Computational Chemistry Studies of Proteins and Organic Reactions

By

YUE ZHANG
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____

Dean J. Tantillo, Chair

____

Justin B. Siegel

____

Vladimir Yarov-Yarovoy

Committee in Charge

2021

Acknowledgements

During my five years of computational chemistry studying at UC Davis, I have received support from supervisors, mentors, peers, families, and friends.

I would like to thank my supervisors, Professor Dean Tantillo and Professor Justin Siegel. I always feel I’m very lucky to work in these two labs. Back in undergrad when I was first exposed to computational chemistry and some of Ken Hauk and David Baker’s work, I never thought I’d have a chance to do research in both areas. I learned a lot about how to approach scientific questions from Dean and Justin. I wish one day I could have as much organic chemistry knowledge and chemical intuition as Dean does and I tried my best to explore different projects and learn more from Dean and other lab mates. I really enjoy working with Justin and doing projects related to environment and health is always my interest. I learned not only how to be a good scientist but also, I had a chance to learn about how to be a good presenter and get knowledge on business related things. Without their support I can never achieve what I’ve accomplished.

I want to thank Professor Vladimir Yarov-Yarovoy for being my dissertation committee and my QE committee. I've got a lot of valuable feedbacks from him and I learned so much from the Rosetta brainstorm meetings. I would like to thank my mentors and lab mates in Tantillo lab and Siegel lab. Terry O'Brien has been mentoring me for my first two years in graduate school. Switching from doing experiments to computational research can be challenging at the beginning, and Terry offered me so much help. I'm very grateful I had him as my mentor. Steve Bertolani and Stephanie Contreras helped me a lot on my homology modeling project. I learned so many scripting and coding from them. I'd like to thank people from Tantillo lab and Siegel lab for their support: Innes Tsui, Croix Laconsay, Amy Bellinghiere, Volga Kojasoy, Zhitao Feng, Phil Gringrich, Shunyang Wang, Gil Fernandez, Wentao Guo, Teresa Ortega, Yusef Ahmed, Kevin Kong, Supreeth Prasad, Will Desnoo, Ryan Caster, Veronica Keys, Ashley Vater, Tim Coulther, Youtian Cui, Sambid Adhikari, Augustine Arredondo, Simon Chu, Morgan Connolly, Pam Denish, Peishan Huang, Xiaoqing Huang, Alex Kehl, Amir Lechner, Emma Luu, Tram Nguyen, Teresa Ortega, Duy Bao Pham, and Mary Riley. I really enjoyed discussing questions with them and having fun together. I would also like to thank my collaborators. Professor Reuben Peters, Meirong Jia and Mark Schmidt-Dannert has been my collaborators for terpene related projects, and I had great time working with them and achieving goals together. I'm grateful to have the chance to collaborate with Bogdan Barnych in Hammock's lab, Erin Doherty in Beal's lab, Nanhao Chen in Wang's lab, Jongmin Baek in BASF. I appreciate all the support from my parents and my friends. I'm so thankful that they are always there for me.

Abstract

Computational chemistry has a variety of applications, from understanding a phenomenon of an experiment, to broader applications such as material science, drug discovery, biocatalysis, and many other fields. The following chapters are focused on using computational chemistry tools to understand the mechanisms of enzymes or organic reactions, and to engineer enzymes that produce novel products. Different methods have been applied to achieve these goals, such as molecular mechanics methods, ab initio molecular dynamics, classical molecular dynamics, quantum mechanics (density functional methods), monte carlo methods and machine learning methods. The first chapter is focused on computational modeling of terpene synthase family, to understand the catalysis process of terpene synthase, and further engineer terpene synthase to produce unnatural products. The second chapter is focused on studying organic reaction mechanisms and catalysis using quantum mechanics and molecular dynamics methods. The third chapter is focused on protein structure prediction of terpene synthase.

Chapter 1. Mechanistic studies and enzyme engineering of terpene synthase

Terpene or terpenoid is the largest class of natural product, with more than 80 000 members. All of which derives from the simple five-carbon isoprene unit. The five-carbon unit can be ligated by prenyltransferases to produce C_{10}, C_{15}, C_{20}… linear structures. These linear structures can then be cyclized by terpene synthase (TPS) to produce multi-ring, multi-stereocenter, cyclic structures. This process usually involves multiple highly reactive carbocation intermediates, brings challenges for modeling and identifying productive binding orientation. *Terdockin* method is applied to study the catalysis mechanism of the diterpene synthase Rv3377c, as well as *ent*-kaurene synthase BjKS. Results reveal the mechanism for catalysis, which can be further applied to engineer terpene synthase to produce unnatural products. A few engineer efforts listed in the chapter was made to alter the product outcome for BjKS.

Chapter 2. Quantum mechanics studies of organic reactions

In the first section of this chapter, the source of the rate acceleration for carbocation cyclization in a biomimetic supramolecular cage is studied using quantum mechanics methods, molecular dynamics, and QM/MM methods. Previous experimental results indicate that the supramolecular cage increases the nazarov reaction rate by roughly 1 000 000 times. The electrocyclization step is slightly enhanced by the catalyst, suggested by computational modeling. The major contribution, indicated by the QM/MM studies, is the activation of the leaving groups, similar as terpene synthase activate the diphosphate leaving group.

Chapter 3. Protein structure prediction of terpene synthase

Terpene cyclases catalyze one of the most complex chemical reactions in biology, converting simple acyclic oligo-isoprenyl diphosphate substrate to complex polycyclic products via

carbocation intermediates. Many computational studies were carried out to illuminate the structural-function relationship of terpene cyclases, which generally rely on a crystal structure. However, among 15,000 terpene cyclases sequences, there are only about 30 types of terpene cyclase crystal structures. To fill in this gap, we proposed a comparative modeling approach to generate high resolution models of class I terpene cyclase, which can be further used as a starting point to predict the productive binding mode of the substrate and potentially set stage for the rational engineering of terpene cyclases.

Table of Contents

Chapter 1 Mechanistic studies and enzyme engineering of terpene synthase.

Portions of this chapter have been published in 1 and 2 and are reproduced with permission.
1 Zhang, Y.; Prach, L. M., O'Brien, T.E.; Dimaio, F.; Prigozhin, D. M.; Corn J.E., Alber, T.; Siegel, J.B.; Tantillo, D. J. Crystal Structure and Mechanistic Molecular Modeling Studies of Mycobacterium tuberculosis Diterpene Cyclase Rv3377c. Biochemistry 2020, 59, 47, 4507–4515.
2 Jia, M.*; Zhang, Y.* Siegel, J.B.; Tantillo, D. J.; Peters, R. J. Switching on a Nontraditional Enzymatic Base—Deprotonation by Serine in the ent-Kaurene Synthase from Bradyrhizobium japonicum. ACS Catal.2019, 9(10), 8867-8871.

1.1 Crystal Structure and Mechanistic Molecular Modeling Studies of Mycobacterium tuberculosis Diterpene Cyclase Rv3377c

This chapter is a publication, all crystal structural experimental data is provided by Lisa Prach.

1.1.1 Abstract

Terpenes are the largest class of natural product, with extensive chemical and structural diversity. Diterpenes, mostly isolated from plants and rarely prokaryotes, exhibit a variety of important biological activities and valuable applications, including providing antitumor and antibiotic pharmaceuticals. These natural products are constructed by terpene synthases, a class of enzymes that catalyze one of the most complex chemical reactions in biology: converting simple acyclic oligo-isoprenyl diphosphate substrates to complex polycyclic products via carbocation intermediates. Here we obtained the second ever crystal structure of a class II diterpene synthase from bacteria, tuberculosinol pyrophosphate synthase (i.e. Halimadienyl diphosphate synthase, Rv3377c, MtHPS, or Rv3377c) from Mycobacterium tuberculosis (Mtb). This enzyme transforms (E, E, E)-geranylgeranyl diphosphate (GGPP) into tuberculosinol pyrophosphate (Halimadienyl diphosphate). Rv3377c is part of the Mtb diterpene pathway along with Rv3378c, which converts tuberculosinol pyrophosphate to 1-tuberculosinyl adenosine (1-TbAd). This pathway was shown to only exist in virulent Mycobacterium species, but not in closely related avirulent species, and

was proposed to be involved in phagolysosome maturation arrest. To gain further insight into the reaction pathway and the mechanistically relevant enzyme substrate binding orientation, electronic structure calculation and docking studies of reaction intermediates were carried out. Results reveal a plausible binding mode of the substrate which can provide the information to guide future drug design and anti-infective therapies of this biosynthetic pathway.

1.1.2 Introduction

Terpenes and terpenoids comprise a class of natural products with over 80,000 members,[1] whose structures are extremely varied and display a wide range of functions. These compounds range from essential metabolites, such as sterols, to unique secondary metabolites involved in communication and defense by various organisms. Terpene biosynthesis involves the isoprenyl diphosphate synthase-catalyzed condensation of isopentyl diphosphate and dimethylallyl diphosphate[2] to form linear precursors such as geranyl diphosphate (C10), farnesyl diphosphate (C15) and geranylgeranyl diphosphate (C20). These linear substrates are then converted by terpene cyclases into complex, often polycyclic, monoterpenes, sesquiterpenes and diterpenes, respectively.[1,3,4], These terpenes are then functionalized, often by the addition of oxygen atoms, by additional biosynthetic enzymes.

While most known terpenes have been isolated from plants and fungi, various monoterpenes, sesquiterpenes, and diterpenes have been isolated from streptomycetes,[5–9] Cyanobacteria,[10] and Myxobacteria.[11] Mycobacterium tuberculosis, the causative agent of tuberculosis, also contains a gene encoding a diterpene cyclase,[12–14] and a broad lipidomics screen recently identified an abundant secreted diterpene, 1-tuberculosinyl adenosine (1-TbAd), as the end product of the Mtb diterpene pathway.(See detail in SI)[15] Two genes, Rv3377c and Rv3378c, are necessary and sufficient to produce 1-TbAd in the non-pathogenic related mycobacteria, Mycobacterium smegmatis.[16] A previous study has shown that survival of Mtb in macrophages worsened with artificial damage to the Rv3377c gene.17 Rv3378c was thought to function as a pyrophosphatase,18 but the crystal structure revealed that it is a diverged cis-prenyl transferase that links adenosine and tuberculosinol pyrophosphate to form 1-TbAd.15 (See SI for complete pathway mechanism) Purified Rv3377c, in the presence of Mg^{2+}, rapidly converts geranylgeranyl diphosphate (GGPP) into tuberculosinol pyrophosphate (see figure 1.1.1a for reaction mechanism), indicating that Rv3377c is a class-II diterpene cyclase enzyme.[14,19] Class II terpene cyclases generate a carbocation by protonation of the C=C π-bond in the terminal isoprene unit, which involves a DXDD motif. These cyclases frequently have a multi-domain architecture. In contrast, Class I terpene cyclases promote the disconnection of a pyrophosphate to generate an allylic carbocation. These enzymes contain a DDXXD motif20 that binds the substrate diphosphate with the support of Mg^{2+} ions, and are structurally related to isoprenyl diphosphate synthases.

Recent crystal structures of plant and bacterial Class II diterpene cyclases[21–23] have revealed the domain architecture of these enzymes, the fold of each domain, the nature of the substrate-binding sites, and the arrangement of catalytic residues in this protein family, although questions remain regarding the specific catalytically competent orientations of substrates. The plant enzymes share three domains (α, β and γ) with the active site in a deep cleft between the β and γ domains.[24–26] The catalytic acid in the characteristic DXDD motif shared by these enzymes is located at the bottom of the active-site cavity. The origin of product specificity for this widespread enzyme class remains a major unanswered question.

Compared to plant homologs, the sequence of Rv3377c is deeply diverged, (the sequence identity is 23% between Rv3377c and taxadiene synthase from Pacific Yew, see figure 2.1.1b) and Rv3377c is predicted to contain only the β and γ domains. To expand the structure-function correlates of this enzyme class, define the architecture of the substrate-binding site and promote inhibitor development, we determined the crystal structure of Mtb Rv3377c. This structure proved elusive for almost a decade after production of the initial crystals, and was finally solved using Rosetta-based molecular replacement.

The fold of Rv3377c matches that of the β and γ domains of the plant enzymes and includes an interdomain loop with a conserved ion pair at the tip. The general acid that initiates cyclization interacts with H341 at the bottom of a large substrate-binding cavity. Many aromatic residues line the cavity, as is expected for terpene synthases.[1–3] These amino acids include two tryptophan residues that may switch rotamers as the reaction proceeds (vide infra).

Electronic structure calculations demonstrate that a reaction mechanism that involves cyclization and a series of hydride shifts and a methyl group migration is energetically viable. The results of extensive docking studies provide a model for the binding of intermediate carbocations and a template for structure-aided inhibitor design.

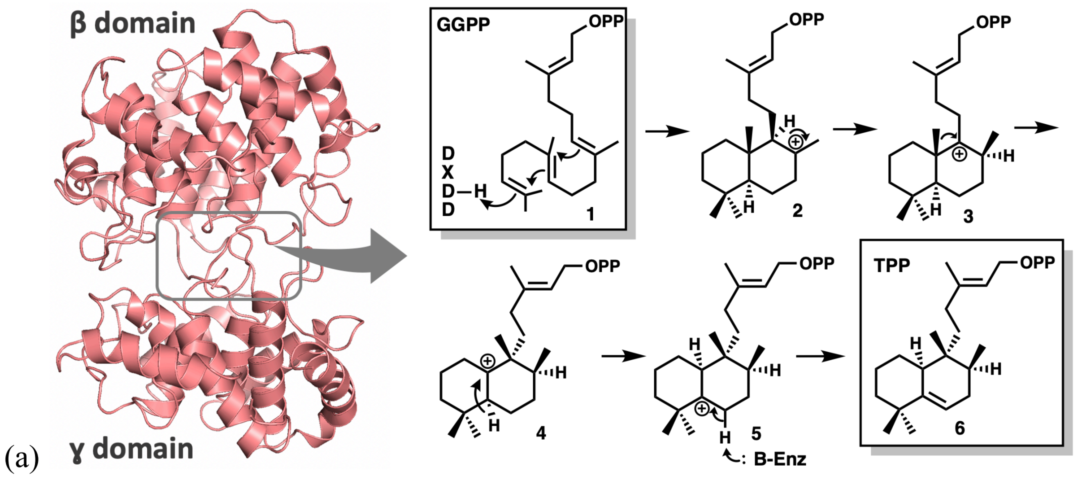
1.1.3 Results and discussion

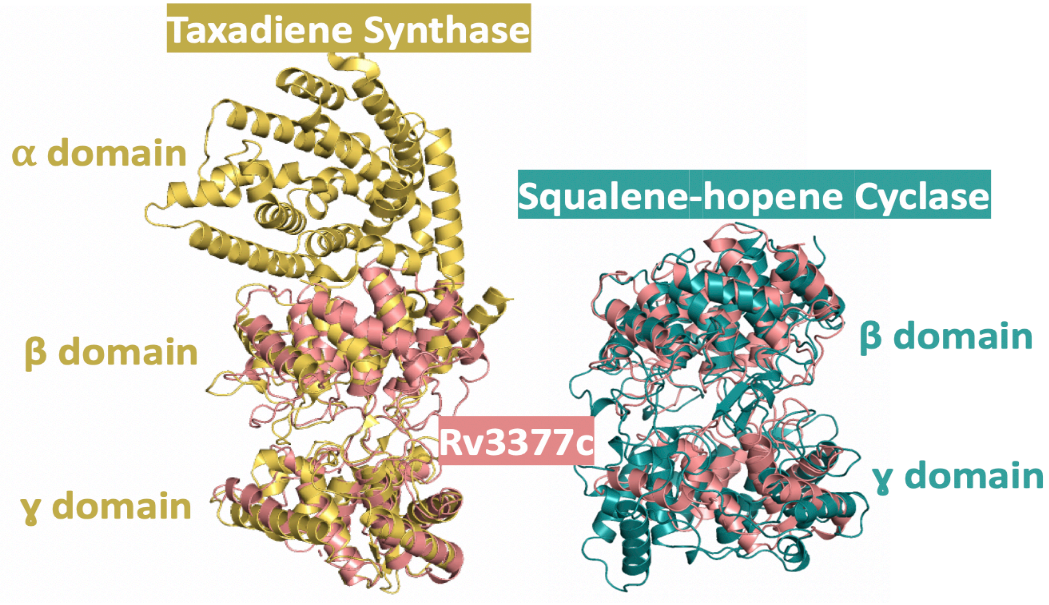
Rv3377c crystal structure

Rv3377c was expressed in E. coli and the purified protein was crystallized in 20% polyethylene glycol (PEG) 8000 and 0.1 M (cyclohexylamino) ethanesulfonic acid (CHES) pH 9.5 with 2 mM MgCl2. Despite repeated efforts over seven years and tests with mutant proteins suggested by the surface-entropy reduction server[27], crystals were obtained only once. Native data were collected to 2.75-Å resolution, but the lack of a heavy-atom derivative or sufficiently related structures precluded crystallographic phase calculations. The structure was determined using molecular replacement with Rosetta-MR. The search model was a hybrid constructed from several different distant templates (PDB 2sqc[28] and 3p5p[25]). An initial solution with one molecule per asymmetric unit was obtained and refined using data processed in space group C2. The refined model (Rfree/R = 0.2392/0.2139) contains two molecules in the asymmetric unit. Residues 1, 103-113, 139-143, and 501 were not modeled due to lack of electron density. Although 2.0 mM $MgCl_2$ was added to the mother liquor of the crystal, with the intention of determining the catalytically-important Mg^{2+} binding site, no bound Mg^{2+} ions are interpretable in the electron density map of the crystal structure. Data collection and refinement statistics are shown in SI. The Rv3377c structure can be obtained from Protein Data Bank (PDB code: 6VPT).

Rv3377c contains two helical domains which form a dumbbell shape and correspond to the β and γ domains of other class II diterpene cyclases (Figure 1.1.1a). Overlays of Rv3377c and close diterpene synthases taxadiene25 and squalene-hopene cyclase28 are shown in Figure 1.1.1b. The β and γ domains are conserved among different diterpene synthases as shown. Domain 1, or the β domain, forms an α_6–α_6 barrel of two concentric rings of parallel α chains and consists of α1 and α11-α20. The N-terminal and C-terminal helices are in close proximity to each other. Domain 2, or the γ domain, contains an α-α barrel consisting of helices α2-α10. The inner α helices of both domains are positioned so that the amino ends point toward an interdomain region, which consists of long loops from both domains and encloses a cavity of about 738.4 Å3, measured by CAVER analyst with 1.4 Å radius probe[29]. An extended loop between the last two helices extends away from the β domain, packs against the interdomain connections and inserts into a cleft in the γdomain. Lys476 at the tip of the loop forms an ion pair with Glu174 in the γ domain (see SI for detail).

Like other terpene cyclases, Rv3377c contains the QW motif, which has the general consensus of QX2-5GXW. There are two QW motifs in Rv3377c and both are composed of QxxDGSWG. One QW motif is in the β domain and the other is in the γ domain. Both motifs are on the outer surface of the protein, connecting two α helices, and the side chains of the Gln and Trp stack, as seen in other terpene cyclases.1,3 (see SI for detail)





(b)

Figure 1.1.1 (a) Left: Global structure of Rv3377c (PDB code: 6VPT), dumbbell shaped β (top) and γ (bottom) domains of the class II diterpene cyclases. Right: Reaction mechanism of Rv3377c, which converts geranylgeranyl pyrophosphate (GGPP) to tuberculosinyl pyrophosphate (TPP). (b) Comparison of Rv3377c structure with closely related homologs. Left: Overlay of Rv3377c (salmon) with taxadiene synthase (Yellow, PDB code: 3P5P), which adopts αβγ domain assembly. Right: Overlay of Rv3377c (salmon) with squalene-hopene cyclase (Green, PDB code 2SQC).

1.1.2 Results and discussion

Electronic structure calculations

Rv3377c catalyzes the conversion of GGPP to tuberculosinyl diphosphate. A reasonable mechanism for this transformation involves protonation and cyclization to form a carbocation with the relative configuration of copalyl diphosphate (CPP) (2 in Figure 1.1.2),14,19 followed by a sequence of 1,2-hydride shift (to form 3), 1,2-methyl group migration (to form 4) and another 1,2-

hydride shift (to form 5). Deprotonation would then yield tuberculosinyl diphosphate. Dickschat and co-workers have shown through isotopic labeling studies that the relative configuration of the product is consistent with a chair/chair conformation of the precursor and initial protonation on the si face of the C=C π-bond of the terminal isoprene unit, i.e., the relative configuration of CPP.[30]

This putative mechanism was subjected to scrutiny using density functional theory (DFT) calculations.31 All DFT calculations were performed with Gaussian09.[32] All geometries were optimized using the mPW1PW91 method32,33 and the 6-31+G(d,p) basis set.[34–38] All stationary points were characterized as minima or transition state structures using frequency calculations. Intrinsic reaction coordinate (IRC) calculations were used to confirm that transition state structures were connected directly to the intermediates shown.[39,40] Relative free energies of intermediates and transition state structures along the reaction pathway are shown in Figure 1.1.2. Given the flexibility of the isoprenyl diphosphate tail in the absence of the enzyme, this group was truncated to a methyl group for DFT calculations, as shown in Figure 1.1.2.

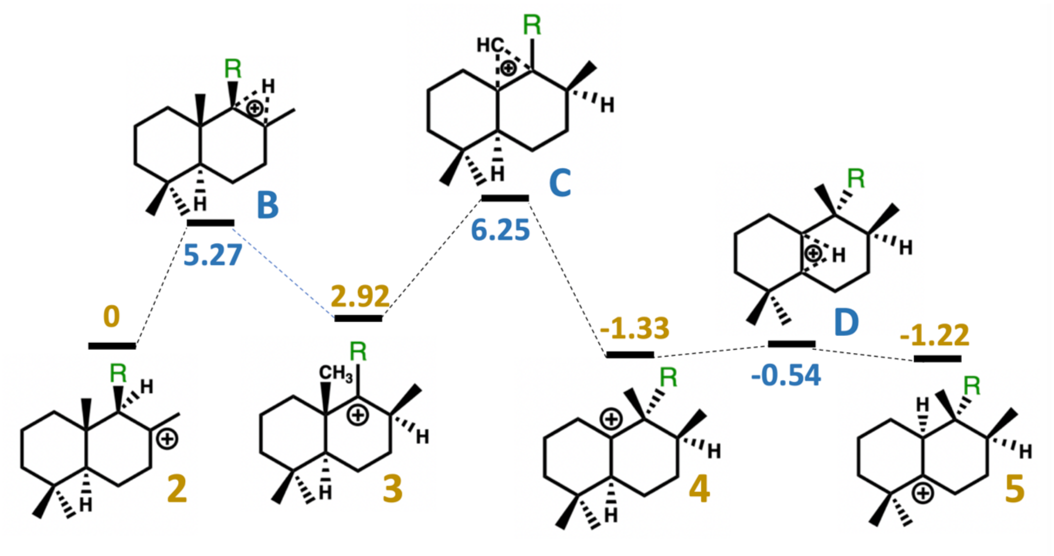


Figure 1.1.2. Relative free energies of intermediates and transition state structures in kcal/mol, calculated with mPW1PW91/6-31+G(d,p) (see SI for details).Yellow numbers are relative energies for minima, blue numbers are relative energies for transition state structures. Note that the isoprenyldiphosphate tail was truncated to a methyl group for these calculations (R = CH3 rather than R = CH2CH2C(CH3)=CHCH2OPP was used).

Docking simulations

Conformers of each intermediate structure were identified with the MMFF force field using Spartan 10.[41] For generating conformers of diphosphate-containing structures, fluorine was used in place of OPP, an approach analogous to that described previously for another terpene synthase (to avoid unnecessary sampling).[42] All conformers generated were then fully optimized using Gaussian09 at the wB97X-D/6-31+G(d,p) level of theory43. Resulting structures within 5 kcal/mol of the lowest energy conformer for each intermediate were kept and combined into conformational libraries for docking (see SI for additional information).

Docking simulations of intermediate structures into the Rv3377c protein structure were performed with the TerDockin approach[42,44,45] using the Rosetta Molecular Modeling Suite.[46,47] The crystal structure of Rv3377c was relaxed with the FastRelax48 procedure. The conformational libraries for each intermediate were then separately docked into the relaxed protein structure. Chemically meaningful constraints were applied during the docking simulations.[42,45] Specifically, since the first step of the reaction mechanism is proton donation from Asp295 to the substrate to generate 2,1 the protonated alkyl carbon was constrained to Asp295 (distance: 2.5 ± 0.5 Å; see SI for additional details). 25,000 docking runs were carried out for each intermediate, i.e., 100,000 total docking runs were performed (all input and output files are attached as SI).

Results from docking runs for each structure were combined and then filtered based on the satisfaction of the constraint, total protein energy (the lowest 25% were retained) and interface energy (only structures that were one standard deviation or lower from the mean were considered). The resulting docked structures were then aligned using TMalign49 and RMSD calculations were performed on each carbon in the skeleton between intermediates. Resulting bound structures are shown in Figure 1.1.3(a) (see SI for additional details about RMSD calculations). All intermediates are converged into a single binding mode. To confirm that docking truncated versions of carbocations does not lead to spurious results, docking of intact intermediate 2 was performed. As shown in Figure 3(b), similar binding modes are predicted (see SI for details). However, it has been observed that upon ligand binding there are often significant structural changes in the region interacting with the diphosphate with limited structural changes surrounding the core of the substrate. Therefore, as these changes in structure upon phosphate binding are likely beyond our current modeling capabilities, we emphasize here that our predictions on binding modes are relevant only to the bicyclic core of the substrate.

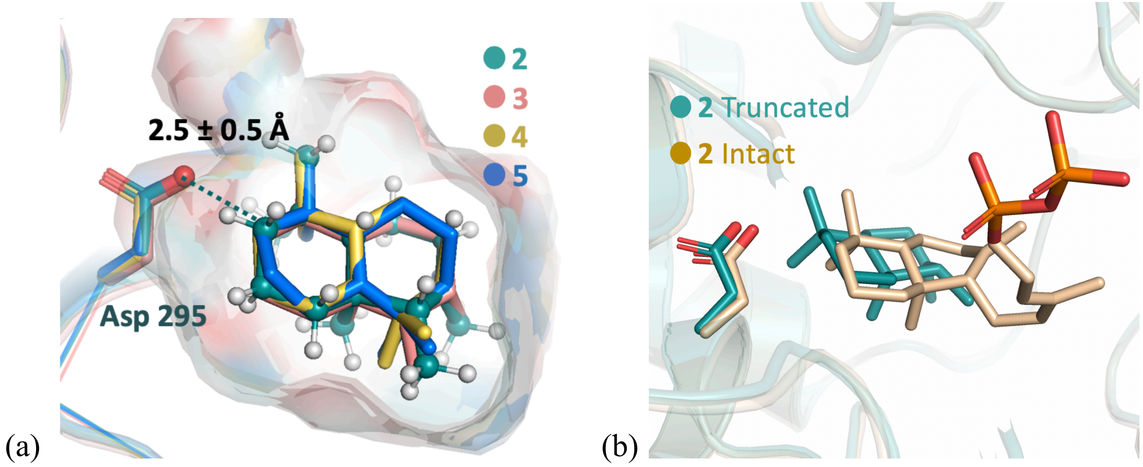


Figure 1.1.3 (a) Overlay of lowest RMSD structures for intermediates 2, 3, 4 and 5 (shown in green, salmon, yellow, blue respectively). Asp295 is highlighted in sticks. The protons of intermediate 2 are shown in white. The protonated carbon is highlighted in dark blue ball, which is constraint to the Asp295 oxygen during the docking simulation. Surface indicates the active site cavity. (b) Overlay of the lowest energy docking result of intact 2 (beige) and truncated 2 (green, same structure as the 2 in figure 1.1.3a).

The crystal structure of Rv3377c shows that the active site cavity sits at the interface of the β and γ domains, which agrees with structures of other class II diterpene synthases. The general acid, Asp295, is positioned deep in the cavity and forms hydrogen bonds to His341 and a potential water molecule. This corresponds to the general acid Asp313 and His 359 in ent-copalyl diphosphate synthase (CPS) from Streptomyces platensis, the first bacterial class II diterpene synthase with a solved crystal structure[50]. In the plant ent-copalyl diphosphate synthase from Arabidopsis thaliana, the general acid, Asp379, is hydrogen bonded with Asn425 instead.51 An overlay of these three structures is shown in Figure 1.1.4(a). The binding mode of the co-crystallized substrate analog in CPS agrees the docked intermediate 2 in Rv3377c. The hydrogen bonding orients the general acid

Asp and facilitates the initial proton transfer. Packing against each other, Trp285 and Trp380 block access to the general acid. The arrangement of the Trp285 and Trp380 after the docking suggests that they move to create space upon substrate binding, as shown in Figure 1.1.4(b). Similarly, studies on ent-CPS suggested that the loop containing Trp380 undergoes a conformational change upon substrate analogue binding[21]

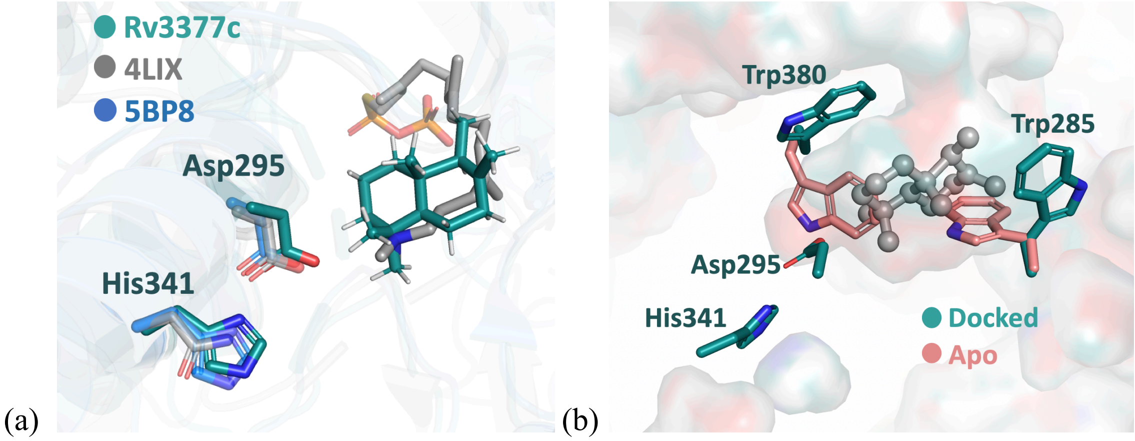


Figure 1.1.4 (a) An overlay of Rv3377c, CPS from bacteria and plant. Rv3377c structure is shown in green, bacteria CPS from Streptomyces platensis is shown in blue (PDB code: 5BP8), plant CPS from Arabidopsis thaliana (AtCPS) is shown in grey (PDB code: 4LIX). Grey stick represents the (S)-15-aza-14,15-dihydrogeranylgeranyl thiolodiphosphate that is co-crystallized with 4LIX. The highlighted residues in sticks represent key catalytic residues: Asp295 and His 341. (b) Comparison between docked Rv3377c structure (green) with apo crystal structure (salmon), with the grey ball and stick representing the predicted binding mode of intermediate 2. Based on the modeling both Trp380 and Trp285 are predicted to serve as a gate for binding.

Diterpene synthases bind a flexible substrate in an orientation relevant to product formation, trigger carbocation formation, shield reactive carbocation intermediates from premature quenching by water molecules, and allow inherent carbocation reactivity to be expressed[52–55]. Elucidating how carbocations—minima, transition state structures and species between them along rearrangement reaction coordinates—bind to terpene active sites is a frontier area of research.[42,45,56–58]AtCPS shares the same substrate, GGPP, and first carbocation intermediate, 2, with Rv3377c. Thus the sequence divergence of AtCPS and Rv3377c provides a unique opportunity to identify which residues mediate the formation of the common cyclic intermediate 2 and which distinguish the following reaction pathways.

Identical residues are clustered around the pyrophosphate binding site, the DXDTT motif, and the bottom of the active site cavity (see SI for detail). A reasonable hypothesis is that the shared residues mediate common steps in the AtCPS and Rv3377c reactions, while residue differences are responsible for the formation of the distinct products. In AtCPS, a base abstracts a proton from intermediate 2, resulting in carbon-carbon double bond formation, but in Rv3377c, hydride and methyl group migrations occur before deprotonation. Consequently, an active site base must be positioned differently.

The structure of the Rv3377c active site is compared to that of AtCPS in Figure 5. Previous studies on AtCPS have shown that a water molecule activated by ligation of a His/Asn pair could constitute the base that deprotonates the carbocation intermediate.59 In AtCPS, the double mutation H263F/N322L results in 75% of product being (−)-kolavenyl diphosphate, consistent with this pair functioning as the base. In Rv3377c, these residues are replaced by hydrophobic residues: F154 and A217, allowing rearrangement to proceed past carbocation 2. Similar behavior was observed in OsCPS4 (syn-copalyl diphosphate synthase from Oryza sativa) where mutation of H501F leads to rearrangement instead of immediate quenching of the initial carbocation intermediate[60]. Additional residue differences between AtCPS and Rv3377c (within 6 Å of the ligand) are shown in Figure 1.1.5(a). Changes to aromatic residues, including Tyr328, Phe235 and Tyr386, may lead to favorable interactions with carbocation produced by rearrangement of 2.

On the basis of our docking results, the identity of a possible residue that may deprotonate carbocation 5 was identified. The oxygen atom of Tyr462 is predicted to be ~2.7 Å away from the proton that is ultimately removed. This group may function directly as a base or may activate a water molecule for deprotonation. The pKa of an O-protonated phenol is around -6 and the pKa of a typical carbocation lacking conjugation is less than −10.[61] This indicates that proton transfer from a carbocation to the tyrosine is energetically feasible. Similarly, a serine was reported to accept proton from a carbocation intermediate in ent-kaurene synthase.[62] This Tyr is conserved in class II diterpene synthases, as shown in Figure 1.1.5(b). Previous studies suggested that the corresponding Tyr in AgAS might be the base for the deprotonation of the carbocation intermediate,[22] while the equivalent Tyr in squalene-hopene cyclase (PDB code: 2SQC) was suggested to stabilize carbocations formed early along the polycyclization reaction coordinate.28

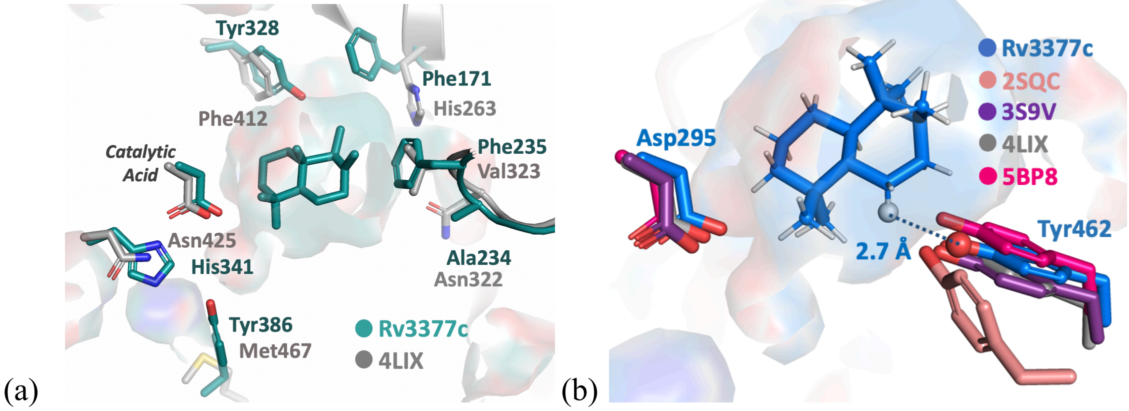


Figure 1.1.5 (a) Comparison of Rv3377c (green) and AtCPS (grey) active sites. (b) Tyr is conserved among class II diterpene synthases: blue is intermediate 5 docked into Rv3377c structure, salmon is squalene-hopene synthase (PDB code: 2SQC), purple is abietadiene synthase from Abies grandis (AgAS, PDB code: 3S9V), grey is ent-copalyl diphosphate synthase from Arabidopsis thaliana (AtCPS, PDB code: 4LIX), and pink is ent-copalyl diphosphate synthase from Streptomyces platensis (PDB code: 5BP8).

We describe the 3-dimensional crystal structure of Rv3377c, the diterpene cyclase from Mycobacterium tuberculosis. The results of docking calculations using this new structure suggest that the substrate need not move much during rearrangement (Figure 1.1.3(a)). Thus, given the energetics predicted for the enzyme-free rearrangement (Figure 1.1.2), upon carbocation formation the conversion of 2 to 5 is likely extremely rapid and does not require active manipulation by the surrounding enzyme.52 The results shown in Figure 1.1.2 also indicate that 2, 3, 4 and 5 may all be in equilibrium, with the identity of the final product of Rv3377c determined by the position of the base that terminates the cascade reaction.63 On the basis of our docking results, we propose that Tyr462 is likely this base.

1.1.3 Methods

Electronic structure calculation. The electronic structure calculations (specifically, density functional theory, DFT) were performed with Gaussian0932 with mPW1PW91/6-31+G(d,p) functional/level of theory33–39. The isoprenyl diphosphate tail of the carbocation structures were truncated in the gas phase calculation to avoid unnecessary sampling. The truncated carbocation intermediates 2-4 and transition state structures B-D were characterized in the manner describe above (see figure 1.1.2 for relative energies, figure s4 for the description of truncated structure). Intermediates 2-4 were then subjected to conformational search using Spartan 10 with MMFF force field41. All conformers were then optimized using wB97X-D/6-31+G(d,p)[43]. The conformers within 5 kcal/mol of the lowest energy conformer were kept for each intermediate, resulting in 2 conformers of 2, 2 conformers of 3, 4 conformers of 4, and 5 conformers of 5. (Relative energies of the conformers are shown in table s3. Structures are attached in SI.)

Docking simulation. The optimized conformers of intermediates 2-5 were docked into Rv3377c structure respectively with Rosetta Molecular Modeling Suite47,48 using Ref2015 scoring function68. The crystal structure of Rv3377c was relaxed with the FastRelax48 procedure prior to docking simulations. During the docking simulation, a set of constraint was applied between Asp295 and the carbocation intermediates. A proton of Asp295 was donated to the substrate in the first step of the reaction (see figure 1.1.s4), therefore, distance, angle and dihedral constraints were applied, details about the constraints were summarized in table s4. 25,000 docking simulations were performed for each intermediate. The resulting poses for each intermediate were then combined and filtered based on the following criteria: (1) Constraint satisfaction. Only poses that with 1 or lower constraint score were kept for the next step filtering. (0 constraint score means a perfect satisfaction of the constraints.) (2) Total protein score. The lowest 25% poses were kept for the next step filtering. (3) Interface energy. Poses that were one standard deviation or lower from the mean interface energy were kept. Poses that passed the filtering undergo a pair-wise RMSD calculation of carbon atoms, i.e. RMSD were calculated between each pose of intermediate 2 to 3, 2 to 4 and 2 to 5. Results of the RMSD calculation were summarized in figure 1.1.s5.

1.1.4 Supporting Information

RV3377c protein sequence

>Rv3377c Transcript

METFRTLLAKAALGNGISSTAYDTAWVAKLGQLDDELSDLALNWLCERQLPDGSWGA
EFPFCYEDRLLSTLAAMISLTSNKHRRRRAAQVEKGLLALKNLTSGAFEGPQLDIKDAT
VGFELIAPTLMAEAARLGLAICHEESILGELVGVREQKLRKLGGSKINKHITAAFSVELA
GQDGVGMLDVDNLQETNGSVKYSPSASAYFALHVKPGDKRALAYISSIIQAGDGGAPA
FYQAEIFEIVWSLWNLSRTDIDLSDPEIVRTYLPYLDHVEQHWVRGRGVGWTGNSTLED
CDTTSVAYDVLSKFGRSPDIGAVLQFEDADWFRTYFHEVGPSISTNVHVLGALKQAGYD
KCHPRVRKVLEFIRSSKEPGRFCWRDKWHRSAYYTTAHLICAASNYDDALCSDAIGWIL
NTQRPDGSWGFFDGQATAEETAYCIQALAHWQRHSGTSLSAQISRAGGWLSQHCEPPY
APLWIAKTLYCSATVVKAAILSALRLVDESNQ

Highlighted residues are not in model (residues 1, 103-114 [these are residues in an interdomain loop], 139-143, and 501)

QW motifs (Q at residue 49, 190, 414, 453)

40 ALNWLCERQLPDGSW 55 (Q and W stack like reported; in a loop region)

405 AIGWILNTQRPDGSW 420 (Q and W stack like reported; in a loop region)

Bold=classic QW motif elements

Active site (D294, D295, T296)

Active site volume

The active site volume is measured by CAVER analyst, with probe: 1.4 Å radius. The volume is 738.4 Å3. The orange sphere indicates the active site cavity.

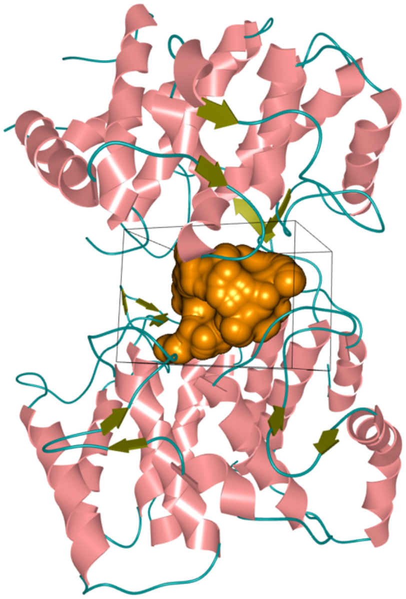
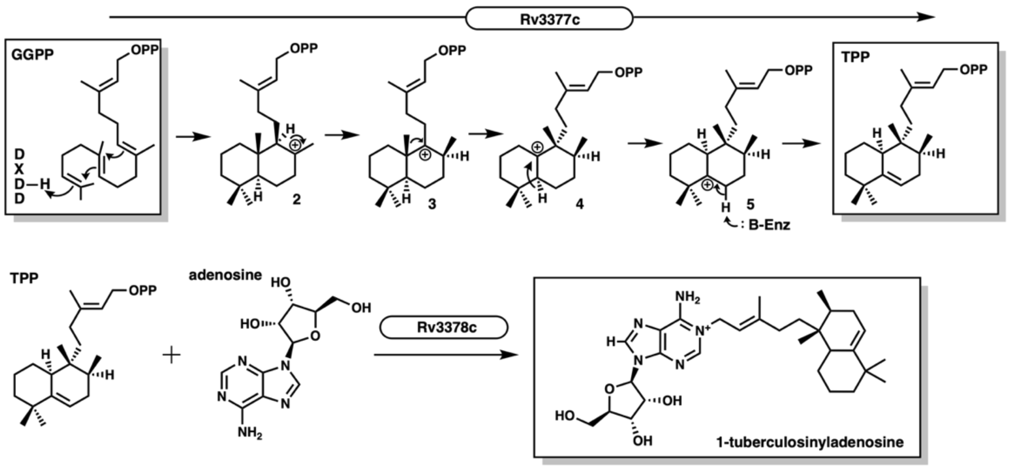


Figure 1.1.s1. Active site cavity. The active site between two domains are shown in orange. Active site volume is 738.4 Å3. Measured by CAVER analyst.

Complete pathway from GGPP to 1-tuberculosinyladenosine



Proc. Natl. Acad. Sci. U. S. A. 2014, 111 (8), 2978–2983.

Rv3377c crystal structure

1. Crystal structure determination

Table 1.1.s1 Data collection and refinement statistics. Statistics for the highest-resolution shell are shown

	Rv3377c
Resolution range(Å)	40.01 - 2.718 (2.815 - 2.718)
Space group	C 1 2 1
Unit cell(Lengths: Å, Angles: degrees)	124.900 76.428 50.123 90 91.644 90
Unique reflections	12605 (1086)
Completeness (%)	98.46 (85.43)
Wilson B-factor(Å^2)	47
Reflections used in refinement	12602 (1085)
Reflections used for R-free	611 (48)
R-work	0.2139 (0.3017)
R-free	0.2392 (0.3491)
Number of non-hydrogen atoms	3789
macromolecules	3757
solvent	32
Protein residues	483

RMS deviation, bonds (Å)	0.005
RMS deviation, angles (degrees)	0.97
Ramachandran favored (%)	97.90
Ramachandran allowed (%)	2.10
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.26
Clashscore	2.29
Average B-factor(Å^2)	46
macromolecules(Å^2)	46
solvent(Å^2)	40

2. Interdomain loop

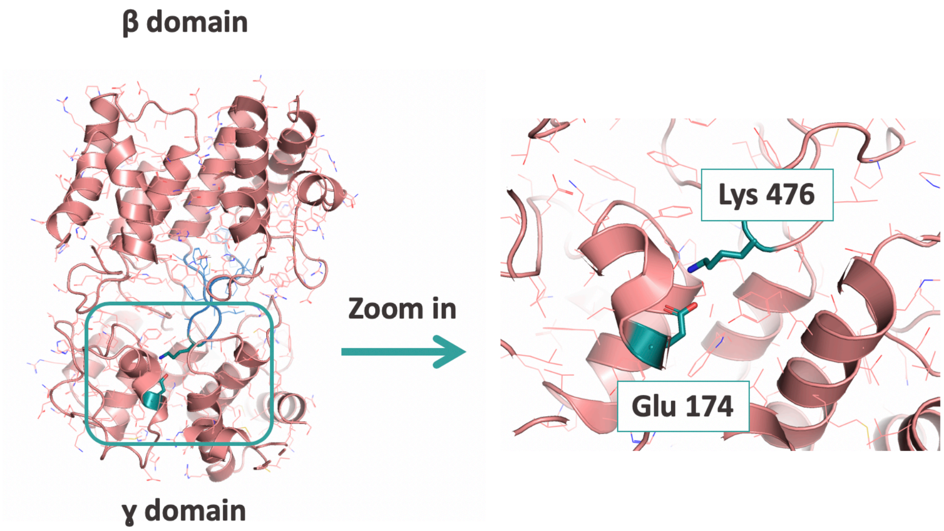


Figure 1.1.s2. Lys476 at the tip of the loop forms an ion pair with Glu174, indicated in green colored residues. The interdomain loop is indicated in blue, with the conserved ion pair at the tip.

3. Conserved QW motif

Rv3377c contains two QW motif, which has the general consensus of QX2-5GXW. Both QWmotifs has Gln and Trp stack, as shown in sticks.

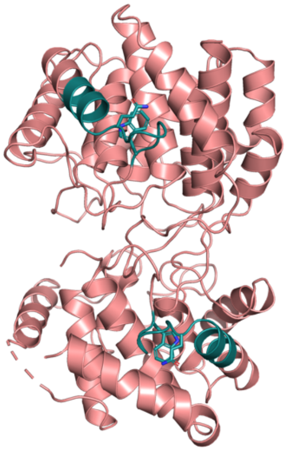
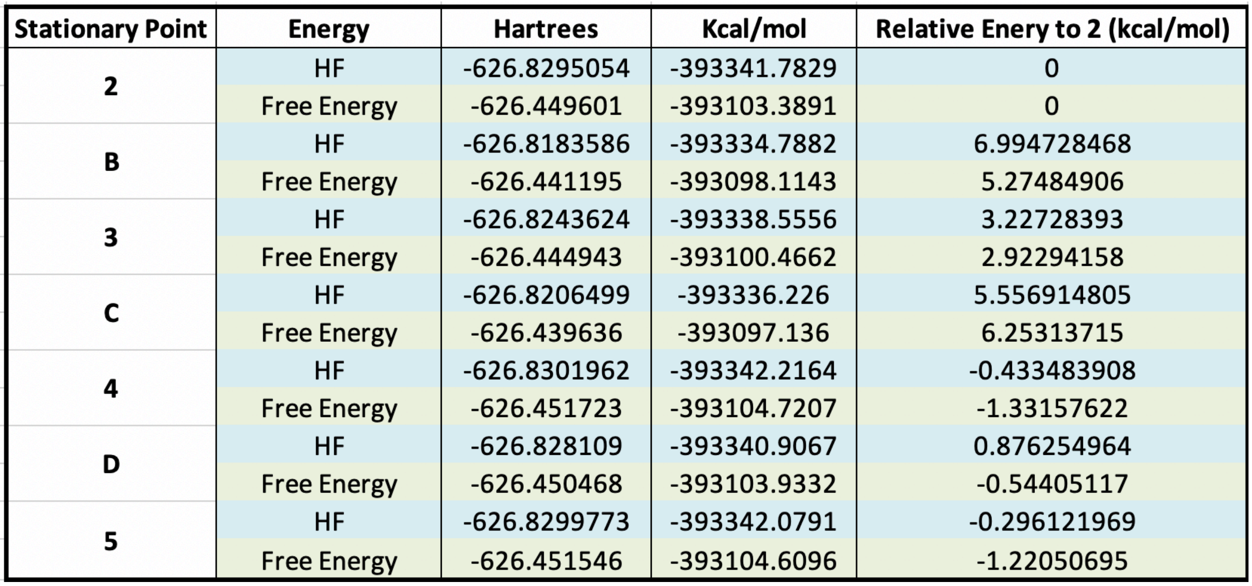


Figure 1.1.s3 QW motif of Rv3377c, highlighted in green. There are two QW motifs present in Rv3377c, with Gln and Trp stack highlighted in stick.

Electronic structure calculation (Density functional theory)

1.Energies of intermediates and transition state structures



Stationary Point	Energy	Hartrees	Kcal/mol	Relative Enery to 2 (kcal/mol)
2	HF	-626.8295054	-393341.7829	0
	Free Energy	-626.449601	-393103.3891	0
B	HF	-626.8183586	-393334.7882	6.994728468
	Free Energy	-626.441195	-393098.1143	5.27484906
3	HF	-626.8243624	-393338.5556	3.22728393
	Free Energy	-626.444943	-393100.4662	2.92294158
C	HF	-626.8206499	-393336.226	5.556914805
	Free Energy	-626.439636	-393097.136	6.25313715
4	HF	-626.8301962	-393342.2164	-0.433483908
	Free Energy	-626.451723	-393104.7207	-1.33157622
D	HF	-626.828109	-393340.9067	0.876254964
	Free Energy	-626.450468	-393103.9332	-0.54405117
5	HF	-626.8299773	-393342.0791	-0.296121969
	Free Energy	-626.451546	-393104.6096	-1.22050695

Table 1.1.s2. Relative energies of intermediate and transition state structures

2.Intermediates and transition state structures

Intermediates and transition state structures are optimized with DFT mpw1pw91/6-31+g(d,p) level of theory. Optimized structures coordination are shown below.

2 Coordination mpw1pw91/6-31+g(d,p)

	C1	-0.4259	-1.9202	-0.1761	C
2	C2	0.1160	0.5211	0.2693	C
3	C3	1.8557	-1.1155	-0.0492	C
4	C4	1.0419	-2.1549	-0.6852	C
5	H5	-0.4924	-2.1983	0.8790	H
6	H6	-1.0489	-2.6241	-0.7298	H
7	H7	1.0365	-2.0359	-1.7723	H
8	H8	1.3712	-3.1620	-0.4234	H
9	C9	-0.2616	1.9575	-0.1234	C
10	C10	-2.3774	-0.2225	-0.1754	C
11	C11	-2.6562	1.2523	-0.5201	C
12	C12	-1.7373	2.2413	0.1800	C
13	H13	0.3639	2.6783	0.4091	H
14	H14	-1.9591	3.2589	-0.1560	H
15	H15	-0.0896	2.1020	-1.1968	H
16	H16	-1.9155	2.2417	1.2601	H

17	H17	-2.5548	1.3817	-1.6058	H
18	C18	0.2388	0.4065	1.7899	C
19	H19	-0.6149	0.8729	2.2787	H
20	H20	0.2940	-0.6213	2.1543	H
21	H21	1.1228	0.9392	2.1482	H
22	C22	-2.8661	-0.5558	1.2406	C
23	H23	-2.6299	-1.5845	1.5266	H
24	H24	-2.4687	0.1053	2.0105	H
25	H25	-3.9551	-0.4629	1.2735	H
26	C26	-3.1812	-1.0893	-1.1590	C
27	H27	-4.2268	-0.7698	-1.1593	H
28	H28	-2.8100	-0.9934	-2.1844	H
29	H29	-3.1754	-2.1480	-0.8866	H
30	C30	2.7447	-1.5011	1.0538	C
31	H31	3.0673	-0.6795	1.6897	H
32	H32	2.3264	-2.3237	1.6409	H
33	H33	3.6436	-1.9202	0.5708	H
34	C34	-0.8482	-0.4801	-0.4165	C
35	C35	1.6253	0.2514	-0.4468	C
36	H36	-0.7285	-0.2937	-1.4947	H
37	H37	1.3646	0.2921	-1.5088	H
38	C38	2.6917	1.2959	-0.1273	C
39	H39	2.2901	2.2775	-0.3833	H

40	H40	2.9123	1.3286	0.9434	H
41	C41	3.9783	1.0759	-0.9231	C
42	H42	4.6909	1.8769	-0.7175	H
43	H43	4.4715	0.1324	-0.6732	H
44	H44	3.7846	1.0777	-1.9993	H
45	H45	-3.7012	1.4776	-0.2824	H

3 coordination mpw1pw91/6-31+g(d,p)

1	C1	-0.4296	-1.9795	0.1802	C
2	C2	0.1390	0.4949	0.3647	C
3	C3	1.9729	-1.1254	-0.3326	C
4	C4	1.0784	-2.2687	0.1966	C
5	H5	-0.8352	-2.1328	1.1823	H
6	H6	-0.9328	-2.7039	-0.4611	H
7	H7	1.3132	-3.1515	-0.4011	H
8	H8	1.4240	-2.4952	1.2103	H
9	C9	-0.2616	1.9308	-0.0474	C
10	C10	-2.3160	-0.2782	-0.3003	C
11	C11	-2.5564	1.1862	-0.7068	C
12	C12	-1.7553	2.1831	0.1163	C
13	H13	0.3095	2.6718	0.5177	H
14	H14	-1.9699	3.2034	-0.2146	H
15	H15	-0.0103	2.0661	-1.1057	H
16	H16	-2.0412	2.1454	1.1729	H

17	H17	-2.2934	1.3120	-1.7659	H
18	C18	0.1634	0.3792	1.9436	C
19	H19	0.6064	-0.5476	2.3138	H
20	H20	0.6829	1.2248	2.3938	H
21	H21	-0.8757	0.3928	2.2699	H
22	C22	-3.0261	-0.5649	1.0330	C
23	H23	-2.8610	-1.5859	1.3841	H
24	H24	-2.7519	0.1191	1.8388	H
25	H25	-4.1042	-0.4526	0.8893	H
26	C26	-2.9722	-1.1719	-1.3663	C
27	H27	-4.0196	-0.8854	-1.4950	H
28	H28	-2.4821	-1.0668	-2.3396	H
29	H29	-2.9634	-2.2293	-1.0905	H
30	C30	3.4516	-1.4941	-0.2563	C
31	H31	3.6032	-2.4653	-0.7304	H
32	H32	4.1050	-0.7822	-0.7608	H
33	H33	3.7772	-1.5872	0.7835	H
34	C34	-0.7761	-0.5677	-0.3066	C
35	C35	1.5700	0.2087	0.1405	C
36	H36	-0.5080	-0.5139	-1.3737	H
37	H37	1.6902	-0.9991	-1.4047	H
38	C38	2.5998	1.2457	0.3530	C
39	H39	2.2286	2.0560	0.9795	H

40	H40	3.4750	0.7996	0.8324	H
41	C41	3.0447	1.8496	-1.0105	C
42	H42	2.2287	2.3836	-1.4962	H
43	H43	3.8460	2.5619	-0.8094	H
44	H44	3.4266	1.0921	-1.6952	H
45	H45	-3.6268	1.4027	-0.6313	H

4 coordination mpw1pw91/6-31+g(d,p)

1	C1	-0.3742	-1.9849	-0.3061	C
2	C2	0.0385	0.5395	0.1072	C
3	C3	1.9740	-1.0261	-0.3146	C
4	C4	1.0845	-2.1951	0.0913	C
5	H5	-1.0240	-2.5117	0.3947	H
6	H6	-0.5616	-2.4398	-1.2812	H
7	H7	1.4556	-3.1103	-0.3766	H
8	H8	1.1721	-2.3605	1.1700	H
9	C9	-0.4806	1.9212	0.2454	C
10	C10	-2.3844	-0.3268	-0.1862	C
11	C11	-2.7547	1.1369	-0.4645	C
12	C12	-1.9839	2.1068	0.4133	C
13	H13	0.0901	2.4535	1.0108	H
14	H14	-2.2361	3.1400	0.1634	H
15	H15	-0.1477	2.3987	-0.6944	H
16	H16	-2.2548	1.9750	1.4643	H

17	H17	-2.5693	1.3641	-1.5231	H
18	C18	1.4284	0.1611	1.9168	C
19	H19	0.6217	-0.4739	2.2893	H
20	H20	2.3728	-0.2922	2.2257	H
21	H21	1.3460	1.1413	2.3867	H
22	C22	-2.7796	-0.7107	1.2484	C
23	H23	-2.6586	-1.7792	1.4372	H
24	H24	-2.2027	-0.1722	2.0057	H
25	H25	-3.8352	-0.4781	1.4110	H
26	C26	-3.1433	-1.2164	-1.1777	C
27	H27	-4.2177	-1.0451	-1.0735	H
28	H28	-2.8731	-0.9887	-2.2140	H
29	H29	-2.9684	-2.2801	-1.0044	H
30	C30	3.4427	-1.3469	-0.0421	C
31	H31	3.7085	-2.2697	-0.5628	H
32	H32	4.1234	-0.5717	-0.3973	H
33	H33	3.6422	-1.5151	1.0199	H
34	C34	-0.8472	-0.5191	-0.3845	C
35	C35	1.4828	0.2922	0.3450	C
36	H36	-0.6472	-0.1906	-1.4463	H
37	H37	1.8664	-0.8848	-1.3993	H
38	C38	2.4037	1.4862	-0.0014	C
39	H39	2.0454	2.3929	0.4918	H

40	H40	3.3718	1.2806	0.4625	H
41	C41	2.5990	1.7552	-1.4911	C
42	H42	1.6548	1.9213	-2.0206	H
43	H43	3.2031	2.6544	-1.6280	H
44	H44	3.1182	0.9377	-1.9952	H
45	H45	-3.8300	1.2653	-0.3109	H

5 coordination mpw1pw91/6-31+g(d,p)

1	C1	-0.3721	-1.9204	-0.3762	C
2	C2	0.0228	0.5762	-0.1061	C
3	C3	1.9888	-1.0213	-0.3019	C
4	C4	1.0690	-2.1915	0.0308	C
5	H5	-1.0734	-2.6456	0.0473	H
6	H6	-0.4907	-2.0674	-1.4663	H
7	H7	1.4091	-3.0923	-0.4868	H
8	H8	1.1158	-2.4236	1.0992	H
9	C9	-0.5222	1.9669	0.2852	C
10	C10	-2.3698	-0.3550	-0.1933	C
11	C11	-2.7523	1.0978	-0.5119	C
12	C12	-2.0436	2.0980	0.3824	C
13	H13	-0.0751	2.2469	1.2401	H
14	H14	-2.3395	3.1139	0.1116	H
15	H15	-0.1339	2.6835	-0.4423	H
16	H16	-2.3707	1.9604	1.4168	H

17	H17	-2.5236	1.2976	-1.5664	H
18	C18	1.4786	0.1601	1.9057	C
19	H19	0.7784	-0.5997	2.2644	H
20	H20	2.4721	-0.1043	2.2728	H
21	H21	1.2131	1.1044	2.3856	H
22	C22	-2.7467	-0.7107	1.2890	C
23	H23	-2.6551	-1.7811	1.4769	H
24	H24	-2.1581	-0.1728	2.0340	H
25	H25	-3.7940	-0.4251	1.4187	H
26	C26	-3.1236	-1.3050	-1.1363	C
27	H27	-4.1933	-1.1057	-1.0439	H
28	H28	-2.8447	-1.1291	-2.1789	H
29	H29	-2.9717	-2.3603	-0.9076	H
30	C30	3.4388	-1.3881	0.0125	C
31	H31	3.7156	-2.2930	-0.5344	H
32	H32	4.1406	-0.6072	-0.2843	H
33	H33	3.5902	-1.5949	1.0749	H
34	C34	-0.8972	-0.5535	-0.1755	C
35	C35	1.4885	0.2852	0.3734	C
36	H36	0.0888	0.5800	-1.2357	H
37	H37	1.9265	-0.8603	-1.3885	H
38	C38	2.3940	1.4777	-0.0080	C
39	H39	2.0219	2.3786	0.4884	H

40	H40	3.3747	1.2982	0.4420	H
41	C41	2.5706	1.7597	-1.4964	C
42	H42	1.6182	1.9395	-2.0095	H
43	H43	3.1715	2.6607	-1.6370	H
44	H44	3.0821	0.9490	-2.0207	H
45	H45	-3.8370	1.1920	-0.4104	H

TS B coordination mpw1pw91/6-31+g(d,p)

1	C1	0.4270	1.9639	0.2795	C
2	C2	-0.1262	-0.5030	0.4146	C
3	C3	-1.9785	1.1692	-0.1153	C
4	C4	-1.0008	2.3245	-0.1073	C
5	H5	0.5430	2.0021	1.3657	H
6	H6	1.0996	2.7203	-0.1283	H
7	H7	-1.0361	2.7925	-1.0968	H
8	H8	-1.4229	3.0629	0.5827	H
9	C9	0.2157	-1.8844	-0.1831	C
10	C10	2.3201	0.2762	-0.2645	C
11	C11	2.5397	-1.1532	-0.7923	C
12	C12	1.7034	-2.2036	-0.0760	C
13	H13	-0.3664	-2.6697	0.3063	H
14	H14	1.8839	-3.1880	-0.5202	H
15	H15	-0.0637	-1.8874	-1.2435	H
16	H16	2.0041	-2.2835	0.9745	H

17	H17	2.2864	-1.1739	-1.8613	H
18	C18	-0.0455	-0.5621	1.9635	C
19	H19	-0.3069	0.3874	2.4344	H
20	H20	-0.7164	-1.3291	2.3583	H
21	H21	0.9635	-0.8263	2.2738	H
22	C22	3.0260	0.4554	1.0877	C
23	H23	2.7794	1.4134	1.5543	H
24	H24	2.7968	-0.3367	1.8029	H
25	H25	4.1098	0.4395	0.9320	H
26	C26	2.9867	1.2404	-1.2601	C
27	H27	4.0220	0.9310	-1.4371	H
28	H28	2.4693	1.2412	-2.2258	H
29	H29	3.0173	2.2691	-0.8903	H
30	C30	-3.4253	1.5893	-0.1753	C
31	H31	-3.5284	2.4617	-0.8231	H
32	H32	-4.1023	0.8134	-0.5261	H
33	H33	-3.7315	1.8851	0.8331	H
34	C34	0.7889	0.5758	-0.2432	C
35	C35	-1.5908	-0.1537	0.1969	C
36	H36	0.5084	0.5516	-1.3098	H
37	H37	-1.7267	0.4642	-1.0956	H
38	C38	-2.6468	-1.2201	0.3432	C
39	H39	-2.2393	-2.0152	0.9686	H

40	H40	-3.5000	-0.7986	0.8796	H
41	C41	-3.1244	-1.8361	-0.9798	C
42	H42	-2.3138	-2.3470	-1.5023	H
43	H43	-3.9045	-2.5699	-0.7663	H
44	H44	-3.5457	-1.0836	-1.6514	H
45	H45	3.6060	-1.3977	-0.7205	H

TS C coordination mpw1pw91/6-31+g(d,p)

1	C1	-0.3801	-1.9267	-0.6108	C
2	C2	0.1238	0.5556	0.0135	C
3	C3	2.0094	-1.1326	-0.2387	C
4	C4	0.9608	-2.2001	0.0587	C
5	H5	-1.1326	-2.5720	-0.1542	H
6	H6	-0.3198	-2.2449	-1.6550	H
7	H7	1.3412	-3.1662	-0.2822	H
8	H8	0.8480	-2.3012	1.1444	H
9	C9	-0.3783	1.9862	-0.0867	C
10	C10	-2.3504	-0.2831	-0.1487	C
11	C11	-2.7528	1.1954	-0.2929	C
12	C12	-1.8089	2.1575	0.4132	C
13	H13	0.2902	2.6885	0.4104	H
14	H14	-2.1197	3.1899	0.2328	H
15	H15	-0.3417	2.2327	-1.1561	H
16	H16	-1.8477	2.0192	1.4997	H

17	H17	-2.7861	1.4512	-1.3601	H
18	C18	0.6246	0.2656	1.6988	C
19	H19	-0.2210	-0.3660	1.9566	H
20	H20	1.5098	-0.2710	2.0554	H
21	H21	0.5936	1.2466	2.1614	H
22	C22	-2.6332	-0.7639	1.2830	C
23	H23	-2.2312	-1.7627	1.4785	H
24	H24	-2.2543	-0.0839	2.0507	H
25	H25	-3.7134	-0.8271	1.4380	H
26	C26	-3.2218	-1.1006	-1.1143	C
27	H27	-4.2781	-0.8940	-0.9229	H
28	H28	-3.0184	-0.8387	-2.1576	H
29	H29	-3.0798	-2.1779	-1.0019	H
30	C30	3.3590	-1.4946	0.3820	C
31	H31	3.6195	-2.5170	0.1016	H
32	H32	4.1735	-0.8516	0.0454	H
33	H33	3.3285	-1.4636	1.4768	H
34	C34	-0.8568	-0.4560	-0.5802	C
35	C35	1.5094	0.2639	0.1030	C
36	H36	-0.8473	-0.1041	-1.6247	H
37	H37	2.1391	-1.0866	-1.3316	H
38	C38	2.5506	1.3591	0.1673	C
39	H39	2.1477	2.2631	0.6235	H

40	H40	3.3786	1.0339	0.8001	H
41	C41	3.0710	1.6926	-1.2372	C
42	H42	2.2682	2.0244	-1.8997	H
43	H43	3.7989	2.5026	-1.1672	H
44	H44	3.5699	0.8413	-1.7045	H
45	H45	-3.7729	1.3245	0.0827	H

TS D coordination mpw1pw91/6-31+g(d,p)

1	C1	0.3141	-1.9534	0.2670	C
2	C2	-0.0214	0.5329	-0.1827	C
3	C3	-2.0078	-1.0080	0.2443	C
4	C4	-1.1111	-2.1515	-0.2172	C
5	H5	1.0063	-2.6192	-0.2548	H
6	H6	0.4054	-2.2256	1.3250	H
7	H7	-1.4885	-3.1007	0.1719	H
8	H8	-1.1341	-2.2400	-1.3080	H
9	C9	0.5319	1.9251	-0.4240	C
10	C10	2.3662	-0.3755	0.1984	C
11	C11	2.7490	1.1037	0.3548	C
12	C12	2.0359	1.9953	-0.6452	C
13	H13	-0.0053	2.3359	-1.2811	H
14	H14	2.3643	3.0317	-0.5397	H
15	H15	0.2436	2.5569	0.4235	H
16	H16	2.2817	1.6983	-1.6689	H

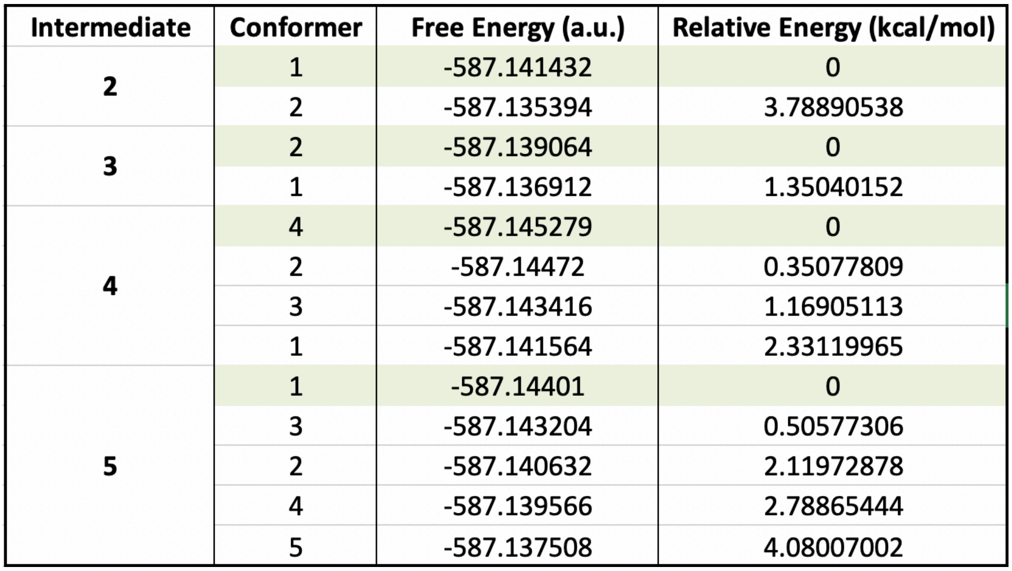
17	H17	2.5119	1.4335	1.3757	H
18	C18	-1.7941	0.3968	-1.8828	C
19	H19	-1.2576	-0.3810	-2.4301	H
20	H20	-2.8610	0.2660	-2.0696	H
21	H21	-1.5159	1.3616	-2.3107	H
22	C22	2.9590	-0.9444	-1.1120	C
23	H23	2.8081	-2.0218	-1.1997	H
24	H24	2.5463	-0.4708	-2.0050	H
25	H25	4.0372	-0.7676	-1.1058	H
26	C26	2.9339	-1.1622	1.3901	C
27	H27	4.0100	-0.9833	1.4493	H
28	H28	2.4953	-0.8348	2.3379	H
29	H29	2.7901	-2.2397	1.2945	H
30	C30	-3.4777	-1.3352	-0.0158	C
31	H31	-3.7545	-2.2297	0.5473	H
32	H32	-4.1527	-0.5374	0.2992	H
33	H33	-3.6691	-1.5475	-1.0707	H
34	C34	0.8496	-0.5522	0.0997	C
35	C35	-1.5390	0.3447	-0.3597	C
36	H36	0.2762	0.2080	1.0357	H
37	H37	-1.8843	-0.9260	1.3342	H
38	C38	-2.2959	1.5287	0.2943	C
39	H39	-1.9795	2.4614	-0.1810	H

40 H40	-3.3500	1.4244	0.0249 H
41 C41	-2.1767	1.6600	1.8097 C
42 H42	-1.1356	1.7044	2.1535 H
43 H43	-2.6523	2.5847	2.1432 H
44 H44	-2.6654	0.8405	2.3409 H
45 H45	3.8342	1.1897	0.2536 H

Docking simulations

1.Conformational search and optimization

All intermediates are subjected to conformational search using Spartan 10 with the MMFF force field. All conformers generated were then fully optimized using Gaussian09 at the wB97XD/6-31+G(d,p) level of theory. Resulting structures within 5 kcal/mol of the lowest energy conformer for each intermediate were kept.



Intermediate	Conformer	Free Energy (a.u.)	Relative Energy (kcal/mol)
2	1	-587.141432	0
	2	-587.135394	3.78890538
3	2	-587.139064	0
	1	-587.136912	1.35040152
4	4	-587.145279	0
	2	-587.14472	0.35077809
	3	-587.143416	1.16905113
	1	-587.141564	2.33119965
5	1	-587.14401	0
	3	-587.143204	0.50577306
	2	-587.140632	2.11972878
	4	-587.139566	2.78865444
	5	-587.137508	4.08007002

Table 1.1.s3. Energies of conformers for each intermediate. Lowest energies of each intermediate are highlighted as green.

2.Docking simulations

Truncated structure and Constraint setup

The first step of the reaction mechanism is proton donation from Asp295 to the substrate to generate 2. In the docking simulation, carbon 7 was constrained to Asp295. The constraint values were summarized in table s4. Angles and dihedrals were also constrained between Asp295 and intermediates to reflect the protonation step.

Constrained atoms	Type	Constraint value
AspO-C7	Distance	2.5 ± 0.5 Å
C6-C7-AspO	Angle	90 ± 10 degrees
C7-AspO-AspCG	Angle	109 ± 10 degrees
C7-AspO-AspCG-AspCB	Torsion	0 ± 20 degrees

Table 1.1.s4. Docking constraints setup

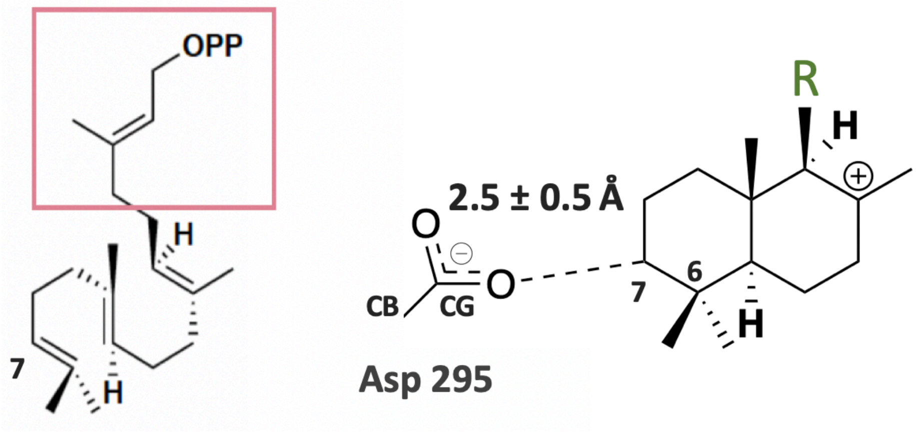


Figure 1.1.s4. Left: truncated structure. Structure in red is substituted by CH3. Right: Constraint setup. The Asp295 is constraint to carbon 7 (protonated carbon).

c.RMSD of docked structures

The RMSD calculation are performed on each carbon in the carbocation skeleton between intermediates 2, 3, 4 and 5. The transition from 2-4 and 2-5 are higher possibly due to the multiple methyl group shifts. Lines and numbers shown on the plot represents the RMSD of the selected model shown in main text figure 3. The RMSD of intermediate 2-3 is 0.377, RMSD of intermediate 2-4 is 1.82, RMSD of intermediate 2-5 is 1.65(see figure 3 for overlap of the intermediates).

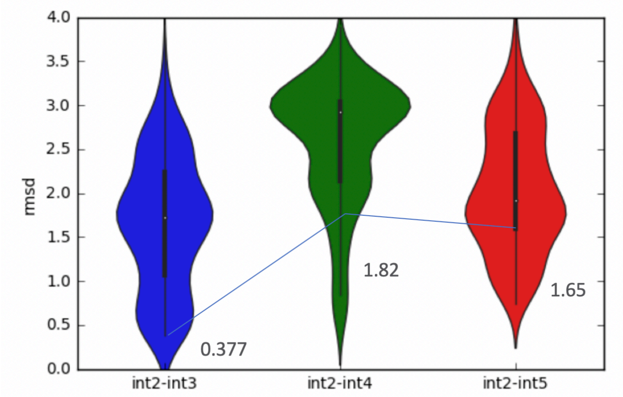


Figure 1.1.s5. Violin plots of the RMSD calculations between each intermediate to 2. Dot represents mean value. Thick black line represents first quartile, thick line represents standard deviation. Lines and numbers shown on the plot represents the RMSD of the selected model shown figure 3.

3.Comparison with AtCPS

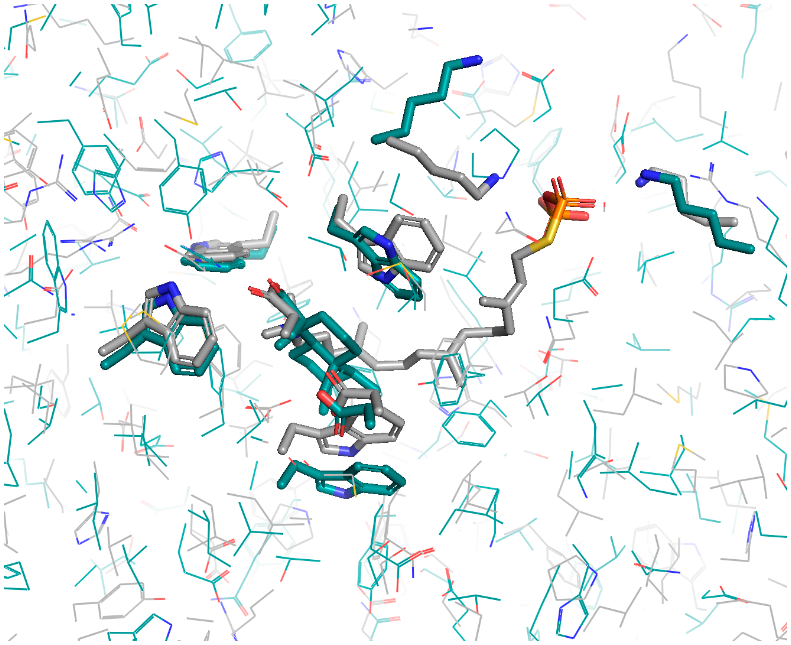


Figure 1.1.s6. Comparison between Rv3377c and AtCPS. Common residues are shown in sticks. Green: Rv3377c. Grey: AtCPS. Docked intermediate 2 shown in green sticks in the middle. Co-crystalized (S)-15-aza-14,15-dihydrogeranylgeranyl thiolodiphosphate in AtCPS shown in grey sticks.

1.1.5 References

1. Christianson, D. W. (2017) Structural and Chemical Biology of Terpenoid Cyclases. Chem. Rev. 117, 11570–11648.

2. Poulter, C. D.; Argyle, J.C.; Mash, E.A. (1978) Farnesyl pyrophosphate synthetase. Mechanistic studies of the 1'-4 coupling reaction with 2-fluorogeranyl pyrophosphate. J. Biol. Chem. 253, 7227-7233.

3. Christianson, D. W. (2006) Structural biology and chemistry of the terpenoid cyclases. Chem. Rev. 106, 3412-3442.

4. Sacchettini, J. C., and Poulter, C. D. (1997) Creating Isoprenoid Diversity. Science (80-.). 277, 1788 LP – 1789.

5. Gerber, N. N. (1969) A Volatile Metabolite of Actinomycetes, 2-Methylisoborneol. J. Antibiot. (Tokyo). 22, 508–509.

6. Dickschat, J. S., Wenzel, S. C., Bode, H. B., Müller, R., and Schulz, S. (2004) Biosynthesis of volatiles by the myxobacterium Myxococcus xanthus. ChemBioChem 5, 778–787.

7. Dickschat, J. S., Martens, T., Brinkhoff, T., Simon, M., and Schulz, S. (2005) Volatiles released by a Streptomyces species isolated from the North Sea. Chem. Biodivers. 2, 837–865.

8. Schulz, S., and Dickschat, J. S. (2007) Bacterial volatiles: The smell of small organisms. Nat. Prod. Rep. 24, 814–842.

9. Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-ya, K., Omura, S., Cane, D. E., and Ikeda, H. (2015) Terpene synthases are widely distributed in bacteria. Proc. Natl. Acad. Sci. U. S. A. 112, 857–862.

10. Izaguirre, G., Hwang, C. J., Krasner, S. W., and McGuire, M. J. (1982) Geosmin and 2-methylisoborneol from cyanobacteria in three water supply systems. Appl. Environ. Microbiol. 43, 708–714.

11. Dickschat, J. S., Nawrath, T., Thiel, V., Kunze, B., Müller, R., and Schulz, S. (2007) Biosynthesis of the off-flavor 2-methylisoborneol by the myxobacterium Nannocystis exedens. Angew. Chemie - Int. Ed. 46, 8287–8290.

12. Mann, F. M., Xu, M., Chen, X., Fulton, D. B., Russell, D. G., and Peters, R. J. (2009) Edaxadiene: A new bioactive diterpene from Mycobacterium tuberculosis. J. Am. Chem. Soc. 131, 17526–17527.

13. Mann, F. M., Prisic, S., Hu, H., Xu, M., Coates, R. M., and Peters, R. J. (2009) Characterization and inhibition of a class II diterpene cyclase from Mycobacterium tuberculosis. Implications for tuberculosis. J. Biol. Chem. 284, 23574–23579.

14. Nakano, C., and Hoshino, T. (2009) Characterization of the Rv3377c gene product, a type-B diterpene cyclase, from the Mycobacterium tuberculosis H37 genome. ChemBioChem 10, 2060–2071.

15. Layre, E., Lee, H. J., Young, D. C., Martinot, A. J., Buter, J., Minnaard, A. J., Annand, J. W., Fortune, S. M., Snider, B. B., Matsunaga, I., Rubin, E. J., Alber, T., and Moody, D. B. (2014) Molecular profiling of Mycobacterium tuberculosis identifies tuberculosinyl nucleoside products of the virulence-associated enzyme Rv3378c. Proc. Natl. Acad. Sci. U. S. A. 111, 2978–2983.

16. Prach, L., Kirby, J., Keasling, J. D., and Alber, T. (2010) Diterpene production in Mycobacterium tuberculosis. FEBS J. 277, 3588–3595.

17. Pethe, K., Swenson, D. L., Alonso, S., Anderson, J., Wang, C., and Russell, D. G. (2004) Isolation of Mycobacterium tuberculosis mutants defective in the arrest of phagosome maturation. Proc. Natl. Acad. Sci. U. S. A. 101, 13642–13647.

18. Nakano, C., Ootsuka, T., Takayama, K., Mitsui, T., Sato, T., and Hoshino, T. (2011) Characterization of the Rv3378c gene product, a new diterpene synthasefor producing tuberculosinol and (13R, S)-isotuberculosinol (nosyberkol), from the mycobacterium tuberculosis H37Rv genome. Biosci. Biotechnol. Biochem. 75, 75–81.

19. Nakano, C., Okamura, T., Sato, T., Dairi, T., and Hoshino, T. (2005) Mycobacterium tuberculosis H37Rv3377c encodes the diterpene cyclase for producing the halimane skeleton. Chem. Commun. 1016–1018.

20. Ashby, M. N., and Edwards, P. A. (1990) Elucidation of the deficiency in two yeast coenzyme Q mutants. Characterization of the structural gene encoding hexaprenyl pyrophosphate synthetase. J. Biol. Chem. 265, 13157–13164.

21. KÖksal, M., Hu, H., Coates, R. M., Peters, R. J., and Christianson, D. W. (2011) Structure and mechanism of the diterpene cyclase ent-copalyl diphosphate synthase. Nat. Chem. Biol. 7, 431–433.

22. Zhou, K., Gao, Y., Hoy, J. A., Mann, F. M., Honzatko, R. B., and Peters, R. J. (2012) Insights into diterpene cyclization from structure of bifunctional abietadiene synthase from Abies grandis. J. Biol. Chem. 287, 6840–6850.

23. Janke, R., Görner, C., Hirte, M., Brück, T., and Loll, B. (2014) The first structure of a bacterial diterpene cyclase: CotB2. Acta Crystallogr. Sect. D Biol. Crystallogr. 70, 1528–1537.

24. Cao, R., Zhang, Y., Mann, F. M., Huang, C., Mukkamala, D., Hudock, M. P., Mead, M. E., Prisic, S., Wang, K., Lin, F. Y., Chang, T. K., Peters, R. J., and Oldfield, E. (2010) Diterpene cyclases and the nature of the isoprene fold. Proteins Struct. Funct. Bioinforma. 78, 2417–2432.

25. Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011) Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. Nature 469, 116–122.

26. Gao, Y., Honzatko, R. B., and Peters, R. J. (2012) Terpenoid synthase structures: A so far incomplete view of complex catalysis. Nat. Prod. Rep. 29, 1153–1175.

27. Goldschmidt, L., Cooper, D. R., Derewenda, Z. S., and Eisenberg, D. (2007) Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. Protein Sci. 16, 1569–1576.

28. Wendt, K. U., Lenhart, A., and Schulz, G. E. (1999) The structure of the membrane protein squalene-hopene cyclase at 2.0 Å resolution. J. Mol. Biol. 286, 175–187.

29. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J., and Damborsky, J. (2012) CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. PLoS Comput. Biol. 8, 23–30.

30. Citron, C. A., Rabe, P., Barra, L., Nakano, C., Hoshino, T., and Dickschat, J. S. (2014) Synthesis of isotopically labelled oligoprenyl diphosphates and their application in mechanistic investigations of terpene cyclases. European J. Org. Chem. 2014, 7684–7691.

31. Tantillo, D. J. (2011) Biosynthesis via carbocations: Theoretical studies on terpene formation. Nat. Prod. Rep. 28, 1035–1053.

32. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Keith, T., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, O., Foresman, J. B., Ortiz, J. V., Cioslowski, J., Fox, D. J., Montgomery Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N.,

Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J., and Fox, D. J. (2013) Gaussian 09, Revision D.01. Gaussian Inc. 1–20.

33. Matsuda, S. P. T., Wilson, W. K., and Xiong, Q. (2006) Mechanistic insights into triterpene synthesis from quantum mechanical calculations. Detection of systematic errors in B3LYP cyclization energies. Org. Biomol. Chem. 4, 530–543.

34. Becke, A. D. (1993) A new mixing of Hartree-Fock and local density-functional theories. J. Chem. Phys. 98, 1372–1377.

35. Becke, A. D. (1993) Density-functional thermochemistry. III. The role of exact exchange. J. Chem. Phys. 98, 5648–5652.

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

37. Stephens, P. J., Devlin, F. J., Chabalowski, C. F., and Frisch, M. J. (1994) Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. J. Phys. Chem. 98, 11623–11627.

38. Tirado-Rives, J., and Jorgensen, W. L. (2008) Performance of B3LYP density functional methods for a large set of organic molecules. J. Chem. Theory Comput. 4, 297–306.

39. Gonzalez, C., and Schlegel, H. B. (1990) Reaction path following in mass-weighted internal coordinates. J. Phys. Chem. 94, 5523–5527.

40. Fukui, K. (1981) The Path of Chemical Reactions - The IRC Approach. Acc. Chem. Res. 14, 363–368.

41. Shao, Y., Molnar, L. F., Jung, Y., Kussmann, J., Ochsenfeld, C., Brown, S. T., Gilbert, A. T. B., Slipchenko, L. V., Levchenko, S. V., O'Neill, D. P., DiStasio, R. A., Lochan, R. C., Wang, T., Beran, G. J. O., Besley, N. A., Herbert, J. M., Yeh Lin, C., Van Voorhis, T., Hung Chien, S., Sodt, A., Steele, R. P., Rassolov, V. A., Maslen, P. E., Korambath, P. P., Adamson, R. D., Austin, B., Baker, J., Byrd, E. F. C., Dachsel, H., Doerksen, R. J., Dreuw, A., Dunietz, B. D., Dutoi, A. D., Furlani, T. R., Gwaltney, S. R., Heyden, A., Hirata, S., Hsu, C. P., Kedziora, G., Khalliulin, R. Z., Klunzinger, P., Lee, A. M., Lee, M. S., Liang, W., Lotan, I., Nair, N., Peters, B., Proynov, E. I., Pieniazek, P. A., Min Rhee, Y., Ritchie, J., Rosta, E., David Sherrill, C., Simmonett, A. C., Subotnik, J. E., Lee Woodcock, H., Zhang, W., Bell, A. T., Chakraborty, A. K., Chipman, D. M., Keil, F. J., Warshel, A., Hehre, W. J., Schaefer, H. F., Kong, J., Krylov, A. I., Gill, P. M. W., and Head-Gordon, M. (2006) Advances in methods and algorithms in a modern quantum chemistry program package. Phys. Chem. Chem. Phys. 8, 3172–3191.

42. O'Brien, T. E., Bertolani, S. J., Zhang, Y., Siegel, J. B., and Tantillo, D. J. (2018) Predicting Productive Binding Modes for Substrates and Carbocation Intermediates in Terpene Synthases - Bornyl Diphosphate Synthase As a Representative Case. ACS Catal. 8, 3322–3330.

43. Chai, J. Da, and Head-Gordon, M. (2008) Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. Phys. Chem. Chem. Phys. 10, 6615–6620.

44. Das, S., Shimshi, M., Raz, K., Nitoker Eliaz, N., Mhashal, A. R., Ansbacher, T., and Major, D. T. (2019) EnzyDock: Protein–Ligand Docking of Multiple Reactive States along a Reaction Coordinate in Enzymes. J. Chem. Theory Comput. 15, 5116–5134.

45. O’Brien, T. E., Bertolani, S. J., Tantillo, D. J., and Siegel, J. B. (2016) Mechanistically informed predictions of binding modes for carbocation intermediates of a sesquiterpene synthase reaction. Chem. Sci. 7, 4009–4015.

46. Lange, O. F., and Baker, D. (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins Struct. Funct. Bioinforma. 80, 884–895.

47. Alford, R. F., Leaver-fay, A., Jeliazkov, J. R., Meara, M. J. O., Dimaio, F. P., Park, H., Shapovalov, M. V, Renfrew, P. D., Mulligan, K., Kappel, K., Labonte, J. W., Pacella, M. S., and Bonneau, R. (2018) HHS Public Access 13, 3031–3048.

48. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., and Baker, D. (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. 23, 47–55.

49. Zhang, Y., and Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309.

50. Rudolf, J. D., Dong, L. Bin, Cao, H., Hatzos-Skintges, C., Osipiuk, J., Endres, M., Chang, C. Y., Ma, M., Babnigg, G., Joachimiak, A., Phillips, G. N., and Shen, B. (2016) Structure of the ent-Copalyl Diphosphate Synthase PtmT2 from Streptomyces platensis CB00739, a Bacterial Type II Diterpene Synthase. J. Am. Chem. Soc. 138, 10905–10915.

51. Köksal, M., Potter, K., Peters, R. J., and Christianson, D. W. (2014) 1.55 Å-Resolution Structure of Ent-Copalyl Diphosphate Synthase and Exploration of General Acid Function By Site-Directed Mutagenesis. Biochim. Biophys. Acta - Gen. Subj. 1840, 184–190.

52. Tantillo, D. J. (2017) Importance of Inherent Substrate Reactivity in Enzyme-Promoted Carbocation Cyclization/Rearrangements. Angew. Chemie - Int. Ed. 56, 10040–10045.

53. Cane, D. E. (1990) Enzymatic Formation of Sesquiterpenes. Chem. Rev. 90, 1089–1103.

54. Pemberton, R. P., Ho, K. C., and Tantillo, D. J. (2015) Modulation of inherent dynamical tendencies of the bisabolyl cation via preorganization in epi-isozizaene synthase. Chem. Sci. 6, 2347–2353.

55. Hammer, S. C., Syrén, P. O., and Hauer, B. (2016) Substrate Pre-Folding and Water Molecule Organization Matters for Terpene Cyclase Catalyzed Conversion of Unnatural Substrates. ChemistrySelect 1, 3589–3593.

56. Freud, Y., Ansbacher, T., and Major, D. T. (2017) Catalytic Control in the Facile Proton Transfer in Taxadiene Synthase. ACS Catal. 7, 7653–7657.

57. Major, D. T., and Weitman, M. (2012) Electrostatically guided dynamics-the root of fidelity in a promiscuous terpene synthase? J. Am. Chem. Soc. 134, 19454–19462.

58. Driller, R., Janke, S., Fuchs, M., Warner, E., Mhashal, A. R., Major, D. T., Christmann, M., Brück, T., and Loll, B. (2018) Towards a comprehensive understanding of the structural dynamics of a bacterial diterpene synthase during catalysis. Nat. Commun. 9.

59. Potter, K. C., Zi, J., Hong, Y. J., Schulte, S., Malchow, B., Tantillo, D. J., and Peters, R. J. (2016) Blocking Deprotonation with Retention of Aromaticity in a Plant ent-Copalyl Diphosphate Synthase Leads to Product Rearrangement. Angew. Chemie - Int. Ed. 55, 634–638.

60. Potter, K. C., Jia, M., Hong, Y. J., Tantillo, D., and Peters, R. J. (2016) Product Rearrangement from Altering a Single Residue in the Rice syn-Copalyl Diphosphate Synthase. Org. Lett. 18, 1060–1063.

61. McCormack, A. C., McDonnell, C. M., More O'Ferrall, R. A., O'Donoghue, A. M. C., and Rao, S. N. (2002) Protonated benzofuran, anthracene, naphthalene, benzene, ethene, and ethyne: Measurements and estimates of pKa and pKR. J. Am. Chem. Soc. 124, 8575–8583.

62. Jia, M., Zhang, Y., Siegel, J. B., Tantillo, D. J., and Peters, R. J. (2019) Switching on a Nontraditional Enzymatic Base - Deprotonation by Serine in the ent-Kaurene Synthase from Bradyrhizobium japonicum. ACS Catal. 9, 8867–8871.

63. Hong, Y. J., and Tantillo, D. J. (2011) The taxadiene-forming carbocation cascade. J. Am. Chem. Soc. 133, 18249–18256.

64. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951–960.

65. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674.

66. Dimaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwaï, H., Pokkuluri, P. R., and Baker, D. (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 473, 540–543.

67. Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., Read, R. J., and Adams, P. D. (2007) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr. Sect. D Biol. Crystallogr. 64, 61–69.

68. Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., and Dimaio, F. (2016) Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J. Chem. Theory Comput. 12, 6201–6212.

1.2 Switching on a Nontraditional Enzymatic Base - Deprotonation by Serine in the *ent*-Kaurene Synthase from Bradyrhizobium japonicum

This section is a published paper. This is a collaboration project with Professor Reuben Peters. Author is listed as the co-first author for this publication. All experiments were performed by Meirong Jia. Data permission acquired.

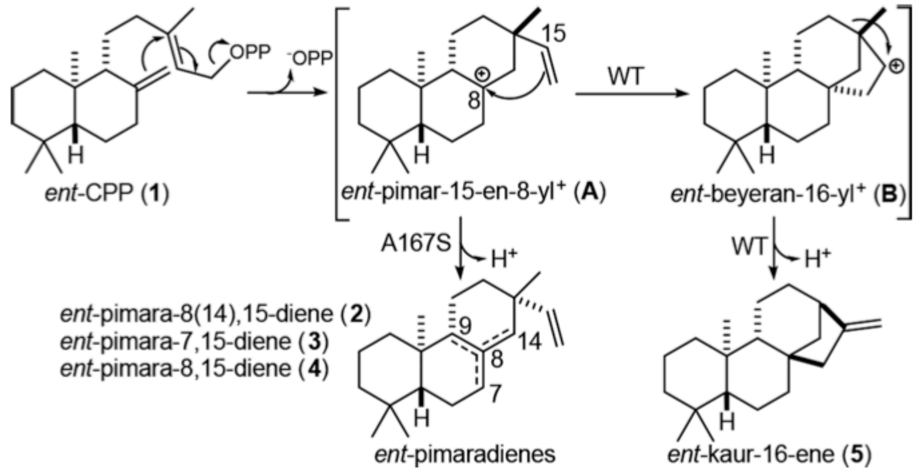
1.2.1 Abstract

Terpene synthases often catalyze complex carbocation cascade reactions. It has been previously shown that single-residue switches involving replacement of a key aliphatic residue with serine or threonine can "short-circuit" such reactions that are presumed to act indirectly via dipole stabilization of intermediate carbocations. Here a similar switch was found in the structurally characterized *ent*-kaurene synthase from Bradyrhizobium japonicum. Application of a recently developed computational approach to terpene synthases, *TerDockin*, surprisingly indicates direct action of the introduced serine hydroxyl as a catalytic base. Notably, this model suggests an alternative interpretation of previous results and potential routes toward reengineering terpene synthase activity more generally.

1.2.2 Introduction

Terpene synthases produce intricate hydrocarbon backbones that underlie the structural diversity of the extensive family of terpenoid natural products. [1] This feat is accomplished by magnesium-assisted lysis of the allylic diphosphate ester in their isoprenyl substrates, which often triggers complex carbocation cascade reactions that are eventually terminated by deprotonation (or, occasionally, carbocation trapping by a nucleophile). To accommodate such reactive intermediates, the relevant portion of terpene synthase active sites have been observed to be largely nonpolar, composed of aliphatic and aromatic residues. Indeed, the perceived lack of side chains with suitable basicity has led to the hypothesis that the pyrophosphate anion coproduct (-OPP) generally serves as the catalytic (general) base. [2]

Previous work has demonstrated that single residue changes can switch the product outcome in certain plant diterpene synthases. [3–11] Arguably the most interesting changes are those involving a key position that controls the complexity of the catalytic reaction. These enzymes are involved in labdane-related diterpenoid biosynthesis. Hence, they react with already bicyclic labdadienyl/copalyl diphosphate (CPP), carrying out initial cyclization to pimarenyl$^+$ intermediates, which can be followed by further cyclization and/or rearrangement (e.g., Scheme 1.2.1). Strikingly, the presence of an aliphatic residue, typically alanine or isoleucine, leads to more complex reactions, while serine or threonine at the relevant key position "short-circuits" the carbocation cascade, leading to production of pimaradienes. The key residue is hypothesized to be proximal to the carbocation in the pimarenyl$^+$ intermediate, which continues to react in the presence of the aliphatic residue, but undergoes deprotonation when this is serine or threonine instead. However, based in large part on the perceived difficulty for such a nonactivated hydroxyl group to act as a catalytic base, these have been suggested to act via dipole stabilization of the initially formed pimarenyl + intermediate, enabling deprotonation (presumably by reor-ientation with respect to -OPP). [12]


OPP
-OPP
15
8
WT
ent-CPP (1)
ent-pimar-15-en-8-yl+ (A)
ent-beyeran-16-yl+ (B)
A167S
H+
WT
H+
ent-pimara-8(14),15-diene (2)
ent-pimara-7,15-diene (3)
ent-pimara-8,15-diene (4)
9
14
8
7
ent-pimaradienes
ent-kaur-16-ene (5)

Scheme 1.2.1. Reactions Catalyzed by BjKS and A167S Mutant

A number of labdane-related diterpene synthases also have been identified from bacteria. Of particular interest here is the ent-kaurene synthase from Bradyrhizobium japonicum (BjKS), [13] which has been shown to be involved in production of gibberellin phytohormones by this rhizobium. [14] Notably, high-resolution crystal structures have been determined for BjKS. 15 This revealed the expected nonpolar binding pocket for the hydrocarbon portion of its substrate, ent-CPP (1). While other residues were suggested to play particularly important roles in the catalyzed reaction, here alanine-167 was noted to exhibit intriguing parallels to a previously identified single-residue switch. In particular, A167 is located at a widely conserved helix-break (G1/2), just as observed for the critical alanine in the only plant diterpene synthase in which both a product switch (alanine to serine) has been identified 4 and that has a crystal structure currently available, 16 that is, the abietadiene synthase from Abies grandis (AgAS).

To investigate the hypothesis that A167 might be important in the (bi)cyclization and rearrangement reaction catalyzed by BjKS (Scheme 1.2.1), specifically continuation beyond initial cyclization of 1 to an ent-pimara-15-en-8-yl + intermediate (A) [e.g., to form the ent-beyeranyl +

intermediate (B)], this residue was mutated to serine. The resulting BjKS:A167S mutant was observed to predominantly produce a roughly equal mixture of ent-pimara-8(14),15-diene (2) and ent-pimara-7,15-diene (3), resulting from immediate deprotonation of A (although no ent-pimara-8,15-diene, 4, which also could be formed by deprotonation of A), along with small amounts of ent-kaurene (5), rather than the exclusive production of 5 exhibited by wild-type BjKS (Figure 1.2.1).

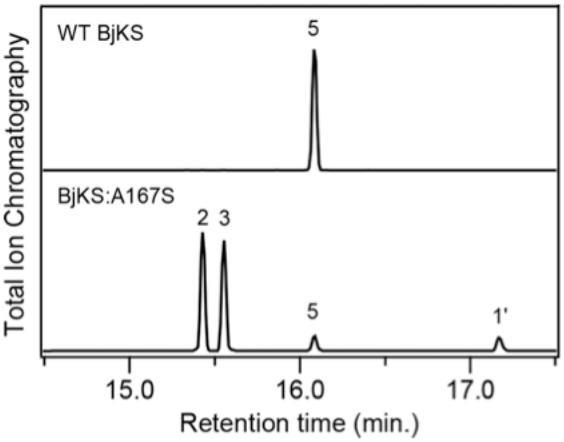


Figure 1.2.1. Chromatograms from GC-MS analysis of BjKS, either wild-type (WT) or A167S mutant, as indicated. Extracts from E. coli cultures engineered to produce 1 and coexpressing the indicated BjKS. Peaks are numbered as in the text, with 1′ indicating the dephosphorylated derivative of 1 produced by endogenous phosphatases (1′, ent-copalol; 2, ent-pimara-8(14),15-diene; 3, ent-pimara-7,15-diene; 5, ent-kaur-16-ene), as identified by comparison to authentic standards.

Intriguingly, this single residue switch in BjKS differs from that found in plant ent-kaurene synthases (KSs), where the analogous residue is an isoleucine, with threonine substitution leading to predominant production of 2 and only small amounts of 3. 3,6,8 Moreover, sequence alignment with AgAS suggests that this isoleucine does not fall into the G1/2 helix-break, but rather on the first turn of the G2 helix (i.e., four residues later). 4 This difference in location of the critical aliphatic residue between plant KSs and AgAS (which is representative of the family of diterpene synthases involved in conifer resin acid biosynthesis that are distinct from plant KSs [17]), has been attributed to their use of enantiomeric forms of CPP. [12] Regardless, it appears that A may be differentially oriented in BjKS than plant KSs, at least relative to the G1/2 helix, which is perhaps not surprising given that these share <15% sequence identity. [13]

1.2.3 Results and discussion

To gain further insight into the role of the single-residue switch in BjKS, computational modeling was undertaken. First, density functional theory (DFT) calculations (PCM(water)-ωB97XD/6-311+G(d,p)) [18] were carried out to compare the energies of the three possible deprotonation products of carbocation A. No significant difference in energy was found, however (relative energies in kcal/mol: 2, +0.54; 3, +0.73, 4, 0.00), indicating that the observed product distribution is not the result of thermodynamic equilibration, nor its manifestation in transition state structures (TSSs) for deprotonation.

To gain further insight, the recently described TerDockin approach [19,20] was employed, using the Rosetta Molecular Modeling Suite. [21,22] To perform docking, all available X-ray crystal structures of BjKS were examined. [15] The structure with PDB code 4XLX was used because it had the most complete active site density (see Supporting Figure S1 for comparison). Hydrocarbon (carbocation) structures and the diphosphate-magnesium complex were docked into BjKS simultaneously. As no available BjKS structure contains a diphosphate-magnesium complex, the diphosphate conformation was extracted from the crystal structure 3P5R, the closest homologue of BjKS with such a complex present. Some conformations of carbocation structures were previously optimized using DFT calculations by Hong and Tantillo. [23] Carbocation conformers were identified using Spartan 10 with the MMFF force field. [24] All conformers generated were then fully optimized using Gaussian09 [25] with ωB97XD/6-31+g-(d,p). TerDockin was applied to both the wild-type BjKS and to the A167S mutant; results for the latter are discussed below, while results for the former can be found in the Supporting Information.

The conformer library of A, along with the diphosphate-magnesium complex, was docked into the BjKS:A167S structure to examine the relative positions of the carbocation center and S167. The first ionization step involves bond breaking between a diphosphate oxygen and the terminal carbon of 1, leading to two possible carbocation-diphosphate ion pair orientations the terminal carbon near to one or the other oxygen since only two diphosphate oxygens protrude into the active site; these were examined separately during the docking simulation (Figure 1.2.2A; see the Supporting Information for details on the chemically meaningful constraints applied during docking).

As described above, simple alkene stability arguments do not rationalize the distribution of pimaradiene isomers observed. Moreover, the more selective production of 2 by the functionally analogous Ile → Thr mutation in the plant KSs argues against any significant effect from relative stability. Preliminary docking results suggested that S167, rather than diphosphate, may act as the base for the deprotonation step to form the pimaradienes. While an introduced histidine has been suggested to act as the catalytic base for production of cembrene A by the relevant mutant of taxadiene synthase, [26] it does not appear to have been previously suggested that a hydroxyl containing residue can act as the catalytic base in terpene synthases. Nevertheless, the pK a of a protonated alcohol is typically around −1 to −4,[27] while that of a typical carbocation lacking conjugation is less than −10,[28] suggesting that proton transfer from a carbocation to an alcohol is energetically reasonable. In addition, hydrogen atoms at C7, C9, and C14 all appeared to be reasonably close to the S167 oxygen. Consequently, we suspected that the hydroxyl group of S167 acts as a base and that certain C carbocation −H−O and H−O−C Ser angles in the deprotonation of TSS (Figure 1.2.2B) were preferred. Optimal angles for proton transfer during deproto-nation were identified with DFT calculations on a model system (Figure 1.2.2B−C): ∼120°for H−O−C Ser and ∼180°for C carbocation −H−O. Constraints favoring these angles were then applied to the docking simulation (see Supporting Information for details and previous papers on terpene docking for the philosophy underpinning this approach and potential limitations[19,20]). Docking simulations (25 000) for each of the two ion pair orientations and each of five possible deprotonation sites (C7 and C14 bear two hydrogen atoms, while C9 bears one) were then carried out (i.e., 250 000 total docking runs were performed). All docking results were combined and then filtered on the basis of satisfaction of constraints, total protein energy (the lowest 10% were kept), and interface energy (the lowest 5% were kept). Results are summarized in Figure 3 (see Supporting Information for details). In total, deprotonation at C7 (to form 2) is predicted to be the most likely, with deprotonation at C14 (to form 3) next most likely and deprotonation at C9 (to form 4) unlikely (Figure 1.2.3A). The predicted 59:36:5 ratio for 2:3:4 is consistent with the experimental observation that 2 and 3 are formed in comparable amounts, with slightly more 2 than 3, while 4 is not observed, suggesting that the ability to approach the ideal TSS geometry during deprotonation plays a major role in product selectivity. Note also that the backbone carbonyl oxygen of I166 can hydrogen bond with the S167 hydroxyl group, further increasing the basicity of the Ser side chain (Figure 1.2.3B).

In summary, we suggest that the shortening of the BjKS carbocation cascade induced by the A167S substitution is due to direct action of the introduced alcohol as a catalytic base mediating premature deprotonation. Even beyond the implications for BjKS, our results further suggest that the previously identified analogous single residue product switches in plant diterpene synthases may operate in the same fashion; that is, the introduced serine or threonine may act as a catalytic base to terminate the carbocation cascade reaction. More importantly, appreciation of this ability to directly deprotonate carbocation intermediates immediately indicates that incorporation of hydroxyl containing side chains at appropriate locations provides a means to alter product outcome in enzymatic engineering of terpene synthases more generally, which will be explored in future work.

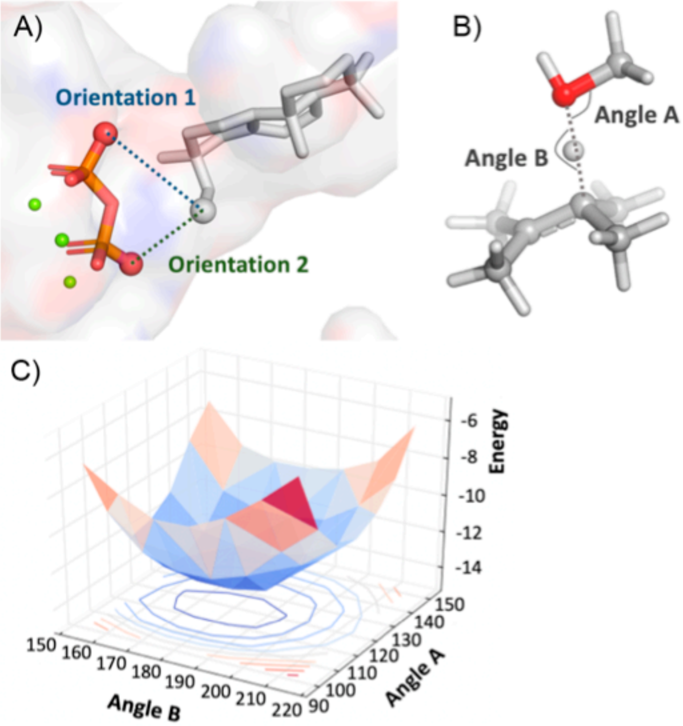


Figure 1.2.2. (A) Two diphosphate oxygen atoms to which the terminal carbon of the substrate may have been connected. (B) Model system to identify optimal angles of deprotonation by the S167 hydroxyl group: methanol and 2,3-dimethyl-2-butene. (C) 2D potential energy scan (vertical axis corresponds to relative electronic energies in kcal/ mol; other axes correspond to angles from panel (B) in degrees) showing that the optimal angles are ~120°for A and ~180°for B.

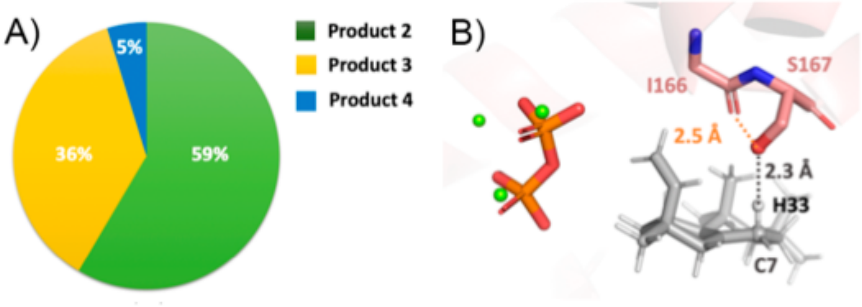


Figure 1.2.3. (A) Predicted relative amounts of pimaradiene products. All docking results are combined and filtered on the basis of satisfaction of constraints, total protein energy (the lowest 10% were kept), and interface energy (the lowest 5% were kept; see SI for additional details). (B) A representative pose predicted by docking (C7/ Orientation 2/H33). The distance between the S167 oxygen and H33 is 2.3 Å. The distance between the I166 backbone carbonyl oxygen and the S167 oxygen is 2.5 Å.

1.2.4 Supporting information

1.2.1. Computational Study

1.2.1.1 DFT calculations

1.2.1.1.1 Relative energies of three possible deprotonation product

Level of theory	Solvent	2	3	4
ωB97XD/6-311+G(d,p)	PCM/Water	0.54	0.73	0
ωB97XD/6-311+G(d,p)	PCM/Chloroform	0.32	0.9	0
mPW1PW91/6-311+G(d,p)	PCM/Chloroform	2.07	2.13	0
BB1K/6-311+G(d,p)	PCM/Chloroform	0.92	0.89	0

Table 1.2.s1 Relative energies(in kcal/mol) of three possible deprotonation product in different level of theory and solvent.

1.2.1.1.2 DFT calculation of optimal angles for proton transfer of model system

Angle B	155	155	155	155	155	155	155
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220766.33	-220769.58	-220771.73	-220771.48	-220770.62	-220768.49	-220766.98
Angle B	165	165	165	165	165	165	165
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220767.54	-220771.34	-220772.9	-220773.07	-220772.39	-220770.85	-220769.04
Angle B	175	175	175	175	175	175	175
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220768.44	-220771.94	-220773.26	NA	-220772.76	-220771.45	-220769.54
Angle B	185	185	185	185	185	185	185
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220768.39	-220771.94	-220773.26	NA	-220772.84	-220771.49	-220769.52
Angle B	195	195	195	195	195	195	195
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220767.54	-220771.34	-220772.9	-220773.05	-220772.38	-220771.24	-220769.16
Angle B	205	205	205	205	205	205	205
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220766.2	-220769.61	-220771.08	-220771.54	-220771.31	-220770.05	-220768.3
Angle B	215	215	215	215	215	215	215
Angle A	90	100	110	120	130	140	150
Energy(kcal/mol)	-220764.61	-220768.03	-220769.64	-220769.96	-220769.39	-220768.25	-220766.08

Table 1.2.s2 Energy of DFT 2D scan with wB97XD/6-31+g(d,p) level of theory.

1.2.1.1.3 Conformational search and energies of A

Energy of each conformer is shown in Table 1.2.s3, only the conformers that are within 5kcal/mol of the lowest energy conformer were kept for docking. All structures are available as mol2 file.

Conformers	Free Energy(kcal/mol)	Relative Energy(kcal/mol)
1	-490173.9459	0
2	-490173.8185	0.12738453
3	-490173.7602	0.18574296
4	-490173.7369	0.20896083
5	-490173.7281	0.21774597
6	-490173.5135	0.43235439
7	-490171.6781	2.26782114
8	-490171.6291	2.31676692
9	-490171.2238	2.72213838
10	-490171.038	2.90788134
11	-490170.8071	3.13880502
12	-490170.6653	3.28062228
13	-490170.2875	3.6583833

Table 1.2.s3 Free energies and relative energies of each conformer.

1.2.1.2 Docking Simulation

1.2.1.2.1 Comparison of four crystal structures.

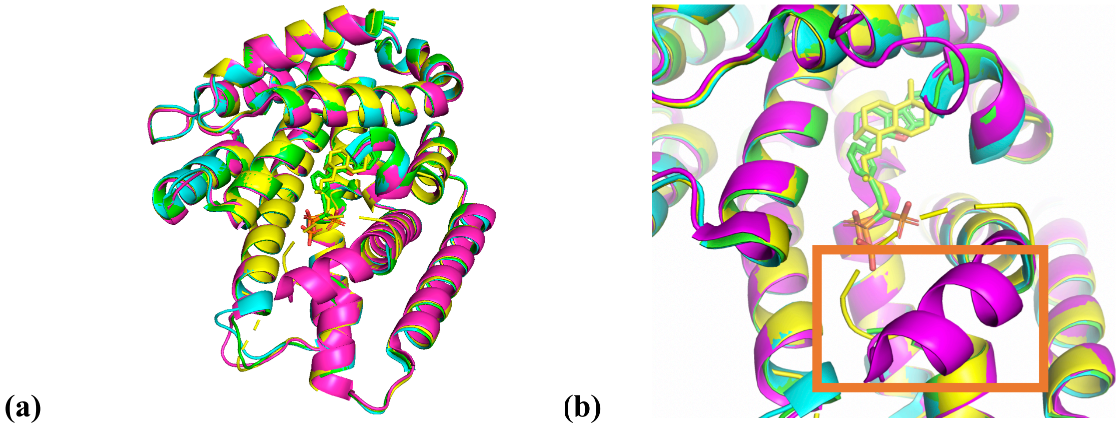


Figure 1.2.s1 (a) Comparison of four crystal structures of BjKS. 4W4R: apo, cyan. 4W4S: with B29, green. 4XLX: apo, magenta. 4XLY: with CPP, yellow. (b) Comparison of the active site of structures in (a). The orange box indicates the region that is complete in 4XLX but is incomplete in other three structures.

1.2.1.2.2 Docking constraint Setup

During the docking simulation, chemically meaningful constraints were applied during the sampling. Specifically, three types of constraints were applied.

1. Chemistry constraints

During the first step of the mechanism, diphosphate-Mg^{2+} complex leaves carbon 14 to form first carbocation intermediate. C14 can leave from two possible oxygens. Since there’s no experimental result confirm which one it is, here we apply constraints to examine both possibilities. Shown in Figure 1.2.S2.

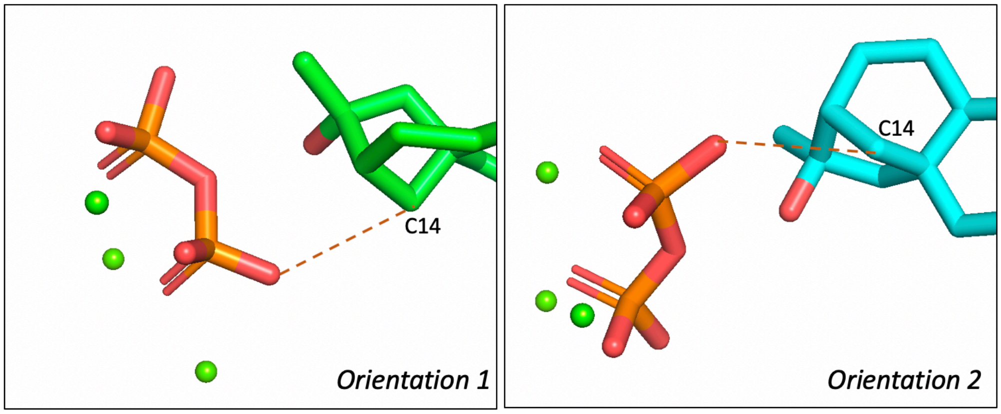


Figure 1.2.s2 Two possible orientations of the DPP complex leaving in the first step of the reaction mechanism.

2. Coordination constraints

Diphosphate-Mg^{2+} complex is coordinating with the aspartate rich motifs. This set of constraints define the distance/angle/dihedrals to be similar with orientations observed in crystal structures with the diphosphate and all three magnesiums, shown in figure 1.2.S3. All constraine values are summarized in Table 1.2.s4. Two magnesiums on the left were constraint to the DDXXD motif while the magnesium on the right was constraint to N205. The typical NSE motif is NGD in BjKS. MG1 is not constraint to the last aspartate because the absence of the diphosphate-Mg^{2+} complex in the crystal structure might lead to the H helix slightly shift away from the active site. During the docking simulation we allow active site backbone movement and extensive side chain conformation sampling to maximumly overcome this issue.

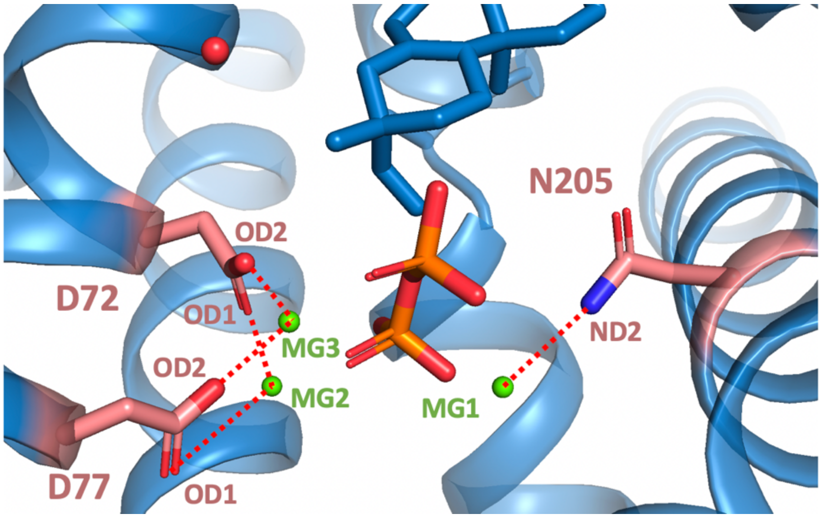


Figure 1.2.s3 Coordination constraints.

All constraint values were obtained from crystal structures. Weight for distance is 10 times of angle or dihedral angles since distance range is less likely to be outside of the set range. All numbering is from the mol2 file numbering. For water constraint, weight for distance is set to 50, angle B and torsion B is set to 30.

3. Water constraints

Initial docking results using these constraints indicated that E28 and F72 sidechains frequently adopted conformations different from those observed in the crystal structure (Figure 1.2.S4). It does not seem likely that a nonpolar group would vacate space in the carbocation (hydrocarbon) binding site to be replaced by a polar group. Close examination of structures from docking indicated that the problem might be that a water molecule (Figure 1.2.S4 left, red) that is present in the crystal structure was absent in the docking simulation (removal of water molecules is a common step in docking procedures). pyWater analysis confirmed the conservation of this water molecule in related crystal structures. (Figure 1.2.S4, right.) The only water that shows conservation in active site is the highlighted large sphere shown in Figure S4, right. Thus, to ensure a reasonable conformation of the important residues in the active site, this water molecule was constrained to the D76 and E28 residues during subsequent docking.

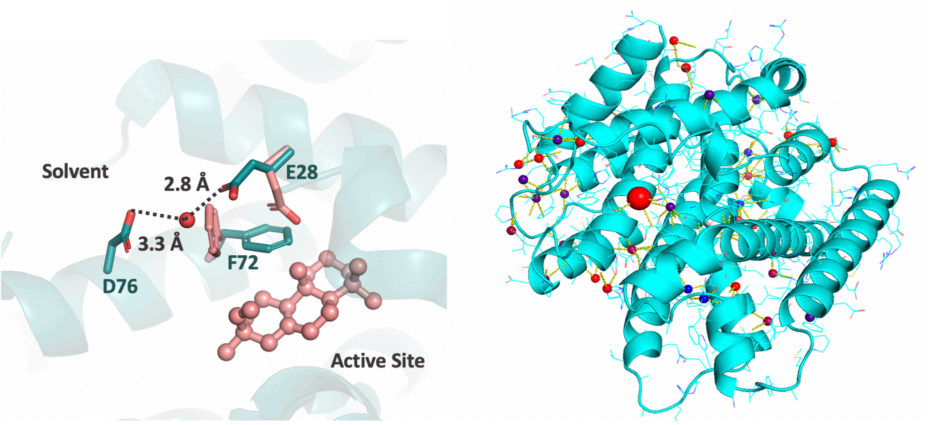


Figure 1.2.s4. Left: Preliminary docking results with angle constraints. Green: crystal structure. Salmon: docking result for A167S mutant. Docked carbocation A shown as a ball-and-stick model.

Right: Pywater analysis: All water molecule in the crystal structure were examined and water with conservation over 0.7 were shown in the figure. Darker color suggests higher conservation. The conserved water used in docking is highlighted (red large sphere).

The constraints are set between E26 and D74 with water shown in figure below. Values are shown in table 1.2.s4. These values are obtained from crystal structures.

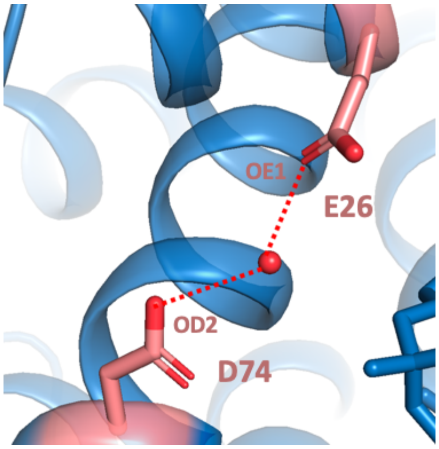


Figure 1.2.s5 Water constraints between D74 and E26.

4. Angle constraints on Ser167

Angle constraints are applied as QM result suggests. Angle A are constrained at 120 degree with 10% deviation while angle B are constrained at 180 with 10% deviation. All the angles are applied to 3 different carbons (5 different proton positions since C7 and C14 has 2 protons attach to them while C9 has one).

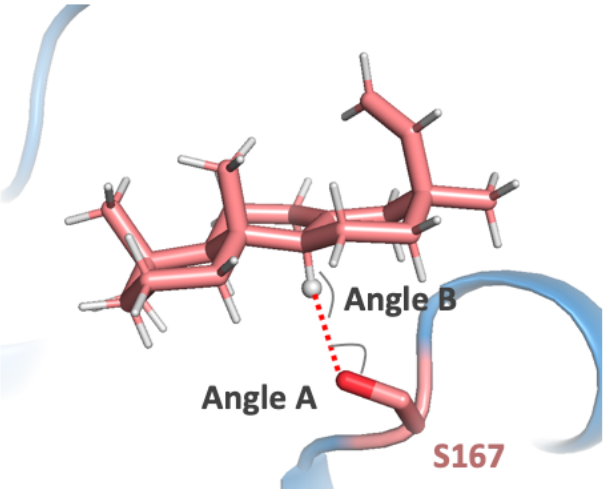


Figure 1.2.s6 Angle constraints between S167 and intermediate.

Chemistry Constraint	Distance		
DPP - C14	3.0 ± 0.5		
Coordination Constraint	Distance	Angle	Torsion
D73_OD2 - MG3	2.5 ± 0.3	145.3 ± 20	128.7 ± 20
D73_OD1 - MG2	2.5 ± 0.5	129.3 ± 20	290.3 ± 20
D77_OD2 - MG3	2.5 ± 0.3	107.2 ± 20	155.1 ± 20
D77_OD2 - MG2	2.5 ± 0.3	129.4 ± 20	220.6 ± 20
N205_ND2 - MG1	2.5 ± 0.5	147.3 ± 20	95.4 ± 20
Water Constraint	Distance	Angle	Torsion
26E_OE1 - Water O	2.8 ± 0.5	103.8 ± 10	191.8 ± 19
74D_OD2 - Water O	3.3 ± 0.5	100.7 ± 10	230.6 ± 23
Angle constraint	Distance	Angle	Torsion
A167S - angle A	NA	120.0 ± 12	NA
A167S - angle B	NA	180.0 ± 18	NA

Table 1.2.s4 Docking constraint values for A167S structure.

1.2.1.2.3 Diphosphate orientation

In the docking study, diphosphate-Mg2+ complex geometry was taken from the the closest homologue. To validate the conformation, we examined all class I terpene synthase crystal structure with complete diphosphate-Mg2+ complex bound. Including 2ONG,1N21,5IKA,1JFG,2OA6,3KB9,4OKZ. As shown in figure S7, the diphosphate-Mg2+ complex has very conserved conformation adopted by the enzyme catalytic residues (DDXXD motif and (N,D)D(L,I,V)X(S,T)XXXE motif, bold residues are catalytic residues).

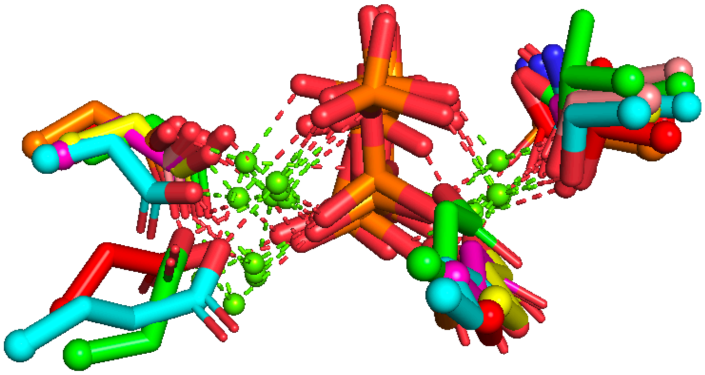
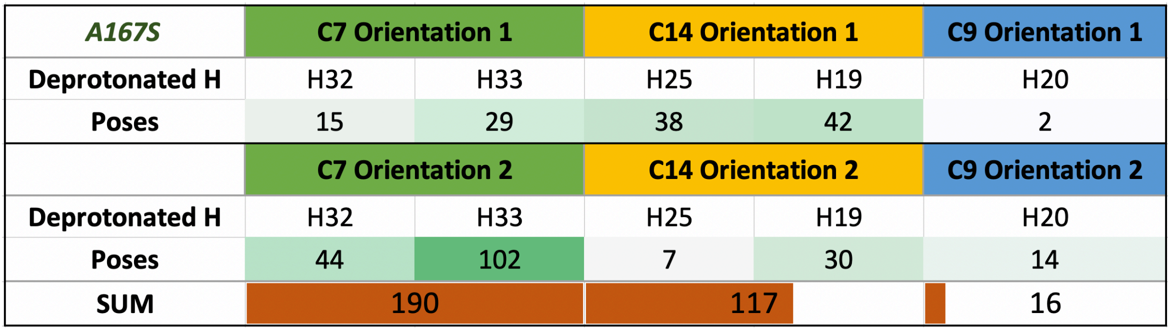


Figure 1.2.s7 Class I terpene synthase diphosphate-Mg2+ conformation overlap with catalytic residues. (Salmon: 2ONG, Green: 1n21, Red: 5ika, Orange: 1jfg, Meganta: 2OA6, Yellow: 3KB9, Cyan: 4OKZ)

1.2.1.2.4 Docking result for A167S

The number of poses passed filters are shown in Table 5. C7 and C14 bear two hydrogen (C7: H32, H33, C14: H25, H19), while C9 bears one (H20). Docking runs were carried out individually for each possible deprotonation position and combined for filtering. Result poses for each deprotonated H are available as PDB files, labeled as the proton name. Representative pose shown in Figure 3b is available as m2_c10_h33_a167s.pdb.

For each of the 10 possibilities, 25,000 docking runs were performed, adds up to 250,000 docking runs in total. All docking results were combined and then filtered based on satisfaction of constraints, total protein energy and interface energy. On the first stage of filtering, all docking results are taken into consideration. If a pose satisfies the constraint described in last section with minimal deviation, it'll be kept for further docking. This step left with 82912 poses. Secondly, poses are filtered based on total protein score. Lowest 10% were kept (lower than -738.25 REU, Rosetta energy unit). Thirdly, poses are filtered based on interface energy and lowest 5% were kept (lower than -17.10 REU). That leads to the final poses pass through the filter, summarized in chart below. All poses are examined and few poses with Ser-H distance greater than 4Å was discarded.



A167S	C7 Orientation 1		C14 Orientation 1		C9 Orientation 1
Deprotonated H	H32	H33	H25	H19	H20
Poses	15	29	38	42	2
	C7 Orientation 2		**C14 Orientation 2**		**C9 Orientation 2**
Deprotonated H	H32	H33	H25	H19	H20
Poses	44	102	7	30	14
SUM	190		117		16

Table 1.2.s5 Docking results. The darker green a cell is, the more poses passed filter. The sum of poses for deprotonation on each carbon is shown in the last row, with the brown bars representing relative amounts of deprotonation at C7, C14 and C9.

1.2.1.2.5 Wild type docking

Four ion-pair orientations were examined individually during the wild-type docking simulation. Carbon 16 and 17 were constraint to either of the two possible oxygens on diphosphate, resulting in four orientations.

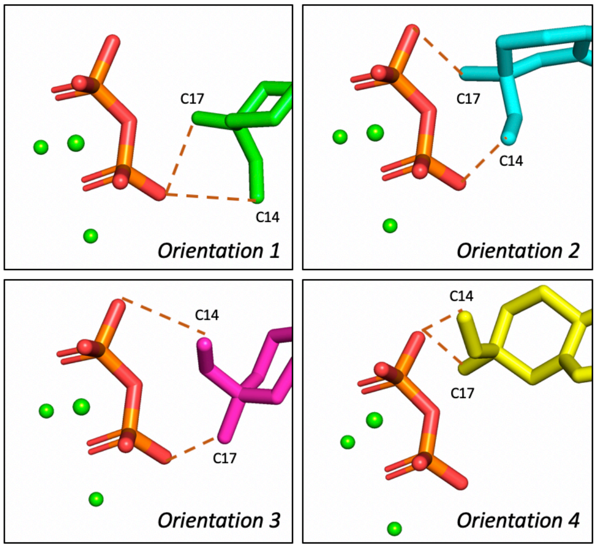


Figure 1.2.s8 Four possible ion-pair orientations during wild-type docking simulation.

Three hydrocarbon structures: A, C, and transition state structure A to C TS(A-C) were docked in wild-type BjKS. Overlay of representative low energy structure of A, TS(A-C) and C are shown in Figure S9. No significant rotational or translational movement between each structure was observed. Detailed information will be reported in future publication. The result poses are available as PDB files.

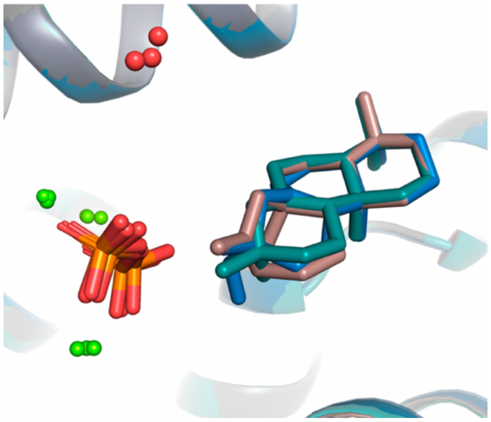


Figure 1.2.s9 Overlay of A(navy), TS(A-C) (dark green), C (light brown) low energy structures.

1.2.5 References

(1) Christianson, D. W. Structural and Chemical Biology of Terpenoid Cyclases. Chem. Rev. 2017, 117, 11570−11648.

(2) Pemberton, T. A.; Christianson, D. W. General Base-General Acid Catalysis by Terpenoid Cyclases. J. Antibiot. 2016, 69, 486−93.

(3) Xu, M.; Wilderman, P. R.; Peters, R. J. Following Evolution's Lead to a Single Residue Switch for Diterpene Synthase Product Outcome. Proc. Natl. Acad. Sci. U. S. A. 2007, 104, 7397−7401.

(4) Wilderman, P. R.; Peters, R. J. A Single Residue Switch Converts Abietadiene Synthase into a Pimaradiene Specific Cyclase. J. Am. Chem. Soc. 2007, 129, 15736−15737.

(5) Morrone, D.; Xu, M.; Fulton, D. B.; Determan, M. K.; Peters, R.J. Increasing Complexity of a Diterpene Synthase Reaction with a Single Residue Switch. J. Am. Chem. Soc. 2008, 130, 5400−5401.

(6) Jia, M.; Peters, R. J. Extending a Single Residue Swtich for Abbreviating Catalysis in Plant ent-Kaurene Synthases. Front. Plant Sci. 2016, 7, No. e1765.

(7) Keeling, C. I.; Weisshaar, S.; Lin, R. P. C.; Bohlmann, J. Functional Plasticity of Paralogous Diterpene Synthases Involved in Conifer Defense. Proc. Natl. Acad. Sci. U. S. A. 2008, 105, 1085−1090.

(8) Zerbe, P.; Chiang, A.; Bohlmann, J. Mutational Analysis of White Spruce (Picea glauca) ent-Kaurene Synthase (PgKS) Reveals Common and Distinct Mechanisms of Conifer Diterpene Synthases of General and Specialized Metabolism. Phytochemistry 2012, 74, 30−9.

(9) Kawaide, H.; Hayashi, K.; Kawanabe, R.; Sakigi, Y.; Matsuo, A.; Natsume, M.; Nozaki, H. Identification of the Single Amino Acid Involved in Quenching the ent-Kauranyl Cation by a Water Molecule in ent-Kaurene Synthase of Physcomitrella patens. FEBS J. 2011, 278(1), 123−33.

(10) Irmisch, S.; Muller, A. T.; Schmidt, L.; Gunther, J.; Gershenzon,J.; Kollner, T. G. One Amino Acid Makes the Difference: The Formation of ent-Kaurene and 16α-hydroxy-ent-Kaurane by Diterpene Synthases in Poplar. BMC Plant Biol. 2015, 15, 262.

(11) Jia, M.; O'Brien, T. E.; Zhang, Y.; Siegel, J. B.; Tantillo, D. J.; Peters, R. J. Changing Face: A Key Residue for the Addition of Water by Sclareol Synthase. ACS Catal. 2018, 8, 3133−3137.

(12) Zhou, K.; Peters, R. J. Electrostatic Effects on (Di)Terpene Synthase Product Outcome. Chem. Commun. 2011, 47, 4074−4080.

(13) Morrone, D.; Chambers, J.; Lowry, L.; Kim, G.; Anterola, A.; Bender, K.; Peters, R. J. Gibberellin Biosynthesis in Bacteria: Separate ent-Copalyl Diphosphate and ent-Kaurene Synthases in Bradyrhi-zobium japonicum. FEBS Lett. 2009, 583, 475−480.

(14) Nett, R. S.; Montanares, M.; Marcassa, A.; Lu, X.; Nagel, R.; Charles, T. C.; Hedden, P.; Rojas, M. C.; Peters, R. J. Elucidation of Gibberellin Biosynthesis in Bacteria Reveals Convergent Evolution. Nat. Chem. Biol. 2017, 13, 69−74.

(15) Liu, W.; Feng, X.; Zheng, Y.; Huang, C. H.; Nakano, C.; Hoshino, T.; Bogue, S.; Ko, T. P.; Chen, C. C.; Cui, Y.; Li, J.; Wang,I.; Hsu, S. T.; Oldfield, E.; Guo, R. T. Structure, Function and Inhibition of ent-Kaurene Synthase from Bradyrhizobium japonicum. Sci. Rep 2015, 4, 6214.

(16) Zhou, K.; Gao, Y.; Hoy, J. A.; Mann, F. M.; Honzatko, R. B.; Peters, R. J. Insights into Diterpene Cyclization from the Structure of the Bifunctional Abietadiene Synthase. J. Biol. Chem. 2012, 287, 6840−6850.

(17) Chen, F.; Tholl, D.; Bohlmann, J.; Pichersky, E. The Family of Terpene Synthases in Plants: A Mid-Size Family of Genes for Specialized Metabolism that is Highly Diversified Throughout the Kingdom. Plant J. 2011, 66, 212−29.

(18) Chai, J. D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Correc-tions. Phys. Chem. Chem. Phys. 2008, 10, 6615−20.

(19) O'Brien, T. E.; Bertolani, S. J.; Tantillo, D. J.; Siegel, J. B. Mechanisticallly Informed Predictions of Binding Modes for Carbocation Intermediates of a Sesquiterpene Synthase Reaction. Chem. Sci. 2016, 7, 4009−4015.

(20) O'Brien, T. E.; Bertolani, S. J.; Zhang, Y.; Siegel, J. B.; Tantillo,D. J. Predicting Productive Binding Modes for Substrates and Carbocation Intermediates in Terpene Synthases-Bornyl Diphosphate Synthase as a Representative Case. ACS Catal. 2018, 8, 3322−3330.

(21) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan,V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley,P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme,T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. 2017, 13, 3031−3048.

(22) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. E.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popovic, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. Methods Enzymol. 2011, 487, 545−74.

(23) Hong, Y. J.; Tantillo, D. J. Formation of Beyerene, Kaurene, Trachylobane, and Atiserene Diterpenes by Rearrangements that Avoid Secondary Carbocations. J. Am. Chem. Soc. 2010, 132, 5375−86.

(24) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G J.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Van Voorhis, T.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath,P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T.R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C. P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y.M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik,J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K.; Chipman, D. M.; Keil, F. J.; Warshel, A.; Hehre, W. J.; Schaefer, H. F., III; Kong, J.; Krylov, A. I.; Gill, P. M.; Head-Gordon, M. Advances in Methods and Algorithms in a Modern Quantum Chemistry Program Package. Phys. Chem. Chem. Phys. 2006, 8, 3172−3191.

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson,G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.;

Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J.V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson,T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V.N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; RendellA.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 09; Gaussian, Inc.: Wallingford, CT, 2016.

(26) Ansbacher, T.; Freud, Y.; Major, D. T. Slow-Starter Enzymes: Role of Active-Site Architecture in the Catalytic Control of the Biosynthesis of Taxadiene by Taxadiene Synthase. Biochemistry 2018, 57, 3773−3779.

(27) Deno, N. C.; Turner, J. O. The Basicity of Alcohols and Ethers. J. Org. Chem. 1966, 31, 1969−1970.

(28) McCormack, A. C.; More O'Ferrall, R. A.; O'Donoghue, A. C.; Rao, S. N. Protonated Benzofuran, Anthracene, Naphthalene, Benzene, Ethene, and Ethyne: Measurements and Estimates of pK(a) and pK(R). J. Am. Chem. Soc. 2002, 124, 8575−8583.

1.3 Enzyme engineering of *ent*-Kaurene Synthase from *Bradyrhizobium japonicum*

1.3.1 Introduction

Enzyme engineering is an important tool for enhance the enzyme catalytic activity, broaden the substrate, improve the selectivity, and increase the stability. [1] A few common applied methods are directed evolution, genome mining, and rational design. Recent developments of machine learning methods provide another alternative way for engineer enzymes. [2] Enzyme engineering has a variety of applications, from enhance the production for drug discovery, to increase the efficiency for producing products such as food and detergent. Specifically, terpene synthase engineering and metabolic engineering methods has been developed for producing perfume, pharmaceuticals, sweeteners, biofuels, flavors and other unnatural scaffolds etc. [3,4,5] In this section, we focus on rational design of diterpene synthase. Diterpenes and diterpenoids have been reported to have anticancer, antibacterial functions or is possible to be used as biomarkers. Previous section illustrated that the Serine mutation can serve as the catalytic base for deprotonation. To validate the method, another single mutation, Y134H is tested computationally, and the results are compared with experimental results. To explore the possibility of altering product outcome and potentially produce unnatural products, a few sequences of BjKS are designed computationally and tested by experiments.

1.3.2 Results and discussion

Method validation – Y136H

Mutagenesis studies of BjKS were carried out, one specific mutation, Y136H, produce 16a-hydroxyl-*ent*-kaurane. The epimer (shown in red, Figure 1.3.1) was not detected. Docking simulations were employed of the BjKS Y136H mutant. Preliminary results suggested that a water molecule coordinated with the His mutation possibly involves in the hydroxylation. Substrate docking studies suggest that a water molecule could be in presence along with the substrate ent-CPP in the active site, shown in Figure 1.3.2. Two ion-pair orientations of DPP with the substrate hydrocarbon were tested in docking simulations. Results shows both orientations are possible, although statistical analysis indicates that the salmon ion-pair orientation is preferred.

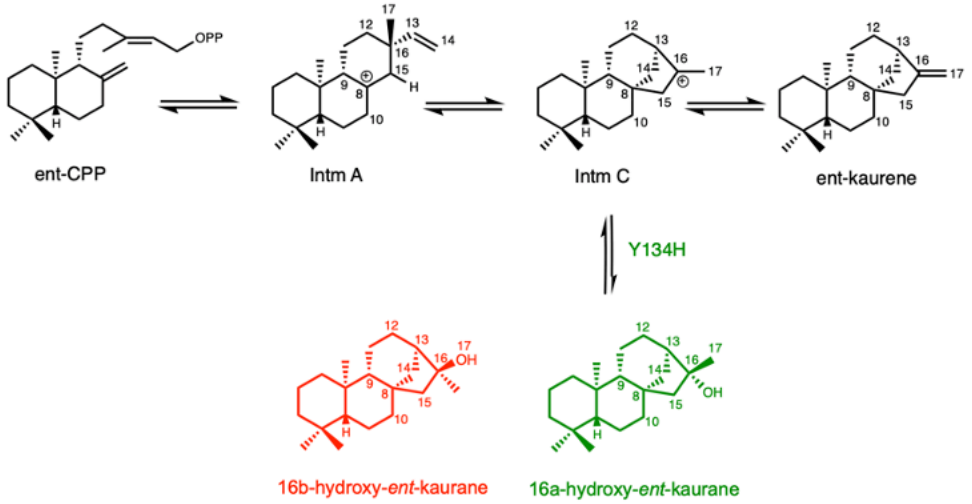


Figure 1.3.1 Enzyme catalysis mechanism of wild type BjKS(black) and Y136H(green).

Enzyme engineering of BjKS

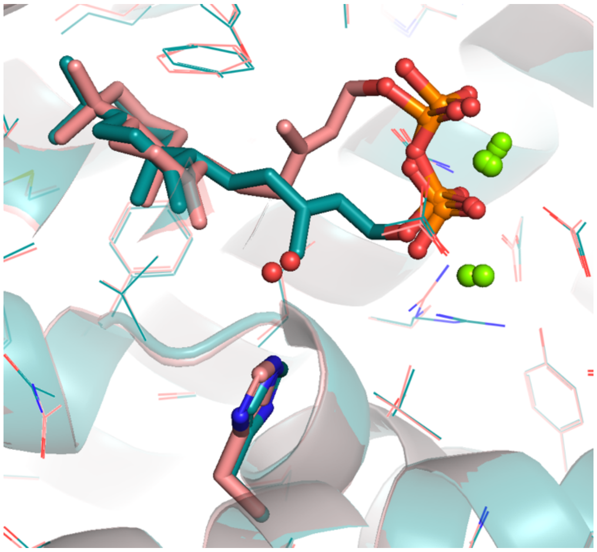


Figure 1.3.2 Binding orientation of ent-CPP with water molecule. His mutation is highlighted in sticks.

Enzyme engineering – BjKS

To explore the possibility of increasing the yield of **2** and **3**, or produce the **4** by BjKS, all residues within 6A were mutated individually to Ser or Thr in silico. Specifically, I166T mutation was predicted to be producing 2 as the major product computationally, as shown in 1.3.4. Mutations were then tested experimentally, results suggest that I166T produce only **2** and some wild-type product **5**.

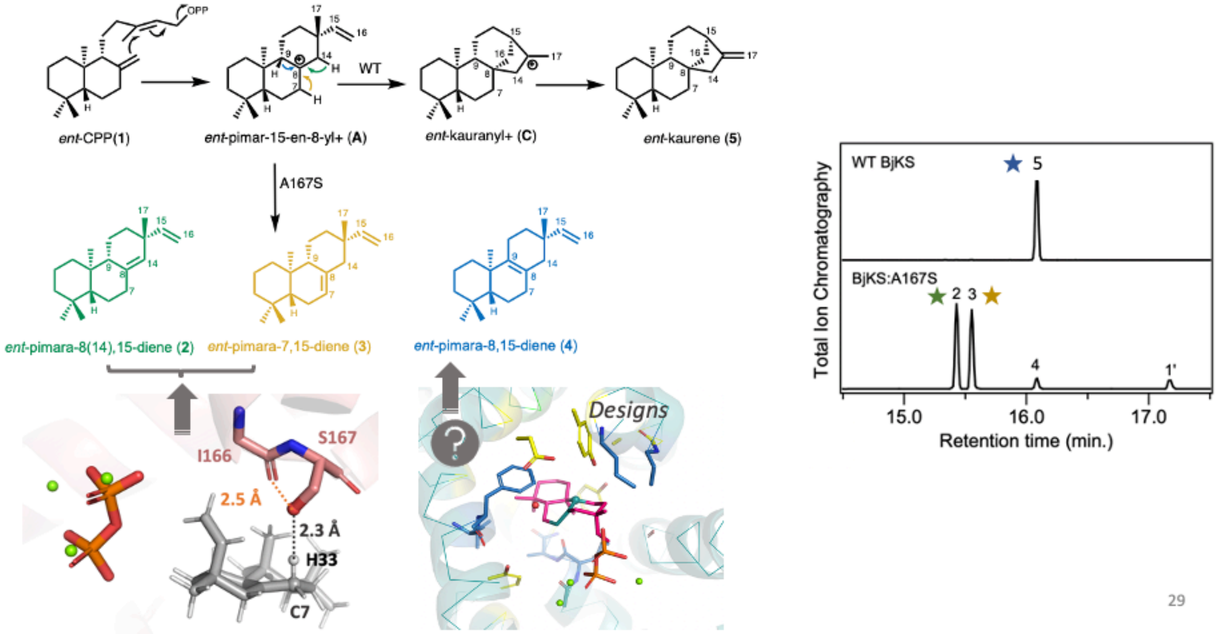
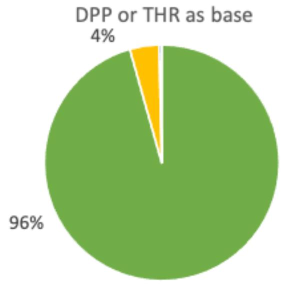


Figure 1.3.3 Design strategy.

WT	C7		C14		C9	
Deprotonated H	all		all		all	
Poses	707	2	35	11	4	0
SUM	1091		46		4	



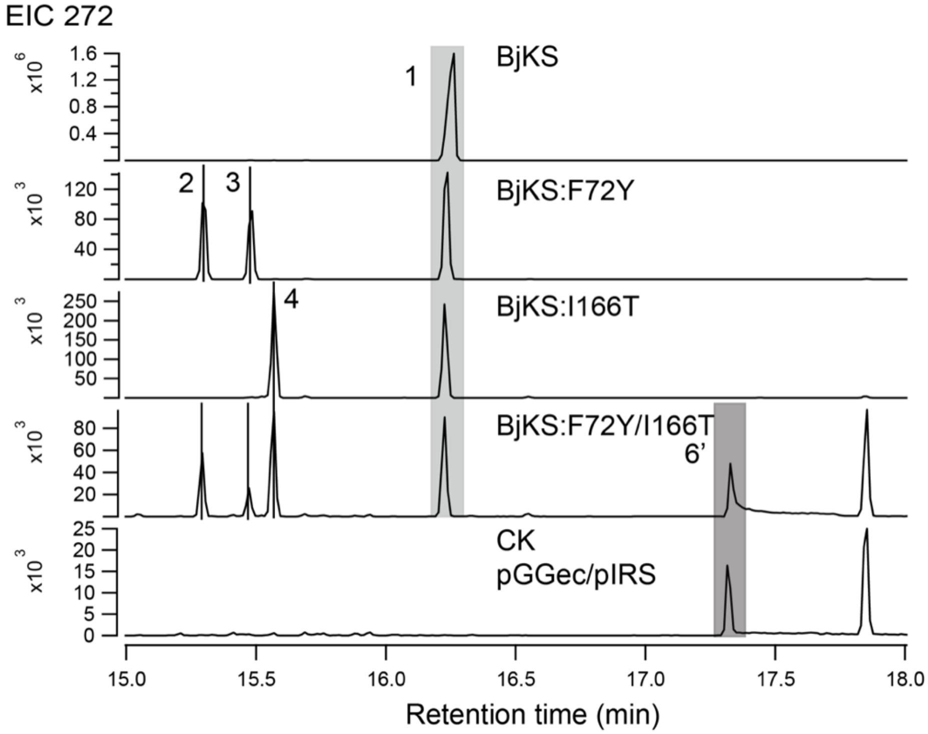
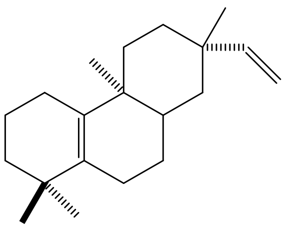
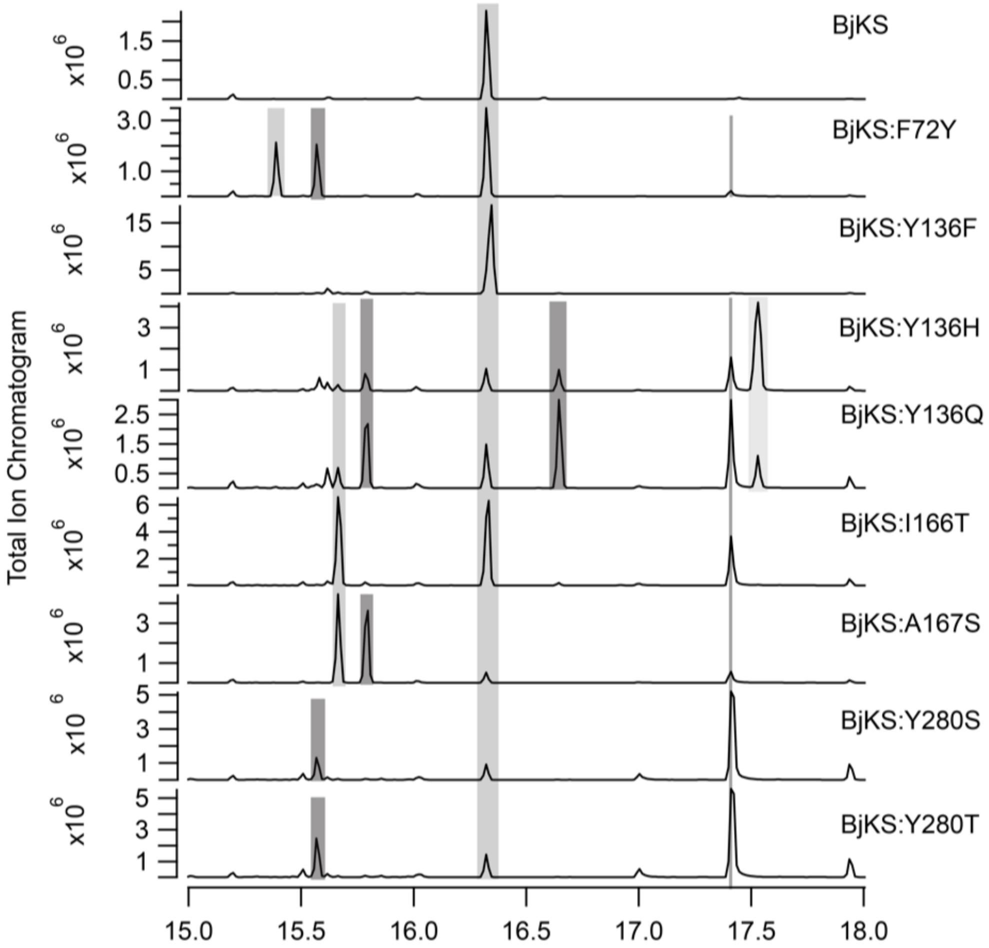
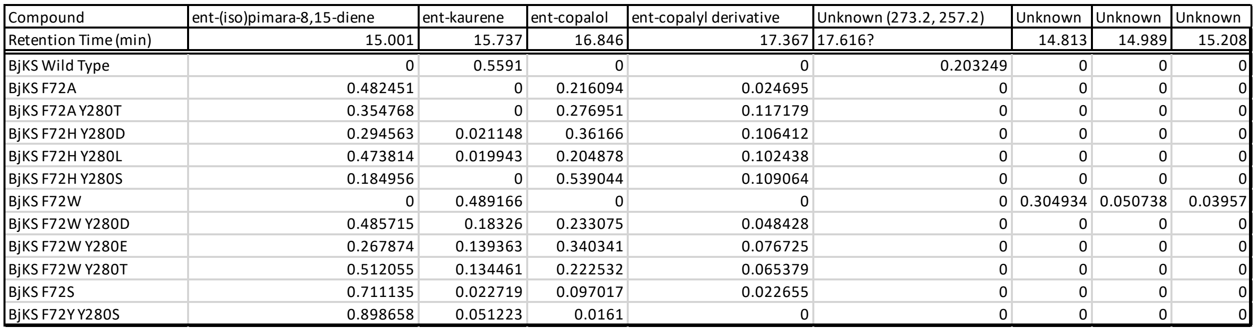


Figure 1.3.4 Top and middle: computational results of I166T. Bottom: Experimental testing of designs.

The other design effort made is to design BjKS to produce **4.** F72Y was shown to have some activity with some side product. Other single mutants and double mutants were explored computationally to increase the yield of **4**. Interestingly, Y280S or Y280T increase the specificity while decrease the activity of BjKS. A double mutant of Y280S/F72Y gives 90% yield of **4**. Noe the experimental result is currently the metabolic engineering data. Kinetic experiments are undergoing.







Compound	ent-(iso)pimara-8,15-diene	ent-kaurene	ent-copalol	ent-copalyl derivative	Unknown (273.2, 257.2)	Unknown	Unknown	Unknown
Retention Time (min)	15.001	15.737	16.846	17.367	17.616?	14.813	14.989	15.208
BjKS Wild Type	0	0.5591	0	0	0.203249	0	0	0
BjKS F72A	0.482451	0	0.216094	0.024695	0	0	0	0
BjKS F72A Y280T	0.354768	0	0.276951	0.117179	0	0	0	0
BjKS F72H Y280D	0.294563	0.021148	0.36166	0.106412	0	0	0	0
BjKS F72H Y280L	0.473814	0.019943	0.204878	0.102438	0	0	0	0
BjKS F72H Y280S	0.184956	0	0.539044	0.109064	0	0	0	0
BjKS F72W	0	0.489166	0	0	0	0.304934	0.050738	0.03957
BjKS F72W Y280D	0.485715	0.18326	0.233075	0.048428	0	0	0	0
BjKS F72W Y280E	0.267874	0.139363	0.340341	0.076725	0	0	0	0
BjKS F72W Y280T	0.512055	0.134461	0.222532	0.065379	0	0	0	0
BjKS F72S	0.711135	0.022719	0.097017	0.022655	0	0	0	0
BjKS F72Y Y280S	0.898658	0.051223	0.0161	0	0	0	0	0

Figure 1.3.5 Top: Side product of F72Y. Middle: experimental results of single mutants. Bottom: experimental results of some single mutants and double mutants.

(1) Sharma, A.; Gupta, G.; Ahmad, T.; Mansoor, S. Enzyme Engineering : Current Trends

and Future Perspectives Enzyme Engineering : Current Trends and Future Perspectives. *Food Reviews International*. 2019, pp 1–34.

(2) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* 2020, *10*, 1210–1223.

(3) Zhang, Y.; Nielsen, J.; Liu, Z. Engineering Yeast Metabolism for Production of Terpenoids for Use as Perfume Ingredients, Pharmaceuticals and Biofuels. *FEMS Yeast Res.* 2017, *17*, 1–11.

(4) Leferink, N. G. H.; Scrutton, N. S. Predictive Engineering of Class I Terpene Synthases Using Experimental and Computational Approaches. *ChemBioChem*. 2021.

(5) Ignea, C.; Pontini, M.; Motawia, M. S.; Maffei, M. E.; Makris, A. M.; Kampranis, S. C. Synthesis of 11-Carbon Terpenoids in Yeast Using Protein and Metabolic Engineering. *Nat. Chem. Biol.* 2018, *14*, 1090–1098.

Chapter 2 The Source of Rate Acceleration for Carbocation Cyclization in a Biomimetic Supramolecular Cage

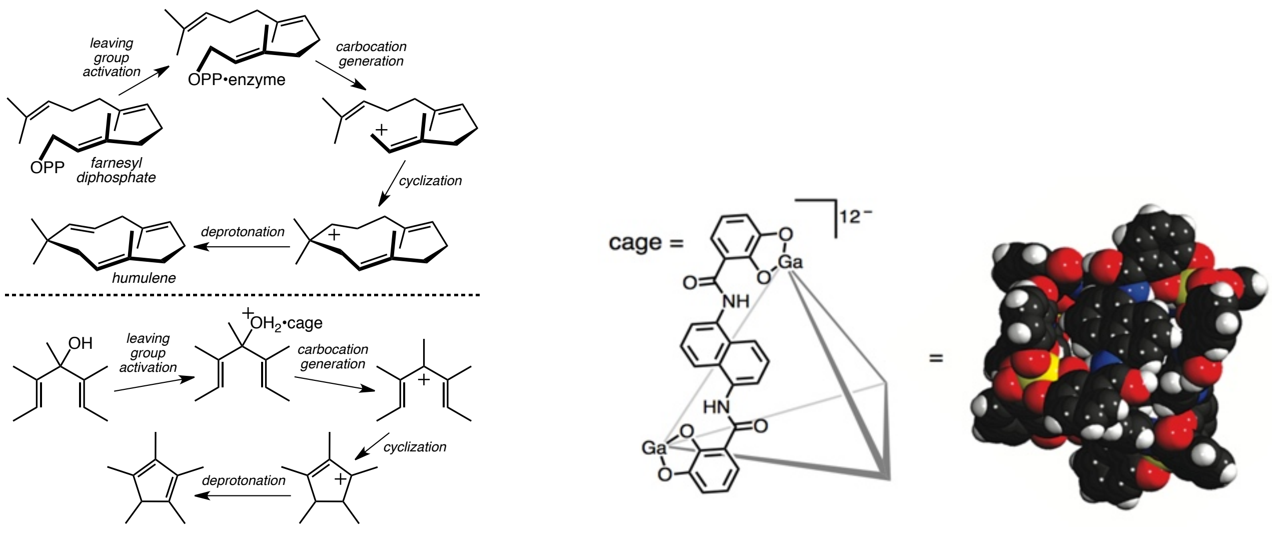
This project is a collaboration with Professor Leeping Wang, Professor Robert Bergman, Professor Ken Raymond, and Professor Dean Toste. Effects of cage walls is done by Nhu Nguyen, QM/MM and part of the MD is done by Nanhao Chen. Data permission is acquired.

2.1.1 Abstract

The results of quantum chemical and molecular dynamics calculations reveal that polyanionic gallium-based cages accelerate cyclization reactions of pentadienyl alcohols as a result of

substrate cage interactions, preferential binding of reactive conformations of substrate/ H_3O^+ pairs and increased substrate basicity. However, the increase in basicity dominates

Carbocations are involved as intermediates in catalytic reactions accomplished by both synthetic chemists and biological systems. The latter is most prominently found in the realm of terpene biosynthesis, where terpene synthase enzymes generate carbocations in productive conformations, allow them to rearrange while protecting them from premature deprotonation or addition by nucleophiles, and allow them to be deprotonated at specific positions (Scheme 1, top, OPP = diphosphate).[1] Synthetic chemists have developed systems that mimic these characteristics.[2,3] In both the synthetic and biosynthetic realms, however, the relative importance of each step in the reaction mechanism is still up for debate. Extensive computational work has targeted this issue for terpene synthasesm,[1c-d,4] but much less attention has been afforded to synthetic systems.[2,3,5] Here we describe molecular dynamics (MD) and quantum chemical computations on a 4-electron electrocyclization of pentadienyl carbocations promoted by the tetrahedral Ga4L4 cage studied by Raymond, Bergman, Toste and co-workers (Scheme 1,



bottom).[2] Our results indicate the essential role of the catalyst cage in promoting carbocation formation primarily via leaving group basicity enhancement.

Scheme 2.1.1 Left: A typical, albeit simple, terpene-synthase promoted reaction. Right: Electrocyclization promoted by the Ga4L4 cage.

2.1.2 Methods

Finding a computational approach for modeling such a large system that strikes the right balance between accuracy and efficiency is a challenge. Inspired by the success of Nitschke and co-workers in computing geometries for related cage structures,[6] we first examined the feasibility of employing various density functionals with relatively small basis. Ultimately, we found that the B3LYP/3-21G level of theory, [7] with the LANL2DZ basis set for Ga,[8] provided reasonable results, ssuitable for ssurveying a wide range of structures. Optimizing geometries and computing vibrational frequencies with ostensibly “better” levels of theory proved, in our hands, to be impractical at best and intractable at worst. However, single point energies on B3LYP/3-21G-LANL2DZ geometries with various other functionals and larger basis sets showed that the results described below were not very sensitive to the method used; all led to the same qualitative conclusions. In addition, for substrates in the absence of the cage, we performed optimization and frequency calculations with a variety of methods and larger basis sets; results were again very similar to those obtained with B3LYP/3-21G. All calculations were conducted using the CPCM continuum model for water (the solvent used experimentally),[9] which allowed the cage, which bears a 12– charge, not to expand significantly beyond its reported crystallographic geometry,[2,10] despite the absence of counterions in these calculations. All reported energies are free energies at 298 K unless stated otherwise.

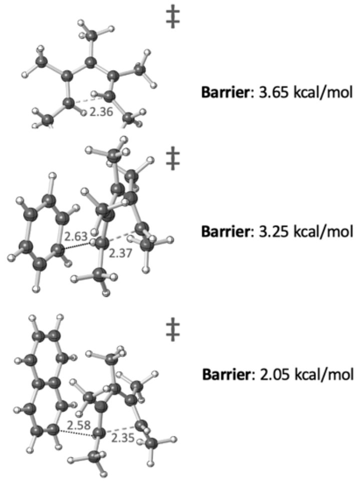
In addition to these static calculations, we carried out molecular dynamics (MD) calculations to address conformational and configurational flexibility. These calculations were set up based on the initial QM optimization results. The organic portions of the cage and the ligand molecules were described by the AMBER general forcefield (GAFF) and their charges were calculated by the RESP method based on ESP results obtained with HF/6-31G*. Since there are no parameters for Ga in AMBER, we substituted Ga by Al for our molecular mechanics (MM) calculations. All system preparations were done using the tleap program in AMBER16. In the MM/MD simulations, a multi-step strategy was used to heat and equilibrate the system gradually. First, energy minimization was carried out, followed by a 100 ps heating process (NVT ensemble). Then, another 100 ps MD simulation was carried out to balance the system (NPT ensemble with the help of the Berendsen barostat). Finally, a 50 ns MD simulation (NVT ensemble) was carried out. During the heating and equilibrium processes, an extra force was first added to the cage to prevent its contraction and the SHAKE5 algorithm was applied to constrain covalent bond lengths involving hydrogen. All MD calculations were carried out using the OpenMM package.

In addition, to estimate energy barriers for leaving group activation and departure, combined quantum mechanics (QM)/MM calculations were carried out using the Q-Chem and AMBER software packages. The QM region was set as the guest molecule, the H_3O^+ and two H2O molecules that engage in hydrogen-bond interactions with the H_3O^+. All atoms in the QM region were treated with B3LYP/6-31G*, and all the non-QM atoms were described by the forcefield mentioned above with a 12 Å cutoff for non-bonded interactions. QM/MM free energy calculations made use of umbrella sampling along the defined reaction coordinate (vide infra) and the weighted histogram analysis method.

2.1.3 Results

Effects of Cage Walls on Cyclization

Raymond, Bergman and co-workers observed rate accelerations of approximately six orders of magnitude for the substrate shown in Scheme 2.1.1 (and various isomers) in the presence of the cage. [2] First, we address the issue of whether or not direct interactions with the Ga4L4 cage lower the barrier for electrocyclization. The transition state structure (TSS) for electrocyclization in the absence of the cage (but in a water continuum) is shown at the top of Figure 2.1.1. A barrier of only ~4 kcal/mol was computed for cyclization from a productive, i.e., preorganized, conformer of reactant. Model calculations (Figure 2.1.1 middle and bottom) in the presence of a benzene or naphthalene molecule indicate that the aromatic walls of the cage can provide rate acceleration, although the effect is not large, i.e., while the walls interact with the substrate, [2,3d] selective binding of the transition state structure over the pentadienyl cation reactant is predicted

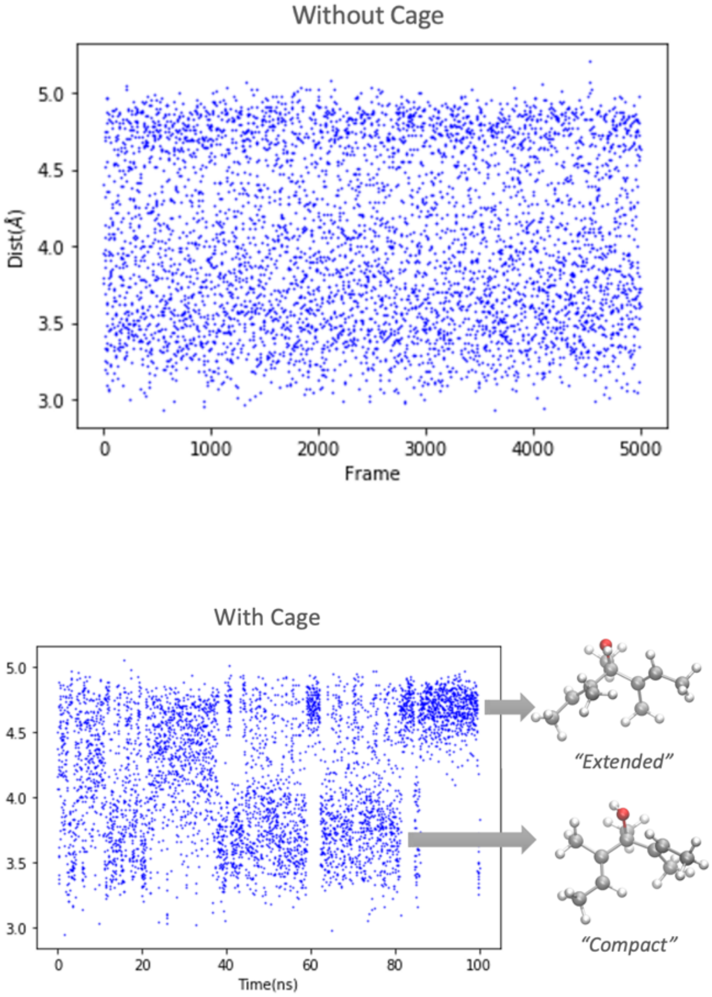


to amount to <2 kcal/mol, indicating that, at best, the large observed rate acceleration must originate primarily from another source.

Figure 2.1.1. Top to bottom: TSS of the model substrate, TSS with a benzene ring nearby, TSS with a naphthalene nearby. Each energy barrier was calculated (ωB97XD/6-311+G(d,p)) relative to the respective reactant with a similar environment, i.e. reactant alone or with a benzene or naphthalene in close proximity.

Conformational Preorganization

Another possible contributor to the observed rate acceleration is reactant preorganization, i.e., adopting the productive conformation in solution is accompanied by an energy penalty that is "paid" by the catalyst upon binding. Our DFT calculations suggest, however, that conformers of the pentamethylpentadienol reactant, its O-protonated form and the pentamethylpentadienyl cation, that are all coiled such that the two end carbons of the pentadienyl system are near to each other (all optimized in a water continuum), are within ~2 kcal/mol of the lowest energy conformer. For the carbocation, this is the lowest energy conformer. Our MD simulations indicate that pentamethylpentadienol in explicit water rapidly explores both extended and compact conformations (Figure 2.1.2, top). Upon complexation, it still explores these conformations, although transitions between them are not as smooth (Figure 2.1.2, middle). The barrier for transitioning between extended and compact forms inside the cage is estimated to be ~3 kT (roughly 2 kcal/mol at room temperature), while outside the cage, the barrier is close to 0. In general, as water molecules flow into the cage, the conformation of the guest changes from extended to compact. Overall, there appears to be no significant preorganization induced by the catalyst for the alcohol. Pentamethylpentadienol along with an explicit hydronium ion next to the hydroxyl group of the substrate was also modeled with MD (Figure 2.1.2, bottom). In this case, the substrate mainly adopts a relatively compact conformation, suggesting that that the alcohol/H_3O^+ pair is preferentially accommodated in orientations productive for reaction.



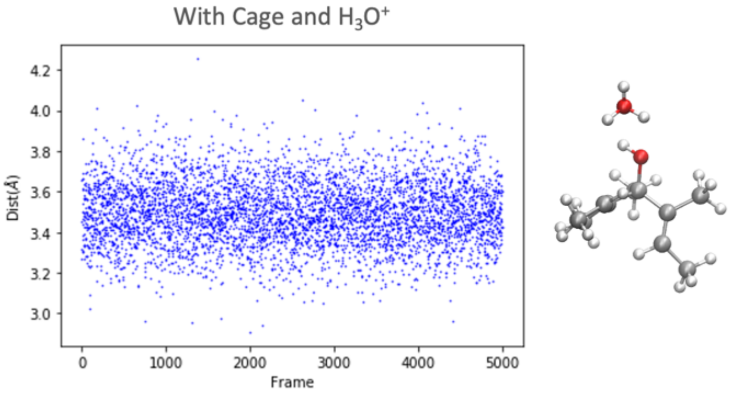


Figure 2.1.2 Distance between alkene carbons that will form the new C–C bond versus time during MD simulations. Top: pentamethylpentadienol in water. Middle: pentamethylpentadienol in the cage. Bottom: pentamethylpentadienol with hydronium ion in the cage.

Leaving Group Activation – Setting the Stage

That leaves us with leaving group activation. However, an important issue to address in this regard is the number of water molecules contained within the cage, along with any substrates. Warshel and co-workers previously examined the binding of four water molecules plus one H_3O^+ molecule within the same cage examined here. [5] Based on preliminary DFT calculations, we find that additional water molecules can be encapsulated without major distortions to the cage geometry. [11] We also addressed this issue with MD simulations of the cage surrounded by water molecules that can enter and leave its interior. Without the substrate present, a range of 0 to 16 water molecules were observed inside the cage over the length of the MD simulation, while 10 to 13 water molecules were the most common numbers (the cage was restrained to prevent shrinking). Recent experimental evidence pointed to 9 ± 1 water molecule in the cage in question. [5e] With the pentamethylpentadienol substrate present in the cage, 0 to 4 water molecules were observed in the MD simulation (Figure 2.1.3, left, “substrate” bars). When a hydronium ion was included during the MD, 2 to 5 water molecules were present in the cage along with the substrate (water count does not include the hydronium ion; Figure 2.1.3, left, “substrate + hydronium” bars). With the O-protonated pentamethylpentadienol ligand present, 2 to 6 water molecules were observed inside the cage (Figure 2.1.3, left, “cation” bars). A representative snapshot showing the substrate with H_3O^+ and three water molecules in the cage is shown at the right-hand side of Figure 3. The hydronium ion shown is poised to protonate the substrate, while other water molecules inside the cage form a H-bond network.

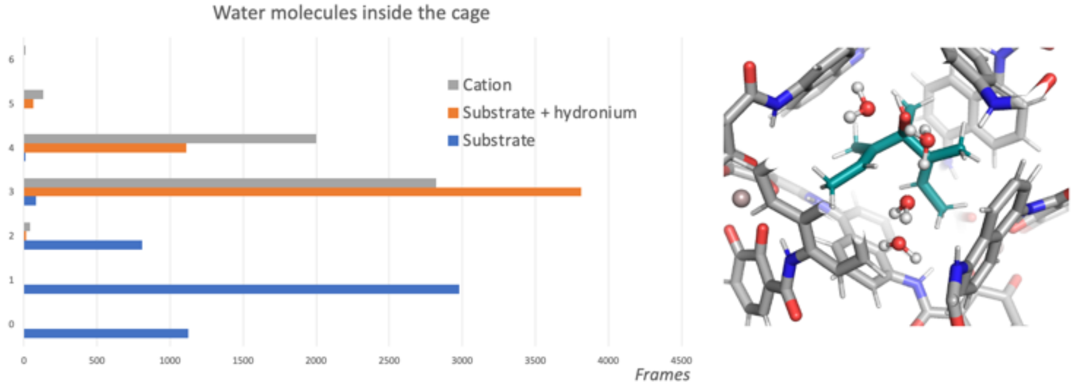


Figure 2.1.3. Left: water molecules inside the cage in MD. Right: A representative ssnapshot of substrate with H3O+ and three water molecules in the cage (water molecules outside the cage are not shown in this picture for clarity).

Leaving Group Activation – Protonation

We begin by discussing DFT (B3LYP/3-21G-LANL2DZ) results on an encapsulated pentamethylpentadienol substrate (Scheme 1), along with various numbers of water molecules and one H3O+ molecule. For these calculations, waters were placed manually near to the substrate’s hydroxyl group and allowed to relax into the nearest potential energy surface minimum; a consistent solvent configuration was used for all structures. Calculations with different numbers of water molecules and alternative orientations produced qualitatively similar results.

First, we focused on the degree to which the substrate protonation event is affected by the Ga4L4 cage. To do so, we calculated the free energies associated with protonation of the pentamethylpentadienol reactant inside and outside the cage (both in a water continuum), with various numbers of explicit water molecules present. In all cases, the pentamethylpentadienol reactant was predicted to be ≥15 kcal/mol more basic when encapsulated. We also selected configurations from MD simulations using GROMACS clustering with awareness of symmetry for DFT computations, and for these, encapsulated species are again predicted to be ~15 kcal/mol more basic (see SI for details). Warshel and co-workers also argued that the cage interior provides a "remarkable case of a low 'local pH'," [5a] i.e., encapsulated species are readily protonated. While our computed values here may be overestimates of the pKa modulation—previous experiments on amine, phosphine and ester guests indicate that a change in pKa of 4-5 units, corresponding to 5-7 kcal/mol can be expected upon encapsulation [10]—it is clear that alcohol basicity is markedly enhanced.

What is the origin of the high acidity of the microenvironment within the Ga4L4 cage? The cage used experimentally has walls made of naphthalene rings and, as shown in Figure 2.1.4a, the electrostatic potential within the cage is more negative than that on the cage's outer surface. [12] As a result, it is possible that substrate–cage CH–π interactions [13] strengthen upon protonation, leading to increased substrate basicity. To test the contribution of the cage walls, all naphthalene rings and attached amide groups were removed (and the remaining catechol rings were capped with hydrogen atoms; Figure 2.1.4b; note that this does not change the overall charge of the system), and the substrate protonation energy was recomputed (the geometry of the cage–substrate complex was not allowed to adjust). These changes led to a reduction in predicted protonation energy by 17 kcal/mol when no waters are co-encapsulated, consistent with there being interactions between the π-faces of the cage and the substrate. However, an increase in protonation energy of approximately the same amount is predicted for the complex of substrate and four bound water molecules, highlighting the fact that some bound waters tend to engage in OH–π interactions that, presumably, decrease the strength of their interactions with the protonated alcohol.

The contributions of specific cage oxygen atoms to basicity modulation were also examined. Keeping the naphthalene walls, but deleting all ortho catechol oxygen atoms, all meta catechol oxygen atoms, or all ortho and meta oxygen atoms along with all gallium atoms (and capping the remaining aromatic carbons with hydrogen atoms), all changes which lead to a neutral truncated cage, decreased the basicity of bound substrate, in the absence of bound waters, by approximately 15-20 kcal/mol. This result indicates that both cage charge and π-rich walls contribute roughly equally to the overall effect in the absence of bound water molecules, but are antagonistically coupled. [14] However, in the presence of bound waters, some water molecules will intrude on specific cage-substrate interactions (vide infra).

We also examined the proton transfer process using QM/MM calculations. A representative snapshot from our MD simulations is shown in Figure 2.1.5 (cage removed for clarity). Note that this structure showcases the general observation that encapsulated waters cluster on one side of the hydrocarbon group as a hydronium ion hydrogen bonds to the substrate hydroxyl group and to two encapsulated waters. When a QM/MM scan of the proton transfer reaction coordinate is carried out, protonated pentamethylpentadienol results (the dip in the energy plot around r = 0.5), which is predicted to be an endothermic process by ~2.5 kcal/mol, associated with a barrier of ~3 kcal/mol. For comparison, we examined the protonation of several amines, reported previously to have pKa's shifted by 4-5 units, with the same approach. [10b] However, little to no barrier was observed computationally in each case, likely because these species are more basic compared to the alcohol, precluding meaningful comparisons (see SI for details).

Leaving Group Activation – Actually Leaving

What of leaving group loss? Continuing along the scan shown in Figure 2.1.5, disconnection of the leaving group inside the cage occurs, resulting in the encsapsulated pentamethylpentadienyl cation + 4 H2O. The barrier for this process is predicted here to be ~6.5 kcal/mol, somewhat higher than expected. In the absence of the cage, once the reactant hydroxyl group is protonated, departure of water is expected to occur with little to no barrier. A TSS was not located for this process (modelling such structures is notoriously difficult at best), [15] but a constrained calculation with the breaking bond fixed at 1.9 Å led to a predicted barrier of only 2.5 kcal/mol (note also that in the coiled conformer of O-protonated pentamethylpentadienol, the C–O bond is predicted to be approximately 1.7 Å long). This barrier is similar to the barrier of ~4 kcal/mol estimated previously for this type of reaction. [2b] Nonetheless, the overall barrier from alcohol to pentamethylpentadienyl cation is predicted to be significantly lower (reduced by ~7 kcal/mol) in the cage than in water (Figure 2.1.6).

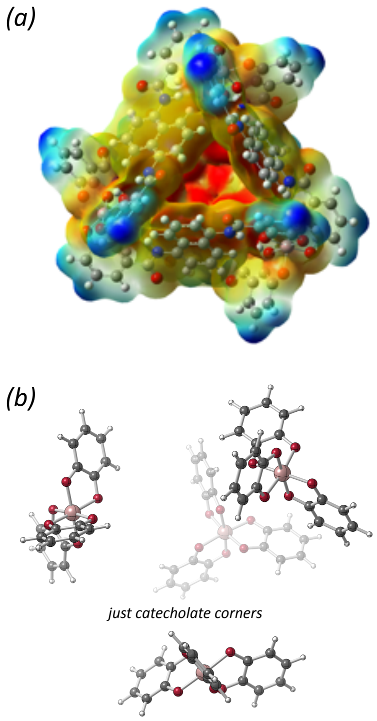


Figure 2.1.4. Electrostatic potential surface for empty cage (isovalue: 0.005, range: -0.83 to -0.5). (b) Just Ga-catecholate corners.

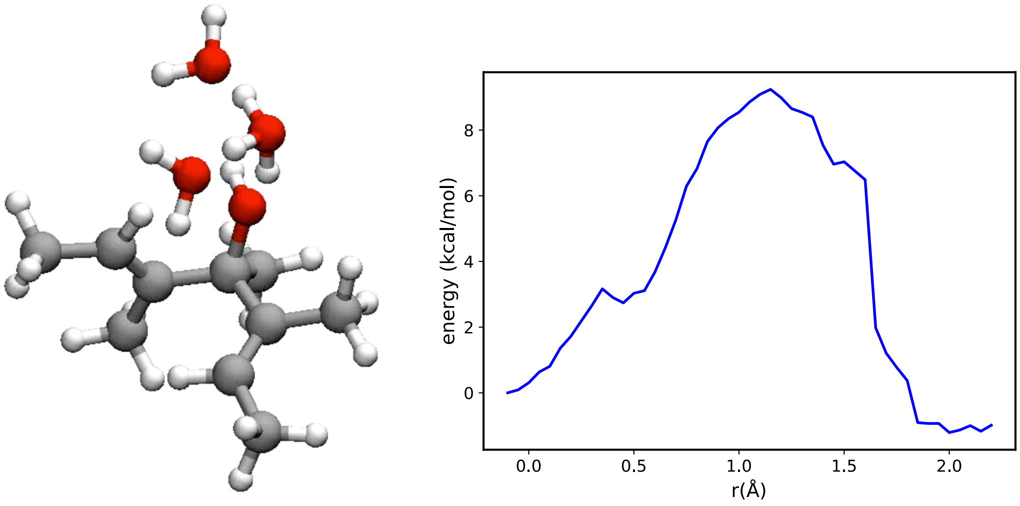


Figure 2.1.5. MD snapshot (cage and external water molecules removed) and proton transfer scan. r = the difference between the substrate C–O bond and the forming O–H bond. QM region

(modeled with B3LYP/6-31G*) includes all atoms shown in figure 4; MM region includes the cage and a water box (not shown in figure).

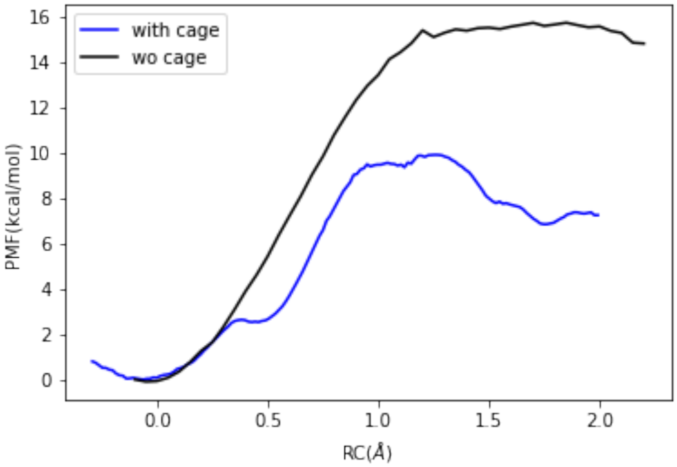


Figure 2.1.6. Proton transfer scan (RC = reaction coordinate, PMF = potential mean force) in water and in the cage.

Overall Model

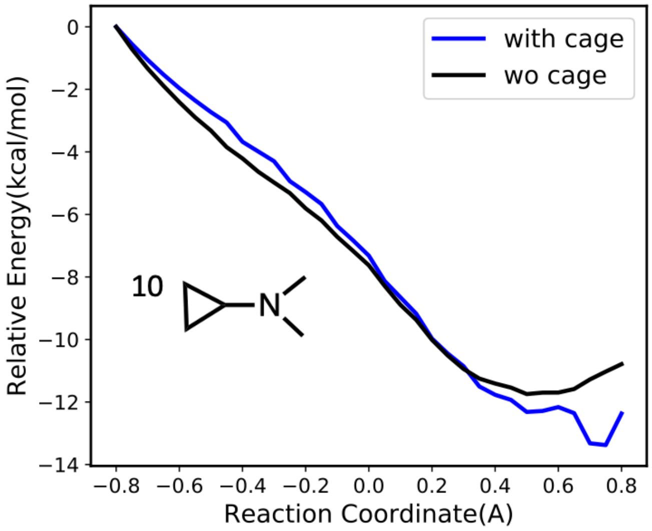
Thus, we arrive at the following model for rate acceleration. Protonation of the reactant alcohol is greatly enhanced upon complexation, leading to a large net acceleration in water loss. Electrocyclization may be enhanced, but only slightly. Warshel and co-workers also concluded that selective transition state stabilization was much less significant than reactant protonation for an orthoformate hydrolysis reaction promoted by Ga4L4 cage. [5a,10a]

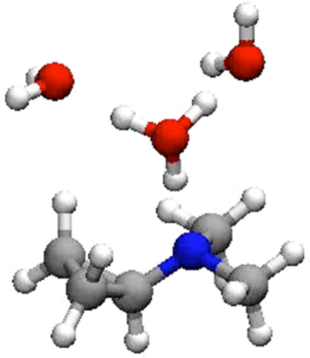
Evidence continues to mount that the reactivity of carbocations generated in the active sites of terpene synthases reflect their energetics in the gas phase. [1d] This was predicted on the basis of results of quantum chemical computations, and multiple experiments with enzymatic systems support this model. [16] Here we have shown that a very similar scenario occurs for the Ga4L4 cage; the cage facilitates formation of a pentadienyl carbocation, but barrier lowering is not at all necessary for the cyclization reaction step. While Type 1 terpene synthase enzymes activate their substrates' leaving groups (diphosphate groups) primarily by binding to Lewis acidic magnesium ions,[1a] the Ga4L4 cage activates its substrates' leaving groups via facilitation of protonation. Protonation is also used in other (Type 2) terpene synthases to generate carbocations from epoxides or alkenes.[17] In addition to utilizing active site amino acid sidechains with enhanced acidity (resulting from hydrogen-bonding arrays in which they participate), these terpene synthases provide active site cavities that are complementary to the carbocations generated upon protonation. Like the Ga4L4 cage, these active sites are lined with aromatic amino acid sidechains that can participate in carbocation–π interactions, [3d,13,18] however the cage also appears to bind several water molecules along with the substrate. [5b-c] While waters do sometimes bind along with substrates for terpene synthases, they tend to be fewer in number.

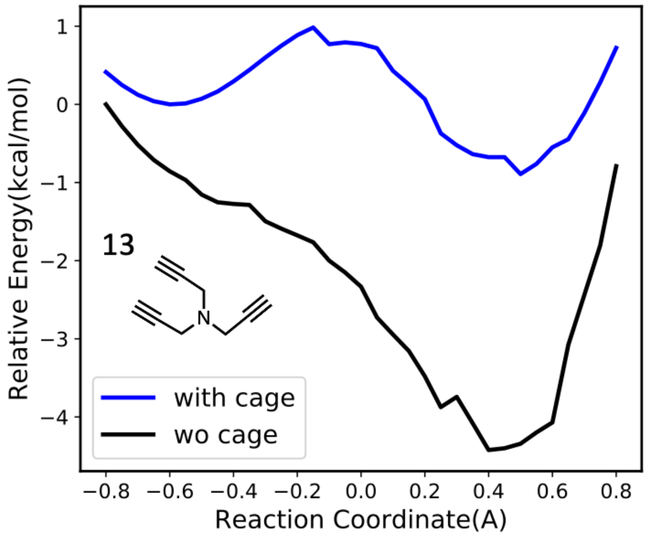
2.1.4 Supporting information

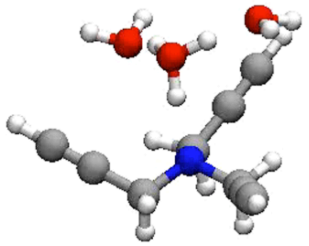
A. Amine protonation scans:

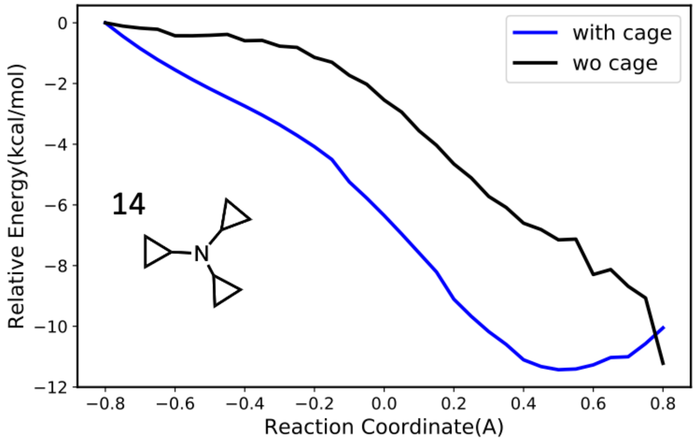
The protonation step of three amines(10, 13, 14, molecules are numbered the same as the experimental paper[10b]) were scanned with QM/MM method described in the main text. The pK_a shifts of 10, 13, 14 were experimentally reported to be 3.3, 3.9, 3.2 respectively. Little or no energy barrier were observed with the simulation as shown in figure below, which might due to the fact that these are highly basic species thus easy to be protonated.

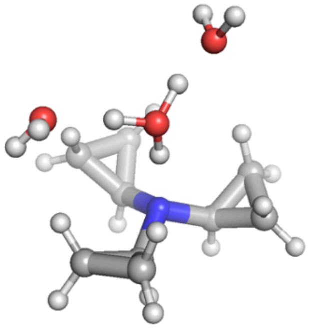

with cage
wo cage
10
N
Relative Energy(kcal/mol)
0
−2
−4
−6
−8
−10
−12
−14
−0.8 −0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8
Reaction Coordinate(A)




13
N
with cage
wo cage
Relative Energy(kcal/mol)
1
0
−1
−2
−3
−4
−0.8 −0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8
Reaction Coordinate(A)

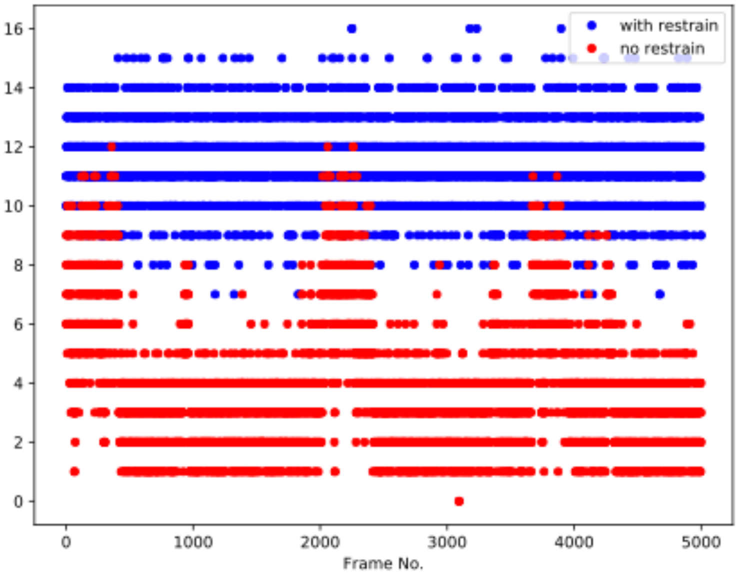



with cage
wo cage
14
N
Relative Energy(kcal/mol)
0
−2
−4
−6
−8
−10
−12
−0.8 −0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8
Reaction Coordinate(A)



B. Solvent sampling:

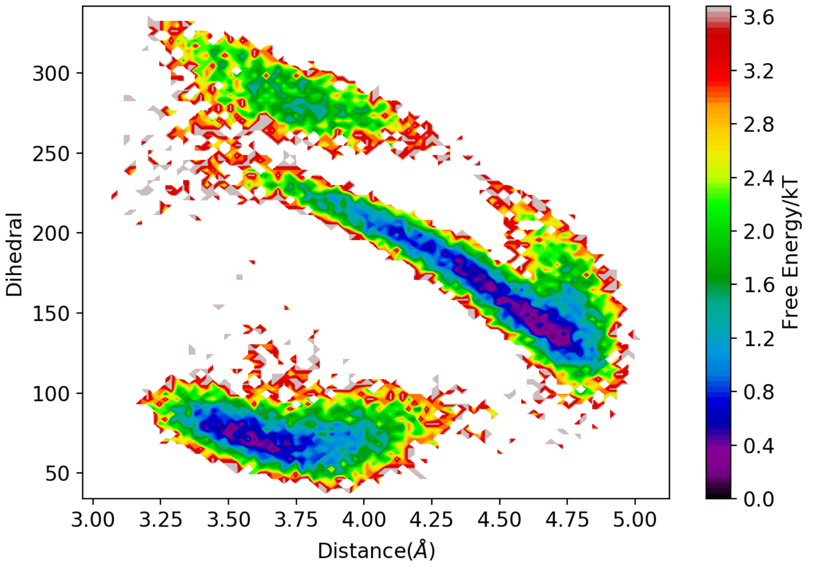
Water molecules “inside” the cage is defined as within 6A of the central carbon of the substrate. MD simulation result in 5000 frames, the amount of water molecules are counted each frame. The water amount without any ligand present in the cage is shown below.

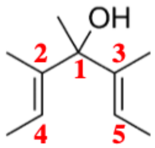


The water amount with ligand present in the cage is shown in the main text. The frames were clustered with GROMACS with awareness of the symmetry. Method is based on the script written by David Ascough.

C. Barrier of extended vs. compact conformation transition

Energy difference between compact and extended conformations are calculated. X axis is the distance of the two carbons that forms the new C-C bond. Y axis is the sum of the absolute values of the dihedrals: | dh (O-C1-C2-C4)| + |dh(O-C1-C3-C5) |.

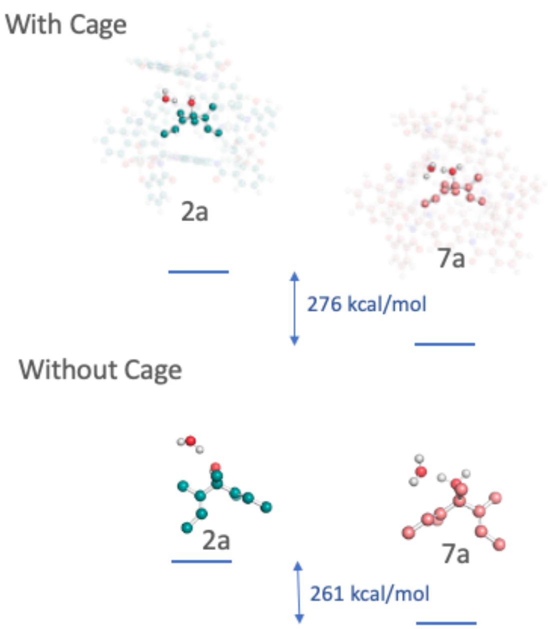




2a

D. Free energy calculation of selected conformation

Free energy differences of substrate and protonated substrate were calculated with and without the cage. A ddG of 15kcal/mol were observed. Note the configurations are selected based on the GROMACS clustering with the symmetry in consideration. In the figure below ligand along with 2 water molecules present in the cage were selected as an example configuration.



2.1.5 References

[1] a) D. W. Christianson, Chem. Rev. 2006, 106, 3412-3442; b) D. W. Christianson, Curr. Op. Chem. Biol. 2008, 12, 141-150; c) D. J. Tantillo, Nat. Prod. Rep. 2011, 28, 1035-1053; d) D. J. Tantillo, Angew. Chem. Int. Ed. 2017, 56, 10040-10045.

[2] a) C. J. Hastings, M. D. Pluth, R. G. Bergman and K. N. Raymond, J. Am. Chem. Soc. 2010, 132, 6938-6940; b) C. J. Hastings, R. G. Bergman and K. N. Raymond, Chem. Eur. J. 2014, 20, 3966-3973.

[3] Recent representative examples: a) S. V. Pronin, R. A. Shenvi, Nature Chem. 2012, 4, 915-920; b) Q. Zhang and K. Tiefenbacher, Nature Chem. 2015, 7, 197-202; c) M. J. Geier and M. R. Gagne, J. Am. Chem. Soc. 2014, 136, 3032-3035; d) C. R. Kennedy, S. Lin and E. N. Jacobsen, Angew. Chem. Int. Ed. 2016, 55, 12596-12624; e) Q. Zhang, L. Catti, J. Pleiss, K. Tiefenbacher, J. Am. Chem. Soc. 2017, 139, 11482-11492.

[4] a) S. R. Hare and D. J. Tantillo, Beilstein J. Org. Chem. 2016, 12, 377-390; b) D. J. Tantillo, Comprehensive Natural Products III: Chemistry and Biology 2020, 1, 644-563.

[5] a) M. P. Frushicheva, S. Mukherjee and A. Warshel, J. Phys. Chem. B 2012, 116, 13353-13360; b) G. Norjmaa, J.-D. Maréchal , G. Ujaque, J. Am. Chem. Soc. 2019, 141, 13114-13123; c) G. Norjmaa, J.-D. Maréchal , G. Ujaque, Chem. Eur. J. 2020, 31, 6988-6992; d) G. Norjmaa, J.-D. Maréchal , G. Ujaque, Chem. Eur., in press, DOI: 0.1002/chem.202102250; e) F. Sebastiani, T. A. Bender, S. Pezzotti, W.-L. Li, G. Schwaab, R. G. Bergman, K. N. Raymond, F. D. Toste, T. Head-Gordon, M. Havenith, Proc. Natl. Acad. Sci. USA 2020, 117, 32954-32961.

[6] T. K. Ronson, A. B. League, L. Gagliardi, C. J. Cramer and J. R. Nitschke, J. Am. Chem. Soc. 2014, 136, 15615-15624.

[7] a) A. D. Becke, J. Chem. Phys. 1993, 98, 5648-5652; b) A. D. Becke, J. Chem. Phys. 1993, 98, 1372-1377; c) C. Lee, W. Yang, R. G. Parr, Phys. Rev. B 1988, 37, 785-789; d) P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, J. Phys. Chem. 1994, 98, 11623-11627.

[8] P. J. Hay and W. R. Wadt, J. Chem. Phys. 1985, 82, 270-283.

[9] J. Tomasi, B. Mennucci and R. Cammi, Chem. Rev. 2005, 105, 2999-3094.

[10] a) M. D. Pluth, R. G. Bergman and K. N. Raymond, Science 2007, 316, 85-88; b) M. D. Pluth, R. G. Bergman and K. N. Raymond, J. Am. Chem. Soc. 2007, 129, 11459-11467.

[11] The structures of water clusters are still actively studied, e.g., C. T. Wolke, J. A. Fournier, L. C. Dzugan, M. R. Fagiani, T. T. Odbadrakh, H. Knorke, K. D. Jordan, A. B. McCoy, K. R. Asmis and M. A .Johnson, Science 2016, 354, 1131-1135.

[12] F.-G. Klarner, J. Panitzky, D. Preda and L. T. Scott, J. Mol. Model. 2000, 6, 318-327.

[13] D. A. Dougherty, Science 1996, 271, 163-168.

[14] S. E. Wheeler, Acc. Chem. Res. 2013, 46, 1029-1038.

[15] a) P. Schreiner, P. v. R. Schleyer and H. F. Schaefer, J. Org. Chem. 1997, 62, 4216-4228; b) P. A. Byrne, S. Kobayashi, E.-U. Wurthwein, J. Ammer and H. Mayr, J. Am. Chem Soc. 2017, 139, 1499-1511.

[16] Representative examples: a) L. Zu, M. Xu, M. W. Lodewyk, D. E. Cane, R. J. Peters and D. J. Tantillo, J. Am. Chem. Soc. 2012, 134, 11369-11371; b) A. J. Jackson, D. M. Hershey, T. Chesnut, M. Xu and R. J. Peters, Phytochem. 2014, 103, 13-21; c) H. Sato, K. Teramoto, Y. Masumoto, N. Tezuka, K. Sakai, S. Ueda, Y. Totsuka, T. Shinada, M. Nishiyama, C. Wang, T. Kuzuyama and M. Uchiyama, Sci. Rep. 2015, 18471; d) J. S. Dickschat, N. L. Brock, C. A. Citron and B. Tudzynski, ChemBioChem 2011, 12, 2088-2095; e) J. Rinkel, P. Rabe, P. Garbeva and J. S. Dickschat, Angew. Chem. Int. Ed. 2016, 55, 13593-13596.

[17] I. Abe, M. Rohmer and G. D. Prestwich, Chem. Rev.1993, 93, 2189-2206.

[18] a) Y. J. Hong and D. J. Tantillo, Org. Lett. 2015, 17, 5388-5391; b) Y. J. Hong and D. J. Tantillo, Chem. Sci. 2013, 4, 2512-2518.

Chapter 2.2

Chapter 3. High resolution modeling of class I terpene cyclases through integration of mechanistic information

3.1 Introduction

Terpenes and their modified derivatives, terpenoids, comprise one of the most numerous and diverse family of natural products. Currently there are more than 80,000 members in the greater family of terpenome[1]. Terpenes are isolated in nature as mammalian, plant, fungal, archaea and bacterial metabolites [2-6], and serve essential roles in ecology, such as protecting autotrophic organisms and communication[7, 8]. Terpenes are common components in human diets for flavor and fragrance, and play important roles in healthcare, notably anti-cancer drug paclitaxel [9, 10].

Despite the large chemical and structural diversity, terpenes are synthesized by terpene cyclases from only a handful simple acyclic precursors, known as C_{5n} isoprenoid diphosphates (n = 2, 3, 4, etc.). These precursors undergo multistep cyclization reactions via highly reactive carbocation intermediates to form complex products. These carbocation rearrangements and cyclization reactions are one of the most complex chemical reactions occurring in nature, makes terpene cyclases a mechanistically intriguing class of enzyme. Many terpene cyclases have high fidelity, generating unique product that has precise structure and stereochemistry. The complicated terpene products are formed in the largely non-polar active site that accommodates the carbocation intermediates. The active site of Class I terpene cyclase is in the middle of the α-helical bundle. Although for many sequences the percent identity can be as low 10%, the evolutionarily distinct α-helical fold is very conserved (Figure 3.1), which make comparative modeling a reasonable approach for generating terpene cyclase illuminate the structural-function relationship of terpene models.

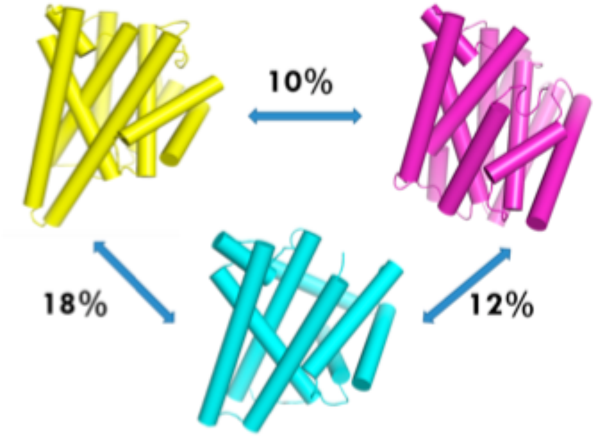


Figure 3.1 The conserved class I terpene cyclase fold. Percentage shows the sequence identity between each pair. Cylindrical Helices shows α-helical fold of three class I terpene cyclases. Yellow: Epi-isozizaene synthase (PDB code: 3KB9). Magenta: Limonene synthase (PDB code: 2ONG). Cyan: Aristolochene synthase (PDB code: 2OA6).

The nature of terpene cyclases produce a variety of complex products with a greasy active site arouse interests to study the structural function relationship of ter-penecyclases. Many computational studies [13-17] were carried out previously to try to uncover the productive binding mode for terpene cyclases. However, most of these simulations have been performed on the enzymes which have been successfully crystalized. Unfortunately, for the overwhelming majority of enzymes in this family there is no crystal structure. Currently there are more than 15,000 non-redundant sequences [18] of terpene cyclases, there're only about 30 non-redundant crystal structures of terpene cyclases available [19] , a 500 fold de-crease. In order to fill in this gap, we proposed a comparative modeling approach combined with iterative loop modeling/docking method to generate high resolution models of class I terpene cyclase. The unique approach used by our lab is a strategy to incorporate structural information from experimental data into the comparative modeling, with the goal of generating high resolution models of class I terpene cyclases. These models can then be the used to predict the productive binding mode of the substrate, perform higher level calculations, such as Quantum Mechanical Molecular Mechanical molecular dynamics (QM/MM MD) simulations and, potentially, the rational engineering of terpene cyclases.

3.2 Results and Discussion

Class I terpene cyclase initiate the carbocation reaction via the Mg^{2+} (or Mn^{2+}) dependent diphosphate group (DPP) ionization, resulting an allylic carbocation [20] . Three Mg^{2+} (or Mn^{2+}) ions are coordinated by the highly conserved metal-binding motifs on helices D, H and by the diphosphate group (Figure 3.2). The first motif on helix D, DDXXD (X: all amino acid is possible, bolded residues coordinate with metals) coordinates two of the ions (see blue helix, Figure 3.2), while the second motif on H helix, (N,D)D(L,I,V)X(S,T)XXXE, coordinates the last metal ion(see salmon helix, Figure 3.2). The arginine colored in grey in Figure 3.2 also interacts with the DPP to assist the metal ions in triggering ionization reaction.

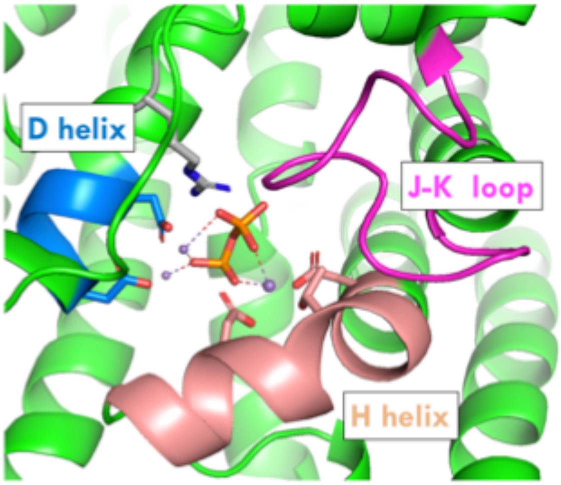


Figure 3.2. Active site of limonene synthase (PDB:2ONG).

Conformational changes in H helix along with the J-K loop are triggered by the metal ion-DPP complexation to enclose the active site from solvent, thus the presence and the spatial position of these regions are very important for catalysis [21]. These regions are highly mobile and consequently, many crystal structures either lacking electron density in these two regions or they are not in the catalytically relevant position. To address this issue and make it possible to predict catalytically relevant pose for unknown structures, here we applied a "catalytic constraint" to the comparative modeling method to help identify the catalytic relevant positions for H helix and J-K loop. The "catalytic constraint" is a set of distance values of six catalytic relevant residues, taken directly from crystal structure data. Two types of distances are measured. First, the distances between metal ion and the binding residues side chain were measured (blue dot lines in Figure 3.3, left). Second, for each two of the residues, distances of Cα - Cβ are measured (orange dot lines in Figure 3.3, right). Two types of constraints adding up 65 measurements for each crystal structure.

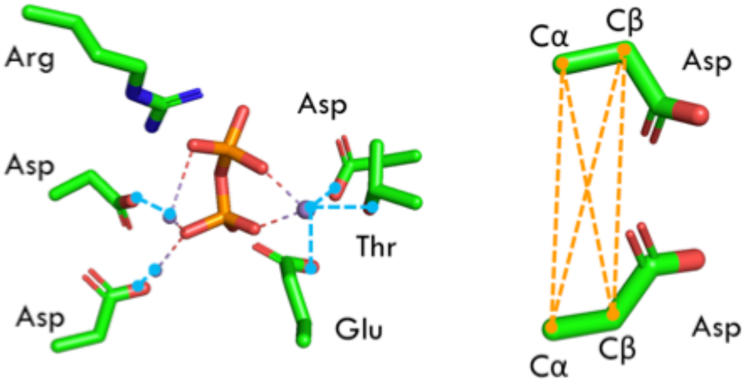


Figure 3.3. Catalytic constraints. Blue and orange dot line indicate the distances that are measured in crystal structures. Left: metal-side chain distance. Right: distances of Cα, Cβ in residues, here use two Asp as an example.

In order to avoid bias towards a specific structure, these 65 measurements were made on all known class 1 terpene cyclases in a closed and theoretically active state. In the protein data bank(PDB) there are 111 crystal structures of class I terpene cyclase available, among all these crystal structures, the ones that are not complete or do not bind with the DPP-metal ion complex in active site were filtered out, since the presence of DPP complex would ensure the catalytic residues be in a productive configuration. After examining all 111 structures, 29 structures were left that meet the requirement. The measurement of the catalytic residues were carried out as outlined above (Figure 3.3). Among the 29 structures, many of them are redundant structures, i.e. the same protein with different ligands etc. There're seven non-redundant types of class I terpene synthase. For redundant structures within the same type of terpene synthase, the mean value of all the measurements are calculated and represent that type of terpene synthase. One example set of Cα, Cβ measurements between the two Asp on D helix is summarized in Figure 3.3. In most cases, the majority of distance values are within 1 Å, regardless what type of class I terpene cyclase they are. This indicates that the catalytic residues are highly conserved in spatial position among class I terpene cyclases, suggests developing a standard modeling method for all class I terpene cyclases is feasible.

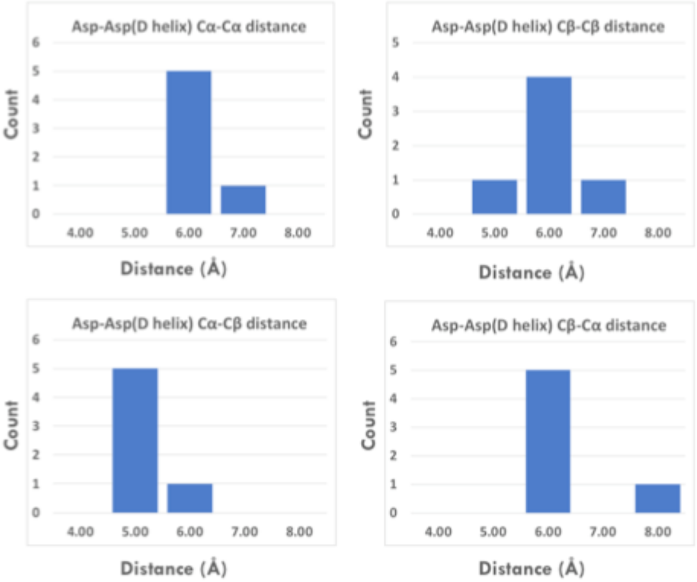


Figure 3.4. Cα-Cα, Cβ-Cβ, Cα-Cβ, Cβ-Cα distance between 2 Asp residues. X axis: distances measured from all types of terpene synthase; Y axis: out of all types, how many falls in certain distance range. (Total count is six instead of seven: one type doesn't have any residues that are equivalent to 2nd catalytic Asp residue on D helix)

To conduct comparative modeling experiment, templates were identified. Comparative modeling constructs models of the target protein from its amino acid sequence and 3D template structures. In this methodology, instead of searching for templates by detecting homologous structures of a target sequence, the templates are manually identified based on if the structure is in a catalytically relevant configuration. As mentioned in the previous paragraph, seven structures are non-redundant structures, compose the template pool for comparative modeling. The PDB codes of seven templates are 2ONG, 1JFG, 1N21, 2OA6, 3KB9, 4OKZ and 5IKA.Templates are identified in this way because 1) this method is intended to set up a protocol for all class I terpene cyclases, not just one sequence. By doing so, the same template pool can be reused for all target class I terpene cyclase sequences. 2) Most crystal structures of class I terpene cyclases are either incomplete or not in catalytically relevant configuration. Manually identify the good quality templates would prevent the modeling from misleading by non-relevant structures.

To validate our methodology, one structure in the templates, (4S)-limonene synthase (PDB code: 2ONG), was pulled out of the template pool and was used to retrospectively predict its known structure as a benchmark study. Figure 3.5 shows the proposed mechanism for (4S)-limonene synthase, a class I terpene cyclase that has the simplest cyclization mechanism.

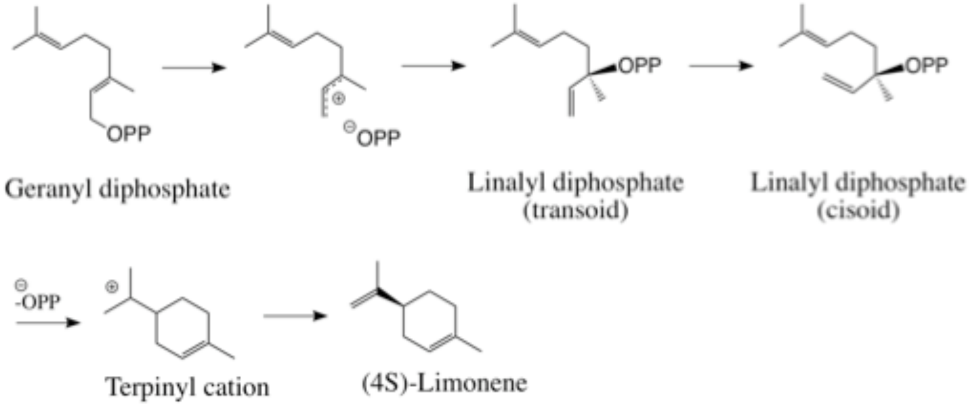


Figure 3.5. Proposed mechanism for isomerization and cyclization of GPP to (4S)-limonene 22

After setting up catalytic constraints and templates, a multiple sequence alignment was generated from the sequences of templates and the query sequence to correlate sequence position to structure using Promals3D 23 . The comparative modeling is then conducted using Ros-setaCM protocol 24 to generate three-dimensional models. To fill in unaligned regions during modeling, structural fragment sets were generated using standard methods 25 . During multi-template fragment based modeling through RosettaCM, the catalytic constraints were applied. If the distance of a certain pair of catalytic residues is within the min-max range of the distances measured previously, this distance is accepted with no penalty. Otherwise a penalty would be assigned to the pose. The Mg and PPi were treated as a single ligand complex and it was incorporated as the homology models were constructed. 5000 models were generated by this procedure.

To compare whether adding the catalytic constraints would help generate a high-resolution model in a catalytically relevant pose, comparative modeling was done both with and without the constraints. The ten lowest (best) score models were shown in Figure 3.6 (magenta), to compare with the known crystal structure(green).

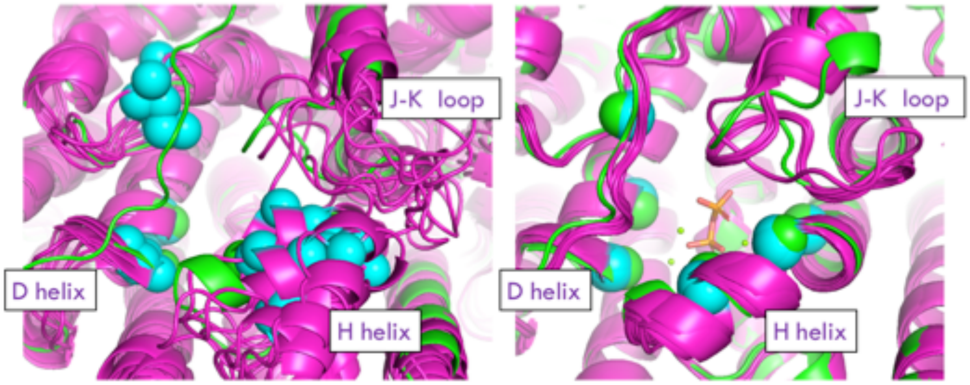


Figure 3.6. Ten lowest score models from comparative modeling. Left: without catalytic constrains and default templates; Right: with catalytic constrains and selected templates. Magenta: models of 4S-limonene synthase. Green: crystal structure of 4S-limonene synthase. Cyan spheres: Cα of catalytic residues of models; Green spheres: Cα of catalytic residues of crystal structure.

Models generated with catalytic constraints converged better to the catalytically relevant pose, indicating incorporating the experimental data help sampling chemically relevant spatial position and significantly reduce the number of samplings needed to generate good quality models. Comparing the H helix region, if no catalytic constraints were applied and generated by default templates, many non-productive poses were predicted, but with constraints and selected templates, all low 10 models predicted were in a relevant pose.

To get the high-resolution models of terpene cyclase, three portions of the active site should be all in catalytically relevant position. Here we demonstrated after incorporating the experimental structural data, the D,H helices of active site are in a relatively productive position. However, the pose of J-K loop still needs refinement. In order to address this issue, we will carry out an iterative loop modeling/docking experiment to mimic the natural “induce-fit” process for enzymes. After establishing this methodology, all templates would be tested as benchmark to validate this method.

To refine the modeling accuracy of the J-K loop of the plant terpene synthases, residue conservation was evaluated in the loop region. On one specific site, aromatic residues are observed frequently, as shown in figure 3.7. During the comparative modeling process, the Cα positions are constrained for this site, to provide a pivot point for the loop sampling and improve the accuracy.

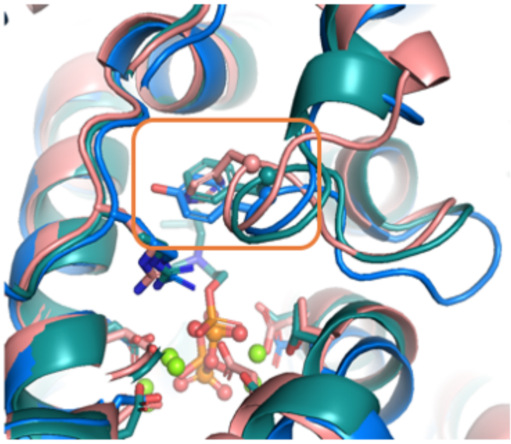
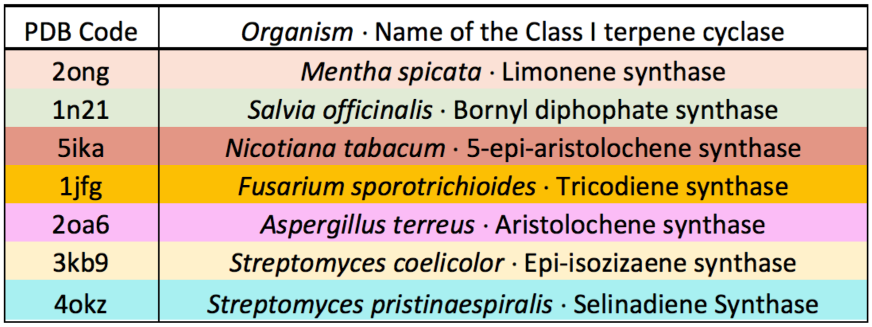


Figure 3.7. Residues on the J-K loop is conserved among homologs.

To validate the method, more class I terpene synthase structures are modeled and compared to its crystal structure. Figure 3.8 shows a list of crystal structures that has the DPP-metal complex bind and the D, H helices and J-K loop (in plant TS) in a catalytic relevant configuration. After trial studies, we discovered to achieve the best accuracy, it's beneficial to model plant terpene synthases and bacterial or fungal terpene synthases separately. The benchmark result is shown in figure 3.9. Along with limonene synthase, two other plant terpene synthases have been tested: bornyl diphosphate synthase and 5-epi-aristolochene synthase. The magenta structures are the 10 top predicted models, while the green structures are the crystal structures for comparison. The helix and loop region of the active site converged better with the constrained describe above.



PDB Code	*Organism* · Name of the Class I terpene cyclase
2ong	*Mentha spicata* · Limonene synthase
1n21	*Salvia officinalis* · Bornyl diphophate synthase
5ika	*Nicotiana tabacum* · 5-epi-aristolochene synthase
1jfg	*Fusarium sporotrichioides* · Tricodiene synthase
2oa6	*Aspergillus terreus* · Aristolochene synthase
3kb9	*Streptomyces coelicolor* · Epi-isozizaene synthase
4okz	*Streptomyces pristinaespiralis* · Selinadiene Synthase

Figure 3.8. List of crystal structures of class I terpene synthase, with catalytic relevant configuration and diphosphate-metal ion ligand bound.

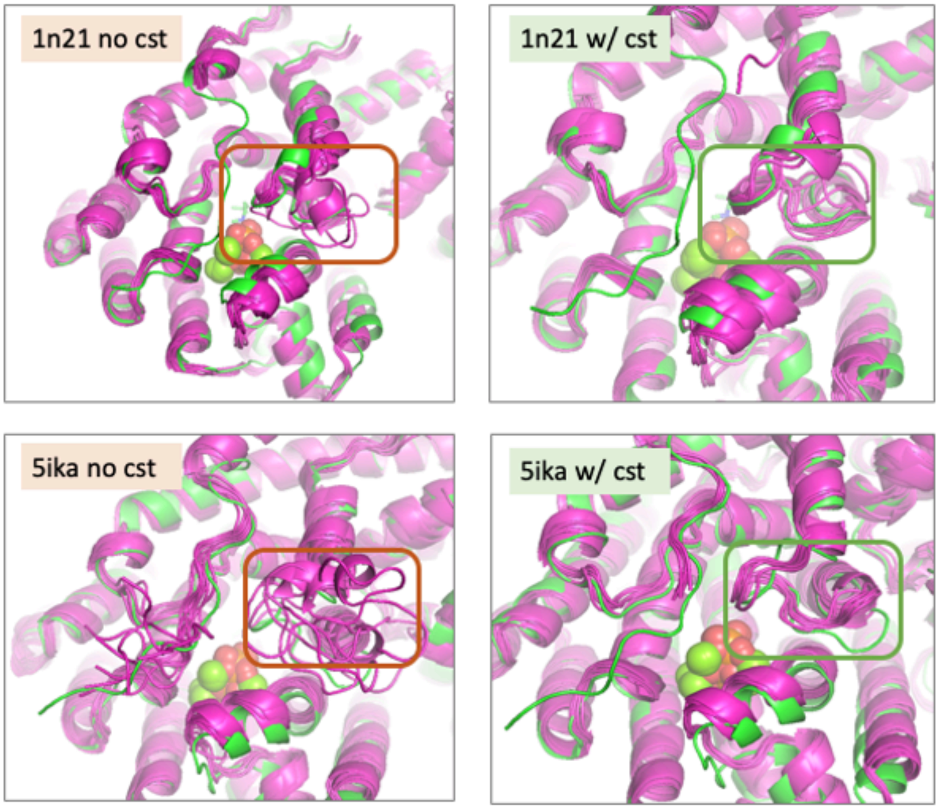


Figure 3.9. Benchmark results. Comparison between Models(generated by RosettaCM, with awareness of terpene mechanistic information, shown in magenta) and crystal structures(green).

To test the ability of predicting terpene synthases without crystal structures or without catalytic relevant structures, taxadiene synthase structures were modeled with both RosettaCM and

AlphaFold 2 [26] for comparison. As shown in figure 3.9, magenta models are modeled with RosettaCM, white structures are modeled with AlphaFold 2, and crystal structure is shown in green. While the J-K loop of the crystal structure is in the “open” conformation, both RosettaCM and AF2 is able to model the “close” conformation, the overall structure predicted are very similar with the two methods(RMSD = ~1.8 for different structure pairs generated by two different methods). The prediction of the J-K loop region by the two methods agrees with each other. The H helix varies a little between the two methods.

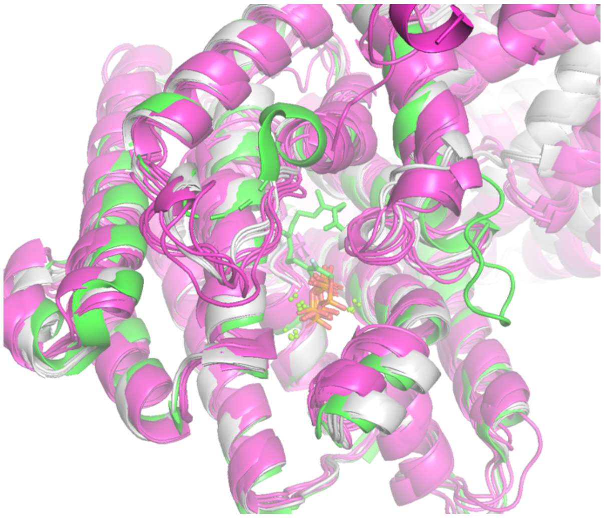


Figure 3.9. Modeling of taxadiene synthase. Top RosettaCM models are shown in magenta, top AlphaFold 2 models are shown in white, crystal structure is shown in green.

3.3 Conclusion

The goal of this project is to generate models that could potentially be used as a starting point to predict productive binding modes of the substrates and intermediates of terpene synthases. Models predicted by this method can be used as the input for molecular dynamics calculations potentially. Previous studies in our lab already shown success by using crystal structures as the starting point [16, 17]. We will apply the predicted models to see if it’s capable reproduce experimental data and potentially set stage for rational design of this class of enzyme. Further studies using machine learning and other statistical methods may set stage for studying terpene synthase sequence – structure – function relationship.

3.4 Reference

1. Philmus, B., Review of Natural Products Desk ReferenceNatural Products Desk Reference. Journal of Natural Products, 2016. 79(11): p. 2982-2982.

2. Yamada, Y., et al., Terpene synthases are widely distributed in bacteria. Proc Natl Acad Sci U S A, 2015. 112(3): p. 857-62.

3. Chen, F., et al., The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J, 2011. 66(1): p. 212-29.

4. Ishida, T., et al., Biotransformation of terpenoids by mammals,microorganisms, and plant. Chemistry and Biodiversity, 2005.2: p. 569-588

5. De Rosa, M.; Gambacorta, A.; Nicolaus, B. Regularity of Isoprenoid Biosynthesis in the Ether Lipids of Archaebacteria. Phytochemistry 1980.19, p. 791−793.

6. Yamada, Y., et al., Terpene Synthases are Widely Distributed in Bacteria. Proc. Natl. Acad. Sci. U. S. A. 2015, 112, 857−862.

7. Pichersky, E., et al. Biosynthesis of Plant Volatiles: Nature's Diversity and Ingenuity. Science, 2006. 311: p. 808-811

8. Schimek, C. and J. Wostemeyer, Carotene derivatives in sexual communication of zygomycete fungi. Phytochemistry, 2009. 70(15- 16): p. 1867-75.

9. Cho, K.S., et al., Terpenes from Forests and Human Health. Toxicol Res, 2017. 33(2): p. 97-

106.

10. Bicas, J. et al., Bio-oxidation of Terpenes: An approach for the Flavor Industry. Chem. Rev. 2009.109: p. 4518-4531

11. Lesburg, C. et al., Crystal Structure of Pentalenene Synthase: Mechanistic Insights on Terpenoid Cyclization Reactions in Biology. Science, 1997.277: p. 1820-1824

12. Koksal M. et al., Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. Nature. 201. 469: p.116-122

13. Yao, J., F. Chen, and H. Guo, QM/MM free energy simulations of the reaction catalysed by (4S)-limonene synthase involving linalyl diphosphate (LPP) substrate. Molecular Simulation, 2018: p. 1-10.

14.Wu, R., et al., Catalytic Promiscuity of Non-native FPP Substrate in TEAS enzyme: Nonnegligible Flexibility of the Carbocation Intermediate. Physical Chemistry Chemical Physics, 2018.

15. Freud, Y., T. Ansbacher, and D.T. Major, Catalytic Control in the Facile Proton Transfer in Taxadiene Synthase. ACS Catalysis, 2017. 7(11): p. 7653-7657.

16. O’Brien, T.E., et al., Predicting Productive Binding Modes for Substrates and Carbocation Intermediates in Terpene Synthases—Bornyl Diphosphate Synthase As a Representative Case. ACS Catalysis, 2018. 8(4): p. 3322-3330.

17.O'Brien, T.E., Mechanistically informed predictions of binding modes for carbocation intermediates of a sesquiterpene synthase reaction. Chemical Science, 2016.

18. Apweiler, R., et al., UniProt: the Universal Protein knowledgebase. Nucleic Acids Res, 2004. 32(Database issue): p. D115-9.

19. Berman, H. et al., The Protein Data Bank. Nucleic Acids Research, 2000.28 (1): p 235-242

20. Christianson, D., Structural Biology and Chemistry of the Terpenoid Cyclases. Chem. Rev. 2006.106: p: 3412-3442

21. Christianson, D.W., Structural and Chemical Biology of Terpenoid Cyclases. Chem Rev, 2017. 117(17): p. 11570-11648.

22. Hong, Y., et al., Quantum chemical dissection of the classic terpinyl/pinyl/bornyl/camphyl cation conundrum—the role of pyrophosphate in manipulating pathways to monoterpenes. Org. Biomol.Chem, 2010. 8: p 4589-4600

23. Pei, J.; Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. Methods Mol. Biol. 2014. 1079:p. 263-71.

24.Song, Y., et al., High-resolution comparative modeling with RosettaCM. Structure, 2013. 21(10): p. 1735-42.

25. Gront, D., et al., Generalized fragment picking in Rosetta: design, protocols and applications. PLoS One, 2011. 6(8): p. e23294.

26. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021)