechnol. 2016;9(5):601-9.

5. Oikawa H, Katayama K, Suzuki Y, Ichihara A. Enzymatic-activity catalyzing exo-selective Diels-Alder Reaction in Solanapyrone biosynthesis. J Chem Soc-Chem Commun. 1995(13):1321-2.

6. Kennedy J, Auclair K, Kendrew SG, Park C, Vederas JC, Hutchinson CR. Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. Science. 1999;284(5418):1368-72.

7. Kato N, Nogawa T, Hirota H, Jang JH, Takahashi S, Ahn JS, et al. A new enzyme involved in the control of the stereochemistry in the decalin formation during equisetin biosynthesis. Biochem Biophys Res Commun. 2015;460(2):210-5.

8. Sato M, Yagishita F, Mino T, Uchiyama N, Patel A, Chooi YH, et al. Involvement of Lipocalin-like CghA in Decalin-Forming Stereoselective Intramolecular 4+2 Cycloaddition. Chembiochem. 2015;16(16):2294-8.

9. Kakule TB, Jadulco RC, Koch M, Janso JE, Barrows LR, Schmidt EW. Native Promoter Strategy for High-Yielding Synthesis and Engineering of Fungal Secondary Metabolites. ACS Synth Biol. 2015;4(5):625-33.

10. Qiao KJ, Chooi YH, Tang Y. Identification and engineering of the cytochalasin gene cluster from Aspergillus clavatus NRRL 1. Metab Eng. 2011;13(6):723-32.

11. Tian Z, Sun P, Yan Y, Wu Z, Zheng Q, Zhou S, et al. An enzymatic [4+2] cyclization cascade creates the pentacyclic core of pyrroindomycins. Nat Chem Biol. 2015;11(4):259-65.

12. Zheng QF, Gong YK, Guo YJ, Zhao ZX, Wu ZH, Zhou ZX, et al. Structural Insights into a Flavin-Dependent 4+2 Cyclase that Catalyzes trans-Decalin Formation in Pyrroindomycin Biosynthesis. Cell Chemical Biology. 2018;25(6):718-+.

13. Hashimoto T, Hashimoto J, Teruya K, Hirano T, Shin-Ya K, Ikeda H, et al. Biosynthesis of Versipelostatin: Identification of an Enzyme-Catalyzed 4+2 -Cycloaddition Required for Macrocyclization of Spirotetronate-Containing Polyketides. Journal of the American Chemical Society. 2015;137(2):572-5.

14. Jia XY, Tian ZH, Shao L, Qu XD, Zhao QF, Tang J, et al. Genetic characterization of the chlorothricin gene cluster as a model for spirotetronate antibiotic biosynthesis. Chem Biol. 2006;13(6):575-85.

15. Kim HJ, Ruszczycky MW, Choi SH, Liu YN, Liu HW. Enzyme-catalysed 4+2 cycloaddition is a key step in the biosynthesis of spinosyn A. Nature. 2011;473(7345):109-12.

16. Fage CD, Isiorho EA, Liu Y, Wagner DT, Liu HW, Keatinge-Clay AT. The structure of SpnF, a standalone enzyme that catalyzes [4 + 2] cycloaddition. Nat Chem Biol. 2015;11(4):256-8.

17. Patel A, Chen Z, Yang ZY, Gutierrez O, Liu HW, Houk KN, et al. Dynamically Complex 6+4 and 4+2 Cycloadditions in the Biosynthesis of Spinosyn A. Journal of the American Chemical Society. 2016;138(11):3631-4.

18. Yang Z, Yang S, Yu P, Li Y, Doubleday C, Park J, et al. Influence of water and enzyme SpnF on the dynamics and energetics of the ambimodal [6+4]/[4+2] cycloaddition. Proc Natl Acad Sci U S A. 2018;115(5):E848-E55.

19. Zheng Q, Guo Y, Yang L, Zhao Z, Wu Z, Zhang H, et al. Enzyme-Dependent [4 + 2] Cycloaddition Depends on Lid-like Interaction of the N-Terminal Sequence with the Catalytic Core in PyrI4. Cell Chem Biol. 2016;23(3):352-60.

20. Yike Z, Houk KN. Computational Investigation of the Mechanism of Diels−Alderase PyrI4. 2020;142:20232-9.

21. Hashimoto T, Kuzuyama T. Mechanistic insights into Diels-Alder reactions in natural product biosynthesis. Curr Opin Chem Biol. 2016;35:117-23.

22. Riedlinger J, Reicke A, Zahner H, Krismer B, Bull AT, Maldonado LA, et al. Abyssomicins, inhibitors of the para-aminobenzoic acid pathway produced by the marine Verrucosispora strain AB-18-032. J Antibiot. 2004;57(4):271-9.

23. Freundlich JS, Lalgondar M, Wei JR, Swanson S, Sorensen EJ, Rubin EJ, et al. The Abyssomicin C family as in vitro inhibitors of Mycobacterium tuberculosis. Tuberculosis. 2010;90(5):298-300.

24. Sadaka C, Ellsworth E, Hansen PR, Ewin R, Damborg P, Watts JL. Review on Abyssomicins: Inhibitors of the Chorismate Pathway and Folate Biosynthesis. Molecules. 2018;23(6):25.

25. Byrne MJ, Lees NR, Han LC, van der Kamp MW, Mulholland AJ, Stach JE, et al. The Catalytic Mechanism of a Natural Diels-Alderase Revealed in Molecular Detail. J Am Chem Soc. 2016;138(19):6095-8.

26. Drulyte I, Obajdin J, Trinh CH, Kalverda AP, van der Kamp MW, Hemsworth GR, et al. Crystal structure of the putative cyclase IdmH from the indanomycin nonribosomal peptide synthase/polyketide synthase. IUCrJ. 2019;6:1120-33.

27. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. De novo enzyme design using Rosetta3. PLoS One. 2011;6(5):e19230.

28. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, Clair JLS, et al. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. Science. 2010;329(5989):309-13.

29. Preiswerk N, Beck T, Schulz JD, Milovnik P, Mayer C, Siegel JB, et al. Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. Proc Natl Acad Sci U S A. 2014;111(22):8013-8.

30. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, Shen BW, et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotechnol. 2012;30(2):190-2.

31. Braisted AC, Schultz PG. An antibody-catalyzed bimolecular Diels-Alder reaction. Journal of the American Chemical Society. 1990;112(20):7430-1.

32. Hilvert D, Hill KW, Nared KD, Auditor MTM. Antibody catalysis of a Diels-Alder reaction. Journal of the American Chemical Society. 1989;111(26):9261-2.

33. Oikawa H. Nature's Strategy for Catalyzing Diels-Alder Reaction. Cell Chem Biol. 2016;23(4):429-30.

34. D.A. Case RMB, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke,, A.W. Goetz NH, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C., Lin TL, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I., Omelyan AO, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails,, R.C. Walker JW, R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman. AMBER 2016. University of California, San Francisco 2016.

35. Zinovjev KvdK, M. W. Enlighten2: Molecular Dynamics Simulations of Protein-Ligand Systems Made Accessible. ChemRxiv. 2020.

36. Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. Journal of Computational Chemistry. 2000;21(2):132-46.

37. Born M, Oppenheimer R. Quantum theory of molecules. Ann Phys-Berlin. 1927;84(20):0457-84.

38. Fock V. Approximation method for the solution of the quantum mechanical multibody problems. Z Phys. 1930;61(1-2):126-48.

39. Cizek J. On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods. J Chem Phys. 1966;45(11):4256-+.

40. Moller C, Plesset MS. Note on an approximation treatment for many-electron systems. Phys Rev. 1934;46(7):0618-22.

41. Zhao Y, Truhlar DG. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor Chem Acc. 2008;120(1-3):215-41.

42. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. The development and use of quantum-mechanical molecular models. 76. AM1 - A new general-purpose quantum-mechanical molecular-model. Journal of the American Chemical Society. 1985;107(13):3902-9.

43. Stewart JJP. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model. 2007;13(12):1173-213.

44. Christensen AS, Kubar T, Cui Q, Elstner M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. Chem Rev. 2016;116(9):5301-37.

45. Porezag D, Frauenheim T, Kohler T, Seifert G, Kaschner R. Construction of tight-binding-like potentials on the basis of density-functional theory - application to carbon. Phys Rev B. 1995;51(19):12947-57.

46. Elstner M, Porezag D, Jungnickel G, Elsner J, Haugk M, Frauenheim T, et al. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. Phys Rev B. 1998;58(11):7260-8.

47. Gruden M, Andjeklovic L, Jissy AK, Stepanovic S, Zlatar M, Cui Q, et al. Benchmarking density functional tight binding models for barrier heights and reaction energetics of organic molecules. Journal of Computational Chemistry. 2017;38(25):2171-85.

48. van der Kamp MW, Mulholland AJ. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry. 2013;52(16):2708-28.

49. Gao JL. Perspective on "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme" - Warshel A, Levitt M (1976) J Mol Biol 103 : 227-249. Theor Chem Acc. 2000;103(3-4):328-9.

50. Field MJ, Bash PA, Karplus M. A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS. Journal of Computational Chemistry. 1990;11(6):700-33.

51. Amara P, Field MJ. Evaluation of an ab initio quantum mechanical/molecular mechanical hybrid-potential link-atom method. Theor Chem Acc. 2003;109(1):43-52.

52. Verlet L. Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Phys Rev. 1967;159(1):98-+.

53. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR. Molecular-dynamics with coupling to an external bath. J Chem Phys. 1984;81(8):3684-90.

54. Petersen HG. Accuracy and efficiency of the Particle Mesh Ewald method. J Chem Phys. 1995;103(9):3668-79.

55. Haydock C, Sharp JC, Prendergast FG. TRYPTOPHAN-47 ROTATIONAL ISOMERIZATION IN VARIANT-3 SCORPION NEUROTOXIN - A COMBINATION THERMODYNAMIC PERTURBATION AND UMBRELLA SAMPLING STUDY. Biophys J. 1990;57(6):1269-79.

56. Shen J, McCammon JA. MOLECULAR-DYNAMICS SIMULATION OF SUPEROXIDE INTERACTING WITH SUPEROXIDE-DISMUTASE. Chem Phys. 1991;158(2-3):191-8.

57. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. Journal of Computational Chemistry. 1992;13(8):1011-21.

58. Roux B. THE CALCULATION OF THE POTENTIAL OF MEAN FORCE USING COMPUTER-SIMULATIONS. Comput Phys Commun. 1995;91(1-3):275-82.

59. Grossfield A. WHAM: an implementation of the weighted histogram analysis method [Available from: http://membrane.urmc.rochester.edu/content/wham/.

60. Miao YL, Feher VA, McCammon JA. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. Journal of Chemical Theory and Computation. 2015;11(8):3584-95.

61. Miller BR, 3rd, McGee TD, Jr., Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J Chem Theory Comput. 2012;8(9):3314-21.

62. Weiser J, Shenkin PS, Still WC. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). Journal of Computational Chemistry. 1999;20(2):217-30.

63. Mongan J, Simmerling C, McCammon JA, Case DA, Onufriev A. Generalized Born model with a simple, robust molecular volume correction. Journal of Chemical Theory and Computation. 2007;3(1):156-69.

64. Beno BR, Houk KN, Singleton DA. Synchronous or Asynchronous? An “Experimental” Transition State from a Direct Comparison of Experimental and Theoretical Kinetic Isotope Effects for a Diels−Alder Reaction. J Am Chem Soc. 1996;118:9984-5.

65. Anslyn Eric DD. Modern Physical Organic Chemistry.

66. Cui CX, Liu YJ. A thorough understanding of the Diels-Alder reaction of 1,3-butadiene and ethylene. J Phys Org Chem. 2014;27(8):652-60.

67. Smyth JE, Butler NM, Keller PA. A twist of nature - the significance of atropisomers in biological systems. Natural Product Reports. 2015;32(11):1562-83.

68. Bister B, Bischoff D, Strobele M, Riedlinger J, Reicke A, Wolter F, et al. Abyssomicin C - A polycyclic antibiotic from a marine Verrucosispora strain as an inhibitor of the p-aminobenzoic acid/tetrahydrofolate biosynthesis pathway. Angewandte Chemie-International Edition. 2004;43(19):2574-6.

69. Zapf CW, Harrison BA, Drahl C, Sorensen EJ. A Diels-Alder macrocyclization enables an efficient asymmetric synthesis of the antibacterial natural product abyssomicin C. Angewandte Chemie-International Edition. 2005;44(40):6533-7.

70. Becke AD. Density-functional Thermochemistry. 3. The role of exact exchange. J Chem Phys. 1993;98(7):5648-52.

71. Goumans TPM, Ehlers AW, Lammertsma K, Wurthwein EU, Grimme S. Improved reaction and activation energies of 4+2 cycloadditions, 3,3 sigmatropic rearrangements and electrocyclizations with the spin-component-scaled MP2 method. Chem-Eur J. 2004;10(24):6468-75.

72. Pieniazek SN, Clemente FR, Houk KN. Sources of error in DFT computations of C-C bond formation thermochemistries: pi ->sigma transformations and error cancellation by DFT methods. Angewandte Chemie-International Edition. 2008;47(40):7746-9.

73. Pham HV, Martin DBC, Vanderwal CD, Houk KN. The intramolecular Diels-Alder reaction of tryptamine-derived Zincke aldehydes is a stepwise process. Chem Sci. 2012;3(5):1650-5.

74. Linder M, Brinck T. Stepwise Diels-Alder: More than Just an Oddity? A Computational Mechanistic Study. Journal of Organic Chemistry. 2012;77(15):6563-73.

75. Linder M, Johansson AJ, Manta B, Olsson P, Brinck T. Envisioning an enzymatic Diels-Alder reaction by in situ acid-base catalyzed diene generation. Chem Commun. 2012;48(45):5665-7.

76. Linder M, Brinck T. On the method-dependence of transition state asynchronicity in Diels-Alder reactions. Phys Chem Chem Phys. 2013;15(14):5108-14.

77. Allinger NL. Conformational-analysis. 130. MM2 - Hydrocarbon force-field utilizing V1 and V2 torsional terms. Journal of the American Chemical Society. 1977;99(25):8127-34.

78. Miertus S, Scrocco E, Tomasi J. Electrostatic interaction of a solute with a continuum - a direct utilization of ab-initio molecular potentials for the prevision of solvent effects. Chem Phys. 1981;55(1):117-29.

79. James NC, Um JM, Padias AB, Hall HK, Houk KN. Computational Investigation of the Competition between the Concerted Diels-Alder Reaction and Formation of Diradicals in Reactions of Acrylonitrile with Nonpolar Dienes. Journal of Organic Chemistry. 2013;78(13):6582-92.

80. Mezei PD, Csonka GI, Kallay M. Accurate Diels-Alder Reaction Energies from Efficient Density Functional Calculations. Journal of Chemical Theory and Computation. 2015;11(6):2879-88.

81. Garcia JI, Mayoral JA, Salvatella L. Is it 4+2 or 2+4 ? A new look at Lewis acid catalyzed Diels-Alder reactions. Journal of the American Chemical Society. 1996;118(46):11680-1.

82. Yamabe S, Dai TS, Minato T. Fine-tuning [4+2] and [2+4] Diels-Alder reactions catalyzed by Lewis-acids. Journal of the American Chemical Society. 1995;117(44):10994-7.

83. Galli C, Mandolini L. The role of ring strain on the ease of ring closure of bifunctional chain molecules. European Journal of Organic Chemistry. 2000;2000(18):3117-25.

84. Page MI, Jencks WP. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. Proc Natl Acad Sci U S A. 1971;68(8):1678-83.

85. Chan L, Morris GM, Hutchison GR. Understanding Conformational Entropy in Small Molecules. Journal of Chemical Theory and Computation. 2021;17(4):2099-106.

86. Swiderek K, Moliner V. Computational Studies of Candida Antarctica Lipase B to Test Its Capability as a Starting Point To Redesign New Diels-Alderases. J Phys Chem B. 2016;120(8):2053-70.

87. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.

88. Sondergaard CR, Olsson MH, Rostkowski M, Jensen JH. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. J Chem Theory Comput. 2011;7(7):2284-95.

89. Olsson MH, Sondergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput. 2011;7(2):525-37.

90. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015;11(8):3696-713.

91. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 2004;25(9):1157-74.

92. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. Journal of Chemical Theory and Computation. 2013;9(7):3084-95.

93. Kuhn M, Firth-Clark S, Tosco P, Mey A, Mackey M, Michel J. Assessment of Binding Affinity via Alchemical Free-Energy Calculations. J Chem Inf Model. 2020;60(6):3120-30.

94. Woods CJ, Malaisree M, Hannongbua S, Mulholland AJ. A water-swap reaction coordinate for the calculation of absolute protein-ligand binding free energies. J Chem Phys. 2011;134(5):13.

95. Pearlman DA. Evaluating the molecular mechanics Poisson-Boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. J Med Chem. 2005;48(24):7796-807.

96. Bea I, Cervello E, Kollman PA, Jaime C. Molecular recognition by beta-cyclodextrin derivatives: FEP vs MM/PBSA study. Comb Chem High Throughput Screen. 2001;4(8):605-11.

97. Camacho CJ. Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets. Proteins. 2005;60(2):245-51.

98. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. In: Johnson ML, Brand L, editors. Methods in Enzymology, Vol 487: Computer Methods, Pt C. Methods in Enzymology. San Diego: Elsevier Academic Press Inc; 2011. p. 545-74.

99. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009;77(4):778-95.

100. Ochoa R, Soler MA, Laio A, Cossio P. Assessing the capability of in silico mutation protocols for predicting the finite temperature conformation of amino acids. Phys Chem Chem Phys. 2018;20(40):25901-9.

101. Li R, Wijma HJ, Song L, Cui Y, Otzen M, Tian Y, et al. Computational redesign of enzymes for regio- and enantioselective hydroamination. Nat Chem Biol. 2018;14(7):664-70.

102. Wijma HJ, Floor RJ, Bjelic S, Marrink SJ, Baker D, Janssen DB. Enantioselective Enzymes by Computational Design and In Silico Screening. Angewandte Chemie-International Edition. 2015;54(12):3726-30.

103. Rocco M, Bender BJ, Allison B, Meiler J. Rosetta and the Design of Ligand Binding Sites. Methods in Molecular Biology. 2016;1414:47-62.

104. Tu JJ, Li ST, Chen J, Song YX, Fu SB, Ju JH, et al. Characterization and heterologous expression of the neoabyssomicin/abyssomicin biosynthetic gene cluster from Streptomyces koyangensis SCSIO 5802. Microb Cell Fact. 2018;17:14.

105. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725-38.

106. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. 2003;31(14):3982-92.

107. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci. 2000;9(9):1753-73.

108. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10(5):449-61.

109. Hummer G, Garde S, Garcia AE, Paulaitis ME, Pratt LR. The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(4):1552-5.

110. Gladstone S LK, Eyring H. The Theory of Rate Processes. New York: McGraw–Hill; 1951.

111. Abascal JL, Vega C. A general purpose model for the condensed phases of water: TIP4P/2005. J Chem Phys. 2005;123(23):234505.

112. Kabsch W, Sander C. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.

113. Miao YL, Sinko W, Pierce L, Bucher D, Walker RC, McCammon JA. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. Journal of Chemical Theory and Computation. 2014;10(7):2677-89.

114. Huang Q, Rodgers JM, Hemley RJ, Ichiye T. Adaptations for pressure and temperature effects on loop motion in Escherichia coli and Moritella profunda dihydrofolate reductase. High Pressure Res. 2019;39(2):225-37.

115. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. Proteins. 2011;79(4):1089-108.

116. Miao YL, Huang YMM, Walker RC, McCammon JA, Chang CEA. Ligand Binding Pathways and Conformational Transitions of the HIV Protease. Biochemistry. 2018;57(9):1533-41.

117. Palermo G, Miao YL, Walker RC, Jinek M, McCammon JA. CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(28):7260-5.

118. Liao Q, Kulkarni Y, Sengupta U, Petrović D, Mulholland AJ, van der Kamp MW, et al. Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. J Am Chem Soc. 2018;140:15889-903.

119. Rozovsky S, McDermott AE. The time scale of the catalytic loop motion in triosephosphate isomerase. Journal of Molecular Biology. 2001;310(1):259-70.

120. Bussi G. Hamiltonian replica exchange in GROMACS: a flexible implementation. Mol Phys. 2014;112(3-4):379-84.

121. Affentranger R, Tavernelli I, Di Iorio EE. A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. Journal of Chemical Theory and Computation. 2006;2(2):217-28.

122. Pan AC, Weinreich TM, Piana S, Shaw DE. Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems. Journal of Chemical Theory and Computation. 2016;12(3):1360-7.

123. Piana S, Laio A. A bias-exchange approach to protein folding. Journal of Physical Chemistry B. 2007;111(17):4553-9.

124. Laio A, Parrinello M. Escaping free-energy minima. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(20):12562-6.

125. Crean RM, Biler M, van der Kamp MW, Hengge AC, Kamerlin SCL. Loop Dynamics and Enzyme Catalysis in Protein Tyrosine Phosphatases. Journal of the American Chemical Society. 2021;143(10):3830-45.