COMPUTATIONAL DESIGN OF A PROTEIN-BASED COENZYME A BIOSENSOR

by
DİLŞAH NUR ELMACI

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
July 2022

ABSTRACT

COMPUTATIONAL DESIGN OF A PROTEIN-BASED COENZYME A BIOSENSOR

DİLŞAH NUR ELMACI

Materials Science and Nanoengineering, M.Sc. Thesis, July 2022

Thesis Supervisor: Prof. Canan Atılgan

Keywords: genetically encoded fluorescent biosensors, protein design, molecular dynamics simulations

Cell environment comprises many small molecules involved in metabolic processes along with proteins to carry out its routine work. Since changes in the intracellular concentrations of these molecules are indicative of cellular abnormalities, it is crucial to monitor and measure these molecules *in situ*. With this regard, genetically encoded fluorescent biosensors (GEFBs) have come into prominence as they enable real-time measurement of dynamic events in cell by using the intrinsic fluorescence property of proteins. The GEFBs, basically consisting of an analyte-sensing domain and fluorescent protein (FP), are based on the principle that the conformational change in the protein upon binding of the analyte of interest triggers the chromophore microenvironment of the FP, thus giving rise to a detectable change in the fluorescence yield. Developing GEFBs requires a long trial-and-error process. However, in theory, it is possible to rationalize these design steps using computational methods and to make effective interventions to the design at the molecular level. With this motivation, in the scope of the hypothesis that the GEFBs design problem can only be optimized with a holistic understanding of the structure-dynamics of proteins, we developed a computational workflow that can provide an initial design idea for GEFB construction. To this end, we selected coenzyme a (CoA) molecule as a target analyte, and developed possible design models for single circularly permuted FP-based CoA GEFBs. Our pipeline not only provides design models for CoA biosensor construction, but also paves the way for the computational design of FP-based biosensors for any generic analyte.

ÖZET

PROTEİN TEMELLİ BİR KOENZİM A BİYOSENSÖRÜNÜN HESAPLAMALI DİZAYNI

DİLŞAH NUR ELMACI

Malzeme Bilimi ve Nanomühendisliği, Yüksek Lisans Tezi, Temmuz 2022

Tez Danışmanı: Prof. Dr. Canan Atılgan

Anahtar Kelimeler: genetik kodlanmış floresan biyosensörleri, protein dizaynı, moleküler dinamik simülasyonları

Hücrenin rutin işlerinin yürütülmesinde proteinler ile birlikte metabolik yolaklara dahil olan birçok küçük molekül de yer almaktadır. Bu moleküllerin hücre içi konsantrasyonlarındaki değişimler, hücresel anormalliklerin göstergesi olduğundan, onları yerinde izlemek ve ölçmek çok önemlidir. Bu bağlamda, genetik kodlanmış floresan biyosensörler (GKFB), kendiliğinden floresan verebilen proteinleri kullanarak, hücredeki dinamik olayların gerçek zamanlı ölçülmesine olanak sağlarlar. Temel olarak analit-algılama alanından ve floresans proteininden (FP) oluşan GKFB'ler, ilgili analitin bağlanmasıyla protein yapısında meydana gelen değişikliğin, FP'nin kromofor bölgesini tetiklemesi ve floresan veriminde bir değişikliğe sebep olması prensibine dayanmaktadır. Maalesef ki, GKFB'lerin geliştirilmesi uzun bir deneme-yanılma süreci gerektirmektedir. Oysaki, hesaplamalı yöntemler kullanarak bu tasarım adımlarını rasyonelleştirmek ve moleküler düzeyde tasarıma müdahaleler yapabilmek mümkündür. Bu motivasyonla, GKFB'lerin tasarım probleminin ancak protein yapı-dinamiğinin bütünsel bir anlayışla iyileştirilebileceği hipotezi kapsamında, GKFB geliştirilmesi için bir başlangıç tasarım fikri sağlayabilecek, hesaplamalı bir iş akışı geliştirdik. Bu doğrultuda, koenzim a (KoA) molekülünü hedef analit olarak seçtik ve döngüsel permütasyonlu floresan protein temelli KoA GKFB'ler için olası modeller önerdik. Sunduğumuz iş planımız yalnızca KoA biyosensör geliştirilmesi için tasarımlar sunmakla kalmayıp, aynı zamanda herhangi bir hedef analit için de FP temelli biyosensörlerin hesaplamalı tasarımının da önünü açmaktadır.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors Canan Atılgan and Ali Rana Atılgan for their great guidance and support throughout my thesis project and graduate studies. I am truly grateful beyond words for allowing me to be a member of MIDST family and for keeping their doors and zoom sessions open whenever I needed help - both in my personal and academic life. Thanks to them, I have learned that academia can also be inclusive and loving.

I deeply appreciate my thesis committee members, Özge Şensoy and Nur Mustafaoğlu Varol, for their time and worthful guidance. Here, I have special thanks to Dr. Özge for introducing me to the field of biophysics in my undergraduate years.

Further, I thank all the current and former MIDST Lab members, especially Dr. S, Melike, Kurt, Işık, Ebru, Gökşin, and Erhan, for making the office environment so colorful, friendly, and sincere.

I would also like to mention my dearest friends, Sinem, Yasin, and Deniz. They have always cheered me up whenever I got into "mırmır" mood. Moreover, I owe special and precious thanks to Metehan for his endless support, encouragement, and priceless friendship. I faithfully appreciate his patience in answering my weird and endless questions. I feel so lucky to have such an amazing colleague in my life and look forward to working with him further.

Last but not least, I owe my deepest gratitude to my dear family, especially my mother Nuynuy. They have always supported me with their unconditional love and made me feel like I am the luckiest person in the world. They have always been there - by my side - in all my absurd decisions and topsy turvy moods. No word would express my gratitude to them, but I definitely know that today would not be possible without their support.

I also thank the Scientific and Technological Research Council of Turkey (with grant number 121Z329) for funding this project.

I would also like to emphasize that the numerical calculations reported in this thesis were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA Resources).

To my loved ones
Most especially to my mother and grandfather

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

C_α : Alpha carbon

α-helix : Alpha helix

β-barrel : Beta barrel

$\alpha 4$: Helix-4

LIST OF ABBREVIATIONS

1. INTRODUCTION

1.1 Genetically Encoded Fluorescent Biosensors (GEFBs)

Designing novel and practical bioanalytical tools to detect and monitor target analytes and biological events *in situ* is still one of the biggest challenges in various fields, including biomedicine and bioengineering. To propose effective solutions to this challenge, researchers from different disciplines have been working on biosensors. Biosensors are defined as integrated tools that can convert the analyte recognition event into a detectable and measurable signal (Okumoto, Jones & Frommer, 2012; Wang, Nakata & Hamachi, 2009).

A biosensor essentially consists of two parts: a sensing unit and a reporter unit. The working principle of biosensors is based on the selective interaction of target molecules with a sensing site and the qualitative and/or quantitative analysis of the change resulting from this interaction with the help of a reporter unit (Wang et al., 2009). As a sensing platform, biomacromolecules such as nucleic acids and proteins are often preferred. Among these structures, proteins are able to selectively recognize and bind to target molecules with high affinity. Upon binding, proteins often undergo conformational changes, and these changes can be easily converted into a signal. In addition, with recent advances in protein engineering, it is also possible to enhance the selectivity and binding affinity of protein interactions as well as the induced conformational change by performing genetic and chemical manipulations (Wang et al., 2009). Due to all these superior properties, proteins are widely used as recognition elements in the construction of biosensors. Depending on the type of generated signal, several transducer methods can be used as reporter elements, including optical, electrical, magnetic, and thermal converters. Detection with fluorescence has gained importance as it provides high temporal and spatial resolution, enhanced sensitivity, and also deep tissue penetration compared to other methods

(Giepmans, Adams, Ellisman & Tsien, 2006; Rao, Dragulescu-Andrasi & Yao, 2007; Wang et al., 2009). Therefore, over the years, protein-based fluorescence biosensor studies have come into prominence.

The discovery of green fluorescent proteins (GFPs) has ignited subsequent crystallography studies. Accordingly, wild-type GFP has a cylinder-like structure in which a β-barrel surrounds its central α-helix with eleven strands (Figure 1.1). The autocatalytic formation of three sequential amino acid residues (S65-Y66-G76) in this barrel forms the chromophore of GFP. This chromophore is located inside the barrel, and thus protected from being accessible to the surrounding solvent (Tsien, 1998; Wang et al., 2009). An external distortion on GFP (e.g., perturbation of GFP due to binding of an analyte to the sensing domain) may disrupt the hydrogen bonding network around the chromophore, thereby altering its protonation state. This disruption can lead to a change in fluorescence intensity and eventually be quantified as a signal (Nifosí, Amat & Tozzini, 2007; Tsien, 1998). Therefore, GFPs are excellent candidates for use in analyte determination, owing to their intrinsic fluorescence abilities. Furthermore, a wide variety of GFP color variants with improved photostability and different absorption/emission spectra have been achieved by engineering the physical and chemical properties of GFPs by means of mutagenesis (Nifosí et al., 2007; Tsien, 1998). With all these breakthrough advances in FPs, the spotlight on biosensor design has turned towards the development of GEFBs.

The construction of GEFBs relies on the genetic fusion of FPs. To end this, the reporter domain is inserted into the analyte-binding protein, namely the sensing domain, or vice versa. This is achieved by recombinant DNA technology (Wang et al., 2009). Hence, it enables complex biological events to be visualized and measured directly under physiological conditions with high spatial and temporal resolution. This non-invasive GEFB construction does not damage cells in which the biosensor is inserted and does not cause any toxicity on these cells. In addition, it allows real-time detection as the light absorption and emission by the fluorophores of FPs are on the order of nanoseconds (Ovechkina, Zakian, Medvedev & Valetdinova, 2021; Wang et al., 2009). As a result of all these advantages, GEFBs are found to be extremely effective tools for real-time monitoring of important biological molecules and phenomena in living cells.

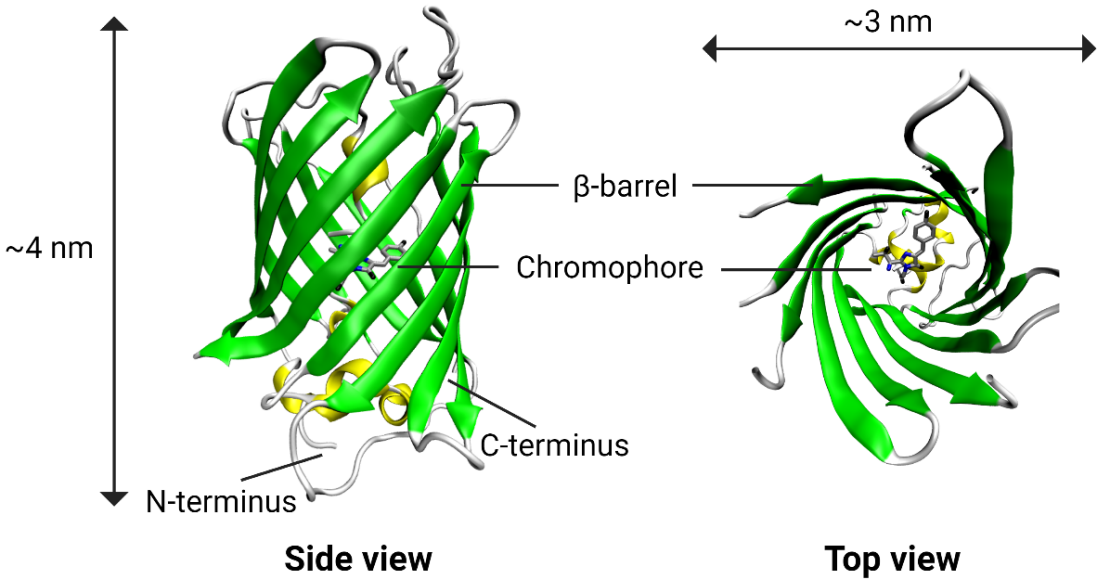

Figure 1.1 Crystal structure of GFP (PDB ID: 1EMA). The secondary structure is shown in cartoon representation with β-strands colored in green, α-helices colored in yellow, and loops in white. The chromophore of GFP is shown in the licorice representation with carbon atoms colored in dark gray, nitrogen atoms colored in blue, and oxygen atoms colored in black.

1.2 Strategies for GEFB Design

Three different strategies are mainly used for designing GEFBs, namely single fluorescent protein (FP)-based biosensors, circularly permuted FP (cpFP)-based biosensors, and Förster resonance energy transfer (FRET)-based biosensors.

1.2.1 Single Fluorescent Protein (FP)-based Biosensors

As its name indicates, single FP-based biosensors utilize a fluorescent protein as both sensing and reporter domains. Such biosensors rely on detecting alterations in fluorescence intensity arising from chemical and/or conformational change in the chromophore environment upon binding of the analyte of interest to the FP (Tamura & Hamachi, 2014) (Figure 1.2). Thus, these biosensors allow to determine the localization of certain analytes in living cells and to assess the change in concentrations.

The foremost advantage of this type of biosensors is that they can exhibit large fluorescent intensity changes in a specific single wavelength. Besides, these biosensors are good candidates for multiparameter imaging as they require only one spectrum window for absorption and emission (Nasu, Shen, Kramer & Campbell, 2021; Shen, Lai & Campbell, 2015).

To selectively increase the sensitivity of single FP-based biosensors, effective mutations can be introduced around the chromophore or in regions interacting with analytes. By the courtesy of mutations, many biosensors have been developed that are able to respond to environmental changes such as oxidation/reduction and pH. For example, Remington *et al.* have mutated specific surface residues of the environmentally insensitive GFP to cysteine residue, so that formation of disulfide bridges in an oxidizing environment could be detected as a signal. This study underlies the design of GEFBs responding to redox reactions (Hanson, Aggeler, Oglesbee, Cannon, Capaldi, Tsien & Remington, 2004). As another known example, Rothman and his colleagues have devised a pH-sensitive single FP biosensor such that by substituting specific amino acids in GFP, they have facilitated the alteration of the protonation state of the chromophore at different pH ranges (Miesenböck, De Angelis & Rothman, 1998). In light of these aforementioned findings, the construction of single FP-based biosensors is promising due to the intrinsic sensitivity of FPs to environmental changes.

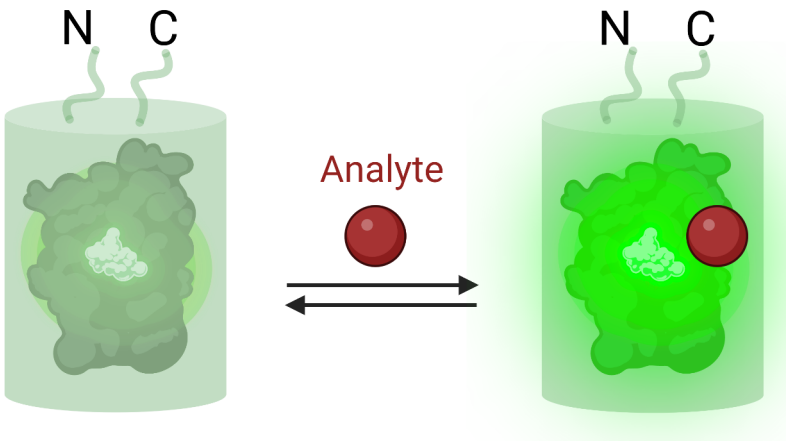

Figure 1.2 Single FP-based biosensor based on its intrinsic sensitivity. (Created with BioRender.com)

1.2.2 Circularly Permuted FP (cpFP)-based Biosensors

Single FP variants reconstructed by circular permutation are another common strategy used in GEFB sensor designs. The N- and C-terminals of FPs are usually spatially distant from the chromophore. Therefore, the mobility of the residues at the terminals does not give rise to any alteration in chromophore protonation or the intrinsic fluorescence of FP. To sensitize the local environment around the chromophore for terminal mobility, the protein sequence is cut at a certain position, and the original N- and C- termini are connected together via a flexible linker. In this way, a cpFP with different amino acid sequence is constructed. The constructed cpFP has new N- and C- termini, which are spatially close to the chromophore (Tantama, Hung & Yellen, 2012; Wang et al., 2009; Zhao, Zhang, Zou & Yang, 2018). Due to the reorganization of FP structure, cpFPs cannot show the fluorescent activity of wild-type FP. For this reason, the N- and C- terminal residues of cpFP are attached to a functional domain(s). Upon the binding of the target substrate, the change in the functional domain causes movement at the cpFP terminals (Figure 1.3). This activity often enhances solvent accessibility to the chromophore so that the binding can be detected as a response in the fluorescence emission intensity and/or spectrum (Germond, Fujita, Ichimura & Watanabe, 2016; Sanford & Palmer, 2017).

An early example of a circularly permuted fluorescent sensor design strategy is the fusion of calmodulin–a calcium-binding protein–into several circularly permuted GFP variants (Nagai, Sawano, Park & Miyawaki, 2001). As a consequence of calcium-binding to the calmodulin-cpFPs complex, the calmodulin domain undergoes conformational changes near the chromophores of cpFPs. This facilitates deprotonation of the chromophores, and subsequently ends up with significantly larger change in fluorescence emission. By using similar strategy, many cpFP-based biosensors have been developed for the *in vitro* and *in vivo* detection of a variety of molecules such as cyclic guanosine 3',5'-monophosphate (Nausch, Ledoux, Bonev, Nelson & Dostmann, 2008), hydrogen peroxide (Belousov, Fradkov, Lukyanov, Staroverov, Shakhbazov, Terskikh & Lukyanov, 2006), and zinc ion (Mizuno, Murao, Tanabe, Oda & Tanaka, 2007).

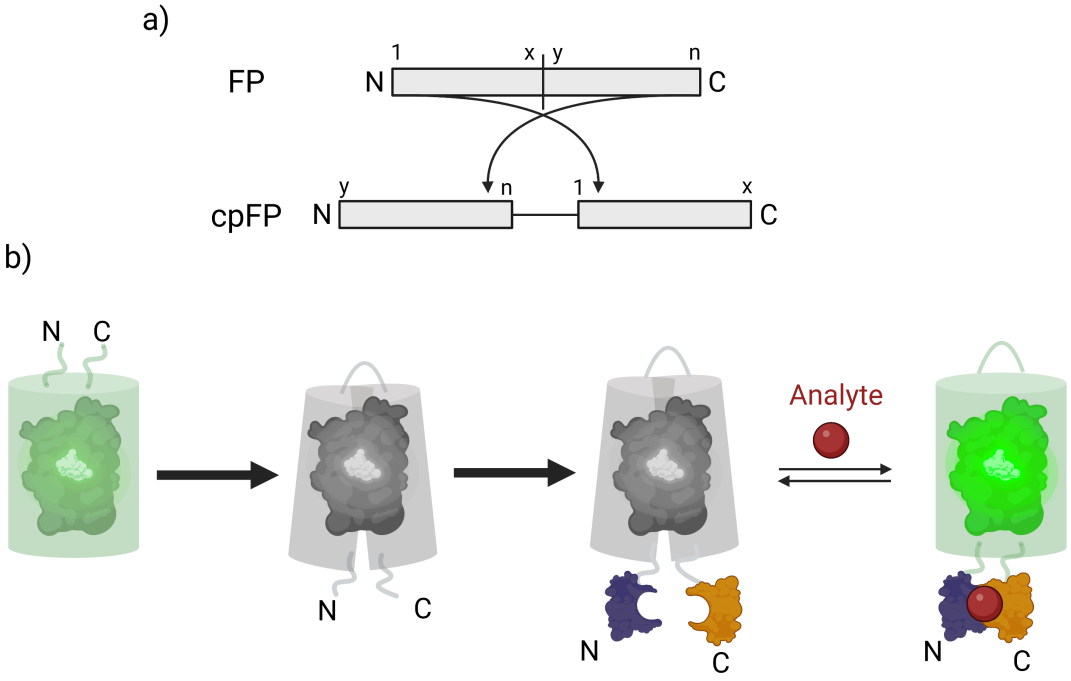

Figure 1.3 Construction of cpFP-based biosensor. a) A scheme of generating a cpFP by switching the N- and C- termini of the original FP. b) Activation mechanism of cpFP biosensor upon analyte binding. (Created with BioRender.com)

1.2.3 Förster Resonance Energy Transfer (FRET)-based Biosensors

FRET-based biosensor strategy takes advantage of the fluorescence resonance energy transfer phenomenon, which employs a distance-dependent non-radiative transfer from the chromophore of one FP (donor) excited at a certain wavelength to the chromophore of another (acceptor) (Okumoto et al., 2012; Sanford & Palmer, 2017). In such biosensors, a sensing domain is usually connected between the donor FP and the acceptor FP. The efficiency of energy transfer between the chromophores of these FPs is highly associated with the conformational change in the sensing domain. Naturally, a fluorophore excited by light absorption tends to return to its ground state by utilizing either fluorescence emission or a non-radiative decay mechanism. If the excited donor is in close proximity (<10 nm) to the acceptor and the emission spectrum of the donor coincides with the absorbance spectrum of the acceptor, the donor can then return to its ground state by transferring its energy to the acceptor via dipole-dipole interaction (Figure 1.4). The acceptor excited via FRET can also produce its own fluorescence emission (Campbell, 2009; Tamura & Hamachi, 2014; Tantama et al., 2012). Therefore, in FRET-based biosensors, donor and acceptor

emission intensities can be compared by taking advantage of the distance and/or conformational change created by the interaction of the analyte with the sensing domain, thus enabling the determination of the analyte and its concentration. This approach is still widely studied today for the determination and real-time imaging of various analytes such as zinc ion (Xu, Zhu, Chen, Bai, Han, Yao, Jiao, Yuan, He & Guo, 2020), formaldehyde (Ding, Yuan, Peng, Zhou & Lin, 2020), and glutathione (Ahmad, Anjum, Asif & Ahmad, 2020) *in vitro* and *in vivo*.

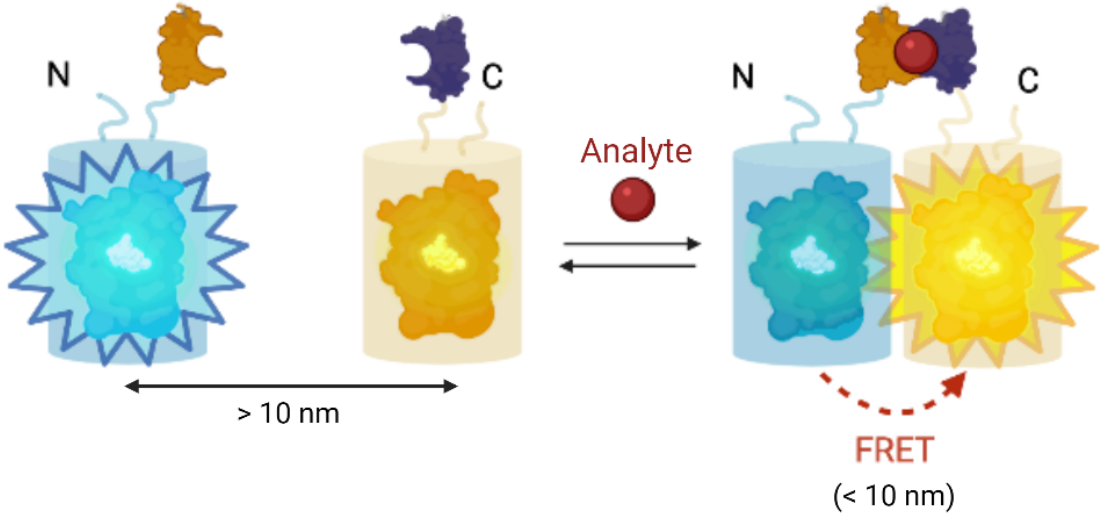

Figure 1.4 Working mechanism of FRET-based biosensor. (Created with BioRender.com)

1.3 Approaching GEFB Design Problem at the Molecular Level

Particularly with increasing demands for monitoring and measuring the change in substrate concentration in living cells, researchers have turned their focus on developing various strategies for GEFB design and optimization. Although a variety of biosensors have been developed for many different purposes leveraging the above-mentioned three approaches, a generalized methodology for designing a unique protein-based fluorescent biosensor for a specific target has not yet been presented in the literature. Although such biosensor applications seem suitable to be standardized in theory, their development is still mainly carried out by trial-and-error approaches. Using these approaches, a design can be only proposed by finding a proper protein that meets the following requirements: (1) a structurally resolved candidate protein with an ability of binding to the analyte, (2) the presence of proper

ligand-driven conformational changes in the candidate protein, (3) production of a measurable fluorescent signal driven by ligand binding to candidate-FP complex, and (4) the capability of calibrating the dynamic range so that the biosensor can be tuned for more accurate measurements.

However, it is theoretically possible to make sense of and manipulate each of these requirements using molecular-level information along with various computational tools, thus offering smarter designs for biosensor construction. For instance, about the first requirement, a protein that binds the desired molecule may not be found in nature. However, it is now known that unrelated proteins with similar binding microenvironments tend to recognize and bind chemically similar ligands (Barelier, Sterling, O'Meara & Shoichet, 2015; Govindaraj & Brylinski, 2018). Therefore, in such a case, proteins that bind similar molecules can be screened and appropriate amino acid substitutions can be made at the binding sites of these proteins. Thus, it is possible to manually create a binding site capable of recognizing the desired substrate. For the second requirement, it is known that, with the advances in the computational field, molecular structures can be engineered so as to trigger their conformational changes with the binding of a ligand of interest (Mizoue & Chazin, 2002). Particularly, some amino acid modifications can be also made in allosteric regions of the proteins to exaggerate the alterations in the protein structure response to the binding. The third condition associated with generating the fluorescent signal may be achieved by linking the fluorescent probe to a mobile site of the protein (Nasu et al., 2021). This eventually allows to induce changes in the chromophore of FP. For this step, experimentalists create a library of fusion proteins containing different FP variants and a sensing module by linking them at the random insertion site with a vast number of linker variants. Consequently, they scan the fluorescent properties of each pair in the presence of the ligand of interest (Patriarchi, Cho, Merten, Howe, Marley, Xiong, Folk, Broussard, Liang, Jang & others, 2018). Thereafter, the fluorescent profiles of each couple are compared, and the one with the best fluorescent results is used to be optimized for further design. As might be expected, this step is quite a time-consuming and blind process. However, at this step, instead of scanning large libraries, computational tools can be utilized to predict the most promising models. Then, the potential of these models to generate fluorescent signals can be revealed by elucidating the change in the conformation of chromophore with molecular dynamics (MD) simulations. The measurements for the fourth requirement are highly related to the binding constant of the ligand. By carrying out thermodynamic calculations at the molecular level, the changes in the binding constant caused by mutations in the binding site can be measured as sensitive energy differences (Mondal, Florian & Warshel, 2019). All of this signifies that the use of

computational methods might help to illuminate these trial-and-error steps, thus more rational designs can be suggested as a starting point for experimental studies.

Regarding the computational studies being performed in the field of protein-based biosensors, several procedures have been proposed by leading research groups working in protein design (Dou, Doyle, Jr Greisen, Schena, Park, Johnsson, Stoddard & Baker, 2017). The procedures are still not as sufficient as demanded. Nevertheless, they are more commonly employed, and particularly, focused on the sensing module of the biosensor. The most up-to-date design steps for sensing domain, where computational methods are listed as follows (1) selecting a protein, which binds to a target of the interest, and searching for other protein structures with similar binding sites, (2) docking the analyte of interest to these scaffold proteins one by one, (3) scanning the side-chain rotamers of the amino acids in the binding site, (4) scoring all candidates with a proper energy function and determining the most suitable candidates accordingly, (5) running simulations to achieve structural modifications in this pre-filtered group of proteins, (6) introducing point mutations to the binding site to improve ligand-protein interactions, (7) rescoring the structures to choose the candidate sensor model(s), and (8) synthesizing candidate model(s) followed by solving their crystal structures, and checking whether the desired design is achieved or not. Whilst these listed steps suggest that a pipeline for biosensor construction has already been established, the designers themselves highlight two major shortcomings in this workflow (Dou et al., 2017). First, the scoring methods used in the fourth and seventh steps are not sufficient for the precise calculations for the design. Hence, therein, there is a need to refine the solvation energy terms, especially for the hydroxyl and carbonyl groups of the ligands. Another drawback is the inadequacy of the sampling methods used in scanning the different conformations adopted by the side groups in the binding site. These two major limitations are identified in the workflow, which by all means appear to be related to the binding site. From this perspective, it is known that although proteins that bind to the same/similar molecules are not structurally similar, they generally have some conserved/analogous residues in their binding sites, which are very crucial for the protein functions and ligand-protein interactions. The mapping and comparison of the binding sites of these proteins would shed light on fine-tuned designs for ligand-binding proteins. In addition, it is worth bearing in mind that amino acids in allosteric sites make a non-negligible contribution to protein-ligand interactions along with protein function (Bhat, Schaeffer, Kinch, Medvedev & Grishin, 2020). Considering these facts, developing a holistic approach involving manipulation of allosteric sites as well as binding site modifications would become a promising strategy for the successful applications of protein-based biosensors.

In addition to having an insight into the construction of the sensing domain, uncovering the underlying mechanism of fluorescent proteins is also crucial for providing better biosensor designs with improved signal resolution. In current studies, to optimize the FP part, very large libraries are first prepared with different combinations of the parameters, and then, these libraries are experimentally screened to find the most optimized version (Fritz, Letzelter, Reimann, Martin, Fusco, Ritsma, Ponsioen, Fluri, Schulte-Merker, van Rheenen & others, 2013; Patriarchi et al., 2018). Fortunately, Campbell and his group have very recently published a guiding work on elucidating the molecular mechanisms of fluorescent-based biosensors. In this study, they have provided remarkable insights into reducing the experimental workload (Nasu et al., 2021). Particularly, two important terms for understanding the molecular basis of FPs were specified: gate post residues and bulge region residues. The bulge region is where the hydrogen bonding of FP is disrupted. This region consists of two consecutive residues and the side chains of these residues are directed towards the surrounding solvent, thus creating a bulge effect. The two residues flanking this region are the gate post residues. In other words, bulge region residues lie between the two gate post residues. Since these residues are able to well tolerate insertions from these sites, they stand out for sensing domain fusion. There are two main reasons why gate post and bulge region residues are privileged sites. First, these sites are the closest areas to the chromophore, thereby allowing the chromophore to isolate itself from the surrounding environment. For this reason, the investigation of the gate post residues' positions, which are in communication with the chromophore, gives information about the protonation state of the chromophore. The second reason is that they allow new termini to be introduced to construct cpFP. Therefore, considering all these, these regions not only illuminate the microenvironment of chromophores but also elaborate on constructing cpFP. Furthermore, the same study also gives information on possible insertion locations within any sensing domain that is likely to undergo a significant conformational change upon ligand binding, and might also tolerate FP insertion. This information suggests choosing a permissive residue with conformational mobility on the sensing domain can be used as an insertion site for FP fusion. To do so, gate post residues of FP can be attached to the insertion site with flexible linkers. Regarding the linker length, it is recommended to keep the linker as short as possible so as to maximize coupling between the fluorescent domain and the sensing domain and also not to disrupt the overall structure folding (Nasu et al., 2021). While this study provides very valuable details about the molecular mechanisms of FPs, it also reveals that the FP optimization process can be rationalized by using molecular information rather than entirely proceeding through trial and error. In conclusion, while all developments in the sensing domain and fluorescent domains provide valuable information on biosensor

design and optimization, understanding GEFB studies at the molecular level with computational methods is indispensable for providing optimal designs.

1.4 Scope of the Thesis

GEFBs are nanobiotechnological tools used to detect conformational changes occurring in a protein upon ligand binding and convert this molecular recognition event into fluorescence signal measurements. Developing functional GEFBs, which can measure selected metabolites in their natural environment, requires trial-and-error processes, and the design, unfortunately, takes years and years to optimize. However, in theory, it is possible to shorten these processes by making use of the structure-function relationships of proteins. In practice, there are various technical problems to be overcome, albeit only those related to designing the ligand-binding site of the biosensor have been discussed by computational biologists.

From this perspective, within the framework of the hypothesis that the GEFBs design problem can only be optimized with rationalizing the structure-function relationship of the proteins, the main purpose of this thesis is set out to develop a modular pipeline that can provide an initial biosensor design idea by taking coenzyme A (CoA) molecule as a case study. In this context, CoA as a chosen target substrate is a cofactor found in all living organisms and plays prominent roles in many cellular activities. CoA is particularly involved in the synthesis and oxidation of fatty acids and also the oxidation of pyruvate in the citric acid cycle for energy production. Furthermore, CoA itself and its derivatives indirectly contribute to the regulation of gene expression, cell cycle, and transcription factors (Daugherty, Polanuyer, Farrell, Scholle, Lykidis, de Crécy-Lagard & Osterman, 2002). For these reasons, undoubtedly, CoA is a critical molecule for vital activities of the cell, and henceforth the importance of controlling the concentration of this molecule in cells comes to the fore. At this point, although FP-based biosensors have been designed and developed for the monitoring and quantification of many metabolites, there is not yet a sensor enabling real-time, high-resolution, spatial measurement of CoA molecules in cell infrastructure. Therefore, presenting a design study for CoA biosensor will come in useful to detect the amount of CoA in cells in real-time, and thus will shed light on the understanding of the function of this metabolite.

Considering the roles of CoA in cellular metabolism, within the scope of this

project, it is aimed to develop a workflow with a holistic approach to propose a design for a single FP-based CoA biosensor by effectively intervening in the design at the molecular level (Figure 1.5). To this end, all proteins bound to CoA and/or its structurally similar derivatives were retrieved from the Protein Data Bank (PDB) (https://www.rcsb.org/) (Berman, Westbrook, Feng, Gilliland, Bhat, Weissig, Shindyalov & Bourne, 2000; Burley, Bhikadiya, Bi, Bittrich, Chen, Crichlow, Christie, Dalenberg, Di Costanzo, Duarte & others, 2021), and a target protein set was created with the goal of finding a suitable recognition module for the sensor design. Then, this set was filtered according to certain criteria (e.g. conformational changes in protein-induced by CoA binding, protein origin or protein chain size), and then, a representative structure was selected that satisfies the desired criteria for the CoA-sensing domain selection. Other proteins sharing similar binding sites with the selected representative protein were fished out. Variations within the binding sites of these proteins were noted as they provide valuable information to enhance ligand binding affinity. Although the active site design of the selected CoA-binding protein has not been studied in detail in this thesis, it was explained how this information can be obtained and utilized as a reference point for fine-tuning biosensor design. In addition, the conformational change of the protein driven by CoA binding was unveiled by means of microsecond-long atomistic simulations. According to the analysis results obtained, two possible insertion sites were determined on the sensing domain for circularly permuted GFP (cpGFP) insertion. After that, the sequence of the cpGFP domain was inserted into the sensing domain sequence from two distinct sites, with or without a linker. Subsequently, the 3D structures of the created sequences were predicted through the AlphaFold2 (AF2) program (Jumper, Evans, Pritzel, Green, Figurnov, Ronneberger, Tunyasuvunakool, Bates, Zídek, Potapenko, Bridgland, Meyer, Kohl, Ballard, Cowie, Romera-Paredes, Nikolov, Jain, Adler, Back, Petersen, Reiman, Clancy, Zielinski, Steinegger, Pacholska, Berghammer, Bodenstein, Silver, Vinyals, Senior, Kavukcuoglu, Kohli & Hassabis, 2021), and these structures were scrutinized in terms of the conservation of the fold of the domains after fusion, the conservation of the positions of the active site residues in the sensing domain, and the orientations of the gate post residues in the cpGFP domain. To sum up, the findings presented in this thesis can provide a useful starting point for studies aiming to design a single FP-based CoA biosensor. Last but not least, the developed workflow can be applied and improved for any genetic biosensor design.

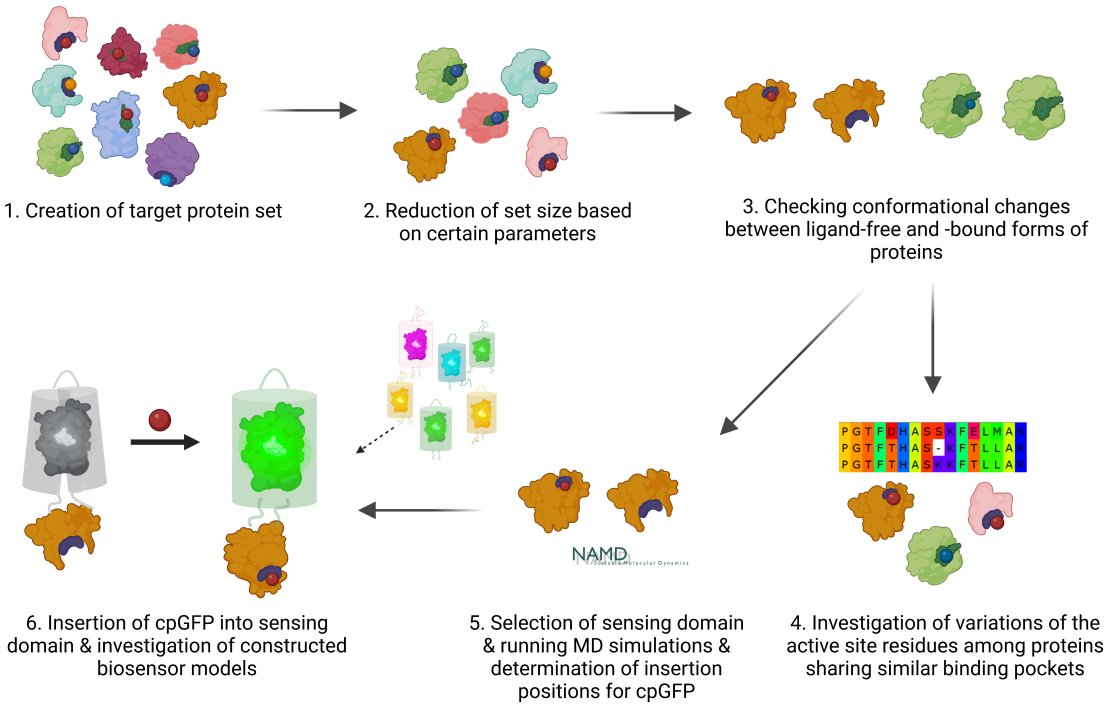

Figure 1.5 Workflow developed for computational design of cpGFP-based CoA biosensor construction. (Created with BioRender.com)

2. SELECTING A REPRESENTATIVE SENSING DOMAIN FOR CPGFP-BASED COA BIOSENSOR

One of the most critical steps in protein-based biosensor design is the selection of protein that specifically recognizes and binds the analyte to be measured. The 3D structures of proteins possessing the selected analyte as a ligand and other related information about the structures are deposited in the PDB. In case there is no protein crystal structure solved for the ligand of interest in the database, other available proteins that bind to similar molecules can be examined, and the active sites of these proteins can be remodeled for the recognition of the interested substrate using cutting-edge computational tools. When picking out a protein from among many possible structures, there are several important criteria that the selected protein must meet to be used as a sensing platform. First and foremost, the protein must be able to undergo a conformational change when bound to the analyte. Otherwise, a fluorescent reporter domain cannot detect this binding event, and thus the signal cannot be generated. Secondly, considering the biocompatibility of the selected protein with the cell conditions, proteins of bacterial origin are generally preferred by experimental groups. Lastly, it might also be beneficial to have an idea about the active sites of other proteins, which bind the same or similar molecules, so that the ligand-binding site of the sensing domain can be more finely modified. In particular, in cases where ligand-protein interaction needs to be improved, it may be a reasonable approach to compare the binding site residues of proteins with similar active sites and take them as a reference while making mutations instead of doing it blindly. To the best of our knowledge, there is no generalized computational methodology that considers these criteria regarding the ligand-sensing part in current biosensor applications. Therefore, this chapter presents a workflow regarding selecting a representative binding domain by taking the CoA molecule as a case study. Afterward, a protein fulfilling the above-mentioned requirements was selected as a representative and proposed to be used as a sensing part in the CoA biosensor model.

2.1 Methods

2.1.1 Extraction of Protein Set Possessing CoA and Its Structurally Similar Derivatives as Ligands from PDB

As a first step, PDB IDs of all bacterial proteins to which CoA and its structurally similar derivatives are bound as ligands, as well as HET codes of the relevant ligands, were extracted from PDB. The Advanced Search Query Builder option of RCSB PDB was used to identify molecules structurally similar to CoA. To this end, search settings were set as follows: The isomeric SMILES notation of the CoA molecule was entered as a query. Accordingly, query and descriptor types were selected as "descriptor" and "SMILES", respectively. As ligand match type, the "similar ligands (substructures including stereoisomers)" option was chosen. This option allows a detailed substructure search for the given query by taking atom type, bond order, formal charge, and aromaticity of molecules into account as matching criteria. As a return option, the "molecular definitions" was selected. This way, the HET codes of CoA derivatives molecules matching the criteria mentioned above were obtained. Afterward, all PDB entries where the ligand CoA and its derivatives are present as standalone ligands in bacteria species were downloaded from PDB. Eventually, a set of bacterial proteins was created from these PDB entries pertaining to these ligands.

2.1.2 Reducing Protein Set by Keeping Only Unique Protein Sequences

Within the created set, there are many crystal structural entries, that bind the same ligand and also have the same sequence, but have been deposited with different PDB IDs due to differences in structures or the methods used to obtain these structures (e.g., different bound ligands/small molecules, mutations, or differences in the experimental conditions such as pH). However, since these proteins are derived from the same organism and have the same sequence, they are identical in origin and function. Therefore, their identifier codes used in the UniProt database (Consortium, 2021), namely their Uniprot IDs, are the same. To eliminate those redundancies in the protein set, only one representative structure of each unique sequence bound to

the same ligand was kept in the set. As such, proteins with the same Uniprot ID were filtered out, and a representative PDB ID for each, preferably wild-type form, was chosen. Hence, the size of the protein set was reduced.

2.1.3 Selection of a Representative Structure for Use as a Sensing Domain

After shrinking the protein set, the suitability of the candidate ligand-bound proteins –holo–for use as a selecting platform in the computational design of the CoA biosensor was evaluated. To this end, proteins that meet the following requirements were prioritized for selection:

1. Candidate proteins should have the potential to undergo structural changes by ligand binding.

For the first criterion, the ligand unbound forms of the proteins, namely apo, were quickly scanned to find out whether the proteins undergo conformational changes upon ligand binding or not. For this search, the Sequence Annotated By Structure (SAS) database of Protein Database summaries (PDBsum) was utilized (Laskowski, Jabłońska, Pravda, Vařeková & Thornton, 2018). SAS scans the fasta sequence of the entered protein chain against all unique protein sequences found in the PDB in a few seconds. As a result, it sorts the protein entries similar in sequence to the given protein in terms of percentage identity and overlapped amino acid numbers. Furthermore, it lists the ligands bound to the corresponding chains of the resulting proteins. This ligand list was convenient in finding the apo form of the protein searched. Later, apo and holo forms of the proteins were superimposed using PyMol (Schrödinger, LLC, 2015). Thus, those with deviations in their apo and holo structures continued to be examined as candidate molecules.

2. Considering computational efficiency, proteins consisting of less than 400 amino acid residues are preferred.

As the second criterion, the computation time to be spent for MD simulations was taken into account. For this reason, small size proteins were favored as a sensing domain.

3. Proteins that are not bound to molecules other than the relevant CoA derivative in their active sites are favored.

In cases where more than one molecule is bound at the active site, the protein's recognition and binding of the ligand of interest may also be associated with other

molecules present in the active site. For this reason, as the third criterion while selecting candidate CoA-binding sites, only proteins that bind the interested CoA derivative as a ligand were interested in keeping the ligand-protein interaction simple.

4. Last but not least, in order to facilitate binding-site design in the following steps, it was checked whether the candidate proteins share similar binding sites with other proteins bound to CoA and its derivatives.

Observing patterns between the binding sites of different proteins that bind similar molecules facilitates effective interventions at these sites and thus sheds light on the fine-tuning biosensor design. Therefore, special attention was paid to selecting a representative structure that shares a similar binding site/pattern with other proteins bound to CoA derivatives, as well as to ensure that differences in this pattern provide sufficiently helpful information for binding site modifications. For this, the Search K mode of the Pocket Similarity Search using Multiple-sketches (PoSSum) server was utilized (Ito, Tabei, Shimizu, Tomii & Tsuda, 2012; Ito, Tabei, Shimizu, Tsuda & Tomii, 2012; Tabei, Uno, Sugiyama & Tsuda, 2010). PoSSum is a web-based database that allows rapid and efficient searching of similar binding sites between protein structures with similar global folds as well as entirely different ones. This database uses a fast fingerprint-based similarity search algorithm and evaluates all ligand binding sites found in the PDB as feature vectors, taking into account these sites' physicochemical and geometric properties. To define ligand-binding sites of proteins, all residues with at least one heavy atom located within 5 Å from one of the heavy atoms of the ligand are chosen. It then applies an ultrafast neighbor search algorithm called SketchSort and calculates the similarity of two binding sites as a value of cosine similarity between their corresponding vectors. If the value is greater than the given cosine cut-off value, these binding site residues are subsequently aligned with each other by employing the TM-align algorithm, and thus the similarity between the sites is also calculated in terms of root-mean-square deviation (RMSD). Therefore, the ligand-binding site of a query protein is compared to those of all other proteins in the database in a pairwise manner, and the similarity between the binding sites is scored in terms of cosine and RMSD values.

To query, it is required to enter the relevant PDB ID and its ligand's HET code as input (Figure 2.1). The necessary settings to group the proteins that bind CoA and its derivatives based on their binding site similarity were set as follows: The target dataset was selected as a known ligand-binding site so that the query was screened against proteins with known and identified binding sites. Sequence redundancy was not removed from the search results in order to keep protein information that have

identical sequences but bind to different molecules. Besides, the cosine similarity cut-off and aligned cut-off values were set to 0.78 and 6, respectively.

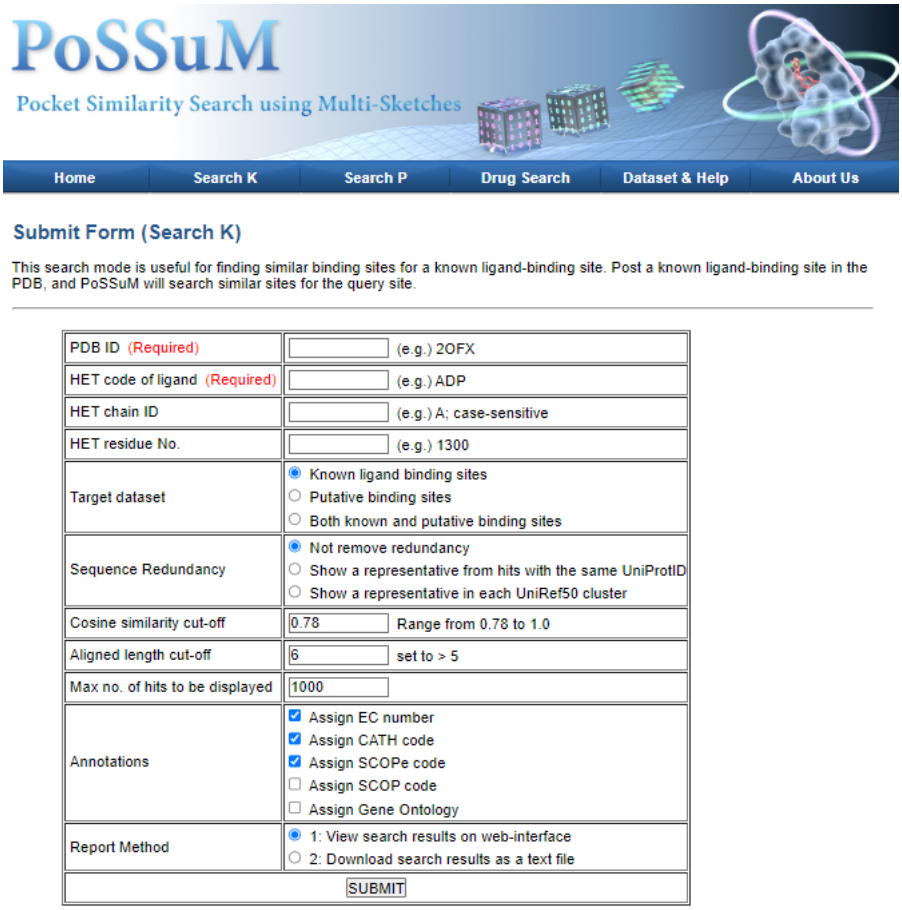

Figure 2.1 Search mode of PoSSum. PDB ID and HET code of ligand must be entered.

With these settings, PDB IDs of the candidate proteins in the set were searched in the database. The results were downloaded in txt file format and then converted into xlxs file. To reduce the redundancy in the downloaded file, the downloaded text file is further processed as follows before the following protein ID in the list is searched as a query:

• The result file may also include the protein IDs to which non-interest molecules are bound as well. To remove these proteins from the downloaded file, the file was shrunk to keep only the results for the proteins to which CoA and its derivatives are bound as ligands.

• There are many homomeric proteins in the PDB, consisting of multiple identical chains bound to the same CoA derivative. In such cases, the PoSSum server separately compares the CoA-binding sites of each chain in these proteins with the given query. Besides, the server aligns two proteins by varying the lengths of their binding

residues in order to find the best fit of the compared binding sites. Therefore, the result file from PoSSum contains different aligned residue lengths for the same query and target pair and their corresponding RMSD values. To eliminate such redundancy in the file, it was processed in a way that only a single representative binding structure was taken for a unique protein ID. To this end, the chain with the highest aligned length and the small RMSD value was chosen as a representative structure for each PDB ID.

• Only one representative of each unique protein bound to the same ligand was kept in the result file, eliminating protein redundancy. For this purpose, among the proteins with the same Uniprot ID and bound CoA derivative, the one with the highest aligned length and the small RMSD value in the result file was selected.

In addition to fishing out proteins with a similar binding site as the given query, the interactions of these proteins with their corresponding ligands were visualized via the LigPlot+ program (Laskowski & Swindells, 2011). As this program can generate two-dimensional ligand-protein interaction diagrams, it is easy to determine which residue in the binding region interacts with which region of the ligand.

2.2 Results and Discussion

To our knowledge, no example of a protein in the literature exhibits a large conformational change in response to CoA binding. Therefore, a target protein set was first constructed to find a candidate sensing protein capable of undergoing such a conformational change. This set consists of different proteins to which CoA or its structurally similar derivatives are bound. There are two main reasons for including proteins bound to CoA analogues for the set. The first reason is that if a protein shows changes in its structure upon binding a molecule similar to CoA, it is possible to remodel its active site for CoA recognition as well with the help of computational tools (Vaissier Welborn & Head-Gordon, 2018). The second reason is related to improving ligand affinity for fine-tuning the biosensor design (Aldeghi, Gapsys & de Groot, 2018; Zhang, Wang, Su, Sun, Zhu, Qi & Wang, 2020). Comparing binding sites of different proteins with similar active sites may provide valuable hints for enhancing protein-ligand interactions. For instance, in the light of such information, sensor selectivity and sensitivity can be tuned by making appropriate amino acid substitutions in the active site.

With this motivation, target protein set was constructed as summarized in Figure 2.2. First, molecules structurally similar to CoA were determined. For this, the exhaustive substructure search option of RCSB PDB was utilized, and consequently, 135 molecules were found. After that, PDB entries of all bacterial proteins that bind these molecules were fished out. Hence the total number of PDB structures that bind these 136 molecules, including CoA, was found as 782. Within these PDB entries are multiple proteins that have the same amino acid sequence as well as bind the same ligand. To eliminate this redundancy from the list, only one representative structure was chosen for each unique protein sequence that binds the same ligand. After these eliminations, the number of unique proteins to be searched for sensing domain selection was reduced to 534 PDB IDs (Table A.1). Then, by prioritizing CoA-binding proteins in this list, a candidate protein meeting the four requirements was searched as explained in detail in the Methods section.

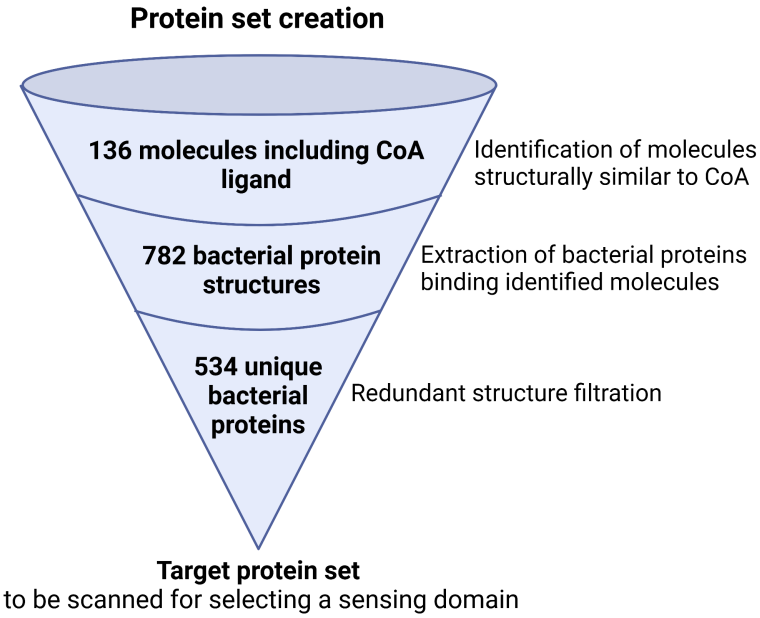

Figure 2.2 Creation of target protein set for sensing domain selection.

As a result of these searches, a CoA-binding protein with PDB accession number 3OTW was chosen as a promising structure as it is found to fulfill all of the above-mentioned criteria. The selected crystal structure belongs to the phosphopantetheine adenylyltransferase (PPAT) protein derived from *Helicobacter pylori* (*H. pylori*). This bacterial PPAT protomer contains five parallel β-strands and eight α-helixes, with a total of 157 residues (Figure 2.3). As it is a small protein, it provides great convenience and efficiency for computational calculations. Furthermore, the protein binds only to one CoA molecule as a ligand, which simplifies sensor design.

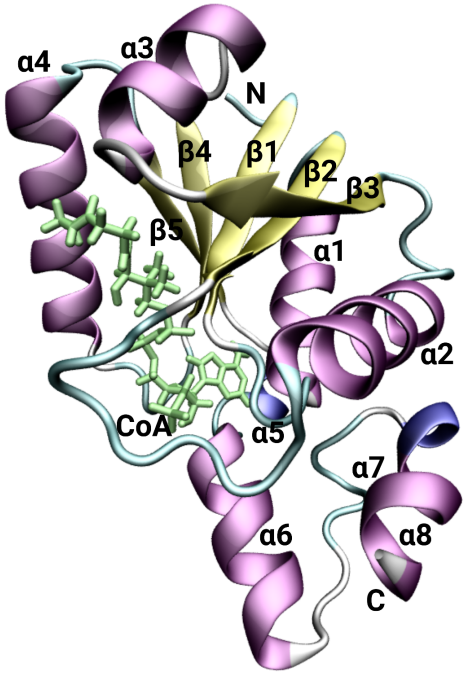

Figure 2.3 Cartoon representation of secondary structures of *H. pylori* PPAT.

In the next step, the criterion of whether the protein induces structural and conformational changes by ligand binding was checked. The CoA-free form of the protein was not found in the PDB databases; however, an I4V/N76Y double mutant form of the protein (PDB ID: 3NV7) was found. When the crystal structure of the mutant protein was compared to that of the CoA-bound protein, it was observed that the structural configuration of the mutant form is unfolded after residue 84, and the secondary structures in this region are also disrupted (Figure 2.4). To elaborate on this large structural deviation, residues 3-84 of both CoA-bound and mutant forms were aligned to each other, and the resulting backbone RMSD was calculated as 0.4 Å. For this fitting, the backbone RMSD value for the rest of the structures (residues 85-157) was calculated as 40.9 Å. Since the double mutation may not presumably drive that much of a drastic change, it is considered that the CoA-free form of the protein might also be in an open conformation and thus undergo folding upon CoA binding. Besides, considering that CoA is a large molecule consisting of 40 atoms, the apo form is highly likely to adapt itself to an open conformational state for CoA binding.

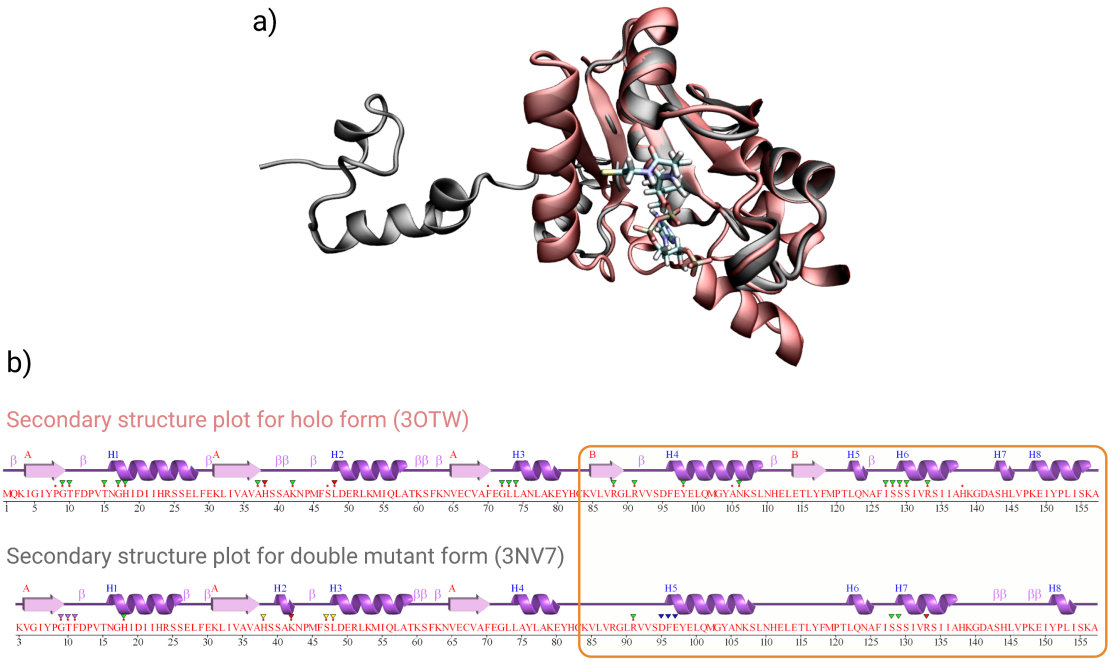

Figure 2.4 Structural and sequential comparison of CoA-unbound *H. pylori PPAT* with the double mutant and COA-bound *H. pylori PPAT*. a) Structural alignment of identified PPAT proteins. b) Comparison of secondary structure plots of PPAT proteins.

As the last criterion, other proteins sharing similar binding sites to the H. pylori PPAT protein were explored. For this, the PPAT protein was searched on the PoSSum server as a query. As detailed in the Method section, the redundancy in the file was removed, and only one representative structure was kept for identical sequences with the same Uniprot ID. After these eliminations, the resulting PDB entries were found as follows (PDB ID_corresponding HET Code): 1B6T_COD, 3PXU_COD, 3X1J_ACO, 4RUK_COA, 5TS2_COD, 5YRR_COA, and 5ZZC_COD (Figure 2.5). It is found that all of these proteins are PPAT proteins originating from different organisms. Their binding site similarities to the given query were found to be around 0.5-1 Å in terms of RMSD. To scrutinize the differences within their active sites, the binding site residues aligned with the query were extracted from the aligned residues column in the result file. The outcomes indicate that, even if many binding site residues are conserved among these proteins, some residues may also vary. Regarding the conserved residues at the binding site of the proteins serving the same biological function, it is known that such residues play crucial roles in the protein binding function as well as its catalytic activity. Bearing this fact in mind, it may be a reasonable approach to remodel residues that can differentiate between organisms rather than those with catalytic importance while modifying the ligand-binding site for a more sensitive recognition design. From this perspective,

the PoSSum results provide alternative options for some residues in the active site of the *H. pylori* PPAT protein.

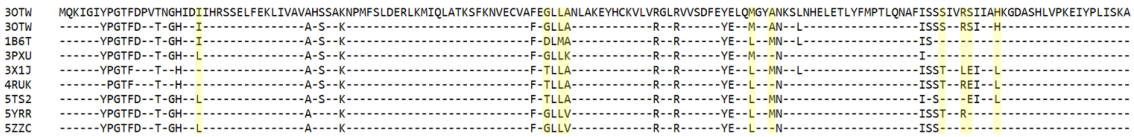

Figure 2.5 Comparison of binding site residues of CoA/analogues-binding proteins. Alternating residues among protein sequences are highlighted in yellow.

Although PoSSum shows alterations in the active site residues among various proteins with similar binding sites, it does not give information about which part of the ligand these residues interact with. Therefore, the protein-ligand interaction scheme of the CoA-bound *H. pylori* PPAT was created via the Ligplot+ program (Figure 2.6). The results reveal that these altered residues correspond to residues L74, S130, and R133, which are involved in hydrogen bonding with CoA, and also residues G72 and A105, which are in hydrophobic contact with CoA. As this thesis primarily aims to establish a generalized methodology to propose a cpGFP-based CoA sensor model, the findings regarding the biosensor's binding affinity and sensitivity have not been studied in detail. Nevertheless, the obtained results offer a potential route to improve ligand recognition of the selected sensing protein.

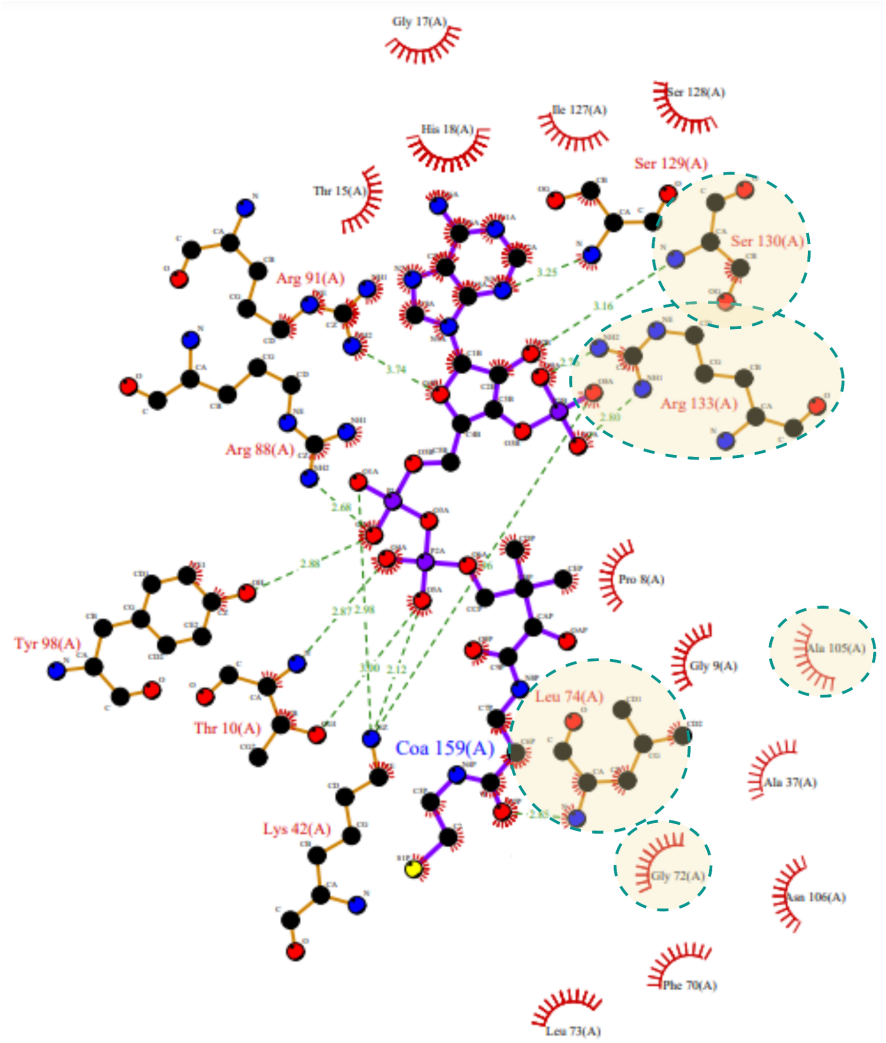

Figure 2.6 Ligplot+ diagram illustrating *H. pylori* PPAT residues that interacts with CoA molecule. Hydrogen bonds are shown by green dashed lines, and nonbonded hydrophobic contacts are represented by red arcs with spokes. Residues circled in light yellow with dashed green lines are alternating binding residues found by PoSSum results for *H. pylori* PPAT.

All in all, a workflow was created to find a candidate binding platform for the CoA biosensor, which can also be used for a generic biosensor design. Using this workflow, *H. pylori* PPAT protein was selected as a representative sensing domain. As a next step, it was decided to scrutinize the dynamics of the CoA-free and -bound forms of the protein by all-atomistic MD simulations to track the conformational changes upon ligand binding and determine possible insertion position(s) for fluorescent domain fusion.

3. STRUCTURAL AND DYNAMICAL INVESTIGATIONS OF COA-BINDING DOMAIN

The main purpose of this chapter is to reveal whether the sensing domain chosen in the previous step exhibits CoA-binding-induced conformational change by using MD simulations. As the crystal structure of the CoA-unbound form of the selected sensing domain is not found in the PDB, the CoA molecule was removed from the bound form, and the resulting structure was then used as the apo form. Afterwards, by running MD simulations, the potential of the prepared apo form to resemble the conformation of the mutant version of the same protein (PDB ID: 3NV7) was investigated, thus the possibility of triggering conformational change as in that mutant version. To this end, apo and holo systems were simulated and analyzed to deeply scrutinize the effect of CoA binding on the protein structure and dynamics. Besides, possible insertion sites on the sensing domain were determined for cpGFP fusion.

3.1 Methods

3.1.1 Preparation of Apo and Holo Systems for MD Simulations

The crystal structure of the COA-sensing domain (PDB ID: 3OTW), namely holo, was downloaded from the PDB. As an apo form, the same structure was used by removing the CoA molecule present in the crystal. The sulfate ions –crystal artifacts–were also removed from both structures, whereas the crystal waters were retained in the structure. Later on, the apo and holo systems were protonated at pH 7.4 using the ProteinPrep and LigPrep modules encoded in Schrödinger's Maestro software (Madhavi Sastry, Adzhigirey, Day, Annabhimoju & Sherman, 2013; Roos,

Wu, Damm, Reboul, Stevenson, Lu, Dahlgren, Mondal, Chen, Wang & others, 2019; Wizard, Epik, Prime & Glide, 2018). After that, the topology and parameter information of protein and water molecules were obtained from CHARMM36m force field (Best, Zhu, Shim, Lopes, Mittal, Feig & MacKerell Jr, 2012; Gutiérrez, Lin, Vanommeslaeghe, Lemkul, Armacost, Brooks III & MacKerell Jr, 2016; Vanommeslaeghe, Hatcher, Acharya, Kundu, Zhong, Shim, Darian, Guvench, Lopes, Vorobyov & others, 2010; Vanommeslaeghe & MacKerell Jr, 2012; Yu, He, Vanommeslaeghe & MacKerell Jr, 2012). However, as this information for the ligand COA was not available in CHARMM36m, the necessary files were generated via the Ligand Reader & Modeler tool in CHARMM-GUI (Jo, Kim, Iyer & Im, 2008; Kim, Lee, Jo, Brooks III, Lee & Im, 2017). These systems were then solvated in a water box with a minimum distance of 15 Å between each atom of the protein and the edge of the box. As a water model, TIP3P was utilized. Finally, following the solvation step, the systems were neutralized with 150 mM NaCl in order to mimic physiological ionic strength conditions.

3.1.2 Simulation Setup for Apo and Holo Systems

MD simulations of ionized apo and holo systems were performed utilizing Compute Unified Device Architecture version of Nanoscale Molecular Dynamics (NAMD), which allows computations to be accelerated through the use of graphics processing units (Phillips, Hardy, Maia, Stone, Ribeiro, Bernardi, Buch, Fiorin, Hénin, Jiang & others, 2020). All simulations were run under the NPT ensemble. To this end, the Langevin piston Nose-Hoover method was used to keep the pressure constant at 1 atm and control fluctuations occurring in barostat (Martyna, Tobias & Klein, 1994). The simulation systems were also controlled with Langevin Dynamics by maintaining a constant temperature of 310 K. Time step was set to 2 femtoseconds (fs) to capture the fastest motions in the systems. Long-range electrostatic interactions were evaluated with the particle mesh Ewald algorithm (Darden, York & Pedersen, 1993; Essmann, Perera, Berkowitz, Darden, Lee & Pedersen, 1995), while non-bonded interactions were calculated using a cut-off distance of 12 Å. After these settings, each simulation system was minimized for 1,000 steps and simulated for 500,000,000 steps. Eventually, each system was simulated for 1 microsecond (μs) in three replicates, each starting with different velocity distributions, thereby for a total of 6 μs for both systems. To enhance sampling, the second replica of holo system were extended by a few hundred nanoseconds (ns).

3.1.3 Trajectory Analyses

Trajectory analyses were performed for both apo and holo systems. Each replicate of the simulations was visualized and scrutinized in the Visual Molecular Dynamics (VMD) software (Humphrey, Dalke & Schulten, 1996; Stone & others, 1998). The analysis figures pertaining to the systems were generated in VMD and rendered with the Tachyon program embedded in VMD. The following analyses were carried out using the 'Groningen Machine for Chemical Simulations (GROMACS)' (Abraham, Murtola, Schulz, Páll, Smith, Hess & Lindahl, 2015; Apol, Apostolov, Berendsen, Van Buuren, Bjelkmar, Van Drunen, Feenstra, Groenhof, Kasson, Larsson & others, 2010).

3.1.3.1 Root-mean-square deviation (RMSD)

For having an insight into the dynamics of the systems, RMSD is widely used to analyze the trajectories obtained from the MD simulations. For RMSD calculations, the frames of a produced trajectory are superimposed to reference structure and subsequently the average distance between atom pairs of these structures calculated. Therefore, it allows to investigate the stability of the system itself and also to reveal the structural divergences between two different systems during the simulation time.

By taking the first frame of each trajectory as a reference, the backbone RMSD of the studied apo and holo systems was calculated throughout the entire trajectory by using the "gmx rms" module of GROMACS as follows (Eq. 3.1) (Abraham et al., 2015; Apol et al., 2010):

$$RMSD(t) = \sqrt{(1/N)\sum_{i=1}^{N}(r_{\mathrm{i}}(t) - r_{\mathrm{i}}{}^{ref})^2} \tag{3.1}$$

where the total number of particles is defined by N. Here, $\mathrm{r_i(t)}$ corresponds to the atomic position of i^{th} atom of the target structure at time t, whereas $\mathrm{r_i}{}^{ref}$ denotes the position of i^{th} atom in the reference structure.

3.1.3.2 Root-mean-square fluctuation (RMSF)

RMSF gives the average deviation of each residue in a protein relative to their reference position throughout a simulation, thereby pointing out flexible residues or regions in the protein.

For RMSF analysis, the backbone atoms of the apo and holo systems were selected, and their corresponding RMSF values were calculated via GROMACS's "gmx rmsf" command utilizing the following formula (Eq. 3.2) (Abraham et al., 2015; Apol et al., 2010):

$$RMSF_{\mathrm{i}} = \sqrt{(1/T)\sum_{t=1}^{T}(r_{\mathrm{i}}(t) - \langle r_{\mathrm{i}}\rangle)^2} \tag{3.2}$$

where T, $\mathrm{r_i}$, and $\langle \mathrm{r_i} \rangle$ correspond to the simulation time to be averaged, the position of i^{th} atom at time t, and the average coordinates of i^{th} atom, respectively.

3.1.3.3 Principal component analysis (PCA)

After gathering a significant amount of data from MD simulations, it is crucial to process the data further to reduce its dimensionality and eliminate the redundancy, thus detecting correlated motions in the trajectories. For this purpose, PCA is a well-accepted method to explore the essential dynamics of the systems. It simply makes use of covariance matrix construction and decomposition.

In this approach, C_α atoms of each frame in a trajectory are first superimposed with those of the reference structure. Using the displacement obtained, a 3Nx3N covariance matrix is constructed for the N number of C_α atoms using the following formula (Eq. 3.3):

$$C_{\mathrm{ij}} = \langle x_{\mathrm{i}} - \langle x_{\mathrm{i}}\rangle\rangle \cdot \langle x_{\mathrm{j}} - \langle x_{\mathrm{j}}\rangle\rangle \tag{3.3}$$

where the generated covariance matrix is denoted by $\mathrm{C_{ij}}$. The displacement of i^{th} and j^{th} atoms from the time-averaged structure is indicated by $\mathrm{x_i}$ and $\mathrm{x_j}$, respectively.

Thereafter, the covariance matrix is diagonalized, and eigenvalues and eigenvectors were calculated from the resulting matrix for each of the interested systems (Eq. 3.4):

$$Cv = \delta^2 \nu \tag{3.4}$$

Here eigenvalues and eigenvectors pertaining to the diagonalized matrix are shown by δ_2 and ν, respectively. Eigenvectors give information on directions of essential motions, whilst eigenvalues provide an insight into the corresponding magnitudes of these motions.

The global dynamics of the studied apo and holo systems were demystified with the help of GROMACS's "gmx covar" and "gmx anaeig" commands (Abraham et al., 2015; Apol et al., 2010) and Python package, ProDy (Bakan, Meireles & Bahar, 2011). The trajectories of the systems were aligned along with the first three eigenvectors. In addition, the 2D principal components of the systems were calculated by projecting the first two eigenvectors of apo in 2D space.

3.1.3.4 Timeline analysis and probability distribution of selected reaction coordinate

During the analysis of the trajectories of apo and holo systems, two distinct regions showed the highest contributions to local and overall dynamics. The distance between the centers of masses of these two regions was selected as a reaction coordinate and measured throughout the trajectories by means of the "gmx distance" command of GROMACS (Apol et al., 2010). Thus timeline data was plotted. To convert the data into probability distributions, the minimum and maximum sampled distances were determined. Considering these distances, the frequency interval was set to 0.5 Å to represent the data in a more sensitive and accurate way. After that, frequencies of the sampled distances were calculated and then converted into probability distribution within a range of 0 to 1.

3.2 Results and Discussion

3.2.1 Having an Insight Into the Stabilities of Apo and Holo Systems

Before getting into detailed MD trajectory analyses, the convergence of the apo and holo systems throughout the simulation time should be checked by means of RMSD. This analysis also provides a rough picture of structural and dynamical changes occurring in the protein during a particular simulation system and/or between different systems. To do so, backbone RMSD profiles of the apo and holo systems were analyzed separately for each simulation replicate (Figure 3.1).

The RMSD values for the first replica of the apo system slightly increased during the first 650 ns and then settled down at around 2 Å until the end of the simulation. Here, between 540-600 ns, the RMSD values reached up to approximately 3 Å. In the second replica of apo, the system fluctuated at about 1.6 Å throughout the first 360 ns, and then the values gradually increased to 3 Å between 360 and 540 ns and flattened there smoothly. A similar RMSD pattern was also observed in the third replica. In this run, the RMSD remained nearly steady at around 1.6 Å for the first 910 ns, and it then reached up to 3 Å. In general, drastic alterations in RMSD plots indicate instability or conformational changes in the systems. To elucidate the changes in the apo structure corresponding to these abrupt deviations, frames where the RMSD increased to 3 Å were visually inspected for all three replicates by VMD. In these investigations, a notable structural change was observed in the fourth (α4) helix of the protein structure, which consists of 95-110 residues. In these frames, a certain portion of the α4 helix (102-107 residues) seemed to unfold and became a mobile loop. Thus, the helix structure was disrupted and turned into two small helices connected by a loop. This flexible loop induced a bending that can be perceived as a distinct conformational state sampled in the apo structure. On the other hand, the RMSD values for three different replicates of the holo structure smoothly fluctuated around 1.6 Å. Only towards the end of the second replica a slight rise in the RMSD was observed. To ensure whether the holo structure could sample another conformational state as in apo, the second simulation was extended for an additional 500 ns. As a result of this extension, it was observed that the RMSD increased to about 1.8 Å and stayed there (Figure B.1). This difference ($\sim$ 0.2 Å) is not drastic enough to be an indicator of conformational transition. In parallel to this, when the trajectories were examined, no disruption in the α4 helix was observed throughout the holo trajectory, unlike in the apo case.

To sum up, during the MD simulations, the apo form started to deviate from its crystal structure after a certain time, and this structural distinction in the protein might indicate a change in its conformational state. In contrast, when CoA was

bound to the protein, it remains fairly stable throughout the trajectory. Therefore, it can be deduced that CoA binding increased protein stability.

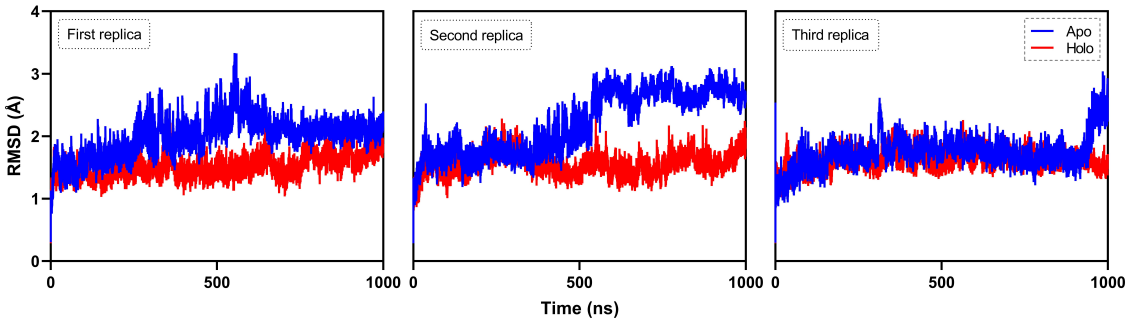

Figure 3.1 Backbone RMSD of apo and holo systems calculated from MD trajectories.

3.2.2 Local and Global Structural and Dynamical Investigations of the Apo and Holo Systems

To understand the local movements in the apo and holo structures and thus unveil which regions of the protein contribute mostly to its motion, RMSF profiles of both systems were compared. For this, the first 40 ns of three trajectories of each system was excluded as the system got equilibrated during this period. The remaining 960 ns of each trajectory was divided into 8 chunks with 120 ns-long segments. The RMSF values of the divided chunks for three replicates of each system were calculated separately and then averaged. Deviations from the calculated mean values were also indicated by standard error of mean.

Comparative RMSF results highlighted significant changes in the fluctuation pattern of the protein residues in response to CoA binding (Figure 3.2). Taking a closer look at the RMSF data, two distinctive peaks were observed in two regions: region 1 (residues 39-47), which corresponds to a loop, and region 2 (residues 94-110), which corresponds to the α4 helix. While these regions were highly dynamic and mobile in the absence of CoA, they became relatively stable in the presence of CoA. Consistent with findings in the visual inspection of the apo trajectories, α4 helix destabilization in the apo structure was also deduced from RMSF results. In addition to these two characteristic regions, two shallow peaks were observed at residues 8-15 and 128-140 that fluctuated slightly in the apo compared to the holo. However, in contrast to regions 1 and 2, the mobility difference in these residues was unnoticeable.

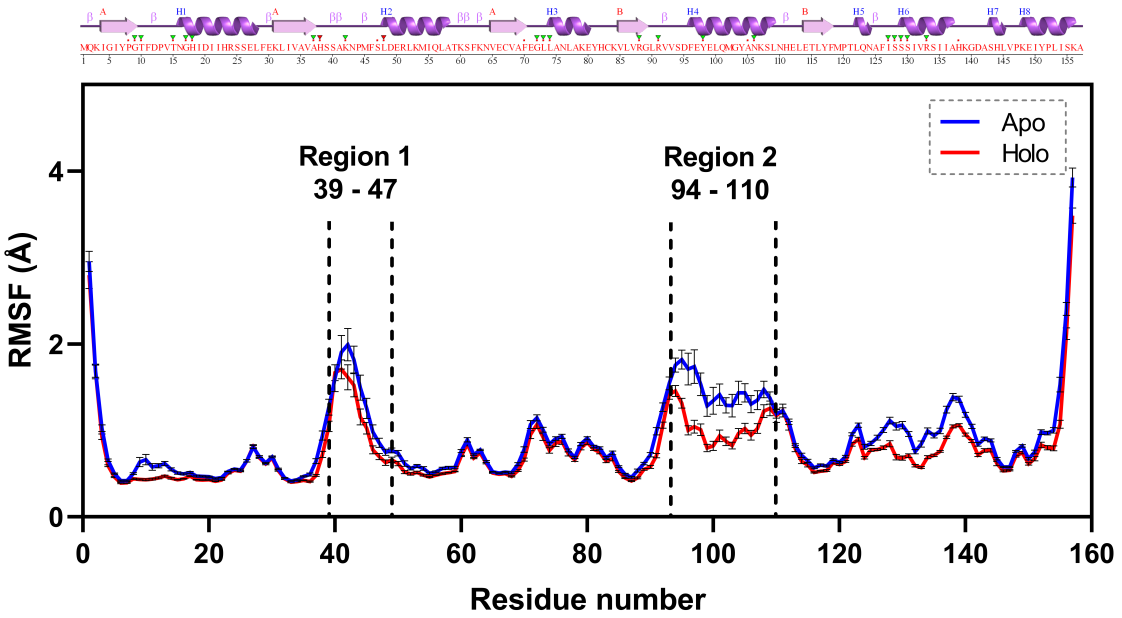

Figure 3.2 Backbone RMSF of apo and holo systems calculated by averaging the corresponding 120 ns-long windows of MD trajectories. Error bars indicate the standard error of the calculated means for fluctuation of each residue.

Besides local analysis, the global motions of the systems were also examined by principal component analysis (Figure 3.3). To this end, the eigenvalues and eigenvectors of both systems were first calculated. Thereafter, the C_α atoms of the trajectories were aligned with respect to the first three essential eigenvectors, which cumulatively accounted for approximately 46% of the overall protein dynamics. As a result of the analysis, it was revealed that the mobility in regions 1 and 2 caused a substantial alteration in the essential dynamics of the protein. The mobility in these two regions decreased by CoA binding. In particular, the fluctuations observed in the $\alpha 4$ helix residues (region 2) significantly contributed to the overall protein dynamics. On the other hand, fluctuations in region 1 were still observed with CoA binding; however, this mobility was less remarkable compared to the CoA-unbound form. Furthermore, it should also be noted that the movements in the residues 8-15 and 128-140 observed in the RMSF profiles did not have a significant impact on the overall protein dynamics. Therefore, as a result of these analyses, it can be pointed out that the local mobility in regions 1 and 2 also significantly dominated the collective dynamic motions of the protein. Furthermore, to better comprehend the changes in protein dynamics induced by CoA binding, 2D projections of the trajectories were taken according to the first two eigenvectors of the apo system, cumulatively corresponding to 23.5% and 12.1% of the overall mobility. The results showed CoA binding confined the conformational space sampled by the protein. To put it another way, the CoA-unbound system may be able to adopt other confor-

mational states, which cannot be seen in the CoA-bound system. Therefore, this result also confirmed that, in the absence of ligand, the protein could exhibit other distinct conformational arrangements in which the $\alpha 4$ region was destabilized, and the protein started to unfold.

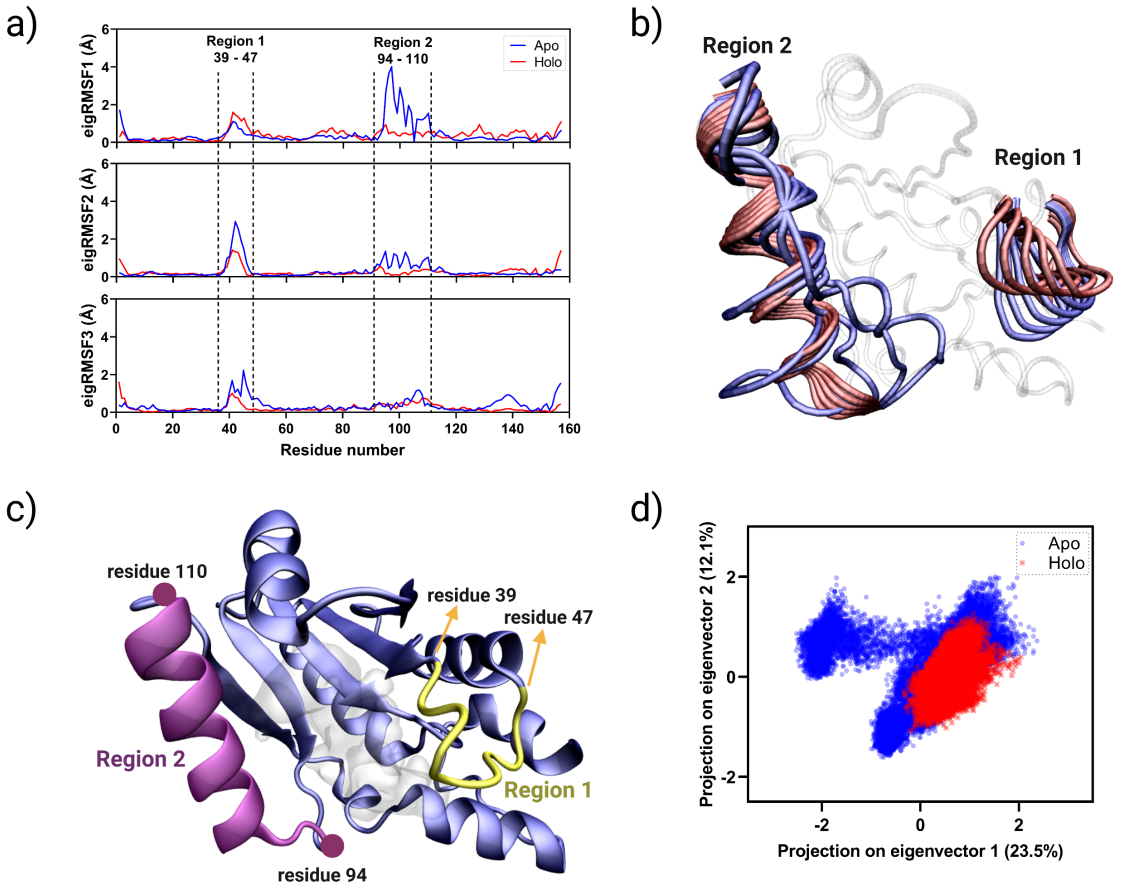

Figure 3.3 Investigation of global dynamics of apo and holo systems. a) RMSF of C_α atoms along with the first three eigenvectors of the systems. b) Obtained extreme structures of apo and holo systems along their first eigenvectors. The apo system is illustrated in metallic blue whereas the holo system is colored in metallic red. c) A representative 3D structure of the apo form, where two regions —regions 1 and 2—display higher flexibility than the holo are colored with yellow and purple, respectively. d) 2D projections of the apo and holo systems with respect to first two eigenvectors of the apo system.

Finally, to quantify the extent of space between regions 1 and 2 during MD simulations, the distance between the centers of mass of these regions was taken as a reaction coordinate (Figure 3.4). To that end, the distance was calculated for each of the three replicates run for apo and holo systems and then averaged. The obtained results indicated that, in all three replicas of the holo system, the distance between these regions deviated around 20-25 Å. However, in the apo system, the distance varied in a broad range between 15-30 Å. The deviations in the reaction distance

observed in the apo system were evidently greater than those in the holo system, as can be seen in its probability plot. Therefore, these results may also validate that the CoA-unbound structure tries to adapt to new conformational states over the course of simulation time, thereby being considered as a trace of the protein unfolding process.

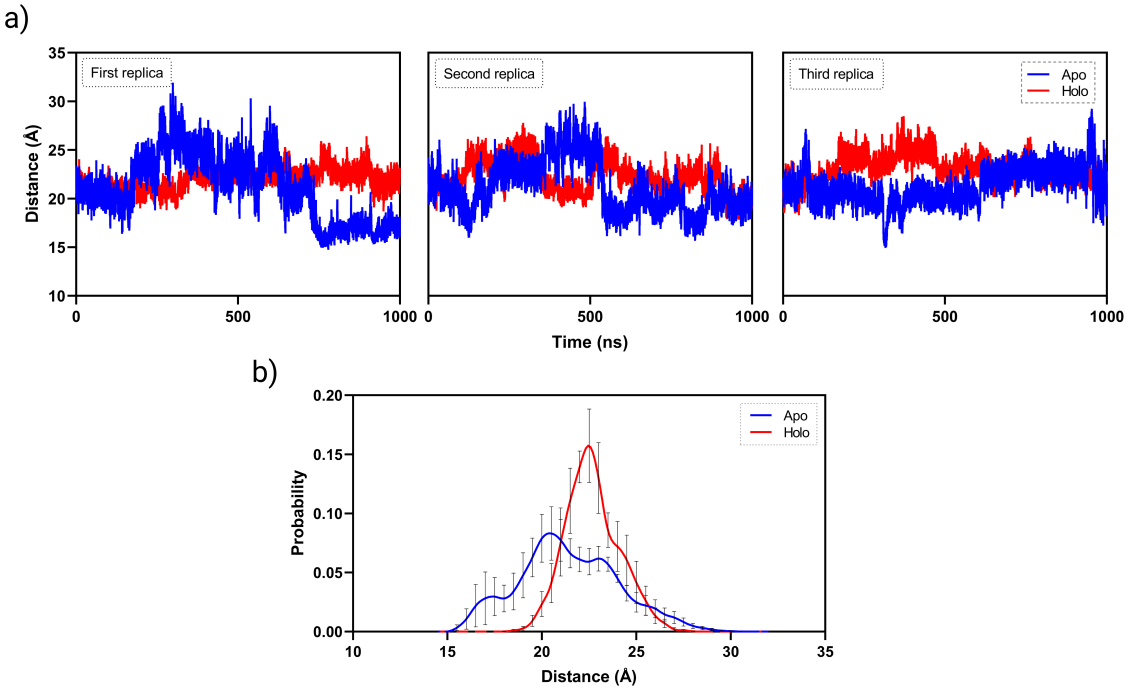

Figure 3.4 Distance measured between a) centers of mass of regions 1 and 2 of apo and holo systems as well as b) their probability distributions.

In conclusion, the ultimate goal in running MD simulations was to see if the selected sensing protein undergoes large conformational changes upon CoA binding, and if so, then to determine possible locations in the protein for fluorescent probe insertion. Here the expected structural change seems to be involved in the unfolding of a particular region within the protein. Since unfolding events occur on the millisecond timescale, it is not possible to access such substantial conformational transitions with classical MD simulations (Fabian & Naumann, 2012). Nevertheless, with the help of MD simulations, fingerprints of this large-scale change in the protein structure can be captured.

In the light of analysis results obtained, two regions (regions 1 and 2) in the protein have come into prominence as the CoA binding to the protein gives rise to a substantial decrease in the mobility of these regions. Considering this, the sites where the fluorescent protein could be inserted are determined as the next step. From the successful applications of biosensors, it is known that in order to propose

plausible single FP-based biosensor designs, the insertion locations should both undergo significant conformational changes by ligand binding and also permit cpGFP fusion. Taking into account these considerations, two possible insertion locations in regions 1 and 2 were selected: residue 41 and residue 94 (Figure 3.5). The first possible insertion site was determined as residue 41 in region 1. This is because the residues 41 and 42 were the most flexible residues in this loop. Since flexible parts of proteins, such as loops, are more likely to be permissive to insertions, it may be promising to attach fluorescent protein at this position. The second insertion site was selected as residue 94, located in region 2. Here residue 94 was not involved in helix-formation but was instead located on the loop connecting the $\beta 4$ sheet and $\alpha 4$ helix. Therefore, this residue might also tolerate the cpGFP fusion to the protein. Overall, these two residues were thought of as promising locations to attach cpGFP into the CoA-sensing domain, and possible single cpFP-based CoA biosensor models were designed by constructing fusion proteins using these insertion positions.

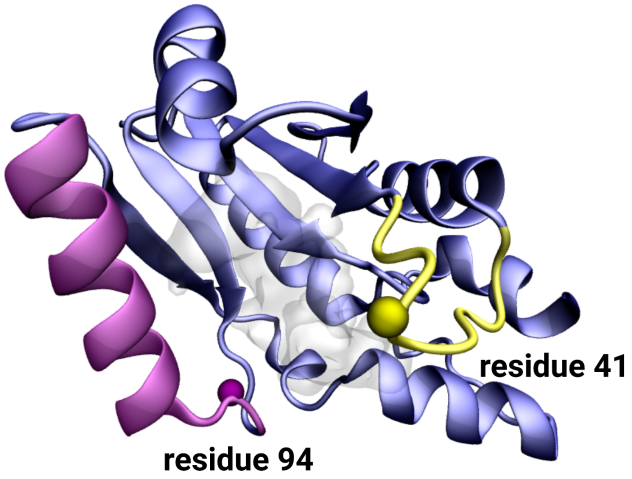

Figure 3.5 Identified cpGFP insertion locations for construction of the chimeric protein consisting of cpGFP and sensing domains.

4. MODELING A CHIMERIC PROTEIN CONSISTING OF CPGFP AND COA-SENSING DOMAIN

In this chapter, the cpGFP sequence was inserted into the sensing domain sequence from two separate insertion sites using GGS flexible linkers, thus obtaining different chimeric protein sequences. Thereafter, folded structures of these chimeric sequences were predicted by AF2, which is an artificial intelligence software that accurately predicts 3D structures of proteins from their primary sequences using multiple sequence alignment (Jumper et al., 2021). As a result of these predictions, for each unique sequence, the top-ranked structures were selected among the predicted models. Then, these structures were analyzed to have an insight into the folding of the domains as well as the positions of some key residues in these domains. Therefore, the most promising ones were proposed as initial design ideas for single cpGFP-based CoA biosensor construction.

4.1 Methods

4.1.1 Sequence Preparation of Chimeric Proteins with Different Linker Lengths

To build CoA biosensor models with a cpGFP inserted into the sensing domain, the crystal structure of cpGFP (PDB ID: 3EVP) was first retrieved from PDB and its sequence was extracted as fasta format using PyMOL. Thereafter, the fasta sequence was aligned with the given sequence on UNIPROT via EMBOSS Needle tool (Madeira, Pearce, Tivey, Basutkar, Lee, Edbali, Madhusoodanan, Kolesnikov & Lopez, 2022). Accordingly, the first four residues (SSLE) were removed from the

sequence. As gate post residues, His (H) amino acid was added to the beginning of the sequence, whilst Phe (F) was added to the end of the sequence. Therefore, a new cpGFP sequence comprised of gate post residues (H and F), threonine-tyrosine-glycine (TYG) chromophore, and cp linker (GGTGGS) was saved to be used as a fluorescent domain for the CoA biosensor model. On the other hand, the fasta sequence of the studied sensing domain was also extracted by PyMOL.

After acquiring the sequences, chimeric proteins consisting of both fluorescent and CoA-sensing domains were prepared by attaching cpGFP to the sensing platform from the two separate insertion sites: residues 41 and 94. For this construction, the cpGFP sequence was added to the sensing domain right after the selected insertion position, then followed by the insertion of the remainder of the sensing domain sequence. In addition, flexible GGS linkers were added from two different linker regions, which correspond to just before the His gate post residue (first linker region) and/or after the Phe gate post residue (second linker region). Hence, the effect of connecting the domains using linkers on the three-dimensional structure of the resulting chimeric protein was be able to evaluate. In this way, a couple of sequences of chimeric CoA biosensor models with various linkers placements were prepared.

4.1.2 Three-dimensional Structure Predictions of Chimeric Proteins via AlphaFold2

Later, three-dimensional structures of the prepared chimeric proteins were predicted via the AF2 program (Jumper et al., 2021). These predictions were run on ColabFold webserver, which combines AF2 with MMseqs2's fast homology search, as well as enables fast prediction by providing free GPU. All settings in Colabfold were kept as default, except relaxation mode. To relax the positions of amino acids' side chains in the predicted structures, Amber relaxation option was selected. Among the resulting relaxed models for each sequence, the one with the highest mean pLDDT value was selected as the model structure for that particular sequence. Finally, to find out whether each domain in the chimeric structures was folded properly or not, these models were superimposed with both the sensing domain and cpGFP separately. For superimposition, the "align" command of Pymol was used. Last but not least, the folding patterns of the models along with the positions of the gate post residues and binding site residues were then examined.

4.2 Results and Discussion

To construct genetically encoded protein-based cpFP biosensor models, it is recommended to insert a FP domain at a location in the sensing domain that shows a large conformational change depending on the presence/absence of ligand of interest (Nasu et al., 2021). The main reason for this is that the conformational mobility in the sensing domain is capable to disrupt the hydrogen network around the chromophore. This disruption then gives rise to a change in the protonation state of the chromophore and its planarity, thereby affecting its fluorescence quantum yield (Pakhomov & Martynov, 2008; Patnaik, Trohalaki & Pachter, 2004). Considering these, two different positions in the sensing domain (just after residues 41 and 94) were determined as cpGFP insertion sites. Thereafter, the cpGPF fasta sequence was inserted into the sensing domain sequence at these two different positions, separately. To do so, GGS flexible linkers were also used to connect fluorescent and sensing domains so as to minimize steric clashes between these domains and also to facilitate the proper folding of each domain. Later on, the relaxed 3D structures of the prepared chimeric sequences were predicted by the AF2 program, and subsequently, these models were examined in terms of (i) the potential of both domains to maintain their 3D structure after fusion, (ii) the preservation of the active site residues' positions, and (iii) the changes in the orientation of the gate post residues.

4.2.1 Constructing Chimeric Proteins with cpGFP Insertion at Residue 41 of the Sensing Domain

In the context of proposing single cpGFP-based CoA biosensor models, the cpGFP probe was attached to the sensing domain at a position adjacent to residue 41. For this fusion, GGS linkers were used at two different linker positions: first linker region and second linker region. Thus, four different models were prepared by attaching cpGFP to the CoA-binding domain (i) without a linker, with (ii) GGS linker at the first linker region, (iii) GGS linker at the second linker region, and (iv) GGS linkers at both linker regions.

As the first model, cpGFP was fused to the CoA-sensing domain after residue 41 of the binding domain without using a linker (Figure 4.1). To identify changes in the three-dimensional structures of domains pertaining to the predicted model

upon fusion, the chimeric protein was aligned with the sensing domain and cpGFP, separately, and the corresponding backbone RMSD values were measured as 0.40 Å and 0.24 Å, respectively. This indicates that the obtained chimeric protein without the use of a linker allowed both domains to maintain their structures after fusion. Afterward, a closer look was taken at these alignments, and it was noted that the binding site residues in the predicted model were also nicely aligned with those in the original sensing domain. Regarding the fluorescent domain, both gate post residues still retained their orientation towards the chromophore after fusion. Therefore, this model seems promising to be proposed as an initial design for a single cpGFP-based CoA biosensor.

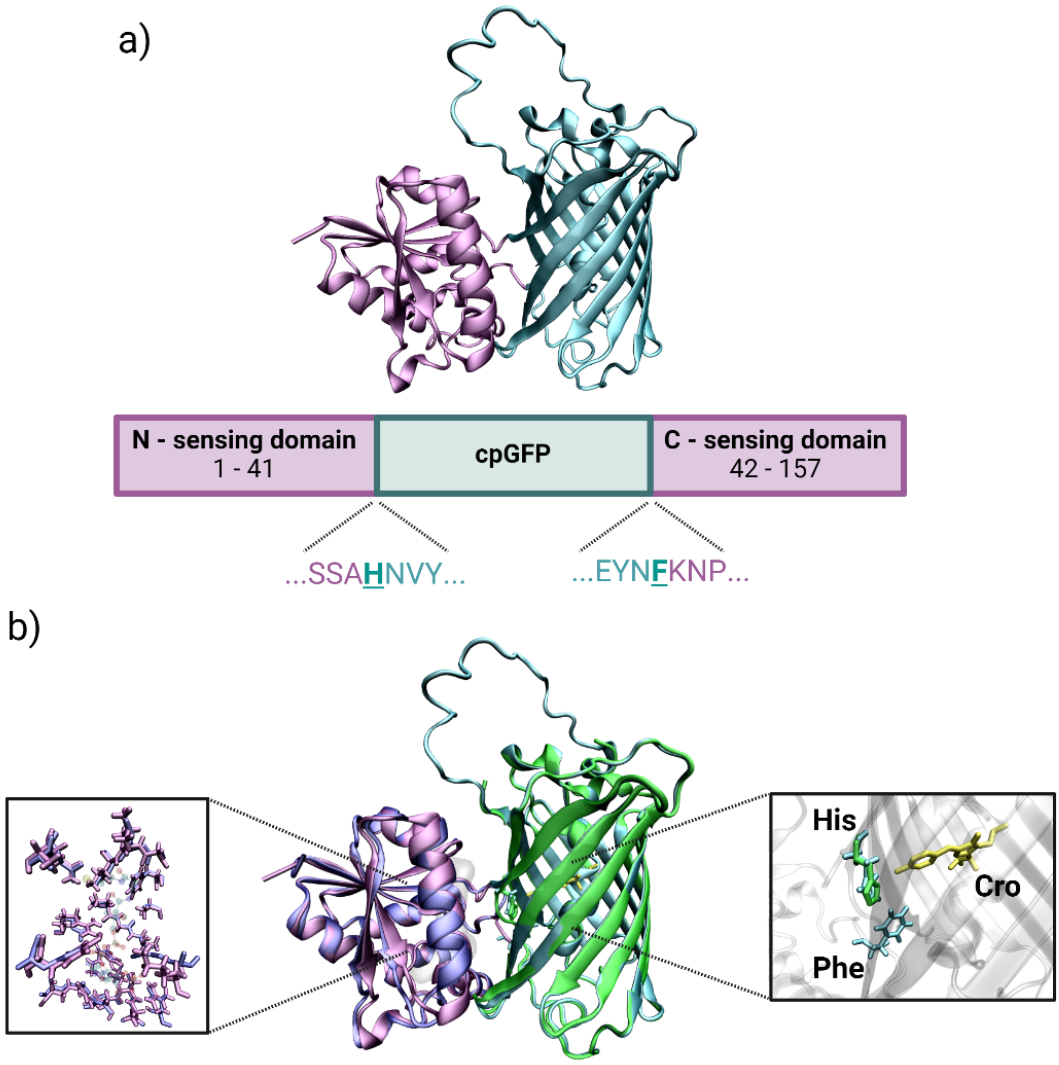

Figure 4.1 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 41. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-41) is linked with cpGFP, followed by C-sensing domain (42-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

For the second model, the chimeric protein was constructed by connecting cpGFP to the sensing domain via the GGS linker at the first linker region (Figure 4.2). Thereafter, the original sensing and cpGFP domains were superimposed onto the chimeric protein separately. Accordingly, the RMSD values were measured as 0.56 Å for structural deviations in the sensing units, and 0.27 Å for deviations in the fluorescence units. When the CoA-binding residues were inspected in detail, a noticeable shift in the position of the residue Lys288 (equivalent to Lys42 of the original sensing domain) away from the CoA molecule was noticed. To understand how this deviation might affect the ligand-protein interaction, the distance between the NZ atoms of residues Lys pertaining to the original sensing domain and the chimeric protein was measured and found to be 4.15 Å. Since the amine group of Lys makes hydrogen bond interactions with all three phosphate groups of the CoA molecule, this deviation might presumably disturb the hydrogen bonding network between residue Lys and CoA. Besides the functional importance of this residue, it is also located adjacent to the insertion site. That is to say that the mobility of this residue might lead to an orientational change in the first gate post residue and ultimately trigger the disruption of the chromophore microenvironment. Therefore, such positional deviation in the residue Lys288 would not be favored. In addition to sensing domain investigations, the gate post residues' orientations in the fluorescent domain were analyzed, and seen that both gate posts were oriented towards the chromophore, thus tending to keep the cpGFP's gate closed.

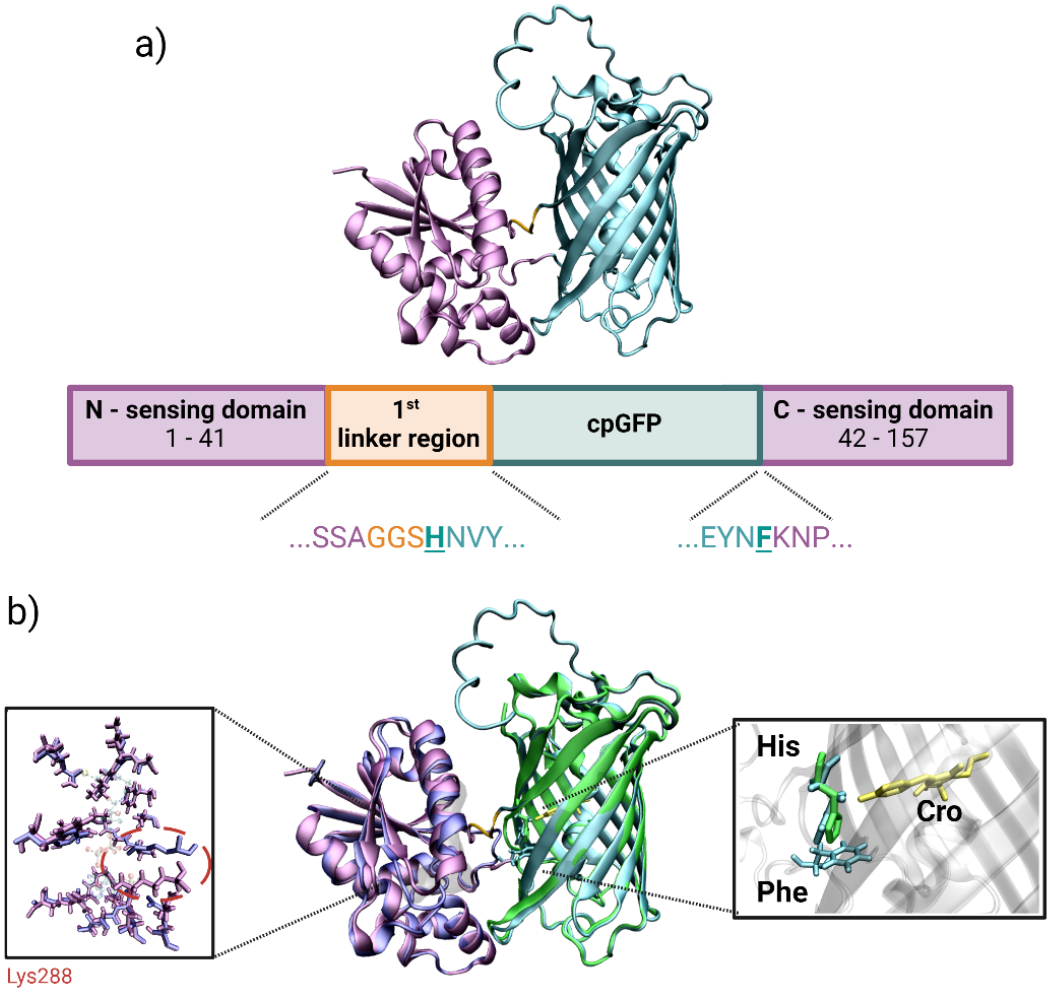

Figure 4.2 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 41 using GGS linker at first linker region. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-41) is linked with cpGFP using GGS linker, followed by C-sensing domain (42-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

As the third model, the GGS linker was used at the second linker region while constructing the chimeric protein sequence (Figure 4.3). The performed structural alignment results revealed that the fluorescence domain of the model was quite nicely overlayed with the original cpGFP domain (RMSD 0.31 Å). Besides, the orientations of the gate post residues in the chimeric protein were preserved as close as possible to those in the original cpGFP. In contrast to the fluorescence domain, noticeable alterations were observed in the CoA-binding domain, which can be measured as a backbone RMSD value of 0.55 Å. Notably, among the active site-forming residues, Lys288 and Arg379 of the chimeric protein (corresponding to residues Lys42 and Arg133 in the original sensing domain) altered their positions when fused to

cpGFP. Considering the fact that both residues are involved in hydrogen bond formation and also the positional shifts increase the distance between hydrogen bond pairs, it may raise a question mark concerning being proposed as an initial design for the CoA biosensor model.

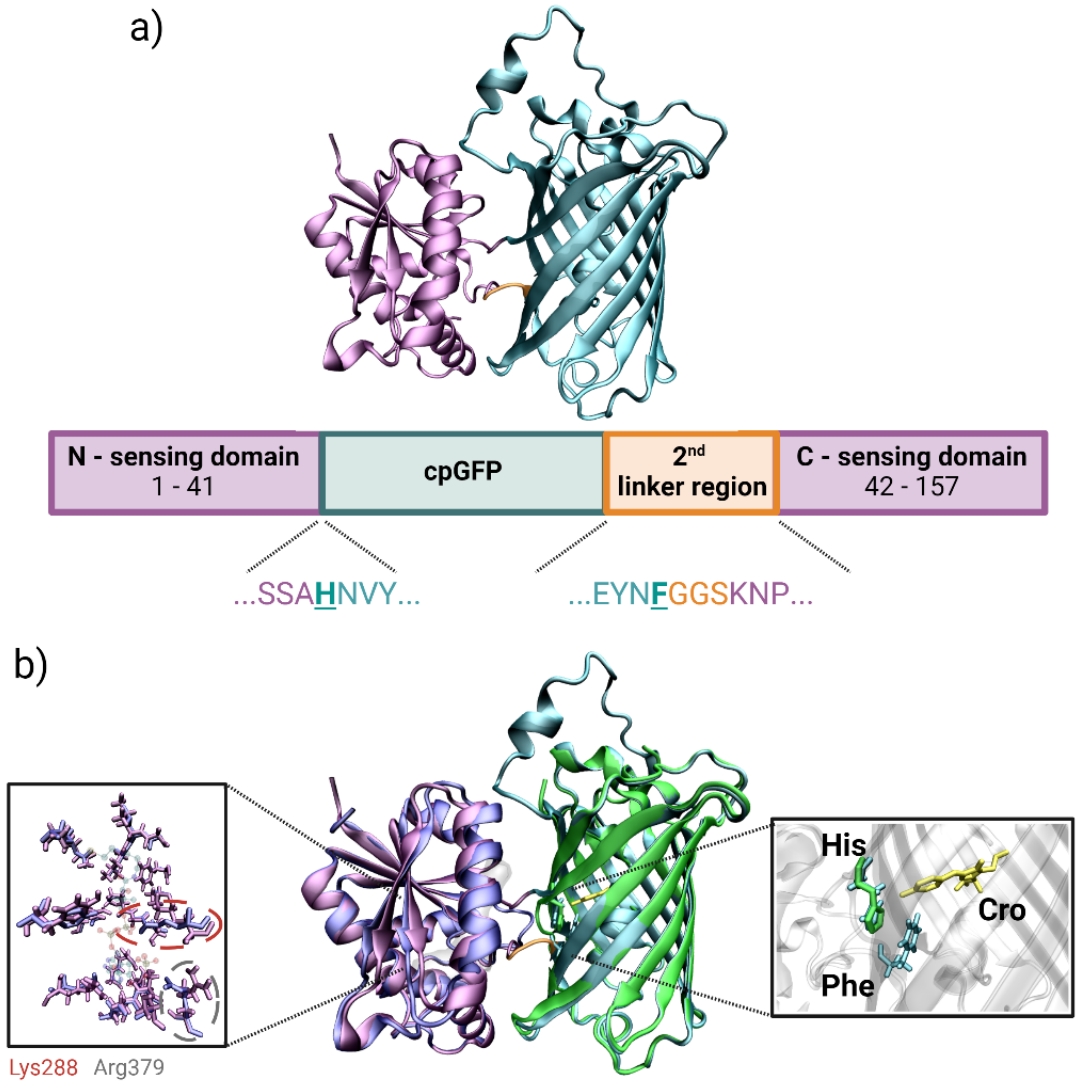

Figure 4.3 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 41 using GGS linker at second linker region. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-41) is linked with cpGFP, followed by GGS linker and C-sensing domain (42-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

For the last chimeric protein model, GGS linkers were added from both the first and second linker regions (Figure 4.4). By individually superimposing the sensing and fluorescence units onto the predicted model, the backbone RMSD values were

calculated as 0.55 Å and 0.30 Å, respectively. As observed in other chimeric models fused with GGS linkers, it can be also noticed that the residue Lys291 of chimeric protein positionally deviated. Regarding the gate post residues orientations, they were observed to be positioned towards the inside of the barrel, thus keeping the gate closed. Although the preservation of the resulting chimeric protein's cpGFP domain structure, as well as its post gate residue positions, seems promising for CoA biosensor design, the positional changes observed in the sensing domain, particularly in the active site region, are not desirable.

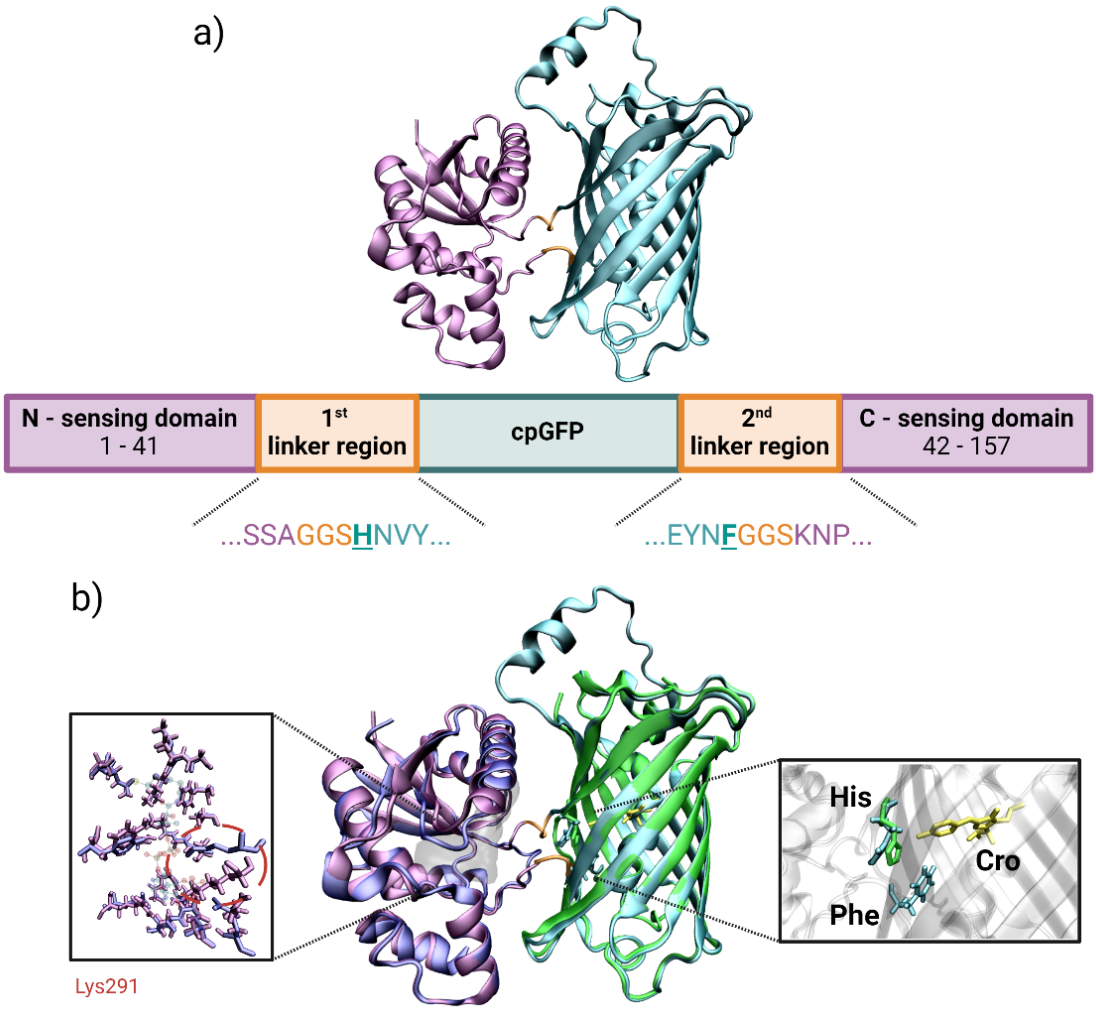

Figure 4.4 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 41 using GGS linker at both linker regions. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-41) is linked with cpGFP via GGS linker, followed by second GGS linker and C-sensing domain (42-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

4.2.2 Constructing Chimeric Proteins with cpGFP Insertion at Residue 94 of the Sensing Domain

Using the same strategy as in the previous constructions, the same linker placements were utilized, and the cpGFP domain was fused to the sensing domain just after residue 94 of the sensing domain.

As the first model, the chimeric structure was generated without using a linker (Figure 4.5). It was superimposed into the sensing and cpGFP domains, and as a result, the backbone RMSD values were found as 0.41 Å and 0.25 Å, respectively. After performing superimpositions, the active site residues were closely examined. Here substantial shifts in the positions of residues Arg91 and Tyr341 of the chimeric protein (corresponding to Arg91 and Tyr98 of the original sensing domain) were observed. Among these residues, Arg makes hydrogen bond interactions with the oxygen in the ribose group of the CoA molecule whereas residue Tyr is involved in the hydrogen bonding with the oxygen atom of one of the phosphate groups in the CoA. Since these residues take part in ligand-protein interactions at the active site, it can conceivable that they might facilitate CoA accommodation to the cleft. Therefore, the weakening or disruption of these interactions may cause the protein to lose its binding property. In addition to sensing domain analysis, gate post residue positions in the cpGFP domains were also examined. As it can be seen from the figure, the gate post residues were prone to stay away from each other rather than cover the bulge region. Although the behaviors of these residues under physiological conditions can be more comprehended by MD simulations, the gate post orientations infer that these configurations may make the chromophore solvent accessible. Having all these in mind, this model may not be suitable to be suggested as an initial design.

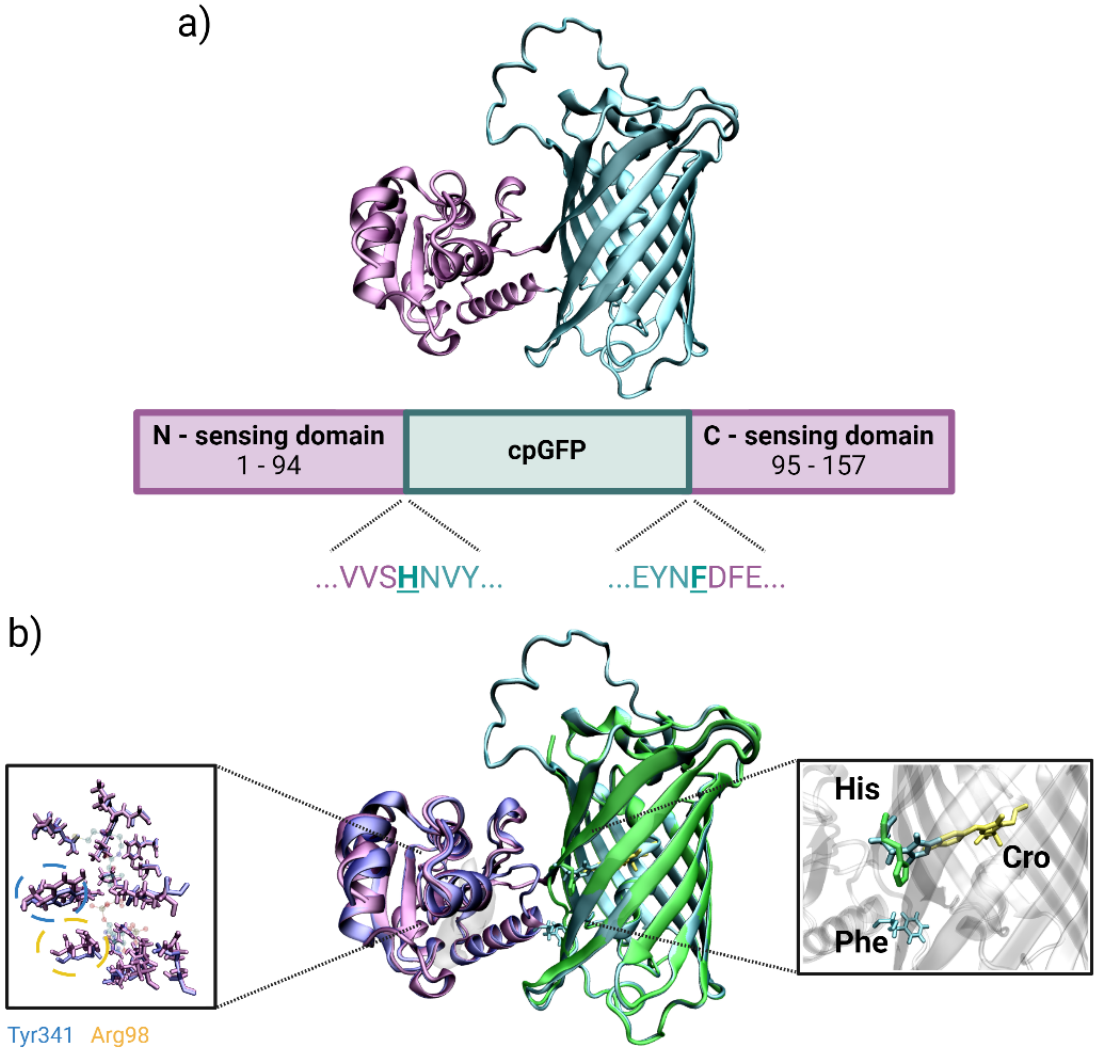

Figure 4.5 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 94. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-94) is linked with cpGFP, followed by C-sensing domain (95-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

As the second model, cpGFP was inserted into the sensing domain using GGS linker at the first linker region (Figure 4.6). After superimposing two domains into the resulting chimeric structure, the RMSD values were measured as 0.55 Å for the alignment of the CoA-binding domain and 0.30 Å for that of the fluorescence domain. From the visual inspections, no serious structural alterations in the chimeric protein were observed except around the linker region. Similarly, active site residues in the chimeric protein also preserved their original positions. Regarding the fluorescent domain part, it is observed that the 3D structure of the cpGFP domain was preserved after fusion, and both gate post residues were spatially positioned to-

wards the chromophore. Considering these findings, this model appears to be worth further investigation as a candidate model for constructing a CoA biosensor.

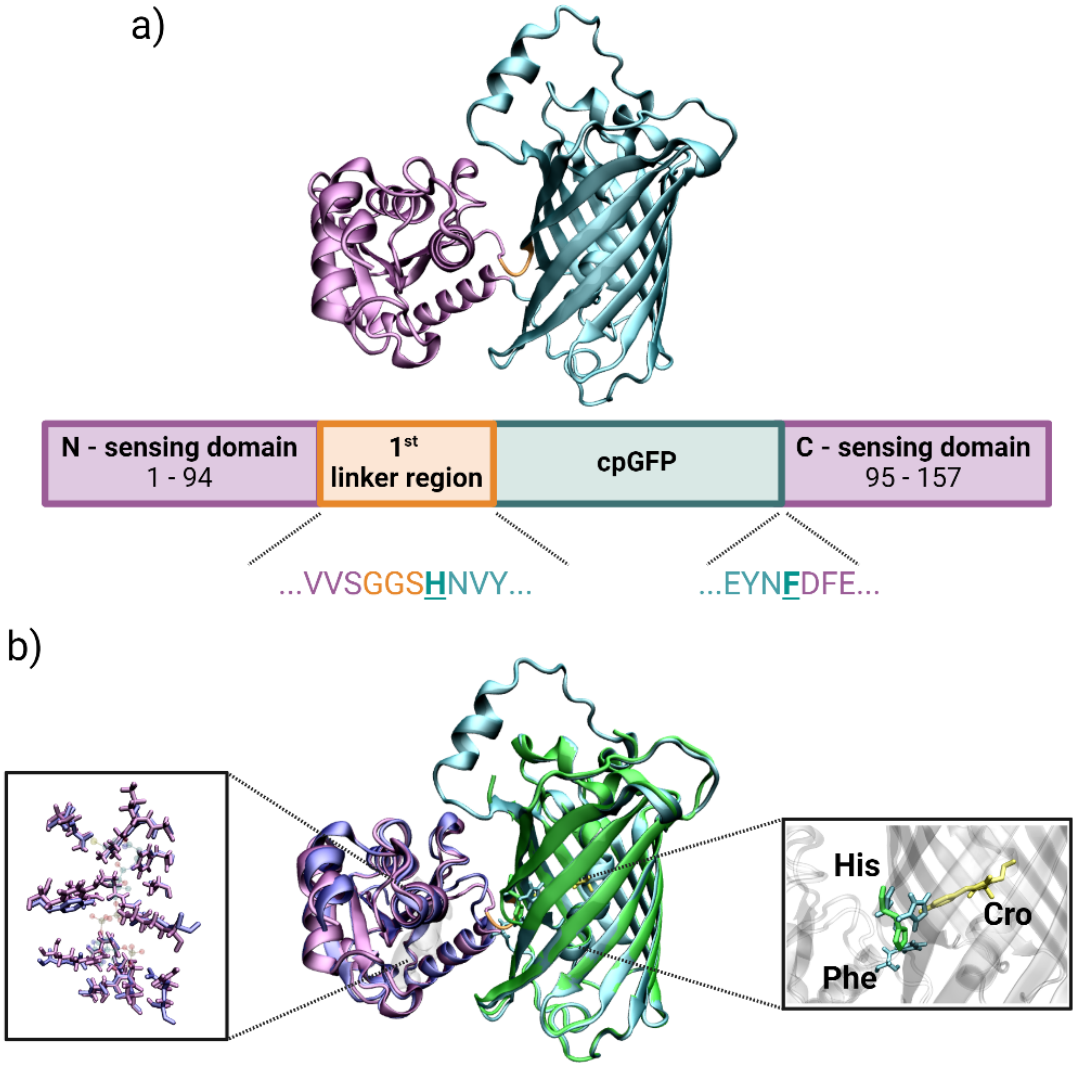

Figure 4.6 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 94 using GGS linker at first linker region. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-94) is linked with cpGFP via GGS linker, followed by C-sensing domain (95-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

As the third model, the chimeric protein was constructed by attaching the cpGPF using the GGS linker from the second linker region (Figure 4.7). The resulting model structure was superimposed onto the sensing domain, and the corresponding RMSD value was calculated as 0.38 Å, which indicates the sensing domain was able to conserve its 3D structure after cpGFP fusion. After alignment, residues around the active site were examined and compared. As a result of this, there is

no drastic change in the positions of the active site residues observed. Thereafter, the deviation between the original cpGFP and fluorescence domain of the resulting chimeric structure was measured and found as 0.25 Å. This also signifies that the fold of the cpGPF domain was predicted as it is nicely preserved after fusion. Besides, it was also found that the gate post residues on the cpGFP of the chimeric protein were pointed towards the inside of the cpGFP β-barrel. Therefore, this chimeric model emerges as a promising candidate for biosensor design.

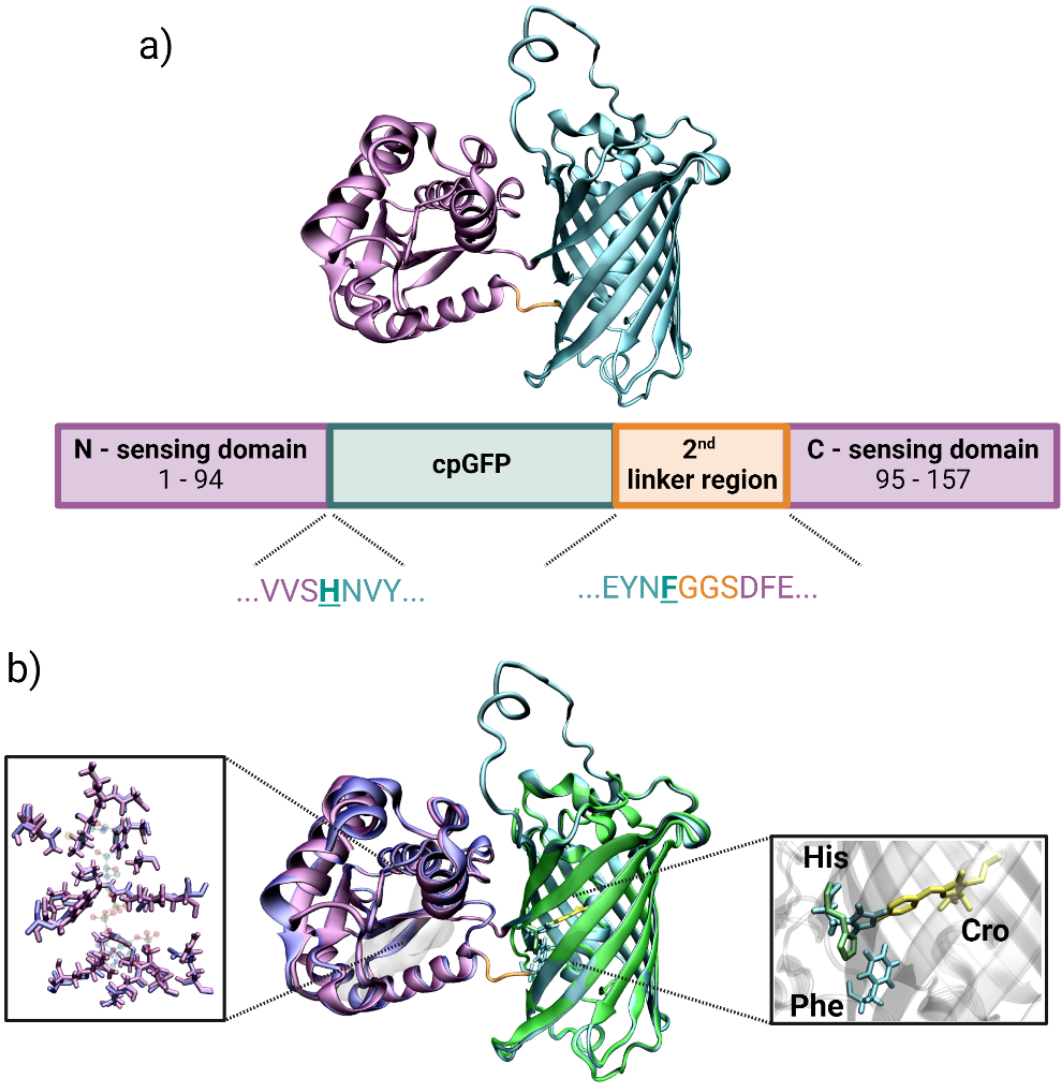

Figure 4.7 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 94 using GGS linker at second linker region. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-94) is linked with cpGFP, followed by GGS linker and then C-sensing domain (95-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

As the final model, cpGFP was connected to the CoA-sensing protein via GGS flexible linkers from both linker regions (Figure 4.8). In this case, the backbone RMSD values were calculated as 0.4 Å for sensing domains and 0.24 Å for cpGFP units. The deviations between the active site residues of the original sensing domain and those of the resulting chimeric protein were examined more closely. As a result of this, similar to the chimeric structure constructed using no linker, significant shifts were noticed in the positions of Arg91 and Tyr347 of the constructed protein. Regarding the FP domain, the gate post residues were positioned towards the chromophore region, keeping the bulge region close. Therefore, this model might not be a good candidate to be proposed as a CoA biosensor design.

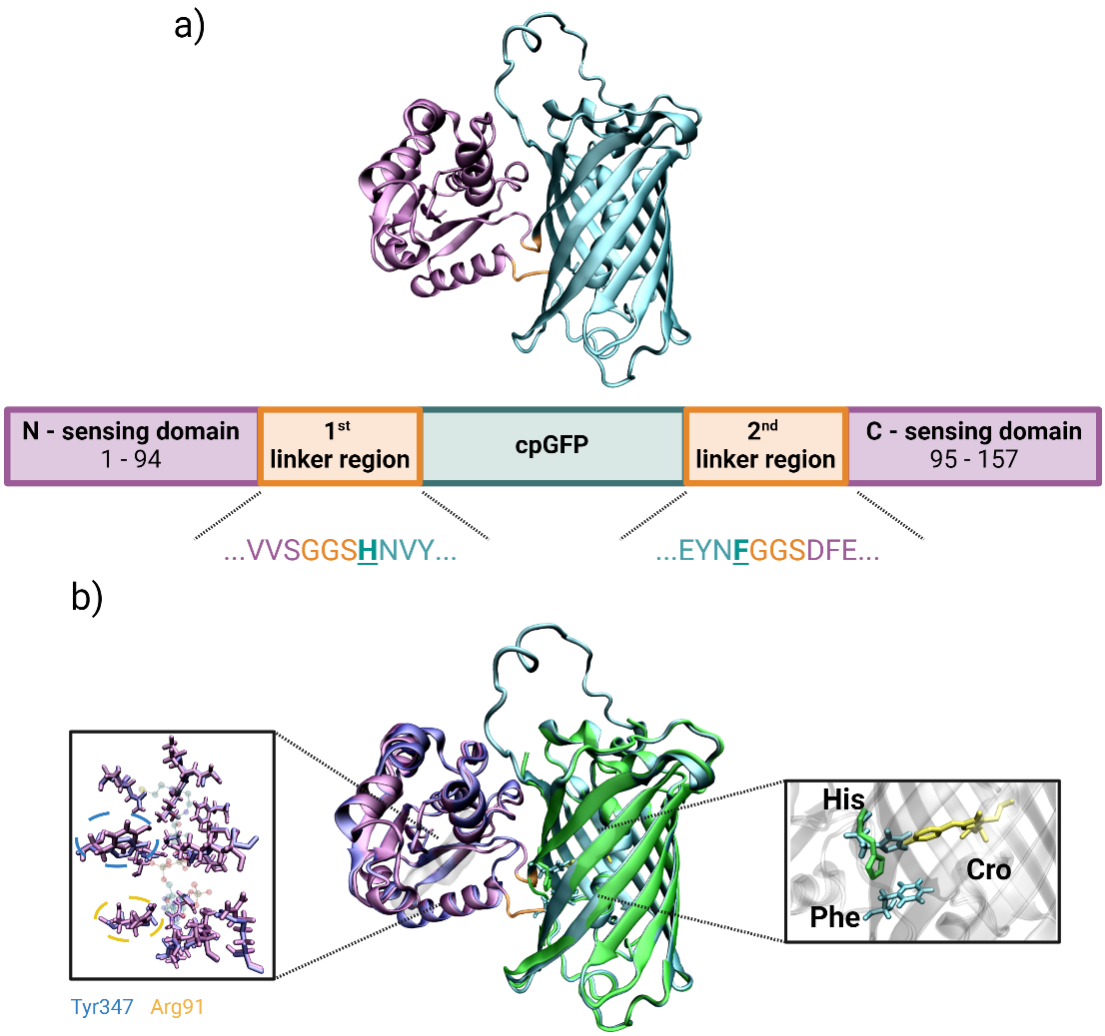

Figure 4.8 Construction of genetically encoded biosensor as chimeric protein inserted cpGFP into sensing domain at residue 94 using GGS linker at both linker regions. a) Schematic representation of construction mechanism of the biosensor. N-sensing domain (1-94) is linked with cpGFP via GGS linker, followed by another GGS linker and then C-sensing domain (95-157). b) Alignment of constructed chimeric protein with sensing domain and cpGFP. Left panel represents a zoom-in view around the active site, while right panel represents a zoom-in view around the chromophore. (Created with BioRender.com)

To sum up, the cpGFP domain was inserted into the CoA-sensing protein from two different insertion sites, namely residue 41 and residue 94. While constructing the sequences, GGS flexible linkers were also used just before and/or after the post-gate residues to attach the domains together. Thereafter, the 3D structures of the relevant sequences were predicted using the AF2 program, thus a total of eight models were created. After obtaining structure predictions, it was investigated whether these models could be selected as a candidate initial design for CoA biosensor construction. According to the results, the chimeric structure constructed

by inserting the domains at residue 41 without a linker stands out as the most promising model. The positional deviations observed in other models, especially in the active site residues and gate post residues, led to raising some question marks for proposing them as an initial design. For the chimeric structures constructed by fusing cpGPF into the sensing domain from the residue 94 insertion site, the most promising models for initial design were the ones obtained with GGS added from the first linker region and also with GGS added from the second linker region. In these models, the 3D structures of the domains were well preserved, and the orientations of active site residues as well as gate post residues were found promising. Thus, three different models out of eight models might be suggested as an initial design for the construction of single cpGFP-based CoA biosensors.

5. CONCLUSIONS AND FUTURE WORK

Cell environments comprise many small molecules involved in metabolic processes along with proteins to carry out their routine work. Since the intracellular concentrations of these molecules vary in different cellular conditions, their precise measurements are paramount in understanding the role of molecules in cells and even in early diagnosis and treatment. From this point of view, a great number of various biosensor models have been developed over the years in order to track these molecules in situ and to quantify the change in their amount as a measurable signal. Among different biosensor models and strategies, protein-based GEFBs have come into prominence as they exploit from superior properties of FPs, such as their intrinsic fluorescence capability. A GEFB consists of two main elements: a sensing domain where the relevant analytes bind, and the fluorescent domain where this binding event can be detected as a fluorescence change. The general working principle of these biosensors is based on the fact that the binding of relevant analyte changes protein conformation in such a way that this can further trigger a change in the fluorescence domain(s), and so that, this binding event can be detected by fluorescence spectroscopy or microscopy. Concerning the construction of GEFBs in practice, their designs are heavily based on trial-and-error process. However, in theory, it is attainable to rationalize these design steps using computational methods and to make effective interventions to these steps at the molecular level.

With this motivation, in this thesis, a computational methodology with a holistic approach including the whole protein structure and dynamics was developed and exploited to have an initial design idea for a single GFP-based CoA biosensor. The proposed method utilizes the following steps: (1) selecting a representative structure that satisfies the certain criteria for use as CoA-sensing domain, (2) finding other proteins sharing similar active sites with the representative protein and comparing the differences between their binding site residues to make mutations/modifications for more sensitive biosensor design with enhanced ligand-binding affinity, (3) examining the changes in the sensing domain conformation through MD simulations and finding suitable insertion locations for fusion of cpGFP, and last but not least

(4) building possible chimeric proteins by adding the cpGFP sequence to that of the sensing domain from these locations and evaluating how the chimeric proteins tolerate the domain insertion with the help of computational methods. Accordingly, first, all proteins that bind CoA or its structural analogues were searched out from the PDB, and subsequently, 728 protein crystal structures were obtained. Among these structures, the Helicobacter pylori PPAT protein, which met specific requirements for the selection of the sensing domain, was chosen as the representative binding protein for CoA biosensor design. Afterwards, for active site design, several protein samples that share similar binding sites with the chosen PPAT (i.e. 1B6T_COD, 3PXU_COD, 3X1J_ACO, 4RUK_COA, 5TS2_COD, 5YRR_COA, and 5ZZC_COD) were found bu utilizing the PoSSum server. Although the binding site modification of the sensing domain is not discussed in this thesis, the obtained differences between the residues in these regions shed light on a more sensitive design. Thereafter, MD simulations were run to uncover whether PPAT protein exhibits CoA-driven conformational change or not. The trajectory analyses showed that the holo system was more stable than the apo system over the course of simulations. Later on, a comparative RMSF analysis was performed to investigate the local changes of both systems and the results pointed out two distinct regions whose mobilities are much higher in the absence of CoA: region 1 (residues 39-47) and region 2 (residues 94-110). In addition to local changes, it was observed that these two regions also have an impact on the overall motions of the systems. By elaborating on the trajectory analyses and visual inspections, the structural and conformational alterations observed in regions 1 and 2 were thought to be a sign that the PPAT protein could undergo a large conformational change in the presence of ligand. With this motivation, two different sites on the sensing domain, namely after residue 41 and residue 94, were identified for cpGFP insertion. After determining these sites, the cpGFP sequence was inserted into the sensing domain from these locations. While preparing these structures, GGS flexible linkers were also used to minimize steric clashes between two domains as well as to facilitate the fold of both domains. In this way, a total of eight chimeric protein sequences were prepared and their corresponding structures were predicted with the aid of AF2. In order to find the possible initial structure(s) among the obtained chimeric structures for CoA biosensor design, the folds of both domains of chimeric proteins, the orientations of the active site residues, and also the positions of the gate post residues were scrutinized. Eventually, three out of eight chimeric structures were proposed as possible starting structures for the CoA biosensor construction. In addition to presenting CoA biosensor models, this study also paves the way for designing computational biosensors for any target analyte.

As future studies, the dynamics of chimeric proteins, which are promising for a single cpGFP-based CoA biosensor design, will be scrutinized via MD simulations. In particular, the hydrogen network around chromophore and the chromophore planarity of both apo and holo chimeric systems will be analyzed to elucidate changes in the fluoresce efficiency. Besides, given the variations in the active site residues among the proteins suggested by the PoSSum results, point mutations will be introduced to the PPAT binding site and their effects on the ligand-binding constant will be investigated. Furthermore, regarding the sensing domain, it will be investigated whether the conformational change of the selected protein could be as large as in the double mutant form. To that end, MD simulations can be extended or enhanced molecular simulation techniques can be utilized. In case the conformational change in the protein could be proven to be as much as in the mutant, this protein might be promising for the FRET-based CoA biosensor strategy.

BIBLIOGRAPHY

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, 19–25.

Ahmad, M., Anjum, N. A., Asif, A., & Ahmad, A. (2020). Real-time monitoring of glutathione in living cells using genetically encoded fret-based ratiometric nanosensor. *Scientific reports*, *10*(1), 1–9.

Aldeghi, M., Gapsys, V., & de Groot, B. L. (2018). Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS central science*, *4*(12), 1708–1718.

Apol, E., Apostolov, R., Berendsen, H., Van Buuren, A., Bjelkmar, P., Van Drunen, R., Feenstra, A., Groenhof, G., Kasson, P., Larsson, P., et al. (2010). Gromacs user manual version 4.5. 4. *Royal Institute of Technology and Uppsala University, Stockholm*.

Bakan, A., Meireles, L. M., & Bahar, I. (2011). Prody: protein dynamics inferred from theory and experiments. *Bioinformatics*, *27*(11), 1575–1577.

Barelier, S., Sterling, T., O'Meara, M. J., & Shoichet, B. K. (2015). The recognition of identical ligands by unrelated proteins. *ACS chemical biology*, *10*(12), 2772–2784.

Belousov, V. V., Fradkov, A. F., Lukyanov, K. A., Staroverov, D. B., Shakhbazov, K. S., Terskikh, A. V., & Lukyanov, S. (2006). Genetically encoded fluorescent indicator for intracellular hydrogen peroxide. *Nature methods*, *3*(4), 281–286.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235–242.

Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., & MacKerell Jr, A. D. (2012). Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation*, *8*(9), 3257–3273.

Bhat, A. S., Schaeffer, R. D., Kinch, L., Medvedev, K. E., & Grishin, N. V. (2020). Recent advances suggest increased influence of selective pressure in allostery. *Current opinion in structural biology*, *62*, 183–188.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., et al. (2021). Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, *49*(D1), D437–D451.

Campbell, R. E. (2009). Fluorescent-protein-based biosensors: modulation of energy transfer as a design principle.

Consortium, U. (2021). Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, *49*(D1), D480–D489.

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, *98*(12), 10089–10092.

Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crécy-Lagard, V., & Osterman, A. (2002). Complete reconstitution of the human coenzyme a biosynthetic pathway via comparative genomics. *Journal of Biological Chemistry*, *277*(24), 21431–21439.

Ding, H., Yuan, G., Peng, L., Zhou, L., & Lin, Q. (2020). Tp-fret-based fluorescent sensor for ratiometric detection of formaldehyde in real food samples, living cells, tissues, and zebrafish. *Journal of agricultural and food chemistry*, *68*(11), 3670–3677.

Dou, J., Doyle, L., Jr Greisen, P., Schena, A., Park, H., Johnsson, K., Stoddard, B. L., & Baker, D. (2017). Sampling and energy evaluation challenges in ligand binding protein design. *Protein Science*, *26*(12), 2426–2437.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh ewald method. *The Journal of chemical physics*, *103*(19), 8577–8593.

Fabian, H. & Naumann, D. (2012). *Millisecond-to-Minute Protein Folding/Misfolding Events Monitored by FTIR Spectroscopy*, (pp. 53–89). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fritz, R. D., Letzelter, M., Reimann, A., Martin, K., Fusco, L., Ritsma, L., Ponsioen, B., Fluri, E., Schulte-Merker, S., van Rheenen, J., et al. (2013). A versatile toolkit to produce sensitive fret biosensors to visualize signaling in time and space. *Science signaling*, *6*(285), rs12–rs12.

Germond, A., Fujita, H., Ichimura, T., & Watanabe, T. M. (2016). Design and development of genetically encoded fluorescent sensors to monitor intracellular chemical and physical parameters. *Biophysical reviews*, *8*(2), 121–138.

Giepmans, B. N., Adams, S. R., Ellisman, M. H., & Tsien, R. Y. (2006). The fluorescent toolbox for assessing protein location and function. *science*, *312*(5771), 217–224.

Govindaraj, R. G. & Brylinski, M. (2018). Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC bioinformatics*, *19*(1), 1–17.

Gutiérrez, I. S., Lin, F.-Y., Vanommeslaeghe, K., Lemkul, J. A., Armacost, K. A., Brooks III, C. L., & MacKerell Jr, A. D. (2016). Parametrization of halogen bonds in the charmm general force field: Improved treatment of ligand–protein interactions. *Bioorganic & medicinal chemistry*, *24*(20), 4812–4825.

Hanson, G. T., Aggeler, R., Oglesbee, D., Cannon, M., Capaldi, R. A., Tsien, R. Y., & Remington, S. J. (2004). Investigating mitochondrial redox potential with redox-sensitive green fluorescent protein indicators. *Journal of Biological Chemistry*, *279*(13), 13044–13053.

Humphrey, W., Dalke, A., & Schulten, K. (1996). Vmd: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33–38.

Ito, J.-I., Tabei, Y., Shimizu, K., Tomii, K., & Tsuda, K. (2012). Pdb-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins: Structure, Function, and Bioinformatics*, *80*(3), 747–763.

Ito, J.-I., Tabei, Y., Shimizu, K., Tsuda, K., & Tomii, K. (2012). Possum: a database of similar protein–ligand binding and putative pockets. *Nucleic acids research*, *40*(D1), D541–D548.

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). Charmm-gui: a web-based graphical user interface for charmm. *Journal of computational chemistry*, *29*(11), 1859–

1865.

Jumper, J. M., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D. A., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*, 583 – 589.

Kim, S., Lee, J., Jo, S., Brooks III, C. L., Lee, H. S., & Im, W. (2017). Charmm-gui ligand reader and modeler for charmm force field generation of small molecules.

Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S., & Thornton, J. M. (2018). Pdbsum: Structural summaries of pdb entries. *Protein science*, *27*(1), 129–134.

Laskowski, R. A. & Swindells, M. B. (2011). Ligplot+: multiple ligand–protein interaction diagrams for drug discovery.

Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., & Lopez, R. (2022). Search and sequence analysis tools services from embl-ebi in 2022. *Nucleic acids research*, gkac240.

Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of computer-aided molecular design*, *27*(3), 221–234.

Martyna, G. J., Tobias, D. J., & Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *The Journal of chemical physics*, *101*(5), 4177–4189.

Miesenböck, G., De Angelis, D. A., & Rothman, J. E. (1998). Visualizing secretion and synaptic transmission with ph-sensitive green fluorescent proteins. *Nature*, *394*(6689), 192–195.

Mizoue, L. S. & Chazin, W. J. (2002). Engineering and design of ligand-induced conformational change in proteins. *Current opinion in structural biology*, *12*(4), 459–463.

Mizuno, T., Murao, K., Tanabe, Y., Oda, M., & Tanaka, T. (2007). Metal-ion-dependent gfp emission in vivo by combining a circularly permutated green fluorescent protein with an engineered metal-ion-binding coiled-coil. *Journal of the American Chemical Society*, *129*(37), 11378–11383.

Mondal, D., Florian, J., & Warshel, A. (2019). Exploring the effectiveness of binding free energy calculations. *The Journal of Physical Chemistry B*, *123*(42), 8910–8915.

Nagai, T., Sawano, A., Park, E. S., & Miyawaki, A. (2001). Circularly permuted green fluorescent proteins engineered to sense ca2+. *Proceedings of the National Academy of Sciences*, *98*(6), 3197–3202.

Nasu, Y., Shen, Y., Kramer, L., & Campbell, R. E. (2021). Structure-and mechanism-guided design of single fluorescent protein-based biosensors. *Nature chemical biology*, *17*(5), 509–518.

Nausch, L. W., Ledoux, J., Bonev, A. D., Nelson, M. T., & Dostmann, W. R. (2008). Differential patterning of cgmp in vascular smooth muscle cells revealed by single gfp-linked biosensors. *Proceedings of the National Academy of Sciences*, *105*(1), 365–370.

Nifosí, R., Amat, P., & Tozzini, V. (2007). Variation of spectral, structural, and vibrational properties within the intrinsically fluorescent proteins family: a density functional study. *Journal of computational chemistry*, *28*(14), 2366–2377.

Okumoto, S., Jones, A., & Frommer, W. B. (2012). Quantitative imaging with fluorescent biosensors. *Annual review of plant biology*, *63*, 663–706.

Ovechkina, V. S., Zakian, S. M., Medvedev, S. P., & Valetdinova, K. R. (2021). Genetically encoded fluorescent biosensors for biomedical applications. *Biomedicines*, *9*(11), 1528.

Pakhomov, A. A. & Martynov, V. I. (2008). Gfp family: Structural insights into spectral tuning. *Chemistry & Biology*, *15*(8), 755–764.

Patnaik, S., Trohalaki, S., & Pachter, R. (2004). Molecular modeling of green fluorescent protein: Structural effects of chromophore deprotonation. *Biopolymers*, *75*.

Patriarchi, T., Cho, J. R., Merten, K., Howe, M. W., Marley, A., Xiong, W.-H., Folk, R. W., Broussard, G. J., Liang, R., Jang, M. J., et al. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science*, *360*(6396), eaat4422.

Phillips, J. C., Hardy, D. J., Maia, J. D., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., et al. (2020). Scalable molecular dynamics on cpu and gpu architectures with namd. *The Journal of chemical physics*, *153*(4), 044130.

Rao, J., Dragulescu-Andrasi, A., & Yao, H. (2007). Fluorescence imaging in vivo: recent advances. *Current opinion in biotechnology*, *18*(1), 17–25.

Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., Dahlgren, M. K., Mondal, S., Chen, W., Wang, L., et al. (2019). Opls3e: Extending force field coverage for drug-like small molecules. *Journal of chemical theory and computation*, *15*(3), 1863–1874.

Sanford, L. & Palmer, A. (2017). Recent advances in development of genetically encoded fluorescent sensors. *Methods in enzymology*, *589*, 1–49.

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.

Shen, Y., Lai, T., & Campbell, R. E. (2015). Red fluorescent proteins (rfps) and rfp-based biosensors for neuronal imaging applications. *Neurophotonics*, *2*(3), 031203.

Stone, J. E. et al. (1998). An efficient library for parallel ray tracing and animation.

Tabei, Y., Uno, T., Sugiyama, M., & Tsuda, K. (2010). Single versus multiple sorting in all pairs similarity search. In *Proceedings of 2nd Asian Conference on Machine Learning*, (pp. 145–160). JMLR Workshop and Conference Proceedings.

Tamura, T. & Hamachi, I. (2014). Recent progress in design of protein-based fluorescent biosensors and their cellular applications. *ACS chemical biology*, *9*(12), 2708–2717.

Tantama, M., Hung, Y. P., & Yellen, G. (2012). Optogenetic reporters: Fluorescent protein-based genetically encoded indicators of signaling and metabolism in the brain. *Progress in brain research*, *196*, 235–263.

Tsien, R. Y. (1998). The green fluorescent protein. *Annual review of biochemistry*, *67*(1), 509–544.

Vaissier Welborn, V. & Head-Gordon, T. (2018). Computational design of synthetic

enzymes. *Chemical reviews*, *119*(11), 6613–6630.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., et al. (2010). Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, *31*(4), 671–690.

Vanommeslaeghe, K. & MacKerell Jr, A. D. (2012). Automation of the charmm general force field (cgenff) i: bond perception and atom typing. *Journal of chemical information and modeling*, *52*(12), 3144–3154.

Wang, H., Nakata, E., & Hamachi, I. (2009). Recent progress in strategies for the creation of protein-based fluorescent biosensors. *ChemBioChem*, *10*(16), 2560–2577.

Wizard, P. P., Epik, I., Prime, L., & Glide, S. (2018). Desmond molecular dynamics system. 2018. *DE Shaw Research, New York, NY. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY*, *2*, 2018–4.

Xu, H., Zhu, C., Chen, Y., Bai, Y., Han, Z., Yao, S., Jiao, Y., Yuan, H., He, W., & Guo, Z. (2020). A fret-based fluorescent zn 2+ sensor: 3d ratiometric imaging, flow cytometric tracking and cisplatin-induced zn 2+ fluctuation monitoring. *Chemical science*, *11*(40), 11037–11041.

Yu, W., He, X., Vanommeslaeghe, K., & MacKerell Jr, A. D. (2012). Extension of the charmm general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry*, *33*(31), 2451–2468.

Zhang, J., Wang, Z., Su, T., Sun, H., Zhu, Y., Qi, Q., & Wang, Q. (2020). Tuning the binding affinity of heme-responsive biosensor for precise and dynamic pathway regulation. *Iscience*, *23*(5), 101067.

Zhao, Y., Zhang, Z., Zou, Y., & Yang, Y. (2018). Visualization of nicotine adenine dinucleotide redox homeostasis with genetically encoded fluorescent sensors. *Antioxidants & Redox Signaling*, *28*(3), 213–229.

APPENDIX A

Selecting a Representative Sensing Domain for cpFP-based CoA Biosensor

Table A.1 Unique protein IDs that binds CoA and/or its structural analogues.

HET Code	PDB ID(s)	Total number of proteins
COA	5HMN 5F48 5HT0 6WQC 6NDS 4HZD 6ZZM 3HO8 5EO2 4M20 5EGJ 3WR7 4ZRB 4XL4 5W40 3X1M 5SZU 7BDW 3ICR 3LNB 3CGD 3FBU 6N2O 4QJL 6IOI 5BZ4 3SXN 6ES9 4IEN 5SUV 3NT6 5GXD 3MQH 5DBV 3UBM 5YRR 6QWU 7Q3A 2VFC 5YH7 4EU4 6YBP 6SK1 6K3C 6BE0 6B2M 3R5C 6HE0 4UBT 5AHS 3Q9N 3Q9U 3RT9 3RTG 3VBI 3VBK 3VBM 3VBN 3VBP 4MFQ 6B0U 6CT5 6P7K 7ED0 7ED1 7S3U 7S43 6CYY 1P5R 5XUK 3OTW 4MRT 3NFD 2NYG 1YVK 6ADD 4QJK 1A59 2PR1 4B3I 6J0P 5TVA 6YUS 4EA7 2JBZ 1DLV 1KQA 3CV2 1H16 1ESM 1H1T 1EBL 1CQJ 5KL9 1HV9 4QVH 1CQJ 1EAB 2REQ 7K0A 1R31 4O9C 5T7E 5XXR 1TIQ 5XUH 1XA4 2K5T 1QR0 2JIB 1SST 1KHR 5G1F 2TDT 1Q6Y 3PVY 6YSW 3RQ5 6RCX 1F7L 3R1K 1P0H 1N8W 3ST8 2QX1 3Q0G 1RJN 3LCJ 2ZSD 6ARB 4JAP 3HQJ 1M4D 1Q4S 6PF1 3S6G 2VHE 4MZU 5HWP 5JBX 3P3I 7L7Z 5JFM 5CUO 2QF7 6CIQ 7PYT 4L9Z 2OZG 6BC3 3R9E 1N71 5VJ1 6MB6 5US1 1BO4 2EIS 7C4G 3B8G 3V4E 5JPH 6MGG 2CYE 2UX9 1IXE 1WLV 5BYU 4BQN 6WN0 2BUE 3FSB 5KLQ 4MFP 6HXQ 6HXP 7CZ3 7CW5 1Y7U 6L3P 6PCD 3PZC 2HQY 4R1L 4R4U 3LD2 3SQZ 3LBE 4HZO 2PRB 6ZZK 2YIZ 2AHV 3U9E 3L92 2CNT 1PG3 1S7N 2WLE 5KF1 4NV7 6HXJ 1YSL 4KUB 7BOR 6EDD 1YRE 3OWC 3U9S 7BCZ 5VJ1 5VJ1 3NWZ 1VPM 4R87 4X0O 3QMN 4NHD 6XBT 1S3Z 6R1E 3QDQ 3S6F 3RT9 4N8I 5KTC 5YO9	226
COZ	2C6X 2G2Z 5F38	3
CAO	1EAC 1EAD 1T3Z 2JDC 6ZZJ 7S3W 7S44	7
COS	4L1F 5OL2	2
FYN	2JI8 7PT4	2
30N	4QC6	1
ACO	1B87 1DM3 1GHE 1HM8 1J4J 1KK4 1KRR 1M3Z 1MR9 1OZP 1P7T 1PT5 1V0C 1WDK 2A81 2C27 2CNS 2CY2 2FIW 2FT0 2GD6 2GE3 2H5M 2I79 2JDD 2OI5 2Q29 2R8V 2R98 2REF 2VQY 2VSS 2WDO 2WLF 2WPW 2XTA 2ZPA 3BLI 3BSY 3EXN 3FS8 3IGJ 3IJW 3IL4 3KVU 3KZL 3MGD 3MQG 3N0M 3NZ2 3PGP 3PP9 3PW8 3R95 3RTA 3RYO 3SLB 3SMA 3SPT 3X1J 3ZJ0 4AVA 4CRY 4HUR 4ISX 4JVT 4JWP 4JXR 4M99 4MY0 4QVT 4R57 4RI1 4UBV 4XPL 4ZBG 5DWN 5FVJ 5IB0 5KF1 5LS7 5T7D 5US1 5W3X 5XUN 5YGE 6AJN 6AO7 6AXE 6BC4 6C32 6EDZ 6G96 6GE9 6GTP 6IOX 6IUF 6MN0 6NZY 6RFT 6U9C 6VTA 6WQB 6YCA 6ZNG 7AK7 7AK8 7B3A 7C4E 7JM1 7K09 7KPS 7KR9 7L7Y 7Q3A 7S45 7TXQ	117

AC8	4MY0	1
1VU	1XNY 4L80 4L9Y 4MZQ 5JFM	5
CMC	4JAE 6VP9	2
SOP	2WLG	1
YAS	6NA4	1
CO6	5CJT 6MFD	2
A1S	5W8A	1
BCO	3Q0G 4XC7 5EGL	1
OXK	2JI6	1
COO	3GF3 5I0K 6JQO	3
3CP	6REQ	1
3KK	4R3U 7PT1 7PT2 7PT3	4
52O	5CJW	1
IVC	5K7H 5W8C	2
CAA	1BUC 1M1O 1Q51 1TXT 1XPK 2GD2 2UZF 3Q0J 3VZS 4FN8 4KUH 4N5M 4NBU 4OMR 4PZE 5HWQ	16
2CP	1EF9 7REQ	2
MLC	1HNJ 2F3X 2Q78 3NYR 4A0Z 5F49	6
3HC	3PVT 3VBJ 4FNB 4R3U	4
1HE	4IZD	1
MC4	2YIM	1
KFV	6N92 6WFI 6X7L 6XBR	4
HXC	1WN3 3V1U 5INF 5T06	4
2KQ	4NNC	1
V0V	6WFH	1
SCA	1KGT 2BWO 2VZZ 3FSY 4REQ 5E3Q 6CYJ	7
MCA	1ON3 3NYQ 4REQ 6WF7	4
BYC	3PM5 3PVR 4Z3Y	3
IRC	3O3N	1
GRA	3GMA 3MPI	2
4KX	4Z3W	1
4CA	1LO8 1Q4U	2
3H9	4FND	1
CO8	4A0S 4Q36 5V0P 5YOA 6IIX 6JQN 7C1L	7
T3D	6SLB	1
CQM	6AQ4	1
BCA	1JXZ 1LO9 1NZY	3
0FQ	4K4A 4K4C 4QD8	3

FAQ	3PW1 4IIT	2
COW	3H77	1
2NE	4I4Z 4QII	2
YXR	6N95 6X7O	2
01A	3CW9	1
LHQ	6SLA	1
4CO	1LO7 1Q4T	2
CCQ	1XVV	1
1CZ	4I56	1
HMG	1QAX 1XPK 4I6A 5HWO 5WPK	5
MFK	4MFK 4MFZ 5T07 6SDA	4
YE2	5KAJ	1
HFQ	4K49 4K4D	2
WCA	5CYV 6C28	2
J5H	6QKR	1
UOQ	4K4B	1
8Z2	5NJI	1
1C4	4I49	1
DCC	1U6S 1VI0 3ANG 4KU5	4
1HA	4I42 4I52 4QIJ	3
SFC	2GCE	1
RFC	2GCE	1
F8G	6CO9	1
MYA	4KU2	1
MDE	1PS9	1
MRR	2GCI	1
MRS	2GD0	1
HD6	5F34 7Q1U	2
UCA	4W97	1
PKZ	5DV5	1
3VV	4PDK	1
ST9	3WHC 6JZZ 6KSA 6KSE	4
5F9	5DTW 6O6N	2
93P	5VD6	1
4BN	4WNB	1
5TW	5CXI	1
93M	5VDB	1
OXT	2JI7 7PT4	2

5JB	5AQC	1
NH9	7TXS	1
JBT	4EA9	1
XQD	7L82	1
5NG	5EGL 6C37	2
0T1	5DW5	1
UT7	6X7R	1
CMX	2H12 6XBQ	2

Total PDB ID: 534

APPENDIX B

Structural and Dynamical Investigations of CoA-Binding Domain

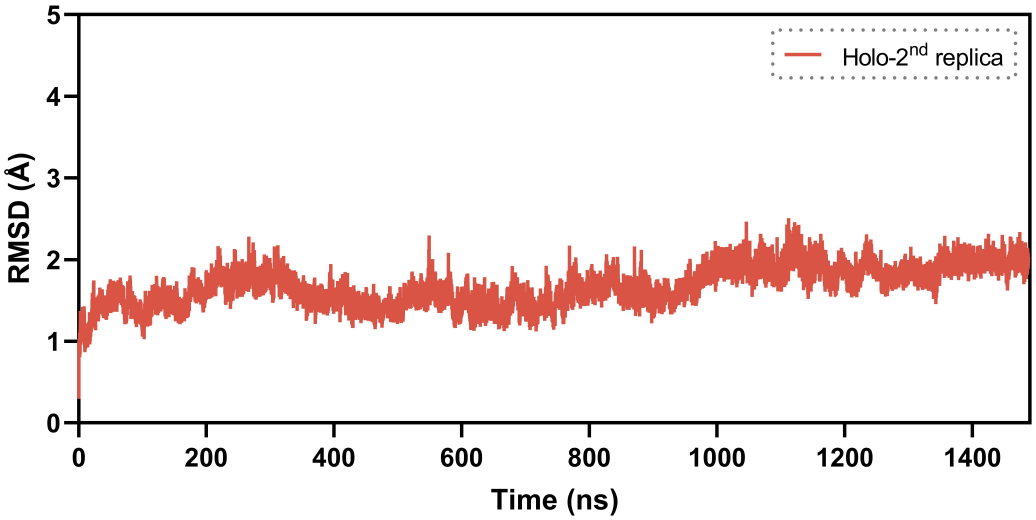

Figure B.1 Backbone RMSD plot for extended second replicate of holo system.