U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
2021

U.PORTO

Enzymatic Technologies for Cleaning Petroleum with Low CO_2 Emissions

Filipa Pinto de Barros Miranda

FC

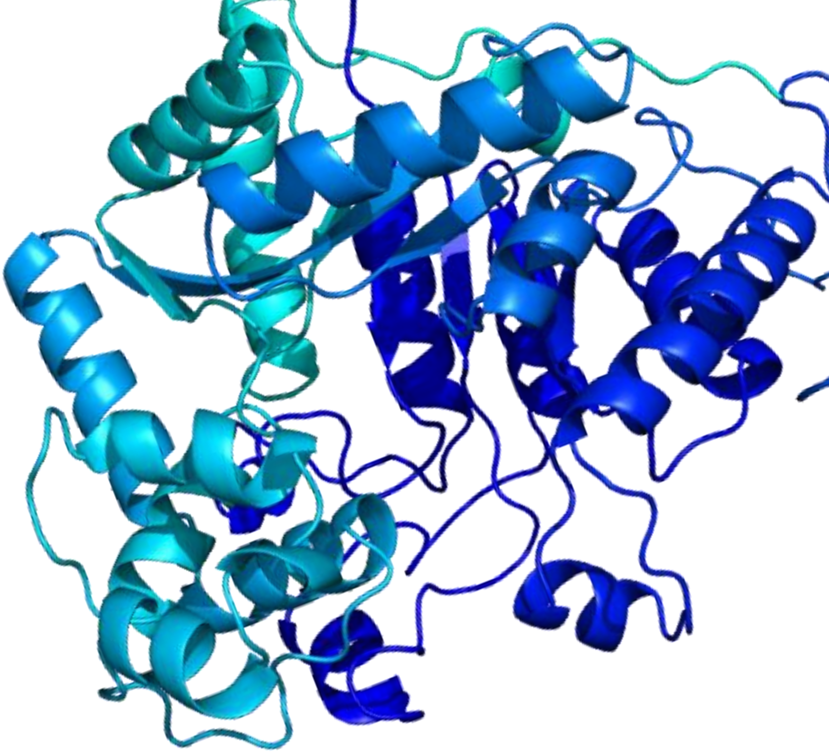
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Enzymatic Technologies for Cleaning Petroleum with Low CO_2 Emissions

Filipa Pinto de Barros Miranda

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Aplicações em Biotecnologia e Biologia Sintética

2021



Enzymatic Technologies for Cleaning Petroleum with Low CO_2 Emissions

Filipa Pinto de Barros Miranda
Mestrado em Aplicações em Biotecnologia e Biologia Sintética
Departamento de Química e Bioquímica; Departamento de Biologia
2021

Orientador
Doutor Pedro Alexandrino Fernandes, Professor Associado,
Faculdade de Ciências da Universidade do Porto

Coorientador
Doutora Maria João Ribeiro Nunes Ramos, Professora Catedrática,
Faculdade de Ciências da Universidade do Porto

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

MSc

Agradecimentos

Aos Professores, Doutor Pedro Alexandrino Fernandes e Doutora Maria João Ramos, agradeço pelas oportunidades que me disponibilizaram. A vossa boa disposição, entusiasmo, simpatia e gosto pela área são contagiantes, contribuindo assim para a minha motivação ao longo deste ano.

Ao Rui, proclamo a minha extrema gratidão por toda a tua incansável dedicação. Pela tua disponibilidade e paciência e por teres partilhado comigo o teu vasto conhecimento. Ensinaste-me a maioria do que sei na área de química computacional e como tal, sem ti, esta tese não seria possível. Aprecio também teres-me encorajado a tentar sempre alcançar mais e melhor através da tua admirável exigência e rigor.

Aos meus colegas de laboratório, com especial destaque à Carola, obrigada pela companhia no dia a dia, simpatia e por partilhares um pouquinho da tua vida comigo. Espero que um dia nos encontremos outra vez, talvez no Equador.

Aos meus pais, tenho que agradecer por tudo. Por me terem privilegiado com tudo o que necessitasse e mais. Por me terem educado, respeitado e suportado, a mim e às minhas escolhas, e por me darem liberdade para ser quem eu quero ser. Nunca vos conseguirei agradecer ou retribuir o suficiente, mas espero algum dia conseguir mostrar-vos o quanto vos aprecio. Sem o vosso constante apoio, preocupação e carinho, esta conquista não seria possível.

Ao meu namorado/melhor amigo/*lab partner*/futuro colega de trabalho/futuro *roommate*/futuro *housemate*/*pain in the ass*/befito/pedacinho de céu/momo/tudo, Bruno Miguel Viveiros Araújo, agradeço por todo o teu amor, constante apoio e por seres *overall* incrível. É um privilégio poder te amar e ser amada por ti. Estou absolutamente convicta que continuaremos a partilhar a nossa deficiência, momentos especiais, danças na rua e miminhos por muitos mais anos. Mal posso esperar para as nossas próximas aventuras e para te ver *taking over the world*!

Abstract

The petrochemical industry negatively affects the environment and health as it releases into the atmosphere harmful substances such as carbon dioxide and sulfur oxides. Due to its detrimental impact, regulations demand oil with low-sulfur levels. However, current desulfurization techniques cannot effectively remove heteroaromatic sulfur compounds, which are abundant in petroleum. Biodesulfurization *via* the 4S pathway of *Rhodococcus erythropolis* IGTS8 enzymes (DszA, DszB, DszC and DszD) is possibly a solution. However, its industrial applicability is dependent on its pathway optimization, through enzyme engineering. To eventually enable such optimization, the unknown structure and mechanism of reaction of DszA needs to be unveiled. Such findings are the aim of this thesis. The DszA structure is attained through homology modeling using as template BdsA, a DszA homolog with a sequence identity of 79%. Such renders a quality model structure in which, upon its transformation into a dimer and energy minimization, the cofactor – hydroperoxyflavin intermediate (C4A) – and the substrate – dibenzothiophene sulfone ($DBTO_2$) – are added to the hydrophobic active center to attain the DszA:C4A:$DBTO_2$ system. C4A is modeled into DszA based on the position of the FMN cofactor seen in BdsA, and $DBTO_2$ is docked through protein:ligand docking. The most promising substrate pose is selected, and the system undergoes molecular dynamics simulations, from which a representative cluster structure is attained. Such structure is then truncated at 12 Å around the reactive system giving a system of 4299 atoms. The reactive system has 137 atoms and is treated at the quantum mechanics theory level, whilst everything around is treated with molecular mechanics. The outer layer is frozen to prevent system expansion. For the first step of the reaction, two main mechanistic hypotheses are attempted in this structure (and in others derived from it): the formation of a peroxyhemiacetal intermediate as proposed by Adak and Begley and the hydroperoxyl group transfer from C4A to $DBTO_2$; however, in all the mechanistic attempts made in this thesis, none renders favorable results. Thus, the results in this thesis suggest that current substrate-binding modes and mechanistic hypotheses for DszA based on the formation of a C4A intermediate do not seem likely to occur. Consequently, this reinforces that new initial structures/hypotheses, such as the existence of a flavin-N5-peroxide as cofactor (as proposed by Matthews et al.) should be the future of mechanistic studies for DszA.

Keywords: Biodesulfurization, 4S pathway, DszA, *Rhodococcus erythropolis,* desulfurization, sulfur, crude oil

Resumo

A indústria petroquímica afeta negativamente o ambiente e a saúde ao libertar substâncias nocivas para a atmosfera, como dióxido de carbono e óxidos de enxofre. Devido ao seu impacto nefasto, normas atuais exigem petróleo com baixos níveis de enxofre. Contudo, as atuais técnicas de dessulfurização não são capazes de remover eficazmente os compostos héteroaromáticos de enxofre, abundantes no petróleo. A biodessulfurização pela via 4S das enzimas de *Rhodococcus erythropolis* IGTS8 (DszA, DszB, DszC e DszD) é possivelmente uma solução. No entanto, a sua aplicação industrial depende da otimização desta via através de engenharia enzimática. Para eventualmente permitir tal otimização, a estrutura e o mecanismo de reação da DszA carecem de ser revelados. Tais descobertas são o objetivo da presente tese. A estrutura da DszA é obtida através de modelação por homologia, utilizando como modelo a BdsA, uma enzima homóloga com a qual partilha 79% de identidade sequencial. Deste modo, obtém-se um modelo de qualidade em que, após a sua transformação em dímero e minimização de energia, o cofator – intermediário de hidroperoxiflavina (C4A) – e o substrato – dibenzotiofeno sulfona ($DBTO_2$) – são adicionados no centro ativo hidrofóbico, obtendo-se o sistema DszA:C4A:$DBTO_2$. O C4A é modelado na DszA com base na posição do cofator da BdsA, e o $DBTO_2$ é acoplado através de *docking* molecular. A pose de substrato mais promissora é selecionada, e o sistema é submetido a simulações de dinâmica molecular, a partir das quais é obtida uma estrutura representativa. Esta estrutura é truncada a 12 Å em torno do sistema reativo, originando um sistema com 4299 átomos. Este sistema reativo tem 137 átomos e é tratado com mecânica quântica, enquanto tudo à sua volta é tratado com mecânica molecular. A camada exterior é congelada para evitar a expansão do sistema. Para o primeiro passo da reação, duas principais hipóteses mecanísticas são estudadas nesta estrutura (e noutras derivadas dela): a formação de um intermediário peroxi-hemiacetal, proposto por Adak e Begley, e a transferência do grupo hidroperoxilo do C4A para o $DBTO_2$; No entanto, em todas as tentativas efetuadas nesta tese, nenhuma produz resultados favoráveis. Desta forma, os resultados desta tese sugerem que o modo de ligação do substrato e atuais hipóteses mecanísticas referentes à DszA não aparentam ser as mais plausíveis. Consequentemente, tal reforça que novas estruturas/hipóteses, nomeadamente a existência da flavina-N5-peróxido como cofator (conforme proposto por Matthews *et al.*) poderão constituir hipóteses futuras para o estudo mecanístico da DszA.

Palavras-chave: Biodessulfurização, via 4S, DszA, *Rhodococcus erythropolis*, dessulfurização, enxofre, petróleo

Table of contents

List of tables

List of figures

List of abbreviations

2-HBP: 2-hydroxybiphenyl

4-MDBT: 4-methyldibenzothiophene

ASP: Astex Statistical Potential

BDS: biodesulfurization

BiCh: bicyclohexyl

$C^{4a}OOH$: hydroperoxyflavin-intermediate

CHB: cyclohexylbenzene

ChemPLP: Piecewise Linear Potential

DBT: dibenzothiophene

DBTO: DBT sulfoxide

$DBTO_2$: DBT sulfone

DFT: density functional theory

DMDBT: dimethyldibenzothiophene

DMF: dimethylformamide

DMSO: dimethyl sulfoxide

DszA: DBT sulfone monooxygenase

DszB: HBPS desulfinase

DszC: DBT monooxygenase

DszD: flavin mononucleotide oxidoreductase

ESP: electrostatic surface potential

FMN: flavin mononucleotide

$FMNH_2$: reduced flavin mononucleotide

GA: genetic algorithm

GAFF: general AMBER force field

GGA: generalized gradient approximation

GMQE: Global Model Quality Estimation

GTO: Gaussian type orbitals

HBPS: 2’-hydroxybiphenyl 2-sulfinic acid

HDS: hydrodesulfurization

HF: Hartree-Fock

LDA: local density approximation

LSDA: local spin density approximation

MD: molecular dynamics

MM: molecular mechanics

NMR: nuclear magnetic resonance

NPT: isothermal-isobaric ensemble

NVT: canonical ensemble

ODS: oxidative desulfurization

PES: potential energy surface

QM: quantum mechanics

QM/MM: quantum mechanics/molecular mechanics

RESP: restrained electrostatic potential

SCF: self-consistent field

STO: Slater type orbitals

THDBT: tetrahydrodibenzothiophene

VDW: Van der Waals

1. Introduction

1.1. Fossil fuels

In a pre-industrial revolution world, energy was attained through the combustion of wood, peat, and crops or by using human and animal force. Then, a powerful source of energy, fossil fuel, was introduced by the first industrial revolution in the XIX century. Since then, this fossil energy has been used for a variety of activities, causing progress in areas such as technology, transportations, food production, and other industries, while also playing its part in changing the economic and social systems of the world.[1,2]

Fossil fuel can be defined as combustible organic matter that is formed in the Earth's crust, from fossils of plants and animals, under conditions of high pressure and temperature applied throughout hundreds of millions of years. These fuels are hydrocarbons that can be present in the form of oil, coal, or natural gas[3] and that, upon combustion, generate a large amount of energy. They are predominantly composed of carbon, hydrogen, oxygen, sulfur, and nitrogen. From which, carbon, hydrogen and sulfur are the most combustible elements.[4] Furthermore, upon combustion, several harmful byproducts are generated, namely carbon dioxide, sulfur oxides, nitric oxides, carbon monoxide[5], and particulate matter.[6] These emissions are detrimental to the environment and thus contribute to global warming[5], with the added burden of also being harmful to human health, with implications on the respiratory and cardiovascular systems, and with nefarious consequences on cancer development and pregnancy.[6,7]

With these detrimental side effects on both the environment and human health, it would be expected that, by now, most fossil fuels would be replaced for low-carbon sources of energy, such as renewable energies, but that is not the case. In fact, oil, gas and coal were still the top three sources of energy in 2019, with a joint share of 84.3% of the primary energy market. When observing these sources individually, oil accounts for 33.1% of the world's energy supply, making it the leading source of energy with more than 97 million barrels consumed in the world, per day.[89] As a result, in 2019, over 34 million tons of carbon dioxide were emitted into the atmosphere[8], along with the other aforementioned harmful emissions, namely sulfur and nitric oxides.[5] As far as projections on fossil fuel usage go, it is predicted that, until 2050, global natural gas consumption will increase by more than 40% and oil global usage will increase by more than 20%. On the other hand, coal consumption and production levels should not differ from the current ones until 2040. However, from 2040 until 2050 their growth is expected.[10]

Crude oil, a type of oil, can be employed for several applications, such as the generation of electricity and heat, for transportation, in the production of synthetic products and plastic, to power combustion engines and even to pave roads.[11] Upon refining and manufacturing processes, crude oil can be transformed into products such as gasoline, diesel oil, jet fuel, asphalt, oils for feedstocks, kerosene, heavy oil, naphtha, lubricating oils, waxes and still gas.[12] With such a broad implementation and diversity, it's reasonable to say that every country in the world needs to use crude oil.[11]

Although assessing the economic value of crude oil involves a full description of its specific features and components, there are two key characteristics upon which crude oil is generally classified: API gravity and sulfur content (the latter will be discussed in the following section). These subsequently determine how the oil is industrially processed, while simultaneously being a quick, but not exhaustive, measure for its economic value. API gravity is a density measure that categorizes oil into light crude oil, crudes with an API value that exceeds 38º, or heavy crude oil, if the API value is 22º or lower, thus emphasizing that the higher the API gravity, the lighter the oil is. If the density is between 22º and 38º, they can be classified as intermediate.[13] These differences in gravity are translated onto the level of complexity needed for the refining process, to obtain the desired products. In lighter crudes, because they include a higher percentage of small hydrocarbons, a simple distillation can be enough to attain high-value products such as gasoline, jet fuel and diesel. For heavier crude oils, due to the presence of greater proportions of bigger molecules, a simple distillation would only yield low-value products like asphalt and industrial fuels. To obtain more profitable products from heavy crudes, it is necessary to process them into smaller molecules, which entails expensive refinery methods, thus ensuring light crudes as the preferable cost-effective option for the refinery industry.[13,14]

1.2. Sulfur in fossil fuels

One of the main constituents of crude oil is sulfur,[13] with concentrations ranging from 0.03 to 7.89 wt%[15], in accordance with its origin and the geographical location of the refinery.[13] Oil that has a sulfur concentration beneath 0.5 wt% is said to be sweet and over 0.5 wt% is considered sour.[13] Sulfur is present in the oil on a variety of different compounds, as it can be found as inorganic substances, such as elemental sulfur, hydrogen sulfide(H_2S) and pyrite; or in organic form, as thiols, sulfides, disulfides, thiolanes, thiophenes, benzothiophenes, dibenzothiophenes (DBT) and benzonaphtothiophenes.[16] In particular, DBT and its derivates, 4-methyldibenzothiophene (4-MDBT), 4,6-dimethyldibenzothiophene (4,6-DMDBT) and 3,7-DMDBT and 2,8-DMDBT can account for 70% of the sulfur present in diesel.[2]

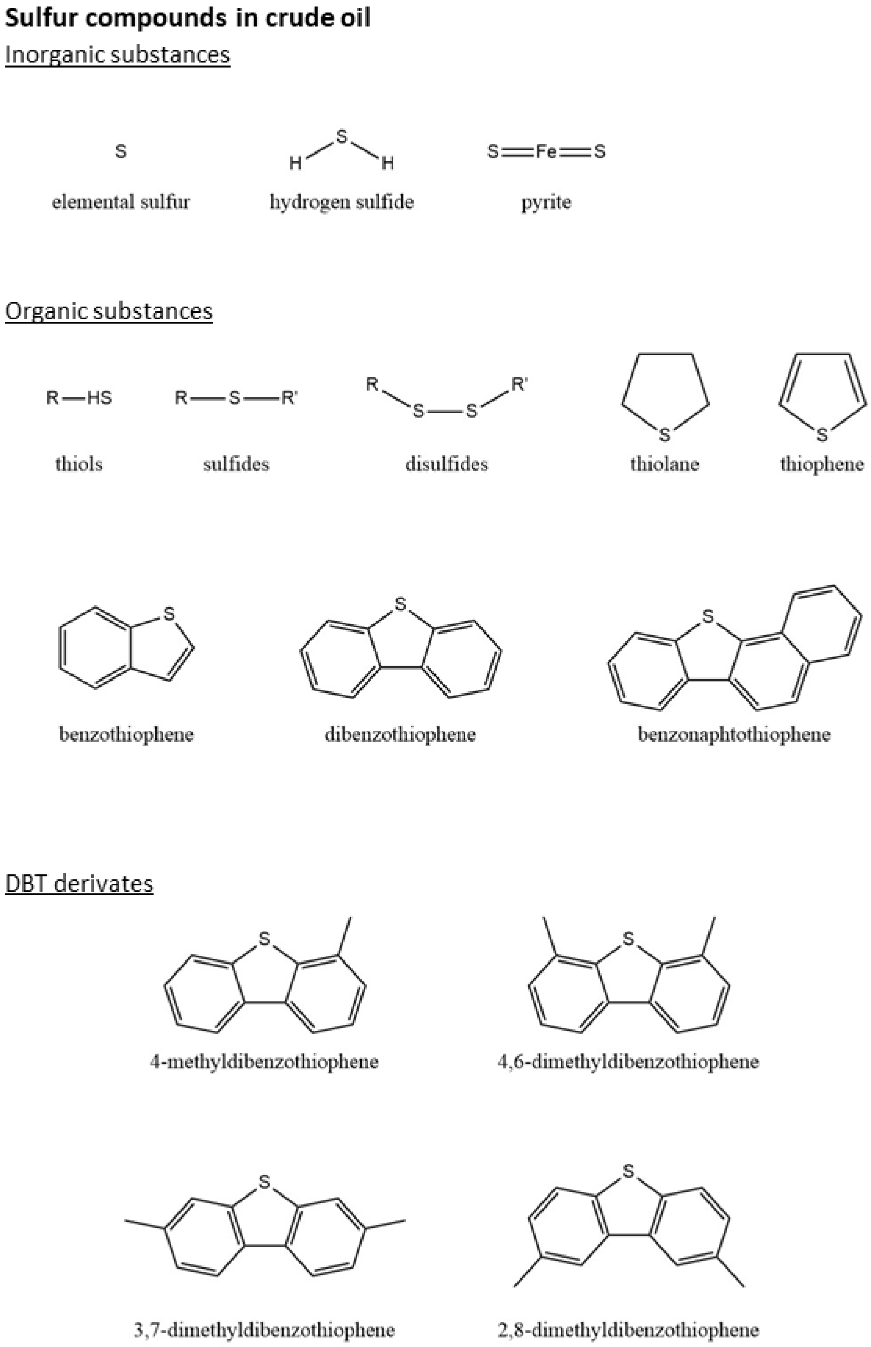


Figure 1. Main sulfur compounds that can be found in crude oil.

As sulfur is present in vehicles fuel, hazardous sulfur emissions are also observed upon combustion. When sulfur oxides are emitted, their combination with rainwater proves to be catastrophic as they react to form sulfurous acid, originating acid rain.[13] This corrosive rain has a pH of 5.2 or lower and when in contact with trees it damages them, making them more prone to pests, droughts, and extreme temperatures. Moreover, acid rain lowers the pH of surface waters, diminishing biodiversity, depletes soils of vital nutrients for plants, releases toxic dissolved aluminum and it corrodes limestone and marble surfaces and buildings, such as monuments. Since fossil fuel burning is the major responsible for acid rain formation, reducing sulfur emission is of great importance to mitigate these devastating effects on the ecological balance and material assets.[17]

Moreover, the sulfur content of crude oil considerably impacts the refining process and the quality of the product, whilst creating detrimental outcomes from their use. High sulfur levels in refinery streams lead to a greater emission of undesirable sulfur oxides onto the atmosphere. Additionally, it can also poison the catalysts that are participating in important chemical reactions, disabling them, or lead to corrosion of the refining equipment.[13] But the problems associated with the high sulfur content do not end there. The industry's visible preference for sweet oil, as it is the most profitable option, entails that the reserves of sweet crude oil are now almost completely depleted. In turn, it is now a necessity to use sour oil reservoirs. In fact, it is estimated that heavy, high sulfur content reservoirs account for 70% of the worlds' known oil reserves. The implications from this are vast, especially when considering that the biggest oil extractions are performed in areas such as the Middle East, Gulf of Mexico and South America, where the oil reservoirs are mostly sour. Hence, the aforementioned undesirable effects (higher sulfur emissions, catalyst poisoning, and deterioration of equipment) are exacerbated by the exploration of this type of reserves, calling for tighter legislation on sulfur emission levels that results in greater costs for the refining industries.

The first legislative regulations were intended to stipulate lead content in gasoline, but from the mid-90s, controlling the sulfur content in automotive fuel has become the main goal.[18] These requirements arose from the necessity to guarantee a higher product quality, which was until then declining steadily as sulfur content was increasing.[19] As the current state of the environment is considered a ticking time bomb, the legislation has been getting stricter, demanding increasingly lower sulfur concentrations in crude oil. In fact, the necessity to reduce hazardous emissions is stated in several documents, such as the Paris Agreement, the 2030 Climate and Energy Policy Framework, and the Kyoto protocol. Currently, the sulfur cap for the majority of transport fuels at the European Union is 10 ppm. This has been the case since 2009 for on-road vehicles and 2011 for non-road vehicles.[18,20] The same limit value is also used in China since 2018.[21] In the United States of America, since 2017, the limit is also 10 ppm when considering gasoline. However, this value is higher for diesel, at 15 ppm, as it has been since 2006 for on-road vehicles, 2010 for non-road vehicles and 2012 for locomotives and marine vehicles. [18,22] Nevertheless, this will not stop here, as the obvious trend is to favor ultra-low sulfur concentrations in diesel and gasoline, especially in developed countries. This is actually visible for example in the USA, where the limit in 2010 was 30 ppm, in the European Union, where 50 ppm fuels were once the norm, and in China, where not so long ago (2009) the maximum sulfur level allowed was 150 ppm. Furthermore, this trend is also seen in countries and areas such as India, Russia, South Africa, the Middle East and Latin America. This reinforces the notion that lowering the limits allowed for sulfur content in fuels is a worldwide goal.[18] Although these legislations amount to an enormous economic cost by refinery industries, these measures are of the utmost importance to guarantee the preservation of the environment and subsequently of the health of every living being.[13]

1.3. Desulfurization techniques

Since crude oil needs to be desulfurized, not only to meet legal requirements but also to attain a higher quality product and to protect the Earth and even ourselves, several techniques to do so are available. The most used method worldwide is hydrodesulfurization (HDS)[18]; however, others are available, namely oxidative desulfurization (ODS), adsorptive desulfurization, extractive desulfurization, ionic liquid extraction, alkylation-based desulfurization, chlorinolysis-based desulfurization, supercritical water-based desulfurization, and biodesulfurization (BDS). Every technique is different and hence, each has its own advantages and disadvantages.[16] Some of these are described in detail in the following sections.

1.3.1. Hydrodesulfurization

Hydrodesulfurization is a chemical process that removes sulfur from refined petroleum products. In refineries, it is employed on distillate streams that either come from the direct distillation of crude oil or from conversion units where petroleum is converted into gasoline and diesel.[23,24] HDS operates with catalysts, under high temperature and pressure, in the presence of hydrogen.[18] In fact, the industries' easy access to hydrogen from catalyst reformers was what initially stimulated the introduction of HDS in refineries. An important factor, considering that this technic uses hydrogen to reduce the different sulfur compounds into H_2S and attain organic compounds free from sulfur. The H_2S byproduct can later be transformed into elemental sulfur by a modified Claus process.[23] The standard used catalysts are $NiMo/Al_2O_3$ and $CoMo/Al_2O_3$; however, others may be relevant accordingly to the desired application. NiMo catalysts are used because they are more hydrogenating, and thus more suitable for fractions that require excessive hydrogenation. This makes them more efficient for refractory compounds such as 4,6-dimethyldibenzothiophene and other dibenzothiophenes. In flow reactors and other situations, where the hydrogen flow is unconstrained and the contact time is restricted, this type of catalyst is recommended. As for CoMo, they are preferred to desulfurize unsaturated hydrocarbon streams, such as the ones from fluid catalytic cracking, since these catalysts are better at hydrogenolysis. They are also more capable in batch reactors.

The harsh conditions that they operate in, from 200 to 425 °C and 1 to 18 MPa, allow the complete elimination of aliphatic sulfur compounds (thiols, sulfides and disulfides), as these are very reactive.[16] Despite effectively removing light organosulfur compounds, in heavier mixtures with aromatic compounds, such as dibenzothiophene and derivates, the sulfur elimination is not as successful, as they are not as reactive as non-aromatic substances.[18,25] The situation is aggravated as 70% of sulfur in diesel is present as DBT and its derivates.[2] Figure 2 depicts the two pathways currently employed in HDS: hydrogenolysis and hydrogenation. The first is the least hydrogen intensive and it removes sulfur without interfering with the aromatic rings. However, It does not happen with ease for the thiophene ring due to their resonance stabilization, which demands high energetic efforts from HDS.[16,23] Despite the hydrogenolysis pathway being faster, the hydrogenation pathway can be enhanced by increasing the H_2S/H_2 concentration, adding methyl groups to the 4 or 4 and 6 positions, or by having a higher concentration of catalysts available. After such optimizations, the hydrogenation pathway is of greater interest when it comes to increasing the desulfurization extent. In particular for organosulfur aromatic compounds, the hydrogenation pathway also decreases their aromatic content, improving their interaction with the surface of the catalyst and the consequent sulfur removal.[23] Nonetheless, this process is still not perfect and the selectivity for ring-opening is not assured,[16] as when the unpaired electrons from the sulfur resonate with the π electrons from the molecule, the energy for C-S and C-C bonds becomes alike. As a result, hydrogenation of C-C bonds is observed and saturated hydrocarbons that lower the product quality are attained. Hence, in HDS, further processing steps may be needed to get a higher-grade fuel. [23]

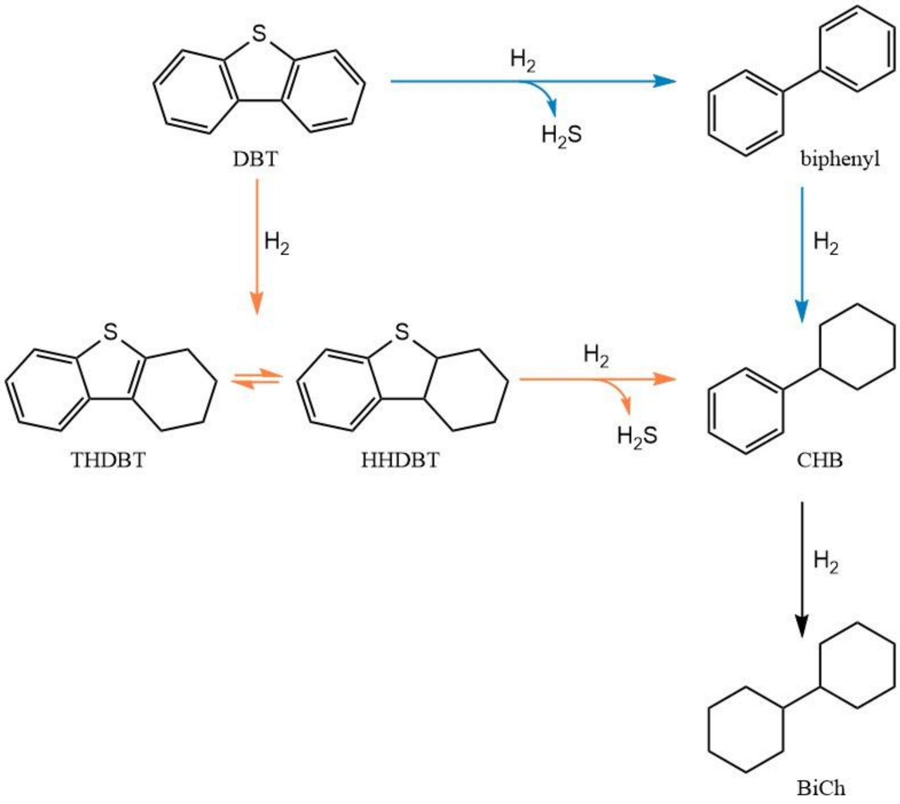


Figure 2. Reaction mechanism of HDS in DBT compounds. Arrows in blue represent the hydrogenolysis pathway, arrows in orange represent the hydrogenation pathway and arrows in black are common to both pathways. Hydrogenolysis: direct hydrogenolysis of the C-S bond of DBT forms biphenyl, that upon hydrogenation, gives cyclohexylbenzene (CHB). Hydrogenation: the sulfur compounds are hydrogenated prior to desulfurization. The aromatic rings are hydrogenated to the intermediary products 4H- or 6H-DBT, which will consequently be desulfurized. In this pathway, the primary desulfurization products from DBT are tetrahydrodibenzothiophene (THDBT) and/or hexahydrodibenzothiophene (HHDBT). These are very reactive and are subsequently transformed into CHB. In both pathways: the slow hydrogenation of CHB gives bicyclohexyl (BiCh) as a tertiary production.

Moreover, HDS is expensive as its costs can rise to around 32€ million to attain 20000 barrels per day, values that can skyrocket when considering the 97 million barrels that are consumed daily in the world.[9,18] Furthermore, the harsher conditions required to remove some sulfur compounds, lead to an octane rating loss and a hydrogen excess, with a consequent decrease in fuel calorific yield.[23] There are also other factors that undermine HDS, such as high metal content that deactivates the catalysts and promotes deposit formation, and the predilection for coking and fouling, which in turn also disables the catalysts. Additionally, the molecular size can hinder the access to the smaller catalysts pores and the steric protection of thiophenic sulfur also negatively impact the technique's effectiveness.[16] In this light, it is important that further research on HDS catalysis and process design is conducted, to achieve sulfur removal while ensuring the quality of the fuel.[23]

1.3.2. Oxidative desulfurization

Oxidative desulfurization is a two-step technology that, as it is deductible by its name, uses sulfur oxidation to remove it. The first part of this process is the addition of one or two oxygen atoms to the sulfur to form either sulfoxide or sulfone, respectively, without breaking any C-S bond. This change in their nature is advantageous for sulfur removal as it decreases the C-S bond strength. Additionally, sulfoxides and sulfones are more polar than the reduced compounds, which, consequently, increases their selectivity for the subsequent separation step.

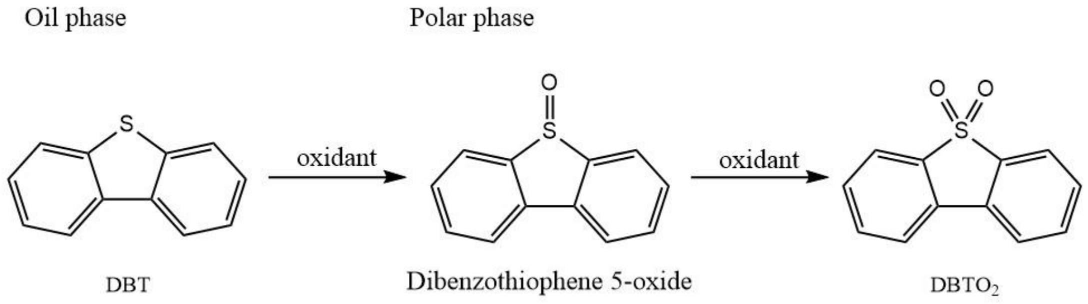


Figure 3.Oxidative step of Oxidative desulfurization.

After the oxidative step, their extraction, distillation, adsorption, or decomposition is carried out to separate the sulfur from the organic phase. A separation method that is commonly employed in the industry is solvent extraction. However, there is not only one approach to achieve any of these steps. Thus, several oxidants and solvents might be used.[16,18,23] Some oxidants used are hydrogen peroxide (H_2O_2), ozone, t-butyl hydroperoxide, t-butyl hypochlorite, nitrogen dioxides, peroxy organic acids (formic, acetic, propionic, performic, pertrifluoro acetic acids, etc.) among others. From this extensive list, H_2O_2 is the most popular choice, due to its environmental friendliness. Oxidation can also be attained with the help of radiation, ultrasounds or light, or electrochemical catalysis. As for the solvent-based extraction, solvents are chosen by their polarity, density, boiling point, freezing point or surface tension. The evaluation of these characteristics is important to assure that the solvents are suitable for the separation process and to ensure their recycling and reuse. This solvent salvaging involves its separation from the solvent mixture and the oxidized compounds by a simple distillation. Some common solvents that ODS uses are dimethyl sulfoxide (DMSO), dimethylformamide (DMF) and acetonitrile.[23]

One of the major benefits of this technique is its cost reduction when compared to other refining technologies, such as HDS. This arises from the fact that this process occurs at lower temperature and pressure conditions and it does not rely on molecular hydrogen, which is usually expensive. In fact, this method easily converts the refractory sulfur compounds, such as DBT and derivates, without necessarily causing a reduction in octane grade. Additionally, as it does not require molecular hydrogen, the refineries' location is not restricted to the proximity of a water pipeline. These reasons are the basis for the interest in this technology since the 60s.[18,23] However, not everything is a plus when it comes to ODS. A disadvantage of this method is that the oxidants are not always as selective as they ought to be, which causes unwanted reactions that jeopardize the product quality/quantity. Another downside is that an inadequate solvent selection may result in the loss of some desirable compounds, they may not extract the desired amount of sulfur or they can difficult their separation for reusage. An example of the latter is seen for two of the most utilized solvents, DMF and acetonitrile, as their boiling points are similar to the boiling point of sulfones, hindering their separation.[23]

1.3.3. Adsorptive desulfurization

Adsorptive desulfurization is a process that utilizes a solid sorbent that selectively adsorbs the sulfur compounds from the petroleum. This implies that the success of sulfur extraction with this method depends on sorbent selectivity to distinguish organosulfur from hydrocarbons. Additionally, the sorbents' adsorptive capacity, durability, porosity, surface area and regeneration are also important factors that should be accounted for when selecting a suitable sorbent.[16] Nonetheless, activated carbon, zeolites, amorphous alumina silicates and zinc oxide are some of the sorbent materials that have been used so far.[16,23] Adsorptive desulfurization can be approached in two ways: physical adsorption or reactive adsorption. The first is not very energetically demanding as there are no chemical alterations to achieve separation. Energy is basically only needed for sorbent regeneration. However, the amount of required energy is dependent on the adsorption strength. When it comes to reactive adsorption, a chemical reaction between the sulfur compounds and the sorbent surface occurs. As a result, the sulfides get attached to the sorbent surface[16] and the hydrocarbons go back to the process stream.[26] Similarly to HDS, this technique is done in the presence of molecular hydrogen and under high-temperature conditions.[16,26] Then, sorbent regeneration is attained either by the action of a desorbent or thermally. The removed sulfur is retrieved either in its elemental form or as H_2S or SOx.[16] However, as some characteristics of this technique are common to HDS, its effectiveness for deep desulfurization of heavy oil also does not entirely meet the expectations. [16,23] Moreover, it is a challenge to develop cost-effective adsorbents that possess not only a reasonable adsorption capacity but also a high degree of compound selectivity. Additionally, the adsorbents' regeneration process is not always as straightforward as it may seem, often requiring some additional steps.[23]

1.3.4. Extractive desulfurization

Extractive desulfurization is a liquid-liquid extraction method that uses two immiscible phases to separate organosulfur compounds from the oil. This separation is based on the solubility of the compounds that enables them to be retained in the solvent. Then, the solvent with the hydrocarbons undergoes a separation process from which the low-sulfur fuel can be further processed into high-value products.[16] Additionally, a distillation allows the separation of the sulfur compounds from the solvent so it can be used again in the desulfurization process.

This possibility for material reusage is one of the reasons why this extraction technique is desirable. But that is not all, as it is performed under mild temperature and pressure conditions and without the need of a catalyst and molecular hydrogen, which inevitably means a reduction in the cost of the refining process. Another advantage is that it has the capacity to selectively extract sulfur compounds without destroying any other constituents which means that there is not a decrease in the products' quality. This lack of a chemical transformation makes it a purely physical process.[16] But as always, some disadvantages also persist. For example, volatile solvents, such as polyalkylene glycol, imidazolidinone, pyrimidinone and dimethylsulfoxide, should not be used as their volatility results in solvent loss and hinders the regeneration of the extractant.[27] For this reason, the correct choice of the solvents is of great importance. This choice should take into account solvent characteristics such as efficiency, recyclability, reusability, sulfur compound solubility, solvent immiscibility and low viscosity. Another aspect to have in consideration when selecting the solvent is that it should have a boiling point that differs from the one of the organosulfur substances.[23,27] Such complexity and proneness to solvent loss, negatively affects the technique's cost-effectiveness. Nonetheless, solvents such as ethanol, acetone and polyethylene glycols were observed to attain a 50 to 90% desulfurization.[23] A greener solvent alternative is an ionic liquid, composed of organic cations (such as imidazolium, pyridiniuim, isoquinolonium, ammonium, phosphonium and sulfonium) and organic or inorganic anions. The ionic liquid properties depend on the cation/anion combination and their individual ionic characteristics. The characteristics that make the use of these solvents interesting are their non-volatility, stability, low viscosity, and a fast phase separation after mixing and extraction. Moreover, the possibility to adjust the cation/anion combination for specific applications, their high ionic conductivity and their capacity to dissolve both organic/inorganic compounds are a very welcome plus.[27] For these added advantages, ionic liquids containing imidazolium, pyridinium or quinolinium, with alkylsulfates, alkylphosphates, or halogen anions, have been employed to extract organosulfur compounds from oils.[23,27] However, in addition to the mentioned high procedure cost, this method does not seem to be viable to apply on a large industrial scale or even to perform desulfurization of heavy oils. Obviously, this substantially hinders the real application of this technique as a desulfurization method.[23]

1.3.5. Biodesulfurization

Biological desulfurization methods can be an alternative to the previously mentioned techniques. This option comes about due to the microorganisms' ability to consume several forms of organosulfur compounds that they consequently employ to sustain their own vital activities. In fact, bacterial cells have a sulfur dry weight of 0.5 to 1.0%. Moreover, their desulfurizing ability is performed by their own enzymes and metabolic pathways.[15,28] Additionally, this sulfur removal process is actually observed in the environment, as there are sulfur-rich soils and desulfurizing microorganisms widespread in nature. Such availability and desulfurization capacity turns them into an interesting subject of research in the area of desulfurization methods. However, their employment is only valuable if this method entails added benefits when compared to the previously mentioned desulfurization processes. Fortunately, that seems to be observed for biodesulfurization.

BDS is a low-cost alternative to financially demanding methods such as HDS.[6] In fact, a reduction of 15% on the cost of the operations was estimated when comparing BDS with HDS. This occurs not only as a result of an energy-saving process, as here sulfur is removed at moderate temperature and pressure conditions, but also because BDS does not require molecular hydrogen as a reactant.[6,16,18,28] Additionally, contrary to HDS, where non-selective ring opening is observed, BDS, due to the involvement of biocatalysts, is highly selective. This avoids the generation of the unwanted products seen in HDS and the need for additional purifying steps, ensuring the quality of the product and the conservation of its energetic value.[15,23,29] Yet, that is not all, as the use of BDS over HDS seems to reduce the emission of greenhouse gases, such as CO_2, into the atmosphere.[6,29] Moreover, the products attained through this method are not only environmentally friendly, but they are also safe, as they do not seem to have a negative impact on human health.[30]

Nonetheless, there are also some limitations associated with this technology. This is seen for example in its dependence on a microorganism that is able to extract a wide variety of sulfur compounds that constitute the crude oil.[6] Additionally, a poor organism choice can lead to the employment of species that not only uses the oil as a sulfur source but also as a carbon source. This is particularly important, as in that case, sulfur will be converted into water-soluble substances that hamper microbial growth and sulfur removal. Other limitations involve a slow sulfur metabolism when compared to the chemical reactions that occur in other methods. This low rate of desulfurization is most likely explained by the fact that these microorganisms require small amounts of sulfur for their growth and their enzymes expression is regulated through feedback inhibition mechanisms. Moreover, the desulfurization capacity of the natural biocatalysts available would need to undergo a 500-fold increase so BDS could be considered for commercial implementation. Yet, the list of disadvantages goes on as there is also the need for a large quantity of biomass (2.5 g of biomass/g of sulfur) and the microbial systems need to be kept alive under specific conditions that usually do not correspond to the ones seen in a refinery. In fact, when BDS was first employed, biocatalysts' longevity was initially of only 1 to 2 days, and since then extended to 8 to 16 days. Moreover, their desulfurization rate also depends on conditions, such as pH, temperature, and oxygen concentration. Additionally, separation of the cells from the oil also seems to present some occasional difficulties and the enzymes are also prone to substrate and product toxicity. Other disadvantages associated with this method are the lack of an optimal reactor design and the rate of transport of sulfur from the oil phase to the organisms' cell membrane.[18,23,29–31] Research on the latter, indicates that the microorganism can better access sulfur compounds when there is a reduced oil droplet size, which can be attained through the costly process of dispersing the oil in the aqueous phase. The use of surfactants to favor this contact between the enzyme and sulfur has also been suggested. However, there are questions about the toxicity of this surfactant to the organism.[18]

As observed, there is a somewhat extensive list of shortcomings and that is why, despite several technique improvements, BDS is often employed as a complementary method to HDS, instead of being used independently. Nevertheless, it can still be used on its own and it still has a lot of advantages that could classify BDS as a breakthrough method. To develop it further so it can be employed as an alternative to HDS, enzyme engineering has been useful to counteract BDS commercial obstacles such as the biocatalysts inhibition by the reaction products, the low rate of desulfurization and the enzymes inability to maintain their activity for prolonged periods of time.[32] Additionally, BDS has already been performed with several microorganisms. These include, but are not limited to, *Agrobacterium*, *Alcaligenes*, *Arthrobacter*, *Bacillus*, *Beijerinkia*, *Brevibacillus*, *Corynebacterium*, *Desulfobacterium*, *Desulfovibrio*, *Gordonia*, *Lysinibacillus*, *Microbacterium*, *Nocardia*, *Paenibacillus*, *Pantoea*, *Pseudomonas*, *Rhodococcus*, *Serratia* and *Shewanella*.[6] From all of these, the application of *Rhodococcus* has been in vogue. *Rhodococcus* are aerobic, non-sporulating, non-motile, Gram-positive bacteria, that have a history of being implemented as biocatalysts in different areas at an industrial scale, due to their robustness. *Rhodococcus* also possess a variety of plasmids and a thick cell envelope, which makes them resistant organisms. Additionally, their cell wall is composed of a peptidoglycan layer of long aliphatic mycolic acid chains that give hydrophobic characteristics to the cell. This hydrophobicity is actually an advantage, as it allows the bacteria to stick to the oil/water interface of the system being processed, which facilitates the capture of the sulfur compounds by the organism. Another advantage of this organism is that it not only can remove 90% of the sulfur compounds of the oil and survive for several months on the bioreactor, but it expresses enzymes that specifically cleave C-S bonds, thus conserving the calorific content of the hydrocarbon products.[15,16,33]

Another subject that has been under the spotlight concerns whether it is more suitable to use the organisms' whole cells or only its enzymes in BDS. Using only the biocatalysts, a minimal amount of water is required in the reactor for catalytic activity, which is an advantage to the refineries. Moreover, this allows a higher enzyme concentration and an easier separation. However, an enzyme-only approach would need a constant supply of cofactors and precursors, and the protection of the enzyme from the denaturing environment of the oil, which entails additional costs. To this adds the fact that BDS is a metabolic pathway that involves a cascade of enzyme-catalyzed reactions, which would be harder to conjugate in a single reactor. As for whole cells, their use may result in unwanted byproducts, since they possess several metabolic pathways, and metabolic repression and inhibition by the product may also occur. Despite these disadvantages, it seems that, currently, using the entire cell is almost mandatory for an efficient BDS. As a matter of fact, the whole-cell desulfurization rate seems to be 40 times higher than the one observed when applying the enzymes-only method.[34–36] Research on BDS has been focused mainly on the desulfurization of DBT and derivates, as these are the most abundant sulfur compounds on crude oil and HDS, the most employed desulfurization technique, fails to remove them.[29]

To remove these substances, several biodesulfurization pathways are available. These intend to transform DBT and derivates into more reactive compounds either through anaerobic or aerobic pathways.[16,18] Anaerobic biodesulfurization was performed with organisms such as *Desulfovibrio desulfuricans* M6, to transform DBT into biphenyl and H_2S. However, the specificity for DBT of several anaerobic strains doesn't seem to be satisfactory and, besides, maintaining an anaerobic system has proven to be difficult.[15] The two main aerobic pathways are the Kodama pathway and the 4S pathway.[23] These are described in the following subsections.

1.3.5.1. The Kodama pathway

The Kodama mechanism is a ring destructive pathway,[6] that usually uses enzymes encoded by the doxABDEFGHIJ genes on *Pseudomonas*' plasmids. This pathway encloses three main steps: the homocyclic rings' lateral deoxygenation, ring cleavage and hydrolysis. The end product of this process is hydroxy-formyl-benzothiophene.[37] The Kodama mechanism is a carbon-carbon cleavage pathway as the sulfur compounds are used as a carbon-source by a transformation of the organosulfur compounds into water-soluble molecules. As previously mentioned, the accumulation of such hydrophilic compounds hinders DBT oxidation and microbial growth, which is undesirable. Another disadvantage of this pathway is that it also targets non-sulfur aromatic rings, which further degrades the energetic content of its resulting products.[15]

1.3.5.2. The 4S pathway

This pathway has been the focus of BDS studies for a while now, as it is sulfur specific, which on the contrary to the Kodama pathway, allows the maintenance of the oils' calorific value.[15,28] The pathway was discovered in 1993, by Gallagher et al., using *Rhodococcus erythropolis* IGTS8. Following this trend, several *Rhodococcus* species were reported as 4S pathway users. However, *R. erythropolis* IGTS8 has been the one under the most scrutiny. This is a 0.5 μm long rod-shaped bacteria, that was first isolated in 1990 by the Institute of Gas Technology in the USA.[37–39] The desulfurizing capacity of *R. erythropolis* is attributed to the *dsz*ABC genes, whose name comes from desulfurization.[31] This operon was previously called *sox*ABC and it is encoded on pSOX, a large linear plasmid of 120 kb, that possesses these genes on a 4 kb fragment of its DNA. Additionally, a chromosome encoded *dsz*D gene is also necessary. The enzymes encoded by *dsz*C and *dsz*A, DBT monooxygenase (DszC) and DBT sulfone monooxygenase (DszA), respectively, are monooxygenases, which agrees with isotopic labeling studies that demonstrated that the oxygen atom from the hydroxyl group of the 2-HBP product comes from molecular oxygen, thus implying the action of one or more oxygenases in this pathway. Additionally, *dsz*D encodes for a flavin mononucleotide (FMN) oxidoreductase, DszD, and the *dsz*B gene encodes HBPS desulfinase, DszB.[28,36,38,39] Moreover, all these enzymes seem to be soluble and found on the cells' cytoplasm.[23]

The 4S name comes from the 4 intermediates formed during its consecutive oxidation mechanism.[15,28] Firstly, DszC oxidizes DBT into DBT sulfoxide (DBTO), followed by the oxidation of DBTO into DBT sulfone ($DBTO_2$), also through the action of DszC. Then, DszA, transforms $DBTO_2$ into 2'-hydroxybiphenyl 2-sulfinic acid (HBPS), which is converted to 2-hydroxybiphenyl (2-HBP) by DszB, resulting in the release of sulfite into the medium. Since the end products, 2-HBP and its derivates, are partitioned back into the oil, no fuel value is lost.[18,38] Additionally, DszD acts in synchrony with DszC and DszA, providing them with reduced FMN, as this is necessary to ensure new desulfurization cycles.[36] A scheme representing the 4S pathway can be seen in Figure 4.

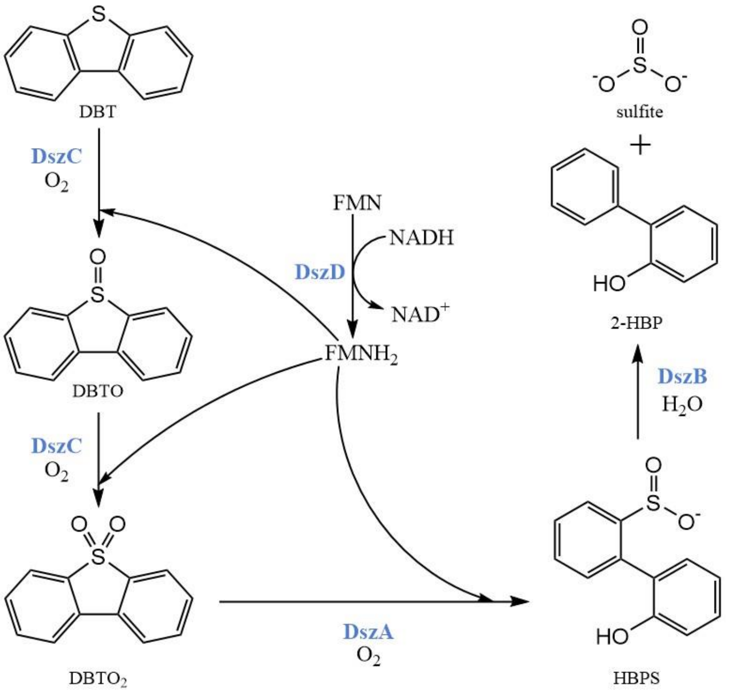


Figure 4. Scheme of the 4S Pathway

Even though *R. erythropolis* IGTS8 enzymes have been used for DBT desulfurization, they present a diminished activity when it comes to thiophenes or benzothiophenes. Thus, for a commercial application where BDS is the sole desulfurization method applied, there is a need to use either a mixture of biocatalysts or engineered enzymes to guarantee the desulfurization of the several types of sulfur compounds that constitute the petroleum.[40] This is particularly true for DszC and DszB, as they are the most inefficient at desulfurization, followed by DszA and then DszD. However, that's not all, as this pathway also suffers from another limitation, the enzyme's feedback inhibition by its products, HPBS and 2-HBP. Inhibition by the latter has proven to have a greater inhibitory strength at the beginning of the pathway, decreasing along it, which implies more prominent effects for DszC, while DszB is the least affected by it. One strategy to diminish this inhibition could involve the decrease in HBP retention; however, that seems unlikely as it has been proposed that the cytoplasmatic membrane is responsible for such retention.[32]

DszC, DszB and DszD are not the scope of this thesis. Thus, some details on their characteristics, their participation in the 4S pathway and their engineering status are described below, however, more focus is given to the enzyme under study, DszA.

DszC is a group C enzyme from the family of flavin-dependent monooxygenases, as it has a TIM-barrel fold and uses reduced FMN ($FMNH_2$) as a cofactor. Moreover, enzymes belonging to this group, need an independent NAD(P)H-dependent reductase, such as DszD, to generate their reduced FMN, as seen in Figure 4.[31,41] The *dsz*C gene from *R. erythropolis* IGTS8 occupies a 1.25 kb portion of the plasmid and its product, DszC, has a length of 417 residues in each monomer, a weight of 45 kDa and it is composed by two tetramers. However, it is active as a dimer of homodimers, each with two catalytic sites that are separated by 50 Å. This information is known as its apo and FMN bounded crystal structures have already been reported.[31,42,43] Moreover, it is also established that each monomer has three distinct regions: the N-terminal region predominantly composed of α-helices, the C-terminal region that is responsible for the tetrameric structure by protruding its four α-helices into the center of the tetramer and the middle region that encompasses the aforementioned TIM-barrel structure composed of six β-sheets. Furthermore, the cofactor binding to the apo enzyme doesn't have a substantial impact on its structure.[42,43] As previously mentioned, DszC catalyzes the first two steps of the pathway, the conversion of DBT to DBTO and subsequently, to $DBTO_2$.[32] This reaction occurs in the presence of $FMNH_2$ and molecular oxygen and it is performed in three stages, starting with the molecular oxygen activation by reduced FMN, forming a hydroperoxyflavin-intermediate ($C^{4a}OOH$). Then, oxidation of DBT into DBTO occurs via a nucleophilic attack of DBT-sulfur to oxygen in $C^{4a}OOH$, and finally, protonation of C^{4a}-hydroxyflavin results in water and oxidized FMN, which is soon reduced by DszD to ensure a new catalytic cycle.[44] As aforementioned, DszC is the most susceptible enzyme to 2-HBP feedback inhibition and it is one of the least efficient catalysts.[31,32] Moreover, it is also inhibited by HPBS through a non-competing mechanism. With such a combination of detrimental characteristics for sulfur removal proficiency, DszC seems to be an appealing target for optimization through genetic engineering. As a matter of fact, DszC mutants that overcome the feedback inhibition and others with an increased hydrophobicity that prompts an improved oil affinity were already successfully attainded.[45,46]

DszB is an oval-shaped monomer with two domains that belong to the phosphate-binding protein family. DszB from *R. erythropolis* IGTS8 has 365 residues and a weight of 39 kDa. Additionally, its gene length is 1 kb. DszB is composed of five stranded β-sheets that are found among α-helix bundles and its' active site is located between its two domains. This active site is where the only Cys residue of the entire enzyme is located, and it seems that this amino acid is essential for activity. Moreover, the binding of the substrate, HBPS, results in the rearrangement of the loops in the region into α-helices that can accommodate the biphenyl rings of the substrate. This subsequently introduces and exposes residues that are relevant for the maintenance of HBPS at the active center, through hydrogen bonds, a salt bridge, and hydrophobic interactions.[31,47] DszB catalyzes the second S-C bond cleavage, transforming HBPS into HBP and SO_3^{2-}, without using a cofactor. As for its mechanism, two have been proposed. However, due to a more favorable activation energy, the most viable mechanism is based on an electrophilic aromatic substitution that inserts a proton on a carbon that binds the SO_2^- group, forming a transient intermediate that further releases the sulfur *via* a water molecule from the bulk solvent.[48] DszB's inhibition by HBP seems to occur through a process of competitive inhibition, as HBP can get trapped on the structure's helices promoting a closed conformation that averts the substrate's access to the active site.[49] This inhibition and the fact that this enzyme has the slowest catalytic rate of the pathway make it a major rate-limiting enzyme of BDS.[47,49] Moreover, DszB has a low expression level.[50] Consequently, attempts to control its inhibition, increase its turnover and enhance its expression are being made. In fact, some successful efforts to increase the production of DszB involved the mutation of its untranslated 5' region or the rearrangement of the operon *dsz*ABC to *dsz*BCA, which resulted, respectively, in an 9- and 12-fold increase in the desulfurization rate.[50,51] Additionally, mutations on Tyr63 and Gln65, produced superior enzymes when considering their thermostability and/or catalytic activity.[52] Further optimizations, could mean a more efficient enzyme that can potentially meet the requirements for an industrial application of BDS.[36]

In contrast to the other pathway enzymes, DszD is chromosome-encoded on a 0.579 kb fragment. DszD is a dimer with 192 residues, a 22 kDa molecular weight and it belongs to the class I flavin reductases.[31,36,53,54] It is suggested that NADH is the first to bind in its active center, as this is seen in similar enzymes, then followed by its second cofactor, FMN.[36,53] This essential enzyme catalyzes the reduction of FMN to $FMNH_2$ through NADH oxidation to NAD^+. Such transformation is achieved firstly through a hydride transfer from NADH to the central nitrogen of FMN's ring, forming FMNH as an intermediate. Then, upon a proton transfer from a water molecule to FMNH, $FMNH_2$ is obtained.[36,55] Although efficient, DszD is inhibited in the presence of excess FMN, which consequently, hinders the efficacy of the entire pathway.[56] However, the overall rate of desulfurization rises upon overexpression of DszD. Thus, as the supply of the reduced cofactor to the first two enzymes of the 4S pathway influences the BDS rate, rational DszD design can amount to a pathway improvement.[36,54] In this light, site direct mutagenesis where Thr62 was substituted by Asn and Ala, revealed an increased desulfurization activity.[57] Additionally, the substitution of Ala79, one of the residues that binds the FMN, with Ile and Asn also resulted in an optimized BDS.[53]

1.3.5.2.1. DszA: dibenzothiophene sulfone monooxygenase

Just as DszC, DszA is also a flavin-dependent monooxygenase family, which entails a TIM-barrel fold, $FMNH_2$ as cofactor, and the dependence on a NAD(P)H-dependent reductase, such as DszD, to provide its cofactor in its reduced form.[58] On *R. erythropolis* IGTS8, the gene *dsz*A, is encoded on a 1.45 kb fragment of pSOX,[31] and it is the most expressed out the three genes in the *dsz*ABC operon, with ratios of 11:3.3:1 of *dszA*, *dszB*, and *dszC* mRNAs in cells.[50] DszA is the second intervenient of the 4S pathway, catalyzing the first C-S bond cleavage to convert the $DBTO_2$ product of DszC to HBPS.[59] As for DszC, it is known that DszA is inhibited by the products HBP and HBPS, through unknown mechanisms. Nonetheless, out of the four enzymes, DszA comes second on the catalytic efficiency scale, as it is three times more efficient than DszC and DszB. Additionally, its reaction rate also surpasses, in 7-fold, the reaction rate of DszC and DszB, ensuring in this way that the consumption of $DBTO_2$ is greater than its generation, but still far from the 500-fold efficiency required to meet industrial requirements. The higher efficiency of DszA relatively to DszC or DszB also entails that more HBPS is being generated than it is being consumed by DszB, which further contributes to DszA's and, especially, DszC's feedback inhibition.[32] Despite this, there is a considerable lack of knowledge when it comes to its structure and mechanism of reaction, hence, efforts to obtain a deeper understanding of this enzyme are necessary.

1.3.5.2.2.1. DszA structure: a comparison with the BdsA homolog

One strategy to generate structural and mechanistic hypotheses is to use analog enzymes (that have similar functions), such as BdsA from *Bacillus subtilis* WU-S2B, to deduce the characteristics and mechanism of DszA.[36,58]

A study where the physicochemical properties and sequences from DszA homologs of 14 different organisms, including from *Bacillus subtilis* WU-S2B, were compared, established that the number of residues per subunit varied from 434 to 474. Moreover, it concluded that at pH 7 the enzymes were negatively charged at an average of -16. Such properties are probable to also characterize DszA. Moreover, it stated that Cys residues are an uncommon component for DszA orthologs, while Ala, Gly, Leu and Val seem to be the most notorious constituents. [60] This is verified to be true for DszA.

The structure of the DszA ortholog, BdsA, is known. BdsA from *Bacillus subtilis* WU-S2B seems to share several similarities with DszA: they possess similar functions, belong to the same group of the same superfamily, participate in the same catalytic step of the 4S pathway, possess FMN as the cofactor, share a high sequence identity of 79% and a residues' conservation at the active center of 90.9%. For these reasons, it is expected that most of the structural characteristics of BdsA are also observable in DszA. Due to all of their similarities, it is also probable that they exploit a similar reaction mechanism.[58] However, the biocatalysts of the 4S pathway of *Bacillus subtilis* WU-S2B have shown a greater activity and stability when in comparison with similar enzymes from other bacteria, including DszA and DszB, when under temperatures within the 300-320K range.[36,59] BdsA studies have shown a conserved scaffold among the other group C enzymes of the flavoprotein monooxygenase family. The enzyme was reported active as a dimer of homodimers, a tetramer with a molecular weight of ~220 kDa, and with an FMN in each subunit, when in its bound form. Its homodimers have a high level of structural similarity and the vast dimers' interface is mainly composed of hydrophobic residues. Furthermore, each monomer is composed of 453 residues, organized in nine β-strands and 13 α-helices that form the characteristic TIM-barrel fold of the group C monooxygenases. Apo BdsA and BdsA:FMN structures' are identical, with slight changes on Phe56, Val137, and Phe246, which may aid the cofactors' binding.

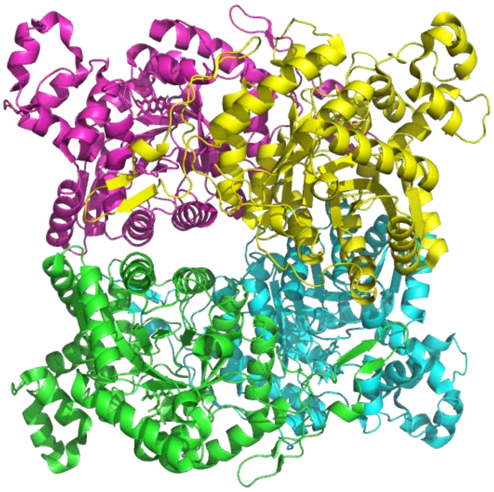


Figure 5. BdsA tetramer, a dimer of homodimers (monomers colored in magenta and yellow form a dimer and monomers in green and blue form another dimer).

The active site of the enzyme has a hydrophobic active pocket with a volume of 5134.9 Å^3, where $DBTO_2$ is proposed to bind through hydrogen bonds and hydrophobic interactions, securing its position on the *si*-face of the FMN isoalloxazine ring, as represented in Figure 6. Moreover, Phe12, His20, Phe56, Phe246, Val248, His316, and

Val372 are thought to be essential for the binding of $DBTO_2$ as single-point mutations on these resulted in a 90% activity loss for His20, Phe56, Phe246 and His316 mutants, an 80% reduction on Phe12 mutant and a 50% decrease for Val248 mutants, when in comparison with the catalytic activity of wild type BdsA. When it comes to the interaction of BdsA with its cofactor, several are seen between the phosphate-ribitol tail of FMN and the sidechains of His156, Arg159, Tyr160 and the backbone of Leu230. A hydrogen bond is also observed between the hydroxyl of Ser139 and N1 of the FMN ring, and the isoalloxazine ring establishes hydrophobic interactions with Val137. In fact, the importance of these residues for FMN binding is further established as single-point mutations on these render the complete loss of catalytic activity by the enzyme, with the exception of mutation on His156, where, nonetheless, a 50% activity reduction was seen. All these residues whose mutation completely abolished or diminished BdsA activity are conserved in DszA.[31,58] A modeled active center of DszA from BdsA can be seen in Figure 6, as proposed by Sousa et al.[36]

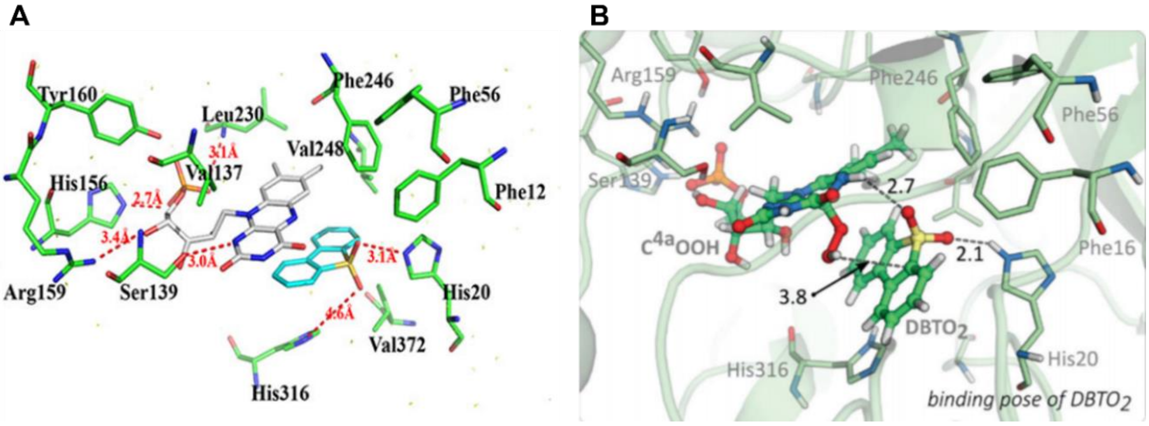


Figure 6. A) Proposed BdsA:FMN:$DBTO_2$ complex, where the substrate binds at the si-face of FMN. Residues that form the binding pocket are represented as green sticks. FMN is shown as gray sticks and $DBTO_2$ as cyan sticks. Image retrieved from Su *et al*.[58] B) Hypothetical reactive complex of modeled DszA from BdsA as proposed by Sousa *et al*.[36], where the image was also retrieved from.

1.3.5.2.2.2. Hypothesis for the reaction mechanism of DszA

The reaction mechanism of DszA, just like the BdsA's mechanism, is still not completely understood. Nevertheless, it is proposed that BdsA is responsible for two roles: providing an environment that facilitates the correct position of reduced FMN and $DBTO_2$, and activating $FMNH_2$ using Ser139 as a nucleophile. Moreover, it is proposed that deprotonation of N1 is a necessary step to activate the reduced FMN, generating $C^{4a}OOH$, and consequently catalyze substrate oxidation. A similar situation may also apply to DszA's mechanism of reaction.[36,58] This activation of $FMNH_2$ with the consequential formation of flavin hydroperoxide can be observed in Figure 7.

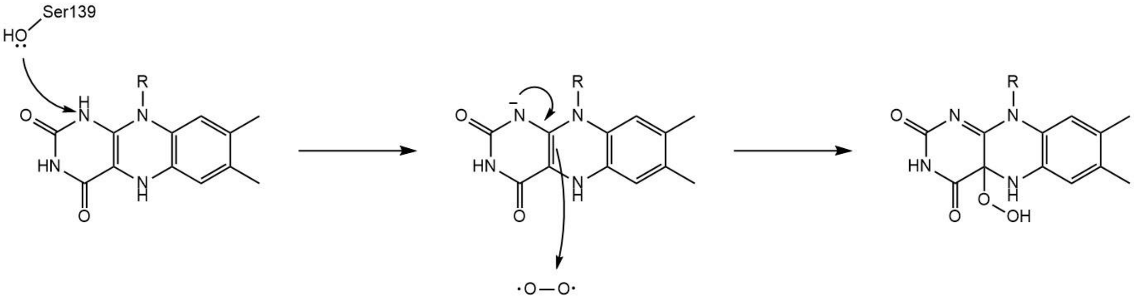


Figure 7. Proposed mechanism for activation of $FMNH_2$ in BdsA.

Nonetheless, a possible mechanism of reaction for DszA was proposed in 2016 by Adak and Begley. When simulating the DszA reaction by incubating DszA, $DBTO_2$, flavin reductase, NADH and FMN, the results seemed to indicate a mechanism where initially the $FMNH_2$ is transformed into $C^{4a}OOH$,[61] a step also seen for the proposed mechanism of BdsA.[58] Then, a base-assisted formation of a peroxyhemiacetal intermediate between $DBTO_2$ and the newly formed $C^{4a}OOH$ occurs. An acid-assisted hydroperoxyl transfer to $DBTO_2$ follows with a subsequent formation of flavin N5-oxide and the product HBPS. Finally, to allow a new catalytic cycle, the cofactor is regenerated by NADH and flavin reductase, releasing NAD^+.[61] This proposed mechanism is seen in Figure 8.

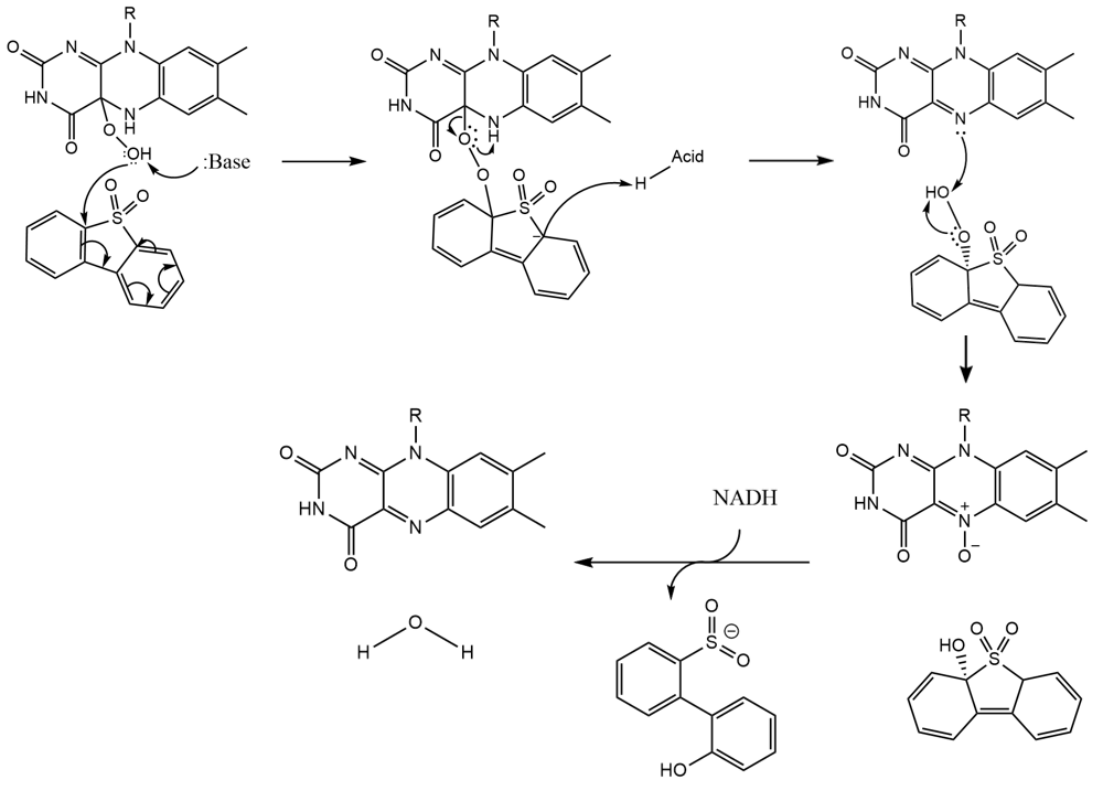


Figure 8. Reaction mechanism of DszA as proposed by Adak and Begley.

1.3.5.2.2.3. Optimization of DszA

Even though DszA does not appear to be as rate-limiting as DszC or DszB, there is still room for improvement. Moreover, if all the pathway enzymes are optimized except for DszA, such can become the rate-limiting enzyme. Hence, DszA is still an interesting target for protein engineering, especially when it comes to reducing the feedback inhibition by the pathway products. Furthermore, if the aforementioned lack of Cys residues seen in ortholog enzymes proves to be true for DszA, creating cysteine richer biocatalysts may be an approach to bring BDS closer to its effective commercial application. This is explained by the fact that some stable proteins exist in nature as cysteine-rich entities due to the disulfide bonds formed between these residues. Such interactions can confer to the proteins an increased resistance to temperature, pH and solvents, characteristics that are of great importance when considering an industrial application. Moreover, it seems that creating enzymes with improved stability may also incite the cell’s necessity for sulfur and hamper its inhibition by HBP. However, the addition of such residues does not always render the mention improvements.[62]

Nevertheless, there is still a lot that is not known about this enzyme, whether when it comes to its structure or its mechanism of reaction. Furthermore, the knowledge on DszA should be further deepened before attempting its engineering into a more efficient and viable commercial biocatalyst for BDS. With that in mind, this thesis will employ computational biochemical methods to obtain hypotheses for its apo and bounded structure, study its interaction with the substrate and test mechanistic hypotheses. The structure will be attained through homology modeling, followed by the docking of the cofactor and substrate into the catalytic center of DszA. Upon energy minimizations and molecular dynamics simulations (MD simulations), hybrid quantum mechanics/molecular mechanics (QM/MM) approaches are applied in efforts to unveil the reaction mechanism of DszA. Such knowledge should facilitate its future engineering and optimization, ensuring its industrial applicability, thus mitigating the current limitations and hazardous effects of HDS and other desulfurization techniques available.

2. Methods

2.1. Theoretical background

2.1.1. Homology Modeling

Some of the most employed techniques to experimentally obtain a proteins' 3D structure are X-ray crystallography and nuclear magnetic resonance (NMR). Yet, these present limitations. Attaining a protein structure through X-ray crystallography can take years, if at all possible, and NMR techniques can only be employed for small molecules of up to 150 residues. Moreover, there is an exponential increase in the number of available protein sequences due to the availability of sequence genome data. Yet, the vast majority has no structural information available. This is where computer-aided protein conformation prediction comes into play.[63] Homology modeling is the most accurate computational method for such predictions. It makes use of the increasing numbers of available sequences to create the tertiary structure of proteins when the sequence of a query protein is similar to a sequence of a template protein with a known structure.[63,64] The first step of homology modeling is a sequence alignment of the query sequence with the sequence of the template structure. If their similarity is above a threshold of 30% of identity, they should fold similarly. Nonetheless, the higher the sequence identity, the more accurate the results. After this alignment, a model is built generating its backbone and side chains. This can be achieved through several techniques, such as fragment assembly (where the coordinates of conserved regions are simply transferred from the template structure to the model structure), segment matching (which divides the target-protein into short segments and then each is matched to its own template. The model is built juxtaposing the atomic positions of the template fragments) and satisfaction of spatial restraints (which employs mathematical methods to satisfy a set of spatial constraints that the target protein must obey as closely as possible). For the gaps seen in the sequence's alignment, loop conformations are inserted either through a knowledge- or energy-based approach. This is then followed by a model validation, through a stereochemical evaluation, checking if bond angles/lengths fall within normal ranges, whilst considering favorable energies, the distribution of the residues in the Ramachandran plot and comparison with other structures existing in nature.[65,66]

The Swiss-Model is a webserver dedicated to performing proteins' homology modeling. When given a query sequence it identifies structural templates, aligns the query sequence with such template structures and builds the model. Then, the model's reliability and quality are evaluated through the values of the GMQE (Global Model Quality Estimation), QMEAN, QMEAN Z-score and through the analysis of the Ramachandran plot. GMQE expresses the expected accuracy of a model, giving it a value between 0 and 1, where the higher the value, the more reliable a model is. QMEAN, is a quality indicator for geometric properties, and values close to 0 indicate that the observed properties are identical to those seen in experimental structures or similar, further reinforcing the models' reliability. A similar quality indicator, the QMEAN Z-score represents how the QMEAN of the model structure is comparable to experimental structures of similar sizes. A good quality model has a QMEAN Z-score near 0, indicating that there is a high similarity between the properties of the model structure and those of experimental structures. Low-quality models have QMEAN Z-scores of –4 or less. Finally, the Ramachandran plot illustrates if the secondary structures of the model residues fall into the common secondary conformations seen for the structures existing in nature.[67]

2.1.2. Protein-ligand Docking

An understanding of the interactions between a specific receptor and a ligand is an essential step to promote the understanding of biological processes, such as the mechanism of reaction under study. This can be accomplished by determining the binding mode of a protein:ligand complex.[64] Even though experimental methods, such as X-ray crystallography, are sometimes able to describe such association, the process is not always straightforward and often implies an excessive time effort. This is a reflection of these methods' incapacity to describe the association between the receptor and the native ligand, often having to resort to receptor inactivation through mutations or ligand mimetics to overcome such limitation. Hence, computational methods, specifically protein-ligand docking are a useful alternative. Protein-ligand docking can predict the binding pose and energy of complexes of at least two molecules: one or more ligands, and a protein with a known structure. [63,64,68] This method can be performed either as a rigid docking, where both the receptor and the ligand are treated as rigid entities, or as flexible docking, where the ligand is considered flexible and the receptor rigid. Although the rigid method entails a smaller search space and thus represents a faster approach, there is a lower chance of finding a reliable binding mode if the conformation of the small molecule is not correct. Moreover, efforts are also being made to employ receptor flexibility, as such allows for a better description of the affinity between protein and ligand. [64,68,69]

Prediction of the binding pose of one or more ligands to its target protein is only possible due to the employment of search algorithms and scoring functions by docking programs.[68]

2.1.2.1. Search Algorithms

A search algorithm allows the generation of several ligand poses on the target protein, given a target search space. It is important to consider its speed and effectiveness at covering the search space, but also a flexible number of degrees of freedom of the protein:ligand system to increase the chances of it including the real binding poses. These algorithms can either be systematic or stochastic. A systematic algorithm samples the search space at predefined intervals until it exhausts every possible combination of rotatable bonds of the ligand. Hence, the application of this algorithm is limited as it generates a massive number of structures.[68,69] Nonetheless, it is still the algorithm used in several docking software (such as LUDI, FlexX and DOCK), usually by applying constraints to the ligand in order to reduce the number of structures generated. The stochastic or random algorithm applies random alterations onto the ligand until it meets the predefined benchmarks. This algorithm is also employed in several docking software (such as GOLD and AutoDock) and a more detailed description of a random algorithm is given in section “2.1.2.3.1. GOLD”.[68]

2.1.2.2. Scoring functions

A scoring function calculates an affinity score for each ligand’s conformation to the protein receptor, suggesting the most probable binding modes. To ensure its applicability and suitability to cover a large number of resulting poses, this function should be fast. However, increases in speed are a result of the function’s simplification, which hinders its accuracy. Thus, the main bottleneck of scoring functions is attaining a function that is suitable both in speed and accuracy.[68,69]

Scoring functions can either be empirical, force field based or knowledge based.[68,69] An empirical approach, uses experimental binding affinity data, which through regression analysis enables the correlation of the free energy of the binding to a set of unrelated variables.[70] A force field based scoring function utilizes molecular mechanics concepts (which will be discussed in "2.1.3. Molecular Mechanics"). It works by quantifying, through van der Waals (VDW) and electrostatic terms, the interaction energy of the protein-ligand complex and the internal energy of the ligand.[68,71] Finally, a knowledge based function uses the statistical data that arises from the analysis of interacting atom-atom pairs between the protein and the ligand(s) with experimentally determined 3D structures. The pairs of atoms that regularly appear in large datasets are given a greater score. Moreover, these atom pairs are the ones who set the criteria to describe their preferred geometry.[68,70]

2.1.2.3. Docking software

Throughout the last decades, a vast amount of docking programs was developed. Some examples are GOLD, AutoDock, AutoDock Vina, FlexX, Dock, ICM and LUDI, which mostly differ from each other on their search algorithms, scoring functions and docking flexibility.[68,72]

2.1.2.3.1. GOLD

GOLD or Genetic Optimization for Ligand Docking is one of the most employed docking software.[64,68] It is a flexible docking program, allowing for full ligand flexibility and rotational flexibility of the polar hydrogens.

It uses a stochastic search algorithm that operates with a genetic algorithm (GA) method.[64,68,73] The GA aims to mimic the evolutionary biological processes that occur in chromosomes (mutations, crossovers and migrations). These are performed onto an initial random population of possible ligand poses (chromosomes), characterized by 'genes', that give information on their translation, orientation and conformation. The poses that arise from such process then replace the least-fit entities of the population, just as the survival of the fittest process observed in nature. Through repetition of this procedure, optimum binding modes arise.[68,74]

As for the scoring function, GOLD allows the choice from several functions: GoldScore, ChemScore, ASP (Astex Statistical Potential) and ChemPLP (Piecewise Linear Potential). GoldScore is force field-based, and ASP is knowledge-based. ChemScore and ChemPLP are empirical scoring functions. The latter is set as the default (from GOLD version 5.1 and upwards), as on average it performed better than the other methods on pose prediction experiments.[68,75,76] ChemPLP is able to model the steric complementarity of the protein and the ligand, and it combines parts of already published force fields and scoring functions. It applies ChemScore angle-dependent terms, to account for both hydrogen bonding and metal binding and, to consider the intraligand interactions, a heavy-atom clash term and the torsional potential from the Tripos force field are included.[77]

2.1.2.3.2. Autodock Vina

AutoDock Vina is also one of the most used docking programs and it enables not only ligand flexibility but also flexibility of the receptor's side chains.[78–80] It is a new generation of the AutoDock docking software, one of the most famous docking software.[64,68,81] However, Vina is faster when it comes to binding pose prediction than its ancestor. Such improvements are achieved by maintaining most of the characteristics from AutoDock 4 and modifying its scoring function, algorithms and source code.[80,81]

The new and improved scoring function of Vina combines knowledge-based and empirical scoring functions, as it uses binding affinity data from a database (PDBbind) and experimental affinity values.[80,82] Its scoring functions account for VDW interactions, hydrophobic contacts, hydrogen bonding and loss of entropy on binding.[79]

As for its search algorithm, it is a stochastic combination of a GA with local gradient optimization. Hence, successive mutation steps with local optimization are made, and each result is either accepted or rejected according to predefined criteria.[80]

2.1.3. Molecular Mechanics

Quantum mechanics (QM) methods treat atoms as particles consisting of nuclei and electrons and, for large systems, due to the complexity of the equations that describe these particles' behavior, calculations performed with QM are extremely time-consuming. An alternative is molecular mechanics (MM) methods. MM treats electrons implicitly and thus when presented with a system with a substantial number of atoms, it is a frequent choice to perform calculations, as it dramatically reduces the calculation time, by using classical physics laws.[83] This also makes it the more suitable method to study solvent effects or crystal packing through simulations.[84]

MM, also known as force field methods, aims to describe the energetic properties (and other related characteristics) of molecular systems. It searches for the local energy minima that correspond to atoms' stable configurations in order to create and analyze multidimensional potential energy surfaces (PES), to follow trajectories of molecular dynamics simulations or to study the system's thermodynamic and geometric characteristics. It is based on simple physics equations, classical mechanics and some assumptions and approximations. MM methods consider nuclei and electrons as a single atom-like particle. Atoms are regarded as spheres connected by springs, and therefore, harmonic potentials based on the elastic properties of Hooke's law are considered for quantifying most of the interaction between bonded atoms when in equilibrium. The implicit treatment of the electrons is done through force field parameters, that allow the generation of potential energy surfaces, thus attaining the energy for a given set of atom coordinates. [83,85,86]

Despite it being able to provide good geometry and energetic predictions for a significant number of molecules in a short time, MM methods are dependent on the availability of suitable parameters. This is especially relevant for uncommon molecules, where usually parameters are not available, making MM applicability more limited or labor-intensive if the missing parameters have to be created. Another limitation of these methods is that assessing the probable error of the results is only possible by comparison with calculations on other similar molecules that possess experimental data.[84]

2.1.3.1. Force fields

A force field is an equation that calculates the systems' PES, as a function of the atom positions and a set of parameters. A simple force field equation can be represented as:

$$E_p = \overbrace{E_{bond} + E_{angle} + E_{dihedral}}^{E_{covalent}} + \overbrace{E_{electrostactic} + E_{VdW}}^{E_{noncovalent}}$$

Equation 1. Simple force field equation for molecular mechanics calculations.

where E_p represents the system's potential energy, which is equivalent to the sum of the energetic contributions of the covalently and noncovalently bonded atoms, $E_{covalent}$ and $E_{noncovalent}$, respectively. $E_{covalent}$ can be further decomposed into: E_{bond}, that describes the energy of the deformation of the bond length between two atoms; E_{angle} representing the energy of the bond angle's variation; $E_{dihedral}$ accounting for the torsional energy of dihedral angles. $E_{noncovalent}$ is frequently represented by the energetic contributions of electrostatic ($E_{electrostactic}$) and Van der Waals (E_{VdW}) interactions.[83,86] A schematic representation of the equation's terms is depicted in Figure 9.

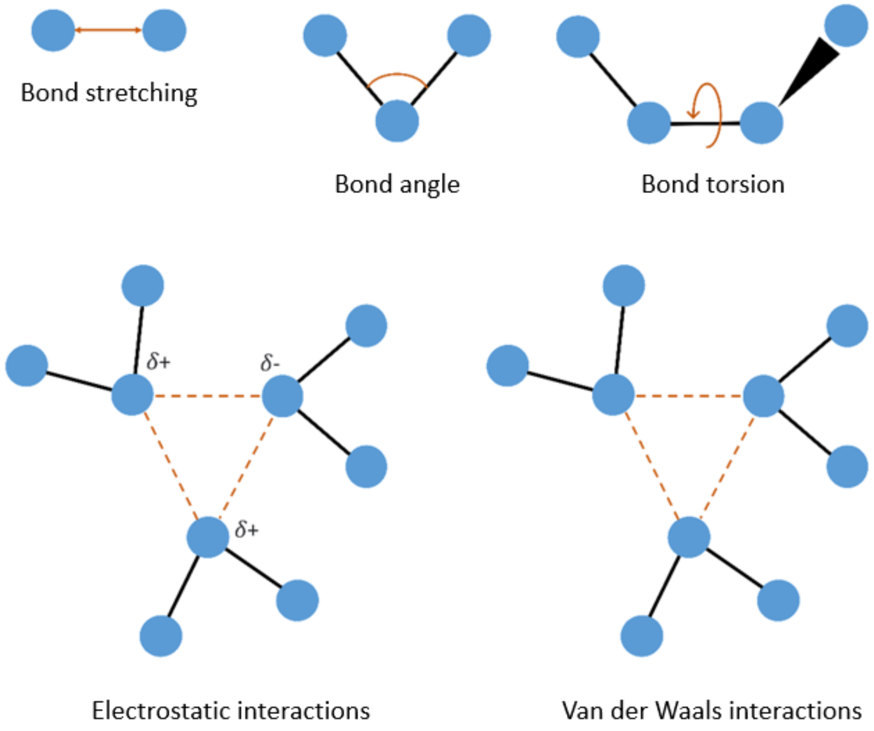


Figure 9. Schematic representation of the terms of a force field

There are several force fields available, however, almost all of them account for at least the terms depicted in Equation 1. Some examples of biomolecular force fields are AMBER, CHARMM, GROMACS and OPLS, which slightly differ on the potential energy function and the employed parameters. Nonetheless, the similarities between them greatly surpass their differences.[83,87]

2.1.3.1.1. AMBER force field

AMBER or Assisted Model Building with Energy Refinement was created in the ’80s by the Kollman group. It is used to describe a wide range of biomolecules, including proteins and nucleic acids, which was made possible upon an improvement of the AMBER force field by Cornell and co-workers that introduced parametrized sets for biomolecules. In this improved force field, charges are determined from an electrostatic surface potential (ESP) calculated with the Hartree-Fock energy function and the 6-31G* basis set (both later explained), and after by a restrained electrostatic potential (RESP) fitting that forces equivalent atoms to have the same charge.[84,86,88]

Several AMBER parameter sets are available. AMBER parameters for proteins are denominated by ‘ff’ followed by the two digits that represent their creation year, for example, ff99 or ff14SB. Small organic molecules, such as ligands, are generally parametrized by GAFF (General AMBER force field) and a common parameter set for water molecules is TIP3P.[83,86,89]

In the Amber force field, Equation 1 takes the following form:

$$E_p = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{\substack{Improper\\dihedral}} K_\Phi(\Phi - \Phi_0)^2 + \sum_{dihedral} K_\phi[\cos(n\phi - \delta) + 1] + \sum_{ij}\left[K_c\frac{q_iq_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right]$$

Equation 2. Potential Energy function equation in AMBER force fields.

Where $\sum_{bonds} K_b(b - b_0)^2$ represents E_{bond} of Equation 1 and K_b is a force constant that is multiplied by the square of the bond elongation (b) relative to the equilibrium state (b_0). $\sum_{angles} K_\theta(\theta - \theta_0)^2$ is equivalent to E_{angle} in Eq.1; K_θ is the force constant of an angle (θ) and θ_0 the equilibrium angle. $\sum_{\substack{Improper\\dihedral}} K_\Phi(\Phi - \Phi_0)^2$ describes out-of-plane angles (Φ) in relation to its equilibrium angle (Φ_0), multiplied by the respective force constant, K_Φ. This term is not depicted in Equation 1, but it accounts for the necessary energy to maintain the planarity of a molecule. $E_{diehdral}$ from Equation 1, now takes the form of a periodic energy function, $\sum_{dihedral} K_\phi[\cos(n\phi - \delta) + 1]$, where K_ϕ is the energetic barrier needed for the dihedral rotation. The number of minima that the function can overcome is given by *n* and δ is the angle phase where the reference minimum is placed. Finally, $E_{electrostactic} + E_{VdW}$ are presented here as $\sum_{ij}\left[K_c\frac{q_iq_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right]$, where the electrostatic terms are calculated with the Coulomb law, the energy associated with the interactions is given as a function of the charges (q_iq_j) and the distance between them (r_{ij}) multiplied by the Coulomb constant (K_c). The energy associated with Van der Waals interactions is calculated by a Lennard-Jones (LJ) potential and it is also dependent on the interatomic distances, r_{ij}, and the experimental determined parameters, A_{ij} and C_{ij}.[86,87,90]

2.1.3.2. Molecular Dynamics

Having an atomistic-level known structure gives a good understanding of its functioning; however, the atoms in a biomolecular system are in constant motion. Hence, a static structure can only be representative of a single moment in time, which is not ideal as both biological function and intermolecular interactions change with the dynamics of a system. Molecular dynamics allows the prediction of how every atom of a given system behaves over time. Hence it gives insight into the system's motion, its conformational changes, folding and the outcome of introduced perturbations (*e.g.*, mutations, protonation, ligand binding, phosphorylation, etc.).[91]

In classical MD, to compute the systems' evolution in time and output the respective system's trajectory that describes the position and velocity of every atom, MD enforces calculations based on Newton's second law of motion: $\vec{F} = m\vec{a}$.[83] Moreover it requires a set of initial positions, generally retrieved from an X-ray structure or modeled structure, and initial velocities that are randomly attributed from a Maxwell-Boltzmann distribution at a desired temperature. The method also demands boundary conditions and a description of how the molecules interact and the existing forces between the atoms, which can be attained from the potential energy given by the force field equations.[92] This reliability on force fields and its parameters poses limitations to the method. Parametrization of all system components is needed; however, their transferability from force field to force field may not be possible, as different force fields may have different parameters and parameterization schemes. Another drawback is that classical MD cannot be generally used to directly study the formation and cleavage of chemical bonds, as a MM method does not account for electronic behavior.[91]

2.1.3.2.1. Boundary conditions

For most protein molecular dynamics simulations explicit solvent molecules are added, typically using the TIP3P water model, to fill a box. However, as these water molecules (located in the limits of the simulation box) are not bounded to each other, contrary to the protein atoms (in the middle of the simulation box), they can undergo significant motions during the simulation. Additionally, the conditions at the box limits differ from the conditions at the center, further influencing its atoms' behavior. To homogenate such conditions, periodic boundary conditions can be applied, where the simulation box is surrounded by infinite replicas of itself. With such approach, when an atom leaves the main cell, another atom enters from the opposite site to replace it, conserving the number of particles in a given unit cell and eliminating the effects of the boundary. Special consideration should be given to the box radius, to ensure that no interactions between the same atoms, present in two adjacent simulation cells, can occur.[92]

2.1.3.2.2. Thermodynamic ensembles

In MD several types of ensembles can be used. An ensemble is a fictional compilation of a vast number of conformations that share macroscopic attributes, such as pressure or density. An *NVE* or microcanonical ensemble, holds the number of particles (*N*), the volume of the simulation cell (*V*) and the total energy of the system (*E*) constant during the simulation. However, such ensembles have some limitations, integration errors and inconsistencies that can be diminished by an ensemble where the temperature (*T*) of the simulation can be controlled, such as *NVT* and *NPT*. *NVT* or canonical ensemble maintains *N*, *V* and *T* constant during the simulation and an *NPT* ensemble (isothermal-isobaric ensemble) does the same but for *N*, pressure (*P*) and *T*. These ensembles can either aim to reproduce the conditions of the systems that are seen in nature, or they can mimic experimental conditions. Reproducing the natural conditions of biosystems usually involves constant temperature and pressure conditions and thus *NPT* ensembles are commonly simulated.[83,92–94]

2.1.3.2.3. Integration step

The integration step, usually depicted as Δt, represents the timestep between each calculation of the motion equations. It assumes that the forces that act up on the system remain constant between each integration step, and it is defined by the user. An appropriate choice of Δt is of the utmost importance, as the accuracy of the system's movement prediction depends on it. A small integration step is associated with improved accuracy; however, it results in a very large number of timesteps and a great computational effort, especially considering that most biochemistry events happen at the timescale of nanoseconds, microseconds or even longer, and there are an immense number of interactions to be evaluated at every timestep. If a large integration step is used, fewer iterations are necessary, but significant energetic drifts and fluctuations are seen, potentially resulting in an unstable simulation. Hence, a compromise is required, and the integration step should be 10 times smaller than the fastest movement of the system. This is done to decompose the oscillatory movement of the atom's vibration into small linear motions, as the description of an oscillatory movement would require far more complex functions. The fastest movement of the system, usually the vibration of C-H, is used as reference as it is one of the fastest and thus would be the most imprecisely described if an inadequate integration step is applied. As the vibration of C-H usually occurs at a frequency of 10 femtoseconds (fs), a time integration of 1 fs is considered suitable to predict the atomic positions and velocities of the system.[83,91,92] A 2 fs time step may also be suitable if the SHAKE algorithm is used. When SHAKE is employed, a constriction of the bonds between hydrogen and heavy atoms is set, thus turning the vibration of second-row atoms into the fastest movement in the system. As these have a frequency of 20 fs, a 2 fs integration step may give accurate results and produce longer trajectories with lower computational efforts.[95]

2.1.4. Quantum Mechanics

Quantum mechanics methods, in opposition to MM, treats the electrons explicitly. This gives it the ability to derive properties of the system that are dependent on electronic distributions, thus transforming QM methods into particularly relevant approaches for the study of chemical reactions, where bonds are being broken and formed. Nonetheless, QM methods are very complex, which limits their application, and their use is usually reserved for systems of about 200 atoms. To characterize all the physical properties of a system, QM methods employ the fundamental postulate of quantum mechanics, which describes microscopic systems as wavefunctions:

$$\vartheta\Psi = e\Psi$$

Equation 3. The postulate of Quantum Mechanics

Where ϑ is an appropriate operator that acts on Ψ (the wavefunction that describes the probabilistic behavior of an electron as a wave) to attain the same wavefunction and the scalar value (e) of the property associated with the operator. [83,95,96]

2.1.4.1. The Schrödinger equation

The Schrödinger equation, developed in 1925 by Erwin Schrödinger, is able to describe molecular, atomic, subatomic and macroscopic systems, which includes the prediction of nuclei and electron behavior. The time-independent Schrödinger can be written as:

$$\hat{H}\Psi(r) = E\Psi(r)$$

Equation 4.Time-independent Schrödinger equation

In the Schrödinger equation, ϑ from the postulate of quantum mechanics (Equation 3) is substituted by the operator $\hat{H}$, the Hamiltonian. This enables the calculation of the system's energy (E).

Despite all the information that solving such equation can give, it can only be solved analytically for one-particle systems. Thus, approximations and simplifications are required for multi-body systems.[83,85,96,97] The Born-Oppenheimer approximation can be here applied to simplify the problem. Such considers the nucleus as having a much greater mass than the electrons, which results in electrons instantaneously adjusting to nuclear movements. Such enables the separation of the Hamiltonian into two equations: the electronic Hamiltonian ($\hat{H}_{elect}$) that considers a system of electrons and nuclear fixed positions, and it is calculated by quantum mechanics; and the nuclear Hamiltonian ($\hat{H}_{nuclear}$), that can be solved by classical mechanics equations.

$$\hat{H} = \underbrace{\hat{H}_{nuclear}}_{aproximately\ classic} + \underbrace{\hat{H}_{elect}}_{quantum\ behavior}$$

Equation 5. The Born-Oppenheimer approximation

Both the nuclear and electronic contributors to the Hamiltonian can be described as a sum of their kinetic and potential energy. The kinetic energy of the nuclei ($E_{k,nuclear}$) is given by the relation between the nuclear mass (m) and their velocity to the second power (v^2). However, as stated by the Born-Oppenheimer approximation, as the electrons readily adjust to changes in nuclei positions, the wavefunction is only dependent on the nuclei positions and not on their momenta. Hence, when calculating new geometries, it is assumed that v^2 does not affect the new positions, consequently

$E_{k,nuclear}$ can be considered as negligible relatively to $E_{k,elect}$, and added *a posteriori* to the energy of the system. Therefore, $\widehat{H}_{nuclear}$ is solely given by the electrostatic repulsion between the positively charged nuclei, depicted as the sum of the potential energies between nuclei ($E_{p,nn}$), where Z_i and Z_j represent their atomic numbers, in relation to the particles' distance, r_{ij}. The first term of the $\widehat{H}_{elect}$ represents the kinetic energy of the electrons ($E_{k,elect}$), where ∇_i^2 is the Laplacian operator. The attraction of the electrons to the nuclei is accounted for by their potential energy, $E_{p,ne}$. Finally, the repulsion between the electrons is also described by their potential energy, $E_{p,ee}$.[83,97]

$$\widehat{H}_{nuclear} = \underbrace{\sum_{i}^{nuclei} \frac{m_i v_i{}^2}{2}}_{E_{k,nuclear}} + \underbrace{\sum_{ij}^{nuclei} \frac{Z_i Z_j}{r_{ij}}}_{E_{p,nn}}$$

Equation 6. Nuclear contribution to the Hamiltonian

$$\widehat{H}_{elect} = -\underbrace{\sum_{i}^{electrons} \frac{\nabla_i^2}{2}}_{E_{k,elect}} - \underbrace{\sum_{i}^{nuclei} \sum_{j}^{electrons} \frac{Z_i}{r_{ij}}}_{E_{p,ne}} + \underbrace{\sum_{i<j}^{electrons} \sum \frac{1}{r_{ij}}}_{E_{p,ee}}$$

Equation 7. Electronic contribution to the Hamiltonian

As in such approximation the nuclei are considered to be fixed, for every nuclei arrangement, the Hamiltonian is solved considering only $\widehat{H}_{elect}$. However, when there is a change in nuclear positions, $\widehat{H}_{nuclear}$ needs to be added to $\widehat{H}$ to attain the system total energy, E of Equation 4.[83]

Yet, the Born-Oppenheimer approximation is still limited as it can only be exactly solved for the simplest molecular species that possess two nuclei and one electron, such as H_2^+.[83] Some of the most common alternative approaches to predict nuclei and electronic behavior can be divided into density functional theory (DFT), *ab initio* and semi-empirical approaches. [83,85,96,97]

2.1.4.2. Hartree-Fock

Hartree-Fock (HF) is the most common *ab initio* method.[97] Moreover, it is also the simplest method to attain the energy and orbital configuration of a polyelectronic system.[97,98] Such method considers each electron spatial distribution to be independent of the instantaneous motions of other electrons. Hence, HF discards the effects of electron correlation, this is, the stabilizing effect resulting from alternative electron distributions that decrease electronic repulsion, and thus HF gives the probability of finding an electron around an atom as dependent on its distance to the nucleus but not on its distance to other electrons.[96,97] However, it includes an average of the effect of electronic repulsion through the integration of a repulsion term.[97] Disregarding the electron correlation, the Hamiltonian can be separated into many simple one-electron equations, where an orbital and an orbital energy is given for each one-electron equation.[97,99] Nonetheless, accounting for electronic repulsion only as an average, often results in large errors, especially when employed to reactive chemical events where electron correlation plays a big role. These errors are translated into energy values that are always greater than the exact energy.[97,98] Hartree-Fock also fails to comply with the antisymmetry principle that states that the spin coordinates should be antisymmetric under the interchange of any pair of fermions (such as electrons). This is, if an electron is switched with another electron, only the sign of the total wave function can change. To satisfy the antisymmetric principle a determinant of spin orbitals, the Slater determinant, is employed, as interchanging two electrons is equivalent to switching two columns of the determinant. This determinant is composed of spin orbitals, where the orbital function is multiplied by the electron spin and a consequence of its application is that all electrons become indistinguishable. Thus, each electron is associated with every orbital. However, this is advantageous, as electronic indistinguishability is a requirement of quantum mechanics. HF method starts with an initial orbital guess, from which the energy is calculated. Then from the energy, a new set of orbitals is derived, and the process repeats itself until convergence is achieved. Hence, the wavefunction is approximately given by the Slater determinant with the lowest energy, which is equivalent to saying that HF determines the set of spin orbitals for which the energy is minimized.[97,99] The Hamiltonian can be described as a sum of the kinetic energy of the electrons, the nucleus-electron attractive potential, and the electron-electron interaction potential, just as seen for the Born-Oppenheimer approximation. However, here the electron-electron interaction potential accounts with the repulsion term and the exchange energy. This latter arises from the required antisymmetry of the spin coordinates and describes the probability of finding two electrons with the same spin close to each other, as they can shift interchangeably among one-electron orbitals.[100]

2.1.4.3. Density functional theory

Density functional theory overcomes some of the limitations of the Schrödinger equation as it can be applied to polyelectronic systems while providing accurate results.[101] It does so by replacing the concept of individual electrons with electron density (ρ). DFT also states that such electron density distribution is the only requirement to calculate the ground state energy of a molecule. Moreover, when in comparison with wavefunction methods, DFT methods require a reduced computational cost, because it depends on solely 3 dimensions (the cartesian coordinates: $x, y \text{ and } z$) and it is independent of the number of electrons of the system, whereas the wavefunction methods are dependent on 3*N* dimensions, where *N* is the number of electrons.

A functional, this is, a function of functions, for DFT can be portraited as the following:

$$E_{DFT} = E_k[\rho] + E_{ne}[\rho] + E_{ee}[\rho]$$

Equation 8. Energy functional equation of DFT

where the energy of the system E_{DFT} is given by $E_k[\rho]$, that describes the kinetic energy, $E_{ne}[\rho]$ that accounts for the interactions between the electrons and the nuclei and $E_{ee}[\rho]$, representing the electron-electron interactions. However, the terms $E_k[\rho]$ and $E_{ee}[\rho]$, can't be explicitly solved. Hence, the addition of Kohn-Sham orbitals is a solution to this problem, as this separates the kinetic energy $E_k[\rho]$ into kinetic energy for non-interaction electrons $E_{k^{ni}}[\rho]$ and residual kinetic energy, $E_{k^r}[\rho]$. $E_{ee}[\rho]$ also suffers from a division, giving two terms, a term for Coulomb interactions $J[\rho]$ and an exchange term $K[\rho]$. Moreover, the sum of $E_{k^r}[\rho]$ and $K[\rho]$, is also known as an exchange-correlation functional ($E_{XC}[\rho]$), which can only be solved by attributing an approximated functional. Hence, the DFT functional takes the following form:

$$E_{DFT} = E_{k_{ni}}[\rho] + E_{ne}[\rho] + J[\rho] + E_{XC}[\rho]$$

Equation 9. Energy functional equation of DFT with the exchange-correlation functional.

The electronic density of the system is assigned to Equation 9 by allocating a guess density from a set of initial electronic orbitals. Then, a new set of orbitals is applied, and a new electron density is generated. If the energy that results from the latter set of orbitals is close enough to the one from the previous guess, within a given threshold, the electron density that corresponds to the ground-state energy is found. If not, new sets of orbitals are generated until the resulting ground-state energy matches the one from the previous guess. When this happens, it is said to be self-consistent. Hence, the ground-state energy is determined by a self-consistent field (SCF) calculation, where the density generates the energy of the system and the energy of the system validates the wavefunction.

DFT has a good relation between the computational effort, accuracy, and the size of the system in study. However, it has some drawbacks as it does not contemplate medium and long-range VDW interactions, and its accuracy is highly dependent on the employed approximation for the exchange-correlation functional.[83,85,96,97,102]

2.1.4.3.1. Exchange-Correlation Functionals

As mentioned, the exchange-correlation functional is of the utmost importance for a successful calculation of an electronic density.

Several exchange-correlation functionals are available and they tend to be simple and to provide promising results, thus increasing the appeal of the DFT approach in computational chemistry. The simplest one is the local density approximation (LDA), in which the electron density is considered homogeneous across all space. Hence, for molecules with heterogeneous electronic densities, such as biomolecules, LDA is often not trustworthy and should not be applied. For systems with high spin, LDA is replaced by the local spin density approximation (LSDA). An approximation that considers the heterogeneity of the electron density is the generalized gradient approximation (GGA). Such approach introduces the dependency on the gradient of the electron density, thus often providing more reliable results for biomolecules. Another popular approach that has produced accurate results is the hybrid GGA, which includes B3LYP, where a percentage of Hartree-Fock exact exchange is blended with the exchange and correlation functionals. As these hybrid functionals have an exact exchange that is lacking on nonhybrid exchange-correlation functionals, these tend to possess an improved energy calculation accuracy. Thus, they are of added value for calculating activation energies.[83,85,96,97] Nonetheless, these hybrid functionals still present some limitations. They possess an ineffective treatment of charge transfer processes, overestimating their interaction, an overpolarization of molecular radicals and they are also not able to account for dispersion interactions. Moreover, they also give wrongful, very low, asymptotic energies to describe the dissociation of simple two-center systems.[103–105]

2.1.4.3.1.1 B3LYP

B3LYP, standing for Becke 3-parameter Lee-Yang-Parr, is the most widespread exchange-correlation functional.[85,97] Such popularity is due to its accuracy for a broad range of compounds, particularly organic molecules.[97] B3LYP includes an LDA for electron-electron and electron-nuclei energy (E_X^{LSDA}), an exact exchange (E_X^{HF}) and GGA corrections (E_X^{Becke}) to the exchange term, a Lee-Yang-Parr non-local correlation functional (E_C^{LYP}), and a standard local correlation function developed by Vosko, Wilk and Nusair (E_C^{VWN}). B3LYP can be described by the following equation:

$$E_{XC} = (1-A)E_X^{LSDA} + AE_X^{HF} + BE_X^{Becke} + CE_C^{LYP} + (1-C)E_C^{VWN}$$

Equation 10. B3LYP exchange-correlation functional

Here, the exchange potential is accounted for by the first three terms of the equation, whilst the remaining two are associated with the correlation potential. In Equation 10, *A*, *B* and *C*, are empirical coefficients determined by fitting experimental data: *A* = 0.20, *B* = 0.72 and *C* = 0.81.[83,85,96,97]

2.1.4.3.2. Basis set

A basis set can be defined as a set of functions (basis functions) that can be combined to describe molecular orbitals. The selection of a proper basis set for such calculations must be carefully considered as it has a great impact on the accuracy of the results. One option is to create a basis set from scratch by merging linear combinations of basis functions and angular functions. However, it is more common to employ already existing basis sets, such as Slater type orbitals (STO) and Gaussian type orbitals (GTO). Slater type orbitals are an accurate mathematical model that can give the exact solution to the Schrödinger equation for the hydrogen atom. However, as they are exponential functions, STO are only effectively applicable to mono or diatomic systems, otherwise they are too computationally demanding. Nevertheless, these tend to be more time costly and thus, only employed when a high level of accuracy is needed. Hence, GTO tends to be the number one choice for routine QM calculations, as they are a simpler set of mathematical functions and are applicable to systems with more than two atoms.

As GTO entails a diminished computational effort, their accuracy in the description of the core and valence atoms is also reduced.[84] To solve such limitations and attain superior results, more complex basis sets are available. The addition of Gaussian mathematical functions enables a better description of the valence electrons' orbitals and the addition of polarization and diffuse functions can improve the flexibility of the orbitals. A widely used GTO basis set that has a favorable compromise between computational cost and accuracy and includes the polarization functions is 6-31G(d). It is a Pople basis set and its notation stands for a contraction of 6 GTO primitives that describe each core electron orbital, a contraction of 3 primitives and another contraction of 1 primitive, both describing each valence electron shell, and d, sometimes also depicted as *, refers to the addition of *d* primitives (polarization functions) to atoms other than hydrogen. [97,102]

2.1.5. Hybrid QM/MM methods

QM methods explicitly treat interacting electrons and nuclei, and thus are suitable to study chemical events. However, their limitation resides in the size of the system in study, where they should usually not surpass 200 atoms. Unfortunately, biomolecules usually greatly exceed that number of atoms. Fortunately, QM/MM methods can solve this problem by exploiting the advantages of both QM and MM methods, describing different parts of the system with different levels of theory. The region of chemical interest where bonds are being broken and formed, the active center, is treated with QM methods that can describe such chemical phenomena. Moreover, as this catalytic center is usually only a small region of the system, the applicability of such level of theory is possible and not too computational demanding. The remainder of the system, despite not being so chemically relevant, still has an impact on the catalytic active site. Therefore, it should also be accounted for but employing a suitable approach that is able to process a large system in a short period of time. Hence, this system remainder is treated at the MM level.[96,106]

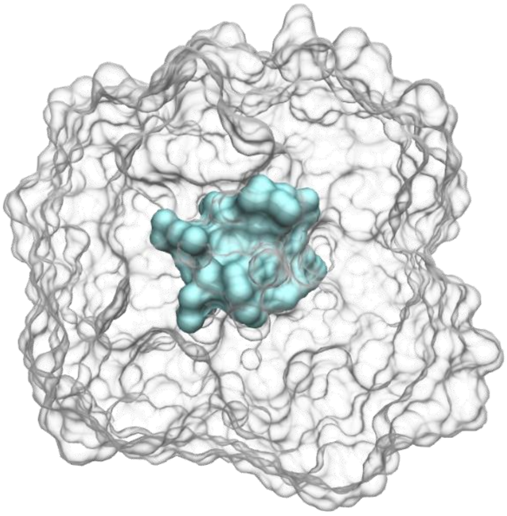


Figure 10. A QM/MM model. In blue is the active center treated with QM and in white is the remainder of the protein, which is treated with MM.

2.1.5.1. QM/MM schemes

There are two approaches to QM/MM methods: additive and subtractive. In additive schemes, the energy of the system is described by the sum of the energy of the QM layer, the MM layer and a term that accounts for the coupling between these two layers. On the other hand, a subtractive scheme proceeds differently. Here the system's energy is obtained by firstly determining the energy of the whole system at a MM theoretical level ($E_{MM,Full\ System}$), followed by the calculation at a QM and MM level of the reactive system ($E_{QM,Reactive\,system}$ and $E_{MM,Reactive\,system}$, respectively), which usually is the active center or part of it. The MM contribution of the reactive system is then subtracted from the MM contribution of the whole system, thus resulting in the total system energy:

$$E_{QM/MM,Full\ System} = E_{MM,Full\ System} - E_{MM,Reactive\,system} + E_{QM,Reactive\,system}$$

Equation 11. Subtractive QM/MM scheme equation

In most cases, these subtractive schemes prove to be advantageous over the additive approaches as they do not require a parametrized expression to calculate the energy of the interlayer region, and the systematic errors in the handling of the reactive layer by MM are canceled out. Nonetheless, one of its major drawbacks is that its results are better when the MM layer is mostly non-polar, as otherwise it may have a significant influence on the electronic properties of the QM layer. Moreover, such influence cannot be accounted for as MM electrostatics are unable to polarize the QM region.

ONIOM, a subtractive QM/MM method, has solved this problem by incorporating an electrostatic embedding scheme.[96,106] The intricacies of the region between both theory levels are described in the following sections.

2.1.5.2. Boundary schemes

The electrostatic interactions between the QM and the MM regions can be delt by either a mechanical embedding approach, in which the electrostatic interactions of such region are only handled at an MM level, or through an electrostatic embedding approach. The latter incorporates the electrostatic interactions of the MM region into the QM equation.[96,106]

2.1.5.2.1. Mechanical embedding

Mechanical embedding is a simpler scheme to account for the effects of the MM environment in the QM/MM calculations. Its simplicity is attributed to the neglecting of the electrostatic effects by the surrounding environment on the QM calculations, which means that the surroundings do not polarize the QM region. Hence, in this scheme, the QM calculation only considers the structural constraints that the MM layer imposes on it. For MM calculations, a fix set of point charges is attributed to every atom in the reactive system.

As in this embedding scheme the boundary interactions are delt by molecular mechanics, one of its limitations is its need for parametrization of the atomic point charges of the reactive region. An additional difficulty is brought by the fact that along the reaction, the charge distribution in the reactive system suffers alterations, thus requiring often reparameterization for more accurate results. [96,106]

2.1.5.2.2. Electrostatic embedding

Electrostatic embedding schemes include the induced polarization of the QM region by the point charges of the MM layer. This is achieved by the addition of specific terms that are built from sets of parameterized MM point charges, into the Hamiltonian. Such addition provides, in theory, better calculation accuracy, as it portraits what occurs in the natural system. This scheme also has the added advantages of not requiring a set of derived charges for the high layer and not needing a charge update, as these charges can be derived from the wavefunction through an electrostatic surface potential computed from the minimized electronic wavefunction.

However, its higher accuracy comes at the expense of a significant increase in computational costs. Additionally, as force fields are not polarizable, the polarization of the MM region by the QM system is not accounted for, which despite being less relevant, can still affect the results to some extent.[96,106]

2.1.5.2.3. The link atom approach

When the QM and MM layers are connected by covalent bonds, which is common, some additional considerations are needed. Cutting through covalent bonds generates unpaired electrons at the boundary region, which can greatly influence the electronic structure of the QM layer. To avoid such complications, a common method (and the one used by ONIOM) is to employ a link atom approach. This entails the introduction of an atom, usually hydrogen, that is added at an appropriate position along the bond vector between both layers. The goal of such addition is to fill the atomic open valences that are generated by the bond cleavage between the QM and MM layers. These link atoms are only going to be accounted for by QM, as they are only required for the subtractive calculation on the reactive system.[96,106]

2.2. Computational methodology

2.2.1. Obtaining the DszA:FMN complex

As the crystallographic structure of DszA from *Rhodococcus erythropolis* (strain igts8) isn't available at the Protein Data Bank, it is necessary to obtain it through homology modeling. To do so, Swiss-Model is employed and the FASTA sequence from DszA (ID: P54995), available on Uniprot, a database of protein sequences, is used as the query sequence. The PDB ID 5XKD, corresponding to the three-dimensional structure of a BdsA:FMN complex from *Bacillus subtilis* WU-S2B, at a resolution of 2.4 Å, is used as the template. This structure is chosen as it is homologous to DszA, sharing a sequence identity of 79% with it. Moreover, as mentioned, DszA and BdsA catalyze the same catalytic step, they have the same function and belong to the same family of proteins.

The attained model is then superimposed with the BdsA structure in Pymol for comparison. DszA is modeled as a dimer, as research indicates that it is its active form.[59] To choose from which homodimer of BdsA, AB or CD, DszA should be modeled, a comparison between these dimers and the structures of other enzymes with similar biological functions is made by superimposing them on Pymol. The chosen enzymes for comparison, which are all dimers, are the following: nitrilotriacetate monooxygenase (PDB ID: 3SDO), alkane monooxygenase (PDB ID: 3B9N) and riboflavin lyase (PDB IDs: 5W4Y; 5W4Z; 5W48). The FMN cofactor is also modeled into the model structure based on its coordinates on BdsA. All these steps are performed on Pymol.

2.2.2. Protonation and parametrization of DszA:C4A

X-ray crystallography cannot resolve hydrogen atoms in most protein crystals. Consequently, the majority of available PDB files don’t show hydrogen atoms and these need to be added, after evaluating where they are required. To do so, information on pK_a and probable protonation states are obtained from Propka for both DszA and BdsA. Since they are closely related enzymes and BdsA has crystallographic waters, comparing the two enzymes can be a valuable way to perceive the influence of water on the protonation state. Careful consideration should be given to residues with pK_a near the physiological pH value (~7), where the protonation state can be uncertain. The residues where this is observed are located on Chain A: Asp59 (pKa = 5.83), Asp125 (pKa = 5.9), Glu86 (pKa = 6.04) and Glu64 (pKa = 6.57). Analysis of possible interactions of these residues showed that they should not need to be protonated at their side chains, hence remaining as deprotonated. Special attention should also be provided to histidine residues, due to their delocalized charge and tautomerization state, which will influence the position where protons should be added. Assignment of protonation states for such residues side chains' can be decided by inspecting the possible interactions between the residue under study and its surrounding residues, at a maximum distance of about 4 Å. This includes the verification of the interaction's nature, the bond/interaction distance and angle, and, for hydrogen bonds, the definition of which atom is the donor and the acceptor. Protonation of the side chains of histidine residues is predicted to be the following, where Hie entails a protonation at the Nε and Hid, protonation at the Nδ:

Table 1. Position of protons on Histidine

ε-protonated	His8, His20, His22, His27, His42, His100, His153, His156, His198, His202, His203, His305, His347, His354, His391, His440
δ-protonated	His116, His316

All the structure's His are replaced by either Hid or Hie, according to Table 1. Where doubts existed, (for His20 and His316), the protonation state is attributed accordingly to the chemical environment seen in BdsA. The other amino acid names remain unchanged as their protonation state corresponds to that expected at the physiological pH of 7. Addition of 6686 hydrogen atoms to the protein is achieved using the Xleap module from the AMBER18 package. The parametrization of DszA is carried out with ff14SB parameters, and the oxidized FMN is parameterized as flavin hydroperoxide (C4aOOH, which is from here on out referred to as C4A for simplicity reasons) with the parameters from Barbosa et al.[44] Figure 11 shows the parametrized C4A with the correspondent atoms name, that will be used in future references along this thesis. The system is then centered in a solvation box of 12.0 Å filled with TIP3P waters (where 26627 water molecules are added) and the total charge of the complex is neutralized with 46 Na^+ and 0 Cl^- counterions. Upon its parametrization, the DszA:C4A system is left with 890 residues and 93574 atoms.

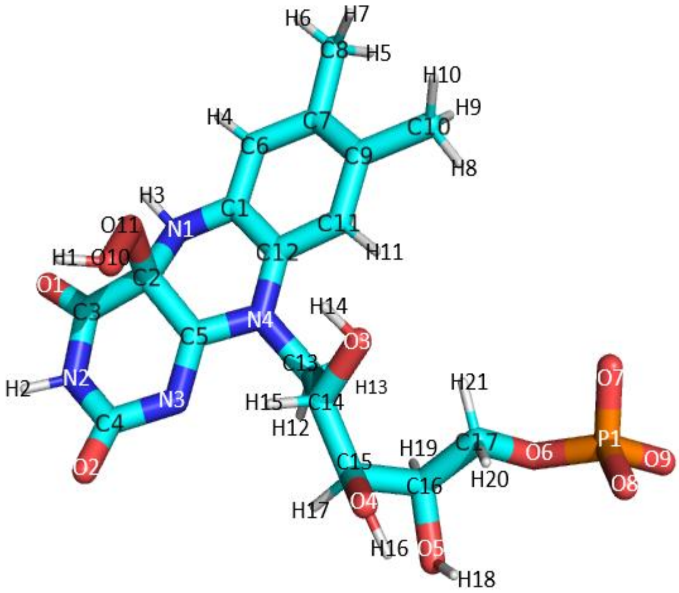


Figure 11. C4A labeled with atom names

2.2.3. Energy minimization of DszA:C4A

After the parametrization of the system, there is a necessity to optimize its initial conformation by energy minimization, to remove any clashes, bad torsions, or bad angles. This is performed with the Sander module, in a sequential process, from higher risk to lower risk of modeling step. The first step is a solvent minimization, followed by a 50 ps molecular dynamics simulation in an *NVT* ensemble designed as a solvent equilibration step. The second minimization step optimizes the side chains of DszA, and the third optimizes the active center's side chains. Finally, the system is minimized without any restrictions. All restraints are applied through an atom-centered harmonic potential with a force constant of 5.0 kcal.mol^{-1}.Å^{-2}. This energetically optimized DszA:C4A structure is then used in future docking studies.

2.2.4. MD simulations of DszA:C4A to validate the system

To validate the system that is used for docking, the minimized structure from "2.2.3. Energy minimization of DszA:C4A", undergoes MD simulations, followed by a clustering protocol and comparison with BdsA.

Equilibration of the system is performed in three sequential phases: heating, solvent equilibration, and final equilibration. This is followed by a production step. The heating is performed from 0 K to 300 K, at constant *NV* for 20 ps, followed by 30 ps in *NVT* (300K). Then follows a simulation in an *NPT* ensemble (1 bar and 300 K) for 2 ns, where the solvent is relaxed to its equilibrium density by restraining the movement of the solute. An *NPT* ensemble of 104 ns for the entire system is performed. The analysis of the resulting trajectories is done with VMD and Cpptraj to evaluate the Rmsd changes during the simulation, the histidine residues, the interactions between essential residues and C4A, and the number of waters in the active center.

To obtain representative MD conformations of the 100 ns production step, a geometric clustering with a *K*-means algorithm is performed, using the Rms of the active center as the clustering metric. On Cpptraj, 2, 3 and 4 clusters are attained for both Chains A and B. To assign the number of clusters that best represents each production run, the uniform distribution of frames per cluster, and the higher quality-value, psF, are taken into consideration. The active center of the dominant structure resulting from the clustering protocol is compared with the active center of BdsA, the template enzyme, in an attempt to validate the attained system.

2.2.5. Building $DBTO_2$

The substrate is built on Gaussview 5.0.8, and its optimal geometry is calculated by an energy minimization at the HF/6-31G* quantum mechanical level, using the Gaussian 09 software. To attain the atom charges necessary for the substrate parametrization, the molecular electrostatic surface potential for the optimized structure is computed from the electronic wavefunction optimized at the HF/6-31G* level. The restrained electrostatic potential method is then applied on the resulting ESP to derive atomic charges for the $DBTO_2$, and the remaining bonded and LJ parameters are derived from the gaff2 force field.

The representation of the substrate with its atom names after parametrization is seen in Figure 12. Such atom names will be used in future references along this thesis.

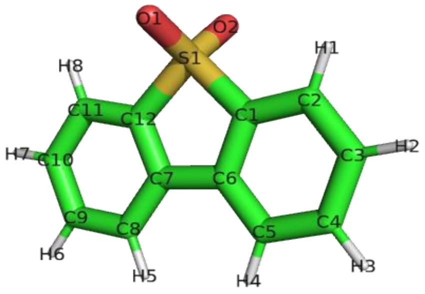


Figure 12. $DBTO_2$ with parametrized atom names

2.2.6. Docking of $DBTO_2$

The docking of the substrate into the minimized structure of DszA:C4A is an important requirement for the eventual study of the enzymes' mechanism of reaction. Since there are no known structures of DszA, or its homologs, in complex with the $DBTO_2$ substrate or any mimetic molecule, protein-ligand docking is performed on both AutoDock Vina and GOLD. Several docking procedures are employed for one of the monomers (Chain A) of our enzyme. Chain's B substrate is added later with the same pose as in the docking results for Chain A.

On GOLD, four docking procedures are employed, using the GA, a search space of 10 Å around the C4A cofactor, and the ChemPLP scoring function:

1. 20 unrestrained docking runs.
2. 20 docking runs using distance restraints, comprising the distance restraint of 3 to 5 Å between the proximal oxygen (O11) of the hydroperoxyl group of C4A and the topologically equivalent carbons (C1 and C12) bonded to the sulfur of $DBTO_2$, in agreement with the mechanistic proposal from Adak and Begley[61], and a spring constant of 5 N/m.
3. 10 unrestrained docking runs, where sidechain flexibility is allowed for the residues located at the *si*-face of the binding pocket: Ser10, Thr15, His16, Gln76, Asn137, Leu244, Thr308, Ser311, His312, Val367.

4. 10 docking runs, where the distance constraints and sidechain flexibility of procedures 2 and 3, respectively, are simultaneously applied.

To perform docking on AutoDock Vina, a rectangular box representing the search space that includes the active center of DszA, especially the *si*-face side, is created with the AutoDock Vina plugin on Pymol. This search space is used in two Vina docking procedures, each resulting in a maximum of 14 ligand poses:

1. Unrestrained docking.
2. A docking accounting for sidechain flexibility, where the flexible residues considered are the same as those of point 3 and 4 of the GOLD docking procedure: Ser10, Thr15, His16, Gln76, Asn137, Leu244, Thr308, Ser311, His312, Val367.

From all the resulting poses, the ones considered to be the most promising are chosen. The applied criteria for such selection are the score or ranking of the pose, a favorable distance and position to interact with the cofactor, recurrence in results from both software, and/or between different docking procedures within the same software, and the extent to which the substrate is buried in the active center. It should be important for $DBTO_2$ to be more buried at the active center, as it would be more protected from interactions with waters that would disrupt its presence in the active center.

2.2.7. Minimization and MD simulations of DszA:C4A:$DBTO_2$

After selection of the desired substrate poses, they are added to both chains of DszA on Pymol. The DszA:C4A:$DBTO_2$ systems are parametrized with Xleap as they were in section "2.2.2. Protonation and parametrization of DszA:C4A", but now also including the substrate parameters. The addition of hydrogens and counterions is also performed as previously, and the complex is centered in a 15 Å TIP3P water rectangular box. Energy minimization of the system with Sander is done following the same protocol as for the section "2.2.3. Energy minimization of DszA:C4A". Equilibration of the system follows a protocol similar to the one described in "2.2.4. MD simulations of DszA:C4A to validate the system": The heating is performed from 0 K to 300 K, at constant *NV* for 20 ps, followed by 30ps in NVT (300 K). Then follows a simulation in an *NPT* ensemble (1 bar and 300 K) for 2 ns, where the solvent is relaxed to its equilibrium density by restraining the movement of the solute, and a subsequent MD simulation of 10 ns is performed for the entire system. The production stage consists of a 100 ns *NPT* simulation. The analysis of the resulting trajectories is done with VMD and Cpptraj to inspect structural changes and to study the hydrogen bonds between essential residues and C4A and $DBTO_2$. To study possible hydrophobic contacts, the hydrophobic residues surrounding the ligands are analyzed with MDAnalysis, a module of Python. Analyzing all these results should enable the choice of the most promising DszA:C4A:$DBTO_2$ conformations to proceed for QM/MM calculations.

To obtain representative MD conformations of the 100 ns production step from the previously assembled DszA:C4A:$DBTO_2$ systems, geometric clustering with the same characteristics as the ones depicted in section “2.2.4. MD simulations of DszA:C4A to validate the system” is done. All the structures resulting from the clustering protocol are compared, and selection of an initial structure for QM/MM studies is done based on: their prevalence in both chains, interactions at the active center, the orientation of possible favorable residues to protect the active center from bulk water and cluster occupation.

2.2.8. System preparation for QM/MM studies

2.2.8.1. Layer assignment

To be studied by QM/MM methods, the DszA:C4A:$DBTO_2$ system needs to be divided into two regions, the one that is going to be treated with a QM level of theory, the high layer, and the one that is going to be treated with a MM theory of level, the low layer. Upon analysis of the selected conformations in the previous steps, the residues C4A, $DBTO_2$, Asp55, Ser135, His16, Phe369, Val367, and two water molecules are considered relevant to the chemical events at the active centers and, hence, are treated with QM methods. The number of atoms in this high layer is reduced by capping the mentioned residues in covalent bonds other than those in ring moieties, polar bonds and/or multiple covalent bonds. As Asp55 establishes hydrogen bonds with C4A through its backbone, it can only be capped at the carboxyl group of its sidechain, hence, its adjacent residues, Leu53, Pro54 and Gly56, are also included in the high layer. The final high layer comprises 137 atoms, a total charge of 0 and a singlet spin multiplicity. The depiction of how the high layer residues are cut is present in Figure 13. At each cut, hydrogen link atoms are automatically added until the valence is full. A protein truncation of 12 Å around the residues in this high layer results in a system with a total of 4299 atoms. Such truncation enables the reduction of computational efforts, without significantly compromising results accuracy, as long as the free residues shell within the high layer is over a 6 Å radius (here is 8 Å).[96,107] The mentioned residues are assigned to the high layer, the surrounding atoms into the low layer to be treated at the MM level of theory, and the coordinates of atoms on an outer shell of 4 Å are set as frozen, to avoid system expansion. The whole QM/MM model exhibits a total charge of -1. All the layer assignments and the introduction of link atoms are done in molUP[108], a VMD plugin.

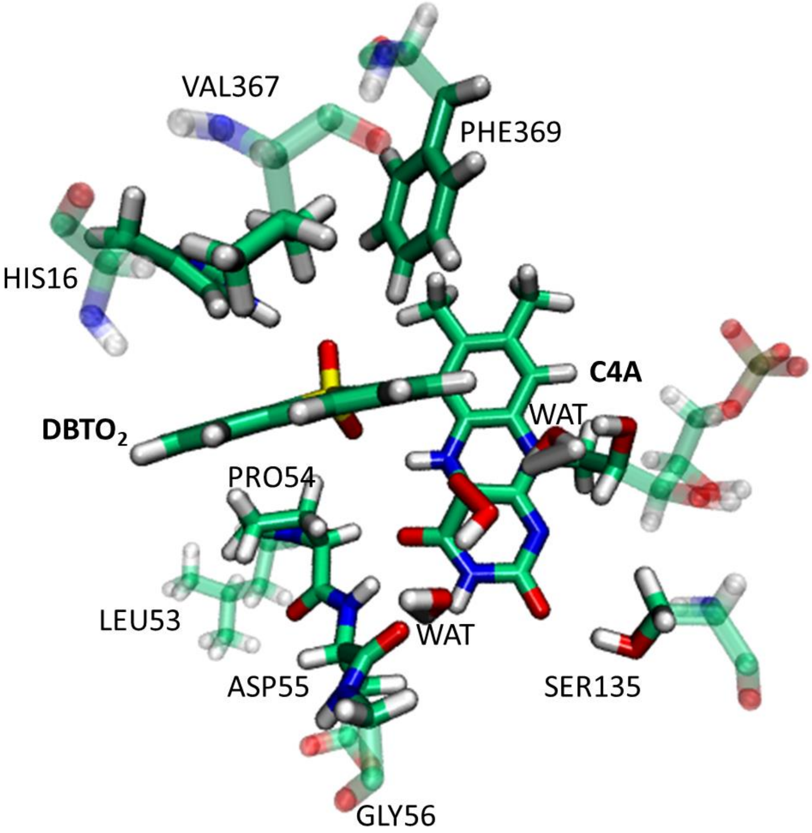


Figure 13. Residues assigned to the QM layer. Atoms represented as transparent are not included in the high layer.

2.2.8.2. Geometry optimization

The energy of the system is firstly minimized on Gaussian 09, using the ONIOM methodology and the mechanical embedding scheme. The high layer is treated with DFT, the exchange-correlation functional B3LYP, 6-31G(d) as basis set and a tight SCF convergence. The rest of the system is treated with the Amber parameters that arose from the "2.2.7. Minimization and MD simulations of DszA:C4A:DBTO2" step. The resultant optimized structure is then subsequently optimized with the electrostatic embedding scheme. For this step, all the waters of the MM layer are frozen using the molUP[108] plugin of VMD.

2.2.9. QM/MM studies

To understand the progress of a reaction from reactant to product, a linear transit scan is made. This allows the observation of how the energy and geometry of the QM/MM model varies along a tentative reaction coordinate, thus enabling the study of mechanistic hypotheses. Using the optimized structure with the electrostatic embedding as the initial structure, several atom pair distances are attempted as tentative reaction coordinates, including the distance between the distal oxygen of C4A and the C1 from $DBTO_2$ that is one of the main reaction coordinates in the first step of the mechanistic proposal by Adak and Begley[61].

The linear transit scans are performed with an increment of -0.12 Å along each reaction coordinate until an appropriate distance for bonding between the coordinates is reached.

More information and other attempted reaction coordinates are described in sections “3. Results and Discussion” and “6. Annexes”. Moreover, in the first, new initial structures for the scans are also depicted. These structures arise from the geometry optimization of relevant scan points, from the addition of residues to the QM layer and/or from MD simulations where distance restraints were once applied.

2.2.10. Other structural attempts

The MD simulations with distance constraints mentioned in the previous section are performed on the equilibrated systems (after the 10 ns unrestraint *NPT* ensemble) from section “2.2.7. Minimization and MD simulations of DszA:C4A:DBTO2”. Several different restraint simulations are performed, in order to search for new conformation and reaction hypotheses (which are described in “3. Results and Discussion”):

1. The distance between O10 of C4A to C1 of $DBTO_2$ is constraint to 3.5 Å, during a 10ns *NPT* MD simulation

2. The distance between O10 of C4A to C12 of $DBTO_2$ is constraint to 3.5 Å, during a 10ns *NPT* MD simulation
3. The distance between O10 of C4A to C1 of $DBTO_2$ is constraint to 3.5 Å, and the distance between H1 of C4A and NE2 of His312 is constraint to 2.1 Å during a 10ns *NPT* MD simulation
4. The distance between O10 of C4A to C12 of $DBTO_2$ is constraint to 3.5 Å, and the distance between H1 of C4A and NE2 of His312 is constraint to 2.1 Å during a 10ns *NPT* MD simulation
5. The distance between O11 of C4A to C1 of $DBTO_2$ is constraint to 3.5 Å, during a 10ns *NPT* MD simulation
6. The distance between O11 of C4A to C12 of $DBTO_2$ is constraint to 3.5 Å, during a 10ns *NPT* MD simulation
7. The distance between H1 of C4A and NE2 of His312 is constraint to 2.1 Å during a 20ns *NPT* MD simulation

The resulting trajectories are observed to ensure that the restraints are successfully applied. To see if these hold up when removed, a 10 ns unrestrained *NPT* MD simulation is done for each of the seven procedures. Analysis of the previously restraint distances and of hydrogen bonds is performed with Cpptraj.

In another attempt, to try to obtain the most similar substrate pose to that modeled in BdsA by Su et al.[58], the chain A from the DszA:FMN system (from section "2.2.1. Obtaining the DszA:FMN complex") is used for a free ligand:protein docking on AutoDock Vina. The pose that is most similar to the one obtained from the docking simulations of Su et al.[58] is selected and added, in Pymol, to both chains of the DszA:FMN system. The DszA:FMN:$DBTO_2$ system then undergoes the same parametrization protocol as in "2.2.2. Protonation and parametrization of DszA:C4A", but with a water model of 15.0 Å (as in previous attempts with a 12.0 Å water model the protein became partially unsolvated as it moved during MD simulations). The system's energy is minimized as in "2.2.3. Energy minimization of DszA:C4A". The equilibration phase is achieved accordingly to "2.2.7. Minimization and MD simulations of DszA:C4A:DBTO2" and the production step is a 141 ns *NPT* MD simulation. Analysis of the trajectories and clustering analysis follows the same procedure as in "2.2.7. Minimization and MD simulations of DszA:C4A:DBTO2" and "2.2.4. MD simulations of DszA:C4A to validate the system", respectively. These conformations are compared on VMD with the structure firstly employed for the QM/MM studies. Such comparison is helpful to see if the initial system in use is similar to the BdsA's system, where mechanistic hypotheses (by Su et al.[58]) that match the Adak and Begley[61] hypotheses are tested.

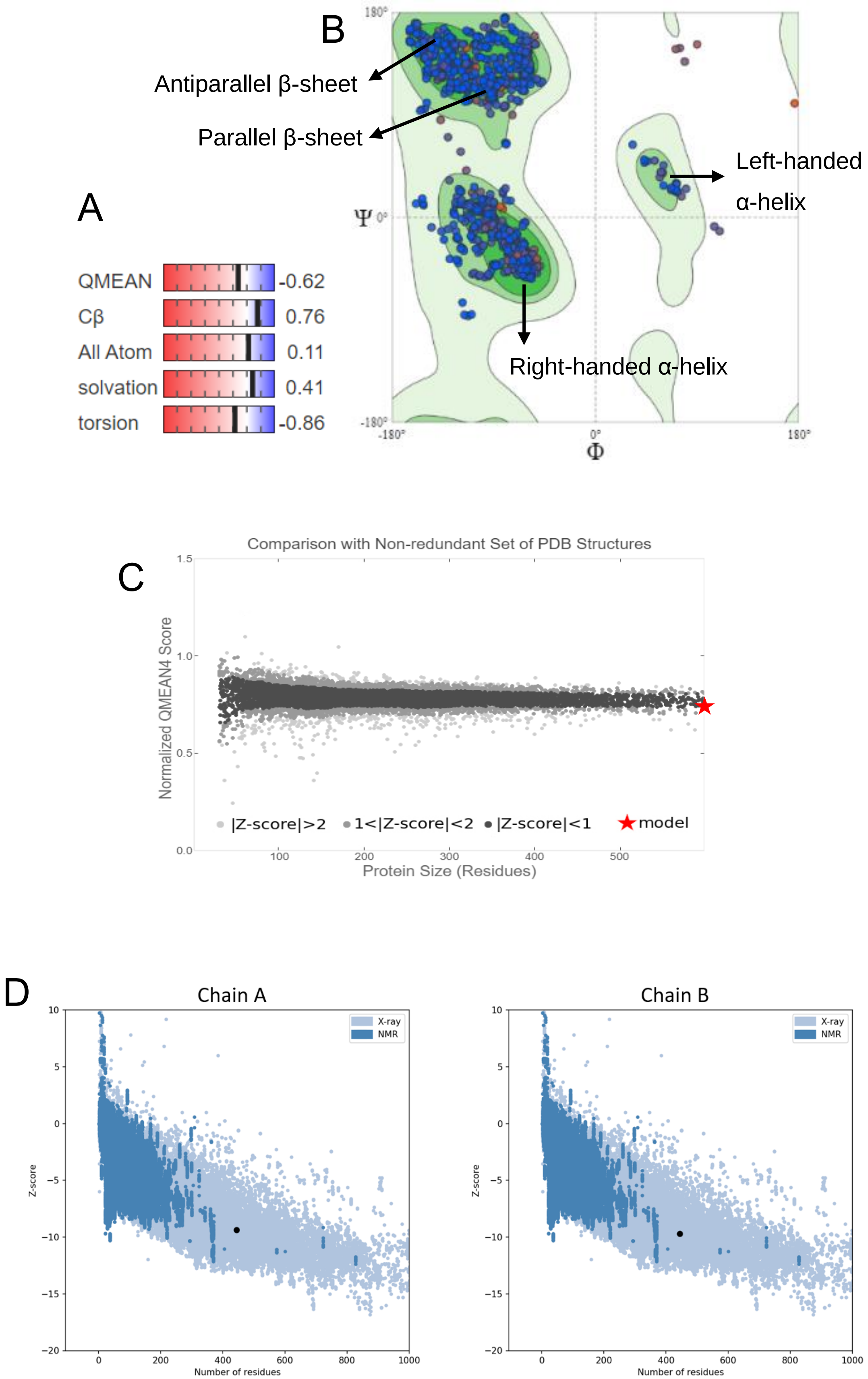
3. Results and Discussion

3.1. Obtaining DszA:C4A

Obtaining the DszA:C4A system involves several steps: production of a quality model of the protein through homology modeling; evaluation of the quality of such model; rational structural transformations; system parametrization; and finally, system validation. All these steps, described in detail in the following subsections, are required so the DszA:C4A system can then be confidently employed in further steps of this thesis.

3.1.1 Homology modeling of DszA:FMN

From Swiss-model, the DszA model is obtained. It is a tetramer with 13772 atoms. Its QMEAN value is -0.62, with a GMQE of 0.91 and a sequence identity of 79.28% in relation to BdsA. Moreover, the Ramachandran plot for such model demonstrates that its secondary structure is comparable to the ones commonly found in nature, with a percentage of outliers of only 0.40% (Val371 of chain A, B and D, Gln338 of chain D, Ser355 of chain A and B, and His354 of chain D). As for the comparison of the model to other naturally existing structures, the QMean Z-Score is less than one, thus falling into the most populated structure area, demonstrating its likelihood of existing in nature. Nonetheless, as it is a tetramer, it falls into the right extremity of the graph in Figure 14C, where there are fewer structures to be compared with due to its size. To further demonstrate that the attained model is similar to other structures existent in nature, each chain is compared to experimentally determined structures in ProSA-web[109]. From such comparison, as seen in Figure 14D, all chains are comparable with existent structures, giving extra credibility to the model. All these good quality results indicate that the obtained model should be of high quality, thus representing a reliable DszA structure. The quality values, Ramachandran plot and the graphs that compare the model to a non-redundant set of structures are depicted in Figure 14.


A
QMEAN -0.62
Cβ 0.76
All Atom 0.11
solvation 0.41
torsion -0.86
B
Antiparallel β-sheet
Parallel β-sheet
Left-handed α-helix
Right-handed α-helix
180°
Ψ 0°
-180°
-180°
0°
180°
Φ
C
Comparison with Non-redundant Set of PDB Structures
Normalized QMEAN4 Score
1.5
1.0
0.5
0.0
100
200
300
400
500
Protein Size (Residues)
|Z-score|>2
1<|Z-score|<2
|Z-score|<1
model
D
Chain A
Chain B
X-ray
NMR
Z-score
10
5
0
−5
−10
−15
−20
0
200
400
600
800
1000
Number of residues

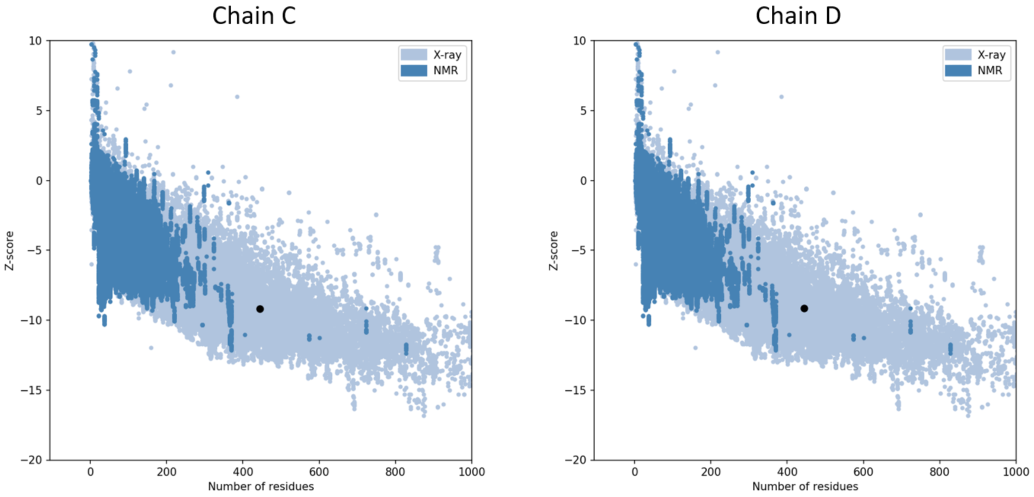


Figure 14. A) Quality values for DszA model, where the bluer, the better quality it represents; B) Ramachandran plot for DszA model; C) Graph for Comparison of DszA model with non-redundant set of PDB Structures; D) Graphs (attained from ProSA-web) for comparison of chain A, B, C, and D of DszA model with set of X-ray and NMR structures.

The superimposition of the model DszA with its template, BdsA, gives an Rmsd value of 0.134 Å, which indicates that the structures are very similar. The residues at the active center of BdsA that, according to Su *et al*., are essential residues for BdsA activity are Phe12, Phe56, Phe246, His20, His316, Val248 and Val372. Observing the active centers from both systems on Pymol (seen in Figure 15), it is noticeable that the position and orientation of the mentioned residues on the active center of the model and of BdsA are aligned and indeed very similar. Nonetheless, there is a residue deletion, as Val372 of BdsA is correspondent to Val371 of the model, and Val248 is substituted by Leu248. However, this latter substitution should not be a problem as both residues have similar sizes and hydrophobic characteristics. Thus, since BdsA and DszA are similar, these residues are also probably essential for DszA activity. Nonetheless, the vast resemblances between the model and FMN-bounded BdsA, further indicate that the attained model is in fact trustworthy and of quality.

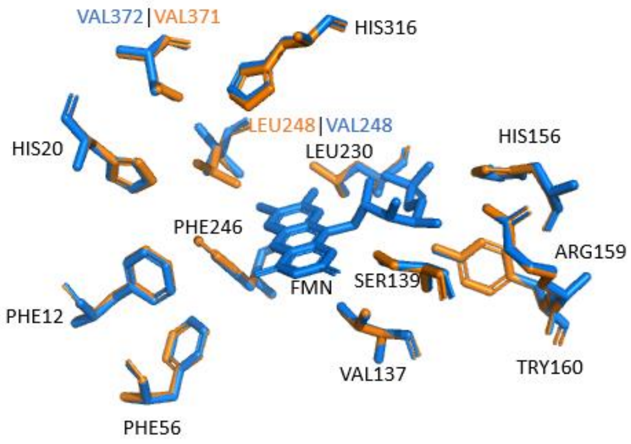


Figure 15. Active center of BdsA (blue) vs active center of model DszA (orange).

Due to their structural similarities, the molecular modeling of the oxidized FMN from BdsA to DszA is a fairly low-risk process that is achieved on Pymol. Moreover, the location for the substrate seen in BdsA (Figure 6A), the so-called *si*-face, should also be the same for DszA. This is also corroborated by observation of the model DszA on Pymol, where it is possible to notice that there is a cavity at this location that should be able to accommodate the DBT sulfone. Moreover, it is also expected for $DBTO_2$ to establish hydrogen bonds with the surrounding residues. Taking this into consideration and observing the DszA structure on Pymol, it can be said that, possibly, the donor residues to the DBT sulfone are the cofactor FMN, His316 (protonated at Nδ) and His20 (protonated at Nε).

Comparison between the homodimers of BdsA with nitrilotriacetate monooxygenase (PDB ID: 3sdo), alkane monooxygenase (PDB ID: 3b9n) and riboflavin lyase (PDB IDs: 5w4y; 5w4z; 5w48), shows that there is no substantial difference in using either dimer AB or dimer CD from BdsA. (Rmsd values for the comparison are depicted in Table 2). Hence, the transformation of DszA into a dimer occurs based on the AB dimer of BdsA. The system DszA:FMN as a dimer, now with 6940 atoms, is finally obtained.

Table 2. Rmsd values for enzymes similar to BdsA and chain AB or CD of BdsA.

PDB ID	5W4Y	3SDO	3B9N	5W48	5W4Z
Dimer AB of BdsA (PDB ID: 5XKD)	1.730 Å	1.696 Å	1.613 Å	1.830 Å	0.788 Å
Dimer CD of BdsA (PDB ID: 5XKD)	1.641 Å	1.726 Å	1.615 Å	1.835 Å	1.730 Å

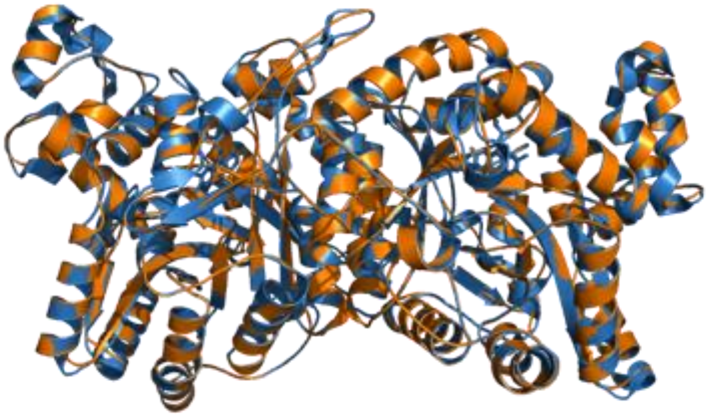


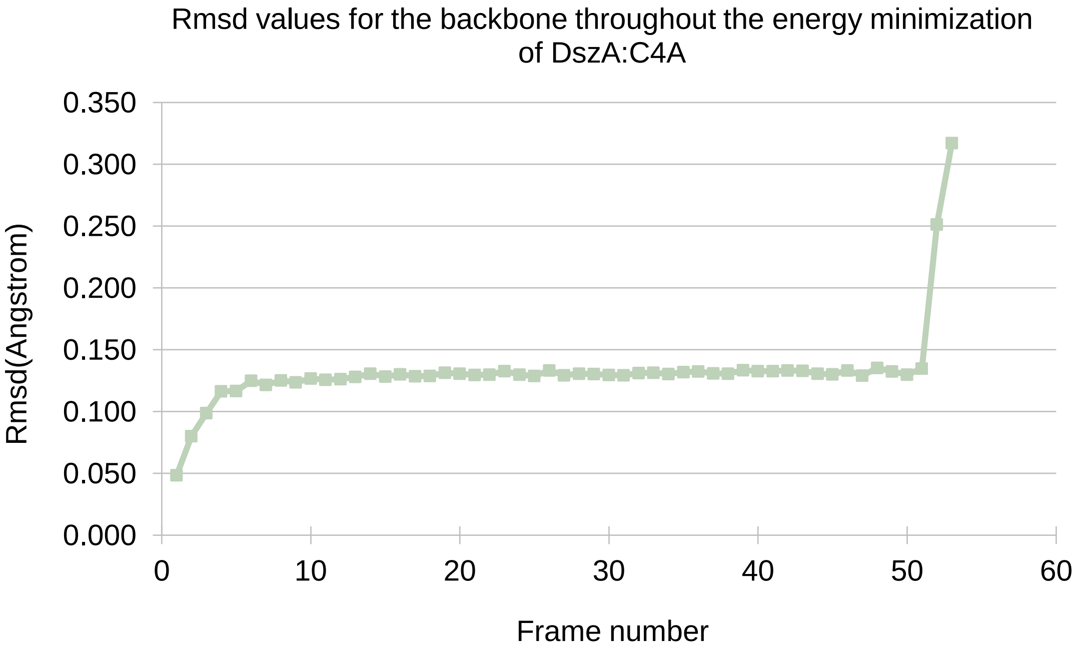
Figure 16. Superimposition of a dimer of BdsA:FMN (blue) with DszA:FMN (orange).

The DszA:FMN system is then parametrized. Hydrogen atoms complete the open valences and counterions neutralize the system’s charge. A solvation box is added and the cofactor FMN is transformed into C4A, by addition of a hydroperoxyl group, attaining the DszA:C4A system. Nonetheless, such system still needs to undergo a geometry optimization (minimizing its energy) and further MD simulations to validate it.

3.1.2. Validation of DszA:C4A

Energy minimization of the DszA:C4A system is achieved. Comparison of the unminimized structure with the minimizations and solvent equilibrium frames gives an average Rmsd of 0.133 Å for the backbone of the DszA:C4A complex. The maximum Rmsd value, which corresponds to comparison with the final minimization step, is 0.317

Å. The evolution of Rmsd values throughout the energy minimization protocol for the backbone can be seen in Graph 1.



Graph 1. Rmsd values for the backbone throughout the energy minimization, where frame 1 is the unminimized system, frame 2 the solvent minimization, frame 2 to 52 the solvent equilibrium, frame 53 the sidechains minimization, frame 54 the active center minimization and frame 55 the entire system's minimization.

Comparison of the active center of the unminimized structure *vs.* minimized is depicted in Figure 17 (note that residue numeration has decreased in 4 units when the system was parametrized; hence, for example, Ser139 is now Ser135). From Figure 17, it is also possible to observe that the structure from the unminimized active center *vs.* the minimized active center is similar, as the position and orientation of the residues are maintained. Still, although not substantial, the major differences seen between the systems are on Arg155 and on the sidechain of Leu244.

As the structure does not go through major changes, the residues taken as essential for BdsA activity and the interactions that are seen at its active center are still expected to be the same for DszA and its active center. Interactions of the cofactor with the protein are discussed later.

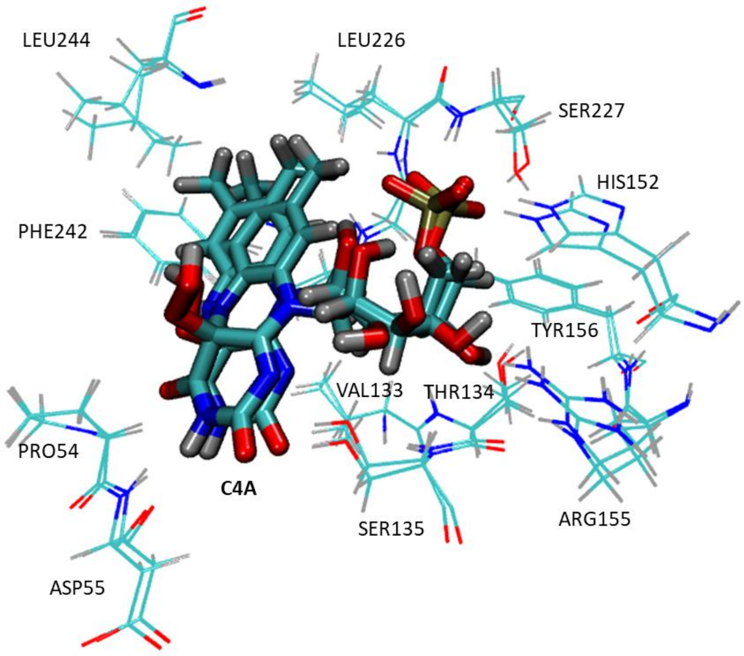
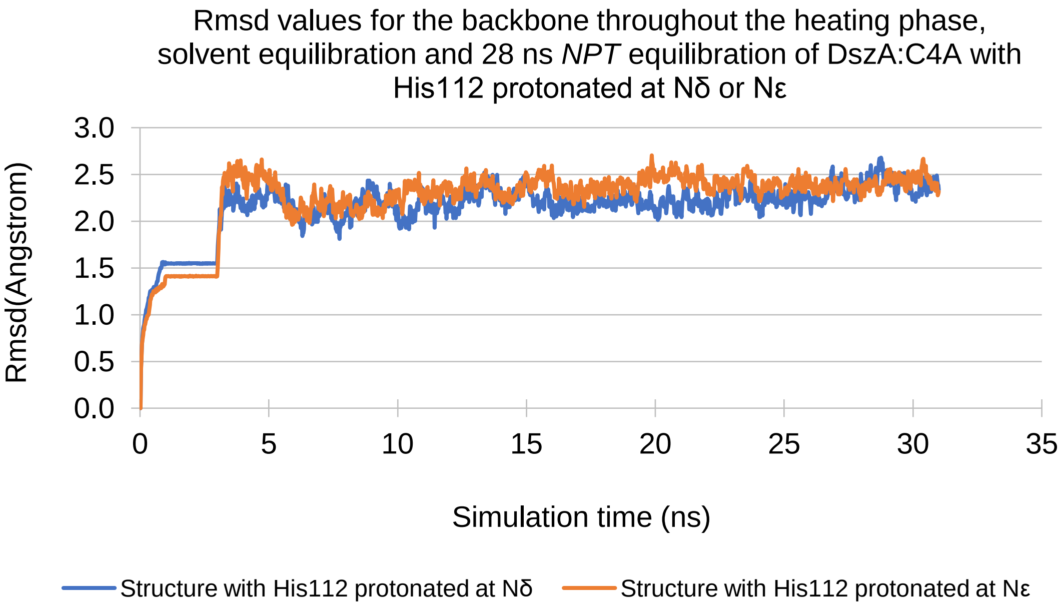


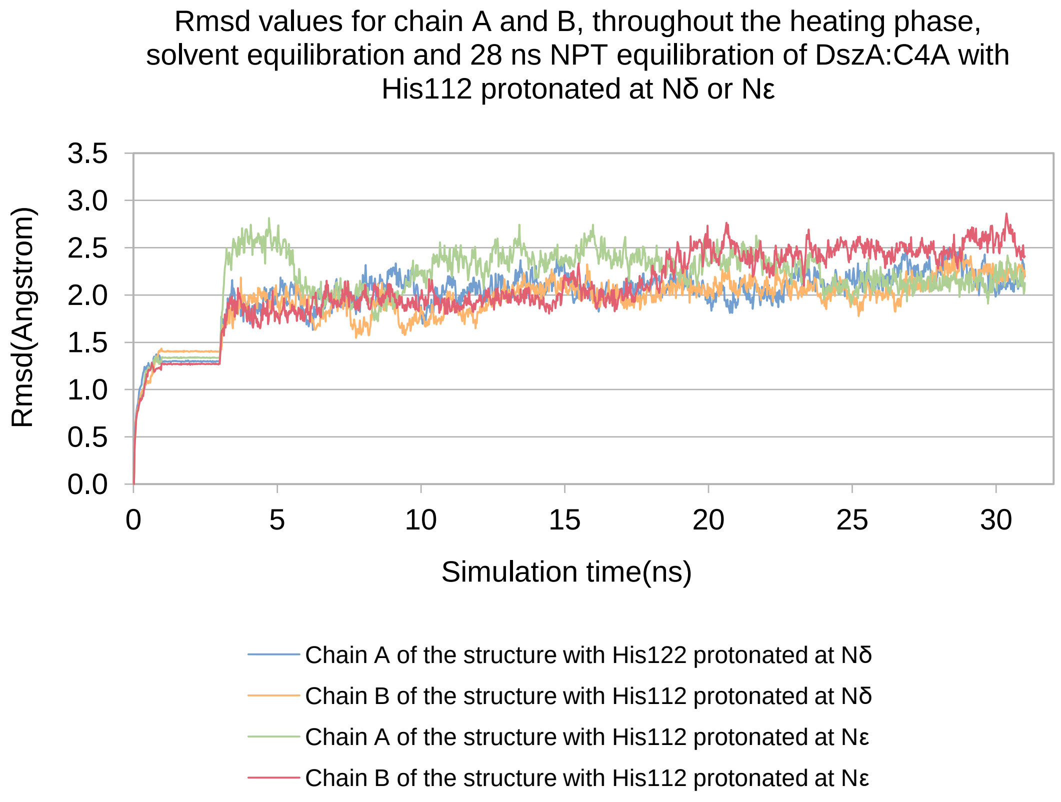
Figure 17. Active center of DszA:C4A unminimized vs minimized.

The optimized DszA:C4A system is then equilibrated in three sequential phases: heating, solvent equilibration, and a final 28 ns *NPT* equilibration. The resulting structures are analyzed, with special attention to the changes in the overall backbone, on dimerization, on histidine residues and on the interactions between essential residues and C4A. An initial analysis, done on VMD, for the mentioned steps shows no major differences on histidine residues, except for His112 (protonated at Nδ) as its ring flips to enable interaction of the donor atom with an oxygen atom that belongs to Glu526. This could be an indication that the system favors a hydrogen bond between these residues. Moreover, this prompts us to believe that an His112 protonated at Nε could be more suitable as the interaction with Glu526 could be formed only with a minor conformational change. This alternative is tested before deciding what system should proceed to the production step. The system with His112 protonated at Nε is parametrized, minimized, and equilibrated following the same procedures as for the system with His112, protonated at Nδ. Observation, on VMD, of the resulting trajectories of the structure with His112 protonated at Nε no longer shows, as expected, this histidine rotation to enable the interaction with Glu526. On Graph 2, the evolution of the Rmsd of the backbone is seen for both systems. Through its analysis is possible to see that, overall, they behave similarly.



Graph 2. Rmsd values for the backbone of DszA system with His112 δ- and ε-protonated, throughout the heating phase (depicted from 0 ns until the Rmsd is about 1.350 Å), the solvent equilibration (from the end of the heating phase until the end of the plateau) and equilibration of the entire system (from the plateau until the end).

Comparison of the Rmsd progress for chain A and B is also done for the two systems and results are depicted in Graph 3. Observation of this graph demonstrates that chain A and B of the structure with His112 protonated at Nδ have a more homogeneous behavior between chains than the one seen for the structure with His112 protonated at Nε. Moreover, upon observation on VMD the major differences between the chains of the system with histidine protonated at Nε are located on alpha-helices and loops on the extremity of the enzyme. As changes on loops are somewhat expected due to the region flexibility, the same cannot be said for movements of alpha-helices, which could have a greater impact on the enzyme.



Graph 3. Rmsd values for chain A and B of DszA system with His112 δ- and ε-protonated, throughout the heating phase (depicted from 0 ns until the Rmsd is about 1.350 Å), the solvent equilibration (from the end of the heating phase until the end of the plateau) and equilibration of the entire system (from the plateau until the end).

As for interactions, Su *et al.* describes that the BdsA cofactor, FMN, is held by hydrogen bonds with residues Ser139, His156, Arg159, Tyr160 and Leu230, and by a hydrophobic interaction with Val137. After the renumeration of the residues on DszA, these correspond to Ser135, His152, Arg155, Tyr156, Leu226 and Val133, respectively. The cofactor of the structure with His112 ε-protonated interacts with Ser227, His152, Tyr156, Thr134, Ser227, Leu226, and Asp55 on chain A and with Ser227, His152, Tyr156, Arg155, Ser227, Leu226, Ser135, Asp55 on chain B. This indicates that the active centers are not exactly the same for both monomers and that some interactions that are necessary to hold the BdsA cofactor are here lacking. Moreover, the missing interactions are not replaced with interactions with other residues. As for the structure with His112 δ-protonated, the existing hydrogen bonds between C4A and DszA can be observed in Figure 18. Despite only one chain being represented (chain A) in Figure 18, the hydrogen bonds seen are conserved in both chain A and B. Upon observation of Figure 18 it is noticeable that the ribitol-phosphate tail of C4A interacts with the sidechain of Ser227, His152, Tyr156, Arg155 and Thr134, and the backbone of Ser227 and Leu226. Moreover, the cofactor's ring has a hydrogen bond with the sidechain of Ser135 and two with the backbone of Asp55. Additionally, Val133 seems likely to interact hydrophobically with the isoalloxazine ring of C4A. Comparison between the interactions seen for the structure with His112 (protonated at Nδ) for DszA:C4A to the ones seen for BdsA:FMN, shows that those that exist for the latter are also present in the DszA structure with His112 protonated at Nδ. Furthermore, additional interactions with Asp55, Thr134 and Ser227 are seen.

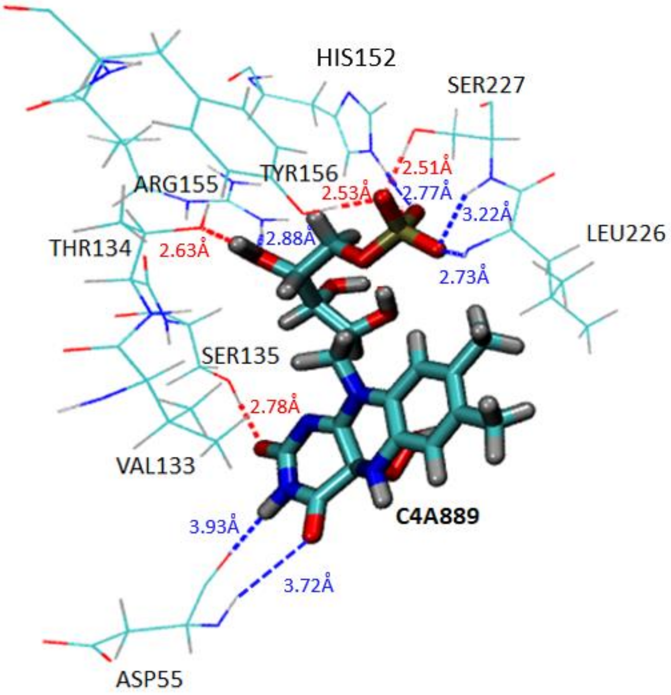
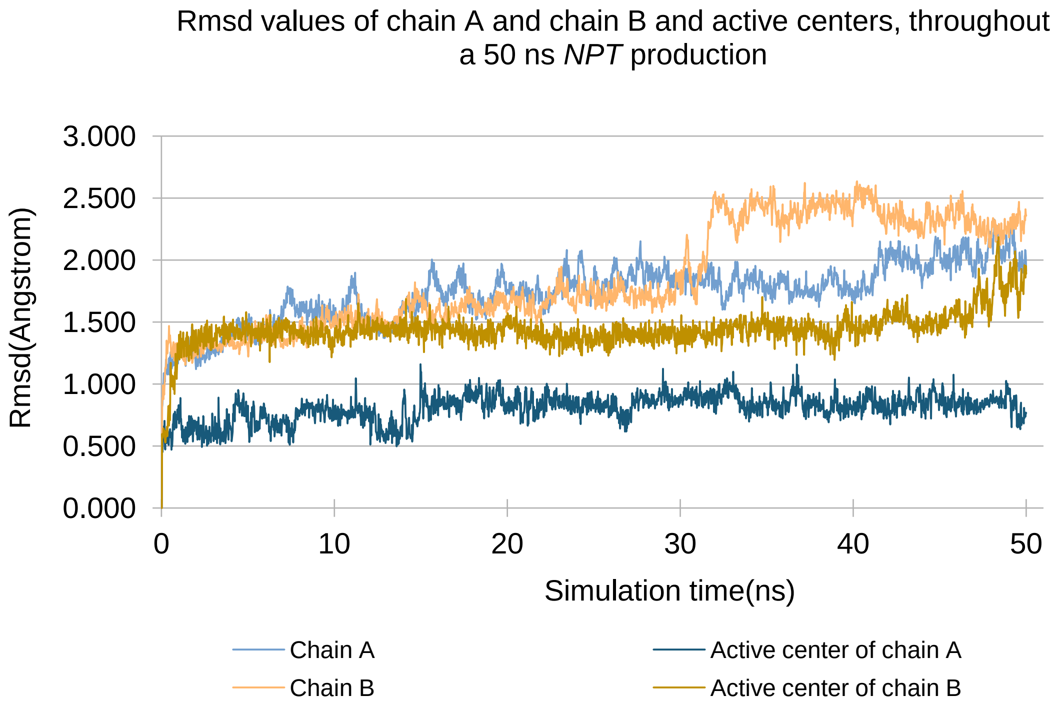


Figure 18. Interactions between C4A and the protein DszA on chain A. Hydrogen bonds marked in blue identify N as the donor atom and bonds marked in red identify O as the donor atom.

Despite the similarities between both systems when it comes to their overall Rmsd values, when comparing the Rmsd for each chain separately, chain A and B are more similar for the structure with His112 δ-protonated. Additionally, more interactions and similarities are seen at the active centers of the structure with this structure than at the structure with His112 ε-protonated. For these reasons, the system's production is done with His112 protonated at Nδ, even though this represents a larger conformational change for this residue. Hence, the results for this system are now discussed.

A 50 ns production for the system with His112 protonated at Nδ is analyzed. Assessing the Rmsd plot for chain A and B (Graph 4), an Rmsd peak is seen for chain

B at around 31 ns. On VMD, it is visible that this difference between chains is a result of a terminal loop, and for that reason, it shouldn't be a concern as it won't have much of an influence on the enzyme's activity. Nonetheless, in the last 20ns of simulation, a Rmsd value of around 2.3 Å for the entire structure is seen, maintaining its value from the equilibration phase. Graph 4 also depicts the active center of chain A and chain B. In this graph it is possible to observe that the Rmsd values of the active centers at equilibrium are different, which might result in some different interactions with the cofactor, as corroborated next.



Graph 4. Rmsd values for chain A and B and active centers, throughout the 50 ns NPT production.

Analyzes of the cofactor's hydrogen bonds with DszA are made. At the active center of chain A, the interaction of Asp55 with the C4A's ring is maintained for the majority of frames (2376/2500). However, the ring interaction with Ser135, is not as present as it is only observed in 1534 frames out of 2500 frames. Nonetheless, this isn't necessarily bad. In fact, the proposed reaction mechanism for BdsA in Su *et al*. states that the interaction of Ser135 with N of the reduced FMN ring can lead to the deprotonation of $FMNH_2$, prior to the oxygen activation mechanism. So, the fact that this serine is sometimes interacting with N of C4A, could be a sign that this process is likely to activate C4A. The interactions with the other parts of C4A are more transient, even though, the Thr134 interaction with C4A's ribitol and the interaction between His152 and the C4A's phosphate are stable. In spite of that, this intermittence is not a concern as it is located on flexible parts of the cofactor, and it is also noticeable that when these interactions disappear, new similar interactions are formed with the same region of C4A. On the active center of chain B, more interactions are maintained, especially the ones with the phosphate group. The Asp interaction with the ring of the cofactor is seen in the majority of the frames (2205 out of 2500). Additionally, the ring interaction with Ser135 is more present on this chain as it can be seen on a total of 1800 frames. Overall, the results indicate that the hydrogen bonds seen for the equilibrium phases are maintained over the system's MD simulation. Only the interaction between the ribitol of C4A with Arg155 seems to get weaker. This is an indicator that Arg155 may not be crucial to maintain the position of the cofactor in the active center, which makes sense since this interaction is on the ribitol tail of C4A, a mobile region of the cofactor. The contrary can be said for the most stable interactions as these are probably important for C4A binding.

To attain a representative structure, geometric clustering, using the active centers' rms as clustering metric, is performed for a 100 ns *NPT* MD simulation. Three clusters for each chain are attained. To further validate the attained DszA:C4A system, the active centers of the dominant clusters are superimposed with the active centers of BdsA. Superimposition of the active centers of chain A (from the cluster and from BdsA:FMN (PDB ID: 5xkd)) gives an Rmsd of 0.691 Å and superimposition of the active centers of chain B (from the cluster of chain B and chain B of BdsA:FMN (PDB ID: 5xkd) yields an Rmsd of 0.546 Å. The aligned active centers are depicted in Figure 19. Moreover, for chain A the most prominent differences are seen for Leu226, Ser135 and their equivalents in BdsA (Leu230 and Ser139, respectively). Some smaller differences are also seen between His156, Asp55, C4A and their analogous: His152, Asp59 and FMN, respectively. For chain B the most notorious changes are the same as seen for chain A, as residues Leu226 and Ser135 are the ones that most differ from their equivalents in BdsA. Additionally, His152 presents a slight variation when compared with the reference residue in the template enzyme. Nevertheless, none of these modifications hinders the formation of hydrogen bonds and thus, they do not compromise the validity of the attained system. Moreover, some of the alterations seem to occur so the residues could be better orientated to established hydrogen bonds, which was probably necessary as the addition of the hydroperoxyl group to the cofactor provoked slight changes at the active center.

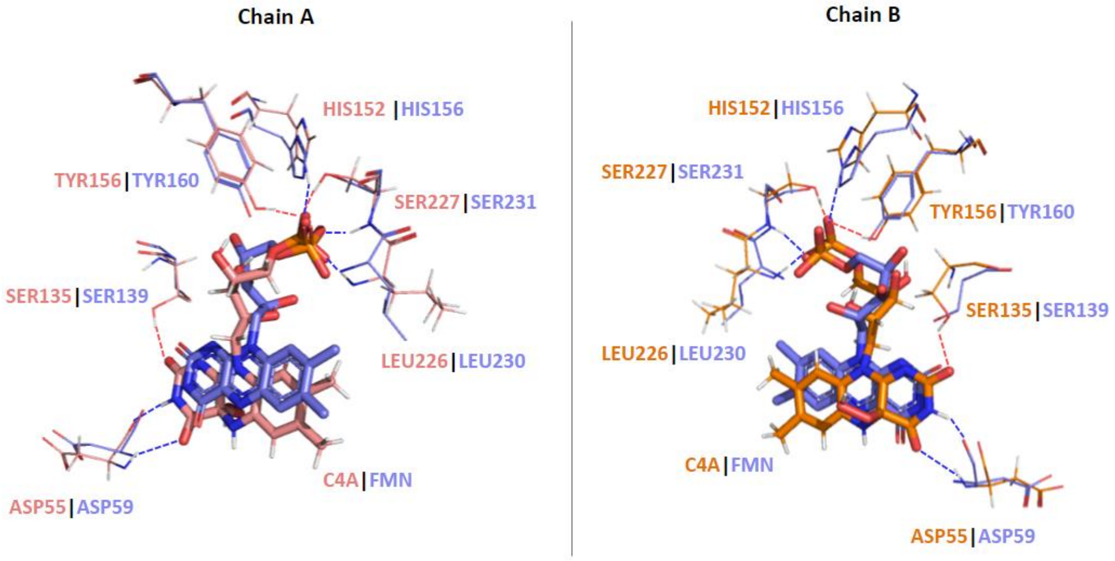


Figure 19. Superimposition of the active centers from the dominant cluster of the active centers of chain A and B from the DszA:C4A system (colored in pink for chain A and in orange for chain B), with the active centers of chain A and B from the BdsA:FMN system (colored in purple). Not all residues that establish hydrogen bonds with the cofactor are depicted for image clarity reasons. Dotted lines in blue show a hydrogen bond where N is the donor atom and red dotted lines represent a hydrogen bond where O is the donor atom.

As is shown along the last subsections, a DszA:C4A system that is optimized, equilibrated, and simulated along time, is attained. Through all the steps the system has shown to be of quality as the initial model obtained from homology modeling should be reliable and upon energy minimizations and MD simulations, it achieved its density equilibrium and maintained the interactions though to be important for cofactor binding. Moreover, comparison with the crystallographic structure of its homolog, BdsA, shows that no impactful structural modifications occur at the active center, which further increased the viability of the obtained DszA:C4A system. Hence, the system can be considered as validated and reliable to be employed in further studies, such as the docking of the substrate, $DBTO_2$, which is discussed in the next section.

3.2. Obtaining DszA:C4A:$DBTO_2$

Now that a validated DszA:C4A system exists, to eventually study mechanistic hypotheses there is still one requirement missing: the docking of the substrate at the active center. Thus, $DBTO_2$ is built, parametrized and then protein:ligand docking occurs to attain the DszA:C4A:$DBTO_2$ system. Afterward, just as before, the new system, or in this case systems, undergo through steps of energy minimizations and MD simulations, so they can be analyzed and validated to proceed into QM/MM studies.

3.2.1 Docking of $DBTO_2$

$DBTO_2$ is built and its structure undergoes a geometry optimization at the HF/6-31G* quantum mechanical level, which is completed in 5 optimization steps. Its atomic charges upon parameterization with the antechamber module of AMBER18 and the gaff2 force-field can be seen in Table 3.

Table 3.Atomic charges of $DBTO_2$

Atom name	Atom type	Charge (A.U)
O1	o	-0.571507
S1	sy	1.04999
O2	o	-0.571507
C1	ca	-0.129437
C2	ca	-0.111185
H1	ha	0.188626
C3	ca	-0.153137
H2	ha	0.147687
C4	ca	-0.08279
H3	ha	0.141729
C5	ca	-0.20418
H4	ha	0.156811
C6	cp	0.092389
C7	cp	0.092389
C8	ca	-0.20418
H5	ha	0.156811
C9	ca	-0.08279
H6	ha	0.141729
C10	ca	-0.153137
H7	ha	0.147687

C11	ca	-0.111185
H8	ha	0.188626
C12	ca	-0.129437

The modeled substrate is then used to obtain several poses from the different docking protocols. From the 20 GA runs of the unrestraint docking on GOLD, 20 poses are attained, from which only three are considered to be distinct. The same outcome is seen for the GOLD distance constraint docking. GOLD docking with side chain flexibility yields 10 results, from which only four are different docking poses. In the last GOLD docking protocol (with side chain flexibility and distance constraint), from the 10 attained poses, 3 different poses are seen. Hence, in total GOLD produced 13 docking poses to be analyzed.

From AutoDock Vina docking protocols, the unrestraint docking gives 14 poses, from which 10 are different. In the docking with side chain flexibility, from the 14 attained poses, 9 represent distinct docking positions. Hence, a total of 19 poses are analyzed.

Comparing all the 32 docking poses, and applying selection criteria that includes the pose ranking, the favorable distance (distance between O11 of C4A and C1/C12 of DBTO$_2$ between 3 and 5 Å) and position for interaction with C4A, its recurrence in both software and/or in several docking protocols, and its accessibility to solvent, three most promising poses, depicted in Figure 20, are chosen. The pose in Figure 20A results from a GOLD distance restraint docking and it is chosen as it is a result from both docking programs, it has a favorable distance between the hydroperoxyl group of C4A and the C1 of DBTO$_2$ (3.60 Å) and it is buried at the active center. This DszA:C4A:DBTO$_2$ system is from now on denoted as conformation A. Figure 20B shows the pose attained from the unrestraint AutoDock Vina docking. Its selection is due to its presence in several AutoDock Vina runs from protocols with and without restraints, its favorable distance between the hydroperoxyl group of C4A and the C1 of DBTO$_2$ (3.53 Å), and it is also buried at the active center. Such system will be named conformation B. At last, Figure 20C, depicts the pose attained from the GOLD protocol with distance constraints and sidechain flexibility. This pose is buried at the active center and it also has the favorable distance seen for the previously mention atoms (4.38 Å). This latter system is designated as conformation C along this thesis.

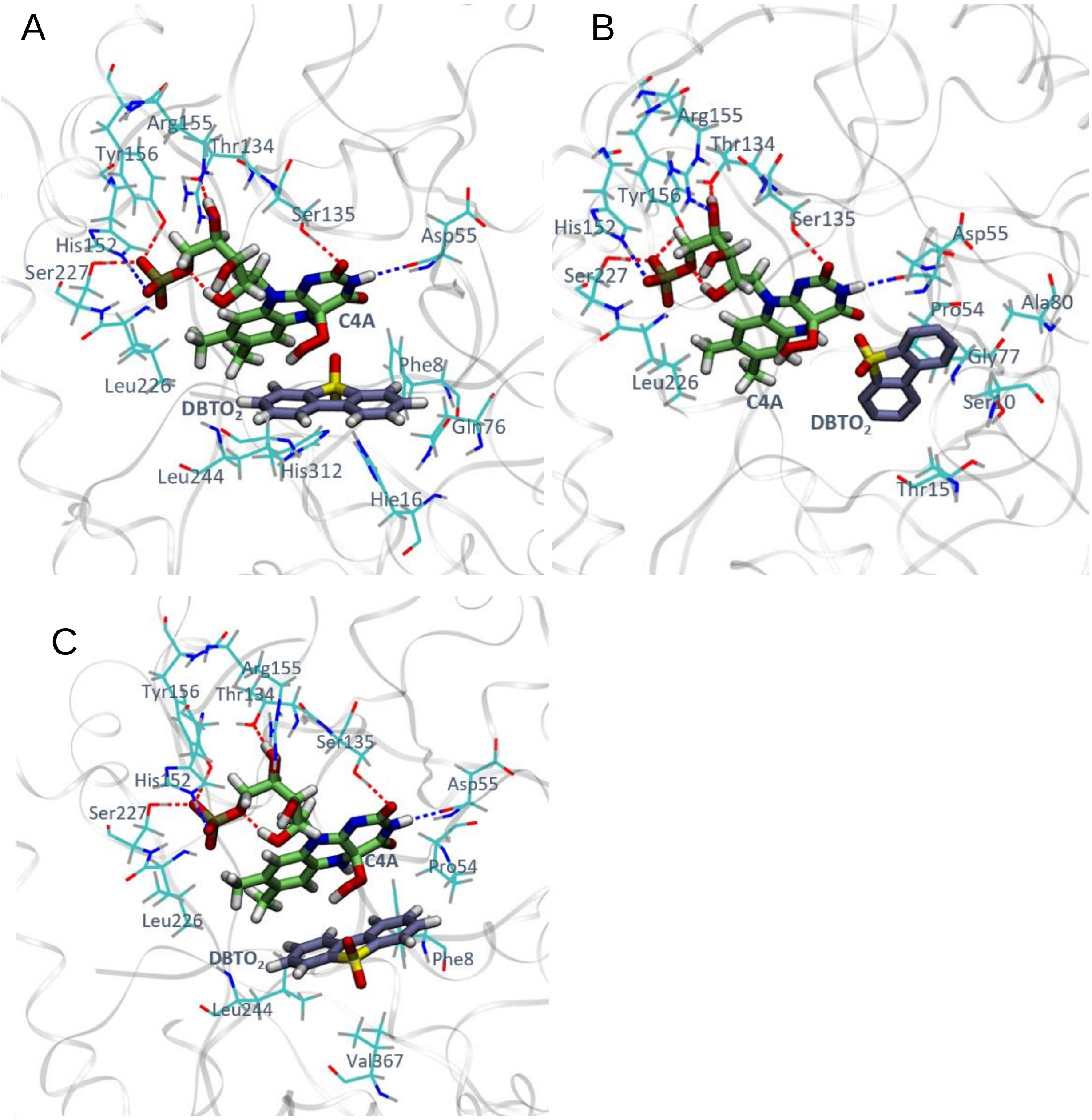


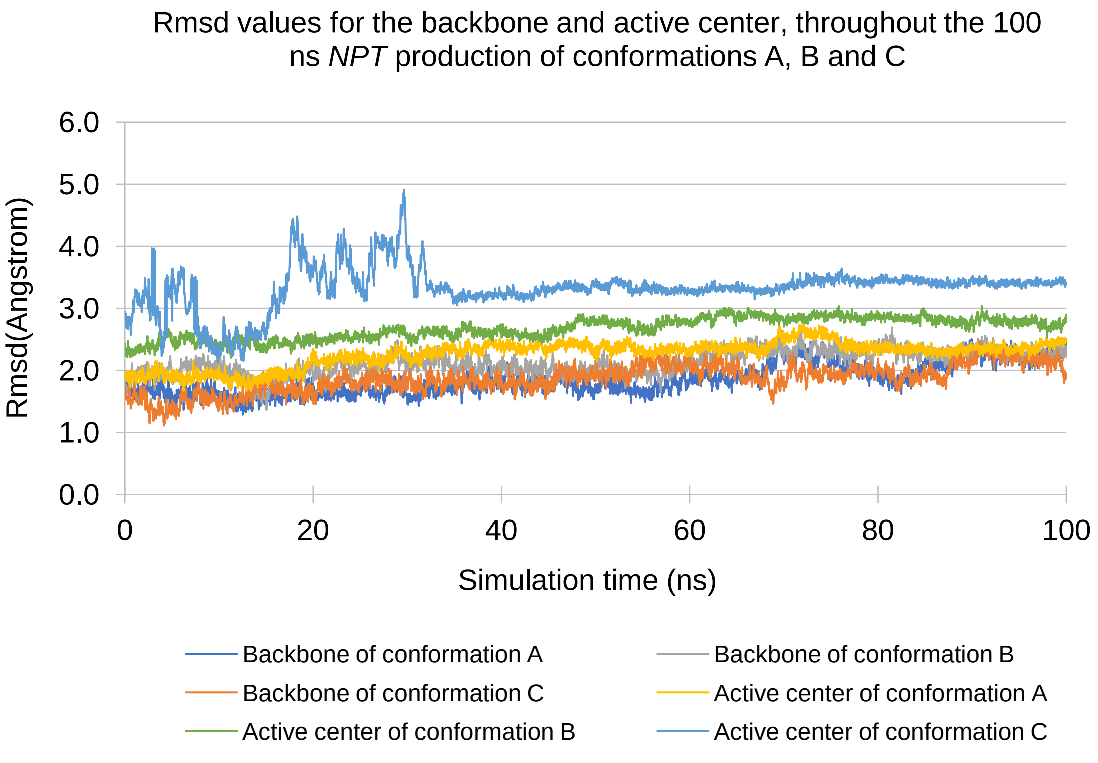
Figure 20. Most promising $DBTO_2$ docking poses. A) Docking pose attained from GOLD distance restraint docking; B) Unrestraint AutoDock Vina pose; C) Docking pose attained from GOLD distance restraint docking and side chain flexibility

As the active center of chain A and B are similar (Rmsd = 0.192 Å), the docking pose in chain A should be transferable to chain B. Hence, three DszA:C4A:$DBTO_2$ systems with the selected poses at both chain A and B are created.

3.2.2 Validation of DszA:C4A:DBTO$_2$

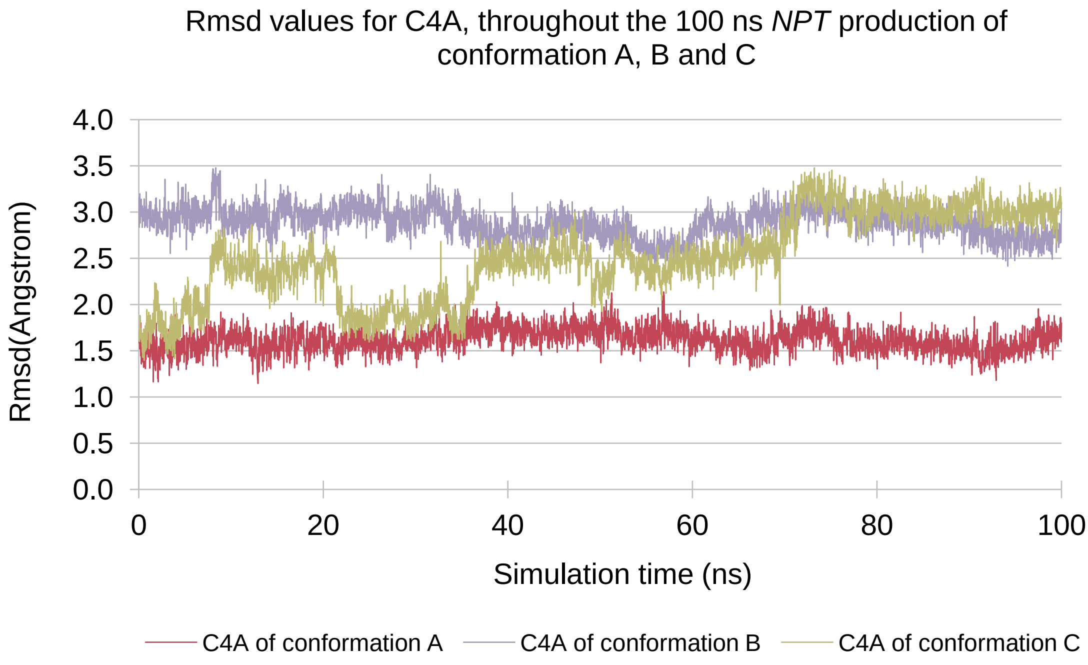
All the three systems created in the previous section undergo an energy minimization protocol, followed by equilibration and a 100 ns production. Analysis of these 100 ns allows the study of the system's behavior during such time to eventually decide which one is the most promising to be employed in QM/MM studies. Such evaluation is done through analysis of the Rmsd variations and of the interactions at the active center during the simulation time.

The evolution of the backbone and active centers' Rmsd for the three systems during the 100 ns production step is depicted in Graph 5 . Upon observation of such graph, it is noticeable that the backbone's Rmsd evolution is similar on all the systems and that all structures are equilibrated at around 2.0 Å. When considering the active centers of the systems, it is clear to see that conformation C not only takes a longer time to stabilize as more structural variations occur in it, resulting in a higher Rmsd value of around 3.3 Å. As for the remaining active centers, they stabilized in slightly different Rmsd values (2.8 Å for conformation B and 2.3 Å for conformation A).

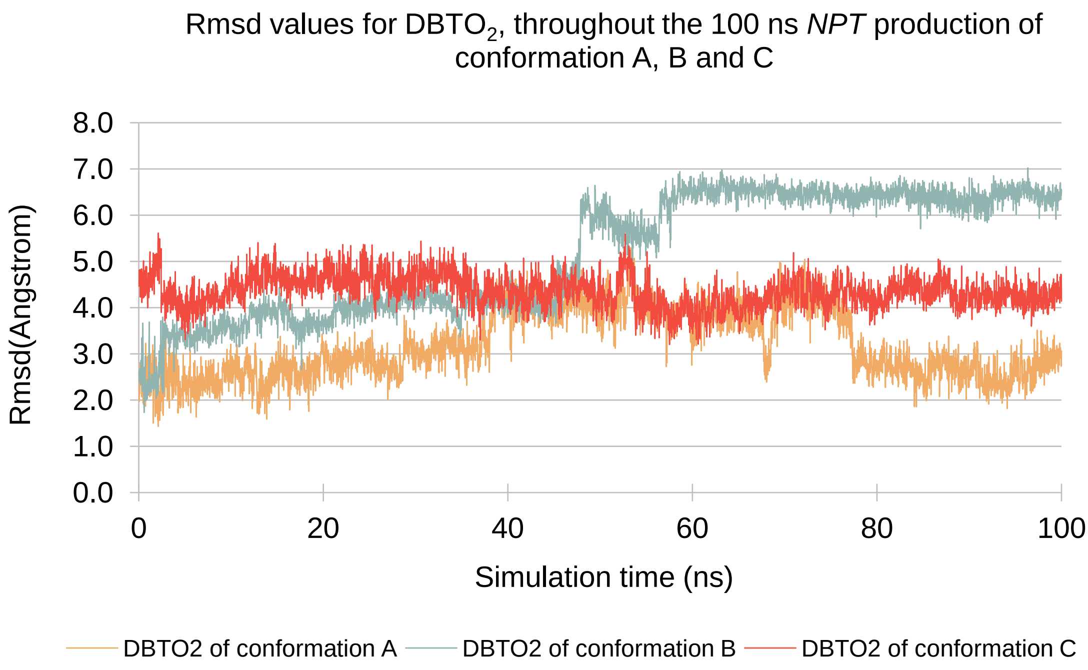


Graph 5. Rmsd values for the backbone and active centers for conformation A, B and C, throughout the 100 ns *NPT* production.

Comparison of the Rmsd variations of the cofactor and substrate of the systems also occurs and results are demonstrated in Graph 6 and Graph 7, respectively. Once again, the same behavior seen for the active center of the system is seen for the cofactor of conformation C as it has a more erratic behavior than the other systems. Moreover, C4A in this system only seems to be stabilized in the last 20 ns of the MD simulation, where its Rmsd fluctuates around 3.0 Å. As for the cofactor of conformation A and B, they both seem stable throughout the simulation. Nonetheless, the Rmsd of the cofactor from conformation A stays around 1.5 Å, which is less than for conformation B, where the cofactor Rmsd fluctuates around 2.7 Å. As for the substrate, conformation C is now the one with the most stable behavior, with its Rmsd fluctuating around 4.4 Å. Nonetheless, this is still an elevated Rmsd value. As for the substrate in conformation B, at around 50 ns, its Rmsd dramatically increases to later be stabilized at around 6.3 Å. Finally, in conformation A, after 40 ns, the Rmsd of $DBTO_2$ increases to 4.0 Å, only to, after another 40 ns decrease back into its initial value of around 2.5 Å. Nonetheless, the substrate of conformation A is the one with the lowest Rmsd value and only suffers from a momentaneous destabilization.



Graph 6. Rmsd values for the cofactor in conformation A, B and C, throughout the 100 ns *NPT* production.



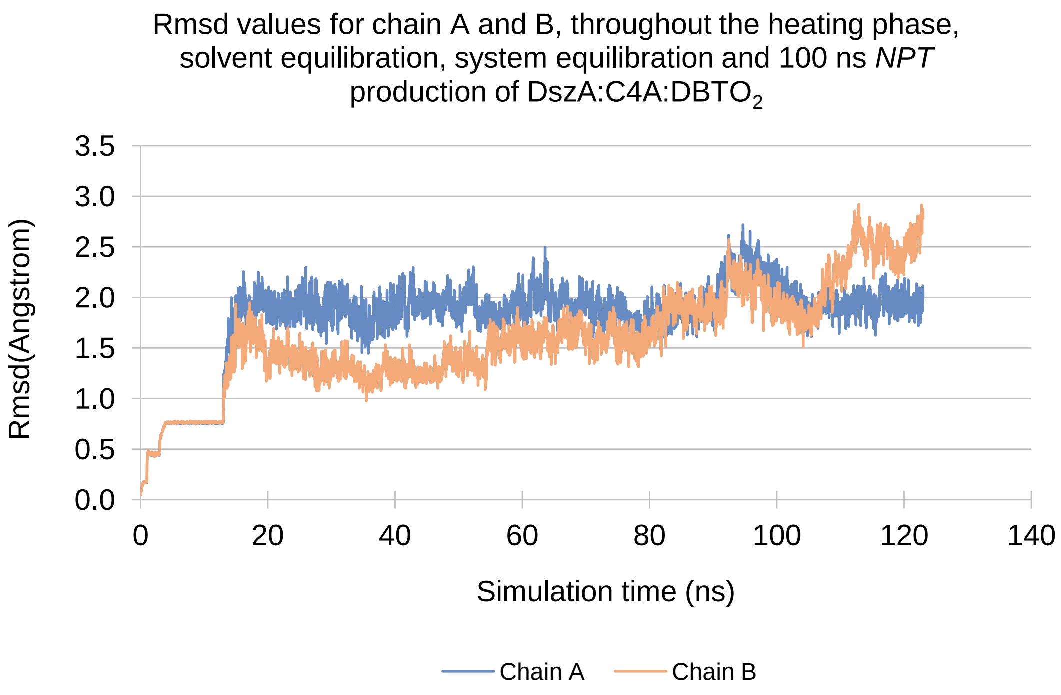
Graph 7. Rmsd values for the substrate in conformation A, B and C, throughout the 100 ns *NPT* production.

Analysis of the ligands' interactions at the active center is also performed for the three systems. When it comes to the cofactor interactions, conformation B and C, lose the majority of the hydrogen bonds that are considered necessary to maintain the cofactor in its position. For conformation B, in at least one of its chains, the cofactor's interactions with Asp55, Thr134, Tyr156, Ser227, His152 and Arg155 are lost. The substrate does not establish hydrogen bonds neither in chain A nor B of this system. In conformation C, C4A no longer interacts with Thr134, Ser135, Tyr156, Leu226, Ser227, His152 and Arg155 in chain A and with Asp55, Thr134, Ser135, His152 and Arg155 in chain B. Moreover, no hydrogen bonds with the substrate are seen. Conformation A, however, is the one that maintains the majority of the cofactors' interactions and the only one where DBTO$_2$ establishes hydrogen bonds with its surrounding residues. C4A of this system only seems to lose its interaction with Arg155, in chain A, and Ser135, in chain B.

Upon the Rmsd and interactions analysis, the results indicate that conformation A has the most stable behavior at the active center, the zone of most interest to this study. Additionally, such also reflects on the interactions that it establishes: as seen, the hydrogen bonds of the cofactor are most preserved in conformation A than in any other and the substrate only establishes hydrogen bonds in this system. For these reasons, conformation A, depicted in Figure 20A, is considered the most promising. Hence, a thorough description of it, detailing the differences between its chains, is given next to further study and validate the system.

3.2.2.1. Rmsd analysis

The evolution of Rmsd for each chain of conformation A is depicted in Graph 8. During the heating phase and restrained equilibrations, both chains behave similarly as their Rmsd values are overlapping. In the same graph, it is visible that chain A seems more stable during the production simulation time, as the Rmsd of chain A stays somewhat steady at around 1.8 Å and the Rmsd for chain B fluctuates more, going from an Rmsd of 1.2 Å to 2.8 Å. Comparison of the Rmsd values for the cofactors and the substrates also indicates that chain A has a stabler behavior (Graph 14 and Graph 15, respectively, in 6. Annexes.)



Graph 8. Rmsd values for chain A and B of DszA:C4A:DBTO$_2$, throughout the heating phase, the solvent equilibration, equilibration of the entire system and 100 ns *NPT* production.

3.2.2.2 Hydrogen bonds

The analysis of hydrogen bonds between the ligands and the surrounding residues is done with Cpptraj. In the production stage, C4A of chain A interacts with Ser227, Tyr156, Asp55, Leu226, His152, $DBTO_2$ and Ser135. With the exception of the latter, which is present in 57% of the 5000 production frames, all the other interactions seem stable as they are detected in over 90% of the production frames. This shows that the majority of the interactions previously seen for the cofactor are maintained, with an additional interaction of the $DBTO_2$ substrate with N1 of C4A. However, it is noticeable that no significant interaction of C4A and Arg155 is seen. Nonetheless, as mentioned before, the interaction with such residue is located at the ribitol of C4A which is very mobile, hence, the lack of interaction with Arg155 should not be impactful. The substrate, in addition to the mentioned hydrogen bond with C4A, also interacts with His16, an interaction that is maintained. The mentioned interactions can be seen in Figure 21.

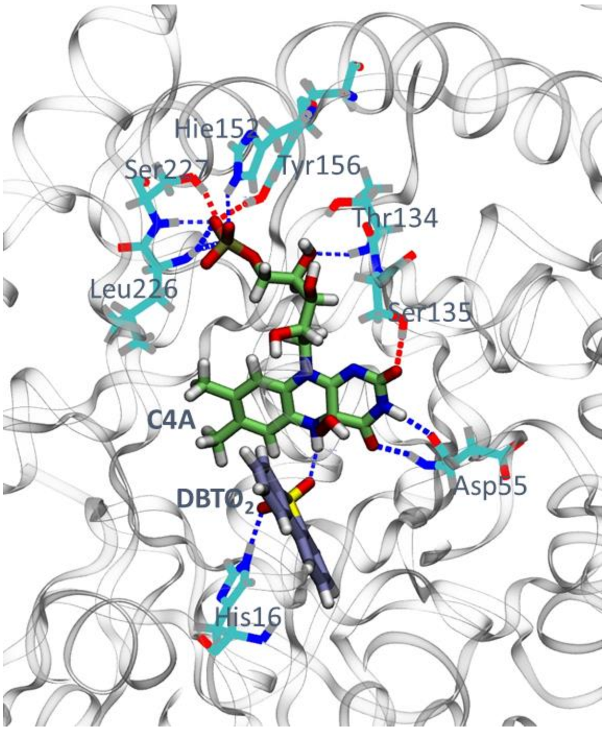


Figure 21. Hydrogen bonds of C4A and $DBTO_2$ at chain A of DszA:C4A:$DBTO_2$ with the surrounding residues.

Analysis of the 100ns production for chain B, demonstrates that, when compared to chain A, it has one less interaction with the cofactor. C4A of chain B not only does not interact with Arg155, as seen for chain A, but it also does not interact with Ser135. Moreover, the other interactions (with Ser227, Tyr156, Asp55, Leu226, His152, $DBTO_2$) are not as stable as they are for chain A, and hence not as consistently present during the MD simulation. As for the substrate, in chain B, $DBTO_2$ only has a hydrogen bond with C4A, at N1.

3.2.2.3. Hydrophobic contacts

Analysis of the hydrophobic residues that possibly can establish hydrophobic contacts with either the cofactor and/or the substrate is done on MDAnalysis. Within the 100 ns MD simulation, residues either within a 3.5 Å or a 6 Å radius of the ligand are observed.

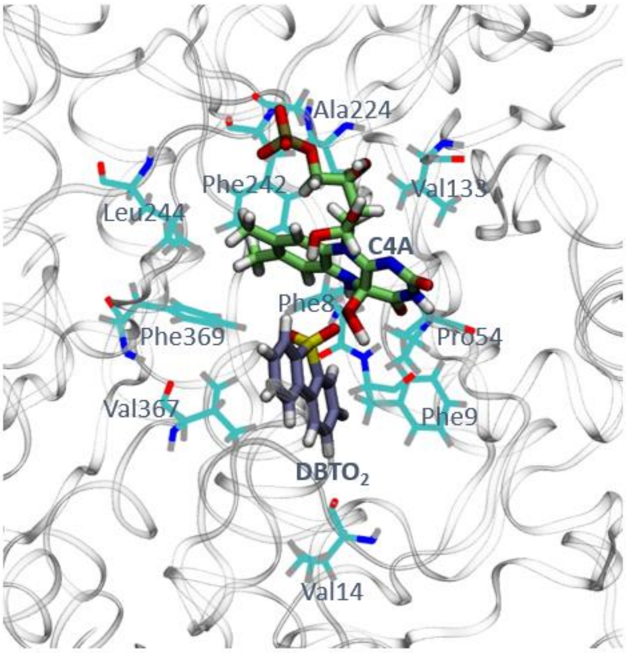


Figure 22. Hydrophobic residues surrounding the C4A cofactor and the $DBTO_2$ substrate on chain A of DszA:C4A:$DBTO_2$.

When assessing, for chain A, the hydrophobic contacts of C4A, the results indicate that the cofactor may hydrophobically interact with Pro54, Val133, Ala224, Gly225, Leu226, Phe242, Leu244 and Phe369 at a 3.5 Å radius and with Phe8, Leu53, Leu136 and Pro228 at a 6 Å radius. As for the substrate in chain A, just as C4A, it also interacts with Phe8, Pro54 and Phe369. Moreover, interaction with Phe367, Phe9 and Phe242 is also seen.

As for chain B, the hydrophobic residues that surround C4A at a maximum distance of 3.5 Å are Pro54, Val133, Ala224, Gly225, Leu226, Phe242 and Leu244. At 6 Å, the present residues are Phe8, Leu136, Pro228. When compared to chain A, the majority of the hydrophobic contacts are maintained, nonetheless, two less interactions are seen for the cofactor. As for the substrate, the tendency continues as fewer contacts are seen for chain B when in comparison with chain A. Here, the substrate at a 3.5 Å radius only interacts with Pro54. At 6 Å, only contacts with Phe8 and Leu244 are seen.

3.2.2.4. Distance between O11 of C4A and C1 or C12 of $DBTO_2$

The distance between the proximal oxygen of C4A and the equivalent C1 and C12 of $DBTO_2$ is monitored during the 100 ns simulation to see if they are close enough so the mechanistic hypothesis from Adak and Berkley can be possible. The mean of such distances is depicted in the following table:

Table 4. Mean distances between O11 of C4A and C1 or C12 of $DBTO_2$

	Chain A		Chain B	
	C1	C12	C1	C12
Mean	3.64 Å	5.10 Å	3.85 Å	3.97 Å
Standard deviation	0.28 Å	0.32 Å	0.49 Å	0.48 Å

From Table 4, it can be said that, in chain A, the most favorable distance is between the oxygen of the hydroperoxide of C4A and the C1, attached to the sulfur atom. In chain B, this distance is adequate for interaction with both of the carbons that are bound to the sulfur atom.

3.2.2.5. Geometric clustering

Finally, clustering is performed to attain a representative structure to use in QM/MM calculations, where hypotheses of the reaction mechanism of DszA are going to be tested. Using the rms of the active center, clusters from the 100 ns simulation are made. Moreover, to observe if the diversity of conformations when only the active center is free is comparable to the diversity of conformations when the entire system is free and thus, get an indication of how stable the attained substrate pose is, geometric clusters are also done for the equilibration step where only the active center is free. The presence of the same pose in both clustering protocols could be an indication of pose credibility. In the DszA:C4A:$DBTO_2$ system, three clusters are obtained for the 100 ns simulation, both for chain A and B (6 in total). From the trajectory where only the active center was kept free, three clusters are obtained from chain A and 2 from chain B. Comparing all of these, in chain A, the clusters from both simulations are not similar. In chain B, the dominant clusters from both origins are similar, which could serve as further validation of the system. Nonetheless, as seen before (Graph 15) during the 100 ns production, the substrate pose is stabler in chain A than in chain B. This, coupled with other reasons, described next, indicates that chain A is a more promising system.

As seen in the previous subsections, both chains are equilibrated, however, as suggested by the MD simulations, chain A has a stabler behavior. As for interactions at the active center, the hydrogen bonds in chain A prove to be more conserved than for chain B. Additionally, the same tendency is observed for the hydrophobic bonds that hold the ligands, as in chain A more contacts with hydrophobic residues are established. Taking into consideration that an adequate distance is seen between the oxygen of the hydroperoxide of C4A and the C12 and/or C1 in both chains, it is possible to validate the system, especially chain A, as suitable to be employed in further studies. Hence, as it demonstrates to be the most promising complex, the dominant cluster from the 100 ns simulation of chain A is the one used in QM/MM studies, to start with.

3.2.2.6. How simulating the substrate pose seen for BdsA:FMN:$DBTO_2$ further validates the selected DszA:C4A:$DBTO_2$ system

Efforts are made to simulate as closely as possible the pose of the substrate seen in BdsA, as it should be similar to the one of DszA. This is originally done to possibly attain a new system to employ in QM/MM studies. However, as it will be described next, this step does not render a substantially different new structure to include in such studies. Hence, its importance relies on its pose similarity to the previously obtained system, as this contributes to its further validation.

To obtain the most similar substrate pose to BdsA, docking in AutoDock Vina is performed in DszA:FMN, instead of C4A, as the substrate pose in BdsA is known only for such cofactor. This is done as a free ligand:protein docking, from which 10 poses are attained. From these, only 4 represent different poses. The best-ranked result is similar to the pose observed for BdsA:FMN:$DBTO_2$, hence, it is modeled into both Chain A and Chain B of DszA. The system is then parametrized on Xleap (transforming FMN into C4A), minimized (solvent minimization, solvent equilibrium, side chains, active center, all) and equilibrated (heating, solvent, active center(10 ns), all(10 ns)). Then, a 141 ns production is performed. To enable the comparison of this system with the DszA:C4A:$DBTO_2$ system from "3.2.2.5. Geometric clustering", 2 clusters from the 141 ns of production are attained for each system's chain. Comparison of the active center of the dominant cluster of chain A from conformation A with the active center of the dominant cluster here attained for chain B, demonstrates a somewhat different substrate pose. Superimposition of these structures gives a 1.065 Å Rmsd value. However, when it comes to interactions with the cofactor there isn't one stable interaction that holds up for at least 50% of the trajectory time. Similarly, the substrate does not establish hydrogen bonds with any residue. As this system is incapable of preserving the ligands interactions, it is not suitable to be further employed in QM/MM studies. When comparing the dominant cluster of the active center of chain A of conformation A with the active center of the cluster here attained for chain A, an Rmsd of 0.686 Å and a somewhat similar substrate pose in both systems is seen (Figure 23). This reinforces that the structure from "3.2.2.5. Geometric clustering" is plausible and should be representative of the reality. Though, when analyzing the cofactors' interactions in the cluster structure here attained, they are not as favorable as in the system selected in the previous section: interaction of C4A with Ser135, Leu226, Ser227 and Arg155 are missing. Moreover, the substrate only interacts with the cofactor. Additionally, a tendency starts to emerge: as seen here and in previous MD simulations of other systems, the interaction of C4A with Ser135 seems to have a propensity to be lost. This can be an indication that its role is possibly more prominent in C4A activation from FMN, than in the reaction mechanism.

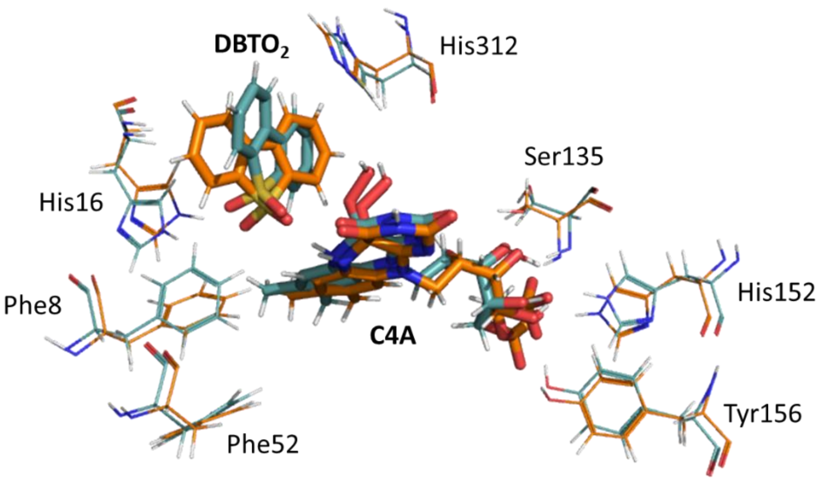


Figure 23. Superimposition of the active centers from the dominant chain A clusters from conformation A (in orange) and from the system intended to simulate the substrate position in BdsA (in blue).

As the clusters here attained failed in maintaining the ligands' interactions and the overall binding pose of the substrate on them does not differ much from that of the structure on "3.2.2.5. Geometric clustering", this latter is still the most promising system to employ in the QM/MM studies, described next.

3.3. Mechanistical reaction studies with QM/MM methods

The system attained in "3.2.2.5. Geometric clustering" is truncated beyond 12 Å of the active center of chain A, the high and low layers are assigned, and an outer shell of 4 Å is set to frozen. The system's geometry optimization with mechanical embedding is achieved in 120 steps and the subsequential optimization through electrostatic embedding is accomplished in 64 steps. After the optimizations, the hydrogen bonds of the cofactor and the substrate with the surrounding residues are preserved. The active center after these optimizations is shown in the following figure:

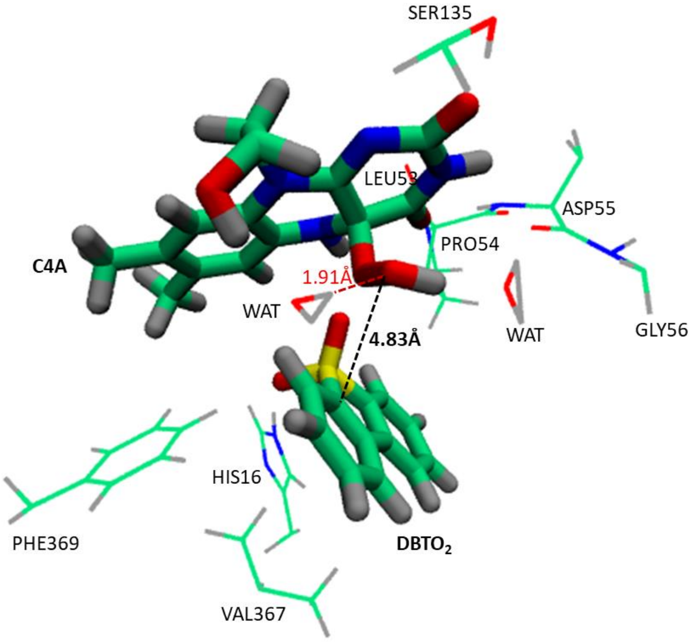


Figure 24. Reactive system of the initial structure used for QM/MM studies.

In this structure several attempts are made. To test the mechanistic hypothesis of Adak and Begley[61], which suggests the formation of a peroxyhemiacetal between the hydroperoxyl of C4A and the C1/C12 of $DBTO_2$, the energy variation of the system as the distal oxygen (O10) and C1 approximate is studied (the initial distance of these reaction coordinates is 4.83 Å as depicted in Figure 24). Another attempt in this structure aims to study the possibility of the transfer of the hydroperoxyl group of C4A to the $DBTO_2$ substrate; thus, the proximal oxygen (O11) of C4A and the C1 of $DBTO_2$ are used as reaction coordinates. The distance between these atoms is depicted in Figure 25. Another set of reaction coordinates that studies the proton transfer from N1 of C4A to O2 of $DBTO_2$ is also used, however, as this does not present promising results and does not give new alternatives to be studied, it is not discussed here. More information about it can be found in 6. Annexes (6.2. Proton transfer from N1 of C4A to O2 of $DBTO_2$).

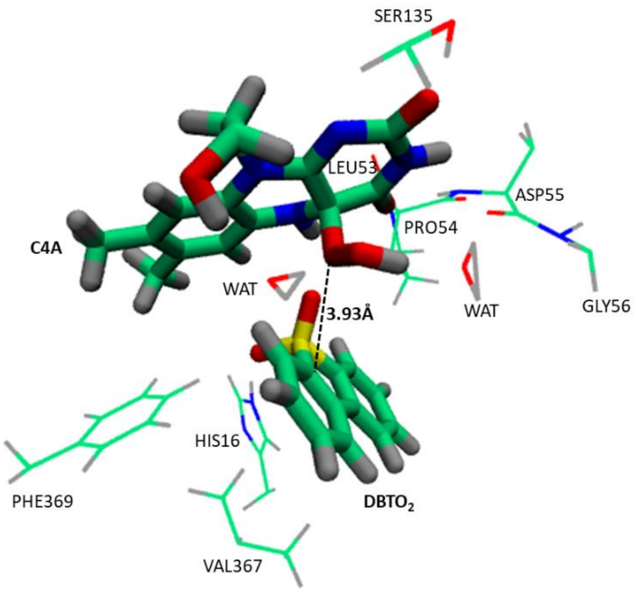


Figure 25. Reactive system of the initial structure used for QM/MM studies, with the distance between O11 of C4A and C1 of DBTO$_2$ marked (3.93 Å).

3.3.1. Attempt to form the peroxyhemiacetal proposed by Adak and Begley[61]

After the system preparation, the first attempt of a hypothetical reaction step is made. Here, the distance between the distal oxygen of the hydroperoxyl of C4A (O10) and the C1 of DBTO$_2$ goes from an initial value of 4.83 Å (as seen in Figure 24) to 1.59 Å in 28 steps, decreasing its distance 0.12 Å in each step.

Several bond events occurred in this attempt. A cleavage of the bond between the distal and proximal oxygen of C4A, O10 and O11, respectively, is seen. The interaction of O10 with the nearby water is also lost as the distal oxygen establishes a hydrogen bond with O11. In this attempt, the C-S bond of the substrate did not break and the formation of the proposed peroxyhemiacetal did not occur. The final structure can be seen in Figure 26 and the energy scan for this attempt is given in Graph 9.

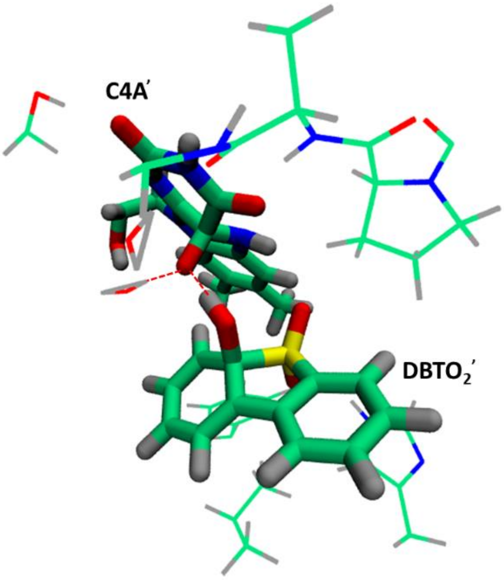
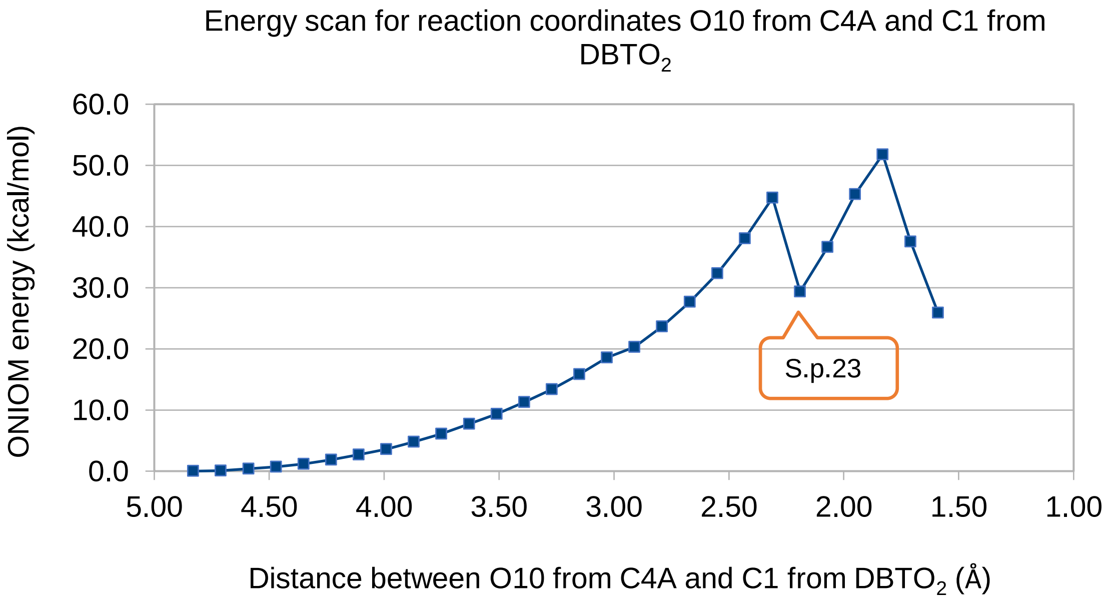


Figure 26. Final structure using as reaction coordinates O10 from C4A and C1 from DBTO$_2$. DBTO$_2$’ and C4A’ represent the modified DBTO$_2$ and C4A, respectively. The red dotted lines mark the hydrogen bonds where the hydroxyl group of C4A participates as a donor.



Graph 9. Energy scan along the distance from O10 from C4A to the C1 of DBTO$_2$.

The energy required for such approximation to occur is too high (51.8 kcal/mol). Nonetheless, such high energy was expected as observation of the initial structure in

Figure 24 indicates a non-favorable high distance between the approaching atoms (4.83 Å). Not only that, but the poor orientation of the oxygen of C4A relatively to the carbon of $DBTO_2$ is also noticeable, as the first is oriented to a water molecule with which it is interacting in a short distance of 1.91 Å.

At scan point 23 (marked as S.p.23 in Graph 9), a local energy minimum can be seen. Such decrease in energy is attributed to the rotation of O11 towards a water molecule, forming a hydrogen bond. This structure sparked attention due to the favorable proximity (2.19 Å) and orientation of the distal oxygen of C4A to the C1 of $DBTO_2$. Hence, such structure undergoes a geometry optimization in electrostatic embedding to establish if it is a true minimum structure that can be considered for further QM/MM studies. This alternative structure, here on out referred to as “structure 23” is further discussed in “3.3.3. ”.

3.3.2 Attempt to transfer the hydroperoxyl group of C4A to the $DBTO_2$ substrate

To study the possibility of the whole hydroperoxyl group of C4A being transferred to the substrate, instead of just the distal oxygen as in the previous section, the distance from the O11 of C4A to the C1 of $DBTO_2$ is used as reaction coordinate. Here the reaction coordinate varies from 3.93 Å to 1.53 Å in 21 steps of 0.12 Å decrements. Observation of the structural changes between the initial (Figure 25) and final conformations (Figure 27) shows, as intended, the full hydroperoxyl group of C4A being transferred to C1 of $DBTO_2$. Upon the group transfer, the C-S bond of the substrate also breaks, resulting in an open ring formation. Additionally, the hydrogen bond between O11 and a water molecule is lost and O2 from the substrate deprotonates the N1 of C4A.

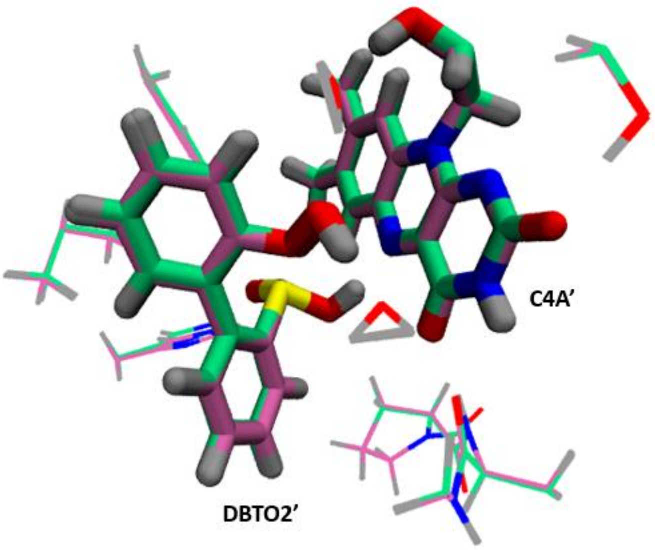
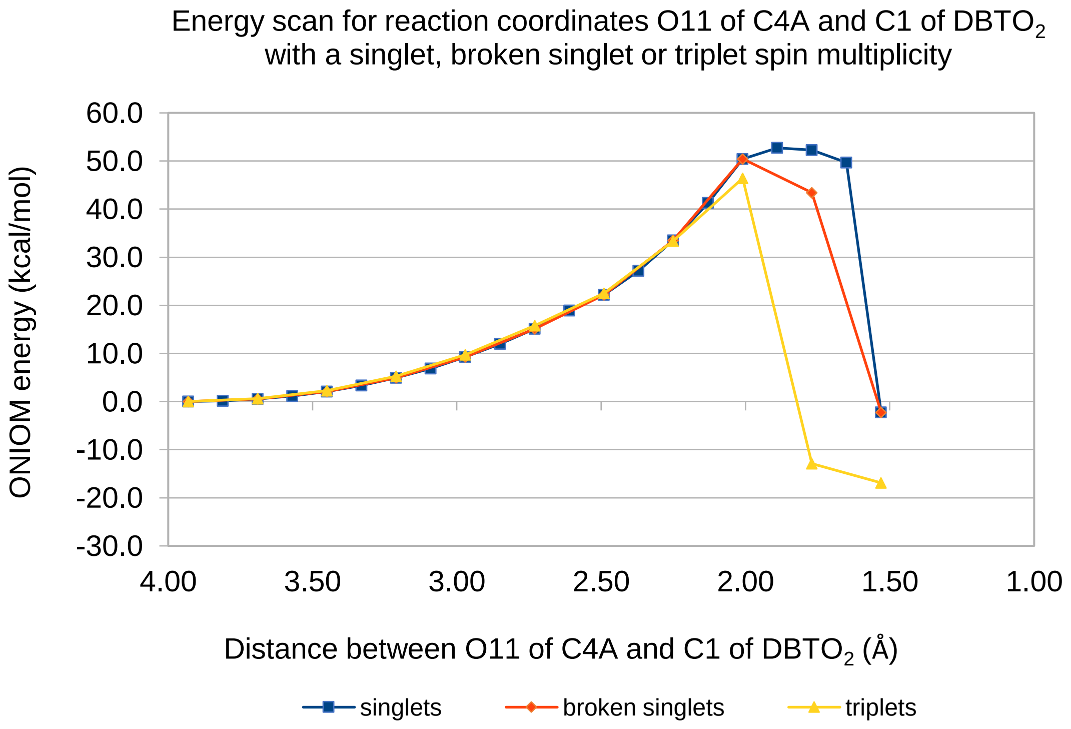


Figure 27. Final structure in singlet (in green) versus final structure in triplet (in pink), using as reaction coordinates O11 from C4A and C1 from $DBTO_2$. $DBTO_2$' and C4A' represent the modified $DBTO_2$ and C4A, respectively.

The energy scan for this attempt can be seen in Graph 10. Such graph indicates a reaction product that is more stable than the reactant, suggesting an exothermic reaction. Nonetheless, a high energy barrier of 52.7 kcal/mol is visible, an energy that can be partially explained by the initial hydrogen bond that O11 established with water that must be lost for the approximation to happen. Due to its high energy, it is not likely that this is a true reaction step in the mechanism of DszA. However, this energetic profile is attained assuming that the reaction proceeds through dative covalent bonding (in a singlet spin multiplicity). To evaluate the possibility of radical formation during the process, the possibility of having unpaired electrons in the system, either as broken singlets or triplets is also studied (Graph 10).



Graph 10. Energy scan along the distance from O11 of C4A to C1 of DBTO$_2$, either in a singlet, broken singlet or triplet configuration.

From this graph, a slight decrease in the energy barrier can be seen when the system exhibits a triplet spin multiplicity. Additionally, when the electrons are unpaired an even stabler intermediate is attained, with a relative energy of -16.9 kcal/mol. Hence, it is a possibility that the system may undergo a change in spin if this intermediate is formed. To evaluate its formation, the final structures attained with a singlet and triplet spin multiplicity are superimposed in Figure 27.

As seen, upon optimization of the formed intermediate in a triplet multiplicity, the active centers of the structures are identical as they superimpose with an Rmsd value of 0.014 Å. As the interactions of the ligands are maintained and the resulting triplet state has a lower energy than the singlet state, the existence of unpaired electrons within a triplet spin multiplicity is possible. Nevertheless, a high energy barrier of 46.4 kcal/mol is still seen. Thus, from an energetic point of view, this reaction step is not likely to occur.

3.3.3. Attempt to form the peroxyhemiacetal proposed by Adak and Begley[61] upon rotation of the hydroperoxyl group of C4A

Upon geometry optimization of the structure from scan point 23 of the section 3.3.1. , the interactions at the active center are maintained, the adequate orientation of the distal oxygen of C4A to C1 of the substrate is also preserved, and the distance between these atoms increases to 3.46 Å. Despite such increase, this is still a favorable distance for interaction and thus, this optimized structure, seen in Figure 28, is used in QM/MM studies.

A linear transit scan is performed in this structure with the same reaction coordinates as before, O10 from C4A and C1 of $DBTO_2$. The distance between these, as mentioned, starts at 3.46 Å and in 18 steps of 0.12 decrements reaches the distance of 1.42 Å. However, the desired product is not attained as observation of the bond events doesn't show the substrate's ring opening. Nonetheless, similarly to the attempt in section "3.3.1. Attempt to form the peroxyhemiacetal proposed by Adak and Begley61", a bond breaking between the proximal and distal oxygen of C4A is seen and the latter, O10, forms a new bond with the C1 of the substrate. Hence, the formation of the peroxyhemiacetal does not occur. The final structure is depicted in Figure 28. As for the energetic profile, an energy barrier of 47.6 kcal/mol is seen in an endothermic reaction, yielding a final structure with an energy of 14.5 kcal/mol. Due to the lack of C-S bond break and the high energy barrier, this attempt also does not seem viable. As this didn't work as expected, a residue increment is done in this structure and described in the following section.

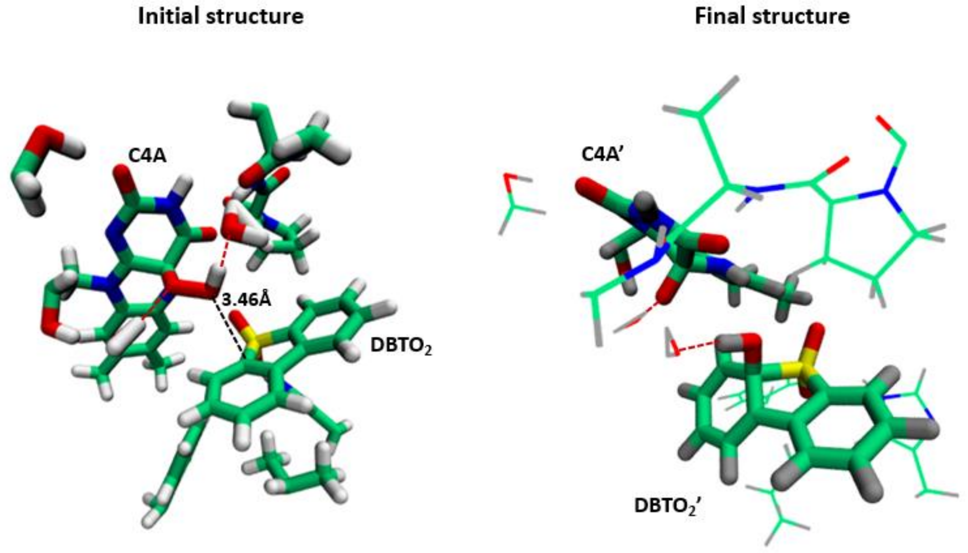


Figure 28. Initial structure 23 employed for the reaction coordinates O10 from C4A and C1 of DBTO$_2$ and respective final structure. The black dotted lines mark the distance between O10 and C1 and the red dotted lines mark the hydrogen bonds where the hydroxyl group of C4A participates as a donor. C4A' and DBTO$_2$' represent the final ligands' structures in this scan.

3.3.4. Attempt to form the peroxyhemiacetal proposed by Adak and Begley[61] using His312 as a base

The mechanism of reaction that is proposed in Figure 8, suggests that a base accepts a proton from the hydroperoxyl group of C4A. Observation of the residues at the active center shows that His312 is the only residue that could serve as this base. Hence this residue is added to the high layer of the previous model. Such addition entails that the high layer now has 148 atoms. The conformation of structure 23 with His312 in the reactive layer, upon geometry optimization, can be seen in Figure 29.

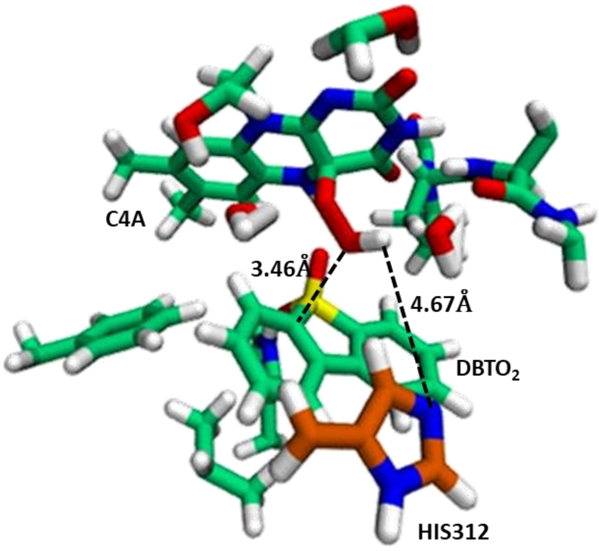
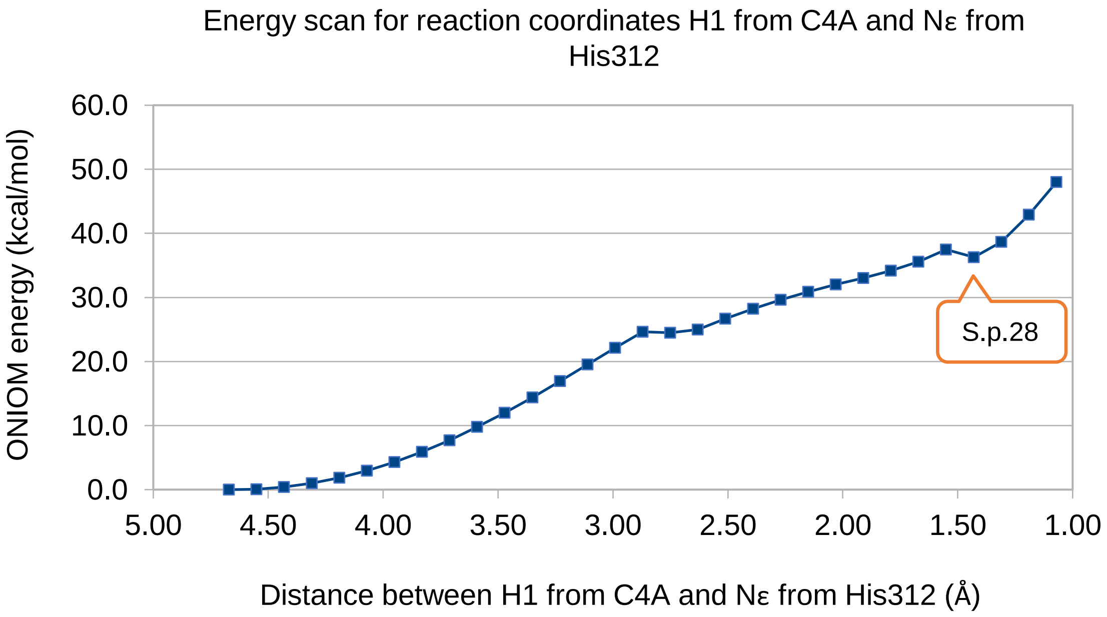


Figure 29. QM layer of structure 23 with the addition of His312, shown in orange.

Two atom pair distances are studied in this structure. One simulates the protonation of His312 by a water molecule. The water would then be restored, as it can accept one proton from the hydroxyl group of C4A, which would facilitate its reactivity toward the C1 or C12 of $DBTO_2$, giving the desired product. However, this attempt rendered unfavorable results and will not be here discussed. More information on it can be seen in 6. Annexes (6.3. Protonation of His312 by a water molecule in structure 23). The other alternative is performing a similar mechanism using His312 directly as a base. Such attempt is described next.

Simulating the deprotonation of the distal oxygen of C4A by the Nε of His312 (the acceptor atom), an energy scan is attained where the coordinates are brought closer, from 4.67 Å to 1.07 Å in 31 steps of 0.12 Å. In this attempt, along the scan, the hydrogen bond formed between the O10 of C4A and the nearby water molecule breaks, so the hydrogen bond can reorientate to the Nε of His312. However, no new minimum is observed, as the energy is still rising at the last scan point. Such energy profile can be seen in Graph 11:



Graph 11. Energy scan along the distance from H1 from C4A to the Nε of His312.

Once again, the energy reaches high values (48.0 kcal/mol). Nonetheless, this poor outcome could have been predicted by observation of the reactive layer (Figure 29) as the distance between the reaction coordinates is somewhat large (4.67 Å) and the residues' orientation is also not favorable for such interaction. Additionally, another factor that contributes to such energetic values is the repulsion forces that occur between the histidine and the surrounding residues as they get too close along the scan. This attempt also does not seem viable as a reaction step of the mechanism of DszA, since no minimum is seen for the protonated His312 and the energy for this protonation to occur is too high. Nonetheless, on scan point 28 (S.p.28 in Graph 11) a hypothetical energy minimum is seen. Upon observation of the optimized structure, a hydrogen bond of 1.80 Å between the reaction coordinates is seen. This structure, depicted in Figure 30, as it maintains its interactions and a favorable distance between O10 of C4A and C1 of $DBTO_2$ (that now are at a distance of 3.48 Å), it is used as a new initial structure for QM/MM studies.

3.3.4.1 Attempt with structure 28

To test the formation of the peroxyhemiacetal proposed by Adak and Begley in the new initial structure (Figure 30), the distance between the O10 of C4A and C1 of

$DBTO_2$ is decreased from 3.48 Å to 1.20 Å in 20 steps of -0.12 Å. This is done to see the influence of having a hydrogen bond between His312 and the O10 of C4A. The hydrogen bond is maintained along the scan until the highest energy point (seen in Graph 12, when the distance between the coordinates is 1.92 Å). After this point, this hydrogen bond is lost as the covalent bond between the O10 of C4A and the C1 of $DBTO_2$ starts to form and the hydrogen of O10 reorients to bond with the O11 of C4A, when the bond between the distal and proximal oxygen of C4A breaks. After the double bond between the O11 and the C1 is formed, the C-S bond of $DBTO_2$ breaks, opening the substrate's ring. Once again, no peroxyhemiacetal is formed, and the hypothesis of its formation seems to be further refuted, at least for the structures studied under the scope of this thesis. The optimized resulting structure after the bond events can be seen in Figure 30.

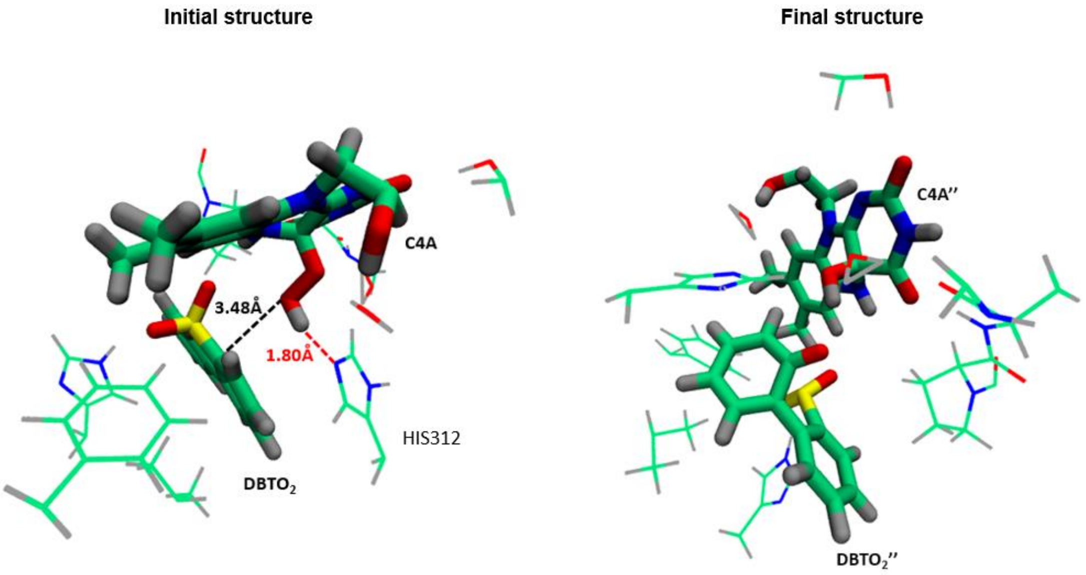
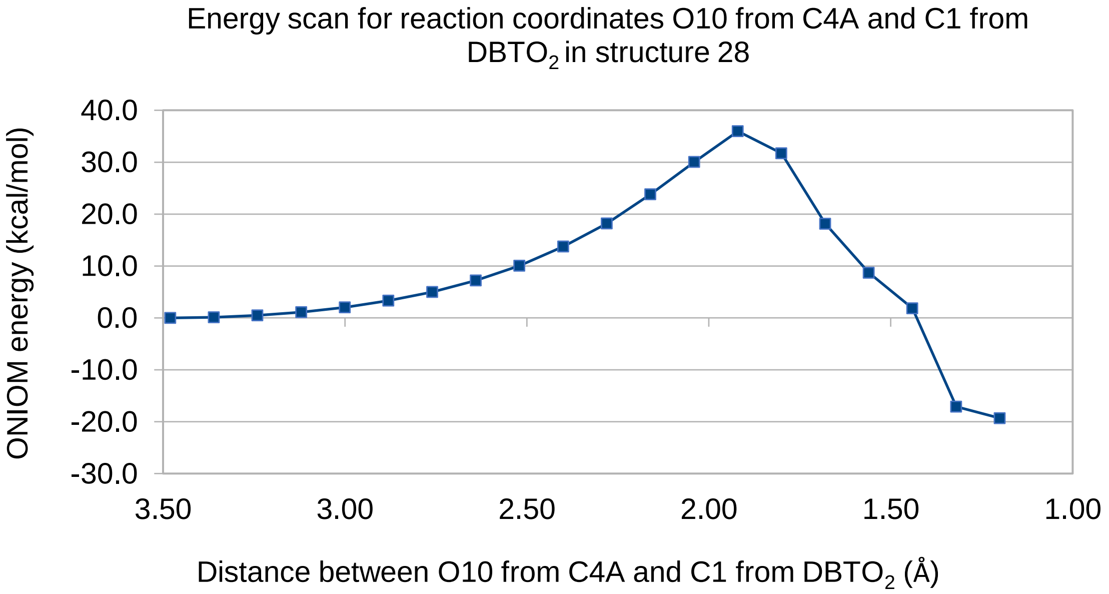


Figure 30. Initial and final structure of the transit scan for reaction coordinates O10 of C4A and C1 of $DBTO_2$ in structure 28. C4A'' and $DBTO_2$'' represent the resulting cofactor and substrate, respectively. The black dotted lines mark the distance between O10 and C1 and the red dotted lines mark the hydrogen bonds where the hydroxyl group of C4A participates as a donor.

As for the energy profile, it follows in Graph 12.



Graph 12. Energy scan along the distance from O10 from C4A to the C1 of DBTO$_2$, in structure 28.

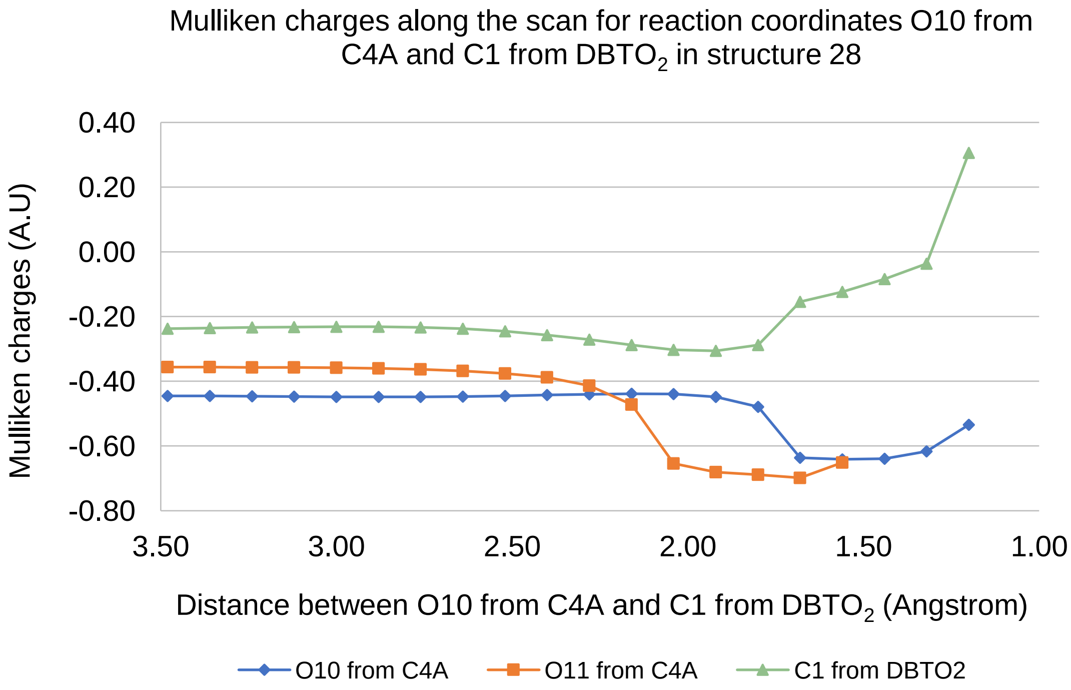
In Graph 12, an exothermic reaction can be seen, rendering a final structure that is more stable than the reactant. Such structure has a total energy of -19.3 kcal/mol and an energy barrier of 36.0 kcal/mol, corresponding to a possible transition state. Such barrier is still too high (even though it is the lowest energy attained so far). Nonetheless, as favorable bond events that get us closer to the intended product occur and the transit scan shows a satisfactory profile (increasing the energy until a possible transition state is achieved followed by an energy slop where a stable structure is attained), this is the best attempt so far and reinforces the idea that a base, such as His312, is important for the reaction mechanism. However, it is still unknown if the position of His312 on this initial structure is viable and possible to exist in a natural system or if it is only a result of the decrement of its distance to the cofactor when it is used as a reaction coordinate. To study this, restricted MD simulations are done and described next.

Several *NPT* MD simulations with restrictions are made on the equilibrated structure from conformation A (section “3.2.2 Validation of DszA:C4A:DBTO2”). In one, to emulate as closely as possible the reactant structure here seen, during 10 ns, the Nε of His312 and the hydrogen of the O10 of C4A are restricted to a distance of 2.10 Å (simulating the hydrogen bond between them), and C1 of DBTO$_2$ and O10 of C4A are set to be apart with a distance of 3.50 Å (emulating the initial distance seen for these reaction coordinates). In another restricted simulation, only the distance between Nε of His312 and the hydrogen of the O10 of C4A is restricted (also to 2.10 Å), during 20 ns, to study if a diminished distance between O10 of C4A and C1 of DBTO$_2$ is a requisite for the establishment of the hydrogen bond seen between Nε of His312 and the O10 of C4A in the initial structure employed in this section. Analysis of both restricted simulations showed that the restraints worked as the restricted atoms are held at the desired distances from each other. Afterward, to assess if such restrictions still hold when released, the systems are set free for another 10 ns *NPT* MD simulations.

In the first restricted MD simulation, where two distances are restraint, when the restraints are lifted an average distance of 4.65 Å and 3.80 Å between C1 of $DBTO_2$ and O10 of C4A are seen for chain A and B, respectively. Thus, for the distance between these atoms, when the system is set free, only the distance in chain B can be considered as maintained. For chain A, the restraints do not hold up. As for the distance between Nε of His312 and hydrogen of the O10 of C4A, in chain A the average distance is now 6.23 Å and in chain B is 5.26 Å. This is a significant change from the 2.10 Å that they were previously restraint at. Hence, the restraints also do not hold up. These results indicate that the position of His312 near the hydrogen of O10 from C4A, as seen in structure 28, it is not likely. Regarding the second unrestricted MD simulation, the hydrogen bond between the Nε of His312 and the O10 of C4A is also not maintained in both chains A and B (distances of 4.13 Å and 5.38 Å). Therefore, whether a diminished distance between C1 of $DBTO_2$ and O10 of C4A is present or not, the result is the same: the position of His312 here in study is unlikely. Instead, the first unrestricted MD simulation suggests that the distal oxygen of C4A (O10) has a better orientation to C12 of $DBTO_2$ than to C1. As these carbons are topologically equivalent, although the distance of the hydroperoxyl to C1 is smaller than to C12, it is also possible that the peroxyhemiacetal could be formed with C12. To test if the position of His312 is viable in such case, the distance between the Nε of His312 and the hydrogen of the O10 of C4A is restricted to 2.10 Å and the distance between the C12 of $DBTO_2$ and the O10 of C4A is set to 3.50 Å, during a 10 ns *NPT* MD simulation, and are then lifted for another 10 ns unrestricted *NPT* simulation. Results from the latter show an average distance of 3.96 Å for the distance between C12 of the substrate and O10 from C4A for both chain A and B. Thus, it is a possibility that C12 may be preferred over C1 (this option will be later discussed in this thesis). As for the hydrogen bond between the Nε of His312 and the O10 of C4A, it is still inexistent (it has a value of 4.10 Å in chain A and a value of 4.15 Å in chain B), which further corroborates that the position of His312 seen in structure 28 is not probable.

3.3.4.1.1 Charge analysis of the most favorable result

Despite unlikely, structure 28 is still the best attempt so far. Hence additional information that can be useful, such as charges analysis, can be retrieved from it. Understanding the charge transfer processes of this reaction step can give some additional information on what happened, including the identification of the nucleophile, which could be either the cofactor or the substrate. This is also important to understand what could be seen in further mechanistic steps, as the protonation of the O10 of C4A (now bonded to the substrate), the deprotonation of the N1 of C4A, and the removal of the hydroxyl group from C4A, are still required to attain the desired product of the catalysis by DszA. To study the charge transfer during the scan, the Mulliken charges for the atoms of the substrate and cofactor are studied along the linear transit scan. Their evolution for the proximal and distal oxygen of C4A, and for C1 of $DBTO_2$ can be seen in the following graph:



Graph 13. Mulliken charges for O10 and O11 from C4A and C1 of $DBTO_2$ along the scan using as reaction coordinates O10 from C4A and C1 from $DBTO_2$ in structure 28.

Analysis of Graph 13 shows a charge decrease on O11 as the distance between the O10 and C1 decreases and the bond between O11 and O10 elongates. When this bond eventually breaks (around 2.00 Å in Graph 13), a decrease in charge is seen for O11. This entails that upon the bond break, O11 hogs the electrons from the bond, making it more negative and potentiating the future deprotonation of O10. This protonation of O11 is represented in the last point of the O11 graph in Graph 13, where a slight charge increase is seen. Moreover, when O10 from C4A and C1 from $DBTO_2$ start forming a bond, at around 1.80 Å, a charge increase is seen for C1, whilst the charge of O10 decreases. This indicates that O10 accepts electrons from C1. Hence, $DBTO_2$ acts as the nucleophile. With such results, predictions of the following reactionary steps can be made. A hypothesis is that the deprotonation of the N1 of C4A and the removal of the hydroxyl group from the cofactor can be attained if the hydrogen of N1 protonates the hydroxyl group, resulting in a water molecule, which is a good leaving group. As these bonds break, N1 hogs the electron from the bond (as it is more electronegative than hydrogen) making it negative, and the carbon that was bonded to the hydroxyl becomes positive, as the oxygen is more electronegative and hogs the electrons. These events could potentiate the formation of a double bond between the carbon and N1. Moreover, as the water molecule is released, it could stay in a favorable position to be deprotonated by the newly formed carbonyl of $DBTO_{2.}$ As the ring bonded to the carbonyl group is electron deficient, protonation of the oxygen will cause the transformation of the double bond between the carbon and the oxygen into a single bond, and the ring is stabilized. However, this hypothesis is not here tested as the energy barrier is too high and the proximity of His312 to the hydroperoxyl group is not likely. Hence, new initial structures and/or hypotheses are still required.

3.3.5. Which is more suitable to study mechanistic hypotheses: C1 or C12 of $DBTO_2$?

Up until now, C1 has been used as reaction coordinate, and not much attention has been given to C12 of $DBTO_2$. This has been the case, as C1 usually presented smaller distances to the hydroperoxyl group of C4A. Nonetheless, $DBTO_2$ is a symmetrical molecule, so the proximal or distal oxygen of C4A can interact with either of its equivalent carbons, C1 or C12. To evaluate which is more likely, restraint MD simulations are performed where the distance between these oxygens and C1 or C12 is restricted to 3.50 Å. To evaluate which carbon atom is more promising to study the peroxyhemiacetal formation, the distance between O10 of C4A and C1 and C12 of $DBTO_2$ is restricted. To study the best carbon atom choice for the hydroperoxyl group transfer hypothesis, the distance between O11 of C4A and C1 and C12 of $DBTO_2$ is restricted.

3.3.5.1. For the peroxyhemiacetal formation

Analysis of the 10 ns restraint MD simulation with C1, shows that the distance restraints are imposed, as the distance between the atoms in study (O10, the distal oxygen of C4A and C1 of $DBTO_2$) is maintained around 3.50 Å during the simulation. To see if this atomic distance is plausible when it isn't being forced, a 10 ns free MD simulation is performed. Its analysis shows, on chain A, that the hydrogen bonds of the cofactor and substrate with the surrounding residues are maintained, except for the interaction between C4A and Arg155. As for the interatomic distance under study, it drifts to an average distance of 4.25 Å. So, for chain A, the imposed restraints did not favor the interaction between the O10 of C4A and the C1 of $DBTO_2$. The same is observed in chain B, where the interatomic distance under study is on average 4.12 Å. When imposing the same distance constraint of 3.50 Å during a 10 ns MD simulation to the distance between O10 of C4A and C12 of $DBTO_2$, the following 10 ns unrestrained simulation indicates that the distance increases to 5.10 Å on chain A and 4.09 Å on chain B. In chain A, the cofactor loses its hydrogen bonds with Ser135 and Arg155, and the substrate conserves its interactions. In chain B, the cofactor only loses its interaction with Arg155 and the substrate maintains its hydrogen bond with C4A.

Comparison of the results above can indicate that, for chain A (the one which is employed in QM/MM studies), the interaction of the distal oxygen of C4A is more likely to occur in C1 than in C12, as when set free the distance between C1 of C4A and O10 of the substrate does not increase as much as when the restraint with C12 is being employed. Additionally, observation on VMD shows that the orientation of the distal oxygen of C4A to C12 of $DBTO_2$ is not favorable. However, despite such results, all attempts made in this thesis with C1 as reaction coordinate do not render favorable results. Nonetheless, as seen in "3.3.4.1 Attempt with structure 28", there is still a slight possibility that C12 may be used instead of C1. Thus, to be sure, energy transit scans with C12 as one of the reaction coordinates are still attempted.

To test the hypotheses of the peroxyhemiacetal formation once again, C12 of $DBTO_2$ and O10 from C4A are used as coordinates in several structures. Structures 23 and 28 are considered because structure 23 presents a better distance and orientation between O10 of C4A and C12 of $DBTO_2$, and structure 28 yields the best results so far with C1. All these initial structures with the respective distance (marked as a dotted black line) between reaction coordinates are seen in Figure 31.

The distance between O10 and C12 in structure 23 is initially 4.32 Å. In 25 steps of 0.12 Å decrements, this distance is shortened to 1.44 Å. When compared with the same structure where C1 and O10 are approximated, structurally, similar results are seen: the bond between the proximal and distal oxygen of C4A breaks and O10 forms a new bond with the C12 of the substrate. Hence, the formation of the peroxyhemiacetal does not occur and the C-S bond does not break. Additionally, the hydrogen bond between the hydrogen of the hydroperoxyl and water is also lost. This scan has a high energy barrier of 50.3 kcal/mol (with C1 as reaction coordinate is 47.6 kcal/mol), and a final structure with an energy value of 33.7 kcal/mol (with C1 as reaction coordinate is 14.5 kcal/mol). When His312 is included in the reactive system, similar results are observed. As for the last attempt with C12 as reaction coordinate, in structure 28, the distance between O10 of the cofactor and C12 from the substrate starts as 5.07 Å and ends as 1.43 Å, upon 32 decrements of 0.12 Å. Nonetheless, the peroxyhemiacetal is also not formed here. Once again, the bond between the O10 and O11 of the hydroperoxyl breaks, and O10 bonds at C12. Ring opening of the substrate does not occur and the hydrogen bonds of the hydroperoxyl and water are lost. This is a very different and worse result than the one observed for the same initial structure when employing C1 as a coordinate. Similarly, this tendency is carried out to the energy scan, as the energy barrier is 49.0 kcal/mol (when using C1 is 36.0 kcal/mol) and the final structure has an energy of 20.8 kcal/mol (when using C1 is -19.3 kcal/mol). All the final structures from these attempts with C12 as one of the reaction coordinates are also depicted in Figure 31.

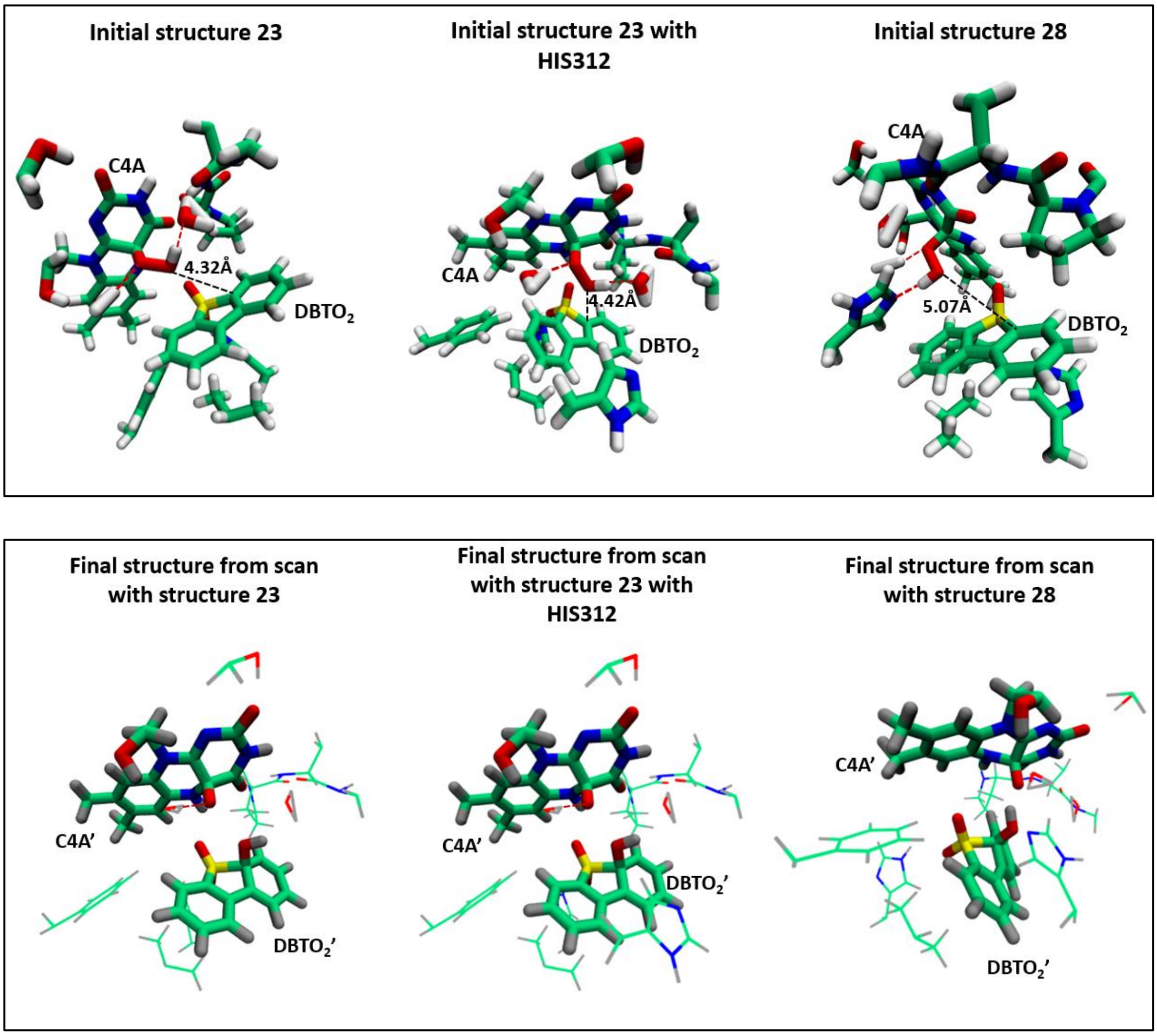

Initial structure 23
C4A
4.32Å
DBTO$_2$
Initial structure 23 with HIS312
C4A
4.42Å
DBTO$_2$
Initial structure 28
C4A
5.07Å
DBTO$_2$
Final structure from scan with structure 23
C4A'
DBTO$_2$'
Final structure from scan with structure 23 with HIS312
C4A'
DBTO$_2$'
Final structure from scan with structure 28
C4A'
DBTO$_2$'

Figure 31. Initial structures employed for the reaction coordinates O10 from C4A and C12 of DBTO$_2$ and final structures attained from such attempts. The black dotted lines mark the distance between O10 and C12 and the red dotted lines mark the hydrogen bonds where the hydroxyl group of C4A participates as a donor. C4A' and DBTO$_2$' represent the final ligand structures in these attempts.

Upon observation of the initial structures here employed, some hypotheses as to why the energy barriers are so high can be made. In all the structures, O10 is at a great distance from C12 of DBTO$_2$, especially in structure 28. Moreover, in all of them the distal oxygen is more favorably oriented to C1 instead of C12 and, additionally, the hydroperoxyl group establishes hydrogen bonds with water molecules and/or with His312. As all of these represent obstacles to the approximation of O10 and C12, they possess an energy cost, which possibly explains the high energy barriers here attained.

As seen, the results are not favorable. In fact, using C12 instead of C1 as reaction coordinate negatively impacts every result. Hence, when studying the hypothesis of a peroxyhemiacetal formation, C1 should be the preferred choice. As so, this means that every plausible pair of reaction coordinates to study the peroxyhemiacetal formation is employed in this thesis, and yet, none rendered the desired results. Consequently, new initial structures or hypotheses are still required.

3.3.5.2. For the hydroperoxyl group transfer

The distance between O11 of the cofactor and C1 of the substrate is restrained in a 10 ns *NPT* simulation around 3.50 Å. When these restraints are no longer being imposed in a following 10 ns simulation, the average distance between the atoms is 3.82 Å and 3.47 Å in chains A and B, respectively. Thus, when the 3.50 Å distance is no longer being imposed, it is still maintained. This is an indication that it is viable for such atoms to be in such proximity. As for the hydrogen bonds of the cofactor, in chain A, two are lost: with Ser135 and Arg155. Nonetheless, as previously seen, the interaction with Arg155 is not essential and the one with Ser135 further indicates that its role should be more evident during C4A activation from FMN and not in the reaction mechanism of C-S bond cleavage. In chain B, the hydrogen bond with Arg155 and Thr134 is also not present. In both chains, the substrates' interaction with N1 of C4A and His16 are maintained. As the essential hydrogen bonds are present for both ligands and the distance between O11 and C1 is maintained near the imposed distance of 3.50 Å, C1 looks like a good reaction coordinate to study the hydroperoxyl group transfer. Nonetheless, the hypothesis of using C12 instead is still on the table. However, when the same distance restraint applied to O10 and C12 is lifted in a 10 ns unrestricted simulation, the distance between the atoms increases. For chain A it has an average of 4.13 Å and for chain B of 3.96 Å. These values are slightly higher than expected, especially for chain A, and as such, the restrained distance does not hold up when no longer being forced. Additionally, all the substrate hydrogen bonds are less frequent, and the interactions with Arg155 and Ser135 are not present.

Comparison between the results when using C1 or C12 demonstrates that worse results are attained in terms of distances and interactions when the distance between O11 and C12 is restrained. This is in accordance with what is seen in Table 4, where at least for chain A, the average distance seen during a 100 ns simulation between O11 and C1 is smaller than the distance between O11 and C12. Thus, using O11 and C1 as reaction coordinates is the most promising choice to study the possible transfer of the hydroperoxyl group to the substrate. However, such attempt as demonstrated in "3.3.2 Attempt to transfer the hydroperoxyl group of C4A to the DBTO2 substrate", also did not render the expected results. Therefore, new initial structures or hypotheses are still required.

4. Conclusion and future perspectives

As the world is on the hunt for more environmentally friendly options, refineries are doomed to have to let go, at least partially, of harmful desulfurization techniques such as hydrodesulfurization. A greener desulfurization approach is biodesulfurization based on the 4S enzymatic pathway. Nonetheless, this alternative suffers from its own limitations, and thus, enzyme engineering is a necessity for its industrial application. However, to enable a rational optimization, the unknown structure and mechanism of reaction of DszA needs to be unveiled. Such discoveries are the aim of this thesis.

The DszA structure is attained as a tetramer through homology modeling, using as template the structure of BdsA (PDB ID:5xkd). Based on sequence identity, quality values and the Ramachandran plot, the model here attained has proven to be of quality and should be a representative and reliable structure of the intended enzyme. Upon rational transformations of this model, a DszA:C4A dimer is attained. Such is then considered as validated, as superimposition with the crystallographic structure of BdsA shows their similarities at the active center. Moreover, analysis of the active center shows hydrogen bonds between the ribitol-phosphate tail of C4A with the sidechain of Ser227, His152, Tyr156, Arg155 and Thr134, and the backbone of Ser227 and Leu226. As for the cofactor's isoalloxazine ring, it forms a hydrogen bond with the sidechain of Ser135 and two with the backbone of Asp55. Additionally, Val133 seems likely to interact hydrophobically with the ring of C4A. Comparison between these bonds with the ones thought to be essential for cofactor binding in the BdsA:FMN system, demonstrates that all the interactions seen for the BdsA's cofactor, FMN, also exist for C4A in DszA. Furthermore, additional interactions with Asp55, Thr134 and Ser227 are also seen. So mechanistic hypotheses could be studied, and the DszA:C4A:$DBTO_2$ system is created through several protein:ligand docking procedures on AutoDock Vina and GOLD. From all the substrate poses given by these, three are chosen as most promising. A number that is reduced to one, conformation A, after the study of their behavior during MD simulations, which identified its chain A as the one which kept most of the expected hydrogen bonds with the cofactor. Additionally, this chain also provides the best anchoring of the substrate, as its oxygens interact with Nε of His16 and N1 of C4A. Along the MD simulations, it is seen that the hydrogen bonds of C4A with Arg155 and Ser135 tend to be lost. Such can be an indication that Arg155 is not essential to hold the cofactor in position and that the role of Ser135 may be more limited to the C4A activation from FMN, rather than the reaction mechanism of C-S bond cleavage. Analysis of hydrophobic contacts seen in this chain demonstrates possible interactions of C4A with Pro54, Val133, Ala224, Gly225, Leu226, Phe242, Leu244 and Phe369 at a 3.5 Å radius and with Phe8, Leu53, Leu136 and Pro228 at a 6 Å radius. As for the substrate, just as C4A, hydrophobic contacts with Phe8, Pro54 and Phe369 are also seen. Additionally, interaction with Phe367, Phe9 and Phe242 is also present. Such hydrophobic contacts should be important to hold the ligands in their correct positions. A structure that is representative of this chain during the MD simulation is attained. As the DszA structure (in its apo and holo configuration) is attained, the first objective of this thesis is accomplished. However, fulfilling the second objective (unveiling the mechanism of reaction of DszA) is not as straightforward.

The above-mentioned representative structure is prepared for QM/MM studies where the two main hypotheses are studied: the peroxyhemiacetal formation proposed by Adak and Begley[61] and the transfer of the hydroperoxyl group of C4A to $DBTO_2$. Although attempted in several structures, the formation of the peroxyhemiacetal does not occur, hence the likelihood of this hypothesis being a true reaction step is low. At least with the structures employed in this thesis. Nonetheless, the best attempt at this mechanism is verified in an exothermic reaction, when a hydrogen bond between the hydrogen of the hydroperoxyl group of C4A and Nε of His312 is seen in the initial structure. Thus, indicating that the presence of a base that can deprotonate the hydroperoxyl group of C4A may be important, as hypothesized by Adak and Begley[61]. However, MD simulations showed that it is unlikely that His312 can be in such proximity to the hydroperoxyl group at the active center. Moreover, as seen for every mechanistic hypothesis in this thesis, the energy barrier is just too high (36.0 kcal/mol). Nonetheless, in this attempt, charge transfer analysis is performed, and the substrate is identified as the nucleophile. The hypothesis of the hydroperoxyl group transfer rendered several conformational changes and a very stable product. However, once more, due to a high energy barrier (52.7 kcal/mol), this should also not be a true reaction step in the attempted structures. Attempt of the same reaction in a triplet spin multiplicity also does not yield a favorable energy (46.4 kcal/mol). After every attempt deemed possible, a suitable energy transit scan that could represent a true step of the mechanism of reaction of DszA could not be attained. Hence, despite that the reaction mechanism of DszA could not be disclosed in this thesis, it was able to identify and clarify the reaction steps that are unlikely to be part of the enzyme's mechanism.

4.1. Future perspectives

As seen, with the structures and hypotheses applied throughout this thesis, none rendered the desired results. Hence, further steps in investigating the reaction mechanism involve either new structures and/or hypotheses. Matthews et al.[110] published an article in 2020, where it is suggested that flavin monooxygenases, which up until now were thought to use exclusively C4A as their oxygen-transferring intermediate, can as an alternative employ a flavin-N5-peroxide as a nucleophile for catalysis. Moreover, in their research on RutA, they concluded that group C flavin monooxygenases (which includes DszA) can in fact use flavin-N5-peroxide. Inclusively, it proposes a reaction mechanism for DszA, shown in Figure 32, where this cofactor enables the transfer of its distal oxygen as OH^- to the substrate, culminating in the cleavage of the C-S bond of $DBTO_2$, and the desired product HBPS, along with the formation of flavin-N5-oxide, as a coproduct.

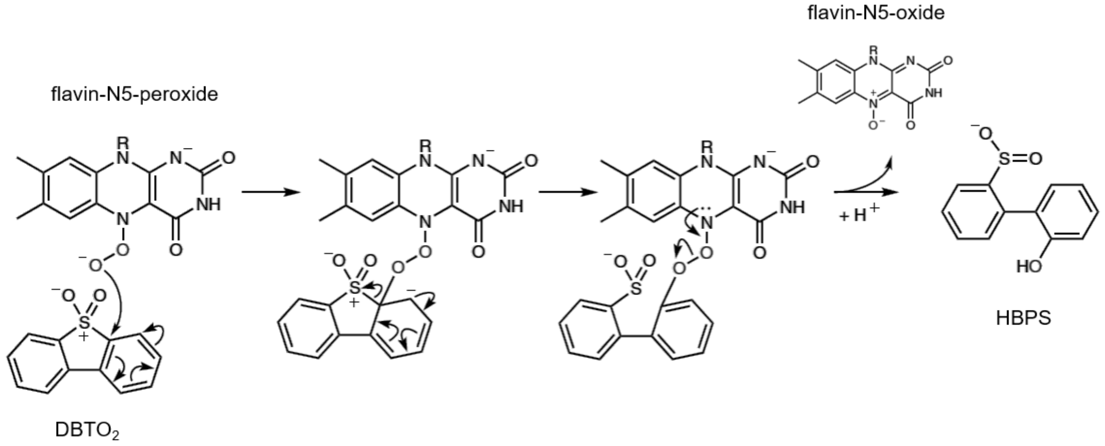


Figure 32. C–S bond cleavage of dibenzothiophene sulfone catalyzed by DszA. The redox-neutral oxygenative cleavage of carbon-hetero bond is enabled by flavin-N5-peroxide, which is then converted into flavin-N5-oxide.

In this thesis, due to time constraints, such hypothesis is not studied. However, considering that all the attempts here made with C4A do not render favorable results, the future of mechanistic studies of DszA should include the hypothesis of using flavin-N5-peroxide as the cofactor.

5. References

(1) Ritchie, H. Fossil Fuels. *Our World Data* **2017**.

(2) Vazquez-Duhalt, R.; Torres, E.; Valderrama, B.; Le Borgne, S. Will Biochemical Catalysis Impact the Petroleum Refining Industry? *Energy & Fuels* **2002**, *16* (5), 1239–1250. https://doi.org/10.1021/ef020038s.

(3) Kiang, Y.-H. Chapter 3 - Basic Properties of Fuels, Biomass, Refuse Derived Fuels, Wastes, Biosludge, and Biocarbons; Kiang, Y.-H. B. T.-F. P. E. and C. P. C., Ed.; Academic Press, 2018; pp 41–65. https://doi.org/https://doi.org/10.1016/B978-0-12-813473-3.00003-9.

(4) Sarkar, D. K. Chapter 1 - General Description of Thermal Power Plants; Sarkar, D. K. B. T.-T. P. P., Ed.; Elsevier, 2017; pp 1–31. https://doi.org/https://doi.org/10.1016/B978-0-08-101112-6.00001-0.

(5) Balachandar, G.; Khanna, N.; Das, D. Chapter 6 - Biohydrogen Production from Organic Wastes by Dark Fermentation; Pandey, A., Chang, J.-S., Hallenbecka, P. C., Larroche, C. B. T.-B., Eds.; Elsevier: Amsterdam, 2013; pp 103–144. https://doi.org/https://doi.org/10.1016/B978-0-444-59555-3.00006-4.

(6) Mohebali, G.; Ball, A. S. Biodesulfurization of Diesel Fuels – Past, Present and Future Perspectives. *Int. Biodeterior. Biodegradation* **2016**, *110*, 163–180. https://doi.org/https://doi.org/10.1016/j.ibiod.2016.03.011.

(7) Ciesielski, T. Climate Change and Public Health: A Small Frame Obscures the Picture. *NEW Solut. A J. Environ. Occup. Heal. Policy* **2017**, *27* (1), 8–11. https://doi.org/10.1177/1048291117691075.

(8) BP. Statistical Review of World Energy, 2020 | 69th Edition. *Bp* **2020**, 66.

(9) https://www.worldometers.info/oil/ https://www.worldometers.info/oil/ (accessed Mar 31, 2021).

(10) U.S. Energy Information Administration - Office for Energy Analysis. *International Energy Outlook 2019 with Projections to 2050*; 2019.

(11) Morales Pedraza, J. Chapter 1 - General Overview of the Energy Sector in the North America Region; Morales Pedraza, J. B. T.-C. E. in N. A., Ed.; Elsevier, 2019; pp 1–87. https://doi.org/https://doi.org/10.1016/B978-0-12-814889-

1.00001-2.

(12) U.S. Energy Information Administration. *Petroleum Supply Annual 2019*; 2020.

(13) Demirbas, A.; Alidrisi, H.; Balubaid, M. A. API Gravity, Sulfur Content, and Desulfurization of Crude Oil. *Pet. Sci. Technol.* **2015**, *33* (1), 93–101. https://doi.org/10.1080/10916466.2014.950383.

(14) U.S. Energy Information Administration. Oil and petroleum products explained, Refining crude oil - Inputs & Outputs https://www.eia.gov/energyexplained/oil-and-petroleum-products/refining-crude-oil-inputs-and-outputs.php (accessed Sep 29, 2020).

(15) Soleimani, M.; Bassi, A.; Margaritis, A. Biodesulfurization of Refractory Organic Sulfur Compounds in Fossil Fuels. *Biotechnol. Adv.* **2007**, *25* (6), 570–596. https://doi.org/https://doi.org/10.1016/j.biotechadv.2007.07.003.

(16) Javadli, R.; de Klerk, A. Desulfurization of Heavy Oil. *Appl. Petrochemical Res.* **2012**, *1* (1), 3–19. https://doi.org/10.1007/s13203-012-0006-6.

(17) J.Butler, T.; E.Likens, G. Acid rain https://www.britannica.com/science/acid-rain (accessed Sep 30, 2020).

(18) Duissenov, D. Production and Processing of Sour Crude and Natural Gas - Challenges Due to Increasing Stringent Regulations. **2013**, No. June, 101.

(19) The International Council On Clean Transportation. An Introduction To Petroleum Refining and the Production of Ultra Low Sulfur Gasoline. *Energy Econ. Appl. Optim.* **2011**, 1–38.

(20) EU fuels:Diesel and Gasoline https://www.transportpolicy.net/standard/eu-fuels-diesel-and-gasoline/ (accessed Dec 28, 2020).

(21) China fuels: Diesel and Gasoline https://www.transportpolicy.net/standard/china-fuels-diesel-and-gasoline/ (accessed Dec 28, 2020).

(22) US fuels: Diesel and Gasoline https://www.transportpolicy.net/standard/us-fuels-diesel-and-gasoline/ (accessed Dec 28, 2020).

(23) Chandra Srivastava, V. An Evaluation of Desulfurization Technologies for Sulfur Removal from Liquid Fuels. *RSC Adv.* **2012**, *2* (3), 759–783. https://doi.org/10.1039/C1RA00309G.

(24) Bai, P.; Etim, U. J.; Yan, Z.; Mintova, S.; Zhang, Z.; Zhong, Z.; Gao, X. Fluid

Catalytic Cracking Technology: Current Status and Recent Discoveries on Catalyst Contamination. *Catal. Rev.* **2019**, *61* (3), 333–405. https://doi.org/10.1080/01614940.2018.1549011.

(25) Li, Y.-X.; Jiang, W.-J.; Tan, P.; Liu, X.-Q.; Zhang, D.-Y.; Sun, L.-B. What Matters to the Adsorptive Desulfurization Performance of Metal-Organic Frameworks? *J. Phys. Chem. C* **2015**, *119* (38), 21969–21977. https://doi.org/10.1021/acs.jpcc.5b07546.

(26) Song, C. An Overview of New Approaches to Deep Desulfurization for Ultra-Clean Gasoline, Diesel Fuel and Jet Fuel. *Catal. Today* **2003**, *86* (1), 211–263. https://doi.org/https://doi.org/10.1016/S0920-5861(03)00412-7.

(27) Abro, R.; Abdeltawab, A.; Al-Salem, S.; Yu, G.; Qazi, A.; Gao, S.; Chen, X. A Review of Extractive Desulfurization of Fuel Oils Using Ionic Liquids. *RSC Adv.* **2014**, *4*. https://doi.org/10.1039/C4RA03478C.

(28) Boniek, D.; Figueiredo, D.; dos Santos, A. F. B.; de Resende Stoianoff, M. A. Biodesulfurization: A Mini Review about the Immediate Search for the Future Technology. *Clean Technol. Environ. Policy* **2015**, *17* (1), 29–37. https://doi.org/10.1007/s10098-014-0812-x.

(29) Sadare, O. O.; Obazu, F.; Daramola, M. O. Biodesulfurization of Petroleum Distillates—Current Status, Opportunities and Future Challenges. *Environ. - MDPI* **2017**, *4* (4), 1–20. https://doi.org/10.3390/environments4040085.

(30) Anteneh, Y. S.; Franco, C. M. M. Whole Cell Actinobacteria as Biocatalysts. *Front. Microbiol.* **2019**, *10*, 77. https://doi.org/10.3389/fmicb.2019.00077.

(31) Borhani, M. S.; Etemadifar, Z. Enhancement/Evolution of Biodesulfurization 4S Pathway by Genetic Engineering and Bioinformatic Approaches. *Adv. Res. Microb. Metab. Technol.* **2019**, *2* (1), 13–23. https://doi.org/10.22104/armmt.2020.3783.1034.

(32) Abin-Fuentes, A.; Mohamed, M. E.-S.; Wang, D. I. C.; Prather, K. L. J. Exploring the Mechanism of Biocatalyst Inhibition in Microbial Desulfurization. *Appl. Environ. Microbiol.* **2013**, *79* (24), 7807–7817. https://doi.org/10.1128/AEM.02696-13.

(33) van der Geize, R.; Dijkhuizen, L. Harnessing the Catabolic Diversity of Rhodococci for Environmental and Biotechnological Applications. *Curr. Opin. Microbiol.* **2004**, *7* (3), 255–261. https://doi.org/https://doi.org/10.1016/j.mib.2004.04.001.

(34) Ayala, M.; Vazquez-Duhalt, R. Chapter 3 Enzymatic Catalysis on Petroleum Products. In *Petroleum Biotechnology*; Vazquez-Duhalt, R., Quintero-Ramirez, R. B. T.-S. in S. S. and C., Eds.; Elsevier, 2004; Vol. 151, pp 67–111. https://doi.org/https://doi.org/10.1016/S0167-2991(04)80144-7.

(35) Kotlar, H. K.; Brakstad, O. G.; Markussen, S.; Winnberg, A. Chapter 1 Use of Petroleum Biotechnology throughout the Value Chain of an Oil Company: An Integrated Approach. In *Petroleum Biotechnology*; Vazquez-Duhalt, R., Quintero-Ramirez, R. B. T.-S. in S. S. and C., Eds.; Elsevier, 2004; Vol. 151, pp 1–27. https://doi.org/https://doi.org/10.1016/S0167-2991(04)80142-3.

(36) Sousa, J. P. M.; Ferreira, P.; Neves, R. P. P.; Ramos, M. J.; Fernandes, P. A. The Bacterial 4S Pathway – an Economical Alternative for Crude Oil Desulphurization That Reduces CO2 Emissions. *Green Chem.* **2020**, *22* (22), 7604–7621. https://doi.org/10.1039/D0GC02055A.

(37) Díaz, E.; García, J. L. Genetics Engineering for Removal of Sulfur and Nitrogen from Fuel Heterocycles BT - Handbook of Hydrocarbon and Lipid Microbiology; Timmis, K. N., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2010; pp 2787–2801. https://doi.org/10.1007/978-3-540-77587-4_206.

(38) Li, S.; Ma, T. The Desulfurization Pathway in Rhodococcus. In *Biology of Rhodococcus*; Alvarez, H. M., Ed.; Springer International Publishing: Cham, 2019; Vol. 16, pp 203–229.

(39) Gupta, N.; Roychoudhury, P. K.; Deb, J. K. Biotechnology of Desulfurization of Diesel: Prospects and Challenges. *Appl. Microbiol. Biotechnol.* **2005**, *66* (4), 356–366. https://doi.org/10.1007/s00253-004-1755-7.

(40) McFarland, B. L. Biodesulfurization. *Curr. Opin. Microbiol.* **1999**, *2* (3), 257–264. https://doi.org/10.1016/s1369-5274(99)80045-9.

(41) Huijbers, M. M. E.; Montersino, S.; Westphal, A. H.; Tischler, D.; van Berkel, W. J. H. Flavin Dependent Monooxygenases. *Arch. Biochem. Biophys.* **2014**, *544*, 2–17. https://doi.org/https://doi.org/10.1016/j.abb.2013.12.005.

(42) Guan, L.-J.; Lee, W. C.; Wang, S.; Ohshiro, T.; Izumi, Y.; Ohtsuka, J.; Tanokura, M. Crystal Structures of Apo-DszC and FMN-Bound DszC from Rhodococcus Erythropolis D-1. *FEBS J.* **2015**, *282* (16), 3126–3135. https://doi.org/https://doi.org/10.1111/febs.13216.

(43) Liu, S.; Zhang, C.; Su, T.; Wei, T.; Zhu, D.; Wang, K.; Huang, Y.; Dong, Y.; Yin,

K.; Xu, S.; Xu, P.; Gu, L. Crystal Structure of DszC from Rhodococcus Sp. XP at 1.79 Å. *Proteins Struct. Funct. Bioinforma.* **2014**, *82* (9), 1708–1720. https://doi.org/https://doi.org/10.1002/prot.24525.

(44) Barbosa, A. C. C.; Neves, R. P. P.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Mechanistic Studies of a Flavin Monooxygenase: Sulfur Oxidation of Dibenzothiophenes by DszC. *ACS Catal.* **2018**, *8* (10), 9298–9311. https://doi.org/10.1021/acscatal.8b01877.

(45) Li, L.; Ye, L.; Lin, Y.; Zhang, W.; Liao, X.; Liang, S. Enhancing the Substrate Tolerance of DszC by a Combination of Alanine Scanning and Site-Directed Saturation Mutagenesis. *J. Ind. Microbiol. Biotechnol.* **2020**, *47* (4–5), 395–402. https://doi.org/10.1007/s10295-020-02274-8.

(46) Torktaz, I.; Etemadifar, Z.; Derikvand, P. Comparative Modeling of DszC, an Enzyme in Biodesulfurization, and Performing in Silico Point Mutation for Increasing Tendency to Oil. *Bioinformation* **2012**, *8* (5), 246–250. https://doi.org/10.6026/97320630008246.

(47) Lee, W. C.; Ohshiro, T.; Matsubara, T.; Izumi, Y.; Tanokura, M. Crystal Structure and Desulfurization Mechanism of 2′-Hydroxybiphenyl-2-Sulfinic Acid Desulfinase*. *J. Biol. Chem.* **2006**, *281* (43), 32534–32539. https://doi.org/https://doi.org/10.1074/jbc.M602974200.

(48) Sousa, J. P. M.; Neves, R. P. P.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Reaction Mechanism and Determinants for Efficient Catalysis by DszB, a Key Enzyme for Crude Oil Bio-Desulfurization. *ACS Catal.* **2020**, *10* (16), 9545–9554. https://doi.org/10.1021/acscatal.0c03122.

(49) Yu, Y.; Mills, L. C.; Englert, D. L.; Payne, C. M. Inhibition Mechanisms of Rhodococcus Erythropolis 2′-Hydroxybiphenyl-2-Sulfinate Desulfinase (DszB). *J. Phys. Chem. B* **2019**, *123* (43), 9054–9065. https://doi.org/10.1021/acs.jpcb.9b05252.

(50) Li, G.; Li, S.; Zhang, M.; Wang, J.; Zhu, L.; Liang, F.; Liu, R.; Ma, T. Genetic Rearrangement Strategy for Optimizing the Dibenzothiophene Biodesulfurization Pathway in &lt;Em&gt;Rhodococcus Erythropolis&lt;/Em&gt; *Appl. Environ. Microbiol.* **2008**, *74* (4), 971 LP – 976. https://doi.org/10.1128/AEM.02319-07.

(51) Reichmuth, D. S.; Blanch, H. W.; Keasling, J. D. Dibenzothiophene Biodesulfurization Pathway Improvement Using Diagnostic GFP Fusions.

Biotechnol. Bioeng. **2004**, *88* (1), 94–99. https://doi.org/10.1002/bit.20220.

(52) OHSHIRO, T.; OHKITA, R.; TAKIKAWA, T.; MANABE, M.; LEE, W. C.; TANOKURA, M.; IZUMI, Y. Improvement of 2′-Hydroxybiphenyl-2-Sulfinate Desulfinase, an Enzyme Involved in the Dibenzothiophene Desulfurization Pathway, from Rhodococcus Erythropolis KA2-5-1 by Site-Directed Mutagenesis. *Biosci. Biotechnol. Biochem.* **2007**, *71* (11), 2815–2821. https://doi.org/10.1271/bbb.70436.

(53) Fallahzadeh, R.; Bambai, B.; Esfahani, K.; Sepahi, A. A. Simulation-Based Protein Engineering of R. Erythropolis FMN Oxidoreductase (DszD). *Heliyon* **2019**, *5* (8), e02193. https://doi.org/https://doi.org/10.1016/j.heliyon.2019.e02193.

(54) Gray, K. A.; Mrachko, G. T.; Squires, C. H. Biodesulfurization of Fossil Fuels. *Curr. Opin. Microbiol.* **2003**, *6* (3), 229–235. https://doi.org/https://doi.org/10.1016/S1369-5274(03)00065-1.

(55) Sousa, S. F.; Sousa, J. F. M.; Barbosa, A. C. C.; Ferreira, C. E.; Neves, R. P. P.; Ribeiro, A. J. M.; Fernandes, P. A.; Ramos, M. J. Improving the Biodesulfurization of Crude Oil and Derivatives: A QM/MM Investigation of the Catalytic Mechanism of NADH-FMN Oxidoreductase (DszD). *J. Phys. Chem. A* **2016**, *120* (27), 5300–5306. https://doi.org/10.1021/acs.jpca.6b01536.

(56) Matsubara, T.; Ohshiro, T.; Nishina, Y.; Izumi, Y. Purification, Characterization, and Overexpression of Flavin Reductase Involved in Dibenzothiophene Desulfurization By&lt;Em&gt;Rhodococcus Erythropolis&lt;/Em&gt; D-1. *Appl. Environ. Microbiol.* **2001**, *67* (3), 1179 LP – 1184. https://doi.org/10.1128/AEM.67.3.1179-1184.

(57) Kamali, N.; Tavallaie, M.; Bambai, B.; Karkhane, A. A.; Miri, M. Site-Directed Mutagenesis Enhances the Activity of NADH-FMN Oxidoreductase (DszD) Activity of Rhodococcus Erythropolis. *Biotechnol. Lett.* **2010**, *32* (7), 921–927. https://doi.org/10.1007/s10529-010-0254-4.

(58) Su, T.; Su, J.; Liu, S.; Zhang, C.; He, J.; Huang, Y.; Xu, S.; Gu, L. Structural and Biochemical Characterization of BdsA from Bacillus Subtilis WU-S2B, a Key Enzyme in the “4S” Desulfurization Pathway. *Frontiers in Microbiology*. 2018, p 231. https://doi.org/10.3389/fmicb.2018.00231.

(59) Ohshiro, T.; Kojima, T.; Torii, K.; Kawasoe, H.; Izumi, Y. Purification and Characterization of Dibenzothiophene (DBT) Sulfone Monooxygenase, an

Enzyme Involved in DBT Desulfurization, from Rhodococcus Erythropolis D-1. *J. Biosci. Bioeng.* **1999**, *88* (6), 610–616. https://doi.org/https://doi.org/10.1016/S1389-1723(00)87088-7.

(60) Karnwal, A.; Kaur, H. Comparative Analysis of Physicochemical Properties and Sequence of DszA Protein in Some Microorganisms through in Silico Approach. *Ecoterra* **2012**.

(61) Adak, S.; Begley, T. P. Dibenzothiophene Catabolism Proceeds via a Flavin-N5-Oxide Intermediate. *J. Am. Chem. Soc.* **2016**, *138* (20), 6424–6426. https://doi.org/10.1021/jacs.6b00583.

(62) Kilbane, J. J. Biodesulfurization: How to Make It Work? *Arab. J. Sci. Eng.* **2017**, *42* (1), 1–9. https://doi.org/10.1007/s13369-016-2269-1.

(63) Bhasin, M.; Raghava, G. P. S. 8 - Computational Methods in Genome Research. In *Applied Mycology and Biotechnology*; Arora, D. K., Berka, R. M., Singh, G. B. B. T.-A. M. and B., Eds.; Elsevier, 2006; Vol. 6, pp 179–207. https://doi.org/https://doi.org/10.1016/S1874-5334(06)80011-0.

(64) Skariyachan, S.; Garka, S. Exploring the Binding Potential of Carbon Nanotubes and Fullerene towards Major Drug Targets of Multidrug Resistant Bacterial Pathogens and Their Utility as Novel Therapeutic Agents. In *Fullerens, Graphenes and Nanotubes*; 2018; pp 1–29. https://doi.org/10.1016/B978-0-12-813691-1.00001-4.

(65) Gromiha, M. M.; Nagarajan, R.; Selvaraj, S. Protein Structural Bioinformatics: An Overview; 2021; pp 445–459. https://doi.org/10.1016/b978-0-12-809633-8.20278-1.

(66) Robinson, S.; Afzal, A.; Leader, D. Bioinformatics: Concepts, Methods, and Data. In *Handbook of Pharmacogenomics and Stratified Medicine*; 2014; pp 259–287. https://doi.org/10.1016/B978-0-12-386882-4.00013-X.

(67) Swiss-Model https://swissmodel.expasy.org/docs/help (accessed May 4, 2021).

(68) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein–Ligand Docking: Current Status and Future Challenges. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (1), 15–26. https://doi.org/https://doi.org/10.1002/prot.21082.

(69) Morris, G. M.; Lim-Wilby, M. Molecular Docking. *Methods Mol. Biol.* **2008**, *443*, 365–382. https://doi.org/10.1007/978-1-59745-177-2_19.

(70) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges . *Frontiers in Pharmacology* . 2018, p 1089.

(71) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-Small Molecule Docking Methods. *J. Comput. Aided. Mol. Des.* **2002**, *16* (3), 151–166. https://doi.org/10.1023/a:1020155510718.

(72) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9* (2), 91–102. https://doi.org/10.1007/s12551-016-0247-1.

(73) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking11Edited by F. E. Cohen. *J. Mol. Biol.* **1997**, *267* (3), 727–748. https://doi.org/https://doi.org/10.1006/jmbi.1996.0897.

(74) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245* (1), 43–53. https://doi.org/https://doi.org/10.1016/S0022-2836(95)80037-9.

(75) GOLD User Guide A Component of the CSD-Discovery Suite. 2019.

(76) What is the difference between the GoldScore, ChemScore, ASP and ChemPLP scoring functions provided with GOLD? https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=5d1a2fc0-c93a-49c3-a8e2-f95c472dcff0 (accessed Jul 11, 2021).

(77) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96. https://doi.org/10.1021/ci800298z.

(78) Vieira, T. F.; Sousa, S. F. Comparing AutoDock and Vina in Ligand/Decoy Discrimination for Virtual Screening. *Applied Sciences* . 2019. https://doi.org/10.3390/app9214538.

(79) Huang, S.-Y. Comprehensive Assessment of Flexible-Ligand Docking Algorithms: Current Effectiveness and Challenges. *Brief. Bioinform.* **2018**, *19* (5), 982–994. https://doi.org/10.1093/bib/bbx030.

(80) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.21334.

(81) AutoDock Vina Manual http://vina.scripps.edu/manual.html (accessed Jul 12, 2021).

(82) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48* (12), 4111–4119. https://doi.org/10.1021/jm048957q.

(83) Leach, A. R. *Molecular Modelling : Principles and Applications*; Prentice Hall: Harlow, England; New York, 2001.

(84) Jensen, F. *An Introduction to Computational Chemistry*, 2nd ed.; John Wiley & Sons, Ltd, 1989.

(85) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models, 2nd Edition*, second.; Wiley, 2004.

(86) Poltev, V. Molecular Mechanics: Principles, History, and Current Status BT - Handbook of Computational Chemistry; Leszczynski, J., Kaczmarek-Kedziera, A., Puzyn, T., G. Papadopoulos, M., Reis, H., K. Shukla, M., Eds.; Springer International Publishing: Cham, 2017; pp 21–67. https://doi.org/10.1007/978-3-319-27282-5_9.

(87) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85. https://doi.org/10.1016/s0065-3233(03)66002-x.

(88) Woods, R. J.; Chappelle, R. Restrained Electrostatic Potential Atomic Partial Charges for Condensed-Phase Simulations of Carbohydrates. *Theochem* **2000**, *527* (1–3), 149–156. https://doi.org/10.1016/S0166-1280(00)00487-5.

(89) AMBER force fields https://ambermd.org/AmberModels.php (accessed Jul 28, 2021).

(90) Sundar, V. C. Vibration Analysis of Heme Porphyrin, Middlebury College, 1997.

(91) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.

(92) González, M. A. Force Fields and Molecular Dynamics Simulations. *JDN* **2011**, *12*, 169–200.

(93) Han, X. Chapter 11 - Mechanism of Nanomachining Semiconductor and Ceramic Blades for Surgical Applications; Grumezescu, A. M. B. T.-E. of N., Ed.; William Andrew Publishing, 2016; pp 329–358. https://doi.org/https://doi.org/10.1016/B978-0-323-41532-3.00011-7.

(94) Oostenbrink, C.; van Lipzig, M. M. H.; van Gunsteren, W. F. 4.25 - Applications of Molecular Dynamics Simulations in Drug Design; Taylor, J. B., Triggle, D. J. B. T.-C. M. C. I. I., Eds.; Elsevier: Oxford, 2007; pp 651–668. https://doi.org/https://doi.org/10.1016/B0-08-045044-X/00268-6.

(95) Sousa, J. P. M. QM/MM Study of DszB Reaction Mechanism for Crude Oil Biodesulphurization, 2018.

(96) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. A Practical Guide to Modelling Enzyme-Catalysed Reactions. *Chem. Soc. Rev.* **2012**, *41* (8), 3025–3038. https://doi.org/10.1039/C2CS15297E.

(97) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; 2001.

(98) Friesner, R. A. Ab Initio Quantum Chemistry: Methodology and Applications. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6648 LP – 6653. https://doi.org/10.1073/pnas.0408036102.

(99) Sherril, C. D. *An Introduction to Hartree-Fock Molecular Orbital Theory*; 2000.

(100) Neves, R. P. P. Metals' Data for Biomolecular Force Fields, University of Porto, 2016.

(101) Śmiga, S.; Buksztel, A.; Grabowski, I. Chapter 7 - Density-Dependent Exchange–Correlation Potentials Derived From Highly Accurate Ab Initio Calculations. In *Proceedings of MEST 2012: Electronic structure methods with applications to experimental chemistry*; Hoggan, P. B. T.-A. in Q. C., Ed.; Academic Press, 2014; Vol. 68, pp 125–151. https://doi.org/https://doi.org/10.1016/B978-0-12-800536-1.00007-1.

(102) Foresman, J. B.; Frisch, A. E.; Gaussian, I. *Exploring Chemistry with Electronic Structure Methods*; Gaussian, Incorporated, 1996.

(103) Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140* (18), 18A301. https://doi.org/10.1063/1.4869598.

(104) Zhang, Y.; Yang, W. A Challenge for Density Functionals: Self-Interaction Error Increases for Systems with a Noninteger Number of Electrons. *J. Chem. Phys.* **1998**, *109* (7), 2604–2608. https://doi.org/10.1063/1.476859.

(105) Becke, A. D. Real-Space Post-Hartree–Fock Correlation Models. *J. Chem. Phys.*

2005, *122* (6), 64101. https://doi.org/10.1063/1.1844493.

(106) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Application of Quantum Mechanics/Molecular Mechanics Methods in the Study of Enzymatic Reaction Mechanisms. *WIREs Comput. Mol. Sci.* **2017**, *7* (2), e1281. https://doi.org/https://doi.org/10.1002/wcms.1281.

(107) Calixto, A. R.; Ramos, M. J.; Fernandes, P. A. Influence of Frozen Residues on the Exploration of the PES of Enzyme Reaction Mechanisms. *J. Chem. Theory Comput.* **2017**, *13* (11), 5486–5495. https://doi.org/10.1021/acs.jctc.7b00768.

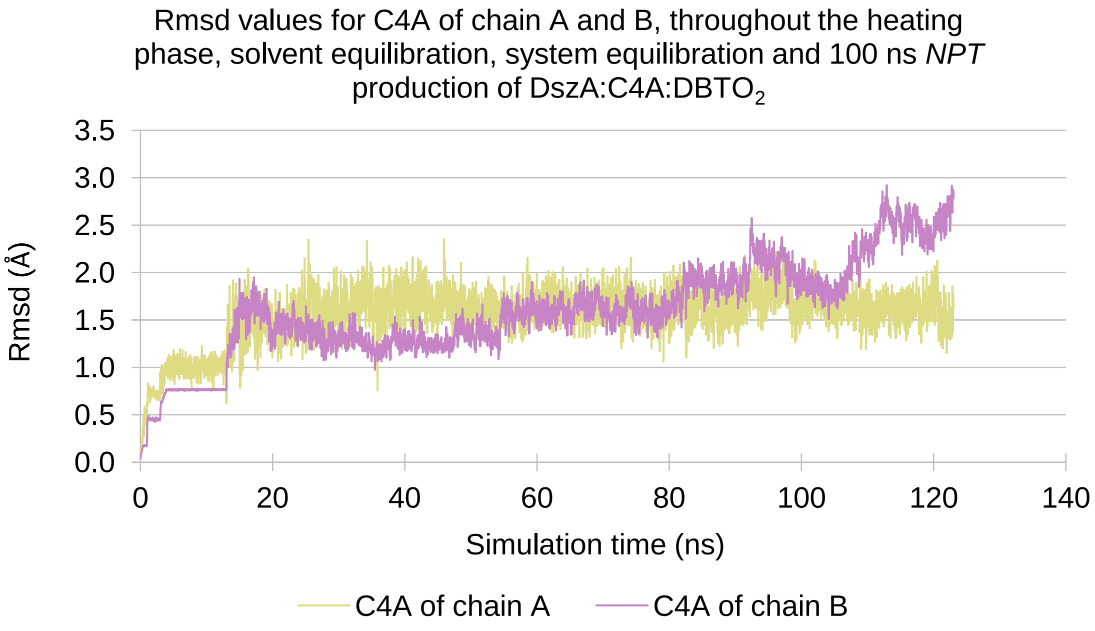
(108) S. Fernandes, H.; Ramos, M. J.; M. F. S. A. Cerqueira, N. MolUP: A VMD Plugin to Handle QM and ONIOM Calculations Using the Gaussian Software. *J. Comput. Chem.* **2018**, *39* (19), 1344–1353. https://doi.org/https://doi.org/10.1002/jcc.25189.

(109) Wiederstein, M.; Sippl, M. J. ProSA-Web: Interactive Web Service for the Recognition of Errors in Three-Dimensional Structures of Proteins. *Nucleic Acids Res.* **2007**, *35* (suppl_2), W407–W410. https://doi.org/10.1093/NAR/GKM290.

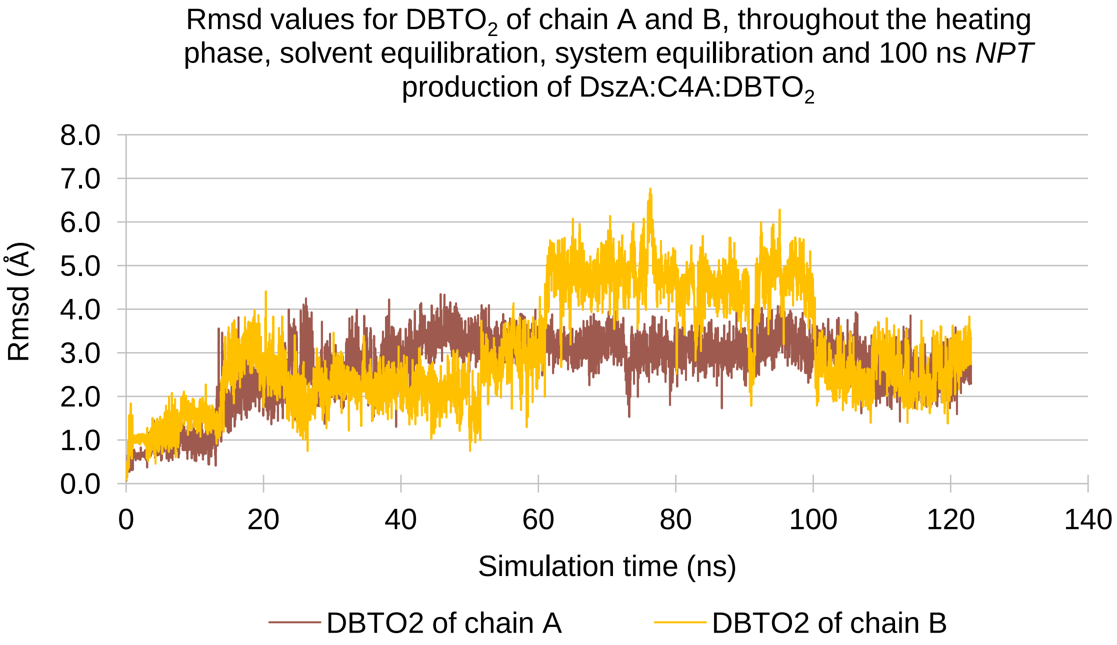
(110) Matthews, A.; Saleem-Batcha, R.; Sanders, J. N.; Stull, F.; Houk, K. N.; Teufel, R. Aminoperoxide Adducts Expand the Catalytic Repertoire of Flavin Monooxygenases. *Nat. Chem. Biol.* **2020**, *16* (5), 556–563. https://doi.org/10.1038/s41589-020-0476-2.

6. Annexes

6.1. Rmsd graphs for conformation A of DszA:C4A:DBTO2



Graph 14. Rmsd values for C4A of chain A and B of DszA:C4A:DBTO$_2$, throughout the heating phase, the solvent equilibration, equilibration of the entire system and 100 ns *NPT* production.



Graph 15. Rmsd values for DBTO$_2$ of chain A and B of DszA:C4A:DBTO$_2$, throughout the heating phase, the solvent equilibration, equilibration of the entire system and 100 ns *NPT* production.

6.2. Proton transfer from N1 of C4A to O2 of $DBTO_2$

Table 5. Energy transit scan characteristics and respective conclusions when attempting a proton transfer from N1 of C4A to O2 of $DBTO_2$.

Energy transit scan characteristics	**Conclusions**
Approximation of the hydrogen of N1 from C4A and O2 from $DBTO_2$, from 2.36 Å to 1.16 Å in 11 steps of -0.12 Å.	A valid final structure is not seen, as the energy is still rising at the last scan point and no transition state is seen along it. Moreover, such energy is already too high at 32.4 kcal/mol (Graph 16).

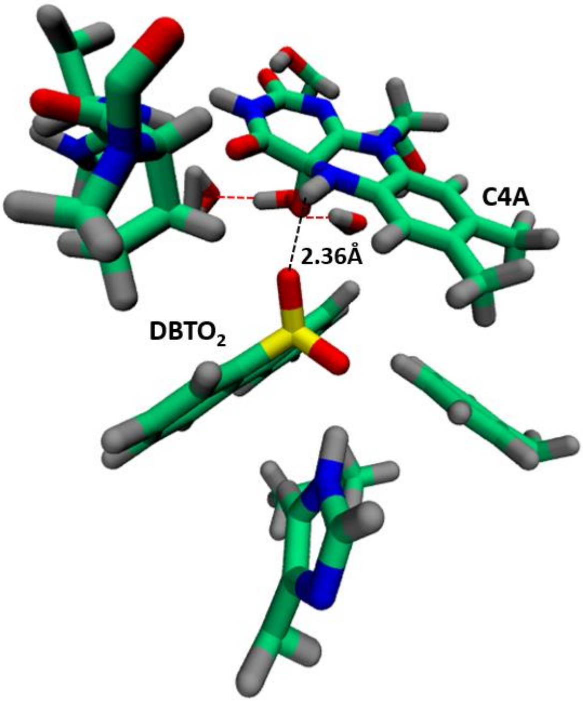
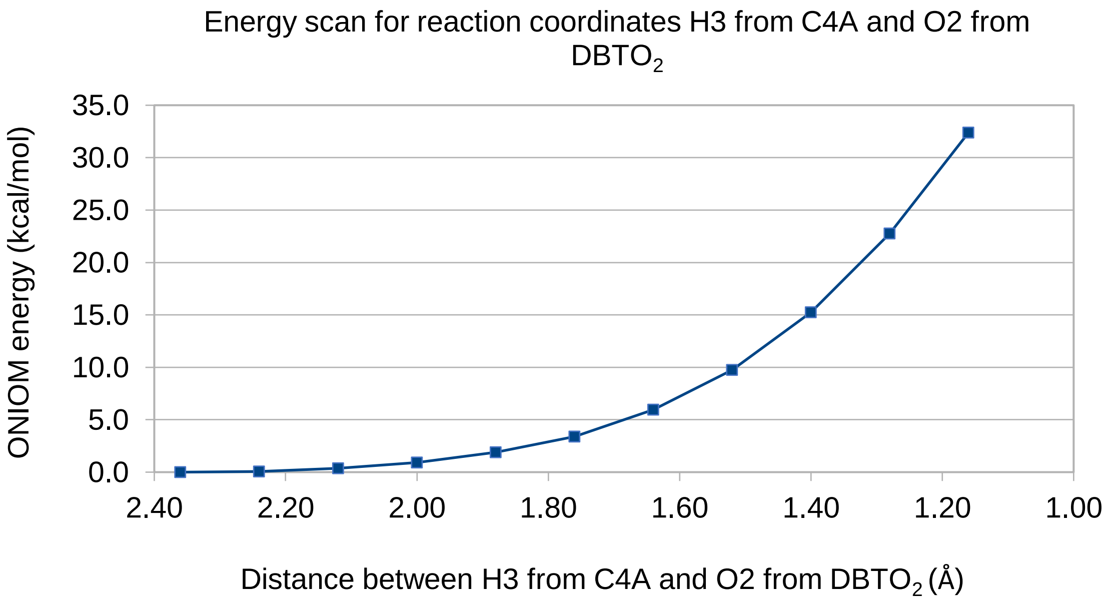


Figure 33. Initial structure of the transit scan for along the interatomic distance formed by the hydrogen of N1 in C4A and the O2 from $DBTO_2$. The initial distance between the reaction coordinates is marked by the dotted black line (2.36 Å). Hydrogen bonds that have O as donor atom are marked in red dotted lines.



Graph 16. Energy scan along the distance from the hydrogen of N1 in C4A and O2 from $DBTO_2$.

6.3. Protonation of His312 by a water molecule in structure 23

Table 6. Energy transit scan characteristics and respective conclusions when attempting protonation of His312 by a water molecule in structure 23.

Energy transit scan characteristics	**Conclusions**
Approximation of hydrogen from Wat26758 and Nε from His312, from 3.74 Å to 0.98 Å in 24 steps of -0.12 Å.	A valid final structure is not seen, as the energy is still rising at the last scan point and no transition state is seen along it. Moreover, such energy is already too high at 77.7 kcal/mol (Graph 17).

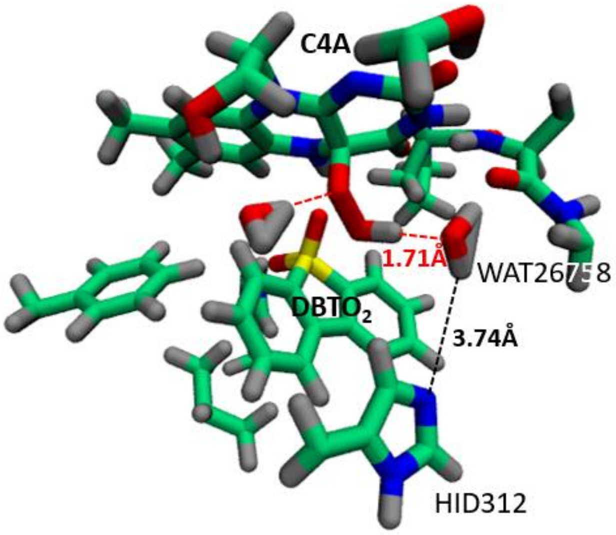
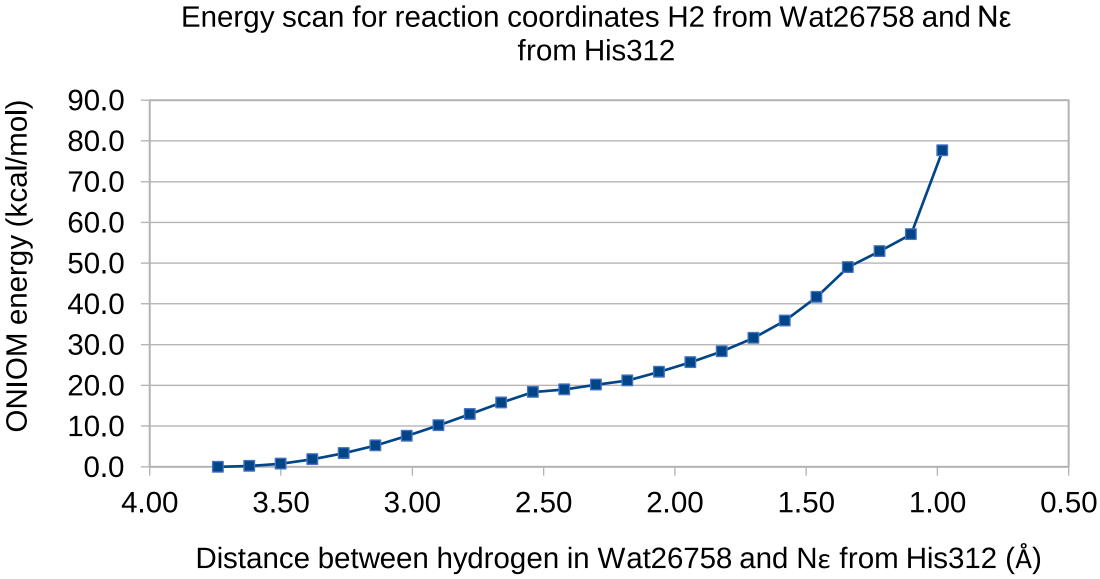


Figure 34. Initial structure of the transit scan for the interatomic distance comprising the hydrogen of Wat26758 and the Nε from His312 in structure 23. The initial distance between the reaction coordinates is marked by the dotted black line (3.74 Å). Hydrogen bonds that have O as donor atom are marked in red dotted lines.



Graph 17. Energy scan along the distance from the hydrogen in Wat26758 and Nε from His312.