Molecular determinants of the health effects of coffee compounds

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Nicolas Pierre Friedrich Müller
aus Köln

Köln, Oktober 2023

from the Institute for Advanced Simulation, Computational Biomedicine (IAS-5/INM-9)
at the Forschungszentrum Juelich GmbH

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor: Dr. Mercedes Alfonso-Prieto
Co-supervisor: Prof. Dr. Patricia Hidalgo

Date of the oral examination: 22.05.2024

Publications

The doctoral thesis is containing parts of an article which has been published. In accordance with §6(3) of the doctoral regulations of the Faculty of Mathematics and Natural Science, those parts are clearly marked and highlighted within the doctoral thesis. First of all, re-used parts are cited within quotation marks and shown in italics. Furthermore, citations, tables or figures were adapted in order to follow the order of the document and the guidelines of the Faculty of Mathematics and Natural Science.

Published article:

[1] Mueller NPF, Carloni P. and Alfonso-Prieto M. “Molecular determinants of acrylamide neurotoxicity through covalent docking”. In: *Front. Pharmacol.* **14**:1125871 (2023). https://doi.org/10.3389/fphar.2023.1125871

Contribution according to §6(3) of the doctoral regulations:

Nicolas Müller and Mercedes Alfonso-Prieto performed research and analysed data. Paolo Carloni and Mercedes Alfonso-Prieto designed research. Nicolas Müller and Mercedes Alfonso-Prieto drafted the initial version of the manuscript. All authors edited and approved the manuscript.

Other research publications where I collaborated during my doctoral period and that were not included in the thesis are the following:

[2] Nin-Hill A, Mueller NPF, Molteni C, Rovira C and Alfonso-Prieto M. “Photopharmacology of Ion Channels through the Light of the Computational Microscope”. In: *Int J Mol Sci.* **21**:12072 (2021). https://doi.org/10.3390/ijms222112072

Eidesstattliche Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist:

1. Diese Arbeit wurde vollständig oder größtenteils in der Phase als Doktorand diese Fakultät und Universität angefertigt;

2. Sofern irgendein Bestandteil dieser Dissertation zuvor für einen akademischen Abschluss oder eine andere Qualifikation an dieser oder einer anderen Institution verwendet wurde, wurde dies klar angezeigt;

3. Wenn immer eigene oder Veröffentlichungen Dritter herangezogen wurden, wurden diese klar benannt;

4. Wenn aus anderen eigenen oder Veröffentlichungen Dritter zitiert wurde, wurde stets die Quelle hierfür angegeben. Dieser Dissertation ist vollständig meine eigene Arbeit, mit der Ausnahme solcher Zitate;

5. Alle wesentlichen Quellen von Unterstützungen wurden benannt;

6. Wenn immer ein Teil dieser Dissertation und der Zusammenarbeit mit anderen basiert, wurde von mir klar gekennzeichnet, was von anderen und was von mir selbst erarbeitet wurde;

7. Ein Teil oder Teile dieser Arbeit zuvor veröffentlicht wie unter Publications aufgeführt.

Düsseldorf, den 4. Oktober, 2023

_____

Nicolas Müller

Abstract

As awareness for health, well-being and a sustainable nutrition has gained more attention, also more people got interested in beneficial or harmful consequences of their diet. In particular, small molecules which are present in foods and beverages can have contrasting effects on human health.
Coffee is one of the most consumed beverages around the world and consists of a diverse mixture of compounds. Determining how those compounds have a positive or negative impact on human health at a molecular level is quite challenging due to their diversity and the wide range of human proteins that they can target. In this thesis, I have used computational methods to address this question for acrylamide and the group of chlorogenic acids.
Acrylamide (ACR) is a small organic compound formed during food processing at high temperatures, for instance, during backing, frying or roasting. Indeed, coffee contains ACR as a result of the roasting process. In addition, ACR is used in different industries, such as water waste treatment and manufacture of paper, fabrics, dyes and cosmetics. Cumulative exposure to acrylamide, either from diet or at the workplace, may result in neurotoxicity.
At the molecular level, ACR is an electrophile which forms covalent adducts with proteins via a Michael addition reaction with nucleophilic cysteine residues. Due to the fact that synaptic proteins are cysteine-rich, they can be particularly affected by ACR exposure, thus explaining the neurological symptoms associated ACR exposure. In order to better understand which cysteine residues are more likely to undergo ACR modification and the impact of covalent adduct formation on protein function, in this thesis I investigated the molecular determinants of ACR reactivity through covalent docking.
My results indicate that acrylamide binding to cysteine is favored in the presence of nearby positively charged amino acids, such as lysines and arginines. For proteins with more than one reactive Cys, docking scores are able to discriminate between the primary ACR modification site and secondary sites modified only at high ACR concentrations. Based on this study, covalent docking is a promising computational tool to predict other potential protein targets mediating acrylamide neurotoxicity.
In contrast to ACR, coffee also contains compounds that have beneficial effects on human health. Among other small molecules present in coffee, chlorogenic acids (CGAs) constitute a group of phenolic molecules considered as nutraceuticals due to their extra health benefits in addition to their basic nutritional value. Such benefits include, for instance, antioxidant and anti-inflammatory properties, modulation of lipid and glucose metabolism, prevention of cardiovascular diseases and neuroprotective effects.
CGAs are a quite diverse group of compounds and their bioavailability depends on coffee strain, growing conditions and post-processing steps (e.g. roasting). In addition, digestion and processing in the human body can increase the chemical diversity of such compounds. From the structural point of view, CGAs are esters of quinic- and hydroxycinnamic acid (HCA). Recently, cinnamic acid, a phenolic precursor of HCAs, showed neuroprotective effects in mouse models (Parkinson's and Alzheimer's disease models) mediated by peroxisome proliferator-activated receptor alpha (PPAR α). This evidence suggested that related compounds, such as HCAs and CGAs, could also act as PPARα activators and explain their proposed neuroprotective effects.
In this thesis, I investigated the molecular determinants of PPARα binding to CGA compounds by means of molecular docking and molecular dynamics. The results

indicate that cinnamic acid can occupy multiple binding pockets of PPARα. Moreover, the predicted binding modes of CGA compounds give insights into their mode of action towards PPARα activation. Nonetheless, further computational and experimental validation is needed to potentially use cinnamic acid, HCAs and CGAs as neuroprotective nutraceuticals.

In summary, I have demonstrated that computational methods, such as docking and molecular dynamics, can give detailed insights into molecular mechanisms through which small molecules present in foods and beverages can have an impact on human health. The results presented in my thesis can pave the way for future computational and experimental studies to further validate and investigate the effects of coffee compounds in human health, as well as other potential nutraceuticals, on human health.

Zusammenfassung

Da das steigende Bewusstsein für Gesundheit, Wohlbefinden und eine nachhaltige Ernährung immer mehr an Bedeutung gewonnen hat, interessieren sich auch immer mehr Menschen für die positiven sowie negativen Folgen ihrer Ernährung.
Wirkstoffe in Lebensmitteln und Getränken können vielfältige Auswirkungen auf die menschliche Gesundheit haben.
Kaffee ist eines der am meisten konsumierten Getränke weltweit und beinhaltet eine Menge verschiedener Wirkstoffe. Die Wirkungsweise dieser Moleküle auf molekularer Ebene, ob negativ oder positiv, sind komplex und benötigen weitere Untersuchungen. In dieser Thesis habe Ich informatische Methoden verwendet, um diese Frage für Acrylamid und der Wirkstoffgruppe der Chlorogensäuren zu beantworten.
Acrylamid (ACR) ist eine organische Verbindung, die bei der Lebensmittelverarbeitung unter hohen Temperaturen entsteht, wie beispielsweise beim Backen, Braten oder Rösten. In der Tat enthält Kaffee auch ACR, welches aufgrund der Röstungsprozesse entsteht. Zudem wird ACR in verschiedenen Industriezweigen verwendet, wie beispielsweise bei der Abwasseraufbereitung, der Herstellung von Papier, in Textilien, in Farbstoffen und in Kosmetikbranche. Eine kumulative Belastung durch Acrylamid, entweder durch die Nahrung oder am Arbeitsplatz kann zu Neurotoxizität führen. Auf molekularer Ebene ist Acrylamid ein Elektrophil, das kovalente Addukte mit Proteinen über eine Michael-Additionsreaktion mit der Aminosäure Cystein bildet. Dadurch, dass Proteine, die eine neurologische Funktion ausüben, oft reich an der Aminosäure Cystein sind, können diese besonders beeinträchtigt werden, sodass Symptome, die durch erhöhten Kontakt mit ACR entstehen, erklärt werden können. Um feststellen zu können unter welchen Bedingungen die Wahrscheinlichkeit am höchsten ist, dass eine ACR-Modifikation stattfindet und welche Folgen diese für die Funktion der Proteine hat, habe Ich in dieser Arbeit die molekularen Determinanten der ACR-Reaktivität durch Molecular Docking untersucht.
Die Ergebnisse deuten darauf hin, dass die Bildung einer kovalenten Bindung zwischen Acrylamid und der Aminosäure Cystein begünstigt wird, wenn positiv geladenen Aminosäuren wie Lysin und Arginin in unmittelbarer Nachbarschaft zu finden sind. Bei Proteinen mit mehr als einem potenziellen Ziel, können Docking Scores zwischen der primären und sekundären Modifikationsstelle unterscheiden, die nur bei höheren ACR Konzentrationen modifiziert wurde. Basierend auf dieser Studie ist kovalentes Docking ein vielversprechendes Werkzeug zur Vorhersage potenzieller Proteinziele, die für die Neurotoxizität von Acrylamide verantwortlich sein können.
Kaffee enthält zudem Wirkstoffe, die sich positiv auf die menschliche Gesundheit auswirken. Unter anderem enthält Kaffee Chlorogensäuren (CGAs), die neben ihren grundsätzlichen Nährwerten auch gesundheitliche Vorteile besitzen und deswegen als Nutrazeutika gelten. Zu diesen Vorteilen gehören beispielsweise antioxidative und entzündungshemmende Eigenschaften, die Modulation des Lipid- und Glukosestoffwechsels, die Vorbeugung von Herz-Kreislauf-Erkrankungen als auch neuroprotektive Wirkungen. CGAs sind eine recht vielfältige Gruppe an Molekülen und ihre Bioverfügbarkeit hängt von der Kaffeesorte, den Wachstumsbedingungen und der Verarbeitung wie zum Beispiel dem Rösten, ab. Darüber hinaus kann die chemische Diversität durch die Verdauung und Verarbeitung im menschlichen Körper steigen. Aus struktureller Sicht sind CGAs Ester der China- und Kaffeesäure (HCA). Kürzlich zeigte die Zimtsäure, ein Vorläufer der Kaffeesäure, neuroprotektive Eigenschaften gegen Parkinson und Alzheimer im Modellorganismus der Maus. Diese Eigenschaften wurden mithilfe des Proteins PPARα vermittelt. Diese Studien legen außerdem nahe,

dass strukturell verwandte Moleküle wie HCAs und CGAs ebenfalls als Agonist von PPARα dienen können und somit die neuroprotektiven Charakteristiken erklärt werden können. In dieser Thesis habe Ich mittels Molecular Docking und Molecular Dynamics die Determinanten von CGAs in Hinsicht auf Peroxisome Proliferator-Activated Receptor alpha (PPARα) Bindung untersucht. Die Ergebnisse deuten darauf hin, dass Zimtsäure mehrere Bindungstaschen besetzen kann, wodurch eine Aktivierung von PPARα ermöglicht wird. Darüber hinaus geben vorhergesagte Bindungsmodi von Chlorogensäuren Einblicke in deren Wirkungsweise bei der Aktivierung von PPARα. Trotz dieser Einblicke sind weitere informatische als auch experimentelle Validierungen von Nöten, um die Zimtsäure, HCAs und CGAs als Nutrazeutika zu verwenden.

Zusammenfassend habe ich gezeigt, dass informatische Methoden wie Molecular Docking und Molecular Dynamics detaillierte Einblicke in molekulare Mechanismen geben können. Die vorgestellten Ergebnisse meiner Thesis können zudem den Weg für künftige Studien ebnen, die die Auswirkungen von Molekülen in Kaffee auf die menschliche Gesundheit untersuchen. Hinzu kommt, dass auf gleiche Weise andere potenzielle Nutrazeutika untersucht werden können.

Acknowledgements

During my research at the Forschungszentrum Jülich I received a lot of support and assistance from different people. First of all, I want to thank my supervisor, Dr. Mercedes Alfonso-Prieto, for giving me support throughout my whole PhD studies. She was providing me useful comments, answered all my stupid questions and taught me Spanish (*Me das la razón como a los locos).*
Furthermore, I would like to thank Prof. Paolo Carloni for both giving me the opportunity to work on such an interesting research topic and the freedom of letting me shape respective projects (*Grazie di tutto*).
In addition, I would like to thank Prof. Dr. Patricia Hidalgo for being the second reviewer of my thesis.
I would like to thank Prof. Dr. Olga Sergeeva for the cooperation and useful comments regarding my research.
I am also thankful for the chance to participate in a variety of courses and workshops which helped me to extend and develop my professional skills.
I would like to thank Dr. Luciano Navarini and the Ernesto Illy Foundation, for suggestions and useful scientific discussions regarding my research projects.
Thank you, Dr. Emiliano Ippoliti, Sabrina Schulte, and Petra Rott for helping me with all the administrative, organizational and computational work. Furthermore, I want to thank all the members of INM-9.
Des Weiteren danke ich meiner Familie und meinen Freuden für die ständige Unterstützung während meiner Promotion, besonders unter den erschwerten Umständen aufgrund der Corona-Pandemie.

CONTENTS

1 Introduction

Acrylamide (CH2=CH-C(O)NH2, PubChemCID 6579) is utilized in variety of industrial processes, encompassing water waste treatment, manufacture of paper, fabrics, dyes or cosmetics [3-5]. Additionally, acrylamide is found in the food industry due to the Maillard reaction between reducing sugars and amino acids [6]. In particular, in foods processed at high temperatures, such as coffee, french fries, and baked/roasted potatoes [7-9]. As soon as its potential harm to the human health was recognized [10, 11], new European Union wide regulations were released [12] in order to mitigate acrylamide formation in food items upon frying, baking or roasting. The cumulative acrylamide exposure, whether from diet or through occupational exposure, may lead to toxicity, especially in the central nervous system. Studies with animals and humans indicate that acrylamide neurotoxicity could mimic (or even contribute to) the symptoms of neurodegenerative disorders like Parkinson's disease [13-16], along with inducing depressive and anxiety-like behavioral effects [17, 18].
The acrylamide (ACR) molecule possesses an α, β-unsaturated carbonyl that acts as an electrophile and is capable of interacting with nucleophilic amino acids (Figure 3). According to the hard and soft acids and bases (HSAB) theory, ACR (here, acting as soft electrophile) prefers to react with soft nucleophiles, such as the thiolate group of deprotonated cysteine residues [19]. The Michael addition reaction between deprotonated Cys residues and acrylamide results in covalent adducts (Figure 3, step 3). In vitro and biochemical studies targeting individual protein targets have identified Cys residues susceptible to ACR modification and thus may influencing protein function. Therefore, in my thesis I aimed at assessing whether covalent docking can serve as a computational tool to characterize ACR reactivity and its (neuro)toxic effects at a molecular level [1].

Chlorogenic acids (CGAs) are esters of quinic- and hydroxycinnamic acid and belong to the group of polyphenols. CGAs are natural compounds present in various plants; however, coffee beans are one of the most abundant and well-known sources. Due to their potential health effects, such as antioxidant and anti-inflammatory properties, modulation of lipid and glucose metabolism, prevention of cardiovascular diseases and neuroprotective effects, CGAs have gained public intention as potential nutraceuticals.
Chlorogenic acids are a diverse group of compounds and their concentration depends on coffee strain, growing conditions and post-processing steps (e.g. roasting). Therefore, it is challenging to pinpoint which CGA molecules are responsible for the earlier mentioned beneficial effects. Nevertheless, in more recent studies cinnamic acid (CNA) (i.e. a precursor of HCAs) has been proven to mediate neuroprotective effects, in particularly activating the nuclear receptor PPARα. PPARα is expressed in different brain areas and involved in glutamate homeostasis, microglia activation in neuroinflammation and anti-amyloidogenic effects. Moreover, experimental evidence in mouse models mimicking either Alzheimer’s or Parkinson’s disease further support the involvement of PPARα in mediating the neuroprotective effects of cinnamic acid. Furthermore, the experimental data on the parent compound cinnamic acid hints at its derivative compounds, such as HCAs and CGAs, having neuroprotective effects by a similar molecular mechanism.
In this thesis I will investigate whether CGAs can bind to PPARα and promote receptor activation via molecular mechanisms similar for known PPARα agonists as described in the literature. The effect of the aforementioned coffee compounds on their

respective protein targets will be examined using computational approaches, in particular molecular docking and molecular dynamics. Furthermore, comparison with experimental data (when available) will help to evaluate whether such *in silico* methods can contribute and enhance our understanding of the health effects of coffee compounds.

1.1 Thesis Structure

Chapter 2 provides a short biological background describing the systems used in this thesis. In particular, I introduce the basics of protein structure and dynamics, as well as the main molecules subject of this thesis, i.e. ACR, CGAs and PPARα. **Chapter 3** gives a general overview about the theoretical background of the computational methods employed in my PhD work, namely molecular docking and molecular dynamics. **Chapter 4** summarizes the results of the ACR project, published in [1], whereas **Chapter 5** describes the workflow and data obtained for the CGA project. **Chapter 6** provides conclusions and future perspectives on the application of computational methods to understand the impact of compounds present in foods and beverages on human health.

2 Biological background of studied protein-ligand complexes

2.1 Amino Acids

During the last decades many theories evolved around the origin of life. Despite quite diverse explanations exists, one of the fundamental steps in order to evolve higher life was the formation of amino acids. Amino acids are one of the essential building blocks in all kingdoms of life.
There are 21 α-amino acids encoded in the human genome. Each amino acid has a central C_α, which is connecting four chemical groups, namely, a hydrogen atom, an amino group, a carboxyl group and an individual sidechain (Figure 1).
The physicochemical properties of each amino acid are defined based on the specific R group of each residue. Although, having two enantiomers in nature (L and D) the human life is exclusively based on L-isomers of amino acids.

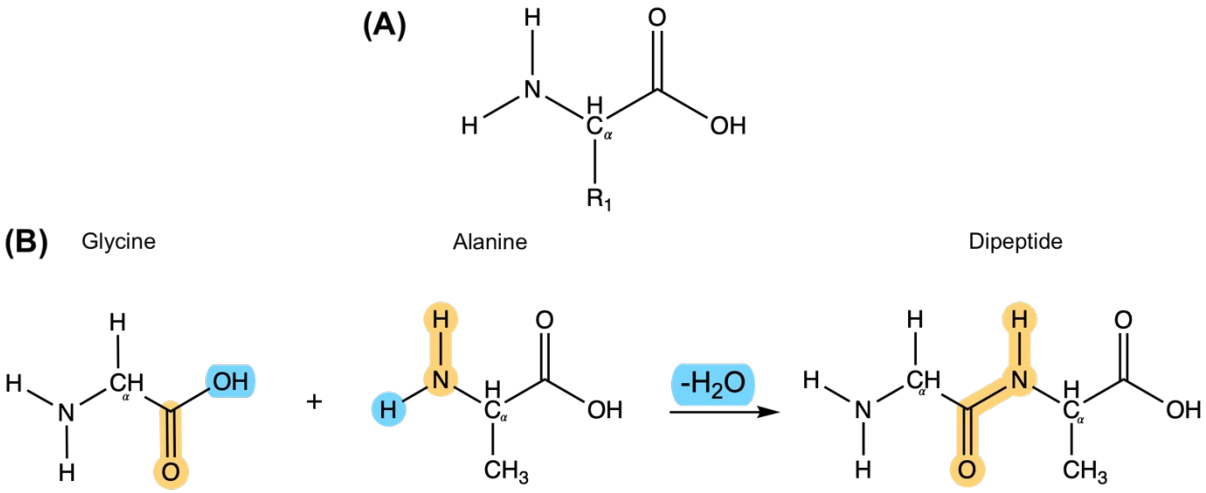


*Figure 1. (**A**) Structure of an amino acid. The side-chain position is indicated with R_1. (**B**) Condensation reaction between two amino acids in order to form a peptide bond (highlighted in yellow).*

2.2 Proteins

One of the most known concepts of biology is the central dogma of molecular biology, which describes the informational flow in Nature. Generally speaking, DNA is transcribed into RNA which is subsequently translated into an amino acid sequence. That amino acid sequence folds into a functional protein.

2.3 Protein Structure and Dynamics

Proteins form forming more complex 3D structures upon amino acid sequence synthesis. Most of the time α-helices and β-sheets are the main secondary structure motifs, which define the three-dimensional shape of proteins. The folding of amino acid sequences into functional protein structures is highly efficient and evolved over millions of years through evolution. Even though amino sequences appear to be random at the first glance, they contain all necessary information about their native

protein structure. This concept is well-known and one of the most extensively studied phenomena of molecular biology [20-23].
The free energy landscape of a protein describes all possible arrangements of a specific amino acid sequence along chosen conformational variables. The folding process takes a pathway through the defined conformational space which results in a native protein structure. That pathway of folding is not engraved in stone - a protein may adapt several different transition states before reaching the global minimum in energy, i.e. the native protein structure.
During the last decades, multiple methodologies have been developed to tackle the above-mentioned issue of protein structure prediction. A remarkable progress was made in recent years by artificial intelligence powered algorithms, which predicted proteins structures to an astonishing accuracy [24, 25]. Accurate prediction of protein structures has given rise to a wide range of application areas and is expected to revolutionized drug development, e.g. development of vaccines. Fast and accurate structure prediction of spike proteins could, for instance, give insights into their mode of action and facilitate, as well as accelerate, therapeutic approaches [26].
As a protein is a dynamic system and undergoes conformational changes in order to perform its biological function, such structural rearrangements are also relevant for pharmacological and biophysical research areas. For instance, in the case of receptors, their activation is a process that often requires conformational changes. Agonists (or more general speaking drug like molecules) are able to bind to their protein target and facilitate such activation mechanism by modifying the free energy landscape. Thus, small molecules can bias the protein conformational state and lower energy barriers through which conformational transitions can be achieved.
Simulating these events can help to understand molecular mechanisms of drug action and facilitate drug development. Therefore, theoretical studies of underlying biophysical phenomena underlying protein function are indispensable.

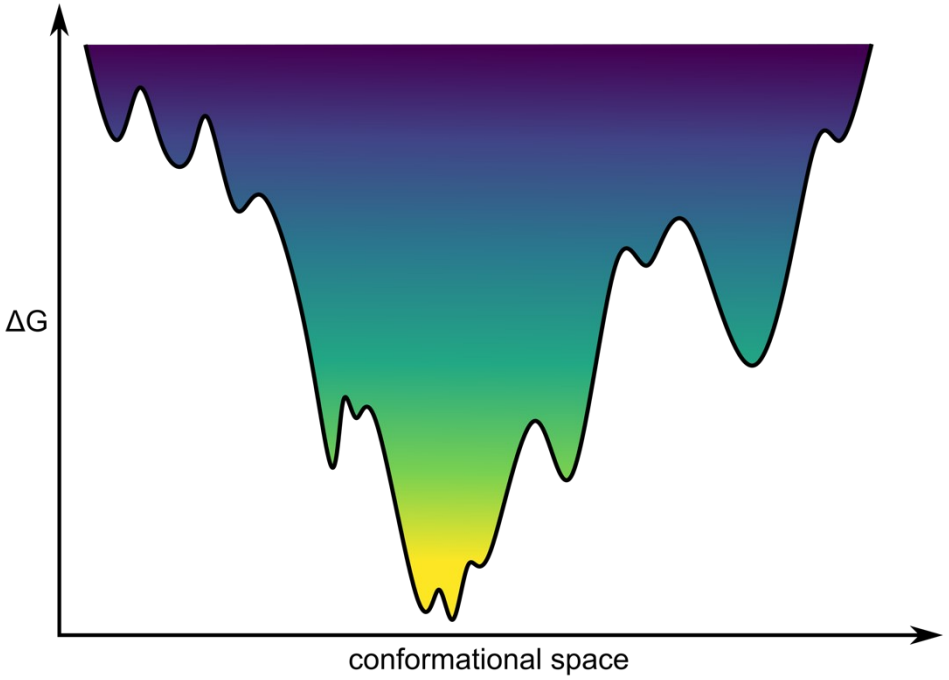


Figure 2. Simplified schematic of a protein folding free energy landscape. The global minimum, corresponding to the protein native structure, is colored in yellow.

2.4 Coffee

Coffee is one of the most consumed beverages worldwide and part of the daily life of millions of people. Therefore, it is not surprising that the coffee business is generating billions in revenue every year. Traditionally, the consumption of coffee was attributed to caffeine. However, during the last decades both the analytical as well as computational methods got more precise and it became apparent that, besides caffeine, multiple other compounds present in coffee have an influence on the human well-being. Together with the growing awareness of leading a healthy lifestyle, coffee has gained even more popularity and companies are adopting a new marketing approach within these characteristics [27, 28].

Coffea arabica (Arabica) and *Coffea canephora* (Robusta) are two of the most prominent coffee species and are cultivated in tropical and subtropical areas, such as Asia, America and Africa. Often coffee is one of the main export products in these areas and processing (roasting, mixing and packaging) of the beans are done in North America and Europe [29].
The final flavor of coffee is not only determined by the respective coffee blend but also by origin, growing conditions and processing procedure. Based on aforementioned circumstances, the amount and composition of coffee compounds can be altered, resulting in the unique flavor of each coffee blend [30].

The most prominent coffee compounds are caffeine followed by a phenolic group of compounds called chlorogenic acids (CGAs) [31]. Despite coffee being a major source of CGA intake, CGAs and their mode of action on the human health is not well understood. Nevertheless, several experimental studies have shown that CGAs have antioxidant, anti-carcinogenic, and anti-inflammatory properties, as well as additional beneficial health effects against type 2 diabetes, obesity, Alzheimer's disease, strokes and blood pressure [32-35].

Besides coffee compounds with beneficial effects on the human health, processing of coffee beans can also result in the generation of unwanted compounds, such as acrylamide (ACR). In particular, chronic exposure to ACR is known to cause toxicity to the human nervous system and reproductive systems, as well as carcinogenicity [5, 10, 11]. More details on ACR and CGAs are given in the following sections (2.5 and 2.6 , respectively)

2.5 Acrylamide

Acrylamide is a small organic compound (CH2=CH-C(O)NH2, PubChem CID 6579), which is present in a variety of foods and beverages. Besides being a by-product of the food industry, ACR is also used in several industrial processes, such as water waste treatment, manufacture of paper, fabrics, dyes or cosmetics.
The first time ACR got increased attention was during the 1980s. At that time, it was noticeable that workers who were exposed to higher ACR concentrations showed neurotoxic symptoms [36]. In particular, Parkinson's disease (PD) like symptoms could be observed after chronic exposure to ACR in the workplace [13-16].
Most people, however, are exposed to much smaller quantities of ACR through their daily diet. In foods or beverages, a so-called Maillard reaction takes place upon roasting, baking or frying, which can produce ACR as a by-product [6].

Based on the earlier mentioned toxic effects of ACR on the human health, new European Union wide regulations entered into force in 2017 [12]. Those rules aimed at keeping acrylamide levels within acceptable boundaries, especially in the food industry.

As shown in Figure 3, ACR possesses an *α, β*-unsaturated carbonyl moiety that can act as an electrophile. Due to the electron-withdrawing nature of the carbonyl group, the adjacent *β*-carbon possesses the lowest electron density, which makes it the most electrophilic site (step 1). According to the hard and soft acids and bases (HSAB) theory, ACR belongs to the group of soft electrophiles. This class of molecules preferably reacts with soft nucleophiles, such as deprotonated cysteine residues. The intrinsic pK_a value of a cysteine side chain is 8.6, so that at physiological pH cysteine residues are mostly present as thiol [37]. Nevertheless, depending on the cysteine location within the three-dimensional protein structure, its microenvironment can shift the intrinsic pK_a value and thus favor the anionic thiolate form (step 2). Water molecules or surrounding residues can act as base catalyst and deprotonate cysteine residues. Michael addition reaction between that negatively charged thiolate and ACR result in the formation of a covalent adduct that can alter protein function (step 3)[38].

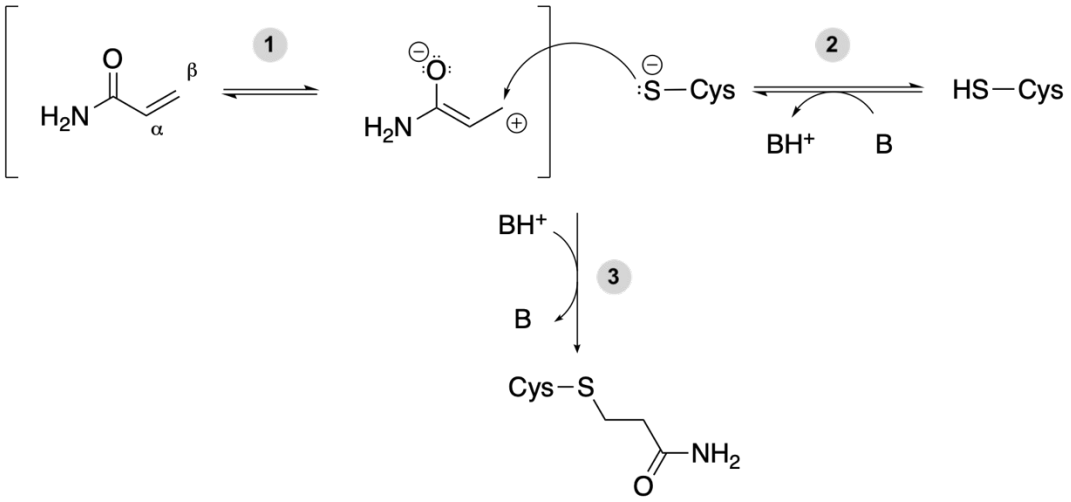
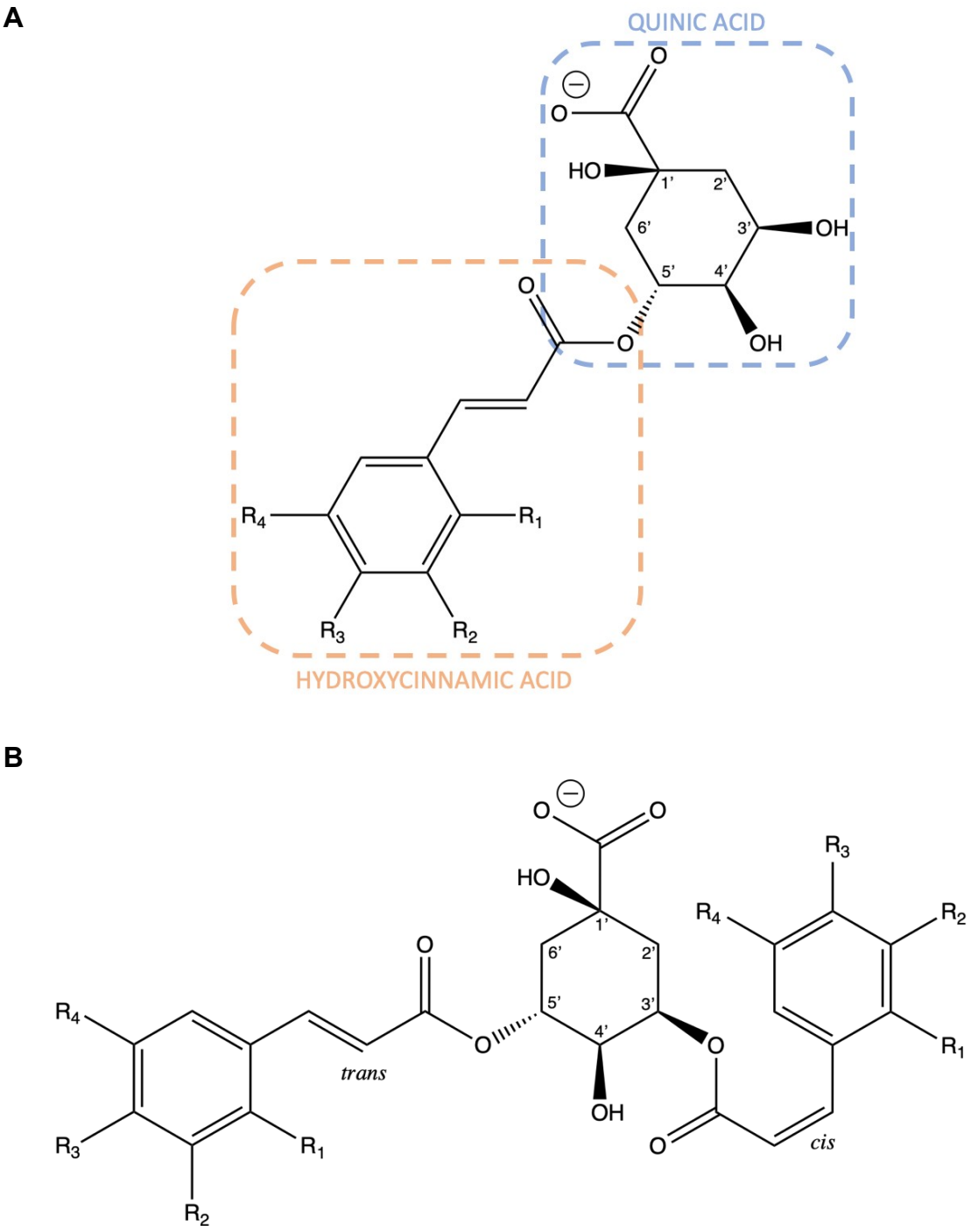


Figure 3. Michael addition reaction between acrylamide and a Cys residue of a protein target. B and BH^+ represent a Brønsted-Lowry acid-base pair, either a protein residue or a water molecule. (1) The electron-withdrawing effect of the carbonyl group makes the *β*-carbon of acrylamide an electrophilic site. (2) The cysteine side chain can act as nucleophile and react with a soft electrophile such as acrylamide. (3) Michael addition reaction between a protein thiolate and acrylamide yields a covalent adduct potentially affecting protein function. Reproduced from [1] under a CC-BY license. © Frontiers.

Besides forming adducts with human proteins, ACR is also metabolized in the human body. The main metabolization pathways of ACR are via epoxidation to glycidamide (GA) or conjugation to glutathione. GA, another small organic molecule, is highly reactive due to the epoxide moiety and, among other things, able to react with proteins as well as DNA. Based on such reactivity, GA is considered to be responsible for the genotoxic effects of ACR. Glutathione (GSH), an antioxidant, is able to scavenge both ACR and GA molecules which results in metabolites belonging to the class of mercapturic acids [39]. These compounds are eliminated through the urinary way and as a consequence cleared from the human body.
Nevertheless, as a result of increasing ACR concentrations, GSH levels are depleted, which is subsequently elevating cellular oxidative stress levels. This depletion of GSH levels is inevitable favoring the formation of ACR adducts with proteins [40].

2.6 Chlorogenic Acids

CGAs are esters of hydroxycinnamic acids and (–)-quinic acid (Figure 4). Among the group of CGAs, the main metabolite is 5-*O*-caffeoylquinic acid, which is a caffeoyl ester at 5’ position of quinic acid (following the IUPAC numbering). However, the chemical toolbox of CGAs is far more diverse. Based on the basic building blocks of CGAs (quinic acid and hydroxycinnamic acids) it is possible to add up to three molecules of hydroxycinnamic acid to different hydroxyl groups of the quinic acid ring via ester bonds. If, for instance, *trans*-caffeic acid is attached to the hydroxyl group at position 3’ to the cyclohexane moiety of quinic acid, the molecule is referred to as 3-*O* -caffeoylquinic acid. If, on the other hand, two molecules of *trans*-caffeic acid are attached to the hydroxyl groups at position 3’ and 5’ of quinic acid, that molecule is called 3,5-dicaffeoylquinic acid. Moreover, it should be mentioned that di- as well as tri-chlorogenic acids can be present as either homo- or heteroesters.



cis-**3**-*trans*-**5**-caffeoylquinic acid

*Figure 4. (**A**) Representative example of a CGA molecule. The hydroxycinnamic acid and the quinic acid fragment are marked in orange and blue, respectively. R-groups of hydroxycinnamic acids are denoted by labels R_1 to R_4. (**B**) A hetero di-CGA is shown with one cis-hydroxycinnamic at position 3' and one trans hydroxycinnamic acid at position 5' is shown.*

Table 1. Examples of hydroxycinnamic acids studied in this thesis and their substituent (R) groups. As shown in Figure 4A, hydroxycinnamic acids can have different substituents at the phenyl ring.

HYDROXYCINNAMIC ACIDS				
R_1	**R_2**	**R_3**	**R_4**	**Name**
H	H	H	H	Cinnamic Acid
OH	H	H	H	*o*-Hydroxycinnamic Acid
H	H	OH	H	*p*-Hydroxycinnamic Acid
H	OH	OH	H	Caffeic Acid
H	OCH_3	OH	H	Ferulic Acid
H	OH	OCH_3	H	Isoferulic Acid
H	OCH_3	OH	OCH_3	Sinapic Acid
H	H	OCH_3	OCH_3	Dimethoxycaffeic Acid
H	OCH_3	OCH_3	OCH_3	Trimethoxycaffeic Acid

In addition, *trans*-caffeic acid can be replaced by other hydroxycinnamic acid molecules (see Table 1), further increasing the diversity of CGAs. In general, most hydroxycinnamic acids compounds are present as *trans* isomers (as the example shown in Figure 4); however, exposure to UV light and other factors like pH value, temperature and time of exposure can shift the equilibrium towards respective *cis* isomers.

As CGA content in coffee beans is affected through different factors, including maturation degree, agricultural cultivation, climate, soil and species, as well as roasting and brewing, also human exposure to CGAs can show large differences. In general, CGAs are degraded during roasting which explains that darker roasted coffee possesses on average lower amounts of CGAs [41].

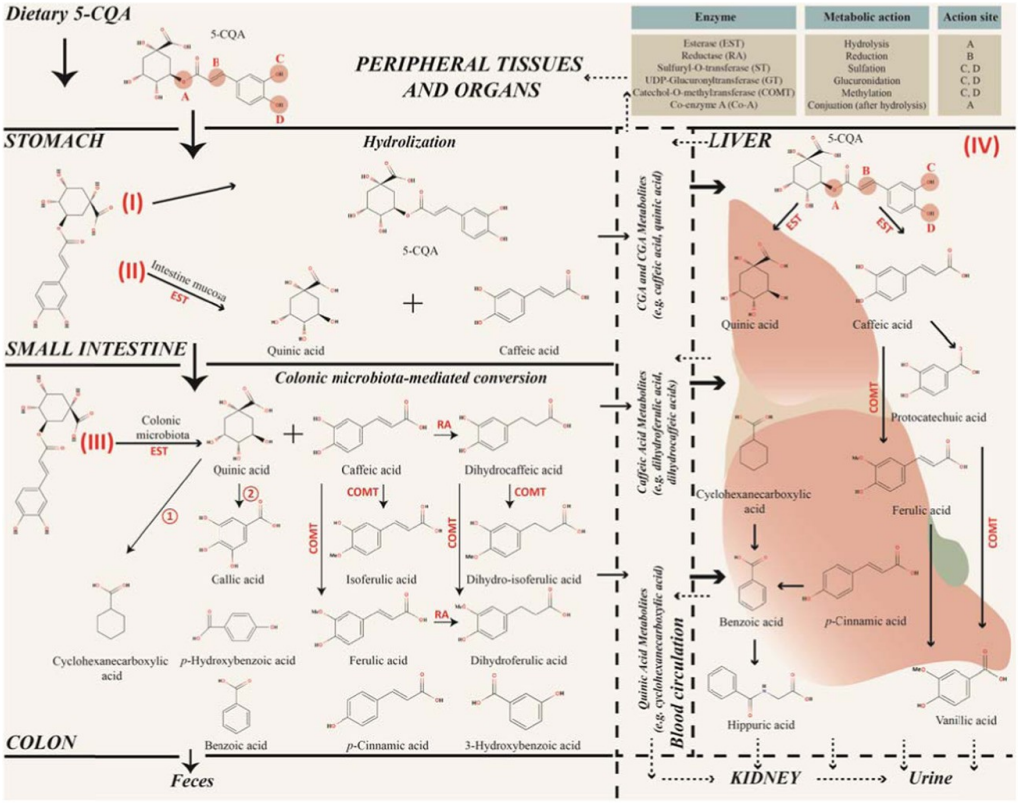


Figure 5. Detailed description of CGA metabolization pathways. Reproduced with permission (order number 5586021374300) from reference [42]. © John Wiley & Sons, Inc.

Another important point is metabolization as well as absorption of CGAs upon consumption. Several detailed studies have been conducted in order to investigate bioavailability of CGA molecules [34, 42-44]. Chlorogenic acids can be processed through multiple pathways in the human gastrointestinal tract. (1) First of all, CGAs can be absorbed without further metabolization through the stomach and/or the upper part of gastrointestinal tract. (2) If not absorbed from the stomach or the upper part of the gastrointestinal tract, CGA compounds can be metabolized via hydrolysis. In addition, untransformed or metabolized molecules can enter the bloodstream and thereby reach the liver, where further modifications are performed by hepatic enzymes (i.e. methylation, lactonization, sulphation, and glucuronidation) (3) Besides CGAs being transformed by endogenous enzymes, also human gut microbiota within the colon are able to utilize such compounds. Afterwards, microbial metabolites can be absorbed and further modified, which further increases the diversity of phenolic compounds (see Figure 5).

Clinical studies, as well as *in vivo* (animal) studies and *in vitro* assays, have provided only some preliminary hints about the cellular and molecular pathways responsible for the positive effects of CGAs in the human body [42]. In particular, a few protein targets of CGAs have been identified so far. However, the molecular determinants of the CGA effect are not fully understood, in particular which CGA compounds are contributing the most to the observed effect on protein function. The protein shown to be targets of CGAs include for instance, peroxisome proliferation-activated receptor alpha (PPARα), matrix metalloproteinases MMP-2 and MMP-9, acetylcholinesterase and butyrylcholinesterase, carbonic anhydrase and alpha-amylase as well as alpha-glucosidase [45-49].

2.7 Peroxisome proliferator-activated receptor alpha (PPARα)

Among CGA targets, PPARα is a protein belonging to the class of nuclear receptors (NRs). This group of receptors is able to regulate gene expression upon ligand binding and thus NRs are also referred to as ligand-activated transcriptional factors[1] [50]. After PPARα activation, heterodimerization with the retinoid X receptor (RXR) takes place. The PPARα-RXR heterodimer can recognize and bind to specific DNA motifs within the promotor region, which are known as peroxisome proliferator response elements (PPREs). Binding of PPARα to those regions modulates expression of the respective target genes.

From the structural point of view, PPARα possesses four structural and functional domains (Figure 6A): The modulator region, the DNA-binding domain (DBD), the hinge region and the ligand-binding domain (LBD) [51]. As apparent from the naming convention, the LBD is the most interesting region from a pharmaceutical point of view, since it is responsible for ligand binding. In addition, the LDB (Figure 6B) also plays a role in nuclear localization and heterodimerization.

[1] Besides being activated through the traditional idea of a protein-ligand mechanism, PPARα can also be activated through phosphorylation. In that case, protein activation is referred to as ligand-independent activation.

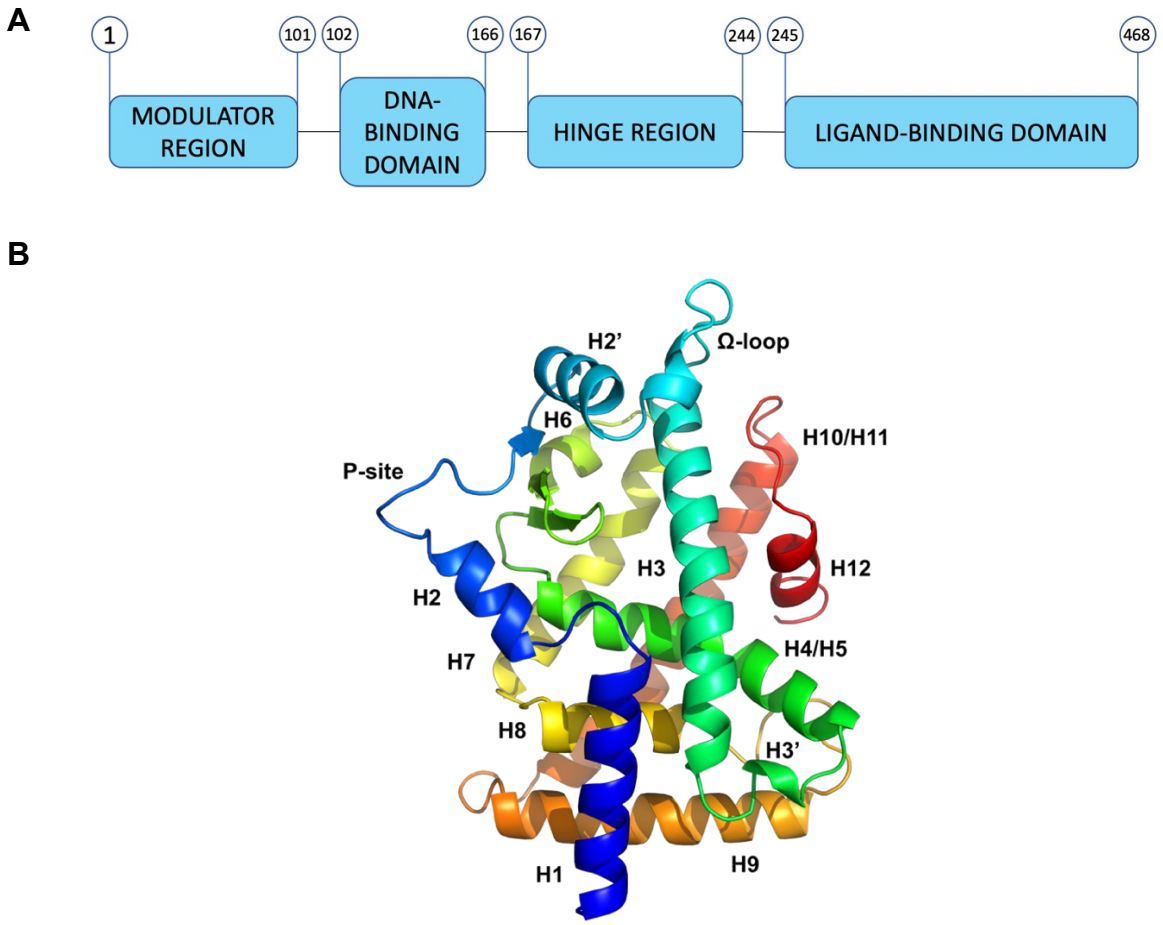


Figure 6. (**A**) PPARα domains. The numbering corresponds to the human PPARα protein sequence (UniProt-ID: Q07869). (**B**) 3D structure of the LBD (PDB code: 6KB3). Helices 1 to 12 (H1-H12) are labeled, as well as functionally relevant loops, such as the Ω-loop and the P-site[2].

PPARα is expressed in a variety of human tissues, such as liver, heart, colon, kidney, intestine and lung cells. The main purpose of PPARα is the regulation of the lipid metabolism; however, PPARα is also involved in plenty of other processes, which explains the diverse expression pattern [52]. In more recent studies, it was also shown that PPARα is located in different brain areas and, among other things, plays a role in glutamate homeostasis, microglia activation in neuroinflammation and anti-amyloidogenic effects [53, 54]. Besides endogenous ligands (i.e. fatty acids) and synthetic ligands (e.g. fibrates), cinnamic acid has also been shown to be able to activate PPARαand thereby attenuated amyloid plaque levels in a mouse model mimicking Alzheimer's disease. Another *in vivo* study demonstrated that cinnamic acid can protect dopaminergic neurons in a mouse model of Parkinson's disease [55-57].

2.7.1 PPARα Ligand Binding Site

The PPARα LBD harbors a Y-shaped binding cavity of 1300-1400 Å^3 buried inside the protein. That cavity can be divided into five sub-pockets, i.e. the Center, Arm I, Arm II, Arm III and Arm X pockets (for their definition, see Table 2).

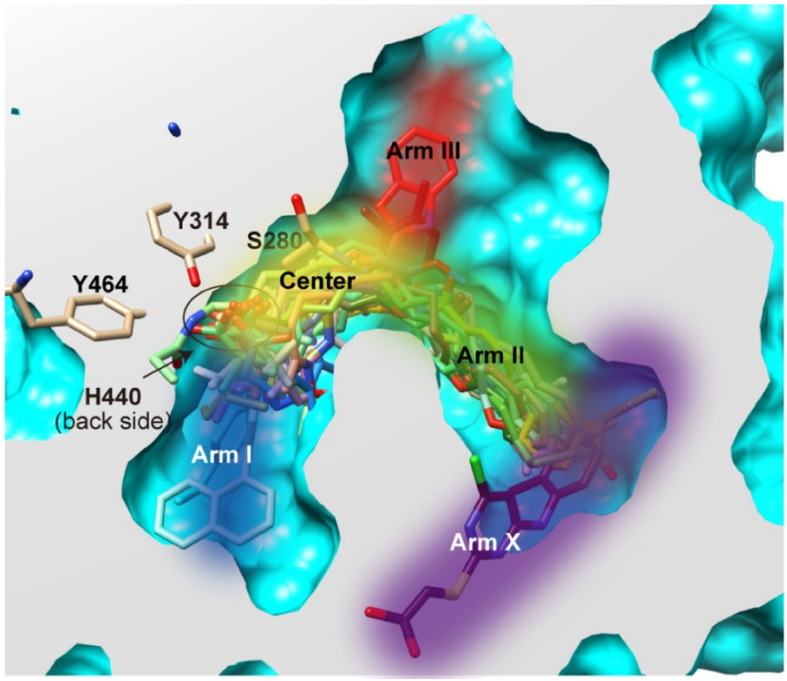
[2] This region is known to undergo post-transcriptional modification, namely, phosphorylation at residue S230, which is required for ligand-induced PPARα transcriptional activity.

Table 2. Ligand binding pockets of PPARα and their corresponding residues. Binding sites were named according to reference [57] and respective residues were identified by inspection of PPARα X-ray structures with co-crystallized ligands. Residues displayed in bold are forming hydrogen bonds (H-bonds) with PPARα ligands irrespective of the occupied binding site, either Center or Arm I pocket

Pocket	Residues
Center	C276, **S280**, **Y314**, F318, I354, M355, K358, **H440**, **Y464**
Arm I	V270, F273, H274, Q277, **S280**, **Y314**, **H440**, V444, I447, A454, A455, L456, L460, **Y464**
Arm II	I241, L247, E251, I272, C275, M330, V332, I339, F343, L344, L347
Arm III	F218, N219, M220, T283, E286, I317, M320, L321, V324
Arm X	L254, V255, L258, T279, E282, A333, Y334

As shown in Figure 7.2A-D, different molecules are able to bind and activate PPARα. Despite having a different shape or extending into different pockets, these agonists all share a carboxylate group that is able to form hydrogen bonds with four protein residues, namely S280, Y314, H440 and Y464 (using the sequence numbering of the human PPARα, UniProt Q07869). These residues are known experimentally to be important for both ligand binding to PPARα and receptor activation [58-61]. In general, stabilization of H12 seems to have an important role in receptor activation. Mutagenesis data, for instance, showed that residues L460, Y464, M467, and Y468 have the biggest impact on H12 stability and thus PPARα activation. A combination of computational and experimental investigations also revealed that non-polar side-chain interactions of Y464 contribute more to protein activation than the HB formed between protein and the carboxylate group of the ligand [62]. Furthermore, a hydrogen bond network among E315 (H4/H5), R388 (loop between H8 and H9), and R434 (H11) and Y468 (H12) favors protein activation through stabilization of H12. In the closely related isoform PPARγ, residues E471 (E462 in PPARα) and K301 (K292 in PPARα) are involved in a charge clamp, which is important for co-activator recruitment. Considering that PPARγ E471 is highly conserved in nuclear receptors (>85%), it is assumed that E462 has a similar role in PPARα. This is supported through experiments in which a E462Q mutantion suppressed transactivation [62].
As H12 stabilization is a key component for PPARα activation, it is assumed that interactions with adjacent residues and/or helices can reduce fluctuations and thus promote protein activation and/or co-activator recruitment. Partial agonists of PPARγ, for instance, stabilize H3 instead of H12, which promotes co-activator recruitment [57, 63, 64]. Besides H3, the Ω-loop is also known to have an impact on protein activation. Wy14643, for instance, is a PPARα agonist that possesses a unique second binding site which involves the Ω-loop. If a second molecule of Wy14643 is bound to that pocket, loop fluctuations are reduced and stabilization of H12 is subsequently improved, contributing to receptor activation [65].

1



2

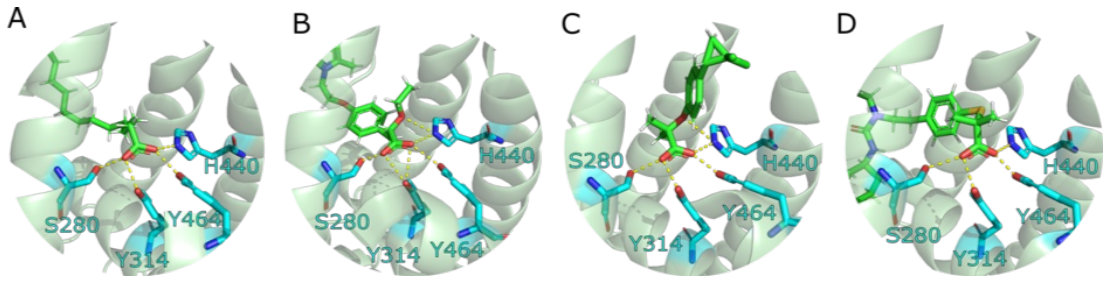


*Figure 7. (**1**) PPARα binding site. Reproduced under CC BY-NC-ND 4.0 license from reference [57]. © Cell Press. (**2**) **(A)** PPARα bound to stearic acid, a fatty acid that acts as endogenous ligand of the receptor (PDB code: 6LX7); **(B)** PPARα bound to saroglitazar, a drug for the treatment of type 2 diabetes mellitus and dyslipidemia (PDB code: 6LXB); **(C)** PPARα bound to ciprofibrate, a lipid-lowering drug (PDB code: 6LX5); **(D)** PPARα bound to GW7647, a potent and selective PPARα agonist used as a positive control in PPARα experiments (PDB code: 6KB3). Residues S280, Y314, H440 and Y464 are displayed as sticks and labelled in bold. Carbon atoms of the ligands and protein residues are colored in green and cyan, respectively. H-bonds are shown with yellow dashed lines.*

3 Methods

3.1 Computational Biomedicine

As more and more structural and genomic information became available during the last decades, computational approaches have also become more popular to understand the details of molecular mechanisms, such as protein activation, protein inhibition or conformational changes upon ligand binding. Another factor that enabled more theoretical investigations was that computational hardware components gained tremendous power. Thus, computation time of the respective simulations got reduced and results are available in more desirable time ranges.

3.2 Artificial Intelligence (AI) based structure prediction

In recent years AI-powered software tools got increasing attention in areas of natural science, due to remarkable performance advantages compared to traditional approaches.

In 2020, AlphaFold2 an AI-powered software application to predict 3D protein structures, won the CASP competition with astonishing precision [24, 25]. Afterwards other research groups followed that example and also focused on protein structure prediction with AI [66].

In basic terms, a neural network is trained with thousands of 3D protein structures through which future unknown structures can be predicted. As the field of AI evolves continuously and different neural network architectures are developed, also more protein structure prediction (PSP) tools are appearing in different flavors. Details of these new PSP architectures can be found in the respective publications [25, 67-70].

In this thesis, I will focus on homology modeling, since missing residues or protein structures were predicted in my PhD work with such bioinformatics approach.

3.3 Homology Modeling (HM) based structure prediction

In a nutshell HM is performed as shown in Figure 8. In a first step, related protein structures are identified and the most promising candidate is selected as template for subsequent modelling steps. After selecting such a template, a pairwise sequence alignment is performed which serves as skeleton for the predicted protein structure. Once a protein model is built, evaluation takes place and if not satisfied either another template can be selected or the pairwise alignment can be repeated with different settings.
The following sections are providing more details to earlier described stages of HM.

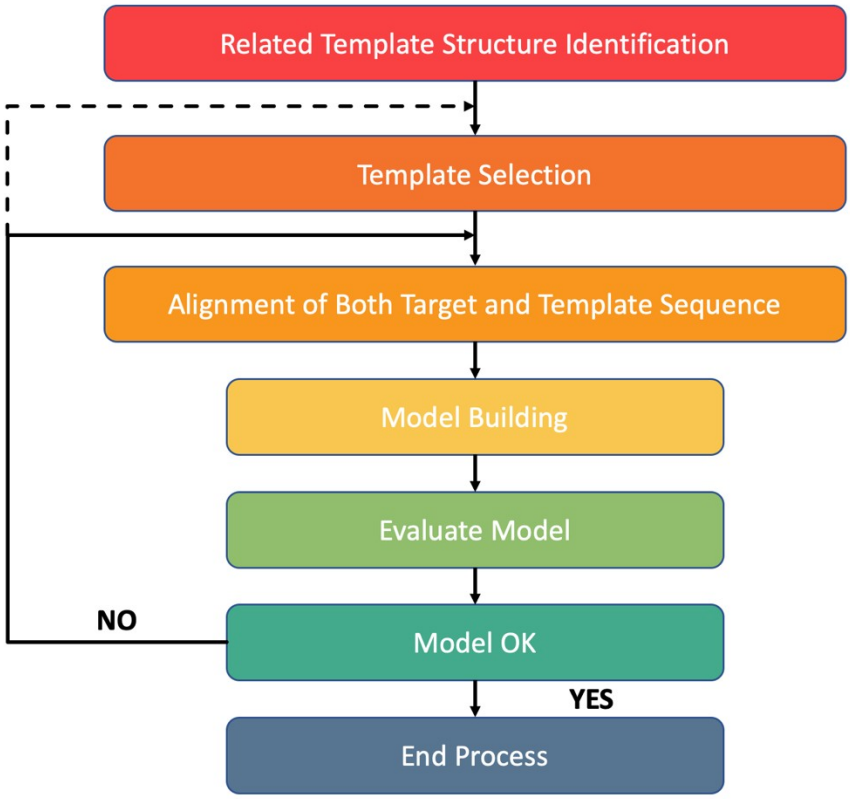


Figure 8. Workflow of homology modelling. Reproduced from [71]

3.3.1 Sequence alignment

In order to predict protein structures, which have not been solved yet experimentally and thus are missing in the Protein Data Bank, many different algorithms are available nowadays. Despite using different methodologies, most algorithms take advantage of multiple sequence alignments (MSAs) to find related amino acid sequences and identify conserved protein regions.
Most of the time sequence alignments are produced by using Dynamic Programming based techniques or Hidden Markov Models (HMMs).

3.3.1.1 Dynamic Programming (DP)

DP is a computational approach that divides a global problem into multiple sub-problems in order to use their step-wise solutions to answer the main question.

A popular approach to compute MSAs is in a progressive manner. In a first step, a so-called guide tree is built which defines the order of alignment. This guide tree is calculated based on a distance matrix of pairwise alignments of respective sequences.

D =

	A	**B**	**C**	**D**
A	-			
B	98	-		
C	32	46	-	
D	35	44	99	-

Scheme 1. Distance matrix of sequences A, B, C and D. Numbers are representing percentages of sequence similarity based on their pairwise sequence alignment.

These pairwise alignments are constructed using DP. An example would be the Needleman-Wunsch algorithm, which computes global sequence alignments following the subsequent scheme:

1. *Initialization:*

$$\text{Mismatches, Deletions, Insertions} = 1$$
$$\text{Matches} = 0$$

$$A_{0,0} = 0$$
$$A_{i,0} = A_{i-1,0} + f(i), \quad 1 \leq i \leq m$$
$$A_{0,j} = A_{0,j-1} + f(j), \quad 1 \leq j \leq n$$

2. *Matrix construction:*

$$A_{i,j} = \min \begin{cases} A_{i-1,j-1} + w(a_i, b_j) \rightarrow \text{Match or Mismatch} \\ A_{i-1,j} + f(\text{Deletion}) \rightarrow \text{Deletion} \\ A_{i,j-1} + f(\text{Insertion}) \rightarrow \text{Insertion} \end{cases}; 1 \leq i \leq m; 1 \leq j \leq n$$

$A =$

		F	V	I	H	D	M	E
	0	1	2	3	4	5	6	7
F	1	**0**	1	2	3	4	5	6
V	2	1	**0**	**1**	2	3	4	5
H	3	2	1	1	**1**	2	3	4
D	4	3	2	2	2	**1**	2	3
M	5	4	3	3	3	2	**1**	2
E	6	5	4	4	4	3	2	**1**

Scheme 2. Example of a distance matrix of the pairwise alignment constructed in step 2.

3. *Backtracking*

After backtracking the following pairwise alignment is obtained:

Sequence 1: FV−HDME
Sequence 2: FVIHDME

3.3.1.2 Hidden Markov Models (HMM)

In bioinformatic applications HMMs are often used to find related protein sequences, which is, among other things, fundamental for protein structure prediction (see AlphaFold2 or more traditional approaches used for homology modeling, as HHpred [25, 72]).
A popular framework which implemented HMMs is HHblits. HHblits uses the so-called HMM profiles in order to align protein sequences and employs the following workflow:

1. Construction of an HMM from a query sequence or multiple sequence alignment. The HMM represents a protein sequence or MSA as transition and emission probabilities, through which a likelihood of each amino acid, gap or insertion at every position can estimated.

2. The HMM can be translated into a so-called 219-letter profile sequence, which is used to prefilter a target database. This prefilter step represents a significant speed up and reduces the amount of subsequent calculations by a factor of ~2500.

3. Since HHblits is an iterative approach, found sequences below a defined threshold can be added to the initial protein sequence or MSA to improve results in later iterations.

3.3.2 Model building

After building a MSA as described before in section 3.3.1.2, the protein modelling part would be the next step [73]. Some common modelling approaches are: (1) Rigid body assembly [74], (2) segment matching [75], (3) spatial restraints [76], (4) or artificial evolution [77]. The resulting modeled protein structures, however, can be inaccurate and a model refinement is often performed to improve the initial protein models. In particular, amino acids which are not conserved or residues which are part of loop regions might need refinement, which can be accomplished, among other approaches through Molecular Dynamics or Monte-Carlo based sampling techniques [78, 79].

3.3.3 Model evaluation

As homology model consists of stages which build up on each other (see Figure 8), errors can accumulate and result in poor structure predictions. Therefore, homology models should be treated with caution especially if low sequence identity (<35%) between target-template alignment is present [80-82]. Hence, model evaluation should be performed in order to assess the quality of predicted protein structures. A first step of the evaluation should be to determine if the correct template was used to build the model itself. If, for instance, a protein undergoes large conformational changes upon protein activation or ligand binding, it is important to select a template structure with the desired conformation. In addition, protein structures should follow stereochemical restrictions, including features such as bond length, bond and torsion angles, side-chain planarity and avoidance of steric clashes.
In this regard, SWISS-MODEL, the webserver I used in this thesis for homology modeling, has implemented model quality evaluation methods called QMEANDisCo for soluble proteins and QMEANBrane for membrane proteins [83, 84].

3.4 Molecular Docking

Molecular Docking is a computational method to predict geometrically and energetically favorable structures of protein-ligand complexes. This approach is considered as a computational “cheap” approach, since it is able to predict a large number of binding poses in a relative short amount of time.
In general, molecular docking follows the basic workflow shown in Figure 9:

Figure 9. Molecular Docking workflow.

1. *Ligand and Protein Structures*

2. *Conformational sampling*
   Ligand and protein structures can be treated as rigid or flexible bodies. Depending on the selected docking approach this can have an enormous impact on the accuracy of the predicted binding poses and on computational performance. The first docking algorithms considered both interaction partners as rigid bodies to speed up calculation; however, this also limits the accuracy of the predicted ligand poses. A system with both molecules treated as rigid bodies consists of six degrees of freedom (translational and rotational). Most of the time rigid body docking approaches only work for very simple problems and fail for more complex systems, because biological systems are always very dynamic and adaptable to their environment. In order to include, at least in part, the dynamics of the protein-ligand complex, more advanced docking algorithms make use other concepts. A popular solution is to integrate a larger conformational space of ligand structures in which the compound is treated as a flexible molecule. These approaches include more conformational degrees of freedom (rotational bonds), which result in increasing complexity and computational costs of those algorithms. Moreover, flexibility can also be introduced to the protein itself. This can either be done for the whole protein or for residues within a certain cutoff with respect to the ligand position. In general, accuracy is often improved when introducing flexibility to more elements within the protein-ligand complex. However, this also implies that a larger conformational space will be sampled and thus calculations require more computational power and time.

3. *Pose evaluation*
   Predicted docking poses are scored and ranked based on a scoring function, a mathematical expression that is used as proxy of the binding affinity or stability of the complex. These scoring functions are one of the main ingredients of a docking implementation and vary for different docking programs. However, this is often also considered as a downside of docking. In general, the outcome of different docking algorithms shows a high variance in precision recovering the true binding pose and, even if the true binding pose is among the docking pose solutions, it might not be ranked first according to the docking score.

Molecular docking with a rigid protein target can also be an advantage if little or no information about possible inhibitors/activators is known. In this case, a large number of compounds can be screened against protein targets in a short amount of time and

narrow down millions of molecules to a few hundreds (i.e. high throughput docking or virtual screening). Nevertheless, molecular docking can also be used in a more individualized scenario (i.e. a few protein-ligand complexes, as done in this thesis) and including flexibility when more information about the binding site is known (as also done in this thesis). Here, available information can be integrated into the docking procedure and bias the predicted binding poses towards solutions compatible with the experimental data. If, for instance, crystal structures with other ligands or experimental evidence, such as mutagenesis data, are available it can be useful to incorporate such information. This can be either done by modifying the underlying docking score or, in a less biased way, by filtering the resulting docking poses.
An illustrative and simple procedure to explore the conformational space of the ligand in flexible docking is through the Metropolis algorithm also known as simulated annealing (SA). An implementation of such an algorithm was already done when Computer-Aided Drug Design (CADD) was still its infancy due to its reasonable usage of computational costs and relative simplicity [85]. At first, the most important values are initialized: (1) An initial temperature as well as (2) number and size of annealing cycles to be performed. (3) Moreover, the selected ligand has to be placed near the binding site. A common approach is to use a minimum conformation of respective ligand as a starting pose.

Once the initial parameters are defined (see the SA pseudocode below), the actual simulated annealing procedure is carried out. In order to let the ligand explore each degree of freedom, a small random displacement is performed. In particular, that means that, in each step positional coordinates are changed in terms of translational, rotational and conformational space with respect to the rigid protein structure.

In order to assess if the resulting docking pose is converging towards a minimum in the binding free energy landscape, an evaluation using a molecular affinity potential (i.e. docking scoring function) is performed. If the new pose has a lower score than the previous one, the docking pose is accepted immediately. If, however, a worse docking score is achieved, i.e. a higher energy value for that particular pose, the acceptance is evaluated in a probabilistic manner, following this equation:

$$P(\Delta E) = \exp\left(-\Delta E / k_B T\right)$$

Here, ΔE is the difference in energy between the current and the previous docking pose, k_B is the Boltzmann constant and T the temperature. The acceptance of higher energy poses is crucial and one of the key components of simulated annealing. By accepting poses with a worse docking score, it is possible to overcome local minima in order to get to the global minima of the free energy surface. In that regard, the temperature plays an important role, since at higher temperatures the acceptance ratio of worst performing docking poses is greater compared to steps with lower temperatures. The temperature is updated each cycle according to

$$T_i = g\, T_{i-1}$$

where T_i is temperature at step i and ga constant variable between 0 and 1. After one iteration, the best docking pose is used as starting point for the next iteration which is repeated until the user-defined number of cycles is reached.

Pseudocode of a simulated annealing cycle:

```
#Initializations:
T                       =     starting temperature
docking_pose            =     coordinates of initial staring pose
number_of_steps         =     size of steps performed in a cycle
best_docking_score      =     score of initial ligand placement

for i in number_of_steps:
      updated_docking_pose = random_displacement(docking_pose)
      docking_score = get_docking_score(updated_docking_pose)
      if (docking_score < best_docking_score):
            docking_pose = updated_docking_pose
            best_docking_score = docking_score
      else:
                if (estimate_probability(updated_docking_pose) >
random.uniform(0, 1)):
                      docking_pose = updated_docking_pose
                      best_docking_score = docking_score
update(T) #update temperature for next annealing iteration
```

When using a flexible ligand, the conformational space is in fact sampled to a larger extent; however, the receptor is still rigid and not able to adjust towards the ligand. Moreover, this lack of flexibility results in a higher chance of false negative docking poses. There are several approaches that aim at addressing these kinds of problems: (1) A very straight forward approach is to decrease the van der Waals radii of the ligand, the protein or both. That procedure eliminates close contacts which would otherwise lead to a higher docking score or in the worst case to a rejection of that docking pose. Nevertheless, since that adjustment did not allow movement of side-chain atoms, it may still result in a wrong conformation of the complex. Moreover, since decreased van der Waals radii also increase the volume of the respective binding site, false positive hits may occur during virtual screenings. (2) Another approach, which is quite popular, is to perform a so-called ensemble docking. If multiple receptor structures are available this might be a reasonable approach to compensate for missing flexibility. Here, multiple docking runs are performed independently with an ensemble of receptor structures. The main problem, however, is often the lack of alternative receptor structures. Also, this approach does not solve the issue of natural side-chain or backbone movement. Here, the movement is imitated through different rotameric states existing in different structures. Hence, new conformations of the binding residues, not present in the available receptor structures, cannot be explored and similar limitations compared to (1) are present. (3) A last, more obvious option would be to introduce flexibility to the whole protein structure (or at least the region near the binding site) using molecular dynamics.

Many different docking algorithms have been developed along the years, some examples include HADDOCK, Rosetta, AutoDock Vina and Glide [86-89]. In the following, I will explain the two approaches I used in this thesis: HADDOCK and Induced Fit Docking (IFD, implemented in the Glide tool of the Schrödinger suite).

3.4.1 HADDOCK

HADDOCK is an open source software package which allows to perform different molecular docking approaches, such as protein-protein, protein-peptide and protein-nucleic acid docking. Furthermore, HADDOCK can be used to perform (covalent) protein-ligand docking (see sections 4.1.2 and 5.1.2).

In general, the Haddock docking protocol consists of three different main stages:

1. Randomization of starting orientations and rigid body docking (it0):
2. Semi-flexible simulated annealing (it1)
3. Flexible final refinement (water)

In steps 2 and 3 the ligand is considered flexible and receptor flexibility is also included, either partially in step 2 or in full in step 3.

HADDOCK ranks the generated docking poses based on a so-called HADDOCK score, which is linear combination of different energy terms. This scoring function is adjusted in each docking stage and can also be modified if desired. In this regard, the default scoring functions for a protein-small molecule docking are defined as follows:

```
it0   = 1.0 Evdw + 1.0 Eelec + 1.0 Edesol + 0.01 Eair - 0.01 BSA
it1   = 1.0 Evdw + 1.0 Eelec + 1.0 Edesol + 0.1  Eair - 0.01 BSA
water = 1.0 Evdw + 0.1 Eelec + 1.0 Edesol + 0.1  Eair
```

Here, Evdw represents the van der Waals intermolecular energy, Eelec the electrostatic intermolecular energy, Edesol the desolvation energy, Eair the distance restraints energy and BSA the buried surface area. If additional experimental restraints are used during a docking stage, further terms are added to the weighted sum above.

Further details can be found in section 4.1.2 and 5.1.2, where specific modifications of the general protocol so far presented are described to tailor it to the problem at hand.

3.4.2 Induced Fit Docking (IFD)

Schrödinger has implemented an Induced Fit Docking protocol in order to include flexibility of the residues surrounding the ligand and thus predict more accurate protein-ligand binding poses. This protocol consists of the subsequent steps:

1. An initial Glide docking step is performed. In order to sample up to 80 initial docking poses, multiple runs are performed with either a trimmed receptor or an untrimmed receptor with an atom-wise softened potential. These structures are afterwards clustered and representative poses are selected based on GlideScore and cluster related parameters.

2. In the next step, Prime is used to predict side chain orientations for every protein-ligand complex. By default, Prime is considering residues within a range of 5Å of the respective ligand. If necessary, additional residues can be included or excluded to make sure that important rotameric side chain conformations remain untouched.

3. Following the side-chains prediction, a minimization of the whole protein-ligand complex is performed.

4. Afterwards, Glide is re-docking each ligand into the induced fitted protein structure. Here, default Glide SP settings are used, which implies that ligands are treated as rigid bodies.

In general, the scoring function of Glide possesses a lipophilic-lipophilic term, a hydrogen bond term, a rotatable bond penalty, and contributions from protein-ligand Coulomb-vdW energies.
If desired additional restraints can be applied during different stages of the IFD protocol. Such restraints are H-bond and metal interaction restraints as well as core restraints (see section 5.1.2 for the type of restraints I used in my PhD work). The GlideScore is a linear combination and described by the following equation:

$$GScore = a \times vdW + b \times Coul + Lipo + Hbond + Metal + Rewards + RotB + Site$$

In this context vdW respresents the van der Waals interaction energy; Coul the Coulomb interaction energy; Lipo the lipophilic contact plus phobic-attractive term; Hbond the hydrogen-bonding term; Metal the metal-binding term; Rewards are resembling various rewards or penalties; RotB the penalty for freezing rotatable bonds; and Site the polar interactions in the active site; the coefficients of the vdW and Coul terms are: a = 0.050 and b = 0.150 for Glide 5.0.

3.5 Molecular Dynamics Simulation

Biological systems, in particular proteins, are flexible and constantly in motion. Instead, most crystal structures show an average of protein states captured in the crystal itself. Deviations can occur for instance, due to artifacts from crystal packing and/or the challenge to distinguish between isoelectronic groups in X-ray crystallography. In addition, solvent exposed residues are often very flexible and multiple rotameric states are present within the protein crystal. This leads to large B-factors and fitting of these residues is not reliable anymore.
Molecular dynamics (MD) can, for instance, help to refine protein structures or show a more dynamic picture of in terms of residue contacts (i.e. H-bond network, solvent exposed surface area or in general protein flexibility) [90, 91]. However, the range of MD applications is far wider and MD can also be useful in order to predict the outcome of controlled changes of a biological system. Such changes include for instance altered protein environment, changes in protonation states, point mutations as well as application of mechanical forces or electric potentials to ligands or certain protein areas.
It is well known that point mutations of specific residues can lead to severe changes in the free energy landscape and thus cause misfolding or malfunction of proteins. Similarly, protein dynamics can influence both ligand binding thermodynamics and kinetics. In this regard, MD simulations can be very advantageous in drug discovery campaigns.

Molecular dynamics is based on Newton's equation of motion to predict future states of a particular system. The MD algorithm requires small finite time steps to ensure numerical stability, in the order of one tenth of the fastest vibration of the system under consideration. In the case of classical, force field-based MD, this means a time step in

the order of femtoseconds (10^{-5}s). The relevant biological processes, however, are observable at longer time scales, i.e. in the range of microseconds or even seconds. Thus, in most simulations, intra- and intermolecular forces have to be (re)-calculated millionfold. Considering that even a simple molecular system, such as a protein-ligand complex in water, consists of thousands of atoms, the resulting computational requirements can be quite demanding.

To tackle those kinds of problems it may be useful to consider:

1. An increase of the computational power used for a simulation. Through an optimized and parallelized codebase, the MD algorithm can make use of multiple CPUs at a time and boost performance significantly [92]. In addition, computations can also be offloaded to GPUs, which can accelerate MD simulations even more. Therefore, MD simulations are often done on supercomputers, which are able to provide adequate hardware resources. In this thesis, I used the supercomputer resources of the Jülich Supercomputing Centre (JSC) and the RTWH High Performance Computing cluster, as well as local computers at IAS-5/INM-9.

2. Another approach to make sure to get sufficient results in an appropriate time window is to reduce the details of the system. Depending on the particular research question it could be advantageous to decrease the granularity and thus reduce the number of interactions to be calculated within a system. Some examples are, for instance coarse-grained (CG) simulations or even mesoscale approaches (Figure 10) [93].

3. If, however, computational offload or reduction of granularity is not possible due to the research question, a third option might be suitable. In that case, the so-called enhanced sampling techniques can allow to explore a larger configuration space in a reduced amount of time [94].

In this thesis, classical (force field-based) all atom simulations were used to investigate molecular determinants of protein-ligand interactions.

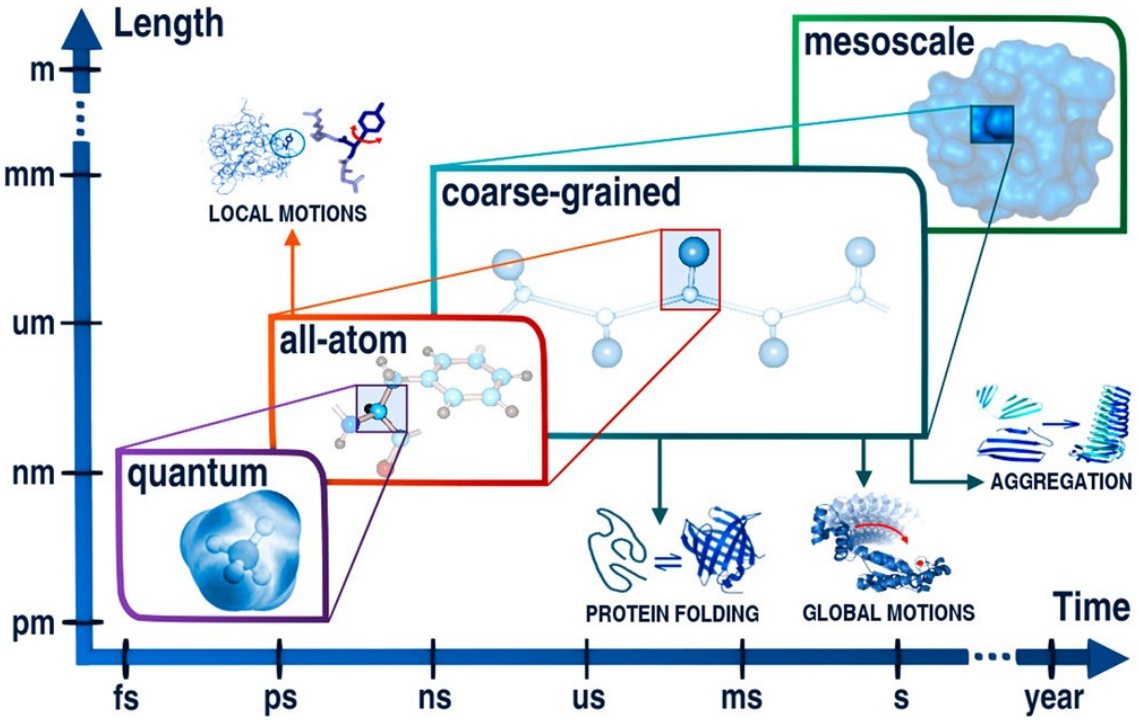


Figure 10. Time scales of different Molecular Dynamics approaches. The plot shows the (spatial and time) resolution for quantum-, all-atom-, coarse-grained- and mesoscale simulations. Depending on the chosen granularity, different time scales as well as biophysical events can be modeled. Reproduced with permission from reference [93] under CC-BY license. © American Chemical Society.

As mentioned earlier, Molecular Dynamics simulations are able to predict future states of a given system based on Newton's equation of motion. Given a set of Cartesian coordinates $x(t)$ and velocities $v(t)$ and time t, future positions at $t + \Delta t$ can be calculated. In particular, Newton's equations of motion are defined in the following way:

$$v_i(t) = \frac{dx_i(t)}{dt}, i = 1 \ldots N$$

$$a_i(t) = \frac{dv_i(t)}{dt} = \frac{F_i(t)}{m_i}$$

Forces acting on each atom are obtained based on a potential energy function $U(x_i(t))$. Energies are represented as a function of their positions $x_i(t)$ and calculated as the negative gradient to retrieve respective forces.

$$F_i(t) = -\nabla U(x_i(t))$$

The potential energy function is described by a so-called force field in classical MD. Most force field are unique in terms of parametrization and/or interaction terms, which can make the selection of an appropriate force field quite challenging (see section 3.7).

3.5.1 Verlet Algorithm

The Verlet algorithm represents a numerical method to integrate Newton's equation of motion. In this case, Taylor expansions can be used to approximate positions at a later point in time $t + \Delta t$. Combining two Taylor expansions, i.e. the forward and backward expansion:

$$x(t+\Delta t)=x(t)+\dot{x}(t)\Delta t+\frac{1}{2}\ddot{x}(t)\Delta t$$

$$x(t-\Delta t)=x(t)-\dot{x}(t)\Delta t+\frac{1}{2}\ddot{x}(t)\Delta t$$

gives us the following equation:

$$x(t+\Delta t)=2x(t)-x(t-\Delta t)+\ddot{x}(t)(\Delta t)^2$$

Considering Newton's second law of motion, meaning that the second derivative of position with respect to time $\ddot{x}(t)$ gives its acceleration $a(t)$, we can retrieve then the Verlet integration algorithm:

$$x(t+\Delta t)=2x(t)-x(t-\Delta t)+\frac{F}{m}(\Delta t)^2$$

However, this approach has some disadvantages. As shown in the equation above, an explicit velocity term is missing and thus velocities can only be calculated after atom positions are available at time $t + \Delta t$, that is:

$$v(t)=\frac{\left[x(t+\Delta t)-x(t-\Delta t)\right]}{2\Delta t}$$

Another approach to overcome that problem is to use half time steps:

$$v\left(t+\frac{1}{2}\Delta t\right)=\frac{\left[x(t+\Delta t)-x(t)\right]}{\Delta t}$$

Nonetheless, this integration scheme requires positions from two previous time steps $x(t)$ and $x(t - \Delta t)$, which are not available at $t = 0$.

3.5.2 Velocity Verlet Algorithm

The Velocity Verlet algorithm solves the problem of the original Verlet integration scheme and includes velocities at each time step, which also ensures that the algorithm is self-starting. In this particular integration scheme, higher order differential equations are reduced to first order equations. Thus, velocities can be expressed as follows:

$$v(t+\Delta t)=v(t)+\frac{a(t)+a(t+\Delta t)}{2}\Delta t$$

New positions are still obtained through a forward expansion of a Taylor series. Replacing $\dot{x}(t)$ with the velocity $v(t)$ and $\ddot{x}(t)$ with the acceleration $a(t)$, we get the following relationship:

$$x(t+\Delta t)=x(t)+v(t)\Delta t+\frac{1}{2}a(t)\Delta t$$

Given positions $x(t)$ and velocities $v(t)$ of the current time step, as well as an expression for $F(x)$, it is thus possible to calculate positions for subsequent time steps.

3.5.3 Leap-Frog Algorithm

Another more commonly used variation of the Verlet algorithm is the so-called leap-frog algorithm. Here, positions and velocities are updated on an alternate basis. New positions are updated on full time steps compared to velocities, which are updated on half time steps. The relevant equations are then:

$$x(t+\Delta t)=x(t)+\Delta x(t);\Delta x(t)=v\left(t+\frac{1}{2}\Delta t\right)\Delta t$$

and

$$v\left(t+\frac{1}{2}\Delta t\right)=v\left(t-\frac{1}{2}\Delta t\right)+\Delta v(t);\Delta v(t)=a(t)\Delta t$$

3.6 Energy Minimization

In order to perform a molecular dynamics simulation, the system of interest has to be prepared accordingly. In both experimental as well as computational predicted protein structures, steric clashes or other unfavorable conformations possibly present in the initial configuration can result in a high potential energy and thus in numerically unstable simulations. Therefore, an energy minimization and equilibration of respective system is needed before starting production MD simulations.
In my thesis, I used the Steepest Descent Algorithm to perform energy minimization. The main equation of that algorithm is:

$$r_{n+1}=r_n+\frac{F_n}{max\left(\left|F_n\right|\right)}h_n$$

At first, forces F_n, or the negative gradient of the potential energy of the current step is calculated. The factor h_n denotes the maximum displacement and $\max(|F_n|)$ indicates the largest scalar force on any atom. This ensures that the maximum step size is only taken by coordinates that have the steepest gradient.

After minimization, the system is brought the temperature of interest, typically 300 K, by running a short simulation in the canonical or NVT ensemble (i.e. constant number of particles, volume and temperature); this requires coupling the system to a

thermostat. Afterwards, the density of the water box surrounding the system is adjusted by running a short simulation in the isothermal-isobaric or NPT ensemble (i.e. constant number of particles, pressure and temperature), which is achieved by coupling the system to both a barostat and a thermostat. For further details on the equilibration protocol I used in my thesis see section 5.1.3.

3.7 Force Fields

Force fields describe inter and intra-molecular interaction energies by using parameters derived through quantum mechanical calculations, from experiments and/or both.

In this thesis, I used the Amber compatible force fields [95, 96] to perform simulations described in section 5.1. The potential energy function of Amber is composed by the following energy terms:

$$V_{AMBER} = V_{BONDS} + V_{ANGLES} + V_{DIHEDRALS} + V_{NON-BONDED}$$

The covalent bond term is represented as harmonic potential with a force constant K_r, a bond length of r and an equilibrium bond length of r_{eq}.

$$V_{BONDS} = \sum_{BONDS} K_r (r - r_{eq})^2$$

Bond angles are also represented by a harmonic potential, with a force constant K_θ, an angle θ and an equilibrium angle θ_{eq}.

$$V_{ANGLES} = \sum_{ANGLES} K_\theta (\theta - \theta_{eq})^2$$

The dihedrals are computed as follows:

$$V_{DIHEDRALS} = \sum_{DIHEDRALS} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$$

where V_n is the dihedral force constant, whereas n is the periodicity and γ the phase angle. The non-bonded energy terms can be further divided into van der Waals and electrostatic interactions.

$$V_{NON-BONDED} = V_{VAN\ DER\ WAALS} + V_{ELECTROSTATIC}$$

Here, van der Waals interactions are represented by a Lennard-Jones potential (1st term of the summation) and electrostatic forces by Coulomb's law (2nd term of the summation).

$$V_{NON-BONDED} = \sum_{i<j}^{N} \left[\left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right) + \frac{q_i q_j}{\in R_{ij}} \right]$$

AMBER ff99SB-ILDN [96] is a force field designed for amino acids, whereas small molecules, such as CGAs and/or other organic compounds, have to be parameterized with the general Amber force field (GAFF2) [97].
Moreover, in order to simulate the system under physiological conditions, water molecules and ions have to be added to the simulation box. The TIP3P model was used for water molecules, since its combination with AMBER compatible force fields showed good performance [98]. Ions parameters were taken from reference [99].

3.8 Clustering

Molecular Dynamics simulations are able to give insights into physicochemical properties of a system over time. To visualize the most populated states of a system, clustering of similar frames is a useful data dimensionality reduction tool. In other words, similar configurations are “grouped” together, which provides information about visited states of a system. Besides clustering of MD trajectories, docking poses can also be clustered in order to find similar protein-ligand complex structures.

Gromos is a simple clustering algorithm, which is often used to classify configurations sampled in MD trajectories of protein-ligand complexes. Furthermore, this algorithm is the default option in the protein-ligand docking protocol of Haddock (version 2.2) used in this thesis (see section 4.1) and available as well in GROMACS, the MD engine employed in this thesis.
Clustering algorithms require a so-called distance matrix, which is calculated as an all-to-all matrix among each data point within the data set. When dealing with MD frames or docking poses, RMSD measurements between each frame/pose are often used as a distance criterion. The RMSD is calculated as follows:

$$RMSD(r,s)=\left[\frac{1}{N}\sum_{i=1}^{N}\left\|r_i-s_i\right\|^2\right]^{\frac{1}{2}}$$

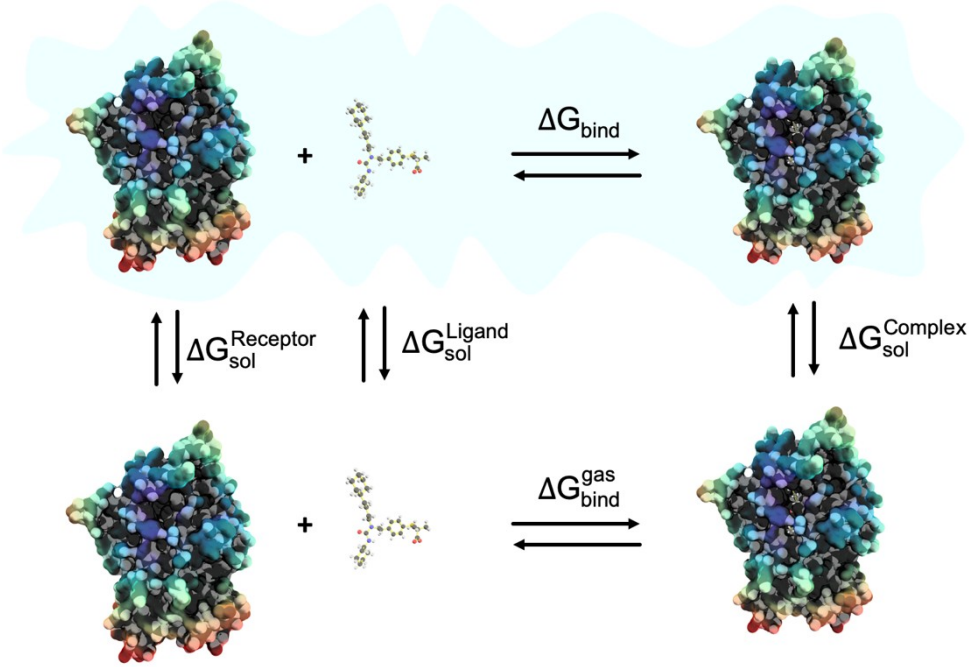
Here, r and s are representing atoms coordinates, whereas N is defining the amount of considered atoms.
At first, the Gromos clustering algorithm searches for the structure with the highest number of neighbors within a user-defined distance cut-off. This structure serves as centroid and forms, with all its neighbors, the first cluster. Afterwards, every pose belonging to that cluster is eliminated from the pool of structures and the search for new clusters is repeated iteratively [100].
Besides RMSD, other metrics can be used for clustering. For instance, in section 5 of this thesis, docking poses were clustered using the Fraction of Common Contacts (FCC), as implemented in Haddock [101].

3.9 Molecular Mechanics with Generalized Born and Surface Area solvation (MM-GBSA)

In order to assess binding energies of small molecules, many different approaches have been developed [102]. The most accurate methods are based on quantum-mechanics calculations (see Figure 10); however, such high-level approaches are quite demanding in time and computational costs. Thus, other approaches are more

popular, since they provide a faster way to calculate free binding energies. It should be noted, that faster calculations always come with the trade-off towards accuracy.

The methodology used in this thesis is called MM-GBSA (molecular mechanics (MM) with generalized Born (GB) and surface area (SA) solvation) [103]. In this approach, and the following thermodynamic cycle is employed:



In general, the binding free energy of a protein-ligand system can be estimated as follows:

$$\Delta G_{bind} = \langle G_{COMPLEX} \rangle - \langle G_{PROTEIN} \rangle - \langle G_{LIGAND} \rangle$$

$$\langle G_X \rangle = \langle \Delta E_{MM} \rangle - \langle \Delta G_{SOLVENT} \rangle$$

The term above can also be rewritten as:

$$\Delta G_{bind} = \Delta H - T \Delta S$$

where ΔH represents the enthalpy, T the temperature and ΔS the entropic term. The latter part of the equation (-TΔS) is often dismissed when dealing with similar ligands, which results in an effective free energy. As in most cases relative free binding energies are compared, it is sufficient enough to rely on ΔH. The enthalpy can be further decomposed into ΔE_{MM} (internal energy within the gas phase), which is retrieved from the force field and $\Delta G_{SOLVENT}$.

$$\Delta H = \Delta E_{MM} - \Delta G_{SOLVENT}$$

$$\Delta E_{MM} = \Delta E_{BONDED} + \Delta E_{NON-BONDED} = \left(\Delta E_{BOND} + \Delta E_{ANGLE} + \Delta E_{DIHEDRAL}\right) + \left(\Delta E_{ELE} + \Delta E_{vdW}\right)$$

$$\Delta G_{SOLVENT} = \Delta G_{POLAR} + \Delta G_{NON-POLAR} = \Delta G_{PB/GB} + \Delta G_{NON-POLAR}$$

The solvation energy can further be split into two parts, namely the polar (ΔG_{POLAR}) and non-polar ($\Delta G_{NON\text{-}POLAR}$) part. The polar contribution of the solvation energy is either estimated through the Poisson–Boltzmann (PB) or the generalized Born (GB) equation and includes electrostatic interactions between solvent and protein; the latter was used in this thesis. Usually, the non-polar part is derived from solvent accessible surface area (SASA) calculations and different methods has been implemented to derive respective energy term. The non-polar term includes protein-solvent van der Waals dispersion interactions, as well as the cost for creating a cavity into the solvent and changes in the solvent entropy, as described in reference [104] and the equation below:

$$\Delta G_{NON-POLAR} = \Delta G_{DISP} + \Delta G_{CAVITY} = \Delta G_{DISP} + \left(CAVITY_{TENSION} \times \Delta SASA \times CAVITY_{TENSION} \right)$$

In this thesis, I used the MM-GBSA implementation in GROMACS called gmx_MMPBSA, described in reference [105].

4 Molecular determinants of acrylamide neurotoxicity

This chapter is based on publication [1]. Citations details, as well as author contributions, can be found on page 4 of this thesis. Text copied directly from the article is indicted in italics and between quotes.

4.1 Computational Details

4.1.1 Ligand and protein structures

"*The product of the corresponding Michael addition reaction, i.e. propionamide (Figure 3), was used as ligand. The respective 3D structure was obtained from PubChem [106] (CID 6578).*
The 3D structures of the human proteins in Table 4 were taken from the Protein Data Bank [107, 108]. When more than one structure was available, the one at the highest resolution was chosen (see Table 4). Protein structures with missing residues were retrieved from the SWISS-MODEL repository [109] or generated with SWISS-MODEL [110], by selecting templates structures with the same sequence as the targets. When experimental structures of the human protein were not available, we generated homology models (see 4.2.2.3 - 4.2.2.7). Target-template sequence alignments were obtained with either BLAST [111] or HHblits [112], as implemented in the SWISS-MODEL webserver. Templates with the highest sequence identity and the highest resolution were selected and models were generated with SWISS-MODEL [110] (see 4.2.2.4-4.2.2.7). Protein structures were processed with MolProbity [113, 114] to add missing hydrogen atoms, assign histidine protonation states and perform His/Gln/Asn flips, if recommended. The reactive cysteines were modeled as already deprotonated [115], as expected for the Michael addition reaction to take place (see Figure 3, step 2). Therefore, our computational protocol does not take into account the energetic cost of Cys deprotonation, i.e. $\Delta G = \ln(10) \times kT \times (pK_a - pH)$. Moreover, we have assumed a default pH of 7, even though the protein targets in our dataset exhibit different optimal pH ranges (see Table S2) and the Michael addition reaction is favored at basic pH [116-118]. However, even if the Cys pK_a (calculated here with the H++ webserver at a default pH of 7) may predict population of the thiolate state smaller than the thiol one, reaction with acrylamide is expected to shift the acid-base equilibrium (step 2 in Figure 3) towards the deprotonated form." [1]

4.1.2 Acrylamide Dockings

"*Covalent docking to the reactive cysteine(s) of each target protein was performed using Haddock (version 2.2.) [87, 119]. We followed the standard covalent docking protocol of Haddock [120]. Such protocol was initially tested for covalent inhibitors of cathepsin K (HADDOCK developer team, 2018) and here we have validated it using experimental protein structures containing Cys-ACR covalent adducts (see Table S3).*
The covalent bond between Cys and the ligand is modeled by scaling down the van der Waals radius of the Cys sulfur atom 10-fold and introducing two distance restraints: (i) between the sulfur atom of the targeted cysteine and the reactive carbon atom of the ligand, set to 1.8 ± 0.1Å (i.e. the average length of a single C-S bond) and (ii) between the cysteine C_β atom and the ligand carbon atom adjacent to the reactive carbon, set to 2.8 ± 0.1 Å (i.e. the same as between the C_γ and C_ϵ atoms of methionine,

to model the proper angular geometry). The docking procedure [121, 122] consisted in the following three different stages: (1) A rigid body docking was performed with all geometrical parameters treated as fixed and allowing 180° rotations to generate 1000 initial poses. After minimization, the best scored 200 poses were selected for further re finement. (2) A semi-flexible simulated annealing simulation (SA) in torsion angle space was applied to introduce gradually flexibility to the system. SA can be further divided into three steps. (2a) First, a rigid body simulated annealing was performed to optimize orientations of the interacting partners. (2b) Then, the system underwent 1000 molecular dynamics (MD) steps from 500K to 50K, with a 2 fs timestep, in which ligand and protein side chain movement was allowed. (2c) Finally, flexibility was introduced to both protein side chains and backbone, besides the ligand. 1000 MD steps (with a 2 fs timestep) were performed with a stepwise temperature decrement from 300K to 50K. It should be noted that flexibility was only applied to the ligand and protein residues within a range of 5 Å. (3) The final stage of the docking protocol was a refinement in explicit water. Namely, three MD-based steps (with a 2 fs timestep) were carried out: (3a) A heating phase of 100 MD steps from 100K to 200K and to 300K, (3b) 1250MD steps at a constant temperature of 300K, and (3c) a cooling down phase of 500 MD steps to a final temperature of 100K. In stage (3), both ligand and protein were fully flexible, with the exception of protein backbone atoms. The HADDOCK score settings recommended for small molecule docking were used across the whole protocol [121, 122]. The obtained 200 docking poses were clustered based on their positional protein-ligand interface root-mean-square deviation (iL-RMSD) with a cutoff of 1.0 Å. The Haddock score of each cluster was calculated as the average of the top four structures, as done by the Haddock webserver [87, 119], using the equation: HADDOCK score = 1.0 Evdw + 0.1 Eelec + 1.0 Edesol + 0.1 Eair, where Evdw and Eelec are the van der Waals and electrostatic intermolecular energies, respectively, Edesolv is the desolvation energy and Eair is the distance restraints energy; the weights of the different terms were parameterized for scoring of protein-ligand complexes [121, 122]. Further analysis was performed for the top cluster (i.e. the one with the best average Haddock score) and, if present, also for other clusters with Haddock scores within the standard deviation of the top cluster. For proteins with more than one reactive cysteine, we performed independent dockings for each of the Cys residues. This approximation is valid provided that these cysteines are far enough apart that they (or their ACR covalent adducts) cannot interact with each other. However, in the case of one of the target proteins, creatine kinase (Table 4), the two Cys are within 7.1 Å (C_α-C_α distance) and thus we also considered the possibility that the two Cys could be targeted simultaneously by ACR (see 4.1.2). In this case, binding of two ligand molecules at the same time was modeled using the multibody docking approach [123] implemented in HADDOCK. Namely, a so-called molecule interaction matrix is used to define partners that interact with each other. In particular, we defined the subsequent interacting pairs: protein-ACR molecule 1, protein-ACR molecule two and ACR molecule 1-ACR molecule 2." [1]

4.1.3 Ligand-Receptor Interactions

"*Hydrophobic interactions were investigated using VMD [124] (version 1.9.3.) and in-house scripts. Namely, such contacts were defined as interactions between either of the two carbon atoms of the ligand and "apolar" protein carbon atoms (i.e. with CHARMM-based point charges below 0.15 electrons) located within the distance cutoff of 4.0 Å. The hydrogen bond (HB) interactions with both the amide and the*

carbonyl group of the ligand were analyzed with ProLIF [125] (version 1.0.0). The donor-acceptor distance cutoff was set to 4.1Å and the donor-hydrogen-acceptor angle tolerance to at least 100°. Each docking was analyzed separately, regardless of whether the reactive cysteines belong to the same protein or different protein targets. The protein-ligand interaction frequency is calculated as the percentage of poses belonging to the top (best scored) cluster that exhibit such interaction. When additional clusters with HADDOCK scores within the standard deviation of the top cluster are present, their poses were also included in the analysis, but a weighted average of the interaction frequencies was calculated, based on the size of each of the clusters analyzed. 2D representations of the protein-ligand interactions for each of the docking clusters considered were generated using ProLIF [125] (version 1.0.0) and are shown in chapter 7. The covalent docking approach used here aims at predicting the most likely configuration or binding pose of the Cys-acrylamide adduct. However, the Michael addition reaction starts with the deprotonation of the reactive Cys. Hydrogen bonding to the Cys sulfur atom is crucial for thiolate formation and stabilization of the transition state of the subsequent reaction [126]. Moreover, the Michael addition reaction involves an enolate-type intermediate in which the ligand oxygen atom acquires negative charge (see step 2 in Figure 3) and thus hydrogen bonding or a positively charged microenvironment could stabilize this intermediate, facilitating adduct formation [127, 128]. Hence, we additionally analyzed protein residues either near the reactive Cys (in the initial X-ray structure of the protein target, i.e. before the Michael addition reaction occurs) or ligand (in the best structure of the top docking cluster, i.e. after covalent adduct formation). First, we checked H-bonded protein residues. These could act as potential proton acceptors to deprotonate the Cys sulfur atom or may stabilize the intermediate and/or product of the Michael addition reaction. Next, we visually inspected other nearby protein residues in the binding cavity that could have favorable, yet longer-range, electrostatic effects on thiolate or adduct formation. In particular, we focused on His, Asp, Glu, Arg and Lys. Histidine is one of the most interesting residues regarding acid-base properties, since its intrinsic pK_a of ~6 is the closest value to the physiological pH of around 7, as well as to the intrinsic pK_a of Cys of ~8.6. Hence, the imidazole side chain can be either singly or doubly protonated and thus serve as both proton acceptor and as positively charged residue stabilizing the thiolate formed upon Cys deprotonation. Aspartic and glutamic acids have lower intrinsic pK_a values (~4.0 and ~4.4); however, their pK_a can shift to higher values depending on their microenvironment. Hence, Asp and Glu can also be potentially responsible for Cys deprotonation in some cases. Instead, the positively charged Lys and Arg are expected to stabilize the negatively charged thiolate (or the enolate-type intermediate formed during the Michael addition reaction), either by forming a salt bridge or electrostatically. The results of this analysis of the Cys microenvironment are presented in Table S1" [1]

4.2 Results

4.2.1 Dataset of experimentally validated acrylamide protein targets

As a first step, a literature search was performed in order to compile a protein dataset (for further details see Figure 11). This resulted in 19 protein targets from which eight proteins had experimental validated attachment sites of either ACR or closely related electrophilic compounds such as N-ethylmaleimide (NEM). For other proteins such

information was missing and based on physicochemical, conservation and functional data potential modification sites were ranked on their likelihood to react with ACR.

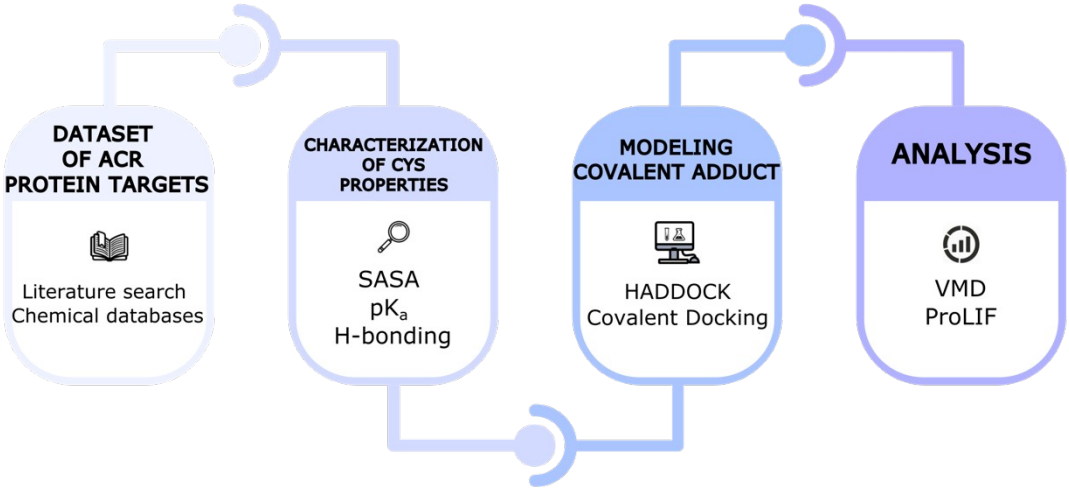


Figure 11. Schematic representation of the computational workflow used in this study. Reproduced from Mueller et al.

4.2.2 Proteins with experimentally verified reactive cysteine

"The effects of ACR modification on the 19 proteins in our dataset were further investigated using covalent docking. Considering that some of the ACR protein targets have more than one potential reactive Cys (Table 4), 34 covalent docking calculations were performed, following the protocol described in section 4.1.2. Below we present the results for the eight ACR protein targets for which the reactive Cys is known (see Table 3), following the alphabetical order of the protein name. […] In all cases, we combined our computational results with previously published experimental data to surmise the possible functional consequences of ACR modification." [1]

Table 3. Covalent docking results for the subset of acrylamide protein targets with experimentally known reactive cysteine. For each considered Cys, the Haddock score and size of the top docking cluster are shown. The latter corresponds to the number of docking poses belonging to the top cluster upon clustering of the total 200 poses.

Protein name		**Reactive Cys**	**Score (a.u.)**	**Size**
Albumin		C34	-32.3	63
Creatine Kinase		C74	-21.3	12
		C141	-31.7	63
		C146	-28.2	22
		C254	-23.3	34
		C283	-32.8	61
Dopamine D3R Receptor		C114	-28.6	130
Dopamine Transporter	outward	C342	-2.0	21
	inward	C342	-11.9	79
		C114	-15.1	87
Glyceraldehyde-3-phosphate dehydrogenase		C152	-12.3	163
		C156	46.1	1
		C247	9.9	14
Hemoglobin		C93	-13.9	74
		C104	-3.3	75
NEM-sensitive factor		C264	-73.4	85
Vesicular proton ATPase		C254	-0.1	110

Table 4. Acrylamide protein targets compiled in this study. Protein structures and cysteine residues considered for the covalent docking calculations are listed. Targets for which homology models of the human proteins had to be generated are indicated by HM and the template structure used between parentheses; further details are provided in sections (sections 3.2.4 and 3.2.7). Physicochemical characterization and location of the candidate cysteines are also included; n.a. indicates cysteines for which no functionally relevant location was identified. Residue numbers highlighted in bold indicate the main reactive cysteine in protein targets with experimental data, whereas an asterisk marks residues predicted to be targeted by acrylamide based on the results of our covalent docking calculations. Reproduced from [1].

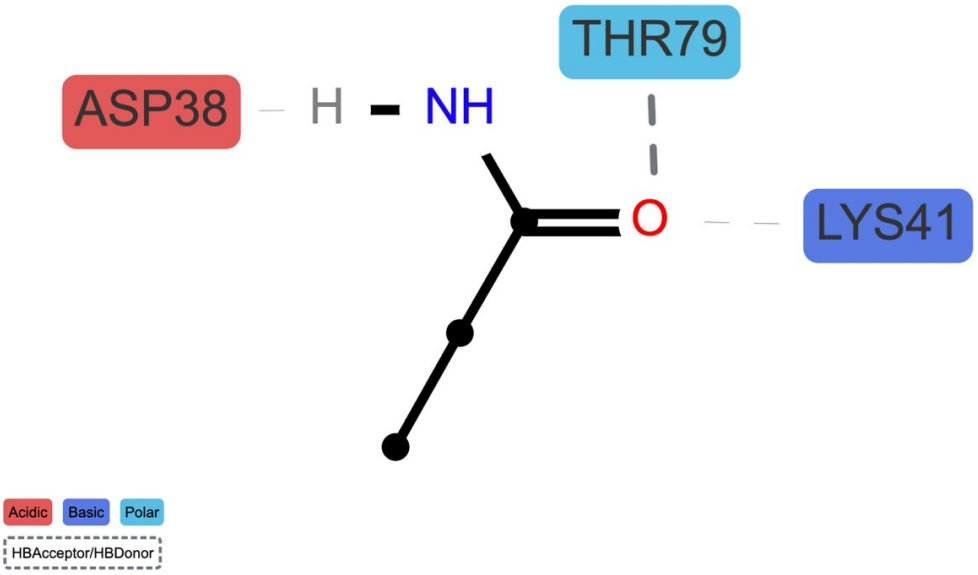
#	Protein name	PDB Code	Resolution (Å)	Cysteine	SASA (A^2)	pK_a	Cys location	References
(1)	Albumin	6HSC	1.9	**34**	8.6	10.2	allosteric site	[129-131]
(2)	Alcohol Dehydrogenase	1U3W	1.45	170 240*	0.0 30.8	>12 >12	n.a. protein surface	[132, 133]
(3)	Aldolase	1QO5	2.5	134* 239* 268 289*	0.0 16.8 2.0 7.6	>12 >12 >12 10.8	n.a. protein surface n.a. protein surface	[134]
(4)	Creatine Kinase	3B6R	2.0	74 141 146 254 **283**	2.6 11.1 2.6 3.6 37.2	>12 >12 >12 >12 ~9	n.a. protein surface n.a. protein surface active site	[135-138]
(5)	Dopamine Receptor	3PBL	2.89	**114**	18.1	>12	binding site	[14]
(6)	Dopamine Transporter (outward) Dopamine Transport (inward)	HM (6M2R) HM (6DZZ)	2.8 3.6	**342** 135 **342** 135	3.7 0.8 3.7 0.8	>12 >12 >12 11.3	intracellular loops	[139]
(7)	Enolase	2PSN	2.2	388 398*	7.7 36.0	>12 ~9	n.a. protein-protein interface	[140]
(8)	Estrogen Receptor	1ERE	3.1	381* 530*	37.5 68.5	>12 9.9	Cys-rich region	[141, 142]
(9)	Glyceraldehyde-3-phosphate dehydrogenase	4WNC	1.99	**152** 156 247	32.8 0.2 0.0	6.6 >12 >12	active site solvent exposed PTM site	[143]

#	Protein name	PDB Code	Resolution (Å)	Cysteine	SASA (A^2)	pK_a	Cys location	References
(10)	Hemoglobin	6KA9	1.4	**93** 104	12.2 5.1	>12 >12	protein-protein interface	[144]
(11)	Immunoglobulin G1 H Nie	6ARP	1.7	395*	-[a]	-[a]	n.a.	[145]
(12)	Immunoglobulin kappa light chain	6N35	1.75	134*	-[a]	-[a]	n.a.	[145]
(13)	Kinesin KIF1C	5WDH	2.25	663*	33.4	>12	protein surface	[146]
(14)	Kinesin KIF2C	4UBF	3.0	260* 287*	9.4 56.4	>12 9.3	dimeric interface dimeric interface	[146]
(15)	NEM-sensitive factor	HM (3J94)	4.2	**264**	85.2	7.7	walker A motif	[139]
(16)	Sex Hormone-Binding Globulin	1KDM	2.35	164* 188*	-[a]	-[a]	protein surface protein surface	[141]
(17)	Topoisomerase IIa (ATPase domain)	1ZXM	1.87	170	0.0	>12	near active site	[147]
(18)	Topoisomerase IIa (Toprim domain)	4FM9	2.90	997*	0.9	11.1	protein-DNA interface	[147]
(19)	Vesicular proton ATPase	6WM2	3.1	**254**	11.0	10.3	walker A motif	[139]

[a]Cysteines involved in a disulfide bridge; hence, pK_a and solvent-accessible surface area (SASA) values could not be calculated.

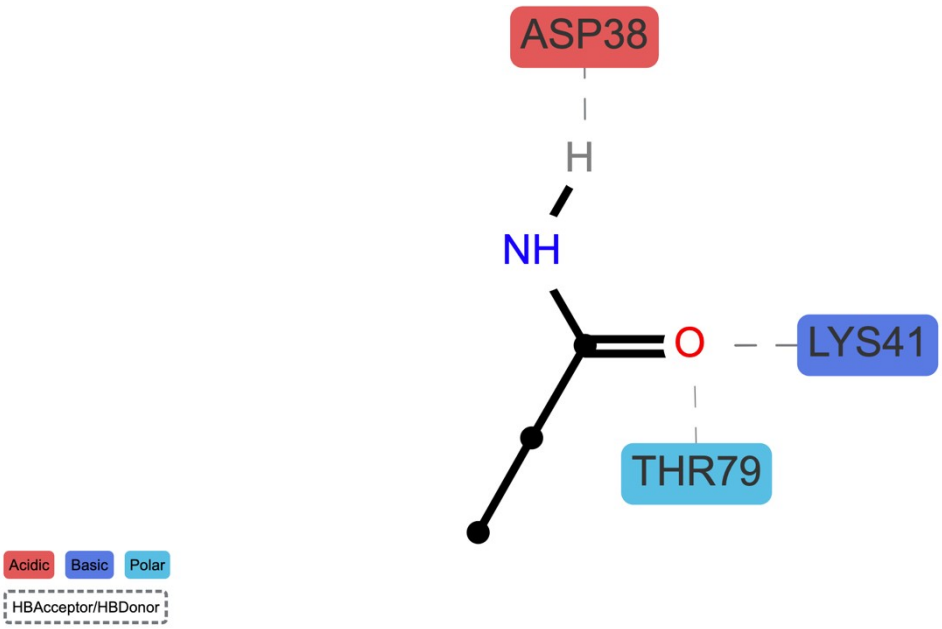
4.2.2.1 Human Serum Albumin (HSA)

"*Albumin is a plasma protein able to bind chemically diverse ligands, from hemin and fatty acids to drugs, acting as their plasma carrier/transporter [148]. Liquid chromatography–tandem mass spectrometry (LC-tandem MS) experiments have shown that C34 binds covalently acrylamide [129, 130]. HSA contains 35 cysteine residues and all form disulfide bridges except C34 [149]. This single free Cys is solvent exposed, with a SASA value of 8.6Å^2, and has a calculated pK_a value of 10.2 (see Table 4). This is in line with spectroscopic measurements showing HSA Cys34 to be more acidic than a normal Cys, with a pK_a around 7 [150]. The difference between the computational and experimental pK_a values can be ascribed to the known limitations of computational pK_a predictors when dealing with Cys residues [37], as well as uncertainties in the experimental estimation of pK_a values using spectroscopic methods. For instance, the pK_a of Cys34 changes by 1.5 pH units depending on the ionic strength of the buffer used [150]. Covalent docking of ACR to C34 resulted in two similar clusters in terms of both score (−31.3 and −32.3 a.u., respectively) and cluster size (69 and 63, see Table 3). Mapping of C34 onto the HSA structure also revealed that this cysteine has two putative proton acceptors in the vicinity (H39 and D38) that can deprotonate the thiol group, as well as a positively charged residue (K41) that could stabilize the transition state and/or product of that reaction (see Figure 15). Comparison with available functional information [148, 151] suggests that acrylamide covalent binding to Cys34 might affect the drug binding properties of albumin. In particular, infrared spectroscopy has shown that Cys34 is linked allosterically with Sudlow's site I for anesthetics such as halothane, propofol and chloroform [151]. Hence, formation of a covalent adduct at Cys34 can be transmitted to this site and modulate anesthetic binding. […].*
Covalent docking for C34 of albumin resulted in two main clusters (numbers 2 and 1) with Haddock scores within standard deviation of each other (see Supplementary Material 2)" [1] The Supplementary Material 2 file is available at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1125871/full#supplementary-material



Albumin (C34) – Cluster 1			
residue	number	interaction	occurrence
Asp	38	HB acceptor	0.04
Lys	41	HB donor	0.04
Thr	79	HB donor	0.48

Figure 12. Modeled covalent adduct between acrylamide (ACR) and the target protein. (Top) Schematic representation of acrylamide and the hydrogen bonds (HBs) formed with the surrounding binding site residues. Residues are colored according to their physicochemical properties, as shown in the figure legend, and the corresponding HB frequency is represented through the width of the dashed line. (Bottom) Table indicating the target protein, reactive Cys and docking cluster considered, together with the list of detected protein-ligand HBs and their respective frequency. Reproduced from [1].



Albumin (C34) – Cluster 2			
residue	number	interaction	occurrence
Asp	38	HB acceptor	0.14
Lys	41	HB donor	0.19
Thr	79	HB donor	0.10

Figure 13. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used. Reproduced from [1].

4.2.2.2 Creatine Kinase (CK)

"*Creatine kinase is an enzyme responsible for converting creatine to phosphocreatine reversibly using adenosine triphosphate (ATP). CK is inhibited by acrylamide and such inhibition exhibits a biphasic behavior with respect to acrylamide concentration [138], suggesting than more than one Cys residue within CK might be modified. C283 has been proposed as the primary site of ACR modification in CK. Based on site-directed mutagenesis, C283 was shown to be essential for enzymatic activity [136]. Furthermore, experimental, studies indicated that C283 has a pK_a around 5.7 and thus this cysteine can be present as thiolate. This is probably required to constrain the position of the guanidinium group of the creatine substrate [152]. The low pK_a of C283 and its role in CK enzymatic activity makes it a good candidate for the main reactive Cys targeted by acrylamide. In contrast, the secondary site of ACR modification is unclear. A combined experimental and computational study [36] suggested that acrylamide can bind to C283, as well as the nearby C74, but the results were not conclusive. Therefore, we performed docking for all five solvent exposed cysteines in CK (C74, C141, C146, C254 and C283; see Table 3). C283 is the most solvent exposed cysteine (37.2Å2) and with the lowest predicted pK_a value (around 9, see Table 4). Although the calculated pK_a of C283 (~ 9) differs from the experimentally measured value (5.7), we ascribed such difference to the known limitations of computational (implicit solvent-based) predictors when estimating the pK_a values of Cys residues, with RMSDs between 3.41 and 4.72 pK_a units [37]. Given this uncertainty, we decided to use the calculated pK_a values only to rank by relative acidity the Cys residues within the same protein, i.e. C283 is the most acidic Cys in CK. In addition, the C283 docking yields the best score (−32.8 a.u.), as shown in Table 3. Furthermore, C283 is located in the catalytic site of CK, whereby nearby residues, such as R96 (distance of sulfur atom to ζ-carbon atom of 7.48Å), R132 (10.71Å) and R236 (10.70Å), can electrostatically stabilize the enolate intermediate of the Michael addition reaction (see Table S2). Instead, docking at C74, previously proposed as acrylamide binding site [36], gives a less favorable docking score (−21.3 a.u.). Together with most of the C74 poses showing distance values between the ligand Cβ atom and sulfur atom outside the defined covalent bond range, this suggests that modification of C283 is preferred over binding to C74. Due to the proximity of C74 to C283, we also explored the possibility of two acrylamide molecules binding simultaneously to both C283 and C74, using a multibody docking approach [123]. The resulting docking poses indicate that adduct formation with one acrylamide molecule already occupies fully the pocket lined by C283 and C74 and thus will preclude binding of a second acrylamide molecule (see Figure 14). Thus, C74 is unlikely to be modified by ACR, either alone or in combination with C283. In contrast, docking to other cysteine residues revealed more suitable candidate for the secondary site of ACR modification in CK. The results for C141 and C146 yielded docking values closer to those of C283 (see Table 3), suggesting that modification of these two cysteines by acrylamide might be possible. Out of these two cysteines, C141 has a slightly more favorable docking score (−31.7 a.u.) than C146 (−28.2 a.u.), as well as a higher solvent exposed surface area (11.1 compared to 2.6Å2), suggesting a slightly higher preference of ACR for C141 over C146. Additionally, C141 has two nearby residues (H145 and E150) that could facilitate thiolate and/or adduct formation (see Figure S5), whereas C146 is hydrogen bonded to P143 (see Figure S6 to Figure S8). Based on our covalent docking results and the biphasic time dependent inactivation of CK observed in enzymatic assays [36], we propose a molecular model in which ACR*

modification of C283 (Figure 15B) occurs first and is the primary site responsible for enzyme inactivation. Adduct formation at C283, located in the enzyme active site [152] , will hinder creatine binding. At longer times, C141 might also be modified by ACR, further contributing to enzyme inactivation by thiol depletion [137, 153]." [1]

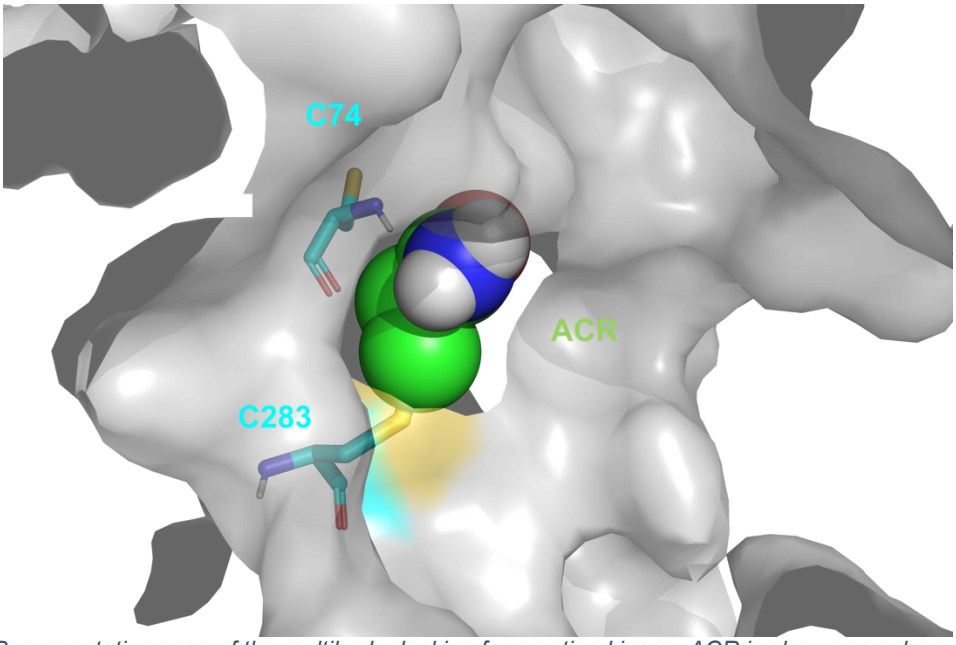


Figure 14. Representative pose of the multibody docking for creatine kinase. ACR is shown as spheres, C283 and C74 in stick representation and the surface of CK is shown in grey. A molecule of ACR is covalently bound to C283; as a result, C74 is no longer accessible to the second ACR molecule coming from the solution. Indeed, the covalent docking calculation placed the second molecule of ACR (not shown) outside the cavity where C74 and C283 are located. Reproduced from [1]

4.2.2.3 Dopamine D3 receptor (D3R)

"*Acrylamide exposure has been shown to result in decreasing dopamine concentrations by altering postsynaptic dopamine receptors [14]. Site-directed mutagenesis data showed that electrophilic compounds, such as NEM, blocked ligand binding to the dopamine D3 receptor (D3R) by modifying C114 [154]. Furthermore, functional data compiled in GPCRdb [155] indicates that C114 is involved in both ligand binding and receptor activation. Taken together the C114 reactivity and functional data, we considered C114 as the most likely candidate for acrylamide modification. Modeling of the covalent C114-ACR adduct further revealed how ACR modification can impair D3R signaling. The ligand interacts with D110 (Figure 15C); this aspartate is essential for ligand binding in aminergic GPCRs [156], such as D3R. Together with the aforementioned functional roles of C114 [155], this indicates that formation of the ACR covalent adduct will hinder ligand binding and/or impair receptor activation. Moreover, Cys at this position (3.36, following the Ballesteros-Weinstein generalized numbering for class A GPCRs) is conserved across dopamine receptors D2, D3 and D4. Given the role of these receptors in dopaminergic neurotransmission, ACR modification of Cys(3.36) might be one of the molecular mechanisms by which ACR intoxication mimics Parkinsonian symptoms.*" [1]

4.2.2.4 Dopamine Transporter (DAT)

"*Dopamine transporters are integral membrane proteins responsible for regulating dopamine neurotransmitter concentrations at the synaptic cleft [157]. Chemicals such as peroxynitrite and 2-aminoethyl methanethiosulfonate (MTSEA), which have in common the potential to modify cysteine sulfhydryls, are known to inhibit DAT [158]. Mutagenesis data has also shown that oxidation of C342 causes a decrease in DAT activity [158]. Furthermore, Cys modification is enhanced if the transporter is in the inward-facing state [159]. To understand this differential reactivity of the two conformational states of DAT, we performed two covalent dockings for C342, using DAT structures in either outward- and inward-facing conformations (hereafter, OF and IF). Since experimental structural information for human DAT is missing, we generated homology models of the two transporter conformations. The templates used for the OF and IF models were the Drosophila melanogaster DAT (PDB code 6M2R) [160] and the human serotonin transporter (PDB code 6DZZ) [161], respectively. The target-template sequence identities are 56.2% (OF) and 52.4% (IF); thus, the models are expected to be medium-to-high quality [80-82]. We further assessed the quality of the models by calculating their Ramachandran plots (Figure S1) and QMEANbrane local quality values (Figure S2). The percentage of residues in favored/allowed regions is 93.3%/98.7% (OF) and 95.6%/99.4% (IF), whereas the predicted local quality scores are above 0.7 (except for loop regions or not resolved in the template structures). Thus, these two quality assessments support the reliability of the DAT homology models used here. SASA calculations show that C342 is more solvent exposed in the IF model, with SASA values four-fold larger than the OF model (see Table 4). Therefore, cysteine accessibility seems to play a role in the observed higher reactivity of ACR with the IF state [159]. Our covalent docking results (see Table 3) further support the enhanced ACR modification in the IF state. The top cluster for the IF model (Figure 15D) had a more favorable score of −15.1 a.u. than the one (−2.0 a.u.) for the OF model (Figure 15E). Such preferential binding of acrylamide to C342 in the IF state could alter the conformational transition between the two states crucial for dopamine transport, resulting in altered neurotransmitter concentrations. Considering the link between DAT and Parkinson's disease, it is tempting to suggest that this might be responsible, at least in part, for the PD-like symptoms of acrylamide neurotoxicity [162, 163]. Besides C342, we also performed covalent docking for C315. Mutagenesis experiments and transport assays upon treatment with sulfhydryl reagents have shown that C342 is the main modification site responsible for transport inhibition in wild-type DAT [159]. However, C315 can also be modified and have a minor contribution to transport inhibition in a DAT C90A/C306A/C319F/C342A mutant construct lacking C342 [164], with higher C135 accessibility in the IF state [159, 165]. The docking score for this alternative C135 site (−11.9 a.u.) in the IF state is less favorable than for the main C342 site (−15.1 a.u.), further supporting our proposal that the docking score can help discriminate the most reactive Cys within a given protein target.*" [1]

4.2.2.5 Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)

"*GAPDH is a housekeeping enzyme involved in both glycolysis, as well as apoptotic cell signaling. Multiple studies, both experimental and computational, have shown that acrylamide can covalently modify GAPDH, inhibiting enzymatic activity [166]. Moreover, such enzymatic inactivation is concentration- and time-dependent, as well as pH sensitive [143]. C152 has been identified as the most reactive cysteine compared to two other solvent exposed cysteine residues, C156 and C247 [143]. At small concentrations of acrylamide, almost only C152 is modified. As C152 is essential for GAPDH catalysis by acting as nucleophile, formation of the Michael adduct will result in enzyme inhibition [143]. However, at higher concentrations, ACR adducts with C156 and C247 are also formed and have been shown to further contribute to enzyme inhibition. To identify features that could explain this differential reactivity, we performed covalent docking for each of the aforementioned Cys residues (see Table 3 ; Figure 15F; Figure S16 and Figure S17). The top docking cluster for residue C152 had score of −12.3 a.u., significantly more favourable than those for C156 and C247 (46.1 and 9.9 a.u., respectively). Moreover, the last two dockings showed poses with C-S distances outside the covalent bond range. Therefore, modeling of the Michael adduct indicates that C152 is the primary binding site of ACR in GAPDH, in agreement with experiments [143]. In addition, both C156 and C247 had a calculated pK_a above 12, and thus are less likely to become deprotonated. Instead, the calculated pK_a value of 6.6 for C152 suggests that the sulfur atom can be present, at least partially, as a thiolate anion. Besides, C152 forms a hydrogen bond with residue H179 (Figure 15F); this could help deprotonate C152. Indeed, H179 activates the thiol group during enzymatic catalysis [167]. Moreover, the resulting doubly protonated H179 and its respective positive charge could electrostatically stabilize the enolate-type intermediate of the Michael addition reaction. Furthermore, the SASA values further support C152 as the most reactive residue, since its predicted accessibility is two orders of magnitude higher than C156 and C247 (see Table 3). Altogether, the computational results are in agreement with the experimental evidence that C152 is the primary site of acrylamide modification of GAPDH. Moreover, the higher reactivity of C152 with respect to C156 and C247 seems to correlate with the more favorable score obtained for the first cysteine.*" [1]

4.2.2.6 Hemoglobin (Hb)

"*Hemoglobin is a heme-containing protein responsible for oxygen transport from lungs to other tissues. Structurally, Hb is a heterotetramer [168] formed by two α and two β subunits that assemble as dimer of dimers ($\alpha_1\beta_1$ and $\alpha_2\beta_2$, respectively). Mass spectrometry showed that C93 within the β chains and C104 in the α chains are modified by acrylamide, with C93 being the most reactive site [144]. Hence, measuring the levels of ACR-modified Hb in plasma can be used to monitor acrylamide exposure [144, 169]. The two aforementioned reactive cysteines have a predicted pK_a value above 12. However, C93 is more solvent-exposed compared to C104 (see Table 4), suggesting that C93 is more accessible to acrylamide. This is in line with the covalent docking results obtained here (Table 3). The top docking cluster for C93 shows a significantly better docking score (−13.9 a.u.) compared to −3.3 a.u. for C104. Indeed, docking simulations with C104 did not result in a properly formed covalent bond between Cys and acrylamide, which could explain the less favorable docking scores compared to C93. Moreover, C93 has three nearby potential proton acceptors, H346,*

H97 and D294 (Figure 15G), belonging to the same β_2 subunit. C93 is also located near K40 of the adjacent α_1 chain, which stabilizes the adduct by forming a hydrogen bond with the ligand oxygen atom (see Figure S18-22). In addition, the nearby positively charged side chain could help stabilize the transient negative charge developed on the ligand oxygen atom during the nucleophilic attack. Instead, C104 only has a single nearby residue, H103, which could deprotonate the thiol group or stabilize the adduct. Taken together, our computational analysis suggests that C93 in the β subunit should be the primary site for acrylamide adduct formation in Hb, whereas binding to C104 in the α subunit is likely to occur only at higher acrylamide concentrations or longer times, in line with the experimentally observed reactivity [144] . Moreover, the location of C93 at the interface between the α1 and β2 suggests that covalent modification of this cysteine by acrylamide could affect Hb function. Oxygen binding to Hb induces changes within this quartenary structure, i.e. a conformational transition from deoxyhemoglobin (T-state) to oxyhemoglobin (R-state). The largest movement occurs between the α1C-helix and the β_2FG corner and a smaller change takes place between the α_1FG corner and the β_2C-helix [170]. Thus, the aforementioned $\alpha_1\beta_2$ intersubunit location of C93 might alter the transition from the T to R state triggered by oxygen binding to Hb and/or its cooperativity mechanism. We propose here that the effect of ACR modification on C93 could be tested experimentally, for instance, by measuring oxygen saturation at different oxygen partial pressures, after incubation with acrylamide. In the absence of such experimental validation, the functional impact of ACR modification is partially supported by a previous experimental study showing that covalent modification of C93 and C104 by other (larger) organic compounds prevented formation of the Hb tetramer [171].“ [1]

4.2.2.7 NEM-sensitive (NSF)

“*N-ethylmaleimide(NEM)-sensitive factor is a homohexameric ATPase [172]. In the presynaptic neuron, NSF, together with SNARE proteins, is involved in fusion of neurotransmitter-loaded vesicles with the cell membrane and vesicle recycling and thus is key for synaptic neurotransmission [173]. Previous studies indicated that thiol reagents (e.g., NEM or NO) inhibit NSF [174]. Mass spectrometry data showed that acrylamide modifies cysteine sulfhydryls, thereby altering the ATPase activity of NSF [175]. Moreover, experimental evidence indicates that C264 is a critical residue for NSF function [176]. This cysteine is located in a so-called Walker A motif, important to ATP binding and thus for NSF ATPase function. Due to the lack of experimental structural information for human NSF, we generated a homology model based on a Cricetulus griseus template (PDB code 3J94) [177], which has sequence identity of 98.4%. Therefore, the model is expected to be high quality [80-82]. Indeed, its Ramanchandran plot (Figure S1) shows 90.4%/96.5% residues in favored/ allowed regions (comparable to the 93.0%/98.0% values, respectively, for the template structure, solved at 4.20Å resolution). Additionally, the QMEANDisCo local quality values (Figure S2) are mostly above 0.7 (except for loop regions or not resolved in the template structures), further supporting the quality of the model. Our computational analysis using this homology model showed that, among the potential attachment points of ACR, C264 is both the most susceptible to deprotonation and the most accessible to acrylamide, with a pK_a value of 7.7 and a SASA value of 85.2Å2. The docking results further support the modification of C264 by acrylamide (Figure 15 H). Structural inspection of the docking poses revealed that ACR would partially block*

access to the ATP binding site, thus hindering ATP binding and decreasing NSF activity." [1]

4.2.2.8 Vesicular proton ATPase (v-ATPase)

"*Filling of the synaptic vesicles with neurotransmitters relies on the proton gradient created by the vesicular proton ATPases (v-ATPases) using ATP. N-ethylmaleimide (NEM), a sulfhydryl reagent similar to ACR, reduces H^+ uptake, as well as decreases v-ATPase activity [178]. The modification site responsible for v-ATPase inhibition by sulfhydryl reagents [139, 179] has been proposed to be C254, which is located in a loop segment of the v-ATPase catalytic subunit, corresponding to the so-called Walker A (GAFGCGKT) motif coordinating ATP binding and hydrolysis. Physicochemical characterization of C254 revealed that this cysteine has a pK_a value of 10.3 and a SASA value of 11.0 Å, further supporting this particular cysteine as ACR target site. Thus, we performed the corresponding Haddock calculation for C254. The covalent docking poses obtained here (Figure 15I) show that the ligand is placed at the entrance of the active site and thus can hinder ATP binding. Nonetheless, the loop where C254 is located exhibits large rearrangements during the conformational cycle of v-ATPase (see Figure S24) and such structural changes cannot be modeled with covalent docking. Hence, we integrated additional experimental data for validation. Our hypothesis is indirectly supported by the experimental observation that NEM inactivation of v-ATPase is associated with exposure of a single cysteine residue that can be protected by incubation with nucleotides [180]. Moreover, another experimental study showed that C254 modification, either through formation of a disulfide bridge with C532 or through adduct formation with NEM, causes inactivation of the v-ATPase [181]. Therefore, we surmise that modification of C254 by ACR might have a similar inhibitory effect.*" [1]

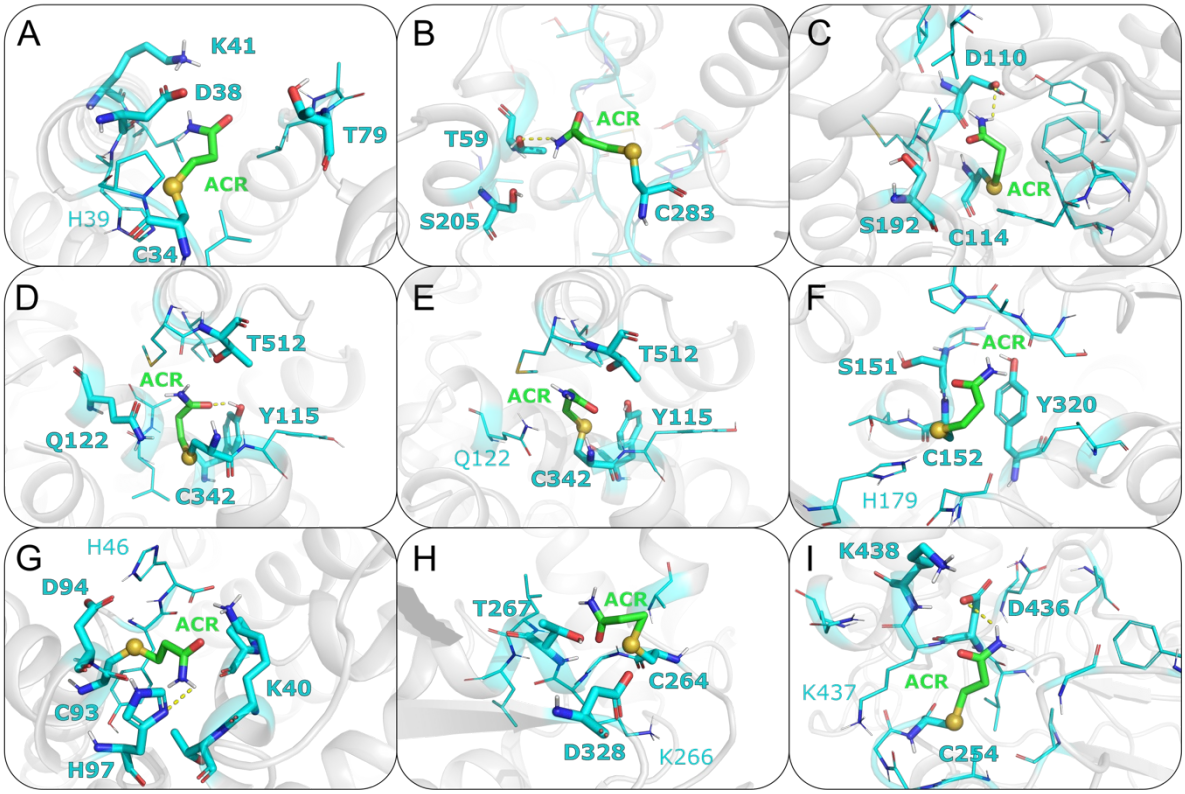


Figure 15. Representative covalent binding poses of ACR for each of the protein targets discussed in the main text. Acrylamide and its surrounding residues are represented as sticks, with carbon atoms colored in green and cyan, respectively. The sulfur atom between the reactive cysteine residue and the adduct is shown as a sphere. Residues forming hydrogen bonds (HBs) with ACR are displayed with thicker sticks and with bold labels. HBs present in more than 60% of the docking poses are shown with a dashed line. Nearby residues (i.e. within 5 Å) potentially favoring the Michael addition reaction are shown with thinner lines, with positively charged residues and putative proton acceptors labelled. (A) Albumin; (B) Creatine kinase; (C) Dopamine D3 receptor; (D) Dopamine transporter (inward conformation); (E) Dopamine transporter (outward conformation); (F) Glyceraldehyde-3-phosphate dehydrogenase; (G) Hemoglobin; (H) NEM-sensitive factor; (I) Vesicular proton ATPase. Reproduced from [1].

4.2.3 Proteins without reactive cysteine experimental information

"Out of the 19 proteins in our dataset (Table 4), the specific Cys targeted by acrylamide is not known for eleven […]. In order to narrow down the most promising candidates to be the reactive Cys, we first checked physicochemical properties and microenvironment, as well as conservation and post-translational modifications, for all the cysteines of the respective protein target." [1]

The results of the covalent docking calculations allowed us to narrow down the most likely Cys candidates for ACR modification. "*Based on our observation that higher Cys reactivity against ACR turned out to correlate with more favorable docking scores, we suggest that"* the Cys residues marked with an asterisk in Table 4 are the most likely sites for ACR covalent adduct formation. Further details are provided in the Appendix and Supplementary Material 2 of the original publication.

4.2.4 Hydrogen Bonds Analysis

"*For each of the considered protein targets in (Table 4), we analyzed the hydrogen bond (H-bond) network between the ligand and its surrounding binding site residues. The representative clusters of each docking were pooled together and hydrogen bond frequencies were calculated as explained in section 4.1.3 shows ligand-protein H-bonds classified by type of amino acid. […].*" [1]

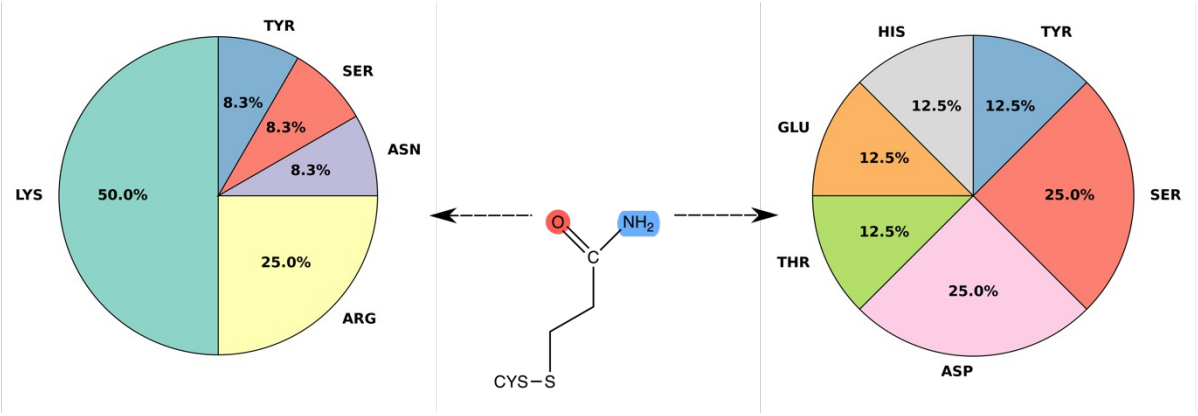


Figure 16. Frequency of hydrogen bonds. Pie chart showing the distribution of binding residues forming H-bonds with acrylamide. Only specific interactions with amino acid side chains and hydrogen bonds with frequency over 60% were considered. Reproduced from [1].

"[…]. Only H-bonds with amino acid sidechains and frequency over 60% are displayed; the use of other thresholds turned out not to significantly change the amino acid ranking. Interactions were grouped based on whether the H-bond was formed with either the carbonyl or the amino group of the ligand. Among the H-bonds formed with the carbonyl group, lysine is the residue with the highest frequency (50.0%), followed by arginine (25.0%). After these positively charged residues, asparagine, serine and tyrosine are next in the ranking, with a frequency of 8.3% each. The ligand amino group formed instead H-bonds with polar residues, such as serine (20%), as well as histidine, tyrosine, aspartic acid, threonine and glutamic acid (14.3% each). By grouping amino acids of similar chemical characteristics, a more clear picture emerges. In particular, positively charged amino acids, i.e. lysine and arginine, act as main H-bond donors to the carbonyl group of the covalent adduct and have a combined frequency of 75%. Therefore, our structure-based analysis suggests a preference of the ligand carbonyl group to interact with positively charged amino acids, in line with a previous sequence only-based analysis with other thiol-reactive electrophiles [182]. The ligand amino group shows a more diverse picture, in that we did not observe any amino acid preference to interact with the amino group. In particular, the ligand nitrogen does not prefer to interact with negatively charged amino acids (aspartate and glutamate, see Figure 16). We surmise that the specificity of the H-bonds with the carbonyl group with respect to the amino group is a remnant of the role of the H-bond donors in the mechanism of the Michael addition reaction. H-bonding to the carbonyl group does not only contribute to stabilize the resulting Michael adduct, but can also help to stabilize the negative charge developed in the enolate-type reaction intermediate." [1]

4.3 Conclusion

"*Cysteine residues are one of the least frequent (3.3%) proteinogenic amino acids [183]. Nevertheless, this residue is disproportionately involved in a variety of important protein functions due to the nucleophilic and redox properties of the thiol group (e.g catalytic or regulatory activities), as well as its ability to form disulfide bridges and bind metal ions. In this regard, proteins within the synaptic vesicle cycle are of interest, since they are considered as cysteine-rich proteins [184]. Acrylamide (ACR) is a toxicant that has been shown to affect protein function by reacting with Cys residues of several protein targets [13, 117, 139]. Such covalent adduct formation proceeds through a Michael addition mechanism and requires the Cys thiol group to be accessible to acrylamide, as well as deprotonated (i.e. thiolate, see Figure 3, step 2), so that it can act as nucleophile (Figure 3, step 3). Here, we first calculated Cys physicochemical properties to assess whether these intrinsic values would help pinpoint the most reactive cysteine(s) in a given ACR protein target. The solvent accessible surface area (SASA) is used as proxy of the Cys exposure to ligands. However, the SASA values can significantly vary depending on the resolution of the crystal structure (which can affect the accuracy of the position of the Cys side chain) and/or the functional state of the structure chosen (for the analysis for proteins undergoing large conformational changes during their functional cycle). For instance, the accessibility of C342 in DAT is larger in the IF state compared to OF (see Table 4). As for the prediction of pK_a values, popular software packages, such as H++ [185], PROPKA [186] or MCCE2 [187], are quite successful in predicting the acidity of aspartic or glutamic acid, but their performance for Cys residues is significantly lower [37]. For instance, C283 in CK has an experimentally validated pK_a value of ~5.6 [188] , yet both H++ and PROPKA estimated values of around 9 [152]. Unfortunately, both experimental and computational approaches show limitations at estimating pK_a values. […] Therefore, in this work we decided to take advantage of a low computational cost approach such as the H++ webserver [185] and then use the calculated pK_a values only to rank Cys residues within a given target protein by acidity, rather than considering the predicted absolute values. Moreover, as thiolate formation is favored in the presence of H-bonding and positively charged residues [189], we further inspected nearby residues in order to identify possible candidates to increase Cys acidity (such as Lys and Arg) or act as proton acceptors (such as His, Asp and Glu). However, this initial filtering of the candidate Cys residues based on physicochemical properties and microenvironment effects is not able to pinpoint a single ACR target site, but rather helps to discard the least likely Cys sites. This is not surprising, because such analysis has been used to predict Cys reactivity in general (see e.g. [190-192]), but does not take into account specific features related to acrylamide reactivity. In this work, we have compiled a list of protein targets associated to ACR toxicity and biomonitoring of ACR exposure and used a covalent docking approach to model the adduct formed upon Michael addition reaction of acrylamide with Cys residues of these proteins (Table 4, Table S1 and Table S2). First, we modeled the Cys-ACR covalent adduct for a set of protein targets for which the most reactive Cys is experimentally verified (i.e. those target proteins with a Cys highlighted in bold font in Table 4). This is the case for (i) C34 of albumin, (ii) C283 of creatine kinase, (iii) C114 of dopamine D3 receptor, (iv) C342 of dopamine transporter, (v) C152 of GAPDH, (vi) C93 of hemoglobin, (vii) C264 of NSF and (viii) C254 of v-ATPase. The covalent docking approach used here, based on scaling down the van der Waals parameters of the Cys sulfur atom and defining two distance restraints,*

allows to streamline the generation of structural models of the protein-ACR adducts, compared to more computationally intensive approaches, such as QM/MM [193-195]. Moreover, in the case of creatine kinase, DAT, GAPDH and hemoglobin, we also applied covalent docking to secondary Cys sites shown experimentally to be modified at increasing ACR concentrations or longer incubation times (see sections 4.2.2.1-4.2.2.8). We found that the docking score was able to discriminate between primary and secondary ACR sites of the aforementioned proteins, as well as the higher reactivity of C342 of the DAT in the IF state compared to the OF one (see section 4.2.2.4). Therefore, we surmised that covalent docking scores can help identify the main Cys reacting with ACR within a given protein and proceeded to apply the same approach to other protein targets associated to ACR toxicity for which the reactive Cys is unknown (see Table 4). Although the aforementioned traditional approaches [190, 191], based on solvent accessibility, pK_a prediction and H-bonding environment, can help pinpoint possible reactive Cys, the combination with covalent docking, as proposed here, can help better discriminate between different cysteines within the protein. Based on our observation that higher Cys reactivity against ACR turned out to correlate with more favorable docking scores, we suggest that the following Cys are modified by ACR (marked with an asterisk in Table 4): (i) C240 of alcohol dehydrogenase, (ii) C134, C239, C268 and C289 of aldolase, (iii), C134 of immunoglobulin kappa light chain, (iv) C398 of enolase, (v) C381 and C530 of estrogen receptor, (vi) C663 of kinesin KIF1C, (vii) C260 and C287 of kinesin KIF2C, (viii) C395 of immunoglobulin G1 H Nie, (ix) C997 of topoisomerase IIa, and (x) C164 and C188 of sex hormone-binding globulin. Nevertheless, further experiments are needed to validate our computational predictions. Analysis of the residues surrounding the ACR covalent adducts modeled in the present study shows that acrylamide binding sites are enriched in positively charged Arg and Lys residues (see Figure 16). Thus, our structure-based study confirms the hypothesis put forward in a previous sequence only-based study with other thiol-reactive electrophiles [182]. In addition, our work shows that residues such as His, Asp or Glu are often found in close proximity of ACR-modified Cys sites. We surmise that the particular amino acid composition of acrylamide binding sites may have catalytic effects on covalent adduct formation, in line with a previous study on model peptide systems [116, 118]. Cys deprotonation (step 2 in Figure 3) may be favored in the presence of positively charged Lys and Arg (which lower the Cys pK_a) and His/Asp/Glu residues (which can either further decrease the Cys pK_a by H-bonding or act as proton acceptors). Moreover, the Michael addition reaction (step 3 in Figure 3) proceeds via an enolate-type intermediate, which may be stabilized in the presence of Lys/Arg/His interacting with the negatively charged oxygen atom, thus decreasing the reaction energy barrier. […] The covalent protocol used here has two potential limitations. As for the SASA and pK_a calculations, the docking results might depend on the input protein structures. To minimize this dependency, we chose the highest resolution structure available for each protein target (to minimize possible errors in the accuracy of the position of the Cys side chain) and employed a fully flexible docking approach (to allow the protein environment to adjust to the presence of the ACR adduct). […] In conclusion, the application of covalent docking to ACR protein targets has provided molecular insights into the binding site where the covalent adduct is formed upon Michael addition. Such sites are enriched in Lys and Arg residues and additionally contain H-bonding residues that stabilize the covalent adduct. Docking scores emerge as a predictive tool to pinpoint Cys residues most likely to be modified by ACR within a given protein. Therefore, the computational workflow presented here (Figure 9) could serve to filter putative ACR

protein targets and candidate reactive Cys resulting from mass spectrometry-based proteomics studies and prioritize those that are more likely to be true positives. However, given the limitations of docking, such ranking should be used to guide follow-up validation studies. Mutagenesis and biochemical experiments would help to assess the impact of ACR on protein function and eventually (neuro)toxicity and computational simulations would provide further insights into the reaction mechanism of ACR modification, as done for other covalent inhibitors [193-195]. The computational workflow presented here is based on experimental structures from the Protein Data Bank [107, 108]. However, recently developed machine learning-based protein structure prediction algorithms [25, 66, 196, 197] could also be used to generate input protein structures. Moreover, here we performed the covalent docking calculations with HADDOCK [198], because its availability as a webserver [87, 119] and the minimal preparation of the protein structures required makes our workflow accessible to both new and experienced docking users. Nonetheless, processing the large number of possible candidate ACR protein targets and reactive Cys sites emerging from mass spectrometry-based proteomics will require automated covalent docking workflows, which could integrate either HADDOCK or other docking programs [199], such as GOLD [200] or Schrödinger [201]. Upon such covalent docking-based initial screening, the most promising protein targets and reactive Cys sites could be further filtered and analyzed using more computationally intensive QM/MM methods [1 93-195], such as empirical valence bond [202, 203]." [1]

5 Molecular insights into the neuroprotective effects of chlorogenic acids

5.1 Computational Details

5.1.1 Chlorogenic Acids and PPARα protein structures

Ligand and protein structures were processed in two different ways depending on the selected docking approach (see section 5.1.2):

1. *Template-based docking*. SMILES strings of chlorogenic acids and known agonists were taken either from PubChem [204] or generated with the JSME editor [205]. Respective 3D structures were afterwards generated using OpenEye OMEGA [206] (version 3.1.2.2) with default parameters. Namely, for each ligand up to 200 conformers were generated, depending on the number of rotatable bonds.
   As mentioned in section 3.4, protein structures may undergo conformational changes upon ligand binding. Thus, different protein structures were considered for dockings into different binding pockets of PPARα. In particular, we retrieved from the Protein Data Bank three protein structures of PPARα (PDB codes: 6KB3, 6LX5 and 6KBA [57]) in complex with agonists GW7647 (bound to the Center/Arm II/Arm III pocket), ciprofibrate (two molecules bound in Arm I and X pockets, respectively) and Wy14643 (two molecules bound in the Center and Arm X pockets), respectively. These three protein structures are almost identical (pairwise backbone RMSD ranges from 0.36 Å to 0.41 Å) and only differ in the rotameric state of F273 (at the boundary between the Center and Arm I pockets) and/or the position of the omega loop (see Figure 6 for its location). These conformational changes are needed for ligand binding to the Arm I and X pockets, respectively, as shown by these and other X-ray structures of PPARα [57].
   Further protein preparation was done via SWISS-MODEL [110] (i.e. adding missing residues) and MolProbity [113] (i.e. adding hydrogen atoms, assessing Asn/Gln/His flips and assignment of protonation states).

2. *Induced Fit Docking***.** Protein and ligand preparation were performed with tools from the Schrödinger's software suite (version 2021-1) [207-209].
   In this case, ligand 3D structures were generated with LigPrep [210], which also takes SMILES strings as input format. LigPrep uses Epik [211, 212] in order to assign ionization states to ligands in the chosen pH range of 7 ± 2.
   The protein structures were the same as for the template-based docking (PDB codes: 6KB3, 6LX5 and 6KBA [57]); however, here they were pre-processed with the Protein Preparation Wizard [213]. This tool prepares protein structures for modelling purposes, including correct bond order assignment, addition of missing hydrogen atoms and assignment of residue protonation states. Furthermore, an optimization of the hydrogen bond network is performed, which takes care of reorienting hydroxyl and thiol groups, water molecules (if present), amide groups of asparagine and glutamine residues, and the imidazole ring in histidine residues. A final minimization step of the protein structure ensures that the nearest local minimum is used as starting point for subsequent modeling

calculations. In the minimization step the protein is restrained, so that the refined protein structure is not allowed to exceed a RMSD value of 0.3Å compared to the input coordinates.

5.1.2 Docking Calculations

As mentioned in the previous two sections (see sections 5.1.1 and 3.4), two approaches were considered in order to predict possible docking binding poses of chlorogenic acids to PPARα.

1. *Template-based docking approach*
   This HADDOCK-based protocol incorporates information of known ligands of the target protein and their crystallographic binding poses (i.e. the template molecules) [122]. In simple words, candidate molecules to be docked are superimposed onto experimental coordinates of one or more template molecule(s), thus defining the spatial location of chemical moieties shared or similar between candidate and known molecules. As result of such an alignment, important protein-ligand interactions are preserved.
   The binding pocket of PPARα is quite large and depending on ligand size and shape different parts of the binding site can be occupied (see section 2.7.1 and Figure 7A). Therefore, the PPARα binding pocket was divided into three different pockets (Center/Arm II/Arm III, Arm I and Arm X) and a template was generated for each of them (Table 5).

Table 5. PPARα ligands used as templates. Chemical structures are displayed in Figure S56.

Pockets	**Ligands**[1]
Center/Arm II/Arm III	GW7647 and Pemafibrate
Arm I	Ciprofibrate and Fenofibrate
Arm X	Ciprofibrate and Fenofibrate or Wy14643

The template generation (based on one or more known ligands, see Table 5), as well as the candidate ligand superposition, was performed via OpenEye ROCS [214]. One or more template molecules define the 3D shape and spatial location of chemical groups (i.e. pharmacophore) in a particular pocket. The superposition aims to align new molecules so that both characteristics are satisfied, as assessed by the TanimotoCombo score [214] of OpenEye ROCS. The best superposition was selected based on the highest TanimotoCombo score and visual inspection. In particular, if a better alignment of the template and ligand carboxylate groups was observed, the corresponding ligand pose was chosen, despite a somewhat lower score. This filtering step was based on the fact that all crystal structures of PPARα bound to agonists containing a carboxylate group show this negatively charged moiety in the same position, next to S280, Y314, H440 and Y464. If no adequate superposition of compound and pharmacophore was found, i.e. the carboxylate moiety could not be placed properly or steric clashes with the protein structure occurred, subsequent docking steps were not performed.

[1] Crystal structures of ciprofibrate (6LX5), fenofibric acid (6LX4) and Wy14643 (6KBA) show that two molecules are bound simultaneously to PPARα.

Ligand parametrization was done using the ATB webserver [215]. An initial optimization is done at HF/STO-3G level of theory. For molecules with less than 50 atoms, a re-optimization with the B3LYP/6-31G* level of theory in combination with the polarizable continuum model (PCM) implicit solvent to represent water takes place. In addition, the Hessian matrix is calculated in order to assign harmonic force constants for bond and angle terms. All-atom CNS parameter and topology files were retrieved from the ATB webserver [215] for subsequent docking. After ligand superposition and parametrization, template-based docking was performed using HADDOCK 2.4; such docking protocol has been described and validated in reference [122]. In particular, a small energy minimization was performed followed by an explicit water refinement stage (step 3 in section 3.4.1), in which the system was solvated with an 8 Å shell of TIP3P water molecules and submitted to three MD-based st eps (with a 2 fs timestep each): (1) A heating phase of 100 MD steps at temperatures of 100K, 200K and 300K, respectively. During these integration steps, weak positions restraints were applied to all atoms not belonging to the protein-ligand interface (5 Å in range of any ligand atoms). (2) Afterwards, 2500 MD steps were carried out at a constant temperature of 300K. Here, positions restraints were imposed only on non-hydrogen atoms that were not part of the protein-ligand interface. (3) A final cooling phase of 500 MD steps at 300K, 200K and 100K, respectively, was performed. In that last stage only, backbone atoms of non-interface residues were restrained [122]. The scoring function used is described in section 3.4.1.

Afterwards, predicted docking poses were clustered using the Fraction of Common Contact (FCC) based clustering with a cutoff of 0.60. Due to the nature of HADDOCK's scoring function, docking scores of different protein-ligand systems cannot be compared among each other. Therefore, in order to compare estimated binding affinities for the docking poses obtained with HADDOCK, a machine-learning based predictor called ΔGscore was utilized, as implemented in the PRODIGY-LIG webserver [216].

2. *Induced Fit Docking approach*

The IFD approach of Glide was performed with the extended sampling protocol as described in section 3.4.2.

In general, Glide uses a grid-based technique in order to generate and score docking poses. These pre-calculated grid maps were placed and defined based on co-crystalized ligands. As mentioned above, PPARα ligands can adopt different binding modes and thus for each pocket of PPARα a separate docking was conducted (see Table 5). In particular, the center of such a grid box was placed at the centroid of the selected ligand. The grid box itself is divided into two different areas, i.e. an inner and an outer box. The outer box defines the area in which all ligand atoms have to be included. The inner part, on the other hand, restricts the ligand centroid to explore a 10Å x 10Å x 10Å volume by default.

Furthermore, core restraints were applied during the dockings to the Arm I and the Center pockets. Namely, the carboxylate group was allowed to deviate only 2Å in comparison to the position of the same group in the crystallographic poses to preserve H-bonds with S280, Y314, H440 and Y464.

Refinement of nearby residues was applied to amino acids within a range of 4Å of the ligand, whereas other parameters were left on default. Here, the

extended sampling protocol in Prime was used to generate possible binding poses as described in section 3.4.2. The scoring function selected to rank the final poses was Glide SP, as indicated in section 3.4.2.

5.1.3 Molecular Dynamics Setup

MD simulations were performed using the GROMACS (version 2020) software package [217]. Starting structures were selected from earlier performed docking runs (see section 5.1.2). In particular, top scoring docking poses from both Haddock and Glide's IFD calculations were taken.
Before MD, the co-activator peptide co-crystallized in the X-ray structures was re-added to each receptor-ligand complex. Moreover, capping groups were added to termini of both receptor and peptide (an amino group for the C-terminus and an acetyl group for the N-terminus). Thus, the final systems consisted of a ternary complex (receptor-ligand-co-activator peptide).
Protein-ligand systems were described using the Amber ff99SB-ILDN force field for the receptor and peptide and the GAFF2 force field with AM1/BCC charges for the small molecules, as calculated with ACPYPE [218]. Afterwards, systems were solvated using TIP3P water molecules using a cubic cell unit with distance between protein and box edge set to a value of 1.5nm. Systems were neutralized with the appropriate number of sodium ions.
For each protein-ligand system four independent replicas were prepared (for details see Table 6).
After set up of respective systems, a subsequent energy minimization using the steepest decent algorithm was performed and considered as converged when the maximum force was below 1000kJ mol^{-1} nm^{-1}. Thereafter, the minimized systems were equilibrated. The first equilibration step was conducted under NVT conditions for 1ns using the leap frog integrator and the Berendsen thermostat [219] to achieve the desired temperature of 300K. The second part of the equilibration was performed under NPT conditions for 2ns using a Parrinello-Rahman barostat [220] to attain 1 bar pressure and velocity-rescaling thermostat in order to preserve a temperature of 300K. In both equilibration steps, heavy atoms of both protein and ligand were restrained with a force constant of 1000 kJ mol^{-1} nm^2 in order to prevent significant changes during the equilibration. Afterwards, restraints were released and production simulations were run with a chosen integration time step of 2fs (for details regarding the total length of the simulation, see Table 6) and using the Parrinello-Rahman barostat and velocity-rescaling thermostat to keep pressure and temperature constant, respectively.

Table 6. Details of the performed MD simulations. The grey highlighted section for gemfibrozil indicates that an alternative starting pose was used for those MD simulations, as explained in section 5.2.2.3.

Ligands	#Molecule(s): Pocket(s)	Time (in ns)	Replica[1]	Starting Pose
GW7647	(#1) Center/Arm II/Arm III	300	R1 R2 R3-R4	Glide Haddock Crystal Structure (PDB code: 6KB3)
Ciprofibrate	(#1) Arm I	300	R1 R2 R3-R4	Glide Haddock Crystal Structure (PDB code: 6LX5)
	(#2) Arm I and Arm X	300	R1 R2 R3-R4	Glide Haddock Crystal Structure (PDB code: 6LX5)
Gemfibrozil	(#1) Center/Arm II	400	R1-R2 R3-R4	Haddock Glide
	(#1) Arm I	400	R1-R2 R3-R4	Haddock Glide
	(#2) Arm I and Arm X	300	R1-R2 R3-R4	Haddock Glide
	(#2) Center/Arm II and Arm X	300	R1-R2 R3-R4	Haddock Glide
	(#2) Center/Arm II and Arm X	300 200	R1-R2 R3-R4	Haddock Glide
Cinnamic Acid	(#1) Center	400	R1-R2 R3-R4	Haddock Glide
	(#1) Arm I	400	R1-R2 R3-R4	Haddock Glide
	(#2) Center and Arm X	300	R1-R2 R3-R4	Haddock Glide
	(#2) Arm I and Arm X	300	R1-R2 R3-R4	Haddock Glide

[1]For each replica a different set of random velocities was generated.

5.1.4 Simulation Analysis

Protein-ligand interactions were analyzed in order to get more insights into molecular determinants of PPARα-ligand complexes.
To this aim, MD trajectories were analyzed with Gromacs, Python3, ProLIF (version 1.0) [125] and MDAnalysis (version 2.2) [221]. The first 50ns of each simulation were considered as unrestrained equilibration phase and were thus not taken into account during subsequent analysis steps.
Hydrogen bond interactions were examined with the two aforementioned software libraries. ProLIF was used to detect direct H-bonds, whereas water-bridged H-bonds (with a maximum of one water molecule) were investigated with MDAnalysis. The donor-acceptor distance cut-off was set to 3.5Å and the donor-hydrogen-acceptor angle tolerance was set between 130° and 180° for both ProLIF and MDAnalysis.
Other considered interactions, i.e. hydrophobic interactions, pi-stacking, halogen bonds, anion-pi interactions as well as cation-anion (salt bridges) interactions, were also investigated through ProLIF with default parameter options, as listed in reference [125].

Clustering of respective MD trajectories was performed with GROMACS and its build-in clustering command. The gromos algorithm, as described in section 3.8, was chosen in order to group respective MD frames into groups of similar structures. The MD trajectories were clustered using frames every 50ps and the RMSD cutoff was selected based on the ligand RMSD distribution.

5.2 Results

5.2.1 PPARα Dockings

5.2.1.1 Validation Tests

In order to predict favorable binding poses of CGA compounds, two docking approaches were carried out, as described in section 5.1.2. Molecular Docking was performed for CGA compounds listed in Table S6. As mentioned in section 3.4, docking approaches can result in poses that might not necessarily resemble the correct pose (i.e. the docking pose is not comparable to the crystallographic one) or might not be ranked properly (i.e. the correct pose is among the predicted docking poses, but it is not scored as the best). Besides predicting incorrectly binding poses (either in terms of 3D structure or score), the false positive rate (FPR, i.e. the probability of incorrectly predicting ligands as binders) can also vary significantly among docking implementations and protein-ligand systems. Therefore, both docking protocols were validated before applying them to CGAs.

In particular, both HADDOCK and IFD docking protocols were validated against available experimental structures of PPARα in complex with known agonists, by redocking the co-crystallized ligands. In general, predicted binding poses and X-ray poses are in good agreement, as shown in Figure S57 (HADDOCK) and Figure S58 (IFD). Out of the fifteen re-docked agonists, fourteen show a good agreement between the predicted HADDOCK poses and the crystallographic poses (Figure S57). The only exception is tesaglitazar, which is predicted to occupy the Center and Arm III pockets, instead of the Center and Arm II, as observed in the crystal structure. This discrepancy could be due to the different conformation of this linear ligand compared to the branched ligands used as template (see Table 5), which bind simultaneously to the Center, Arm II and III pockets. However, the binding poses of other known agonists that span only the Center and Arm II pockets (see Table 7) was correctly predicted. Instead, it seems that the sulfone moiety of tesaglitazar is more likely to be placed by HADDOCK near that Arm III because the template molecules used for the Center/Arm II/Arm III also contain H-bond acceptor/donor groups located near or in Arm III.
In the case of IFD (see Figure S58) more variability compared to the crystallographic poses are observable when compared to the pose with the lowest RMSD among the top 5 poses. Indeed, that behavior was expected due to the nature of the different docking approaches. In particular, HADDOCK only performs a refinement of the initial ligand pose generated by superimposition with known agonists, through which only minor adjustments of that ligand pose are possible. The IFD protocol instead only considers information about the receptor structure and the approximate location of the binding site. Moreover, the Prime refinement and extended sampling algorithm used in the IFD protocol allow the binding site residues to adjust to the ligand, thus resulting in more diverse poses.

Table 7. Agonists of PPARα with available crystallographic structures used for validation of the two docking approaches used in this thesis. Ligands are listed in alphabetical order, together with the PDB code of the corresponding X-ray structure and information about the binding pockets they occupy. Dockings scores for HADDOCK (∆Gscore) and IFD (Glide score) are also included.

Ligand	PDB code (resolution in Å)	Pocket	∆Gscore[1] (in a.u.)	Glide score (in kcal/mol)[2]
Aleglitazar	3G8I (2.20)	Center/Arm II	68.4 (2.2)	-11.5 (0.1)
Bezafibrate	7BPZ (2.43)	Center/Arm II	90.0 (2.2)	-10.6 (0.2)
CHEMBL1089210	3KDU (2.07)	Center/Arm II	61.6 (2.8)	-13.2 (0.2)
CHEMBL1089501	3KDT (2.7)	Center/Arm II	95.4 (3.8)	-11.4 (0.3)
CHEMBL219586	2P54 (1.79)	Center/Arm II	30.2 (5.7)	-13.3 (0.2)
CHEMBL271240	2REW (2.35)	Center/Arm II	53.3 (2.5)	-12.7 (0.4)
Ciprofibrate	6LX5 (1.87)	Arm I	95.2 (3.3)	-9.6 (0.0)
Fenofibrate	6LX4 (2.13)	Arm I	93.6 (2.8)	-10.7 (0.6)
GW409544	1K7L (2.50)	Center/Arm II	43.9 (3.1)	-13.4 (0.1)
GW7647	6KB3 (1.45)	Center/Arm II/Arm III	39.9 (2.0)	-12.0 (0.4)
Pemafibrate	6KB4 (1.42)	Center/Arm II/Arm III	68.3 (1.8)	-12.7 (0.2)
Saroglitazar	6LXC (2.03)	Center/Arm II	61.6 (3.6)	-11.7 (0.1)
Tesaglitazar	1I7G (2.20)	Center/Arm II	90.7 (2.5)	-10.8 (0.3)
TIPP-703	7E5F (1.79)	Center/Arm II	41.5 (3.7)	-14.0 (0.2)
WY14643	6KBA (1.82)	Center	103.0 (2.2)	-10.1 (0.3)

[1]Average score of top 20 poses and respective standard deviation. A better docking score corresponds to a less positive DGscore value. [2]Average score of top 5 poses and respective standard deviation. A more negative Glide score indicates a better docking score.

Another validation test is the comparison of the ranking predicted for the known agonists based on the docking scores against the experimentally measured EC_{50} values. Although, here I am using EC_{50} values, such half maximal effective concentrations can depend both on the ligand binding affinity as well as its ability to trigger the conformational changes needed to activate the receptor. Despite this limitation, I observed that both HADDOCK and IFD (using DGscore and Glide values, respectively) are able to correctly rank the known agonists according to their EC_{50} values (see Table 7).

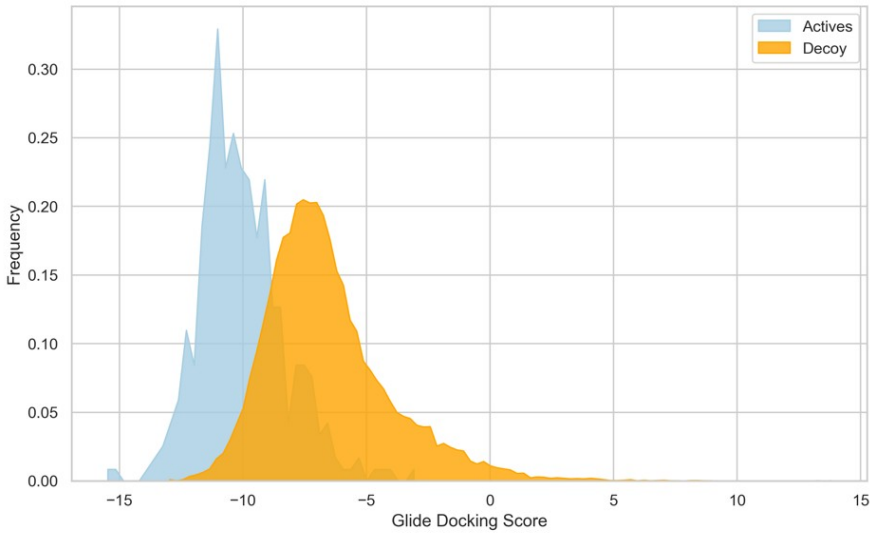
For agonists located in the Center/Arm II/Arm III region of PPARα, GW7647 is predicted to be the most potent PPARα agonist/activator (EC_{50} of 51.8nM), followed by Saroglitazar (5.6 μM) and Bezafibrate (50.5 μM) [57]. As for the Arm I binders, Fenofibric acid (23.2 μM) is followed by Ciprofibrate (23.9 μM) [57]. Nonetheless, it should be noted that the Center/Arm II/Arm III and Arm I dockings presented here do not take into account the contribution of a second molecule binding to Arm X to the affinity of the aforementioned fibrates.

As Glide was designed to perform high-throughput screening in an automatic way, another validation test was applied, besides re-docking respective known agonists (see Table 7 and Figure S58). In particular, active and decoy compounds from the DUD-E database [222] were docked into PPARα to verify that the docking score is able to discriminate binders from non-binders.

The DUD-E database contains pre-compiled libraries of active and decoy compounds for several protein targets [222], including a dataset for PPARα with 19,356 decoy compounds and 373 actives. Actives are known binders of the protein target, whereas decoys are molecules assumed to not bind to the protein target. Decoys are normally chosen so that they exhibit physico-chemical properties similar to actives, but different chemical structures.

The performance of Glide is shown in Figure 17. In general, active compounds perform better (i.e. more negative docking scores) compared to decoy molecules (see Figure 17A). This is more evident considering the receiver operating characteristic (ROC) curve in Figure 17B. This curve represents the true positive rate (TPR, i.e. the probability that an active compound is predicted by Glide as so) with respect to the false positive rate (FPR, i.e. the probability that a decoy is predicted as binder by Glide). The area under the curve (AUC, i.e. the probability that a randomly chosen active has a higher score than a randomly chosen decoy) turns out to be 0.89, which indicates that Glide can discriminate between agonists and decoys of PPARα with high accuracy.

A



B

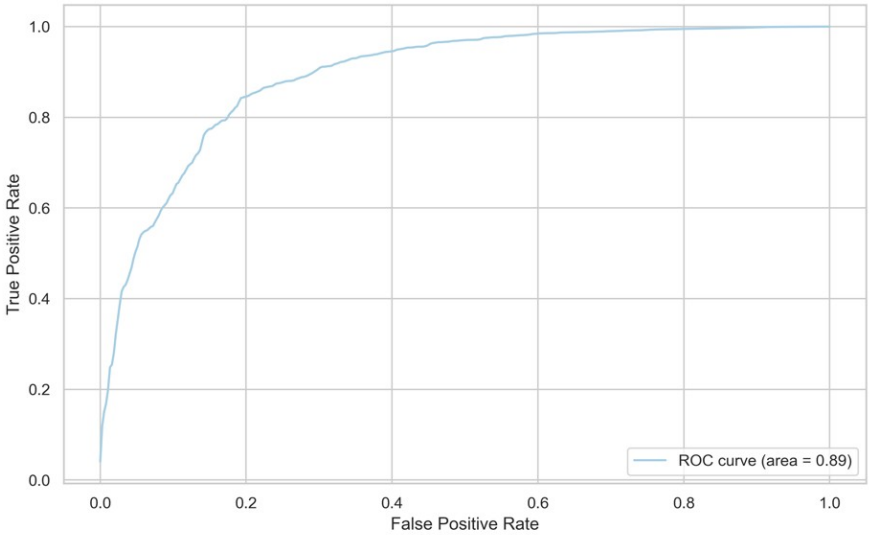


*Figure 17. (**A**) Histogram of docking scores for active and decoy compounds of PPARα. The frequency is normalized such that the total area of the histogram equals 1. (**B**) ROC curve of respective validation dockings.*

5.2.1.2 Haddock

HADDOCK dockings (i.e. the template-based docking approach) were performed as described in section 5.1.2. Results of that docking approach are shown in Figure 18.

As shown in Figure 18 and Table 7, known PPARα agonists bound to either the Arm I or Center/Arm II/Arm III pocket (star-shaped points in Figure 18) show lower ΔGscore values (i.e. more favorable binding to PPARα) compared to CGAs.
To the best of my knowledge, EC_{50} values of CGAs have not been determined despite the one for cinnamic acid (5.08 μM). This value is close to the one of saroglitazar (5.6 μM); however, docking scores of cinnamic acid either bound to Arm I or the Center pocket are not close to the one from saroglitazar. Besides the EC_{50} value, the Hill slope of cinnamic acid was also determined to be 12.89. A Hill coefficient larger than one suggests that multiple molecules can bind to a protein, thus indicating that cinnamic acid has multiple binding sites in PPARα. X-ray structures of PPARα revealed that fibrates can also simultaneously bind additional molecules in other pockets, such as Arm X (see PDB codes 6LX4, 6LX5 or 6KBA). If cinnamic acid can indeed bind to multiple binding sites of PPARα, as hinted by the Hill slope, the different score of cinnamic acid and saroglitazar (in contrast to the similar EC_{50} values) could be explained by the fact that the ΔGscore value only takes one single molecule into account (whereas the experimentally measured value includes the effect of multiple molecules).
Moreover, a correlation between ΔGscore values and ligand properties emerges when considering hydrophobicity and size. The former was computationally estimated using the consensus LogP value, calculated with the SwissADME webserver [223] and plotted in the x-axes of the plots in Figure 18, whereas size was expressed as ligand efficiency (or docking score normalized by the number of heavy atoms) and encoded in the color bars of the same plots. In particular, larger and more hydrophobic compounds tend to have both better EC_{50} and ΔGscore values.
Interestingly, hydrophobicity can influence not only binding to PPARα (which contains mostly hydrophobic pockets) [58], but also bioavailability (i.e. distribution across of the body; indeed, this logP parameter is used by many drug design predictors of drug likeness) [224]. Altogether, the use of two parameters ΔGscore and consensus log P, allows not only to rank compounds according to their predicted affinity towards PPARα, but also to discriminate among compound groups.
Within CGA compounds, di-CGAs (hexagonal points in Figure 18) bound to the Center/Arm II/Arm III pocket appear to have ΔGscore values comparable to those of the lipid-lowering agents bezafibrate and gemfibrozil, with CGAs (rhomboidal points) and hydroxycinnamic acids (triangular points) not far behind. The consensus log P values show that CGAs are less hydrophobic than PPARα synthetic agonists. This is not unexpected, considering the presence of hydroxyl groups in the chemical structures of CGAs and the fact that hydrophobicity is optimized during drug design to improve bioavailability.
If molecules are docked into the Arm I pocket of PPARα, a similar pattern emerges. Di-CGAs have similar ΔGscore values compared to known agonists, such as ciprofibrate and fenofibrate. CGAs, however, perform slightly worse, followed by hydroxycinnamic acids and related compounds.
Furthermore, docking scores showed that di-CGAs with at least one hydroxycinnamic acid group as *cis* isomer often perform better than the counterpart molecules with both hydroxycinnamic acid moieties as *trans* isomers. This observation was not dependent on a particular binding site and, in some cases, also extendable to CGAs as well as HCAs.

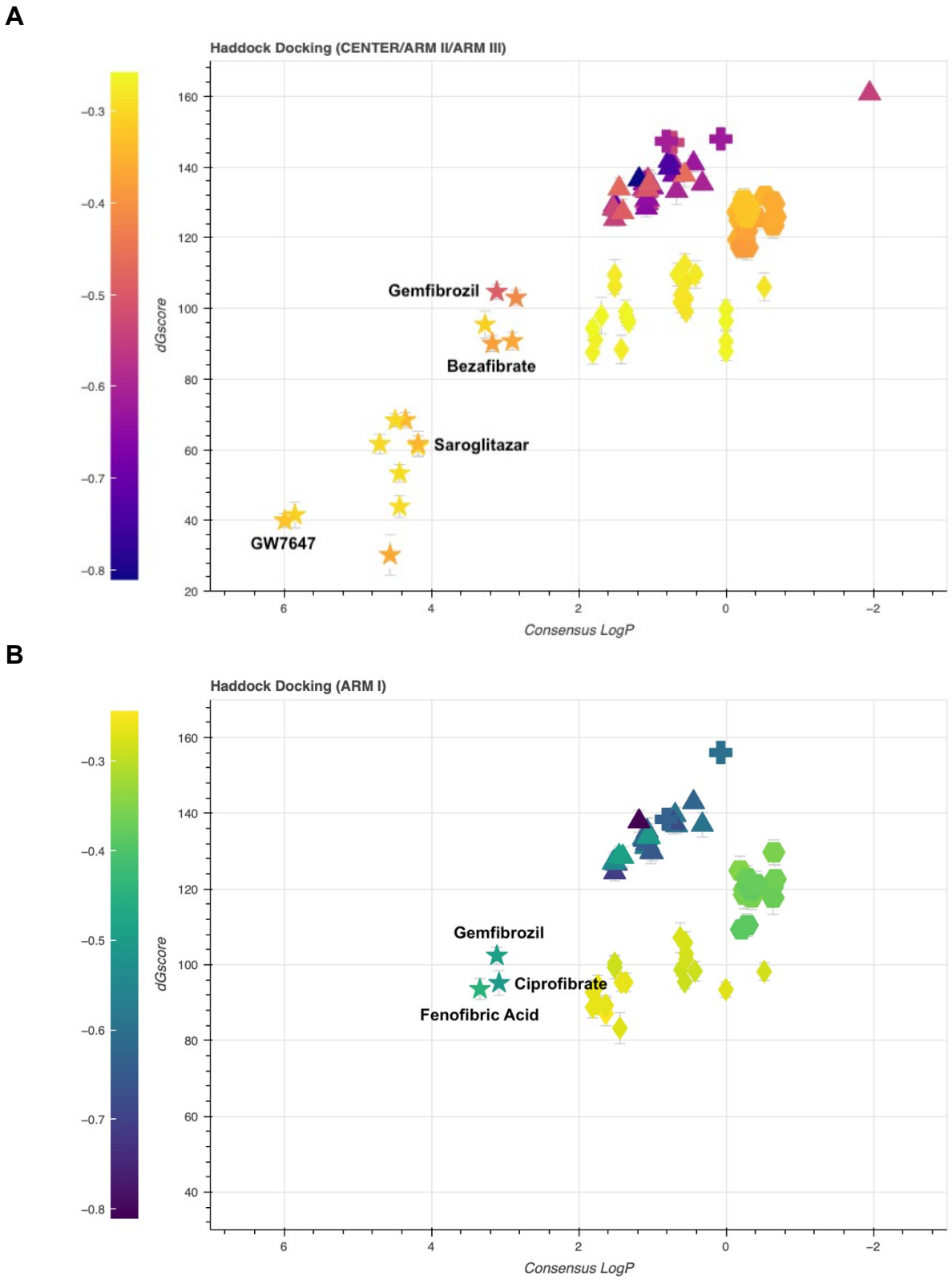
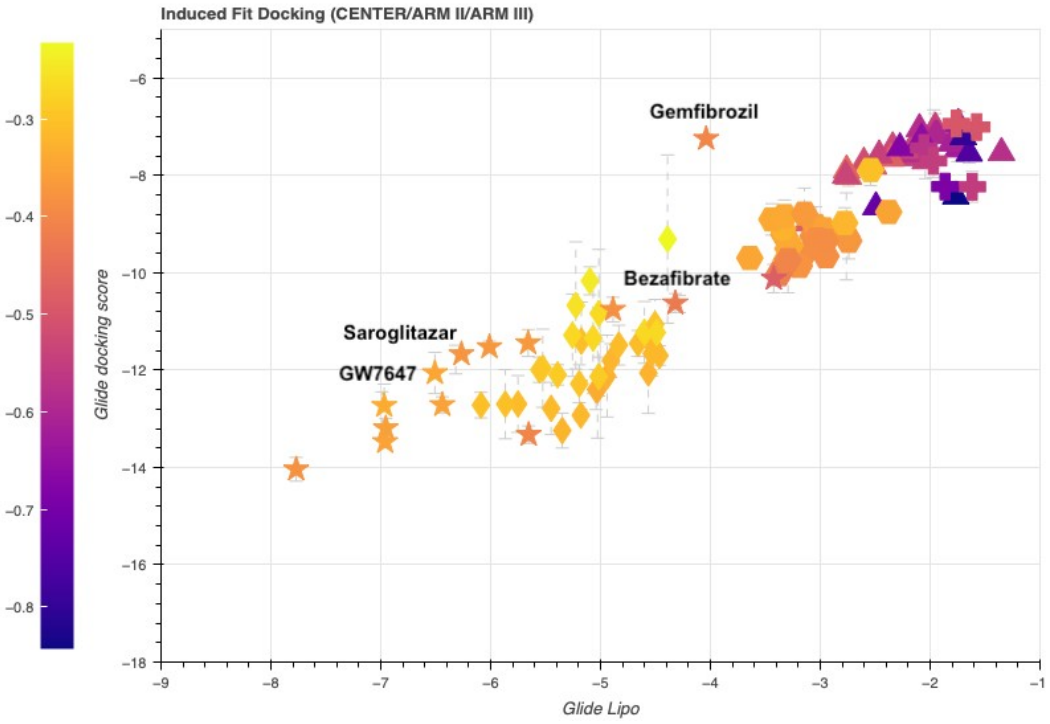


*Figure 18. HADDOCK docking results. Symbols indicate different groups of compounds, i.e. stars represent known agonists, triangles HCAs, cross symbols related compounds, hexagonal shapes CGAs, and diamond shapes di-CGAs. The ligand efficiency is visualized through the color bar. (**A**) Dockings of the Center/Arm II/Arm III pocket (**B**) Dockings of the Arm I pocket.*

5.2.1.3 Induced Fit Docking

The induced fit dockings were performed as described in sections 3.4.2 and 5.1.2 and the results are shown in Figure 19.
Similar to the HADDOCK dockings, di-CGAs display better dockings scores than CGAs, which in turn exhibit better docking scores than hydroxycinnamic acids.
However, a significant difference between the HADDOCK and IFD dockings is that di-CGAs, CGAs and hydroxycinnamic acids have scores closer to known agonists according to the Glide scoring function. For instance, mono-CGAs are ranked better than gemfibrozil independently of the binding pocket (see Figure 19A and B). Moreover, di-CGAs reach docking scores within the range of GW7647 if placed within the Center/Arm II/Arm III pocket. If, however, Arm I agonists were used as a template to dock di-CGAs into PPARα, their docking scores are even surpassing the one for GW7647.

A



B

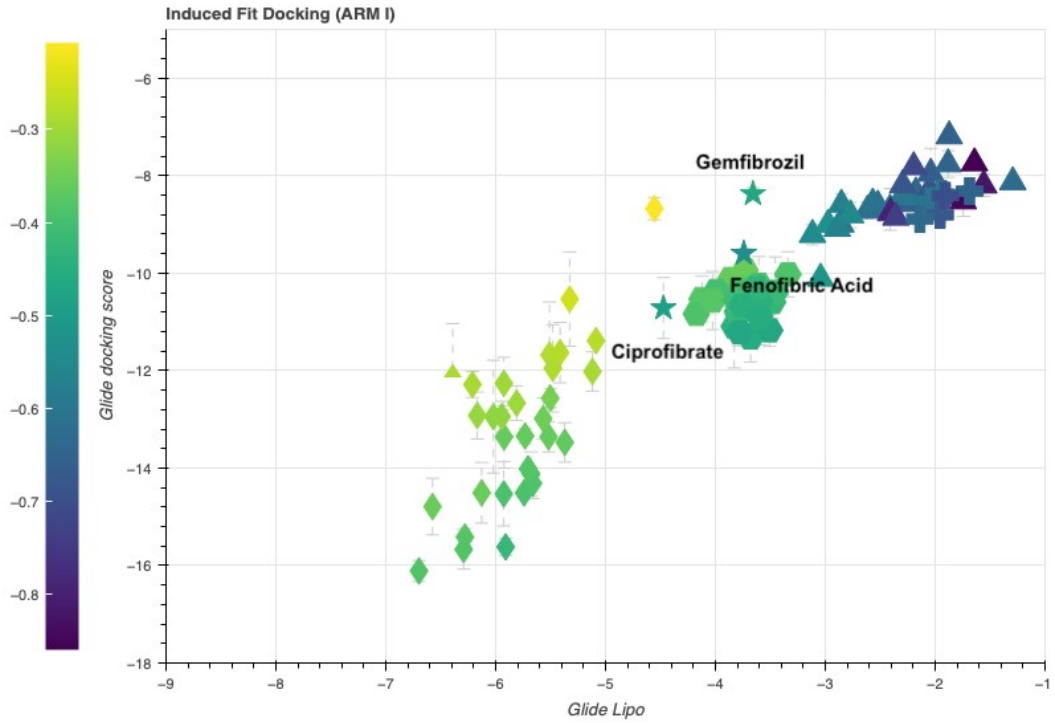


*Figure 19. Glide docking results. Symbols indicate different groups of compounds, i.e. stars represent known agonists, triangles HCAs, cross symbols related compounds, hexagonal shapes CGAs, and diamond shapes di-CGAs. The ligand efficiency is visualized through the color bar. (**A**) Dockings of the Center/Arm II/Arm III pocket (**B**) Dockings of the Arm I pocket.*

5.2.1.4 Molecular details of predicted binding modes

Chlorogenic acids (mono-CGAs) have sizes comparable to fibrates (i.e. ciprofibrate or fenofibric acid) or even synthetic agonists such as Wy14643. Therefore, it was not surprising that docking poses of CGAs had similar orientation within the binding pockets (i.e. Center or Arm I, respectively). As shown in Figure 20 for the prototypical member of this group, 5-CGA, the carboxylate forms two bifurcated H-bonds with S280, Y314, H440 and Y464, independently of the orientation of the hydroxycinnamic acid part.

If CGAs are located in the Center/Arm II/Arm II pocket of PPARα, the 1’ OH group of the quinic acid moiety (see Figure 4) could act as H-bond acceptor for H440[1] and as H-bond donor for Y464. This would also imply that Y464 breaks its H-bond with the carboxylate and most likely acts as H-bond donor for Y314. Another imaginable situation would be that the hydroxyl group is only acting as H-bond acceptor for H440 while the carboxylate is maintaining bonds with S280, Y314 and Y464, similar to ciprofibrate (see PDB code: 6L37). In either case, Q277 would be able to change its rotameric state and form a H-bond with an oxygen atom of the carboxylate group and indirectly stabilize H12. Moreover, hydroxyl groups attached to the hydroxycinnamic acid part would be in range of forming a direct or water-mediated H-bond with T279 as seen for GW7647.

If, however, CGAs are located in Arm I, an additional H-bond could be formed between one oxygen of the ester moiety (connecting the quinic acid and the hydroxycinnamic

[1] H440 is singly protonated in epsilon, as the delta nitrogen of the imidazole ring is forming a H-bond with K358 in all crystal structures of PPARα; this applies also to all performed simulations.

acid groups) and Q277. In addition, pi-stacking interaction with F351 (located in H7) could stabilize 5-CGAs binding pose. Besides, other hydrophobic interactions formed by the hydroxycinnamic acid part could help to stabilize the unstructured part between H10/11 and H12, thus contributing to protein activation by reducing fluctuations in H12. Furthermore, the hydroxyl group at position 1' of the cyclohexane ring can potentially take over the H-bond with S280 from the carboxylate group, so that 5-CGA moves further towards residue Y464, strengthening that H-bond.

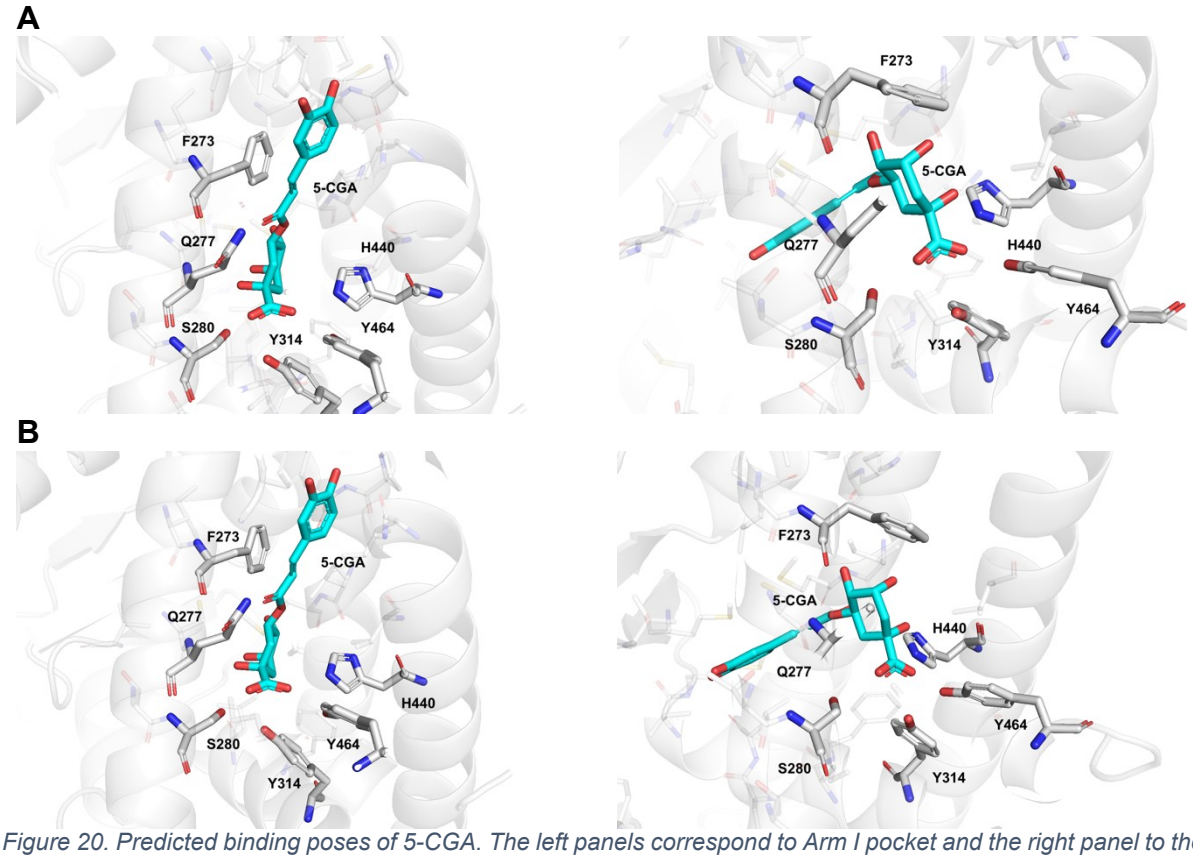


*Figure 20. Predicted binding poses of 5-CGA. The left panels correspond to Arm I pocket and the right panel to the Center/Arm II/Arm III pocket. (**A**) Haddock (**B**) IFD.*

Molecular docking revealed two favorable binding modes for di-CGAs in PPARα (as shown in Figure 21 for the prototypical di-CGA, 3,5-diCGA). Di-CGAs are larger compounds than CGAs and similar in size to GW7647 or pemafibrate. Therefore, it is not surprising that the Center/Arm II/Arm III docking poses (left panels in Figure 21) shows similarities to GW7647, that is, the two hydroxycinnamic acid moieties occupy the Arm II and Arm III pockets, respectively. If such a binding mode would be adopted by di-CGAs, quinic acid could behave in a similar way as for CGAs, i.e. an additional H-bond (besides the ones the carboxylate is participating in) could be formed between H440 and the hydroxyl group at position 1' of quinic acid. Other possible H-bond partners could be side chains of N219, T283 and E283 in case of Arm III. In Arm II, however, only backbone atoms could act as potential candidates for H-bonds, since that pocket is lined by non-polar side chains.

The second binding mode of di-CGAs is more similar to compound CHEMBL271240 (PDB code 2REW; see Figure S57 and Figure S58) in which Arm I and the Center/Arm II pocket are occupied (see Figure 21; left side). In that case possible additional H-bond partners would be T279, A333 or even K257 if the Ω-loop moves closer to the pocket

entrance. In Arm I, Q277 or other backbone atoms could again be considered as potential interactions partners.

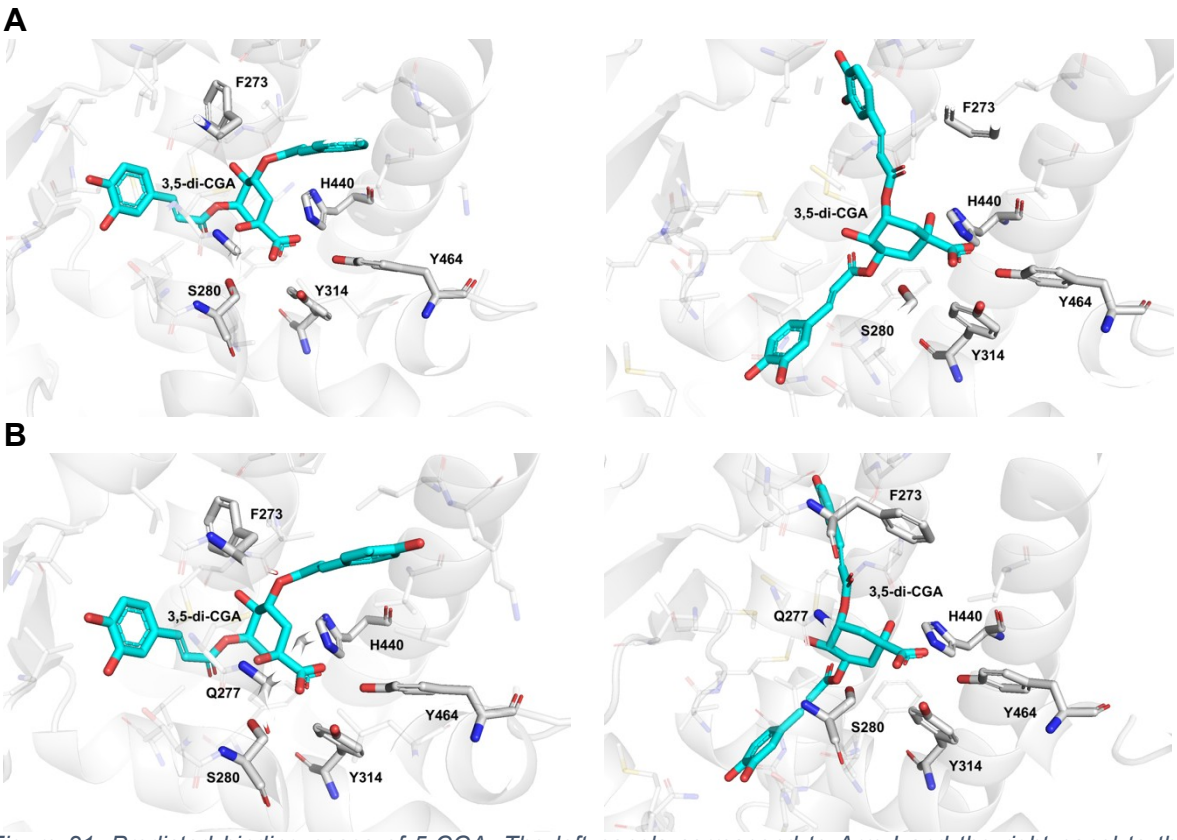


*Figure 21. Predicted binding poses of 5-CGA. The left panels correspond to Arm I and the right panel to the Center/Arm II/Arm III dockings. (**A**) Haddock (**B**) IFD.*

As shown in Figure 22, cinnamic acid is forming four H-bonds to key residues, namely, S280, Y314, H440 and Y464 (in both pockets, i.e. Arm I and Center pocket). While the phenyl group can form an aromatic H-bond in Arm I with Q277 (see Haddock pose in Figure 22A, left), within the Center pocket only favorable hydrophobic interactions can be detected. Considering that the latter mentioned pocket is larger, more movement of the phenyl group would be allowed, which is also reflected in respective binding poses (see Haddock vs Glide binding poses in the right panes of Figure 22).

Molecules belonging to the class of hydroxycinnamic acids, have at least one hydroxyl group more attached to the phenyl ring compared to cinnamic acid (see Figure 4 and Figure 22). Indeed, that can be favorable when more H-bonds are formed, for instance with residue T279.

Lastly, consideration of chemical structures and inspection of respective docking poses, revealed that compounds containing hydroxycinnamic acid group(s) with a *trans* double bond exhibit more constrained geometries than known agonists of PPARα. As mentioned in section 2.6, *trans* as well as *cis* isomers of HCAs can be present in coffee, which has an impact on the 3D structure of respective molecules. That change in double bond configuration can have a positive effect on the docking scores of some compounds (data not shown) and thus on the likeliness to bind to PPARα. In general, that effect became more important the larger the respective ligand became. I ascribed this correlation to the binding pocket of PPARα possessing a curvature around H3. Due to this feature, larger and more rigid molecules may have problems to be able to adapt properly to the shape of the binding pocket. In case of di-

CGAs, *cis* isomers could indeed improve binding poses compared to *trans*, either by better embracing H3 or by finding a better positioning in Arm I. HCAs also showed better docking scores when in *cis* configuration, as this isomer allowed a better placing of the phenyl ring.

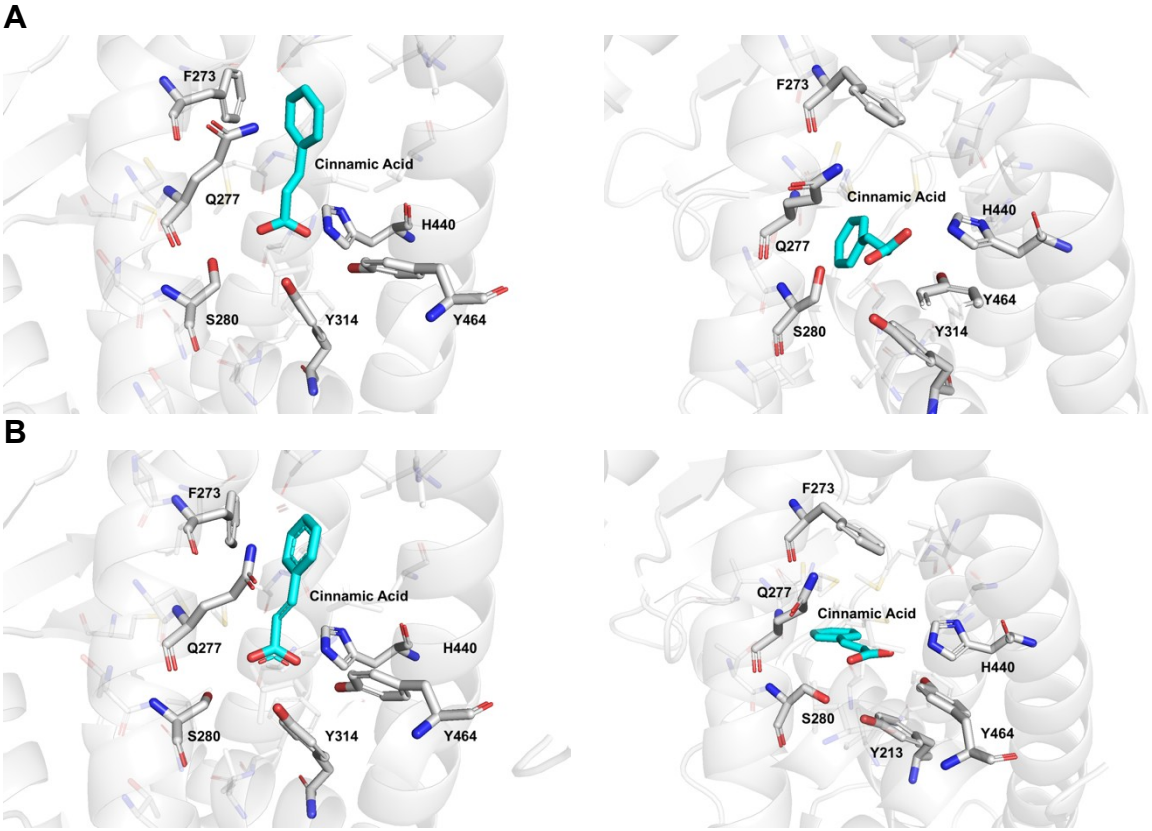


*Figure 22. Predicted binding poses of cinnamic acid. The left panels correspond to Arm I and the right panel to the Center/Arm II/Arm III dockings. (**A**) Haddock (**B**) IFD*

5.2.2 PPARα Simulations

MD simulations give more detailed insights into the dynamics of a protein-ligand complex, which helps to identify important protein-ligand interactions and thus helps to validate respective docking poses. Moreover, some docking poses showed comparable docking scores for ligands bound to different binding pockets (i.e. gemfibrozil or cinnamic acid) and thus it was not possible to discriminate whether binding to one pocket is favored over the other. through which an accurate prediction of binding modes was quite challenging. In that case, MD simulations can be an adequate computational method as it overcomes limitations of traditional docking approaches by refining predicted binding poses. If binding modes are predicted accurately, protein-ligand poses should not experience large movements from respective starting structures. If, however inaccurate binding poses were generated from earlier dockings, the ligand should either leave the binding pocket or adjust accordingly [225].

5.2.2.1 GW7647

GW7647 is one of the most potent PPARα agonists with an EC_{50} value of 51.8 nM [57] . Hence, besides serving as benchmark in comparison to other ligands, MD simulations of latter mentioned complex help identifying and validating important residues for the activation of PPARα. Four simulations were performed to have statistical meaningful results (R1-R4, see Table 6) [226]. In order to assess convergence of the simulations, protein backbone and ligand RMSD values were calculated.

The protein backbone showed values ranging from 1.5 to 3 Å, whereas ligand RMSD values span from 0.5-3 Å (Figure S62). Replicas 1 and 2 showed slightly higher ligand RMSD values compared to R3 and R4, which was expected since HADDOCK and IFD docking poses served as starting structure for these simulations. However, both protein backbone and ligand RMSD values, comparable with previously performed MD simulations of PPARα-ligand complexes [227, 228].

Moreover, root mean square fluctuations (RMSFs) of C_α backbone atoms were calculated for every amino acid of PPARα (Figure S63) to investigate possible changes in protein flexibility upon ligand binding. Residues located in secondary structure elements, i.e. α-helices (H1-H12) or β-sheets (S1-S4), show low fluctuations throughout the whole simulation time. Besides the C- and N-terminal residues, three other protein areas, (namely, the P-site, the Ω-loop and a loop region between H9 and H10/H11) also show larger RMSF values of up to 5 Å. The P-site, a loop region connecting H2 and H2', is known as flexible region of PPARα and also displayed higher RMSF values as observed in previous simulation studies [65, 229].

The Ω-loop is also known to have large RMSF values and is even disordered in some protein structures, which could indicate that the pocket entrance is able to adapt to different ligand sizes [58, 60, 65, 227].

GW7647 is a large and branched compound that occupies four binding cavities of PPARα, namely, the Center-, Arm I, Arm II- and Arm III-pocket.

Large parts of Arm II and Arm III are composed of hydrophobic residues (Table 2 and Figure S64), thus, most atoms of GW7647 are in hydrophobic contact with PPARα.

In addition, GW7647 possesses three oxygen atom and one sulfur atom, which are able to serve as H-bond acceptor, and one nitrogen, which is able to act as H-bond donor (see Figure S56). As mentioned in section 2.7.1, four amino acids, i.e. S280,

Y314, H440 and Y464, are important for protein activation and interact with most PPARα ligands [58, 59, 61, 230, 231]. As evident from the respective X-ray structure, these four residues form H-bonds with the carboxylate moiety of GW7647 (PDB code 6KB3). The performed MD simulations showed that, in all replicas (R1-R4), all four H-bonds are indeed stable and present almost the whole simulation time, as shown in Figure S64. Heatmap of GW7647-protein interactions. Moreover, that X-ray structure (PDB code 6KB3) showed a second conformation of GW7647 in which the carboxylate is rotated and the H-bond with H440 is lost. Such an alternative H-bond pattern of the carboxylate group is also sampled in my simulations.

In addition, GW7647 forms additional water-mediated H-bonds with the backbone of L331 and A333, which appear to be stable in all four replicas (Figure S65), as observed in the crystal structure. Amino acid Q277 is a residue located in H3 and forms a transient H-bond with the ligand. In R3, however, that H-bond is water-mediated and formed with both atoms of the carboxylate group (Figure S65). Interactions with H3 has been proposed to stabilize the activated protein state of PPARα, so that H-bonds with Q277 could help to promote PPARα activation. Thus, the observed interactions agree with the available crystallographic data and the proposed molecular mechanism for PPARα activation (see section 2.7).

In order to assess the sampled binding poses of GW7647, frames of each simulation were clustered using a 1.3 Å RMSD cutoff (as this value corresponds to the main peak of the ligand RMSD distribution, see Figure S66). Out of the four replicas (Table S7), R1, R2 and R4 are represented by one main cluster, whereas R3 sampled two main clusters, which was expected given the ligand RMSD bimodal distribution plots (Figure S67).

As shown in Figure 23A, the simulations of the PPARα-GW7647 complex explored mostly poses similar to conformer A of the crystal structure, regardless of whether they were started from docking poses (R1-R2) or the crystallographic pose (R3-R4).

Nonetheless, the sampling and clustering of R3 also revealed another conformation of the ligand. In particular, rotation of the carboxylate group (Figure 23B) results in GW7647 no longer forming a H-bond with S280 and being able to establish a water-mediated H-bond with Q277. Furthermore, I observed that this change in the H-bonding pattern of the carboxylate is due to water molecules diffusing into the binding site. A further consequence of such water entrance is an enhanced polarity of the cavity, through which a change of the rotameric state of F273 seems to occur (Figure 23B and Figure S66).

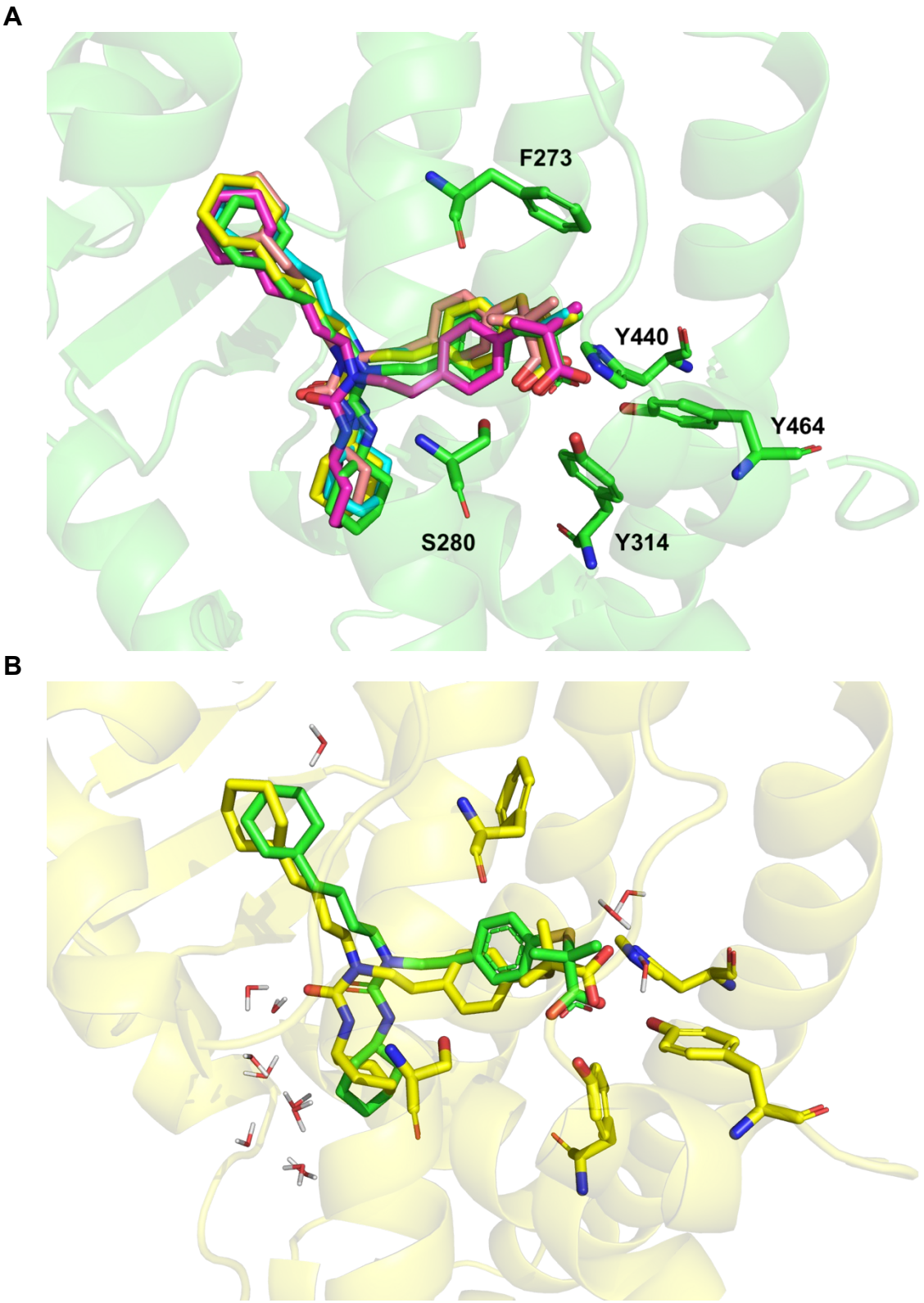


*Figure 23. (**A**) Cluster centroid structures from R1-R4 simulations of the PPARα-GW7647 complex. Ligands and important residues are represented as sticks, with carbon atoms colored as follows: (**Green**) Crystal structure (PDB Code 6KB3), (**Cyan**) R1: Cluster 1 - 93%, (**Pink**) R2: Cluster 1 - 94%, (**Yellow**) R3: Cluster 2 – 26.2%, (**Salmon**) R1: Cluster 1 - 97.2%. (**B**) R3: Cluster 1 - 69.3%. The percentage values indicate the population of the corresponding clusters.*

5.2.2.2 Ciprofibrate

Ciprofibrate is another potent agonist of PPARα with an EC_{50} value of 23.9 μM [57]. This compound belongs to the chemical class of fibrates, which includes several molecules able to activate PPARα (i.e. pemafibrate, bezafibrate, fenofibrate and clofibrate [57]). As ciprofibrate differs in size compared to larger agonists (e.g. pemafibrate or GW7647), the corresponding crystal structure (PDB code 6LX5) shows two molecules simultaneously bound to PPARα, one in the Arm I pocket and the second within the so-called Arm X region of the binding cavity (see Figure 7 for the pocket definition).
In order to examine the putative impact of the second molecule of ciprofibrate on protein activation, two separate protein-ligand systems were investigated. The first protein-ligand complex had one molecule bound to PPARα (Arm I), whereas the second one had two molecules bound (Arm I and Arm X). As done for GW7647, for each system four replicas were run (R1-R4) with independent starting velocities to provide statistically robust data. As shown in Table 6, in both simulation setups R1 used an IFD docking pose as starting position, whereas R2 started from a HADDOCK docking pose and R3-R4 simulations were initiated from the crystal structure containing two ciprofibrate molecules (PDB code 6LX5).
Protein stability and ligand movement were first monitored through the respective RMSD values. RMSD values of the protein backbone of both systems ranged from 1.5 Å to 3 Å, similar to the simulations with GW7647. If one molecule is bound to PPARα, RMSD values for ciprofibrate bound to the Arm I pocket are in range of 1.5 Å to 3.5 Å. If, however, two molecules are bound, RMSD values of the molecule bound to Arm I are slightly lower (1.5 Å-3.0 Å; see Figure S68 and Figure S74). The second molecule bound to Arm X possesses higher RMSD values, from 2 Å to 6 Å. The larger ligand flexibility in Arm X might be ascribed to most residues of in this pocket belonging to the Ω-loop (Table 2), which is known to undergo large conformational changes [65]. RMSF values of both simulations, i.e. with one and two molecules, are comparable (see Figure S69 and Figure S75).
ProLIF and MDAnalysis provided an overview of interactions between PPARα and ciprofibrate (see Figure S70 and Figure S76). In contrast to GW7647, ciprofibrate has three oxygen atoms acting as H-bond acceptors and two chloride atoms which can either serve halogen bond donor or acceptor.
In both simulation setups, H-bonds to Q277, S280 and Y314 are present and appear to be stable. While GW7647 forms a transient H-bond with Q277 in one replica (see previous section), ciprofibrate is able to maintain that H-bond in all replicas independently of whether 1 or 2 molecules are bound. As mentioned earlier, Q277 is located in H3 and interactions with residues in that helix seem to support protein activation [232]. Another clear difference compared to GW7647 is related to the H-bonds with H440 and Y464. While GW7647 showed stable H-bonds with these residues in all four replicas, in simulations with ciprofibrate both H-bonds to H440 and Y464 are less stable as well as more often water-mediated - irrespectively of the number of molecules bound to PPARα (see Figure S70 and Figure S76). The two available X-ray structures of PPARα with ciprofibrate (PDB codes 6LX5 and 6L37) showed different conformations of the ciprofibrate molecule located in Arm I (see Figure 25). The main difference between the two is a displacement of the carboxylate group, so that H-bond with H440 is lost; however, the oxygen atom attached to the phenyl group is able to replace this missing H-bond. The centroid structures obtained upon clustering analysis of the R1-R4 simulations mostly sampled the latter mentioned

binding pose (see Figure 26). R2 explores a second cluster, which contains 22.7% of MD frames (see Table S8; see Figure 24). This cluster representative has a similar position of the carboxylate group; however, the cyclopropane moiety as well as attached chlorine atoms have an orientation comparable to the one observed in X-ray structure PDB 6L37 (see Figure 25).

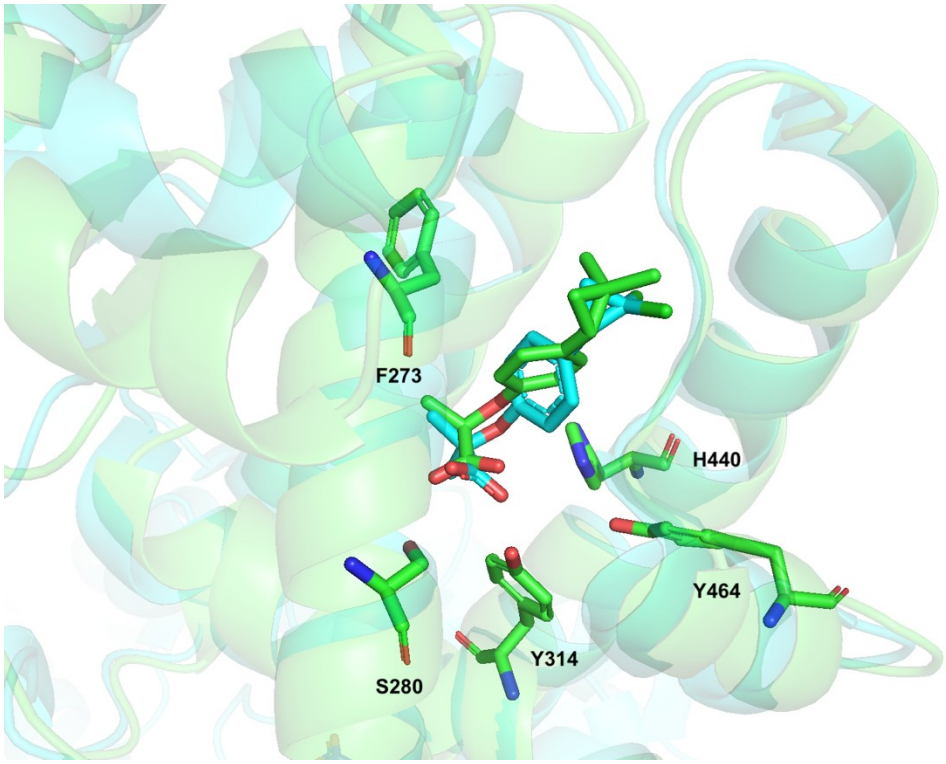


*Figure 24. Cluster centroid from the second cluster of simulation R2 of PPARα with one molecule bound to Arm I. Residues F273, S280, Y314, H440 and Y464 are represented as sticks and with carbon atoms colored as follows: (**Cyan**) X-ray structure (PDB Code 6L37) and (**Green**) R2: Cluster 2 – 22.7%.*

As mentioned above, the most remarkable feature of ciprofibrate compared to GW7647 is that an additional H-bond to Q277 (H3) is formed and direct H-bonds to H440 and Y464 (H12) are mostly water-mediated. As shown by the respective centroid structures (see Figure 26), GW7647 has two methyl groups pointing into the Arm I cavity, which keep residue Q277 with the right orientation to form a H-bond with the ligand. Furthermore, this permits Q277 to form H-bonds with the backbone of H457 and A455 (as observed in PDB code 6KB3), which reduces fluctuation in the loop connecting H10/11 and H12 (see Figure S63, Figure S69 and Figure S75). In addition, such rotameric state of Q277 prevents water molecules from entering the binding site. While ciprofibrate molecules can form a H-bond with Q277, their methyl moieties are pointing into the Center pocket instead and thus they cannot prevent water molecules from entering the binding cavity and intercalate in the ligand H-bonds with H440 or Y464. Since Y464 (H12) plays an important role in PPARα activation [62], this could be one reason that explains that ciprofibrate has a worse EC_{50} value compared to GW7647.

The second molecule of ciprofibrate is bound to the Arm X region of the binding pocket (see PDB code 6LX5). As mentioned above, RMSD values are significantly higher (up to 6 Å) than the ones for the Arm I molecule, which also explains that most interactions are transient (see Figure S77). The Arm X region is more solvent exposed and thus the ligand is more flexible, as well as interactions are more likely to be perturbated by water molecules. Indeed, H-bonds with backbone atoms of L254 (Arm X), A256, K257,

L258 (Arm X), L331 and G335 are either of short nature or water-mediated with an occurrence of up to 20%. Furthermore, L331 forms water-mediated H-bonds with the ligand, as observable for the larger size GW7647 ligand. A333, on the other hand, appears to form more stable interactions, since, besides water-mediated H-bonds, direct ones are also present (R1-R4). T279 is a residue located in H3 and forms direct H-bonds in R1 (70%), R3 (37%) and R4 (82%), but water-mediated ones in R2 (25%) and R4 (35%). Y334 also seem to be an important interaction partner of ciprofibrate, considering that in R1-R4 H-bonds are present between 70% and ≤ 87% of the time (considering both direct and water-mediated H-bonds; see Figure S77 and Figure S79). Y334 is located between S3 and S4, in vicinity to the P-site of PPARα. It seems that H-bonds with this particular tyrosine residue are stabilizing the P-site, since a higher interaction frequency results in lower RMSF values of that region (see Figure S69 and Figure S75). This is also in agreement with GW7647 simulations, where none of the replicas show a H-bond with Y334 and high RMSF values of this protein region are observed.

The crystal structures of PPARα in complex with Arm I binders show that F273 is pointing towards the Arm II region of the PPARα binding pocket, thus adopting an "open" conformation with respect to Arm I (see Figure 23). Analysis of the F273 sidechain dihedral angles (X_1 and X_2) along the MD simulations with ciprofibrate revealed that that the second molecule bound to sub-pocket Arm X reduces rotation of the phenyl moiety of F273 (see Figure S72 and Figure S80), thus hinting at cooperativity between the Arm X and Arm I bound molecules.

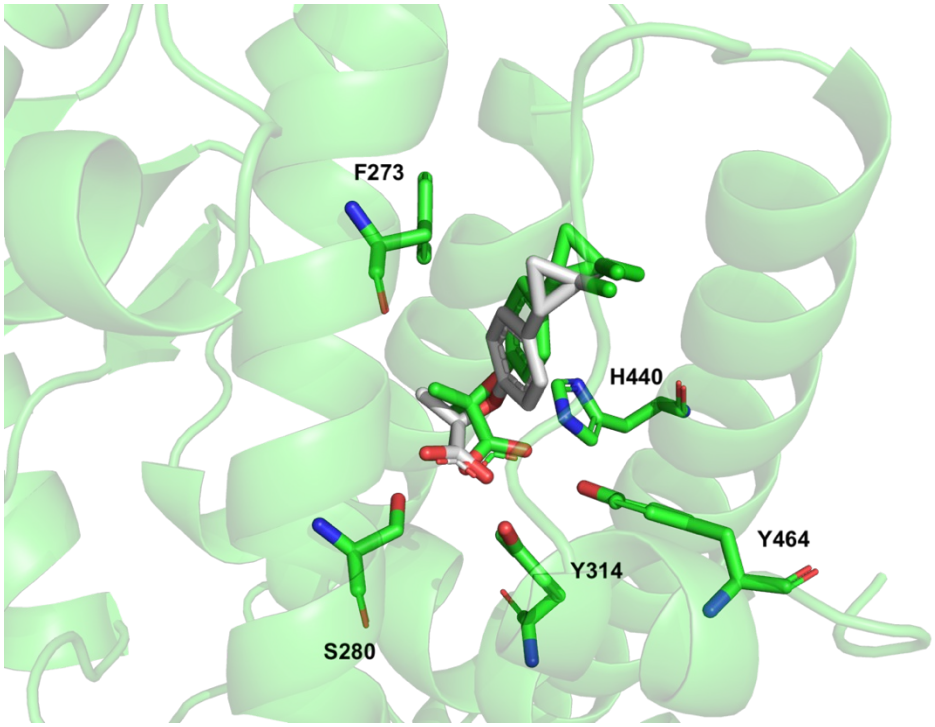


Figure 25. Crystal structures of ciprofibrate. The protein backbone and residues (F273, S280, Y314, H440 and Y464) are shown in green and as cartoon or sticks representation, respectively. The two conformations of ciprofibrate in Arm I are shown with grey (PDB code 6L37) and green (PDB code 6LX5) carbon atoms

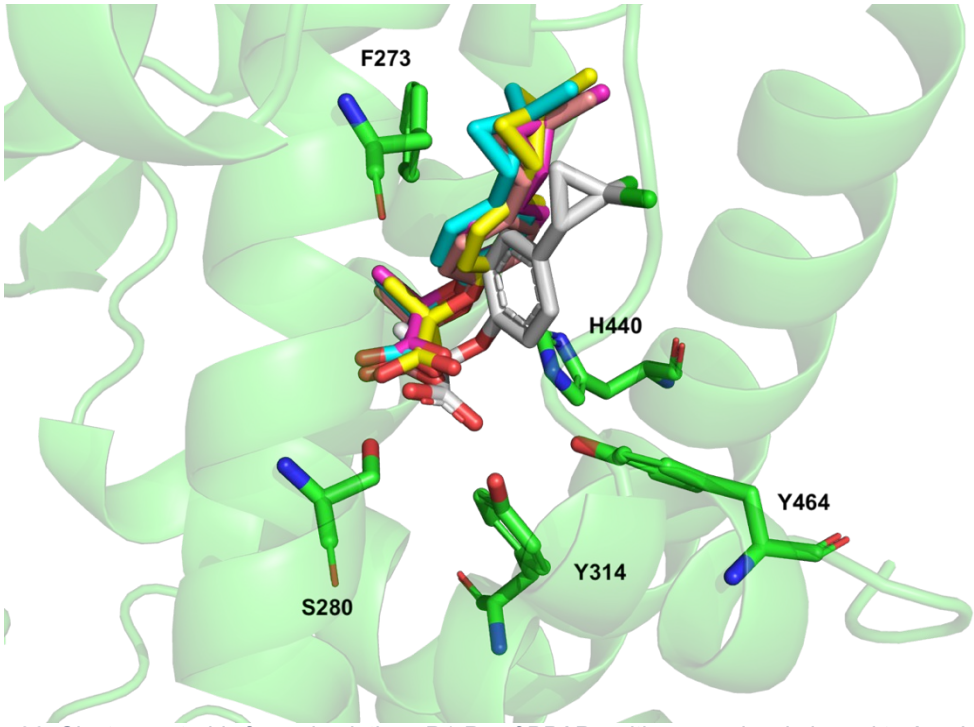


*Figure 26. Cluster centroids from simulations R1-R4 of PPARα with one molecule bound to Arm I. Residues F273, S280, Y314, H440 and Y464 are represented as sticks and with carbon atoms colored as follows: (**Grey**) X-ray structure (PDB Code 6L37), (**Cyan**) R1: Cluster 1 – 98.2%, (**Pink**) R2: Cluster 1 – 71.5%, (**Yellow**) R3: Cluster 1 – 91.6%, (**Salmon**) R4: Cluster 1 – 93.8%.*

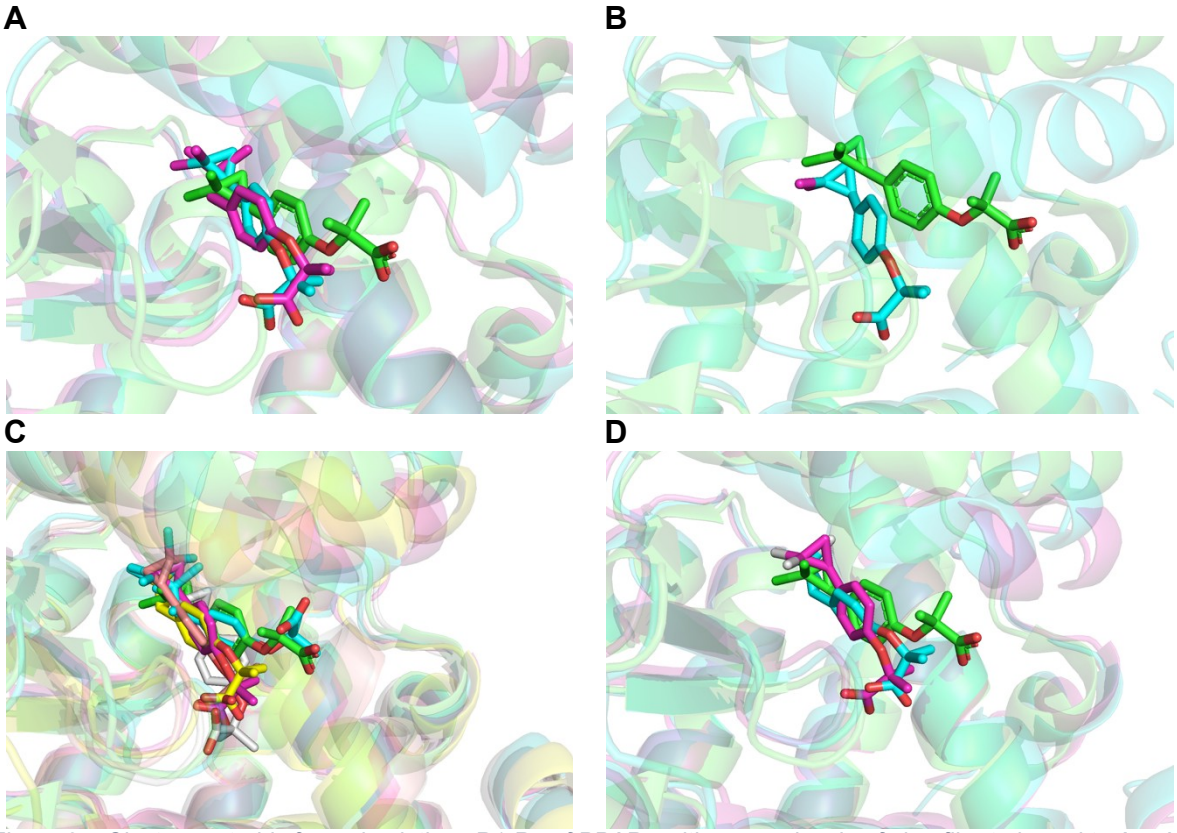


*Figure 27. Cluster centroids from simulations R1-R4 of PPARα with one molecule of ciprofibrate bound to Arm X. Protein structures are shown as cartoon representation and the ligand is shown as licorice, with carbon atoms in green (crystallographic pose, PDB code 6LX5) or cyan/pink (centroid of the first/second cluster sampled in the simulations). (**A**) R1: Cluster 1 – 62.3%, Cluster 2 – 13.7%; (**B**) R2: Cluster 1 – 85.8%, (**C**) R3: Cluster 1 – 15.3%, Cluster 2 – 12.9%, Cluster 3 – 9.4% (yellow), Cluster 4 – 8.5% (orange), Cluster 5 – 8.3% (grey), (**D**) R4: Cluster 1 – 53.8%, Cluster 2 – 11.3%.*

5.2.2.3 Gemfibrozil

Gemfibrozil is also an agonist of PPARα and has an experimental determined EC_{50} value of > 300 μM. To the best of my knowledge, there is no X-ray structure available in which Gemfibrozil is bound to PPARα. Kamata et al. were only able to obtain crystals together with GW9662; however, the electron density of gemfibrozil was not clear enough to determine whether it binds to Arm I or the Center/Arm II/Arm III pockets. Docking was not able to discriminate the preferred binding mode either, as comparable docking scores were predicted for gemfibrozil bound to either pocket. Therefore, MD simulations were performed to ascertain the more favorable binding mode of gemfibrozil. Furthermore, it is not clear if multiple binding pockets of PPARα can be occupied from gemfibrozil, as proposed for fibrates [57]. In experimental structures of agonist similar in size, two possible binding modes were observed for PPARα: (i) one molecule bound to the Arm I and the second to the Arm X pocket (see PDB codes: 6LX5 and 6LX4) or (ii) that one molecule is bound to the Center/Arm II/Arm III and Arm X pocket (see PDB code: 6KBA). Hence, a total of five systems were prepared, as shown in Table 6, with either one or two molecules of gemfibrozil bound to the receptor and occupying either the Arm I or the Center/Arm II/Arm III pockets (alone or together with Arm X). The fifth simulation covered the latter case (observable in PDB structure 6KBA), in which the second molecule of gemfibrozil is adopting an alternative binding mode in Arm X.

Protein backbone RMSD values of the simulations with gemfibrozil is placed into Arm I alone range from 1.5 Å to 3 Å in all four replicas, similar to the ciprofibrate simulations. Ligand RMSD values reach from 1 Å to 4 Å and the ligand poses seem to be converged (as evidenced by the plateau reached by the ligand RMSD, see Figure S83).
If one molecule of gemfibrozil is located within the Center/Arm II/Arm III region of the binding pocket, the protein backbone movement also ranges from 1.5 Å to 3 Å. The ligand RMSD values, however, reach values up to 7 Å in R1-R2 (see Figure S89). The starting structure of these two simulations was the HADDOCK pose, which placed the phenyl moiety into sub-pocket Arm III (Figure S60). During the course of these simulations, gemfibrozil adjusted itself so that the phenyl ring moved towards Arm II, resulting in higher RMSD values. Instead, the starting (Glide) pose of simulations R3-R4 has the gemfibrozil phenyl ring already placed in Arm II, where it remains for the whole simulation (400 ns) time.
RMSF values of both systems with only one molecule bound (to either Arm I or Center/Arm II/ Arm III) exhibit comparable values, with the exception of residues between H10/11 and H12, which show less movement when gemfibrozil in the Center/Arm II/Arm III region. However, that was expected, considering that, if a molecule is located in the Arm I pocket, mutual adaption of residues and ligand is taking place. In particular, spatial space in Arm I is more restricted compared to the Center/Arm II/Arm III pocket so that rearrangements of the protein structure are needed in order to accommodate gemfibrozil.
From the chemical structure point of view, gemfibrozil is able to serve as H-bond acceptor as the molecule possesses three oxygen atoms (see Figure S56). Gemfibrozil is, in contrast to ciprofibrate, a more flexible compound, as an acyl chain is connecting the functional groups (see Figure S56). Thus, it was expected that the ligand would display higher fluctuations compared to other known ligands binding to

the Arm I pocket. Furthermore, an EC_{50} of >300 µM is hinting that interactions towards PPARα are less stable. Despite having structural differences compared GW7647 and ciprofibrate, gemfibrozil is able to form similar H-bonds, respectively. If bound to the Center/Arm II/Arm III region, direct H-bonds to S280, Y314, H440 and Y464 are detectable. The H-bond to residue Y464 seem to have a more transient nature, since the occurrence (0.14 to 0.38) is smaller compared to residues S280, Y314, H440 (>0.90) (see Figure S91).
The same H-bonds are observable if gemfibrozil is located in Arm I; however, those H-bonds seem to be less stable compared to the ones where gemfibrozil is bound to the Center/Arm II/Arm III pocket (see Figure S85 and Figure S91). Nonetheless, binding of gemfibrozil in Arm I allows the formation of an additional direct H-bond between the third oxygen atom next to the phenyl ring and Q277 with an occurrence of >0.5 (in simulations R1 and R3-R4).

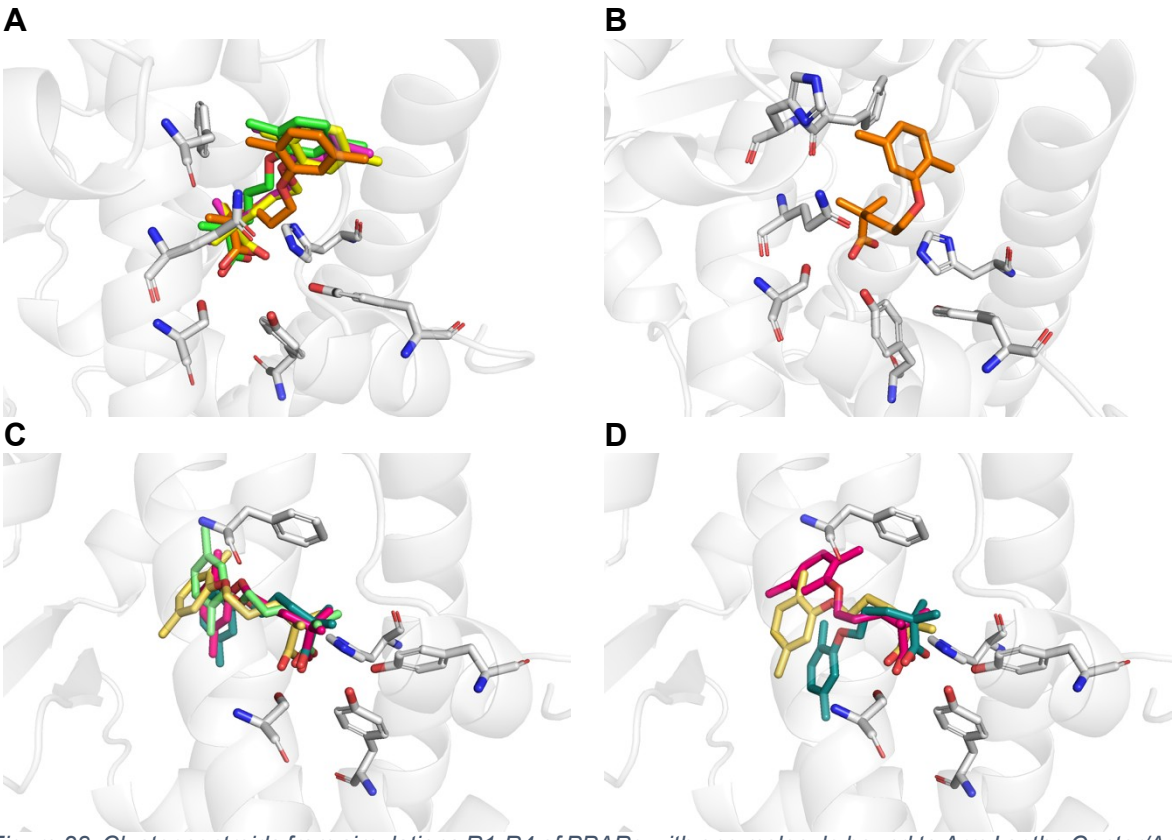


*Figure 28. Cluster centroids from simulations R1-R4 of PPARα with one molecule bound to Arm I or the Center/Arm II/Arm III pocket. Residues F273, S280, Y314, H440 and Y464 are represented as sticks. (**A**) 1st cluster of Arm I (R1-R4), (**B**) 2nd cluster of R2, (**C**) 1st cluster of Center/Arm II/Arm III (R1-R4), (**D**) 2nd cluster of R2-R4.*

Nonetheless, since detectable interactions could not strongly support one pocket over the other one, MM-GBSA calculations were carried out (see Table 8). As shown below, the average ΔH values are similar for both systems with only one gemfibrozil molecule bound.

Table 8. Results of MM-GBSA calculations of Gemfibrozil.

Sub-Pocket	Replica	MM-GBSA (ΔH) in kcal/mol	Standard Deviation (SD) in kcal/mol
Arm I	R1	-48.56	6.39
	R2	-42.22	2.04
	R3	-36.40	3.69
	R4	-38.87	4.49
Center/Arm II/Arm III	R1	-43.73	4.08
	R2	-41.52	4.01
	R3	-41.15	3.69
	R4	-40.98	4.08
Arm I + Arm X	R1	-40.65	5.06
	R2	-43.14	4.71
	R3	-42.24	5.11
	R4	-42.57	1.14
Center/Arm II/Arm III + Arm X	R1	-31.59	4.30
	R2	-33.70	4.57
	R3	-39.89	4.13
	R4	-43.75	3.97

Next, I investigated whether the presence of a second molecule bound to Arm X could shift the binding preferences of the first molecule. In the following, I am referring to gemfibrozil bound to Center/Arm II/Arm III and Arm X as model 1 (M1) and to gemfibrozil bound to Arm I and Arm X as model 2 (M2).
RMSF values of C_α backbone atoms of M1 and M2 are comparable with one exception in which residues belonging to the P-site of M1 show less movement (see Figure S96 and Figure S105).
Both systems with two molecules of gemfibrozil bound showed similar protein backbone movement with RMSD values of up to 3 Å. In general, molecules bound to the primary pocket, i.e. either Arm I or Center/Arm II/Arm III, show lower RMSD values compared to molecules in Arm X. This is not unexpected, since the latter binding pocket is composed of the more flexible Ω-loop (see Figure S95 and Figure S104). In both M1 and M2 sets of simulation, molecules bound to the primary binding site show RMSD values up to 3.5 Å. However, a change in RMSD is seen in R4 of M2 (molecule in Arm I), where after 130ns gemfibrozil rearranges, so that values up to 6 Å are reached at the end of the simulation. This is due to the phenyl ring moving between H10/11 and the unstructured region connecting H10/11 with H12, large RMSD values of up 20 Å are also observed for R3 of M2, due to a significant change within the binding pose, though in this case for gemfibrozil bound to Arm X (see Figure S95). In that case, gemfibrozil is rearranging itself so that is reminiscent of the binding pose of WY14643.
Regarding the protein-ligand interactions, some differences are observed depending on the docking pose used to start the simulations. Among the M1 simulations, R1 and R2 show the phenyl ring occupying Arm III. Both simulations were started from a pose generated through HADDOCK, in which the phenyl group in Arm III. Unlike the previously described simulation with only one gemfibrozil molecule, in which the phenyl group can change pocket during the simulation, the presence of a second molecule in Arm X hinders that movement. As a consequence, important interactions such as H-bonds to S280, Y314, H440 and Y464 are showing a less frequent occurrence (see Figure S107). R3 and R4 in M1, on the other hand, had a starting

pose generated from Glide which placed the phenyl ring in Arm II so that no major adjustment had to take place. Thus, H-bonds had a similar occurrence as if one molecule was bound (see Figure S105 and Figure S89). This is also reflected in the MM-GBSA calculated binding energies in Table 8: R1 and R2 have less favorable values compared to R3 and R4.

In the M2 simulations, direct H-bonds seem to be more stable in comparison to simulations in which only one molecule was bound to Arm I of PPARα (see Figure S83 and Figure S96). As shown in Table 8, ΔH values are indeed more stable, which gives further support the hypothesis that a second molecule bound to Arm X can help to stabilize the primary molecule bound to Arm I.

The molecule in Arm X in both M1 and M2 simulations does form stable ionic bonds with K257, which is located within the Ω-loop. In M2, other residues in the Ω-loop also form transient H-bonds with gemfibrozil, i.e. A256, L258 and N261 (direct and water-mediated). Moreover, E282 is able to form a water-mediated H-bond independently from the primary binding site occupied by gemfibrozil. Y334, a residue which also formed H-bonds with ciprofibrate, forms transient H-bonds in M2 (R1, R2, R4) and a stable H-bond in M1 (R1). T279, however, is only forming a weak H-Bond in M2 (R2) if Arm I is occupied.

Altogether, gemfibrozil show similar binding poses in Arm X whether a molecule is bound to the Arm I or Center/Arm II/Arm III binding pocket. This behavior is not surprising, since gemfibrozil possess a bulky tail, i.e. a phenyl ring with two attached methyl groups, which acts as an anchor to hold the molecule in place.

If no molecule of gemfibrozil is located in the Center/Arm II/Arm III pocket (i.e. molecules are located in Arm I and Arm X), respective molecule in Arm X possesses more freedom to move so that a direct H-bond to T279 can be formed (R2 of M2).

This behavior is not observable in simulations of M1, since ligand-ligand interactions are restricting mutual movement of both molecules through which no direct H-bond towards T279 can be formed. I would like to note that here that, out of the eight simulations with gemfibrozil bound to Arm X (four for M1 and four for M2), one (R3 of M2) exhibits a different behavior. The ligand undergoes significant changes compared to the starting pose. As shown in Figure 28, this outlier binding pose is reminiscent of the conformation adopted by the second molecule of Wy14643 bound to PPARα in the corresponding crystal structure (PDB code 6KBA). In order to investigate this alternative binding mode, further MD simulations were performed, using a starting pose similar to the one of Wy14643. These simulations, however, showed that this pose of gemfibrozil is not stable, resulting in either ligand dissociation or rearrangement to return to the previously described orientation. Thus, I discarded that gemfibrozil could adopt a Wy14643-like binding mode in Arm X.

5.2.2.4 Cinnamic Acid

As mentioned before, it was shown that cinnamic acid is able to activate PPARα [55]; however, structural information about its binding mode(s) is missing. Therefore, here I performed multiple MD simulations with different pocket occupancies, as done for gemfibrozil (see Table 6), as experiments indicated that multiple cinnamic acid molecules can bind PPARα (Hill slope of 12.89) [55], binding of more than one cinnamic acid molecule is likely.
As shown in Figure S111 and Figure S117, MD simulations with one molecule bound to PPARα seem to be converged; in particular protein backbone RMSD values reach a maximum of 3 Å, which is similar to simulations with ciprofibrate (see above). In general, ligand movement is higher compared to ciprofibrate, which is not unexpected considering the smaller size of cinnamic acid. If cinnamic acid is located in Arm I, RMSD values up to 6 Å are observable. Even higher values are seen if cinnamic acid is bound to the Center/Arm II/Arm III region. For this latter set of simulations, R3 shows the lowest deviation with respect to the starting docking pose with values up to 4 Å. In simulations R1 and R4, cinnamic acid adjusts itself more, as RMSD values between 6 Å and 8 Å were detected. In R2 the highest deviation from the starting pose was observed with displacement of up to 15 Å. Further discussion about the flexibility of cinnamic acid is included below in terms of protein-ligand interactions.
RMSF values of both set ups are almost similar values with an exception if cinnamic acid is bound to Arm I. Here, residues belonging to the Ω-loop show less fluctuation with up to 2 Å values, compared to 6 Å if cinnamic acid is located in the Center/Arm II/Arm III pocket.
Cinnamic acid is a precursor of hydroxycinnamic acids, since it is missing hydroxyl groups attached to the phenyl moiety. Nevertheless, H-bonds can still be formed due to the carboxylate group (see Figure 4). Moreover, due to cinnamic acid being a conjugated system a planar geometry is preferred over other geometries. Furthermore, cinnamic acid is small molecule in which the phenyl moiety and the carboxylate group define a significant part of the compound. As a consequence, cinnamic acid's binding pose is either dominated by the carboxylate, i.e. through H-bonds, or by the phenyl ring, i.e. through hydrophobic or pi-stacking interactions. If cinnamic acid is bound in Arm I, favorable interactions are either an aromatic H-bond with Q277, stacked pi interactions with F273 (R4: 41%) or T-shaped pi-stacking interactions with H440 (R1: 86%). Pi-stacking with F273 is only possible when the phenyl ring of cinnamic acid forms the aforementioned aromatic H-bond with Q277, which also makes room for F273 to change its rotameric state and relocate into Arm I (Figure S113 and Figure S115).
H-bonds with the carboxylate group, on the other hand, are formed with S280 and Y314 in all four replicas with an occurrence ≥ 0.7. H-bonds towards H440 and Y464 are formed in R2-R4 and present 41% to 75% or 17% to 73% of the time, respectively (see Figure S113). Due to the phenyl ring preferring to form interactions with Q277 and cinnamic acid to be as planar as possible, the carboxylate is cannot to be placed in an optimal way which allows formation of two bifurcated H-bonds with S280, Y314, H440 and Y464, as observed for the previously described GW7647, ciprofibrate and gemfibrozil.
In simulations in which cinnamic acid is occupying the Center/Arm II/Arm III, H-bonds seem to be less stable. R1, R3 and R4 show H-bonds with S280 and Y314 (with frequencies from 83% to 100%); however, in R2 none of these residues are forming H-bonds with cinnamic acid. I ascribed this difference to cinnamic acid moving towards

Arm II/Arm X in the R2 simulation, so that H-bonds with T279 and A333 are formed (similar to ciprofibrates second molecule bound to Arm X, see section 5.2.2.2). This relocation of cinnamic acid results from the earlier mentioned dualism in which the hydrophobic part of cinnamic acid prefers to be placed into Arm III/Arm II/Arm X, i.e. more hydrophobic regions, whereas the carboxylate rather forms H-bonds with S280, Y314, H440 and Y464. In addition, cinnamic acid does not possess hydrophobic groups pointing into Arm I so that it is easier for Q277 to change its rotameric state and form H-bonds with the respective ligand oxygens (R1: 28% and R3: 41%). Due to that additional H-bond, cinnamic acid seem to be more stable in R1 and R3, which in turn allows more frequent H-bond formation with H440 and Y464 (see Figure S119). Q277 is not forming a H-bond in R4, since cinnamic acid relocates itself into Arm I. Thus, the phenyl moiety of cinnamic acid is preventing Q277 from changing its rotameric state and a similar picture compared to Arm I emerges.
This more stable behavior of CNA in Arm I compared to the Center/Arm II/Arm III is also reflected in the respective MM-GBSA calculated binding energies. As shown in Table 9, ΔH values of cinnamic acid located in this pocket are more favorable and/or show narrower standard deviations than when CNA is bound to the Center/Arm II/Arm III region. When cinnamic acid is bound to the Center/Arm II/Arm III pocket, R1 and R3 have values in the range of simulations considering Arm I. Simulation R2, in which CNA relocates into Arm II/Arm X shows significantly less favorable values. This observation together with the fact that X-ray structures of PPARα and fibrates show either one molecule (in Center/Arm II/Arm III or Arm I pockets) or two molecules (in one of the former pockets and Arm X) strongly suggests that the former site has higher affinity than the latter. Also, lower values are observed in R4 due to the relocation of CNA into Arm I.

Table 9. Results of MM-GBSA calculations of cinnamic acid.

Sub-Pocket	**Replica**	**MM-GBSA (ΔH) in kcal/mol**	**Standard Deviation (SD) in kcal/mol**
Arm I	R1	-24.52	3.08
	R2	-25.14	4.45
	R3	-25.14	6.07
	R4	-30.49	6.24
Center/Arm II/Arm III	R1	-25.62	6.86
	R2	-13.45	4.63
	R3	-26.31	5.62
	R4	-19.11	4.08
Arm I + Arm X	R1	-25.29	5.32
	R2	-25.17	4.98
	R3	-19.68	3.75
	R4	-19.95	4.47
Center/Arm II/Arm III + Arm X	R1	-28.17	0.92
	R2	-28.77	8.65
	R3	-35.44	4.78
	R4	-19.19	3.96

In order to investigate if a second molecule is able to bind to Arm X, thus further contributing to PPARα activation and/or changing the binding preference of the first CNA molecule, further simulations were performed (see Table 6). Hereafter I will refer

to simulations with one molecule of CNA bound to Arm I and one molecule bound to Arm X as model 1 (M1). Simulations investigating the possibility of CNA simultaneously bound to the Center/Arm II/Arm III or Arm X pocket are referred to as model 2 (M2). Protein backbone RMSD values of both set-ups are with values up to 3 Å in the range of the other simulations (see above). RMSD values of CNA bound to different pockets are in fact different, however, simulations seem to be converged (see Figure S123 and Figure S132).
As for other simulations, no significant differences in RMSF values for both set-ups were observed (see Figure S124 and Figure S134).
Interestingly, CNA in Arm X showed stable and similar binding poses in both M1 and M2 (see Figure 29). This behavior is also reflected through detected interactions between this second molecule of CNA and PPARα. Direct H-bonds with T279 and A333 are present in M1 (≥ 56%) as well as in M2 (≥ 79%). In M1 also a transient H-bond with K257 is formed, whereas in M2 transient H-bonds with T334, K257 and T253 are detected (see Figure S126 and Figure S135). Considering that the carboxylate group of CNA is quite solvent exposed in Arm X, it is not surprising that also water-mediated H-bonds are present. In particular, such interactions are formed with the side-chains of T279 and Y334, as well as with backbone atoms of A250, T253 and L331 (see Figure S128 and Figure S137). Moreover, these interactions were also observed for the second molecule of ciprofibrate, which further supports that also cinnamic acid is also likely to bind to the Arm X pocket of PPARα.

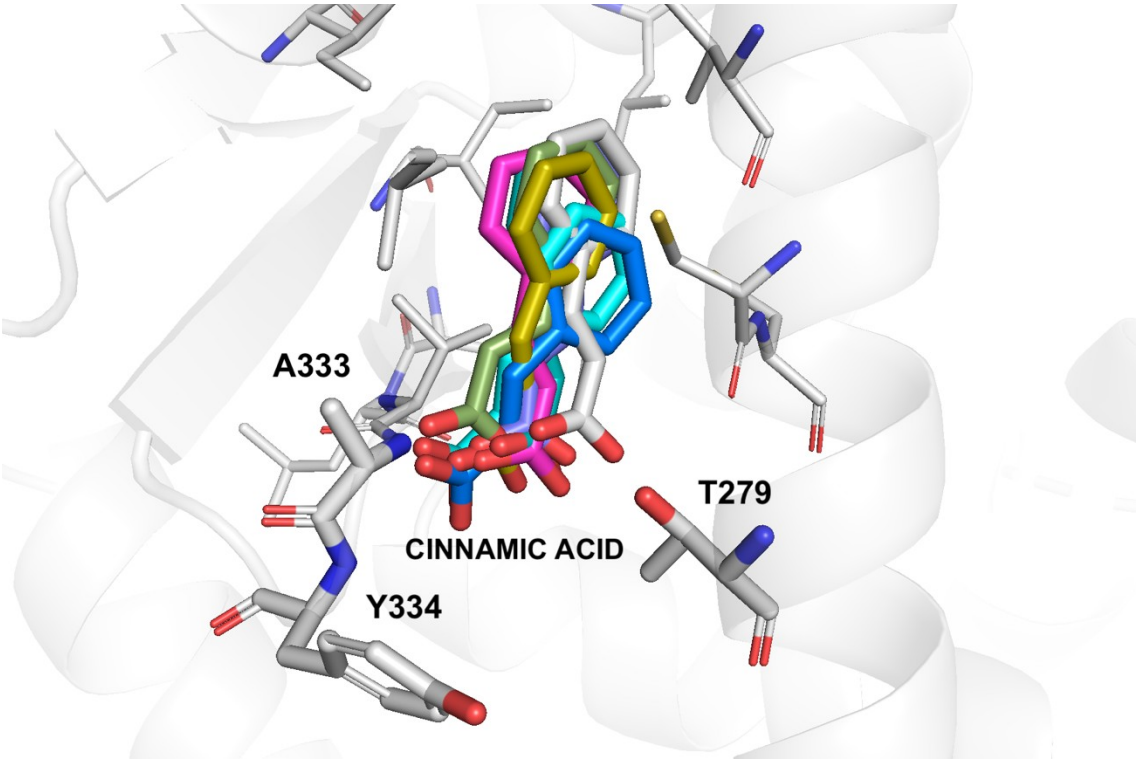


Figure 29. Cluster representatives of cinnamic acid bound to Arm X of M1 and M2. The protein backbone is shown as cartoon and surrounding residues as stick representation. Residues participating in H-bonds during performed MD simulations are shown as bold sticks and labelled accordingly. Cinnamic acid is also visualized in stick representation and different colors for its carbon atoms indicate different replicas; the representative ligand structures for the M1 simulations are in cyan (82.4%), pink (54.9%), grey (84.1%), and purple (84.4%) for R1-R4, respectively, whereas for M2 the colors used are aquamarine (74.2%), dark green (78.7), sand-yellow (86.0%), and blue (96.5%) (R1-R4).

In order to discriminate if Arm I or Center/Arm II/Arm III is favored as first binding site in the presence of a second molecule of CNA at Arm X, formed interactions, binding

poses and MM-GBSA-based energies were investigated in more detail, as done for gemfibrozil.

R1 and R2 of M1 seem to be as preferable as the simulations with one molecule and display more favorable energies than R3 and R4 (see Table 9). In R1 cinnamic acid does not experience much movement and as an aromatic H-bond with Q277 is stabilizing the phenyl ring. As a result, the carboxylate cannot be placed in an optimal way (see Figure 30 and the discussion above) and H-bonds with H440 and Y464 are less frequent than with S280 and Y314 (see Figure S125). This is also the case in R2; however, after around 150ns, F273 changes its rotameric state so that its side-chain points into Arm I (see Figure S129). That can be favorable if cinnamic acid can form pi-stacking interactions with this phenylalanine. In contrast, in R2 (after 150ns) and R4 (whole trajectory) the phenyl moiety of cinnamic acid is pushed backwards, so that it relocated between H10/11 and the unstructured region between H10/11 and H12 (see Figure 30). That movement is in fact making place for Q277 to form a H-bond with CNA; however, that also causes the loss of H-bonds with S280, resulting in less favorable MM-GBSA energy values for R4. In the case of R3, the phenyl ring is not pushed backwards but downwards instead, which also results in loss of H-bonds to H440 and Y464 (see Figure 30).

In M2 the rotameric state of F273 does not change with respect to the starting structure and thus the interactions between protein and the CNA ligand located in the Center/Arm II/Arm II region remain close to the initial docking pose (see Figure S138).

In general, interactions between CNA and PPARα are more stable if a second molecule is present, as shown in Figure S134. In R1-R3 direct H-bonds with Q277 (40% to 82%), S280 (89% to 100%), Y314 (91% to 100%), H440 (36% to 56%) and Y464 (50% to 83%) are observed. This increase in frequency is not surprising, since the second molecule is preventing relocation of the first (see Figure 31). MM-GBSA calculations further support the observation of more stable interactions in the presence of a second CAN molecule in Arm X, as ΔH values for CNA in the Center pocket are more favorable compared to simulation with one molecule (see Table 9). In R4, no H-bonds with residues Q277 and Y464 are detected. Moreover, interactions with S280 (18%), Y314 (36%) and H440 (20%) are reduced in a significant manner. Instead, they are replaced by a new H-bond formed with K358; however, as shown from MM-GBSA calculation, that binding pose is not as favorable as the previously described. In particular, the phenyl ring moves towards Arm III, which results in loss of H-bonds with S280, H440 and Y464. R1 and R2 still have lower MM-GBSA values compared to R3 because the phenyl ring is still able to move upwards, which can cause some H-bonds to break.

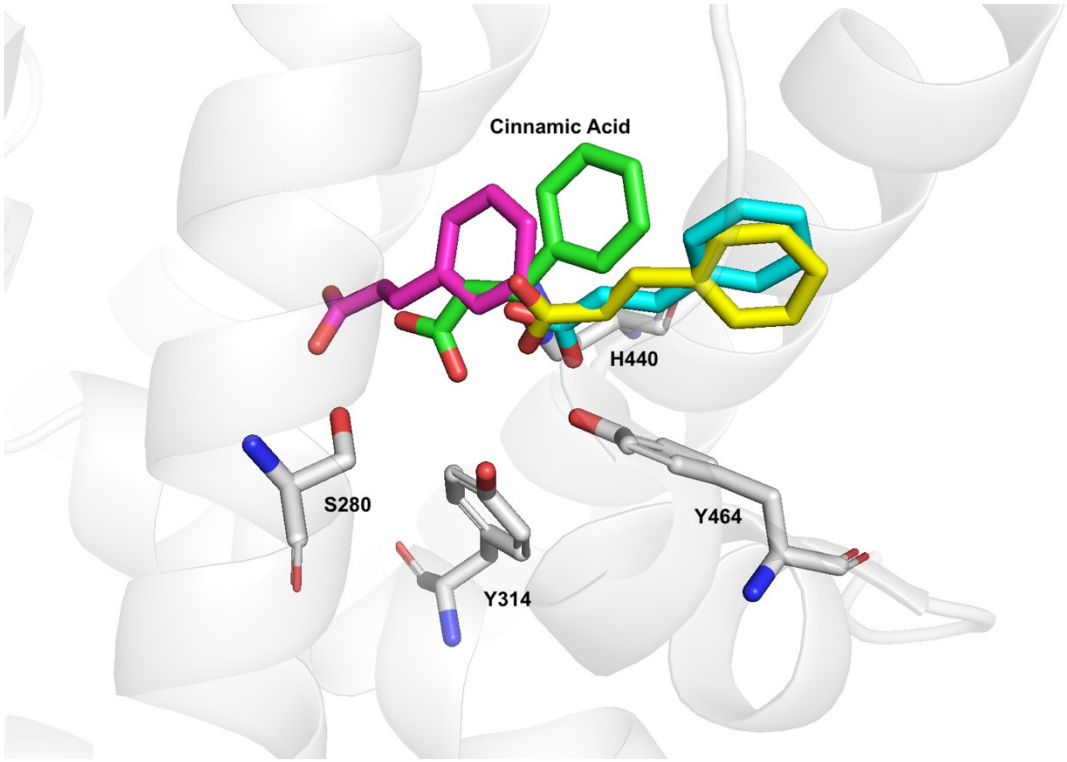


*Figure 30. Cluster representatives of M1 simulations, with CNA in the Center/Arm II/Arm III pocket. The protein structure is displayed as cartoon representation and residues S280, Y314, H440 and Y464 as sticks (**grey**). Cinnamic acid molecules are also shown as sticks and colors are representing respective replicas. (**Green**) R1: Cluster 1 – 74.2%; (**Cyan**) R2: Cluster 1 – 63.9%; (**Pink**) R3: Cluster 1 – 96%; (**Yellow**) R4: Cluster 1 – 93.7%.*

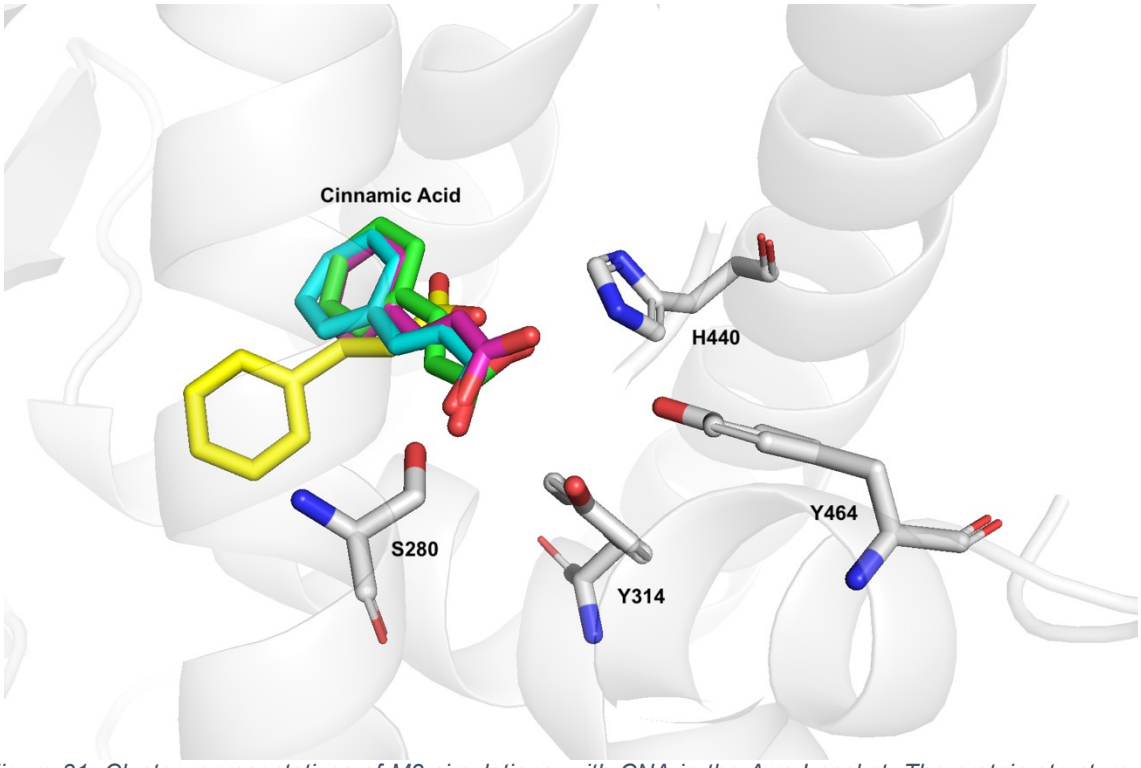


*Figure 31. Cluster representatives of M2 simulations, with CNA in the Arm I pocket. The protein structure is displayed as cartoon representation and residues S280, Y314, H440 and Y464 as sticks (**grey**). Cinnamic acid molecules are also shown as sticks and colors are representing respective replicas. (**Green**) R1: Cluster 1 – 87%; (**Cyan**) R2: Cluster 1 – 76.6%; (**Pink**) R3: Cluster 1 – 99.7%; (**Yellow**) R4: Cluster 1 – 98.4%.*

5.3 Conclusion

Several groups of compounds have been shown to be able to activate PPARα, for instance, fatty acids (e.g. stearic acid), fibrates (e.g. pemafibrate, gemfibrozil, fenofibric acid and ciprofibrate) or even smaller molecules, such as cinnamic acid. Such wide variety of agonists can be explained by PPARα bearing a large binding site subdivided into different pockets. One of the ligand features controlling which pockets are occupied is ligand size. For instance, while GW7647 spans most of the binding cavity (Center, Arm II and Arm II pockets), ciprofibrate (and other fibrates) can only fill either Arm I or Arm X and thus needs two molecules binding simultaneously to activate PPARα (see PDB codes 6KB3 and 6LX5, respectively, as well as simulations above). X-ray structures revealed that one molecule is located within the primary binding site, which means either Arm I and or the Center/Arm II/Arm III pocket, depending on the ligand. Despite the different binding modes, most agonists of PPARα share a common structural group, namely a carboxylate, which can form H-bonds with PPARα residues S280, Y314, H440 and Y464, located at the interface between the Center and the Arm I primary binding sites. Experimental studies indeed provided evidence that these four residues are particularly important for protein activation. Additional H-bonds, for instance, with residues in H3 have been shown to decrease fluctuations within these regions and thus been proposed to facilitate protein activation [57, 65]. The molecule bound in the secondary Arm X site interacts instead with residues in the Ω-loop, in line with protein activation through indirect stabilization of H12 [62].

In the case of gemfibrozil and cinnamic acid, there are no X-ray structures available of the corresponding PPARα complex and thus docking combine with MD simulations can be used to decipher their binding modes. As mentioned above, for gemfibrozil a co-crystal structure with GW9662 revealed electron density compatible with binding to the primary site, but assignment to either the Center or Arm I pockets could not be done. My MD simulations with one molecule of gemfibrozil and respective MM-GBSA calculations showed that binding to either Arm I or Center/Arm II/Arm III appear to be equally likely. Furthermore, gemfibrozil is a compound with similar size to other fibrates (e.g. ciprofibrate or fenofibric acid; see Figure S56) so that binding of multiple molecules cannot be excluded. As shown in PBD structures 6LX4 (PPARα in complex with fenofibric acid) and 6LX5 (PPARα in complex with ciprofibrate), PPARα binding pocket is large enough to accommodate multiple molecules simultaneously; in both cases one molecule is located in Arm I and one molecule in Arm II/Arm X. In X-ray structures of PPARγ even three molecules are bound to the protein structure, with none of them located in Arm I. Considering that small changes in the protein sequence and conformation can result in binding to different pockets, it is not surprising that MD simulations of gemfibrozil did not show any clear preference for one pocket over the other.

Therefore, I also ran MD simulation with two molecules of gemfibrozil bound (in the primary site and in Arm X). When gemfibrozil is bound to Arm I and Arm X, the presence of the second molecule in the secondary site improved H-bond frequencies and ligand stability in the primary site. This is in line with molecules similar in size and structure to gemfibrozil (such as ciprofibrate or fenofibric acid) also preferring Arm I as primary binding pocket when the secondary Arm X pocket is occupied. Similarly, gemfibrozil also showed cooperative binding effects if bound simultaneously to Arm X and the Center/Arm II/Arm III region (data not shown), which resulted in two consequences: (1) Due to the ligand-ligand interaction, gemfibrozil molecules seemed to be more rigid, which had disadvantages in case of R1 and R2 (i.e. molecules cannot

adjust themselves properly to the binding cavity) and (2) can have advantages in terms of dissociations kinetics (assuming that more interactions means longer residence time). However, considering that binding of two molecules may happen either simultaneously or consecutively, the data from simulations R1 and R2 is not enough to understand whether the aforementioned effects (1) or (2) would prevail.

Cinnamic acid is small molecule and an agonist of PPARα with unknown binding mode. The MD simulations I performed in this thesis suggest simultaneous binding of at least two cinnamic acid molecules to PPARα, in line with the experimentally measured Hill coefficient [55]. The molecule located in Arm X showed interaction fingerprints similar to ciprofibrate within the same pocket. Moreover, cinnamic acid in Arm X showed a stable binding pose independently of whether the first molecule is located in Arm I or the Center/Arm II/Arm III. Further evidence of CNA binding to Arm X is given from R2 (where one molecule is present, bound to the Center pocket) in which CNA switches from the Center/Arm II/Arm III region to the Arm X pocket. These observations are in line with *in vitro* experiments showing a Hill coefficient of 12.89 for CNA [55]. As seen for fenofibric acid, a molecule which showed cooperative binding of two molecules in PPARα and of three molecules in PPARγ, cinnamic acid, an even smaller molecule, could also bind to PPARα in a similar manner.

As for gemfibrozil, cinnamic acid showed similar results when located in either of the primary binding sites. Therefore, several possibilities can be considered to achieve activation of PPARα. As discussed, smaller molecules (in particular fibrates) tend to bind to PPARα in both Arm X and Arm I. That may also be the case for cinnamic acid as for instance MM-GBSA calculations showed comparable values between Arm I and the Center/Arm II/Arm III binding pockets. Moreover, in R4 (one molecule in the Center/Arm II/Arm III pocket) cinnamic acid relocated into Arm I, which might indicate a preference for Arm I as primary binding site. Nonetheless, additional replicas and/or longer simulations might be needed to provide further evidence. Instead, binding to the Center/Arm II/Arm III pocket was more preferable if a second molecule was present. In particular, the second molecule of cinnamic acid (located in Arm X) is preventing the first one (in that case present in the Center/Arm II/Arm III pocket) from relocating and thus typical interactions, i.e. H-bonds with S280, Y314, H440 and Y464, have a higher appearance.

As cinnamic acid is a precursor of hydroxycinnamic acids it is likely that also this group of compounds can act as agonists of PPARα. In particular, the attached hydroxyl groups could from additional H-bonds to stabilize respective binding poses, e.g. to T279 if bound to the Center/Arm II/Arm III or to Q277 in case of Arm I. This behavior could be investigated with further MD simulations, for instance, with caffeic acid. In that regard, the presented protocol could serve as blueprint. CGAs also showed favorable dockings scores which were comparable to known agonists of PPARα. Thus, it is plausible to assume that CGAs as well as the larger di-CGA compounds may activate PPARα. Nevertheless, further computational and experimental studies are needed to validate the proposed binding mode for CGAs and their effect on PPARα activation. Besides plain MD simulations, as I used in my thesis, more advanced computational methods, such as metadynamics, could help to obtain more accurate binding free energies and give some insights into dissociations kinetics [233]. Thus, a more complete picture of binding to PPARα could give hints to discriminate whether cinnamic acid (or gemfibrozil) bind to Arm I or the Center pocket.

6 Conclusions

Nowadays, a conscious nutrition is getting more attention and, since Covid-19, the awareness of leading a healthier lifestyle has been brought back into focus. That has not gone unnoticed by companies and thus respective marketing strategies shifted in order to highlight beneficial attributes of their products on human health. As a result, the interest grew in compounds present in food and beverages and their mode of action within the human body.

Coffee is one of the most consumed beverages around the world and a cup of coffee contains a mixture of hundreds of compounds, which can mediate diverse effects on human health. Nevertheless, despite being consumed on a daily basis, for most compounds a detailed mode of action is still missing.

Unfortunately, roasting of coffee beans increases abundance of a small organic molecule named acrylamide. In addition, that molecule is also formed during food processing at high temperatures (i.e. during backing and roasting) and responsible for neurotoxic effects upon cumulative exposure. As a consequence, the European Union established regulations in order to minimize concerns regarding potential risks.

From a chemical standpoint, acrylamide is an electrophile which reacts with nucleophilic residues such as cysteines. Synaptic proteins are especially rich in cysteine residues, which increases chances of acrylamide modification and thus neurotoxicity. Investigation through covalent molecular docking showed that acrylamide modification is favored if positively charged residues, i.e. lysine or arginine residues, are in vicinity of respective cysteines. Moreover, docking scores were able to discriminate between primary and secondary attachment sites of acrylamide characterized experimentally, demonstrating that covalent docking can serve as computational tool to predict acrylamide reactivity. Thus, additional targets could be predicted through covalent docking, which could facilitate identification of putative modification sites and thus increase the dataset of proteins potentially modified by ACR. This task is expected to be accelerated by the integration of recently developed AI-based structural prediction and docking methods. Moreover, information about putative ACR targets can serve as starting point for further computational and/or experimental studies, e.g. mass spectrometry and functional assays.

An example for beneficial health effects of coffee would be the group of chlorogenic acids, esters of quinic- and hydroxycinnamic acid that can promote antioxidant and anti-inflammatory properties, modulation of lipid and glucose metabolism, prevention of cardiovascular diseases and neuroprotective effects. However, a detailed description of the molecular mechanism(s) by which CGAs can have such beneficial effects on human health is absent.

Literature suggested that the peroxisome proliferator-activated receptor subtype alpha (PPARα) could serve as a potential protein target of chlorogenic acids, which could in turn explain the neuroprotective effects of that group of compounds. This is also supported by data available on cinnamic acid (i.e. a precursor of hydroxycinnamic acids), since its neuroprotective effects were demonstrated in AD and PD mouse models to be mediated through PPARα.

In this thesis, molecular docking provided support for chlorogenic acids being binders of PPARα. Both mono- and di-chlorogenic acids showed similar performance to known PPARα agonists, such as gemfibrozil, ciprofibrate or fenofibric acid, among other fibrates. In regard to molecular dynamics, I followed a step-wise approach, i.e. I started with giving molecular insights into the binding modes of gemfibrozil, a known agonist of PPARα, for which no experimental structural information is available. Afterwards, I

investigated cinnamic acid, a precursor of HCAs whose binding mode is also unknown. Both of these molecules showed favorable binding modes comparable to other known PPARα agonists for which crystal structures have been solved, such as ciprofibrate or fenofibric acid. In particular, my results support that multiple molecules of cinnamic acid can bind to PPARα simultaneously to stabilize PPARα and thus promote protein activation in a cooperative manner. However, further computational and/or experimental investigation are needed to investigate ligand cooperativity. In the case of chlorogenic acids, MD simulations with these larger molecules could validate if the predicted docking poses are indeed stable under physiological-like conditions. From the experimental side, it would also be necessary to prove hypothesized mechanisms, for instance, through coactivator recruitment or thermostability assays, as well as mutagenesis studies, which could verify that chlorogenic acids are in fact binding to PPRα as predicted in this thesis.

Besides PPARα, other proteins have been proposed to serve as targets of chlorogenic acids (e.g. matrix metalloproteinases MMP-2 and MMP-9, acetylcholinesterase and butyrylcholinesterase, carbonic anhydrase and alpha-amylase as well as alpha-glucosidase). Thus, the computational workflow presented in this thesis could serve as guideline to study those proteins. Moreover, additional computational techniques could facilitate the discovery of new, additional protein targets of CGAs. In particular, predictors based on ligand information (i.e. machine learning algorithms based on chemical similarity, such as SwissTargetPrediction and PharmMapper), structural data on common protein targets (i.e. reverse docking approaches using a library of known protein drug targets, such as ACID, CRDS or TarFisDock) or combining both types of ligand and structure-based approaches (LigTMap or GalaxySagittarius) could be used.

Altogether, this thesis has demonstrated that computational approaches, such as molecular docking and molecular dynamics, are appropriate tools to get insights into the biophysical and biochemical mechanisms by which coffee compounds have an impact on human health. Insights obtained through such theoretical studies can serve to guide subsequent experiments. Moreover, I showed that it has advantages to incorporate experimental evidence into theoretical approaches, such as molecular docking.

7 Appendix A

Appendix A is equal to the original Supplement Material 1 of reference [1]. Supplement Material 2 of the same publication, however, is not shown in this thesis and can be accessed online. (https://www.frontiersin.org/articles/10.3389/fphar.2023.1125871/full#supplementary-material).

7.1 Dataset of acrylamide protein targets

"*The list of 19 proteins emerging from our literature and chemical database search of acrylamide (ACR) targets (see section 4.2.1) is given in Table 4. Experimental structures are available for 17 out of the 19 target proteins in our dataset; when more than one structure was available, we chose the one at the highest resolution (see Table 4). For the two remaining targets (i.e. the dopamine transporter and the NEM-sensitive factor), we generated homology models of the human proteins, as explained in section 4.1.1. Quality evaluation of these models is discussed in chapter 4 (sections 4.1.1). In particular, Ramachandran plots were calculated to check whether the protein backbone geometry falls within allowed/favored regions; these plots are shown in Figure S1. In addition, we estimated the local quality values (Figure S2) using QMEANbrane for DAT and QMEANDisCo [83] for NSF. We would like to note here that QMEANbrane [84] has been trained to evaluate the quality of homology models for membrane proteins, such as the DAT, whereas QMEANDisCo [83] was developed for soluble proteins, such as the ATPase domain of NSF considered here.*

For each protein target, we have used the structures listed in Table S1 to analyze the physicochemical properties and location for each of the candidate Cys residues, as well as their corresponding microenvironment. Most of the cysteines in our dataset happened to be located in enzyme active sites (see Table 4). Most likely this reflects the more readily available purification methods and functional assays for enzymes compared to other protein functional classes (see also below). Cys residues with higher SASA and lower pK_a values in Table 4 are more accessible and acidic and thus potentially more reactive. However, SASA and pK_a calculations can have limited accuracy due to the dependency on the structure used to represent the protein and the poor performance of pK_a predictors for Cys [37]. Hence, we have also inspected residues in the vicinity of the candidate cysteines that could potentially favor Cys deprotonation (step 2 in Figure 3). His and Asp/Glu (in green and red, respectively, in Table S1) could deprotonate the Cys thiol group, whereas Arg/Lys (in blue) would stabilize the resulting thiolate. In addition, other H-bonding capable residues (in orange in Table S1) could also help make the Cys more acidic [189]."

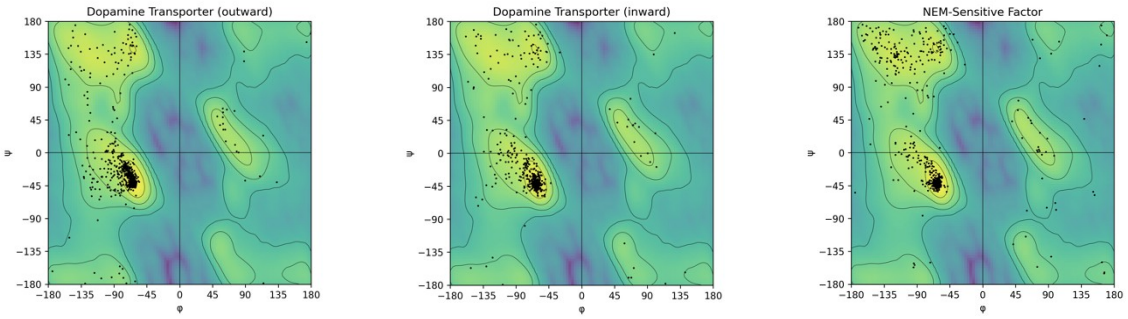


Figure S1. Ramachandran plots of the homology models built for the dopamine transporter, outward and inward conformations, as well as the NEM-sensitive factor (from left to right). The plots were generated using the RamachanDraw tool (https://pypi.org/project/RamachanDraw/), distributed under the MIT license.

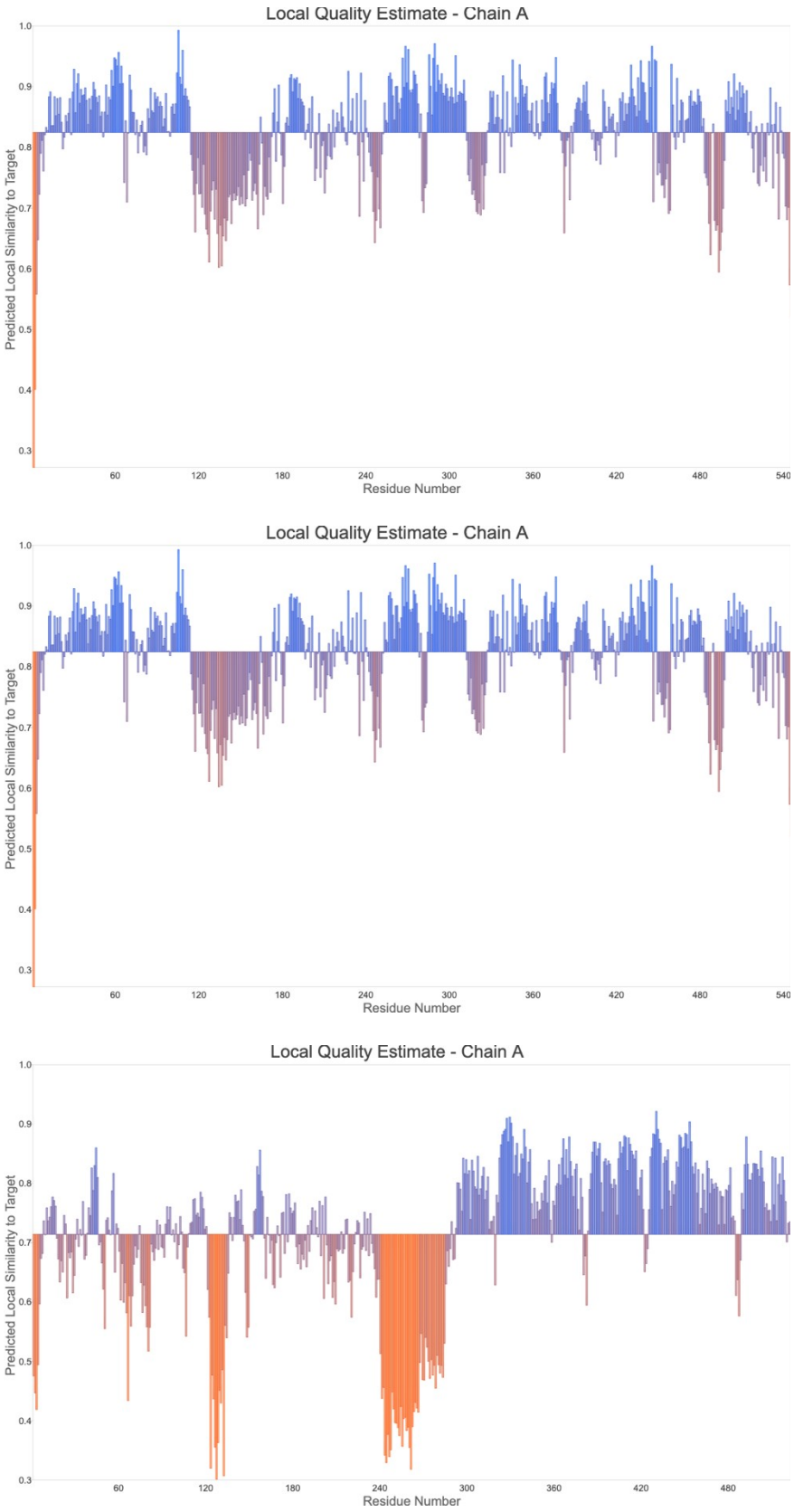


Figure S2. Local quality values of the homology models built for the dopamine transporter, outward and inward conformations, and the NEM-sensitive factor (from top to bottom). The plots were generated with the QMEAN webserver (https://swissmodel.expasy.org/qmean/).

Table S1. Protein microenvironment of the potential reactive cysteines. The analysis was performed using the same protein structures (listed here by their PDB codes) as in Table 4. Protein residues within a distance cutoff of 10 Å from the respective Cys were analyzed. Asp/Glu and Arg/Lys residues are displayed with blue and red shades, respectively, whereas His residues are in green and other H-bonding capable residues in orange. In addition, for residues within 5 Å from the candidate Cys, distances with respect the Cys sulfur atom are specified. For candidate cysteines surrounded only by hydrophobic residues or surface water molecules, a gray shade (and the label n.a., not applicable) was used.

#	Protein name	PDB code	Cysteine	Environment
(1)	Albumin	6HSC	**34**	Lys41
				Asp38 (5.65 Å)
				His39 (4.62 Å)
				Thr79 (4.54 Å), Tyr84 (2.98 Å)
(2)	Alcohol Dehydrogenase	1U3W	170	Lys369
			240*	Glu62 (5.98 Å)
				Lys233 (4.97 Å)
(3)	Aldolase	1QO5	134*	Arg133 (6.83 Å)
				Cys177 (4.07 Å), Asn180 (5.32 Å)
			239*	Lys241
				Asp195 (7.01 Å), Asp197 (5.74 Å)
				Tyr243 (4.43 Å)
			268	n.a.
			289*	n.a.
(4)	Creatine Kinase	3B6R	74	n.a.
			141	Asp78 (5.08 Å), Glu80 (7.77 Å), Glu150 (4.11 Å)
				His145 (4.58 Å)
				Ser49 (2.91 Å)
			146	Arg151 (4.14 Å), Arg209 (5.28 Å), Arg215
				Asn230 (6.27 Å)
			254	Arg135
			283	Thr251 (6.58 Å), Thr258 (6.46 Å)
				Arg96, Arg132, Arg236, Arg341
				Glu232, Asp233 (8.73 Å)
				Ser285 (3.26 Å), Asn286 (5.32 Å)
(5)	Dopamine Receptor	3PBL	**114**	Asp110 (7.57 Å)
				Ser117 (6.52 Å)
(6)	Dopamine Transporter (inward)	HM (6M2R)	342	Tyr115 (5.15 Å), Gln122 (4.94 Å), Thr339 (5.42 Å)
			135	n.a.
	Dopamine Transporter (outward)	HM (6DZZ)	342	Tyr115 (5.48 Å), Asn341 (6.64 Å), Thr512 (5.45 Å)
			135	n.a.
(7)	Enolase	2PSN	388	Lys357
				Glu141 (6.67 Å)
				Asn139 (4.50 Å), Gln360 (7.77 Å)
			398*	Arg399 (4.39 Å), Arg399'
				Lys192, Arg208 (7.18 Å)
				Tyr188 (3.30 Å)

#	Protein name	PDB code	Cysteine	Environment
(8)	Estrogen Receptor	1ERE	381*	Arg515', Lys520'
				Glu380 (7.72 Å), Glu385 (8.14 Å)
				His377 (4.56 Å), His513', His516', His547 (4.92 Å)
				Thr460 (4.67 Å)
			530*	Lys523 (4.42 Å)
				Tyr526 (3.93 Å)
(9)	Glyceraldehyde-3phosphate dehydrogenase	4WNC	**152**	Arg234
				Glu317 (9.16 Å)
				His179 (3.48 Å)
				Thr153 (3.65 Å), Tyr314 (4.37 Å), Asn316 (5.0 Å)
			156	Ser293 (3.57 Å), Ser312 (3.83 Å)
			247	n.a.
(10)	Hemoglobin	6KA9	93	Lys40'
				Asp94 (4.55 Å)
				His146 (4.29 Å) His97 (6.40 Å)
			104	Arg31
				His103 (5.47 Å)
				Ser35 (4.86 Å), Gln127 (4.76 Å)
(11)	Immunoglobulin G1 H Nie	6ARP	395*	Lys251
(12)	Immunoglobin kappa light chain	6N35	134*	Glu161
				Ser177 (6.29 Å)
(13)	Kinesin KIFC1	5WDH	663*	Ser607 (4.18 Å)
(14)	Kinesin KIF2C	4UBF	260*	Asp312 (4.29 Å)
				His257 (5.18 Å)
				Cys262 (5.61 Å), Cys344, Cys560 (3.56 Å)
			287*	Lys286 (6.77 Å)
				Glu244 (5.44 Å), Glu712 (5.11 Å)
				Ser285 (Ser285A), Cys287 (5.61 Å)
(15)	NEM-sensitive factor	HM (3J94)	264	Arg271, Arg446
				Glu329, Glu446
(16)	Sex Hormone-Binding Globulin	1KDM	164*	Arg47 (4.85 Å)
				Asp162 (7.69 Å)
				His17 (4.56 Å)
			188*	Arg47 (6.11 Å)
				His17 (3.64 Å)
(17)	Topoisomerase IIa (ATPase domain)	1ZXM	170	Ser174 (4.08 Å)
(19)	Vesicular proton ATPase	6M2R	254	Lys256, Arg400, Lys437 (6.37 Å), Lys438
				Asp436 (6.97 Å)

[a]For residues within 5 Å from the candidate Cys, distances with respect the Cys sulfur atom were calculated˚ using the following side chain atoms: carboxylate C for Asp/Glu, guanidinium C for Arg, amino N for Lys, closest imidazole N for His, hydroxyl O for Ser/Thr/Tyr, thiol S for Cys, indole N for Trp and closest amide N or O for Asn/Gln.

"*The ACR protein targets were further characterized based on their oligomerization state and protein class (see Table S2). The latter classification (see Figure S3) aims at connecting the corresponding protein target with the subcellular mechanism responsible for ACR toxicity [...]. In addition, Table S2 includes the optimal pH range for optimal function of the corresponding protein, since the Michael addition reaction is favored with increasing pH [116, 118]. Most proteins in our dataset have an optimal pH close the physiological value of 7.*"

Table S2. Additional information for the ACR protein targets listed in Table 4, including oligomerization state and protein class, as well as the pH range for optimal protein function (as indicated in the corresponding reference).

#	Protein name	Oligomerization state	Protein class	Optimal pH	Reference
(1)	Albumin	monomer	plasma protein	7.4	[234]
(2)	Alcohol Dehydrogenase	dimer	enzyme	7-8	[235]
(3)	Aldolase	tetramer	enzyme	7.2	[236]
(4)	Creatine Kinase	dimer	enzyme	6.5-7	[237]
(5)	Dopamine Receptor	monomer	membrane receptor	~7	[238]
(6)	Dopamine Transporter	monomer	membrane transporter	6.0-7.4	[239]
(7)	Enolase	dimer	enzyme	6.8–6.9	[240]
(8)	Estrogen Receptor	dimer	nuclear receptor	6-8	[241]
(9)	Glyceraldehyde3-phosphate dehydrogenase	tetramer	enzyme	7.2-8.3	[242]
(10)	Hemoglobin	tetramer	plasma protein	7.4	[243])
(11)	Immunoglobulin G1 H Nie	tetramer	plasma protein	~6.5	[244]
(12)	Immunoglobin kappa light chain	tetramer	plasma protein	~6.5	[244]
(13)	Kinesin KIFC1	monomer	ATPase*	6.8-7.2	[245]
(14)	Kinesin KIF2C	dimer	ATPase*	6.8-7.2	[245]
(15)	NEM-sensitive factor	hexamer	ATPase	9.0	[165]
(16)	Sex Hormone-Binding Globulin	dimer	plasma protein	>5	[246]
(17)	Topoisomerase IIa (ATPase domain)	dimer	enzyme	7.5	[247]
(18)	Topoisomerase IIa (Toprim domain)	dimer	enzyme	7.5	[248]
(19)	Vesicular proton ATPase	monomer	ATPase*	7	[249]

*For both kinesins and the vesicular proton ATPase, their ATPase domain is the one considered in the covalent docking calculations and thus was used for the protein class and oligomeric state classification.

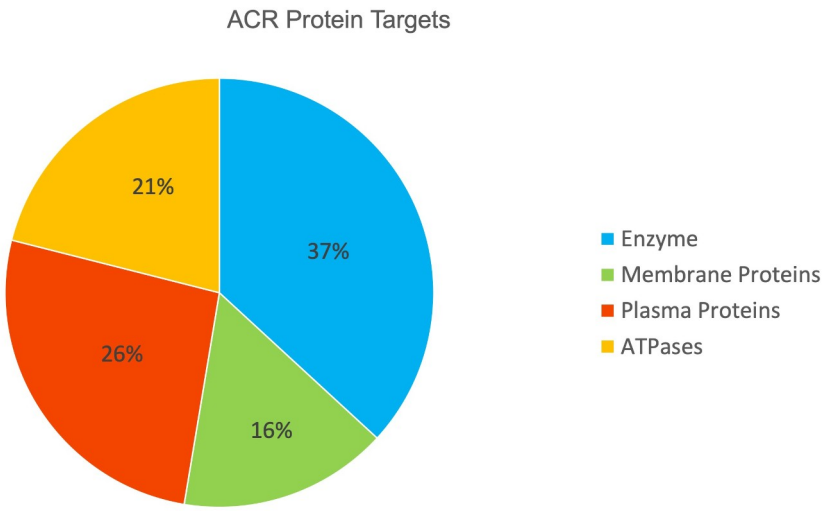


Figure S3. Functional classification of the acrylamide protein targets in the dataset compiled for this study.

7.2 Validation of docking approach

“Covalent docking to the reactive cysteine of each target protein (see Table S1) was performed using Haddock (version 2.2.) [87, 119]. We followed the standard covalent docking protocol of Haddock [120], which was initially developed for covalent inhibitors of cathepsin K [120]. Here we have validated it for Cys-ACR adducts using the available experimental structures for such covalent ligand-protein complexes. In particular, we searched in the Protein Data Bank [107, 108] for experimental structures containing the ligand name ROP (i.e. propionamide, the product of the Michael addition reaction, as explained in the main text, section 2.5); see [250]. We then filtered for entries in which this ligand is covalently linked to the protein, rendering a total of five X-ray structures (PDB codes 3ZVI [251], 4GYL [128], 4IZV, 4IZU, and 4WGF [252], of which one (PDB 4IZU) contains two covalent ACR adducts (with C53 and C145). All these structures correspond to bacterial enzymes and have resolution between 1.4 Å and 2.3 Å. For the redocking calculations, the protein structures were stripped from the ROP ligand and submitted to the same covalent docking protocol described in the main text (see section 4.1.2). The redocking and crystallographic poses are compared in Table S3, in terms of their protein-ligand interactions. Most of the protein-ligand interactions observed in the crystal structures are reproduced in the redocking poses. Additional interactions are present in the redocking poses, which we ascribed to the increased flexibility of acrylamide in the redocking calculations (which include a final refinement molecular dynamic step at 300 K; see step 3b in section 4.1.2 in the main text) compared to the X-ray structures (solved at 100K-110K). For further information on the redocking calculations, we refer the reader to Supplementary Material 2, which includes the Haddock score and size of the redocking clusters, as well as the Haddock score of the top four poses of each cluster.”

*Table S3. Comparison of the protein-ligand interaction fingerprints of the crystal structures containing Cys-ACR adducts and the corresponding redocking poses. For each experimental structure, the PDB code, reactive Cys and resolution (between parentheses) is given. The presence/absence of an interaction is indicated with a **x**/**o** sign, respectively; H-bonds were defined as explained in section 4.1.3.*

PDB Code 3ZVI – C361 (1.90 Å)

Residue	Interaction	X-ray	Redocking
Gln73	HB acceptor	**x**	**x**
Thr360	HB donor	**x**	**x**

PDB Code 4GYL – C166 (1.90 Å)

Residue	Interaction	X-ray	Redocking
Tyr60	HB acceptor	**x**	**x**
Lys134	HB donor	**x**	**x**
Asp167	HB acceptor	**o**	**x**
Tyr192	HB acceptor	**o**	**x**

PDB Code 4IZV – C53 (1.65 Å)

Residue	Interaction	X-ray	Redocking
Pro89	HB acceptor	**x**	**o**

PDB Code 4IZU – C53 (1.40 Å)

Residue	Interaction	X-ray	Redocking
Pro49	HB acceptor	**o**	**x**
Ser50	HB acceptor	**o**	**x**
Pro89	HB acceptor	**x**	**o**

PDB Code 4IZU – C145 (1.40 Å)

Residue	Interaction	X-ray	Redocking
Gln41	HB donor	**x**	**x**
Gln41	HB acceptor	**o**	**x**
Thr47	HB acceptor	**o**	**x**
Thr47	HB donor	**x**	**x**
Lys111	HB donor	**x**	**x**
Glu119	HB acceptor	**x**	**x**

PDB Code 4WGF – C118 (1.40 Å)

Residue	Interaction	X-ray	Redocking
Asp19	HB acceptor	**x**	**x**
Ser59	HB acceptor	**x**	**o**
Ser59	HB donor	**o**	**x**
Asn65	HB donor	**x**	**x**
Trp176	HB donor	**x**	**o**

7.3 Proteins with experimentally verified reactive cysteine

"*Out of the 19 proteins in our dataset, the specific Cys targeted by acrylamide is known for eight. Below we present the results of the covalent docking performed for each of these proteins. In some cases (creatine kinase, glyceraldehyde-3-phosphate dehydrogenase and hemoglobin) experimental evidence suggests more than one cysteine targeted by acrylamide, but with different reactivity; thus we performed a docking calculation for each of the cysteines within the same protein target separately. The outcome of these dockings is presented in part in the main text and in part here below. Namely, the main text includes an overview of the results and their discussion, as well as Table 3, which reports the Haddock score and cluster size of the top (best scored) cluster. Here below we show the protein-ligand interaction analysis of the docking poses of the top cluster; additional clusters are also considered if their Haddock scores fall within standard deviation of the top cluster. Such analysis was carried out with ProLIF [125], as explained in section 4.1.3 of the main text; if no interactions were detected, no scheme is shown. For further details, we refer the reader to Supplementary Material 2 of [1], which includes the full report of the docking results.*" [1]

7.3.1 Creatine Kinase (CK)

"*Because of the the biphasic time dependent inactivation of CK by ACR observed in enzymatic assays [138], we performed covalent docking for several cysteine residues (see Table 4). Covalent docking for the experimentally known primary site of ACR modification in CK, C283, resulted in one main cluster (number 1); the top pose is shown Figure 15B in the main text and the corresponding protein-ligand interaction fingerprints in Figure S9 below. One main cluster (number 1) was also obtained for the secondary ACR binding site (C141) predicted in this study (see Figure S5). The docking results for the alternative cysteine discussed in the main text (C146, see section 4.2.2.2) are shown in Figure S6 to Figure S8. For both C141 and C146, some docking poses showed a distance between ACR and the corresponding Cys too long to be compatible with a covalent bond. Such non-covalently bound ligand poses were excluded when calculating the average Haddock score and cluster size shown in Table 3 in the main text and the protein-ligand interaction fingerprints in in Figure S6 to Figure S8.*"

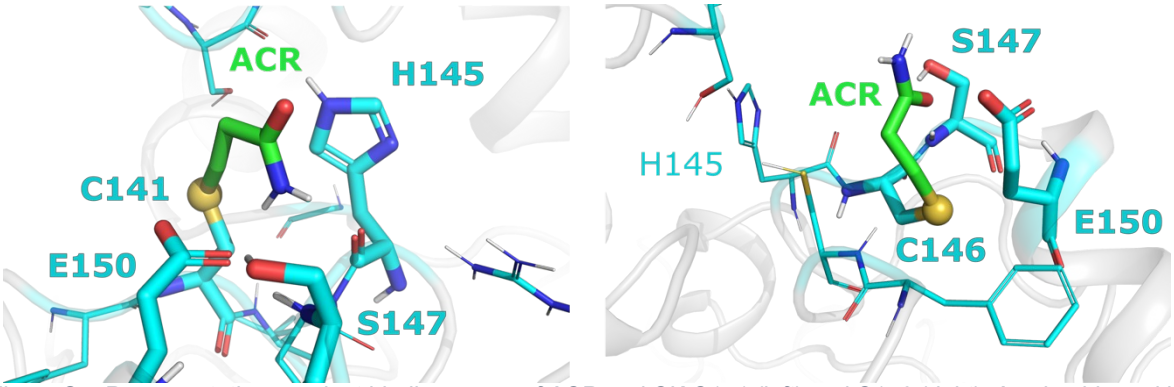
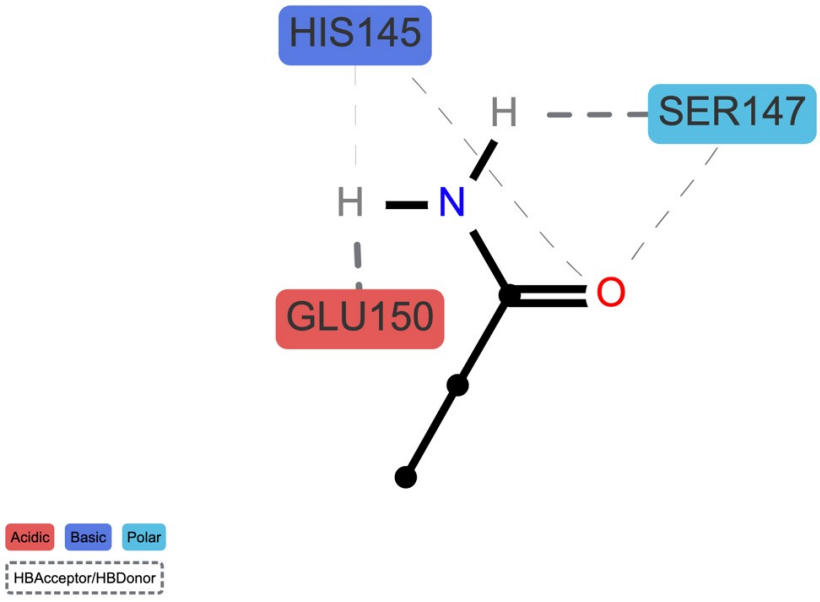
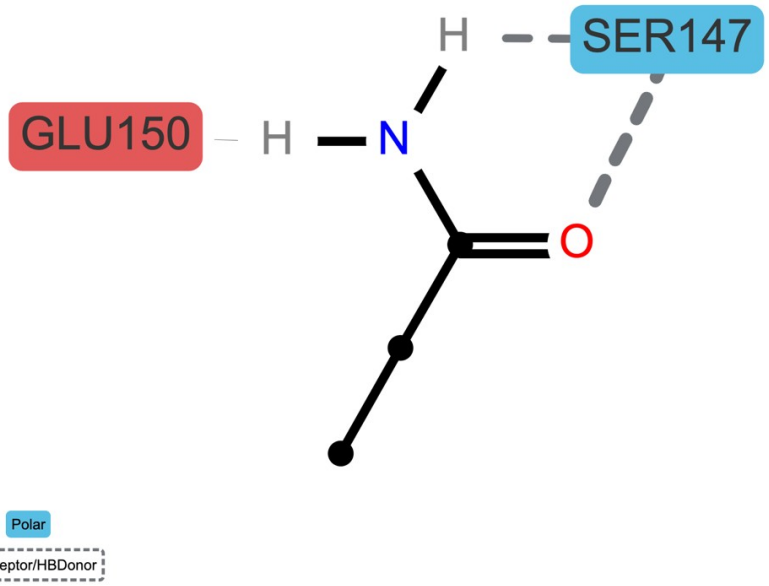


Figure S4. Representative covalent binding poses of ACR and CK C141 (left) and C146 (right). Acrylamide and its surrounding residues are represented as sticks, with carbon atoms colored in green and cyan, respectively. The sulfur atom between the reactive cysteine residue and the adduct is shown as a sphere. Residues forming hydrogen bonds (HBs) with ACR are displayed with thicker sticks and labeled. HBs present in more than 60% of the docking poses are shown with a dashed line. Residues within 5 A˚ of the adduct that can potentially favor the Michael addition reaction are shown with thinner lines.



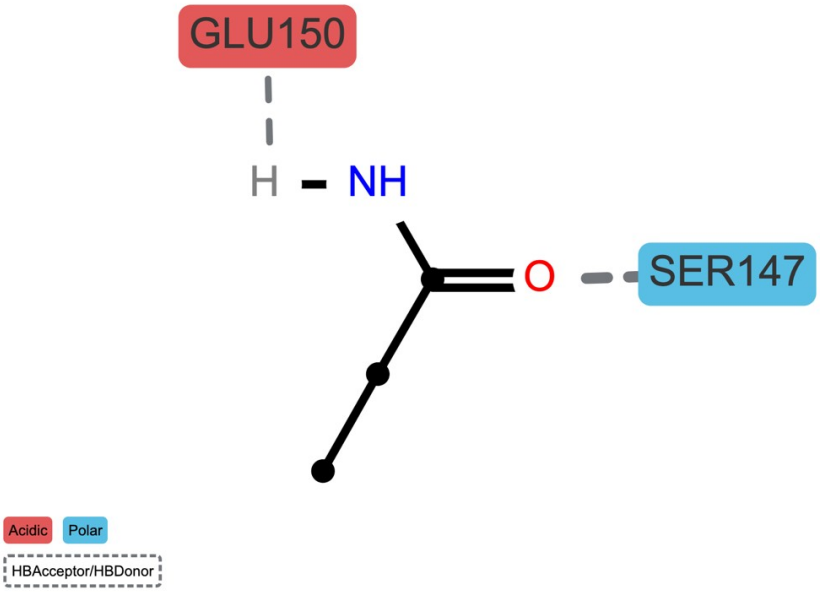
Creatine Kinase (C141) – Cluster 1			
residue	number	interaction	occurrence
His	145	HB donor	0.10
His	145	HB acceptor	0.03
Ser	147	HB donor	0.11
Ser	147	HB acceptor	0.51
Glu	150	HB acceptor	0.57

Figure S5. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



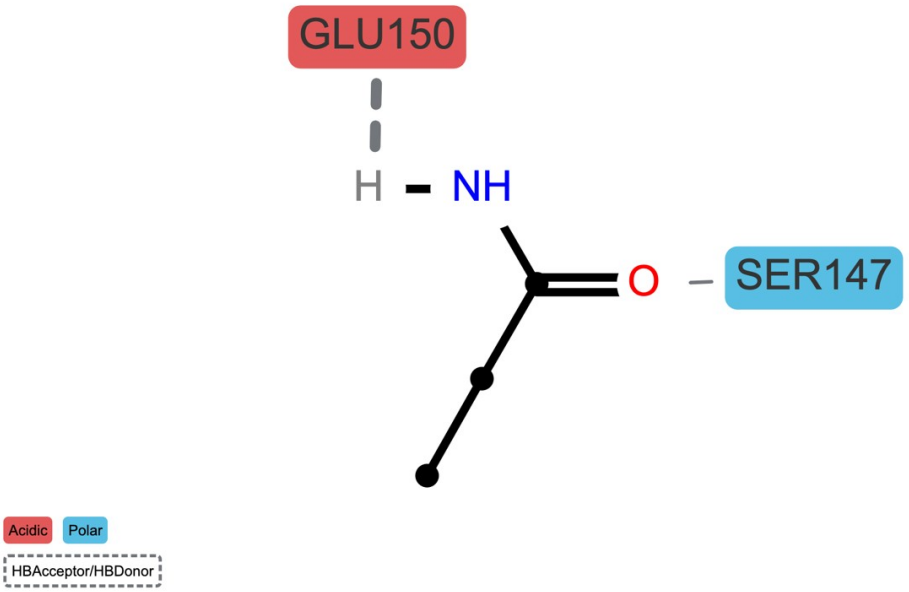
Creatine Kinase (C146) – Cluster 1			
residue	number	interaction	occurrence
Ser	147	HB donor	0.95
Ser	147	HB acceptor	0.59
Glu	150	HB acceptor	0.05

Figure S6. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



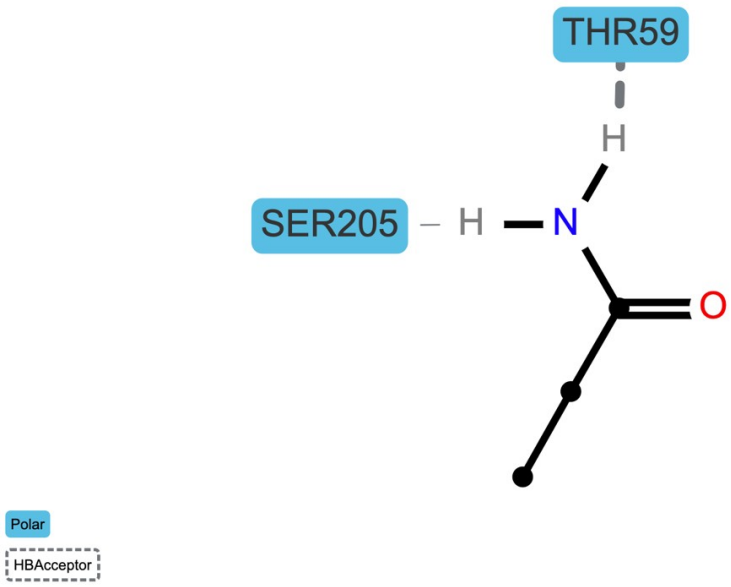
Creatine Kinase (C146) – Cluster 2			
residue	number	interaction	occurrence
Ser	147	HB donor	0.88
Glu	150	HB acceptor	0.53

Figure S7. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Creatine Kinase (C146) – Cluster 5			
residue	number	interaction	occurrence
Ser	147	HB donor	0.29
Glu	150	HB acceptor	0.86

Figure S8. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

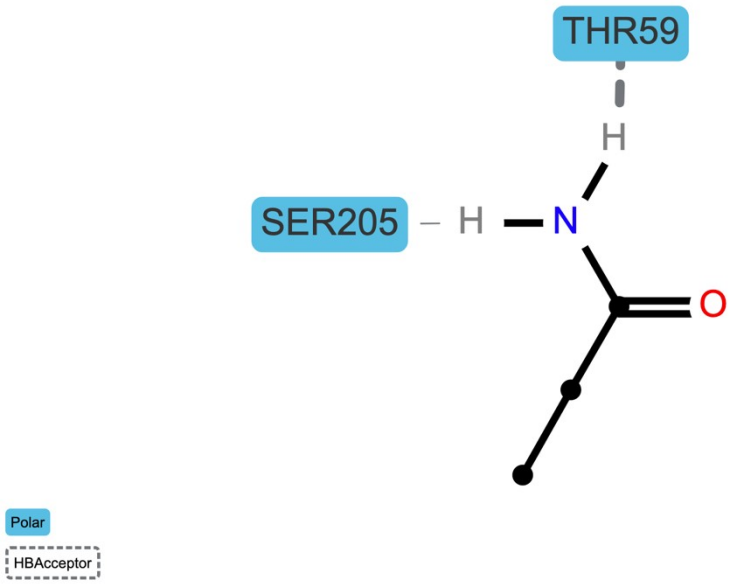


Creatine Kinase (C283) – Cluster 1			
residue	number	interaction	occurrence
Thr	59	HB acceptor	0.89
Ser	205	HB acceptor	0.20

Figure S9. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.3.2 Dopamine D3R Receptor (D3R)

Covalent docking for C114 of D3R resulted in one main top cluster (number 1).

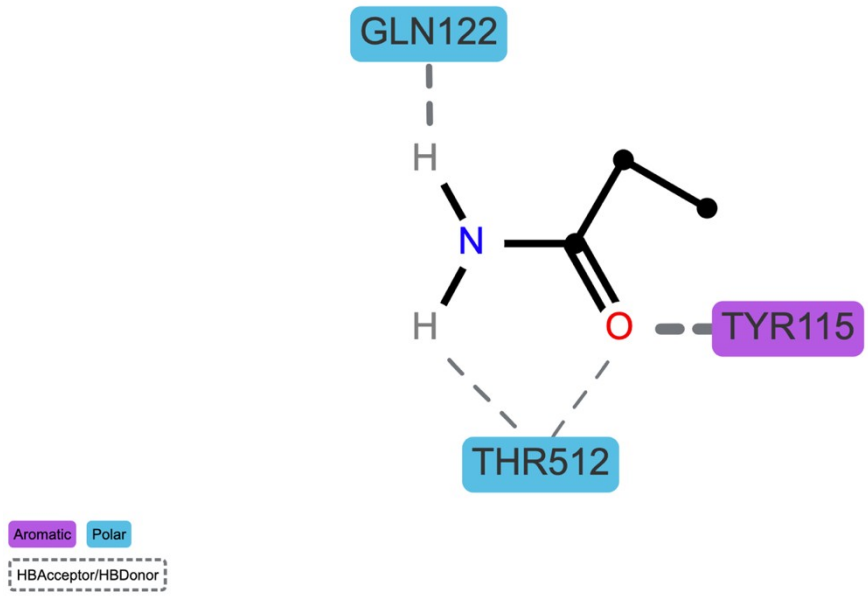


Dopamine D3 Receptor (C114) – Cluster 1			
residue	number	interaction	occurrence
Asp Ser	110 192	HB acceptor HB donor	0.82 0.01

Figure S10. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

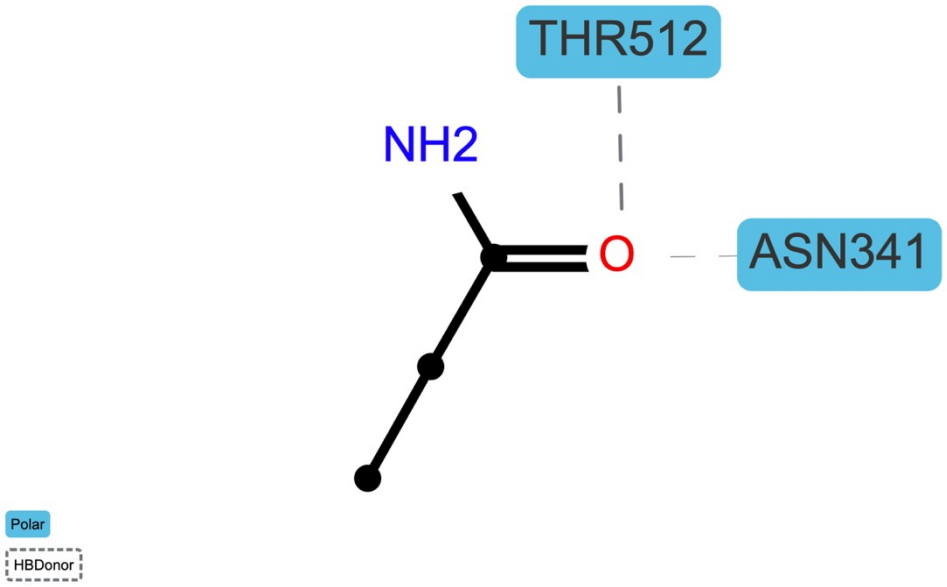
7.3.3 Dopamine Transporter (DAT)

"*Covalent docking for C342 of DAT (i.e. the primary site of ACR modification for the wild-type transporter) was performed for both the inward and outward conformations (IF and OF, respectively). One main top cluster (number 1) was obtained for the IF state, but two (clusters 1 and 4) for the OF state.*"



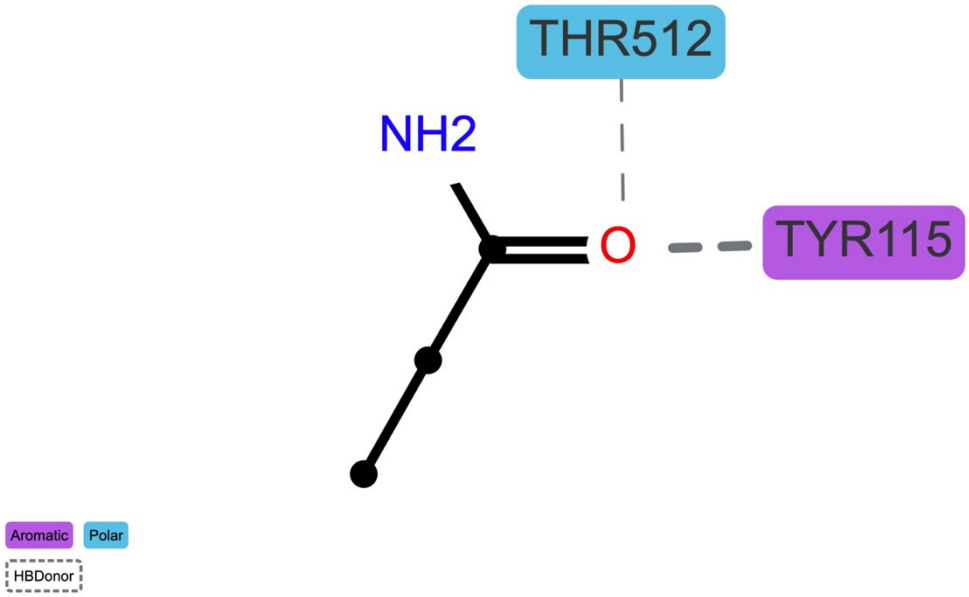
Dopamine Transporter (IF) (C342) – Cluster 1			
residue	number	interaction	occurrence
Tyr	115	HB donor	1.00
Gln	122	HB acceptor	0.59
Thr	512	HB donor	0.31
Thr	512	HB acceptor	0.38

Figure S11. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Dopamine Transporter (OF) (C342) – Cluster 1			
residue	number	interaction	occurrence
Asn	341	HB donor	0.06
Thr	512	HB donor	0.24

Figure S12. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Dopamine Transporter (C342) – Cluster 4			
residue	number	interaction	occurrence
Tyr	115	HB donor	0.67
Thr	512	HB donor	0.22

Figure S13. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.3.4 Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)

"For GAPDH, covalent docking was performed for three cysteines: C152, C156 and C247, which can be modified at increasing ACR concentrations [143]. Docking with C152 resulted in one top cluster (number 1); the top pose is shown in the main text (Figure 15) and the corresponding protein-ligand fingerprint in Figure S15. For C156 and C247, the generated docking poses were more diverse, resulting in either two clusters (numbers 1 and 2) or five clusters (numbers 1-5), respectively, with Haddock scores within standard deviation of the top (best scored) cluster. The top docking poses are shown in Figure S14, and the protein-ligand fingerprints of the corresponding dockings are presented in Figure S16 for C156 and Figure S17 for C247. As explained for CK (see section 4.1.2), docking poses with a distance between ACR and the corresponding Cys too long to be compatible with a C-S covalent bond were excluded from the analysis."

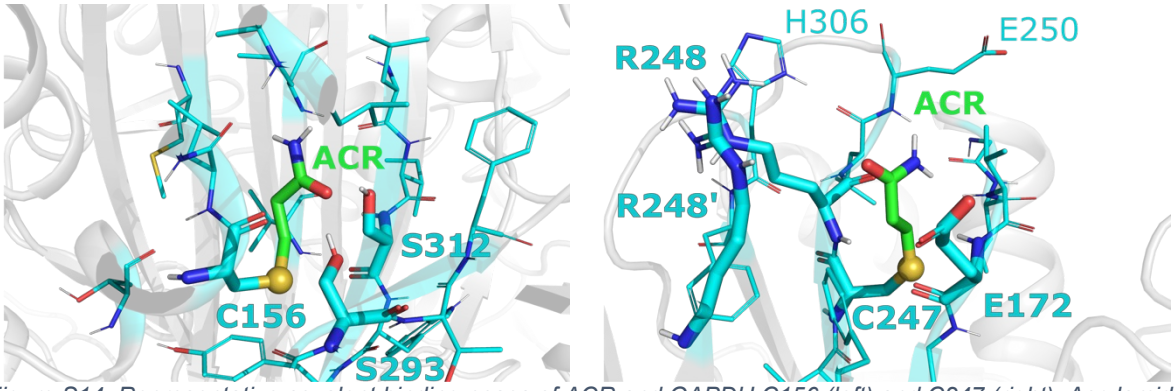
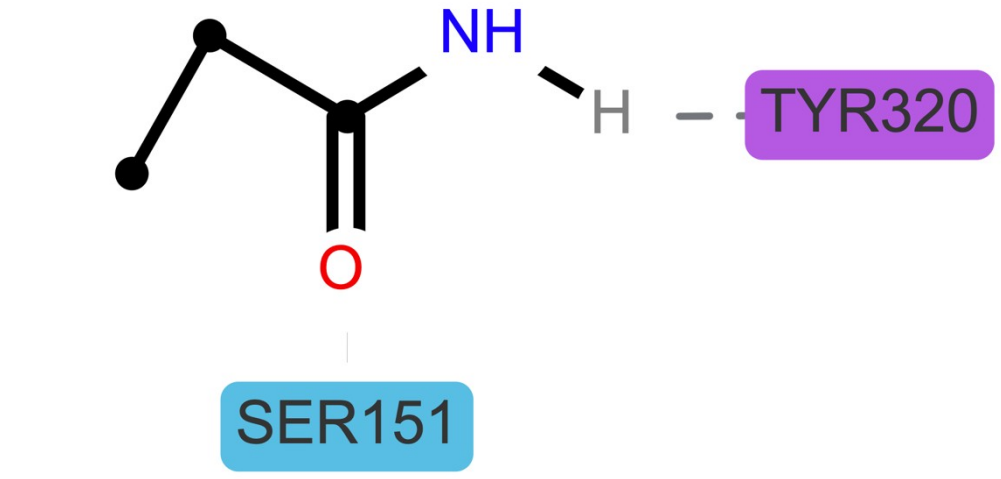
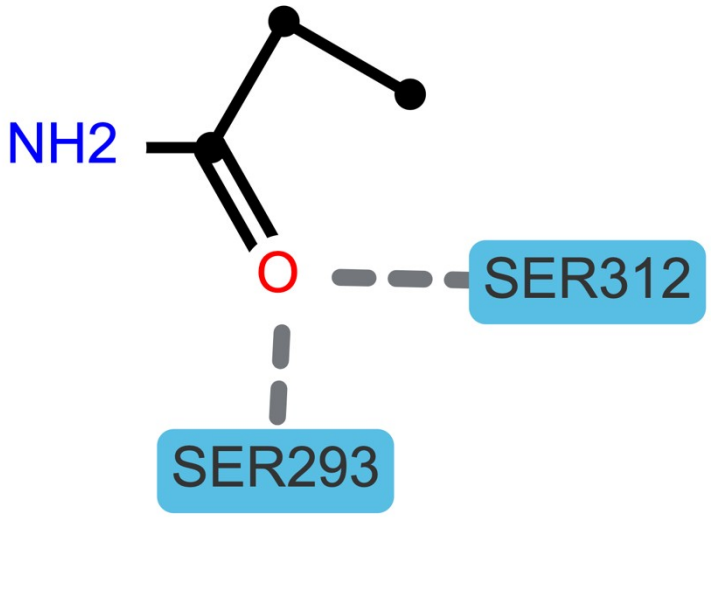


Figure S14. Representative covalent binding poses of ACR and GAPDH C156 (left) and C247 (right). Acrylamide and its surrounding residues are represented as sticks, with carbon atoms colored in green and cyan, respectively. The sulfur atom between the reactive cysteine residue and the adduct is shown as a sphere. Residues forming hydrogen bonds (HBs) with ACR are displayed with thicker sticks and labeled. HBs present in more than 60% of the docking poses are shown with a dashed line. Residues within 5 Å of the adduct that can potentially favor the Michael addition reaction are shown with thinner lines.



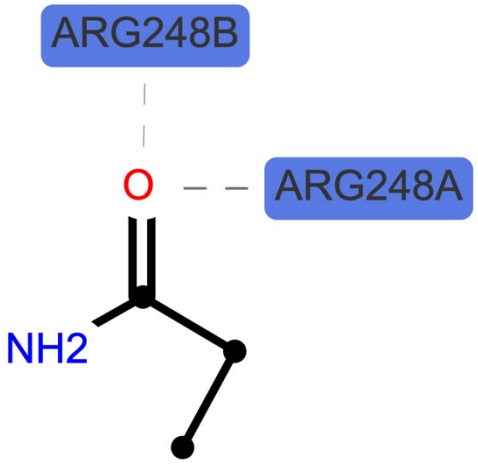
Glyceraldehyde-3-Phosphate Dehydrogenase (C152) - Cluster 1			
residue	number	interaction	occurrence
Ser Tyr	151 320	HB donor HB acceptor	0.01 0.48

Figure S15. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Glyceraldehyde-3-Phosphate Dehydrogenase (C156) - Cluster 1			
residue	number	interaction	occurrence
Ser Ser	293 312	HB donor HB donor	1.00 1.00

Figure S16. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

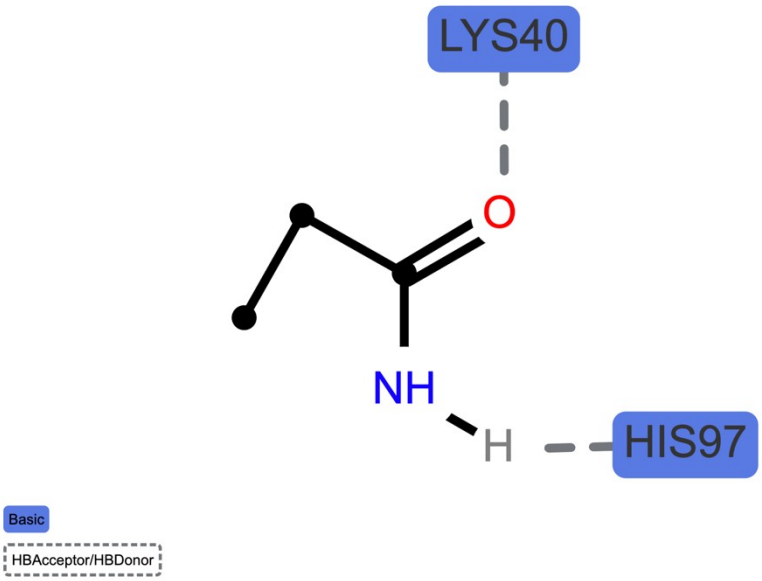


Glyceraldehyde-3-Phosphate Dehydrogenase (C247) - Cluster 1			
residue	number	interaction	occurrence
Arg Arg	248 (chain A) 248 (chain B)	HB donor HB acceptor	0.21 0.07

Figure S17. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

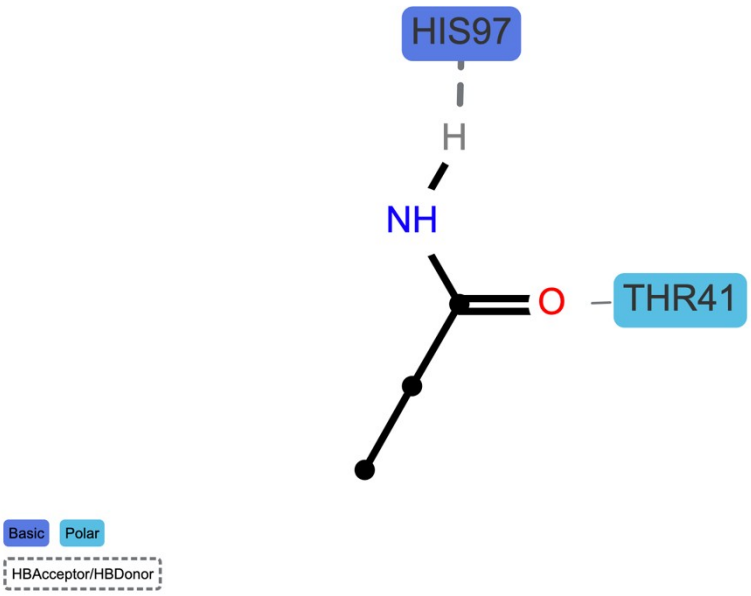
7.3.5 Hemoglobin (Hb)

"*Covalent docking for Hb was performed for two cysteines. For C104 in the α subunit, the resulting poses did not exhibit a properly formed C-S bond, suggesting that adduct formation is less favorable for this cysteine. Instead, for Hb C93 (β subunit), five main clusters (numbers 1-5) were obtained with Haddock score within standard deviation of top (best scored) cluster; their protein-ligand interaction profiles are reported below.*"



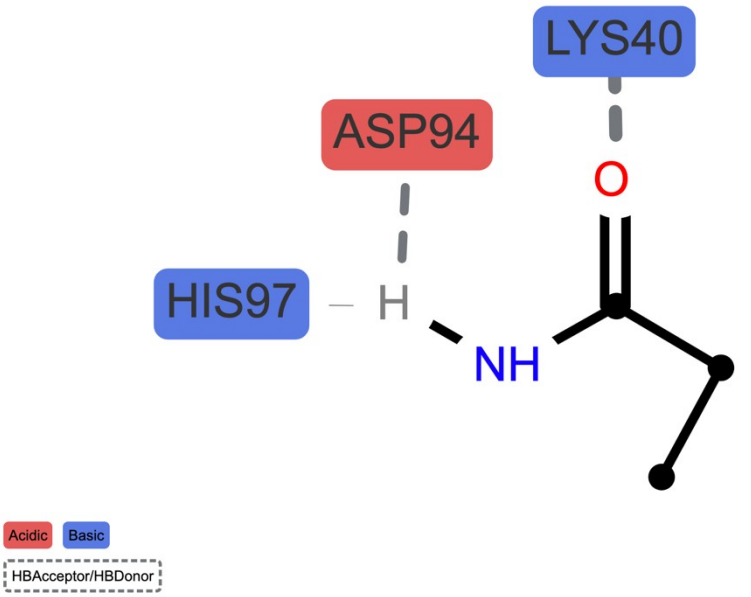
Hemoglobin (C93) - Cluster 1			
residue	number	interaction	occurrence
His Lys	97 40	HB acceptor HB donor	0.77 0.66

Figure S18. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



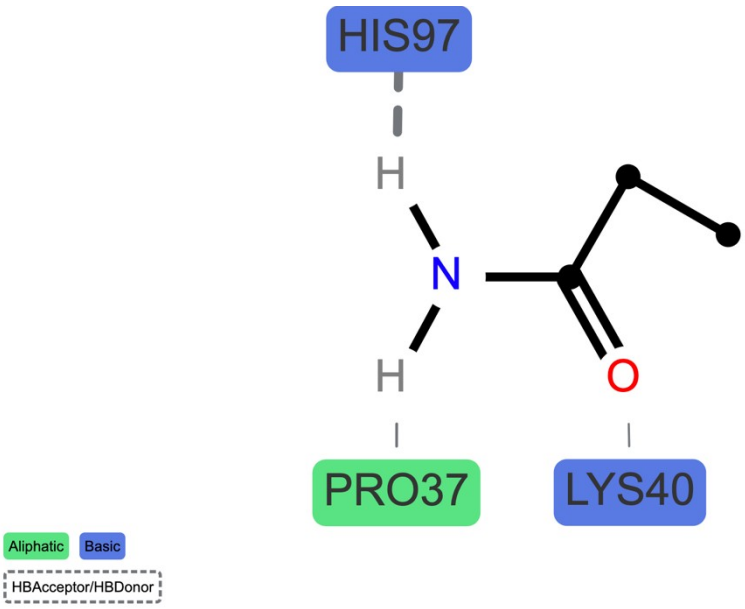
Hemoglobin (C93) - Cluster 2			
residue	number	interaction	occurrence
His	97	HB acceptor	0.90
Thr	41	HB donor	0.23

Figure S19. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



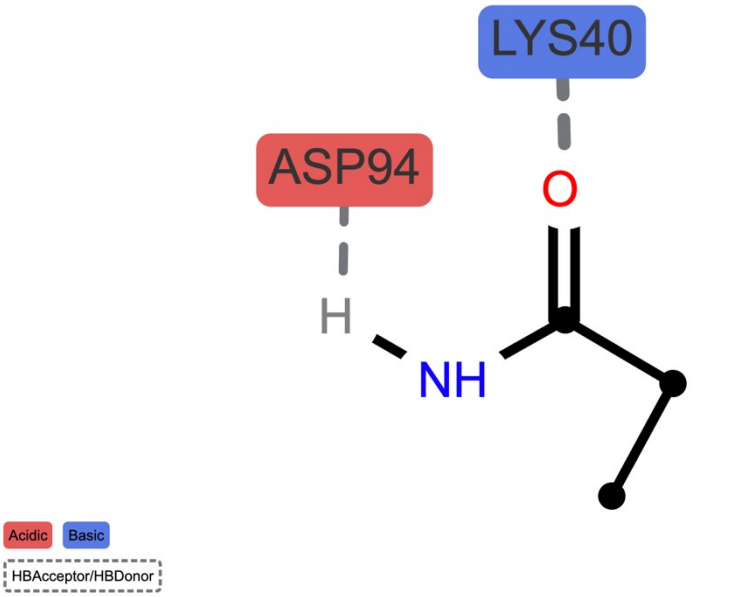
Hemoglobin (C93) - Cluster 3			
residue	number	interaction	occurrence
His	97	HB acceptor	0.07
Lys	40	HB donor	0.93
Asp	94	HB acceptor	0.89

Figure S20. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Hemoglobin (C93) - Cluster 4			
residue	number	interaction	occurrence
His	97	HB acceptor	0.68
Lys	40	HB donor	0.16
Pro	37	HB acceptor	0.21

Figure S21. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

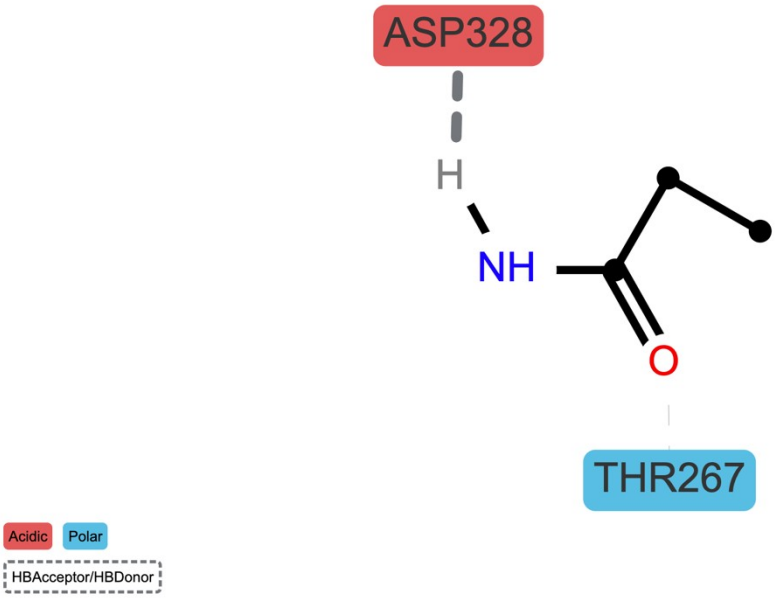


Hemoglobin (C93) - Cluster 5			
residue	number	interaction	occurrence
Asp	94	HB acceptor	0.75
Lys	40	HB donor	0.88

Figure S22. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.3.6 NEM-sensitive factor (NSF)

"*Docking to NSF C264 rendered two main clusters (numbers 1 and 2) with Haddock scores within standard deviation of each other. Analysis of the protein-ligand interactions for cluster 1 did not identify any significant contact, whereas the interactions for cluster 2 are shown below.*"

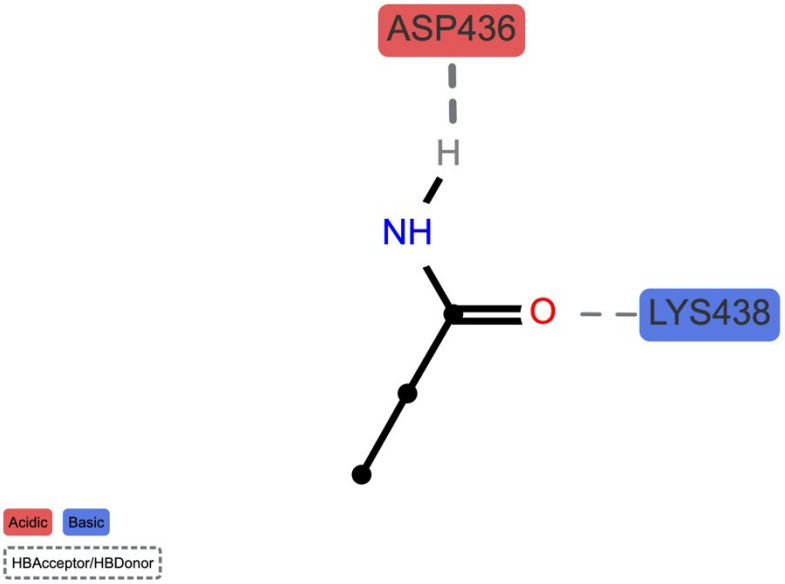


NEM-sensitive factor (C342) – Cluster 1			
residue	number	interaction	occurrence
Thr Asp	267 328	HB donor HB acceptor	0.03 0.88

Figure S23. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.3.7 Vesicular proton ATPase (v-ATPase)

"*Covalent docking for C254 of v-ATPase yielded one main top cluster (number 1); the analysis of the protein-ligand interactions present in the docking poses belonging to that cluster is shown below. As mentioned in the main text, the (only slightly) favorable docking score of this top cluster (−0.1 a.u.) can be attributed to C254 being located in a loop segment of the Walker A motif (GAFGCGKT). This motif is involved in coordinating ATP binding and hydrolysis and hence exhibits large rearrangements during the v-ATPase conformational cycle (see Figure S25). Such structural changes cannot be sampled with covalent docking protocols, thus resulting in lower accuracy of the predicted docking poses.*"



v-ATPase (C254) – Cluster 1			
residue	number	interaction	occurrence
Asp Lys	436 438	HB acceptor HB donor	0.72 0.36

Figure S24. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

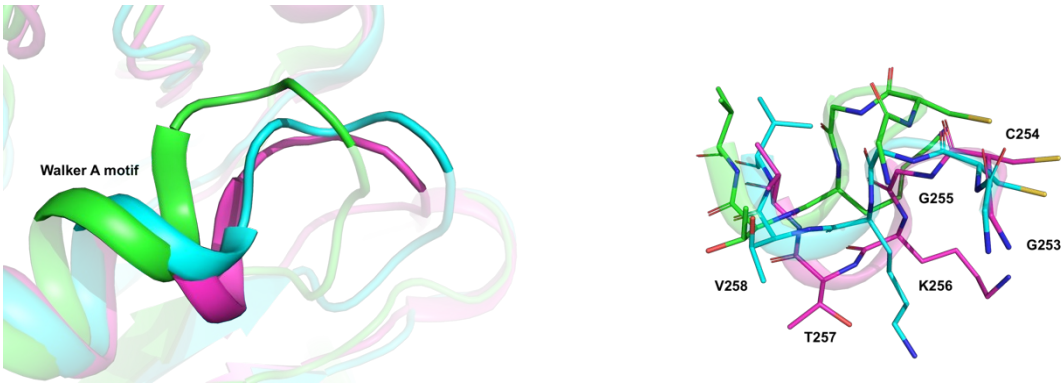


Figure S25. Flexibility of the v-ATPase Walker A motif. Structural superposition of crystallographic chains A (green), B (cyan) and C (magenta) of PDB structure 6WM2; the left and right panels display the secondary structures and the side chain positions, respectively, of the last five residues of the Walker A motif (G253-T257), including the reactive C254.

7.4 Proteins without reactive cysteine experimental information: Selection of candidate residues

"*Out of the 19 proteins in our dataset (Table 4), the specific Cys targeted by acrylamide is not known for eleven (see below; proteins are listed in alphabetical order). In order to narrow down the most promising candidates to be the reactive Cys, we first checked physicochemical properties and microenvironment, as well as conservation and post-translational modifications, for all the cysteines of the respective protein target [..]. Below we present the outcome of these analyses, as well as the main results of the covalent docking calculations performed for each of the selected candidate Cys residues. For further details, we refer the reader to Supplementary Material 2, which includes the full report of the docking results. Here we show the protein-ligand interaction analysis of the docking poses of the top (best scored) cluster, as well as for additional clusters when their Haddock scores fall within standard deviation of the top cluster. If no interactions were detected by ProLIF [125], no scheme is shown.*"

7.4.1 Alcohol Dehydrogenase (ADH)

"*Alcohol Dehydrogenase is an enzyme part of the ethanol metabolism that converts alcohols to aldehydes. ADH possesses multiple isoforms in humans which can be grouped into different sub-classes. The crystal structure with the highest resolution belongs to ADH1C, also known as ADH3 (see Table 4). Hence, the subsequent analyses were performed with this isoform.*
An MSA of the seven human ADH sequences (Figure S26) revealed eleven conserved Cys. However, six of them are part of the two Zn^{2+} binding sites present in ADHs [253], and thus were discarded as potential ACR binding sites. Out of the remaining five conserved cysteines, we selected C170 and C240 (ADH1C numbering) for our covalent docking tests. C170 is located near the enzyme active site and thus its modification is more likely to have an impact on the protein function. In addition, C240 in ADH appears to be the residue equivalent to C247 in GAPDH; the latter has been shown experimentally to be the target of ACR, though at high concentrations.
The results of the covalent docking for each of these two cysteines, using the aforementioned ADH structure, are reported below and in Supplementary Material 2. Based on the more favorable docking score for C240 (−10.8 a.u.) compared to C170 (−4.2 a.u.), we suggest that C240 might be the primary site of ACR modification in ADH."

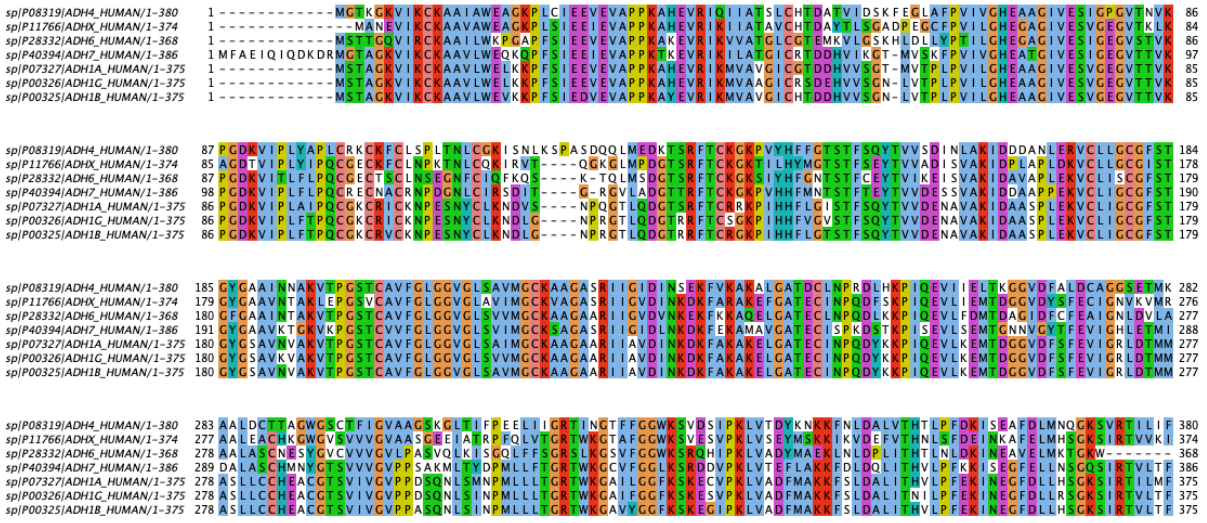
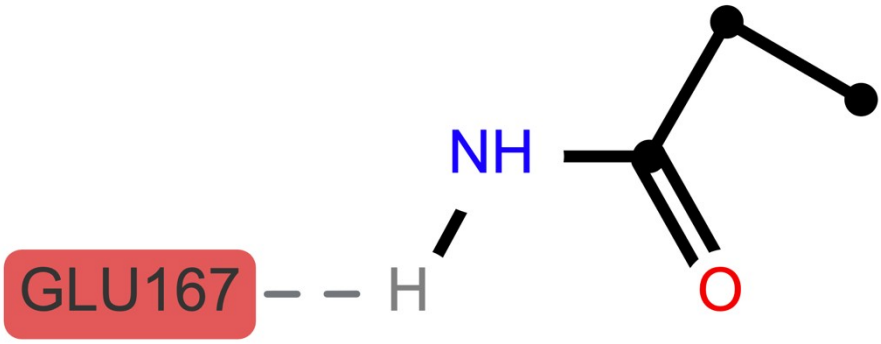
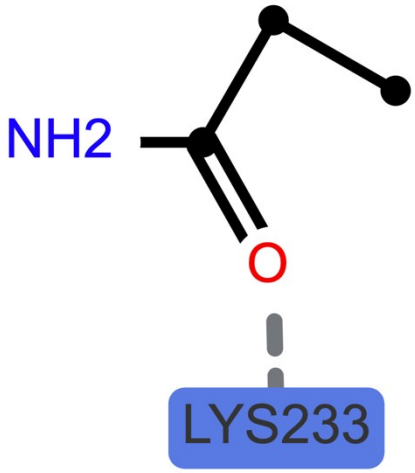


Figure S26. Multiple sequence alignment of human alcohol dehydrogenase isoforms, performed with the MAFFT webserver [254]. The Clustal color code was used, with hydrophobic residues in blue, positively charged in red, negatively charged in magenta, polar in green and aromatic in cyan. Special residues are shown in orange (glycine), yellow (proline) and pink (cysteine) and non-conserved residues are in white.



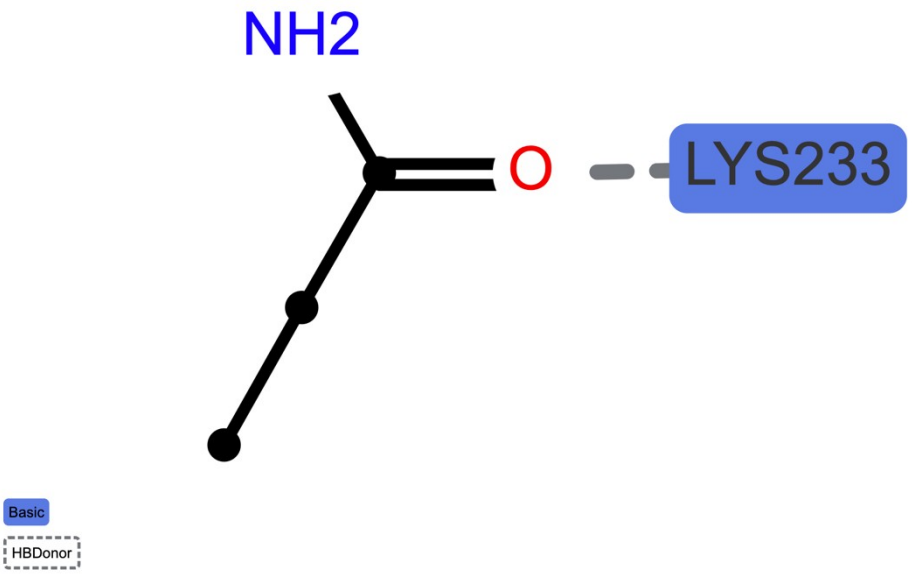
Alcohol Dehydrogenase (C170) – Cluster 1			
residue	number	interaction	occurrence
Glu	167	HB acceptor	0.44

Figure S27. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Alcohol Dehydrogenase (C240) – Cluster 1			
residue	number	interaction	occurrence
Lys	233	HB donor	1.00

Figure S28. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Alcohol Dehydrogenase (C240) – Cluster 2			
residue	number	interaction	occurrence
Lys	233	HB donor	0.86

Figure S29. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.4.2 Aldolase

"*Aldolase (or fructose-bisphosphate aldolase) is a key enzyme within the glycolysis pathway. It is responsible for splitting fructose 1,6-bisphosphate into dihydroxyacetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GADP). Aldolase has three isoforms: aldolase A (expressed in muscle), aldolase B (liver) and aldolase C (brain). In vitro experiments on rabbit muscle aldolase showed a multiphasic inactivation of aldolase by acrylamide [134]; however, there is no mutagenesis data pinpointing the Cys residues targeted by ACR. Hence, we first checked cysteine conservation by generating a multiple sequence alignment of human and rabbit aldolase isoforms (Figure S30). We found that C134, C177, C239 and C289 are conserved and thus can be potential candidates for the reactive Cys targeted by ACR and/or have functional relevance. Next, we used the crystal structure of human liver aldolase (i.e. the highest resolution crystal structure) to perform ACR covalent dockings for each of these cysteines. As shown in Figure S30, this isoform possesses an additional cysteine residue at position 268, which is solvent exposed and thus was also submitted to our docking workflow. As mentioned above, experimental studies showed that aldolase inactivation is dependent on both ACR concentration and time of incubation [134]. This rather complex pattern of inactivation indicates that acrylamide could form multiple covalent adducts at several Cys sites before aldolase activity is completely abolished. Therefore, in this case it is particularly relevant to consider all possible Cys candidates for covalent docking.*"

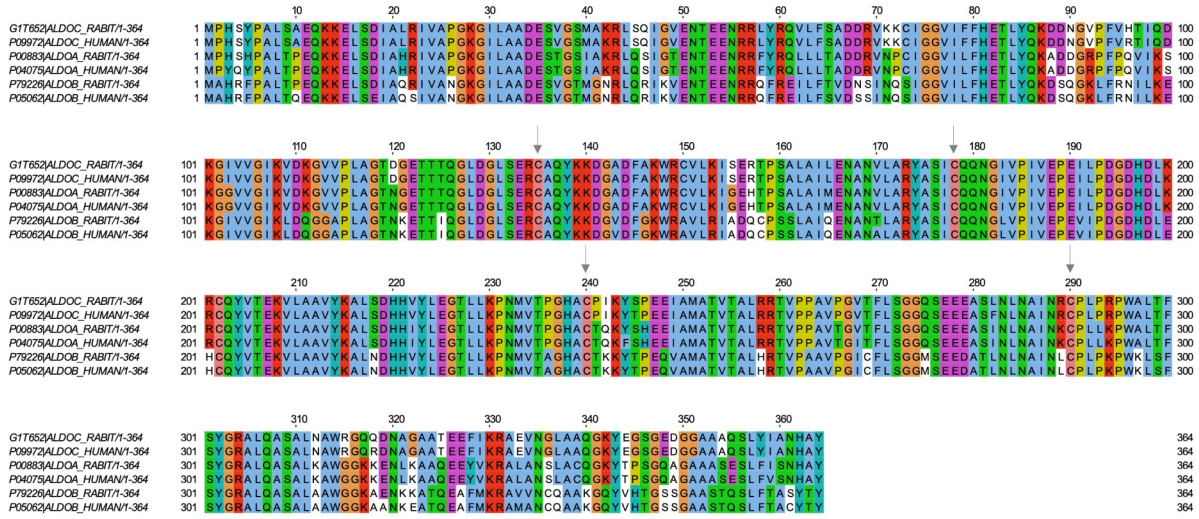
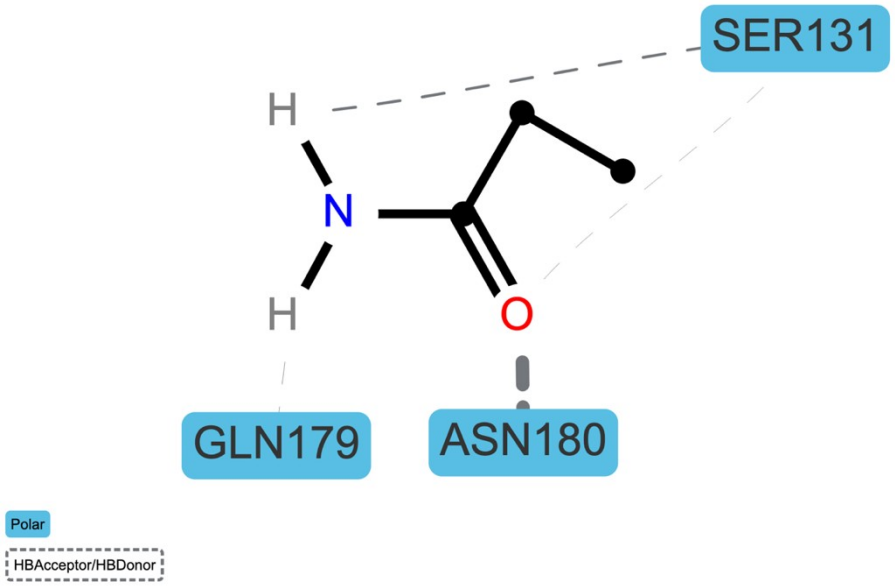


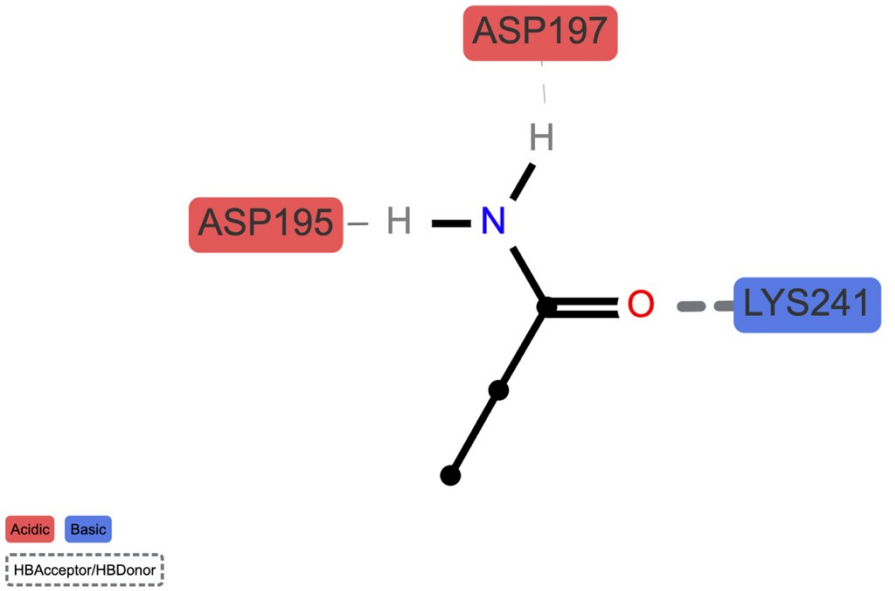
Figure S30. Multiple sequence alignment of human and rabbit aldolase isoforms, performed with the MAFFT webserver [254]. Aldolase isoforms A (muscle), B (liver) and C (brain) were included. The same color code as in Figure S26 was used.

"*Our computational results showed that C134, C239, C268 and C289 exhibit similar favorable docking scores for their top clusters, i.e. −26.1 a.u., −23.1 a.u., −29.4 a.u. and −24.1 a.u., respectively (see Supplementary Material 2). Below we report the protein-ligand interactions for each of the Cys-ACR covalent adducts. In contrast, the less favorable docking score for C177 (−9.7 a.u.), as well as the lack of a properly formed S-C bond between C177 and ACR in our models, suggest that modification of this particular residue is less likely.*"



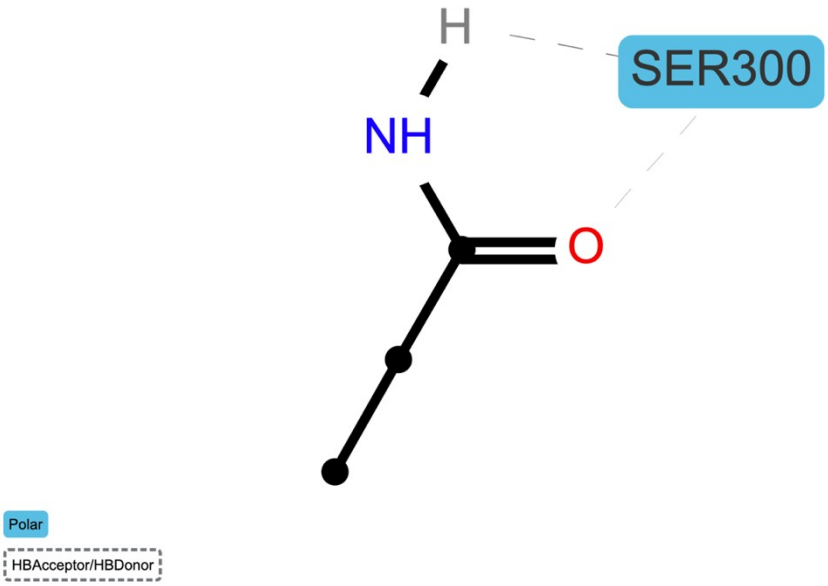
Aldolase (C134) – Cluster 1			
residue	number	interaction	occurrence
Ser	131	HB donor	0.03
Ser	131	HB acceptor	0.21
Gln	179	HB acceptor	0.05
Asn	180	HB donor	1.00

Figure S31. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



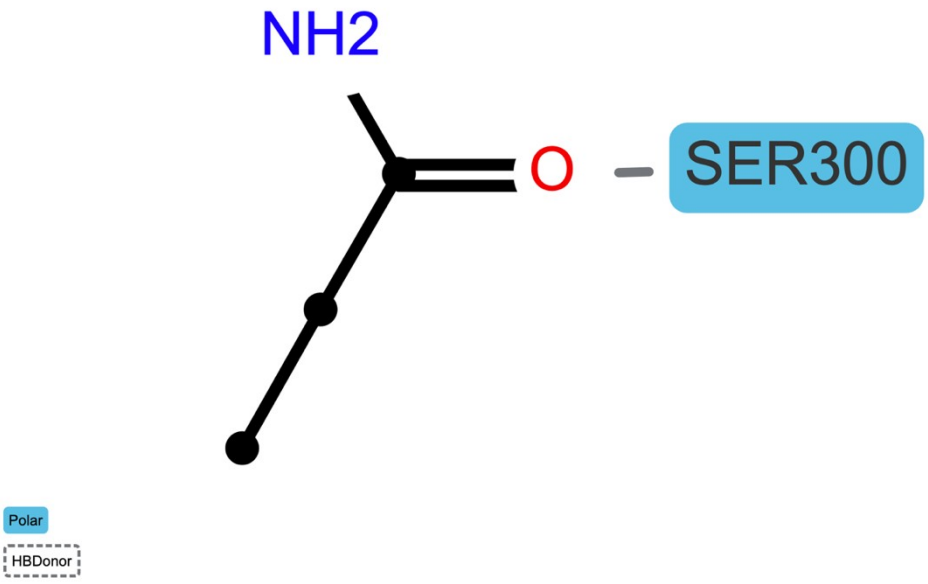
Aldolase (C239) – Cluster 1			
residue	number	interaction	occurrence
Asp	195	HB acceptor	0.24
Asp	197	HB acceptor	0.06
Lys	241	HB donor	0.96

Figure S32. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Aldolase (C268) – Cluster 1			
residue	number	interaction	occurrence
Ser	300	HB donor	0.03
Ser	300	HB acceptor	0.09

Figure S33. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Aldolase (C268) – Cluster 2			
residue	number	interaction	occurrence
Ser	300	HB donor	0.52

Figure S34. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.4.3 Enolase

"*Like aldolase and GAPDH, enolase is an enzyme part of the glycolysis pathway and is responsible for the conversion of 2-phosphoglycerate to phosphoenolpyruvate. Different isoforms of enolase exist, assembled as either homo- or heterodimers. In order to pinpoint possible candidates for subsequent covalent docking, we first identified solvent exposed Cys residues. Out of the six conserved cysteines present in enolase (C118, C336, C356, C338, C356, C388 and C398, see Figure S35), only the last two are significantly solvent exposed and thus accessible to acrylamide. Hence, C388 and C398 were considered for the subsequent covalent docking calculations. Based on the docking scores obtained (−36.3 and −39.0 a.u.) for C388 and C398, respectively), we suggest that both cysteines can potentially be targeted by acrylamide. Modification of C388 by ACR is supported by experimental studies showing that this cysteine undergoes modification under electrophile-induced oxidative stress, resulting in enzyme inactivation [255]. In addition, ACR modification of C398 could further contribute to the experimentally observed protein inhibition [140] by altering the quaternary structure of the enzyme, as C398 is located at the interface between two enolase subunits. In this regard, this example highlights the importance to map the location of the candidate reactive Cys on the protein 3D structure in order to establish a connection between ACR modification and its impact on protein function.*"

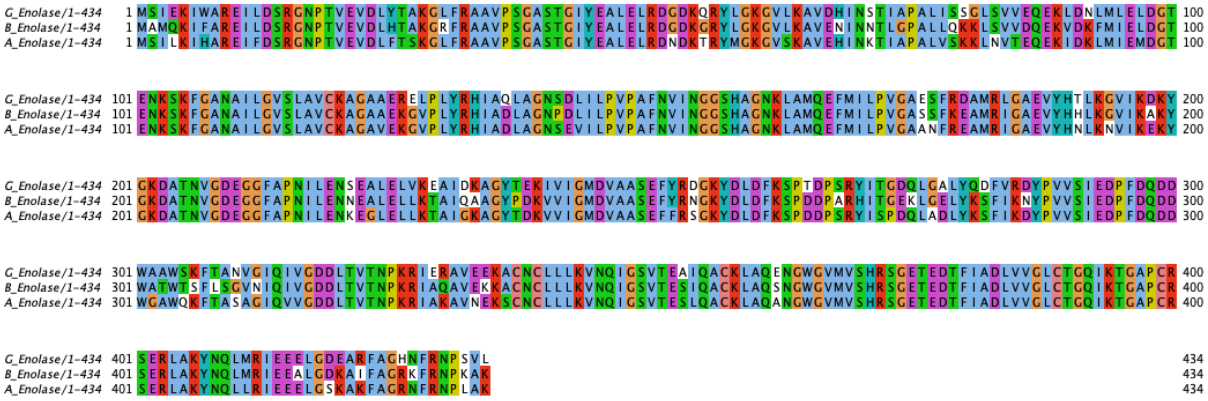
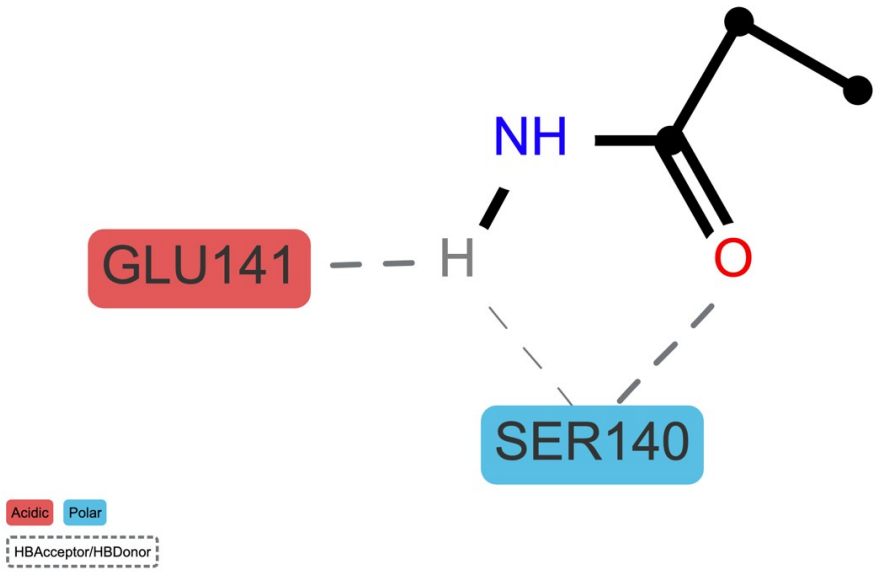
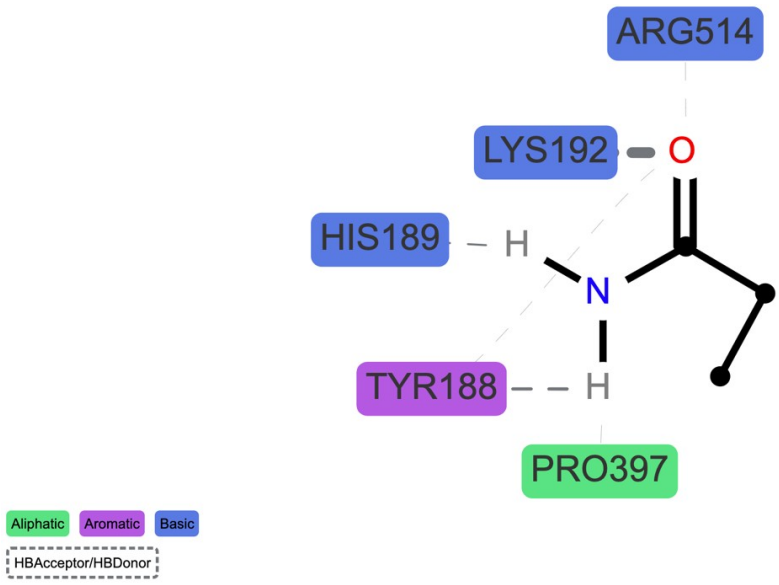


Figure S35. Multiple sequence alignment of human enolase isoforms (α, β and γ) generated using the MAFFT webserver [254]. The same color code as in Figure S26 was used.



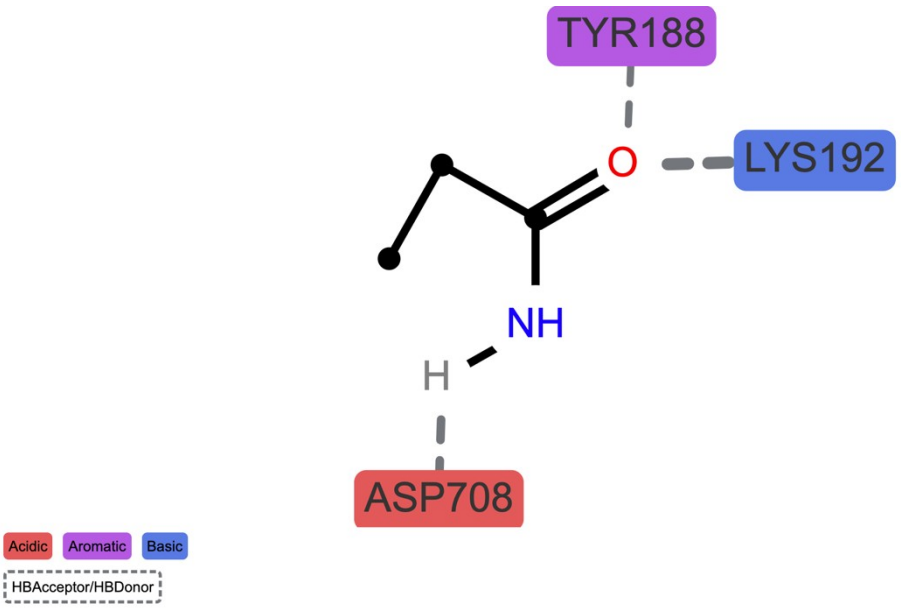
Enolase (C388) – Cluster 1			
residue	number	interaction	occurrence
Ser	140	HB donor	0.36
Ser	140	HB acceptor	0.13
Glu	141	HB acceptor	0.39

Figure S36. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



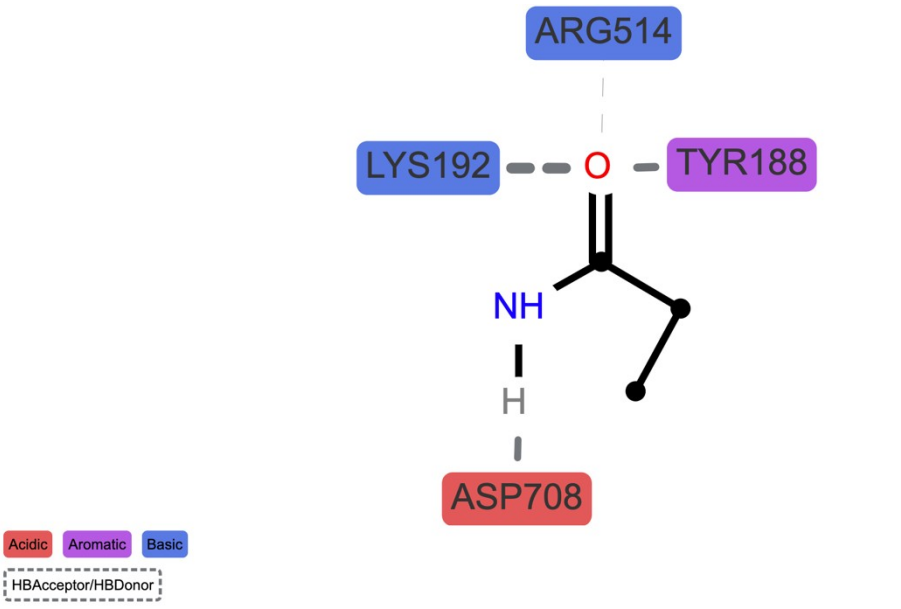
Enolase (C398) – Cluster 1			
residue	number	interaction	occurrence
Tyr	188	HB donor	0.01
Tyr	188	HB acceptor	0.31
His	189	HB acceptor	0.19
Lys	192	HB donor	1.00
Pro	397	HB acceptor	0.01
Arg	514	HB donor	0.03

Figure S37. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Enolase (C398) – Cluster 2			
residue	number	interaction	occurrence
Tyr	188	HB donor	0.60
Lys	192	HB donor	1.00
Asp	708	HB acceptor	0.79

Figure S38. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

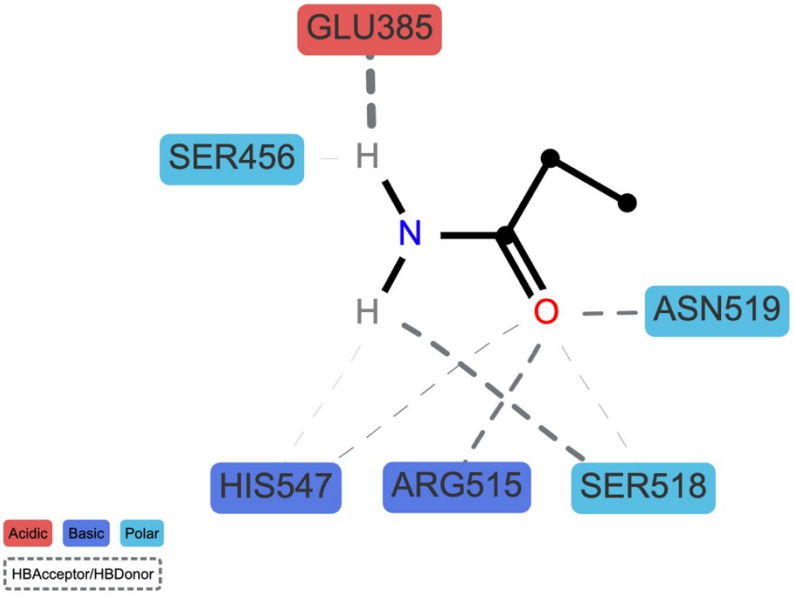


Enolase (C398) – Cluster 3			
residue	number	interaction	occurrence
Tyr	188	HB donor	0.79
Lys	192	HB donor	1.00
Arg	514	HB donor	0.05
Asp	708	HB acceptor	0.70

Figure S39. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

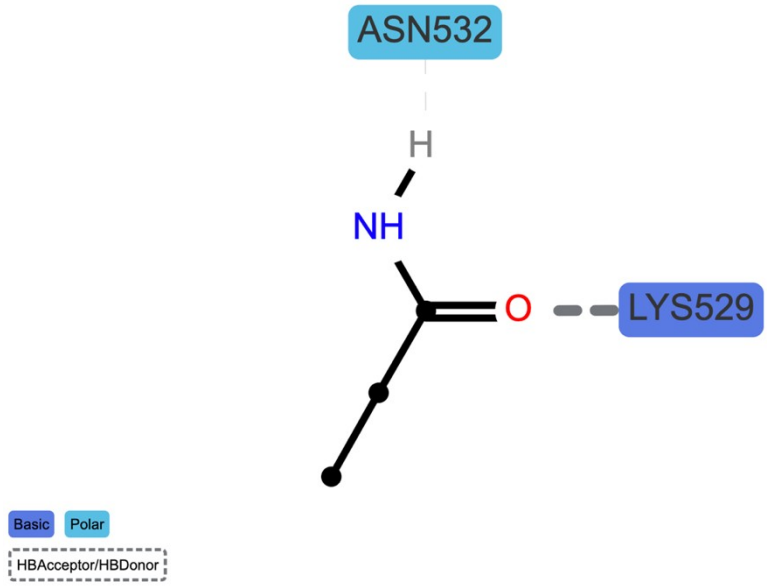
7.4.4 Estrogen Receptor

"*Estrogen receptors are nuclear receptors initially identified as binding female steroids of the estrogen group and regulating pathways related to sexual maturity. Alterations of this tightly regulated hormonal balance can lead to pathological side effects, such as tumor formation. Estrogen receptors have been proposed as potential targets of acrylamide based on the experimental observation that acrylamide induces misregulation of the hormone balance [141]. Moreover, human estrogen receptors contain conserved cysteine-rich binding domains that can act as reactive sites for ACR. In this regard, mutagenesis data has shown C381 and C530 to be involved in covalent hormone binding [142] and thus these two cysteines could potentially form covalent adducts also with acrylamide. Both residues showed favorable docking scores for their top cluster, −36.0 and −22.6 a.u. for C381 and C530, respectively. Considering the more favorable docking score for C381, the presence of more protein-ligand HBs and a slightly lower pK_a value of 9.9 (compared to >12 of C530), it is reasonable to assume that C381 is the primary site of ACR modification. However, we cannot discard that C530 could still be modified by acrylamide at higher concentrations.*"



Estrogen Receptor (C381) – Cluster 1			
residue	number	interaction	occurrence
Glu	385	HB acceptor	0.94
Ser	456	HB acceptor	0.02
Arg	515	HB donor	0.06
Ser	518	HB donor	0.44
Ser	518	HB acceptor	0.54
Asn	519	HB donor	0.39
His	547	HB donor	0.11
His	547	HB acceptor	0.02

Figure S40. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

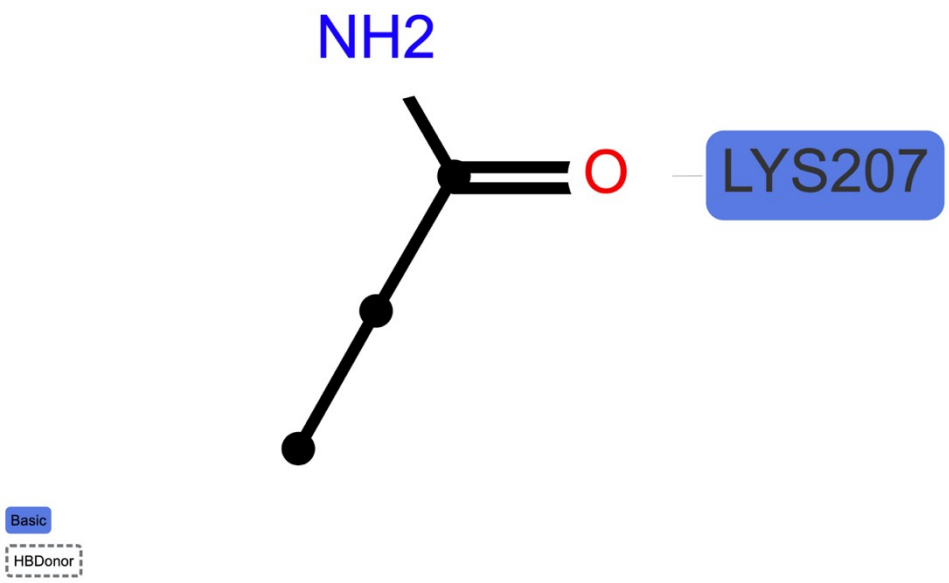


Estrogen Receptor (C530) – Cluster 1			
residue	number	interaction	occurrence
Lys Asn	529 532	HB donor HB acceptor	0.99 0.02

Figure S41. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

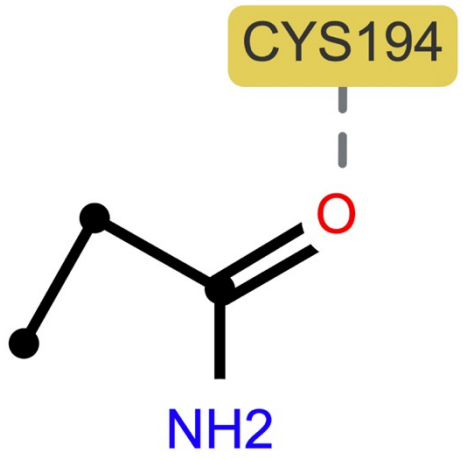
7.4.5 Immunoglobulins (Igs) G1 H Nie and kappa light chain

"*Acrylamide modification of two immunoglobulins was identified through nano liquid chromatography combined with tandem mass spectrometry [145]. Mapping of the modified peptides against the corresponding full protein sequences pinpointed C395 (immunoglobulin G1 H Nie) and C134 (immunoglobulin kappa light chain). Inspection of the respective protein structures (see Table S1) showed that both residues are involved in disulfide bridges. However, free thiols and chemical modification of disulfide bridges have been detected in other Igs [256]. Therefore, we performed covalent dockings for each of these cysteines, assuming that either they are reversibly reduced under certain conditions or acrylamide is able to break and modify the corresponding disulfide bridge. As shown in Supplementary Material 2, docking to C395 of immunoglobulin G1 H Nie showed one main cluster (number 1), whereas docking to C134 of the immunoglobulin kappa light chain rendered 4 clusters with Haddock scores within standard deviation of top (best scored) cluster. Below we report the protein-ligand interactions of those clusters for which HBs were detected. We speculate that ACR modification of the aforementioned cysteines will affect protein structure and stability of these two Igs, since it removes disulfide bridges.*"



Immunoglobin Kappa Light Chain (C134) - Cluster 1			
residue	number	interaction	occurrence
Lys	207	HB donor	0.03

Figure S42. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Sulfur

HBDonor

Immunoglobin Kappa Light Chain (C134) - Cluster 4			
residue	number	interaction	occurrence
Cys	194	HB donor	0.54

Figure S43. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.4.6 Kinesins KIFC1 and KIF2C

"*Kinesins are motor proteins essential to transport cellular cargos along microtubules, powered by ATP hydrolysis. ACR-mediated inhibition of kinesins has been proposed as the molecular mechanism by which acrylamide impairs fast anterograde and retrograde axonal transport [14]. Indeed, experimental studies have confirmed that acrylamide is able to inhibit two kinesin motor proteins [146, 257], namely KIFC5A and KRP2 (later renamed as KIFC1/HSET and KIF2C) [258]. In order to pinpoint candidate reactive Cys, we first generated a multiple sequence alignment of kinesin sequences homologous to KIFC1/HSET and KIF2C, including human and rat isoforms belonging to kinesin families 14 and 13, respectively (Figure S44). Noteworthily, the sensitivity to acrylamide is much higher for KIFC5A/KIF2C than for KRP2/KIFC1/HSET [146]. Therefore, in this case we focused on cysteine residues that are conserved in one kinesin family but not the other.*"

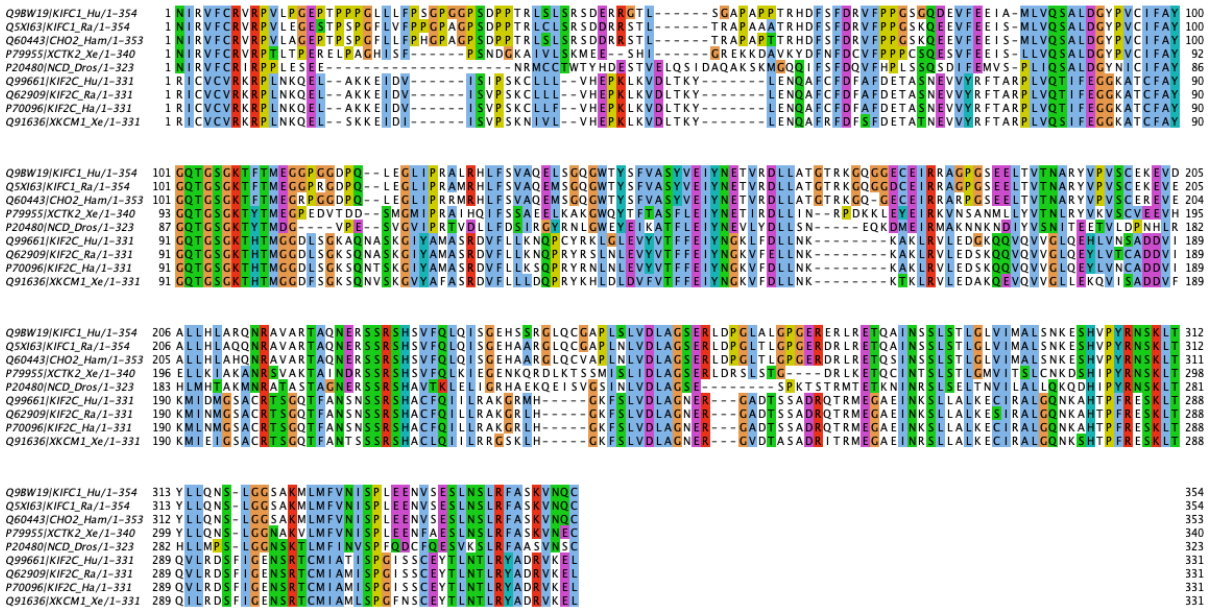
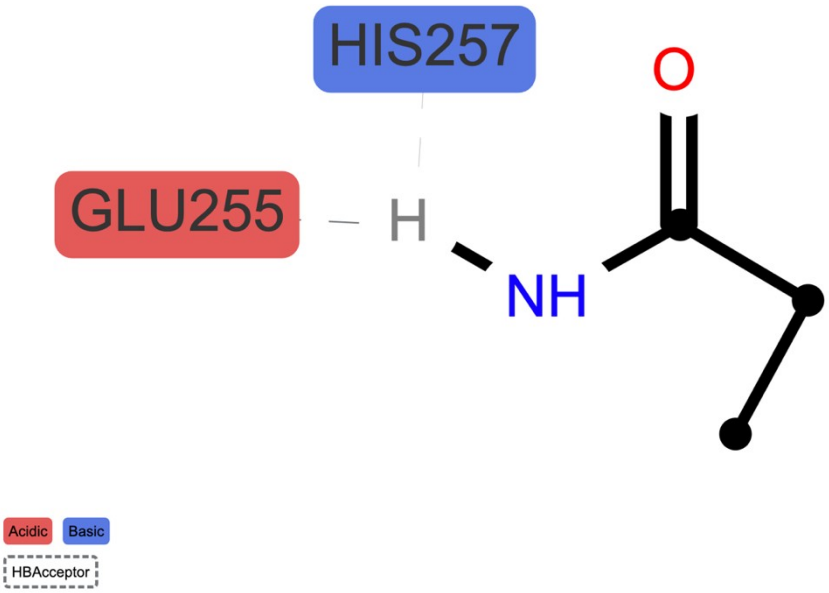


Figure S44. Multiple sequence alignment rat and human isoforms homologous to the kinesins KIFC1/HSET and KIF2C studied here, belonging to the kinesin 14 and 13 families, respectively. The MSA was generated using the MAFFT webserver [254] and the same color code as in Figure S26 was used.

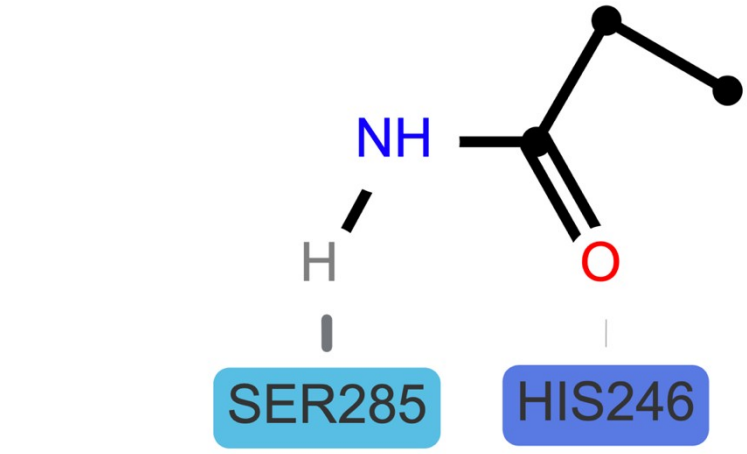
"*For KIF2C, we focused on the kinesin ATPase motor domain, as ACR-mediated inhibition of such catalytic domain has been observed for other neuronal proteins, such as NSF or v-ATPase (see main text). Among the six Cys conserved within the kinesin 13 family, integration of structural data revealed that C260 and C287 are more likely to play a functional role. In particular, a study performed for MACK [259], a hamster homolog of KIF2C, showed that ATPase activity requires dimerization of the motor domain and C260 and C287 are located at this dimeric interface. Moreover, dimerization is promoted by interaction with the C-terminal (CT) domain and C287 is located in the vicinity of CT [259]. Indeed, introduction of a Cys mutation in the CT domain resulted in formation of a disulfide bond with C287 [259]. Therefore, C260 and C287 (−35.3 and −34.1 a.u., respectively). We surmise that attachment of ACR to either C260 or C287 in KIF2C could interfere with motor domain dimerization and/or CT domain binding. For KIFC1/HSET, we again relied on experimental data to identify the best candidate residues among the conserved cysteines within the kinesin 14 family. In particular, a previous study identified C663 as an attachment site for covalent*

inhibitors [260]. Therefore, we selected C663 for the covalent docking of acrylamide to KIFC1, resulting in a favorable docking score of −14.0 a.u.. We speculate that binding of acrylamide to C663 could impair ATPase activity of KIFC1/HSET based on its structural position. This cysteine is at the C-terminal end of KIFC1 and is part of the so-called α6 helix, which together with α4 forms a cleft known to bind inhibitors [261]. Moreover, the α4-α46 cleft is located adjacent to the P-loop (or walker A motif) responsible for ATP binding and hydrolysis. Thus, ACR modification of C663 could allosterically trigger rearrangements of the P-loop through either helix α4 or α6. were selected for covalent docking to KIF2C. Favorable docking scores were obtained for both C260 and C287.”



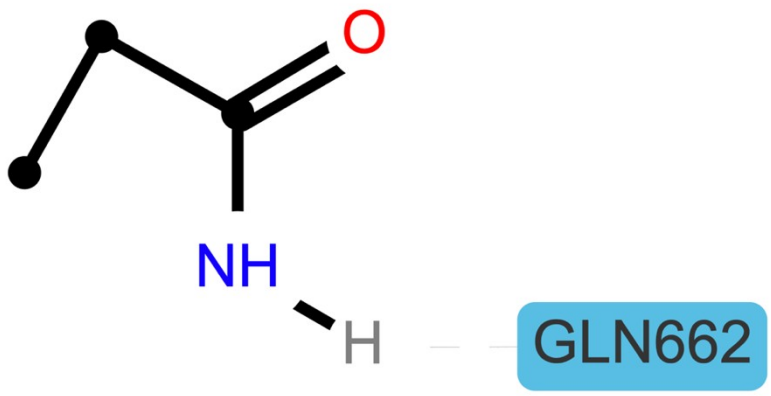
Kinesin KIF2C (C260) - Cluster 1			
residue	number	interaction	occurrence
Glu	255	HB acceptor	0.09
His	257	HB acceptor	0.01

Figure S 45. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



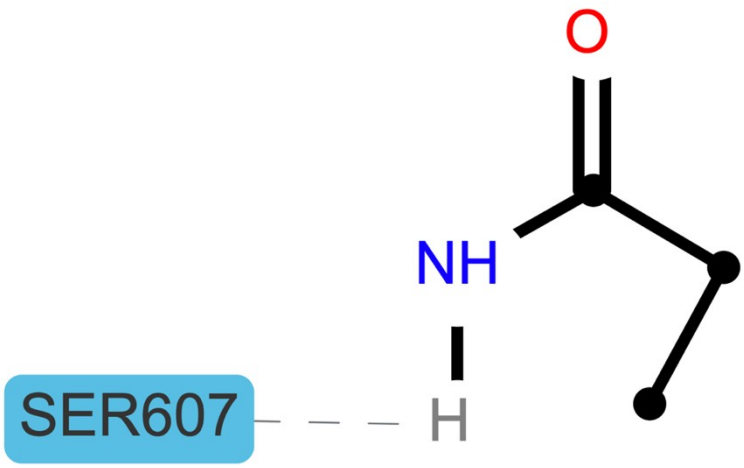
Kinesin KIF2C (C287) - Cluster 1			
residue	number	interaction	occurrence
His Ser	246 285	HB donor HB acceptor	0.05 0.66

Figure S46. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



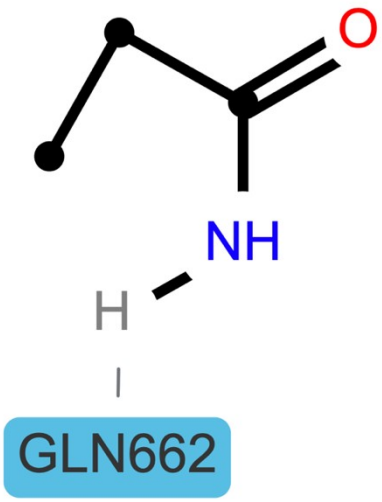
Kinesin KIF1C (C663) - Cluster 1			
residue	number	interaction	occurrence
Gln	662	HB acceptor	0.01

Figure S47. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Kinesin KIF1C (C663) - Cluster 3			
residue	number	interaction	occurrence
Ser	607	HB acceptor	0.11

Figure S48. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

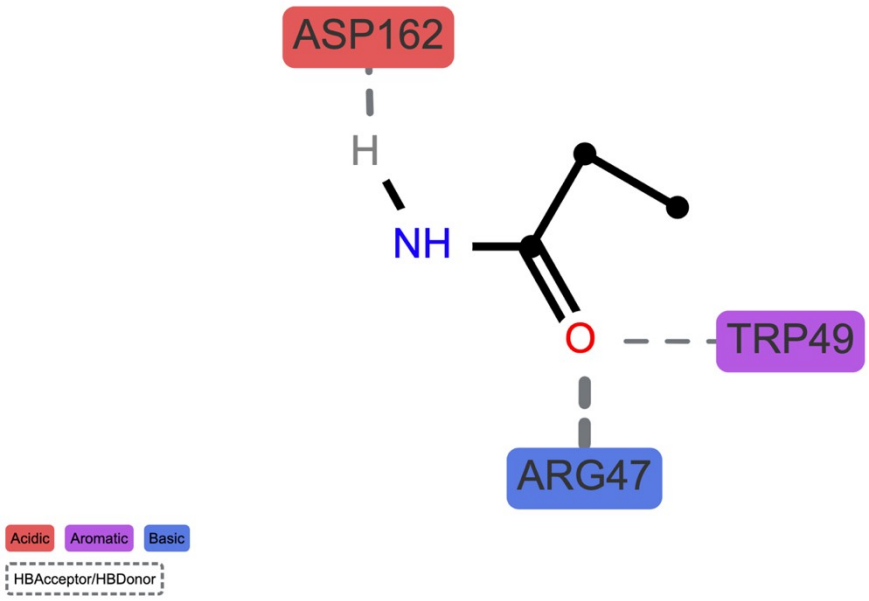


Kinesin KIF1C (C663) - Cluster 6			
residue	number	interaction	occurrence
Gln	662	HB acceptor	0.20

Figure S49. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

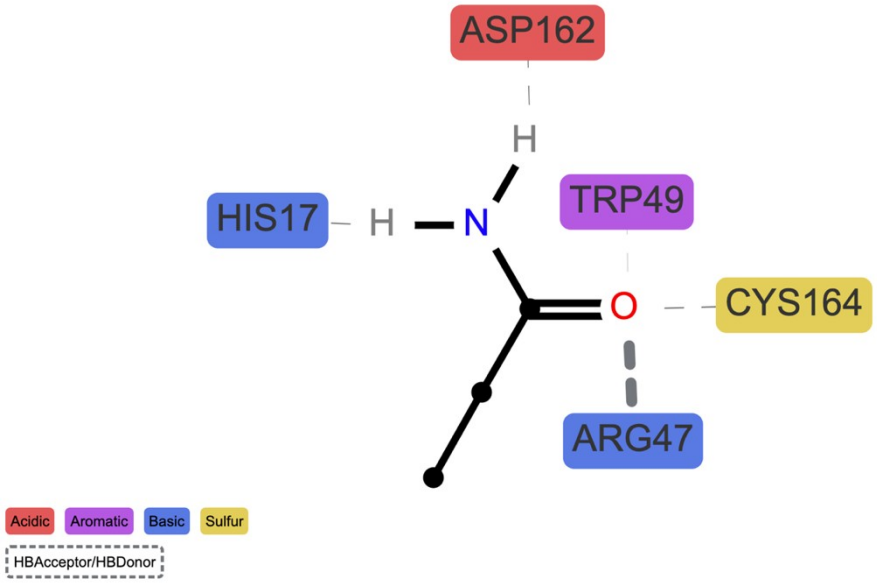
7.4.7 Sex Hormone-Binding Globulin (SHBG)

"*Sex hormone-binding globulin is a glycoprotein that binds steroid hormones to regulate the amount of free steroid molecules in plasma. SHBG contains cysteine-rich regions essential for ligand binding. Indeed, truncated mutants of rabbit SHBG, which did not contain the disulfide bond formed between C164 and C188, were not able to bind androgens anymore [262]. Moreover, these two cysteines are highly conserved among G-domains containing proteins, such as rabbit SHBG, and have been shown to contribute to protein stability [262]. Therefore, we selected C164 and C188 as potential ACR binding sites to perform the subsequent covalent docking calculations. Nonetheless, we would like to note that these two cysteines are involved in a disulfide bridge and thus prior reduction or chemical modification of the disulfide bridge would be needed to react with acrylamide, as discussed in section 7.4.5. The docking results show favorable docking scores for both candidate cysteines, −23.0 a.u. for C164 and −16.6 a.u. for C188, as well as stabilizing protein-ligand interactions (see figures below). Considering the aforementioned changes in ligand binding and protein stability upon mutation or truncation of these cysteines, it is tempting to speculate that ACR modification of C164 and C188 will have a similar effect.*"



Sex Hormone Binding Globulin (C164) - Cluster 1			
residue	number	interaction	occurrence
Arg	47	HB donor	1.00
Trp	49	HB donor	0.38
Asp	162	HB acceptor	0.59

Figure S50. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

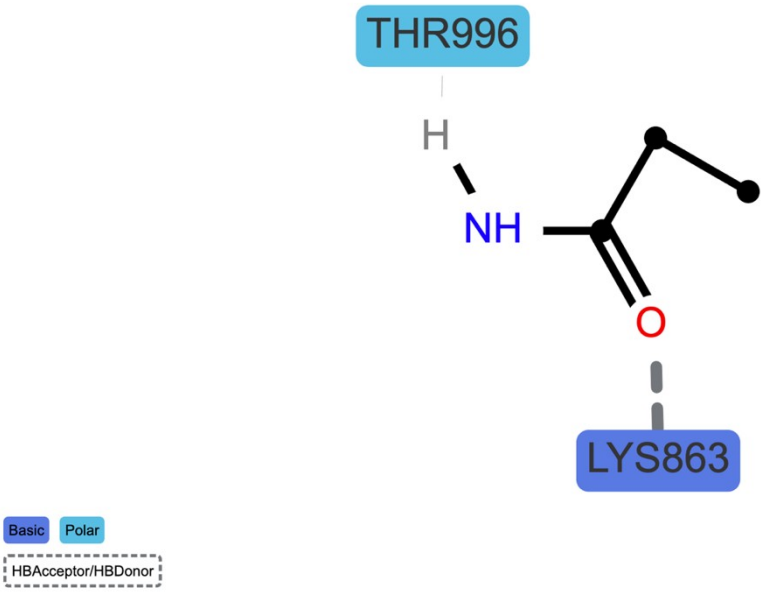


Sex Hormone Binding Globulin (C188) - Cluster 1			
residue	number	interaction	occurrence
His	17	HB acceptor	0.11
Arg	47	HB donor	1.00
Trp	49	HB donor	0.03
Asp	162	HB acceptor	0.09
Cys	164	HB donor	0.13

Figure S51. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.4.8 Topoisomerase IIa

"*Topoisomerases are highly conserved enzymes involved in multiple processes taking place at the cell nucleus, such DNA replication, chromosome condensation and chromosome segregation. Acrylamide has been shown to inhibit topoisomerase II activity in nuclear extracts [147]. However, the molecular mechanism of such inhibition is unclear, as acrylamide does not induce DNA cleavage, unlike other sulfhydryl-reactive agents modifying topoisomerase II. Therefore, we inspected the whole protein structure to pinpoint possible Cys sites targeted by ACR and selected two candidates, C170 and C997. C170 is located near the active site of the ATPase domain and thus its modification by ACR could hinder ATP binding, as shown for other ATPases in this study. Instead, C997 belongs to the Toprim (topoisomerase-primase) domain, the catalytic domain involved in DNA strand cleavage and religation. The location of C997 at the protein-DNA interface suggests that its modification by ACR could interfere with Toprim catalytic activity. Covalent docking to C170 showed that this cysteine is too buried inside the protein to allow formation of the adduct. In contrast, docking to C997 showed a favorable docking score of −8.4 a.u. for the top (best scored) cluster, thus supporting this cysteine as the primary site of ACR modification. This is in line with acrylamide being described as a catalytic inhibitor that reduces the amount of catalytically competent enzyme sensitive to the topoisomerase II poison etoposide [147].*"



Topoisomerase II (C997) - Cluster 1			
residue	number	interaction	occurrence
Lys Thr	863 996	HB donor HB acceptor	0.95 0.03

Figure S52. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

7.5 Analysis of hydrophobic interactions

"*Besides the H-bonds described see section 4.2.4, we also analyzed the hydrophobic interactions between the two methylene carbon atoms of the ACR adduct and nearby protein residues, as explained in section 4.1.3.The results are presented in Table S4. No clear preference for aliphatic/aromatic or small/branched amino acids was identified. We speculate that this is due to the small size and flexibility of the ACR adduct, which allow the ligand to form diverse (and non-directional) hydrophobic interactions.*" [1]

Table S4. Hydrophobic Interactions.

Protein name	Cysteine	Residue	Number	Atom	Occurrence
Albumin	34	Val	77	CG2	7.58
		Val	77	CG1	92.42
		Leu	31	CD1	64.39
		Leu	31	CD2	3.03
		Leu	41	CD1	39.39
		Asp	38	CB	57.58
Alcohol Dehydrogenase	170	Leu	166	CB	100.00
		Val	41	CG1	25.64
		Val	41	CB	69.23
		Ala	42	CB	100.00
		Ala	70	CB	46.15
	240	Lys	233	CG	12.50
		Phe	229	CB	61.11
		Phe	229	CD2	37.50
Aldolase	134	Ile	176	CG2	97.32
		Ala	135	CB	97.32
		Gln	179	CB	64.29
		Asn	180	CB	100.00
	239	Asp	197	CB	64.91
		Lys	241	CB	9.36
		Lys	241	CG	7.60
		Lys	241	CD	76.60
	268	Ile	29	CD1	94.20
		Leu	227	CD2	3.62
		Leu	227	CD1	15.22
		Leu	227	CB	81.16
		Val	23	CG1	91.30
		Val	23	CG2	1.45
	289	Glu	246	CB	18.84
		Ala	249	CB	97.10
		Ala	285	CB	11.59
		Met	250	CB	37.68
Creatine Kinase	141	His	145	CD2	96.83
		His	145	CG	1.59
	283	Val	75	CG1	60.66
		Val	72	CG1	77.05
		Val	72	CG2	9.84

Protein name	Cysteine	Residue	Number	Atom	Occurrence
		Leu	201	CD2	3.28
Dopamine Receptor	114	Phe	418	CZ	100.0
		Phe	417	CE2	85.44
		Val	120	CG1	1.94
		Trp	414	CZ2	0.97
		Trp	414	CZ3	98.06
Dopamine Transporter (inward)	342	Tyr	115	CE2	12.08
		Tyr	115	CE1	55.71
		Ala	119	CB	58.39
		Gln	122	CB	12.08
		Gln	122	CG	23.49
		Leu	118	CB	59.73
Dopamine Transporter (outward)	342	Ala	119	CB	1.19
		Gln	122	CB	2.38
		Let	511	CB	96.43
		Ile	508	CB	3.57
		Thr	512	CG	38.10
Enolase	388	Ile	143	CG1	0.83
		Leu	431	CD2	50.41
		Asn	139	CB	90.91
		Glu	141	CB	57.85
	398	Arg	514	CG	2.40
		Arg	514	CB	20.96
		Phe	212	CE2	2.99
		Val	206	CG1	71.86
		Tyr	188	CE2	41.32
Estrogen Receptor	381	His	247	CD2	7.41
	530	Tyr	226	CE2	31.68
		Tyr	226	CE1	43.56
		Lys	229	CD	2.97
		Lys	229	CB	0.99
Glyceraldehyde-3-phosphate dehydrogenase	152	Tyr	320	CD2	34.97
		Tyr	320	CG	0.61
		Tyr	320	CE2	5.52
		Asn	316	CB	13.50
		Ile	14	CD1	4.29
Hemoglobin	93	Thr	41	CG2	12.08
		Tyr	145	CD2	76.58
		Asp	94	CB	0.63
		His	146	CD2	30.38
		Lys	40	CB	0.63
Immunoglobulin G1 H Nie	395	Pro	378	CB	64.62
		Lys	464	CG	92.31
		Lys	464	CB	1.54
		Val	466	CG2	100.00
		Val	376	CG1	18.46

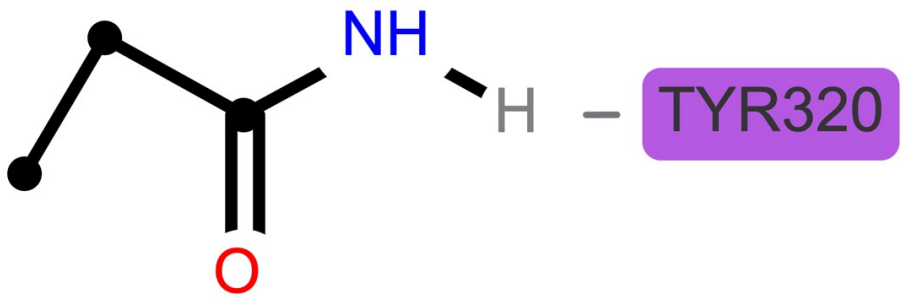
Protein name	Cysteine	Residue	Number	Atom	Occurrence
Immunoglobin kappa light chain	134	Lys	207	CB	28.68
		Ile	117	CB	63.24
		Ile	117	CD1	19.12
		Leu	136	CD1	2.21
		Val	115	CG1	44.12
		Val	115	CB	41.91
		Val	196	CG2	4.41
Kinesin KIFC1	663	Met	317	CB	0.93
		Lys	372	CB	2.78
		Ile	316	CG2	4.63
Kinesin KIF2C	260	Glu	49	CB	11.24
		Glu	49	CG	1.12
		Phe	107	CD2	6.74
		Phe	107	CE2	89.89
		Asp	106	CB	97.75
		His	51	CB	88.76
		His	51	CD2	5.62
	287	Lys	80	CB	1.681
		Lys	80	CG	53.782
		Leu	82	CD1	2.521
NEM-sensitive factor	264	Lys	266	CB	0.86
		Ala	439	CB	2.56
Sex Hormone-Binding Globulin	164	His	17	CG	3.01
		His	17	CD2	46.62
	188	His	17	CD2	10.00
Topoisomerase IIa (Toprim domain)	997	Tyr	907	CE2	0.971
		Leu	995	CB	8.74
		Leu	995	CD1	87.38
		Pro	865	CG	14.56
		Lys	863	CG	99.03
		Lys	863	CB	0.97
Vesicular proton ATPase	254	Ala	251	CB	99.09
		Lys	437	CB	70.91
		Lys	437	CG	1.82
		Pro	249	CB	0.91

7.6 Dependence of the covalent docking results on the input structure

"*As mentioned in section 4.1.2, the docking approach is fully flexible, so that the protein structure can adapt to the ACR covalent adduct. Nonetheless, we checked whether the results were robust with respect to the initial protein structure. As a test case, we chose GAPDH because (i) there are several experimental structures available and (ii) it contains three Cys residues that have been shown experimentally to be modified by ACR (C152, C156 and C247). Besides the protein structure reported in the main text (PDB code 4WNC), we tested two more (PDB codes 1U8F and 6YND), solved at different resolution (see Table S5). Regardless of the input structure, the primary ACR binding site C152 still exhibits a better docking score than the two Cys residues modified only at high ACR concentrations (C156 and 247). Therefore, this test case further supports that covalent docking can be used to pinpoint the most reactive Cys site within a given protein.*" [1]

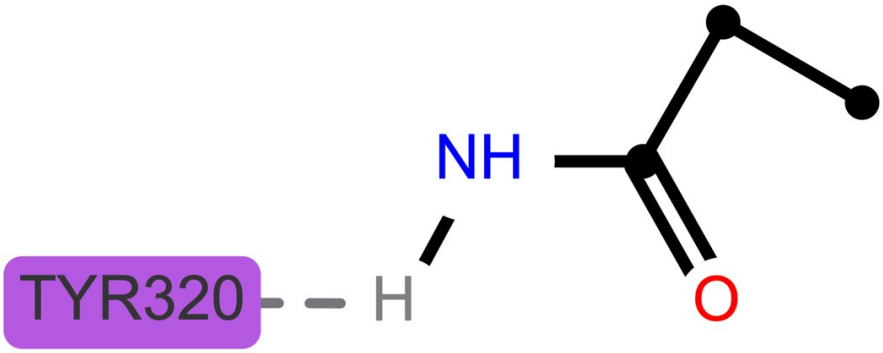
Table S5. Covalent docking results for a subset of GAPDH crystal structures. For each considered Cys, the Haddock score of the top docking cluster is shown.

Protein name	PDB code	Resolution (Å)	C152	C156	C247
Glyceraldehyde-	4WNC	1.99	-12.3	46.1	9.9
3-phosphate	1U8F	1.75	-16.8	-3.1	0.3
dehydrogenase	6YND	1.52	-9.5	8.2	7.8



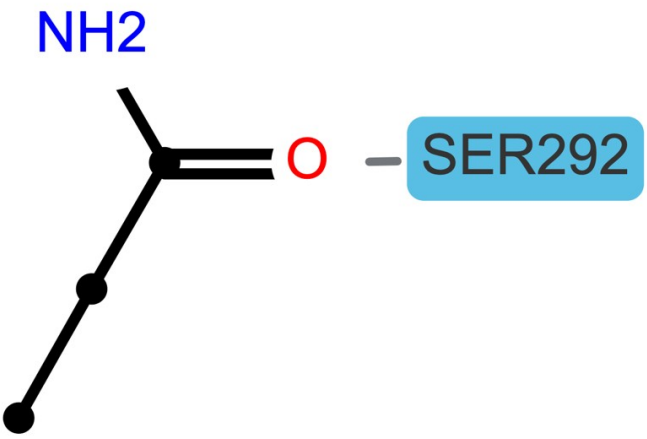
Glyceraldehyde-3-phosphate dehydrogenase (C152) – PDB 1U8F - Cluster 1			
residue	number	interaction	occurrence
Tyr	320	HB acceptor	0.39

Figure S53. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Glyceraldehyde-3-phosphate dehydrogenase (C152) – PDB 6YND - Cluster 1			
residue	number	interaction	occurrence
Tyr	320	HB acceptor	0.60

Figure S54. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.



Glyceraldehyde-3-phosphate dehydrogenase (C156) – PDB 6YND - Cluster 1			
residue	number	interaction	occurrence
Ser	292	HB acceptor	0.50

Figure S55. Modeled covalent adduct between acrylamide (ACR) and the target protein. Here, the same representation as in Figure 12 was used.

8 Appendix B

This part of the appendix includes supplementary material for the PPARα section (see chapter 5).

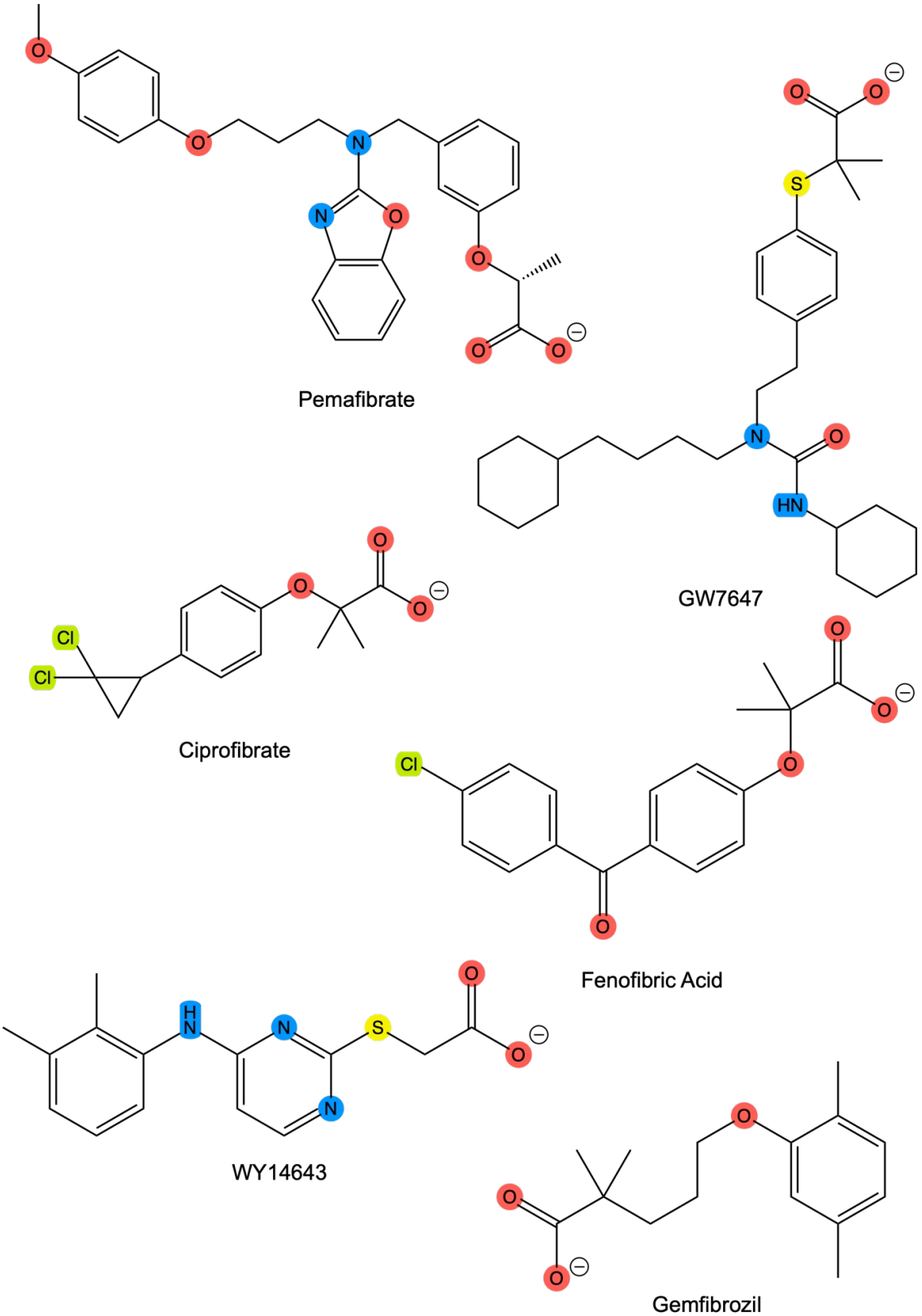


Figure S56. Chemical structures of known PPARα agonists mentioned in this thesis.

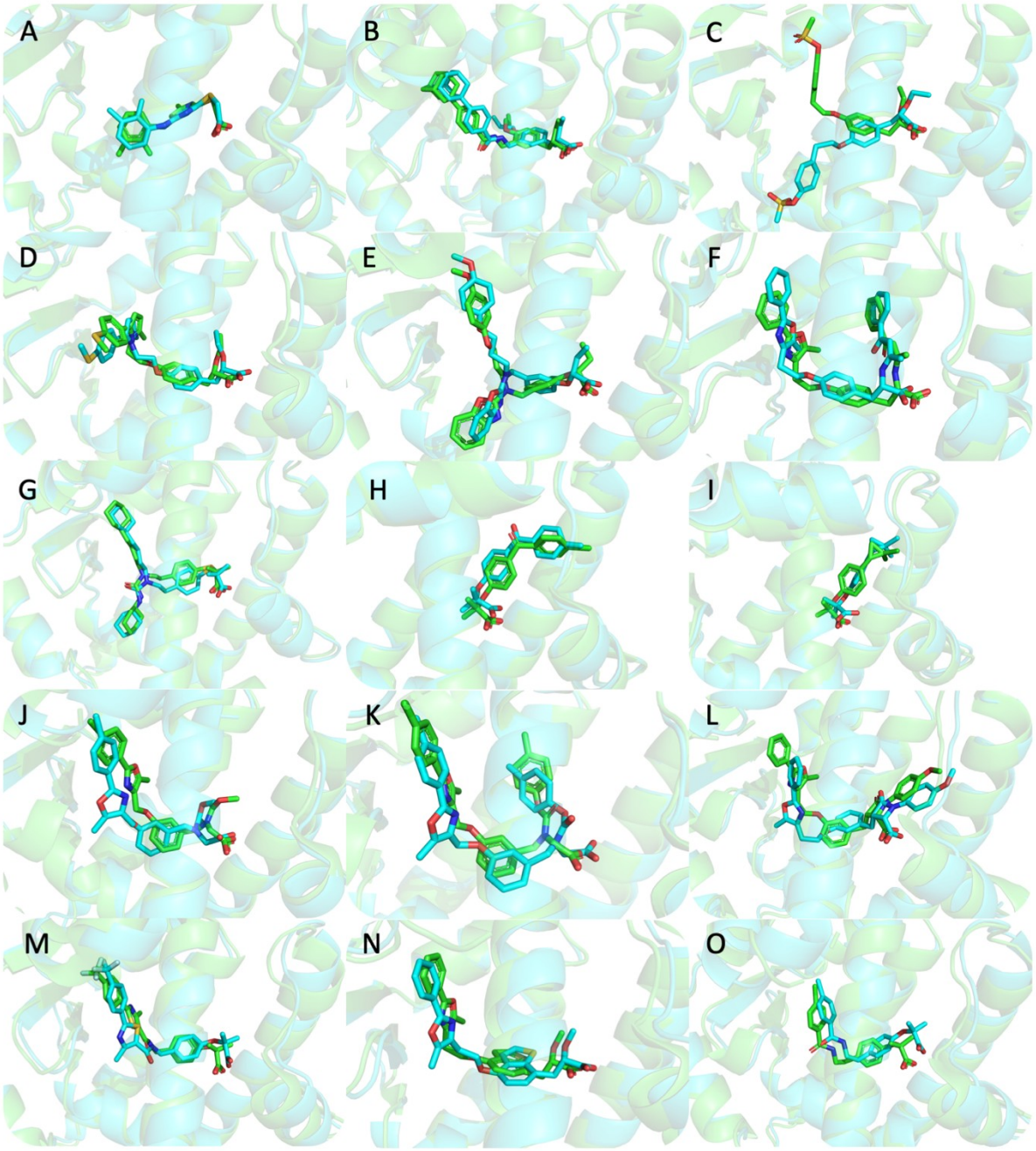


Figure S57. Validation dockings performed with HADDOCK. The top scoring pose according to PRODIGY-LIG was compared to the crystallographic pose. (A) WY14643; (B) TIPP-703; (C) Tesaglitazar; (D) Saroglitazar; (E) Pemafibrate; (F) GW409544; (G) GW7647; (H) Fenofibric Acid; (I) Ciprofibrate; (J) CHEMBL1089501; (K) CHEMBL1089210; (L) CHEMBL271240; (M) CHEMBL219586; (N) Aleglitazar; (O) Bezafibrate.

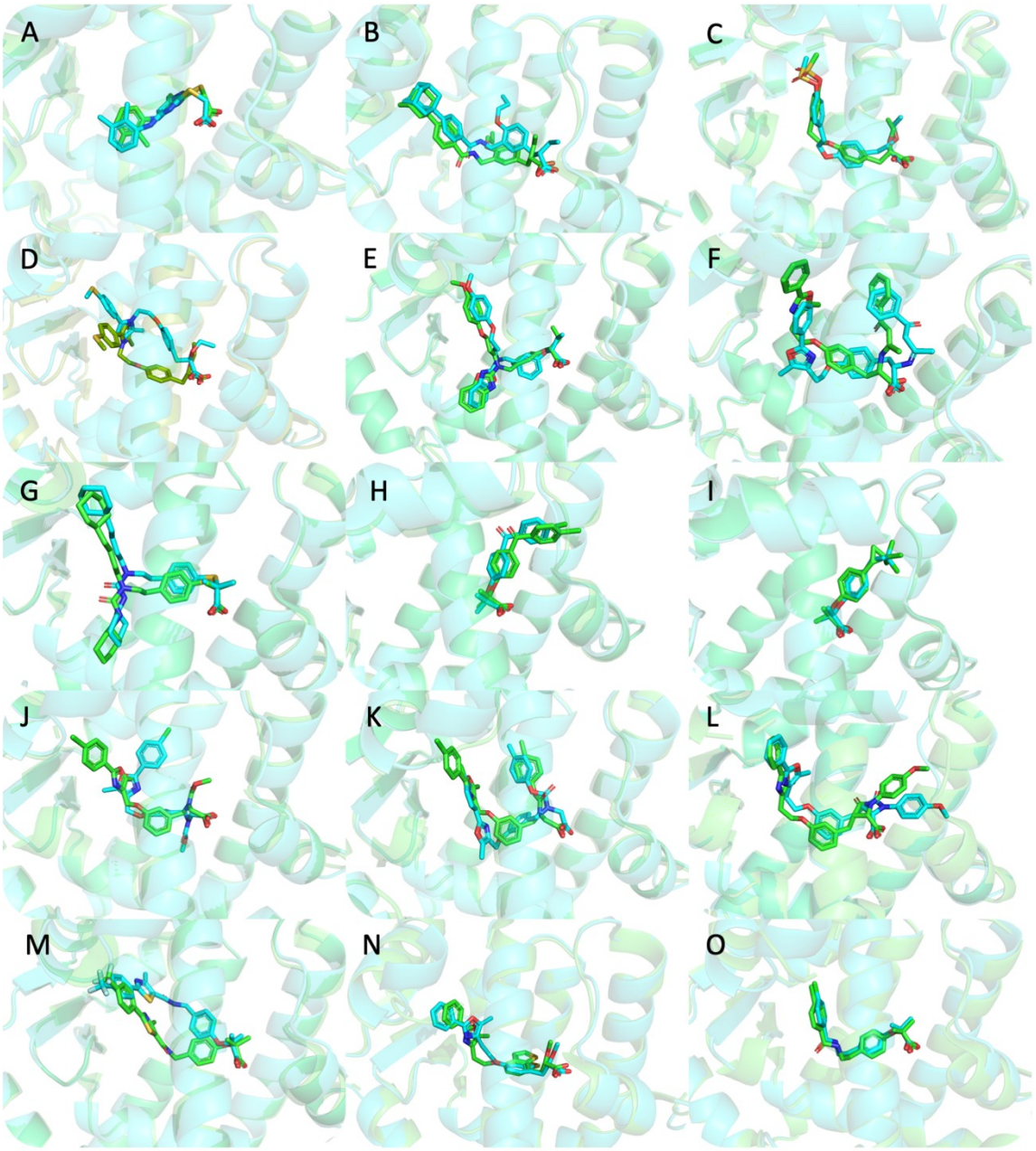


Figure S58. Validation dockings performed with Glide. Among the five best docking poses the one with the lowest RMSD value was compared to the X-ray structure. (A) WY14643; (B) TIPP-703; (C) Tesaglitazar; (D) Saroglitazar; (E) Pemafibrate; (F) GW409544; (G) GW7647; (H) Fenofibric Acid; (I) Ciprofibrate; (J) CHEMBL1089501; (K) CHEMBL1089210; (L) CHEMBL271240; (M) CHEMBL219586; (N) Aleglitazar; (O) Bezafibrate.

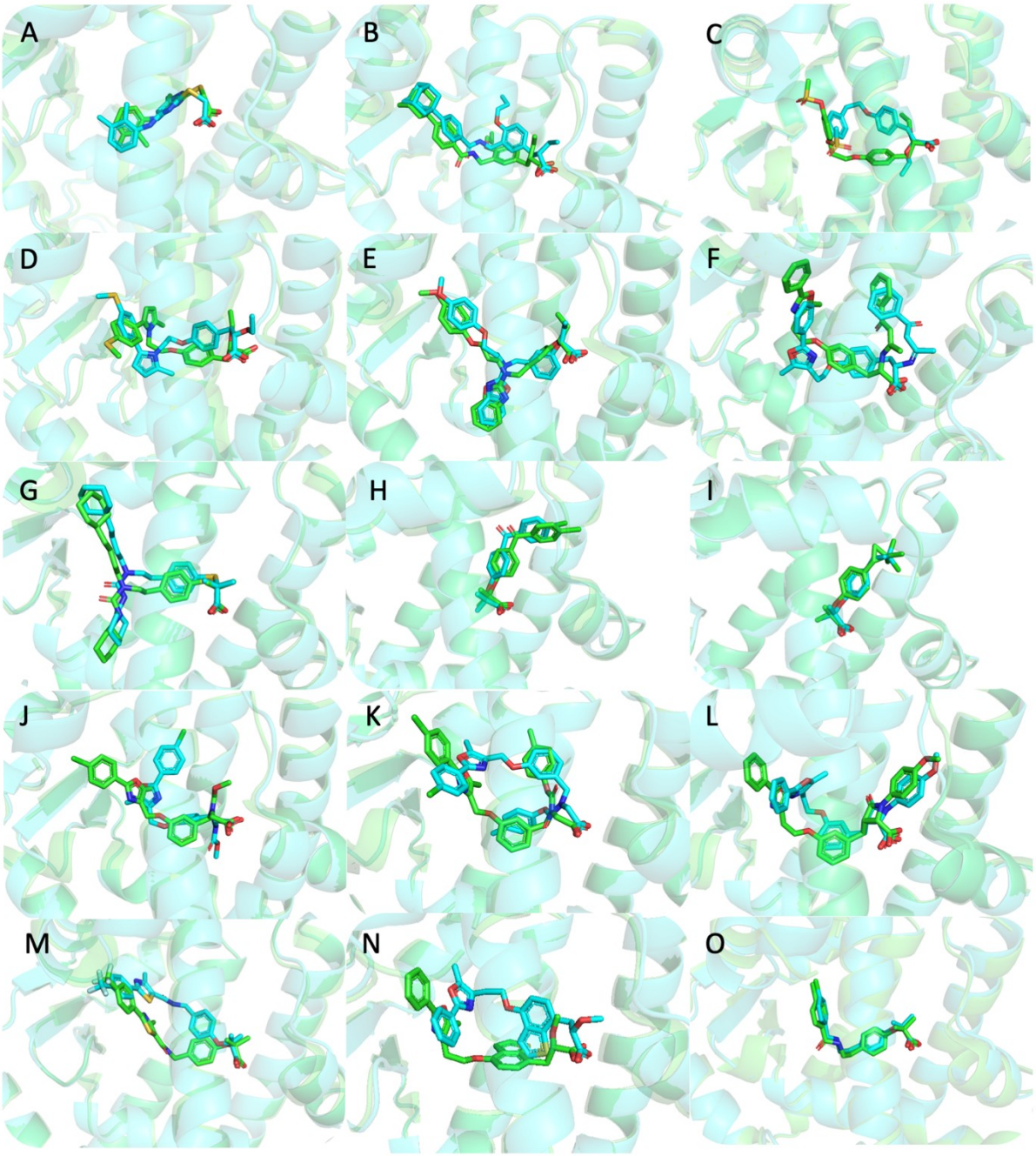


Figure S59. Validation dockings performed with Glide. The top scored docking pose was compared to the X-ray structure. (A) WY14643; (B) TIPP-703; (C) Tesaglitazar; (D) Saroglitazar; (E) Pemafibrate; (F) GW409544; (G) GW7647; (H) Fenofibric Acid; (I) Ciprofibrate; (J) CHEMBL1089501; (K) CHEMBL1089210; (L) CHEMBL271240; (M) CHEMBL219586; (N) Aleglitazar; (O) Bezafibrate.

Table S6. PPARα known agonists, as well as chlorogenic acids and related compounds considered in this study, listed in alphabetical order. If the cis isomer of a compound had no PubChem ID, however, the trans isomer was listed, latter mentioned entry is provided. [1]Numbers in the "class" column correspond to the following groups: (1) Hydroxycinnamic acids, (2) chlorogenic acids, (3) di-chlorogenic acids, (4) tri-chlorogenic acids, (5) known agonists of PPARα and (6) related compounds.

Ligand	**PubChem ID**	**Class[1]**
3-(3-Hydroxyphenyl)-Propanoic Acid	91	1
3,4-Dihydroxyphenylacetic Acid	547	1
3-Hydroxybenzoic Acid	7420	6
3-Hydroxyhippuric Acid	450268	6

Ligand	PubChem ID	Class[1]
3-Hydroxyphenylacetic Acid	12122	6
4-Hydroxybenzoic Acid	135	1
4-Hydroxyhippuric Acid	151012	1
Benzoic Acid	243	1
Bezafibrate	39042	5
cis-3,4-Dimethoxycinnamic Acid	158026	1
cis-3-*cis*-4-Dicaffeoylquinic Acid	5281780	3
cis-3-*cis*-5-Dicaffeoylquinic Acid	13604688	3
cis-3-Hydroxycinnamic Acid	5316112	1
cis-3-Caffeoylquinic Acid	1794425	2
cis-3-Feruloylquinic Acid	131752769	2
cis-3-Sinapoylquinic Acid	72193657	2
cis-3-Trimethoxycinnamoyl-*cis*-4-Feruloylquinic Acid	n.a.	3
cis-3-Trimethoxycinnamoyl-*cis*-5-Caffeoylquinic Acid	n.a.	3
cis-3-Trimethoxycinnamoyl-*cis*-5-Feruloylquinic Acid	n.a.	3
cis-3-Trimethoxycinnamoyl-*trans*-4-Feruloylquinic Acid	n.a.	3
cis-3-Trimethoxycinnamoyl-*trans*-5-Caffeoylquinic Acid	n.a.	3
cis-3-Trimethoxycinnamoyl-*trans*-5-Feruloylquinic Acid	n.a.	3
cis-3-p-Coumaroylquinic Acid	164893	2
cis-3-*trans*-4-Dicaffeoylquinic Acid	5281780	3
cis-3-*trans*-5-Dicaffeoylquinic Acid	13604688	3
cis-4-*cis*-5-Dicaffeoylquinic Acid	5281780	3
cis-4-Hydroxycinnamic Acid	1549106	1
cis-4-Caffeoylquinic Acid	58427569	1
cis-4-Feruloylquinic Acid	10177048	1
cis-4-Sinapoylquinic Acid	72193643	1
cis-4-Trimethoxycinnamoyl-*cis*-5-Caffeoylquinic Acid	n.a.	3
cis-4-Trimethoxycinnamoyl-*cis*-5-Feruloylquinic Acid	n.a.	3
cis-4-Trimethoxycinnamoyl-*trans*-5-Caffeoylquinic Acid	n.a.	3
cis-4-Trimethoxycinnamoyl-*trans*-5-Feruloylquinic Acid	n.a.	3
cis-4-p-coumaroylquinic Acid	5281766	2
cis-4-*trans*-5-Dicaffeoylquinic Acid	5281780	3
cis-5-Caffeoylquinic Acid	1794425	2
cis-5-Feruloylquinic Acid	73210496	2
cis-5-Sinapoylquinic Acid	72193641	2
cis-5-p-coumaroylquinic Acid	90478782	2
cis-Caffeic Acid	1549111	1
cis-Cinnamic Acid	5372954	1
cis-Ferulic Acid	1548883	1
cis-Isoferulic Acid	1549043	1
cis-Sinapic Acid	1549091	1
cis-Trimethoxycaffeic Acid	n.a.	1
Dihydrocaffeic Acid	15847196	1
Dihydroferulic Acid	17865499	1
Dihydroisoferulic Acid	2752054	1
Feruloylglycine	5280527	1
Gallic Acid	370	6
Gemfibrozil	3463	5

Ligand	PubChem ID	Class[1]
GW7647	3392731	5
Hippuric Acid	464	6
Quinic Acid	6508	1
Saroglitazar	495399	5
Syringic Acid	10742	6
trans-3-4-Dimethoxycinnamic Acid	717531	3
trans-3-*cis*-4-Dicaffeoylquinic Acid	5281780	3
trans-3-*cis*-5-Dicaffeoylquinic Acid	13604688	3
trans-3-Hydroxycinnamic Acid	637541	1
trans-3-Caffeoylquinic Acid	5280633	2
trans-3-Feruloylquinic Acid	10133609	2
trans-3-Sinapoylquinic Acid	72193657	2
trans-3-Sinapoyl-*trans*-4-*trans*-5-Dicaffeoylquinic Acid	n.a.	4
trans-3-Trimethoxycinnamoyl-*cis*-4-Feruloylquinic Acid	n.a.	3
trans-3-Trimethoxycinnamoyl-*cis*-5-Caffeoylquinic Acid	n.a.	3
trans-3-Trimethoxycinnamoyl-*cis*-5-Feruloylquinic Acid	n.a.	3
trans-3-Trimethoxycinnamoyl-*trans*-4-Feruloylquinic Acid	n.a.	3
trans-3-Trimethoxycinnamoyl-*trans*-5-Caffeoylquinic Acid	n.a.	3
trans-3-Trimethoxycinnamoyl-*trans*-5-Feruloylquinic Acid	n.a.	3
trans-3-p-Coumaroylquinic Acid	9945785	2
trans-3-*trans*-4-Dicaffeoylquinic Acid	5281780	3
trans-3-*trans*-5-Dicaffeoylquinic Acid	13604688	3
trans-4-*cis*-5-Dicaffeoylquinic Acid	5281780	3
trans-4-Hydroxycinnamic Acid	322	1
trans-4-Caffeoylquinic Acid	58427569	2
trans-4-Feruloylquinic Acid	101024370	2
trans-4-Sinapoylquinic Acid	72193643	2
trans-4-Trimethoxycinnamoyl-*cis*-5-Caffeoylquinic Acid	n.a.	3
trans-4-Trimethoxycinnamoyl-*cis*-5-Feruloylquinic Acid	n.a.	3
trans-4-Trimethoxycinnamoyl-*trans*-5-Caffeoylquinic Acid	n.a.	3
trans-4-Trimethoxycinnamoyl-*trans*-5-Feruloylquinic Acid	n.a.	3
trans-4-p-Coumaroylquinic Acid	5281766	2
trans-4-*trans*-5-Dicaffeoylquinic Acid	5281780	3
trans-5-Caffeoylquinic Acid	5280633	2
trans-5-Feruloylquinic Acid	73210496	2
trans-5-Sinapoylquinic Acid	72193641	2
trans-5-p-Coumaroylquinic Acid	6441280	2
trans-Caffeic Acid	689043	1
trans-Cinnamic Acid	444539	1
trans-Ferulic Acid	445858	1
trans-Isoferulic Acid	736186	1
trans-Sinapic Acid	54710960	1
trans-Trimethoxycaffeic Acid	n.a.	1
Vanillic Acid	8468	6

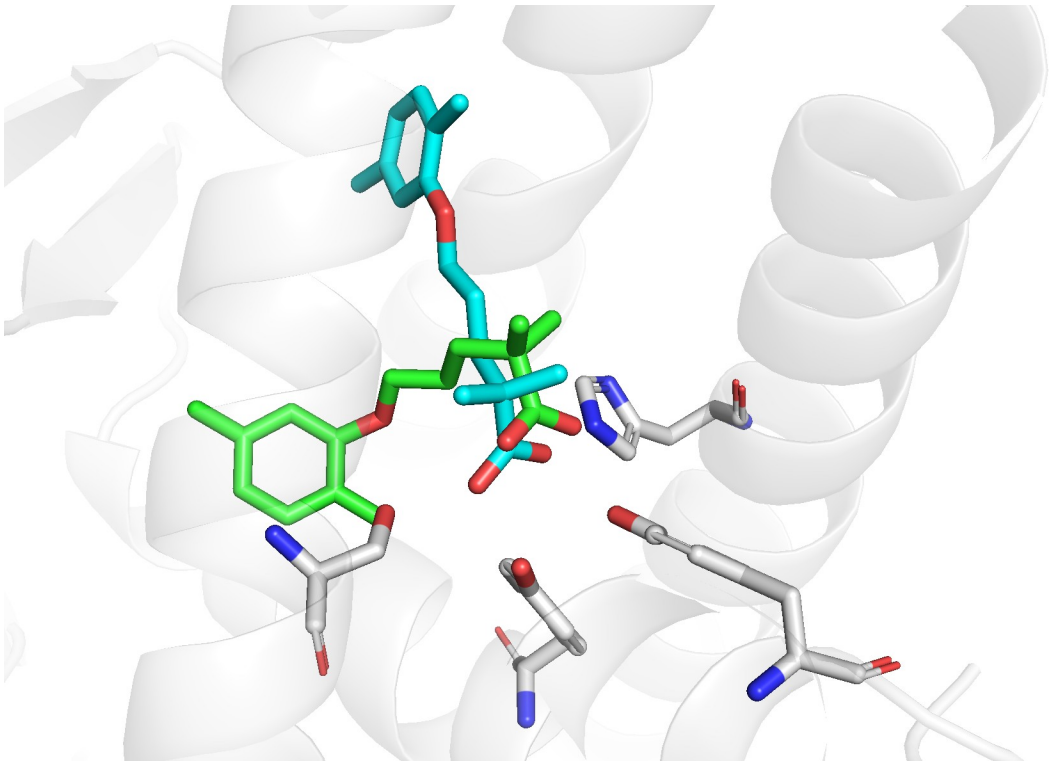


*Figure S60. Starting structures for simulations in which Gemfibrozil is bound to the Center/Arm II/Arm III pocket. The protein structure (PDB Code 6KB3) is represented as cartoon representation, whereas residues S280, S314, H440, Y464 and Gemfibrozil are shown as sticks. (**Green**) Starting pose of simulations R1 and R2 (HADDOCK); (**Cyan**) Starting pose of simulations R3 and R4 (Glide).*

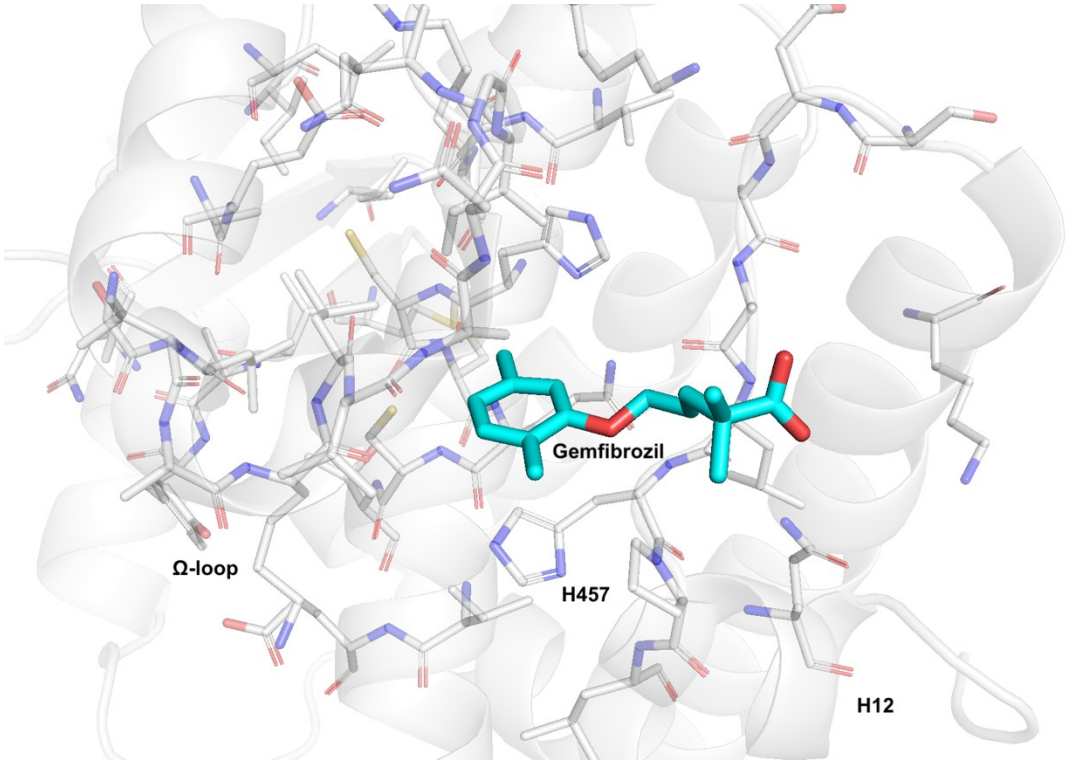


*Figure S61. Cluster centroid of R4 (M1). Protein structure is shown as cartoon representation in grey. Gemfibrozil (**cyan**) and surrounding residues are shown as sticks.*

8.1 GW7647

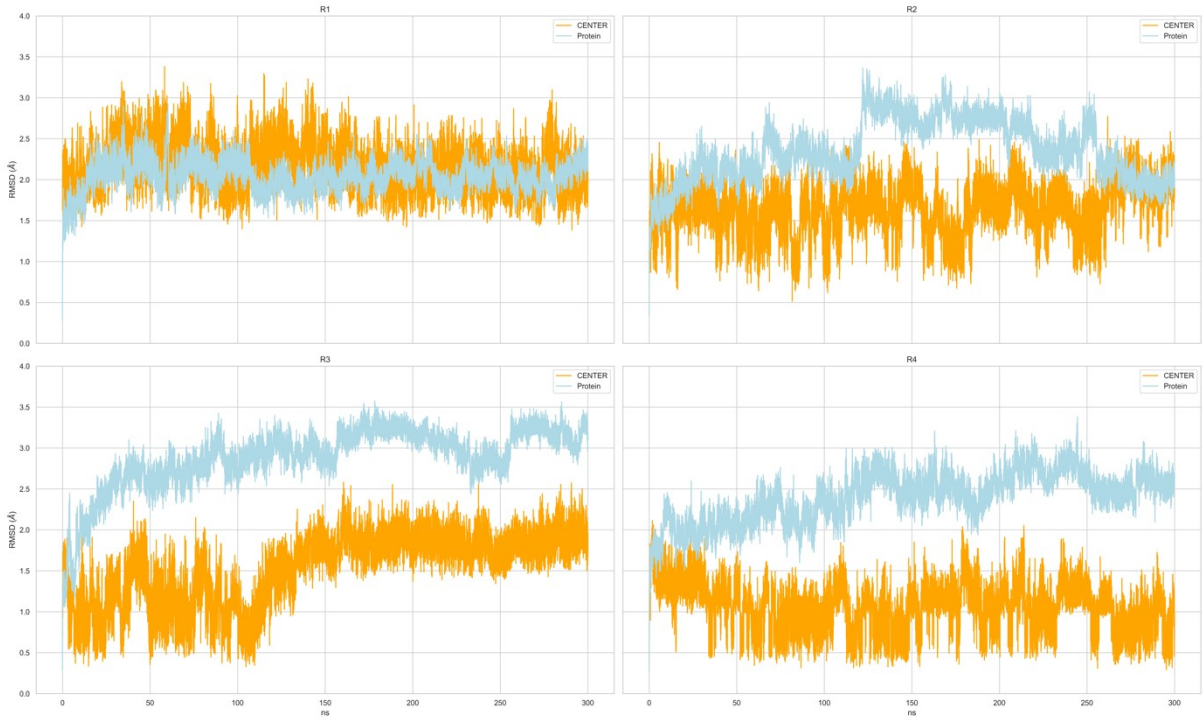


Figure S62. RMSD calculation of the protein backbone (light blue) and GW7647 (orange). The starting structure of the MD simulations is the Glide top docking pose (R1), the Haddock top docking pose (R2) and the crystal structure with PDB code 6KB3 (R3 and R4, each with different initial velocities). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

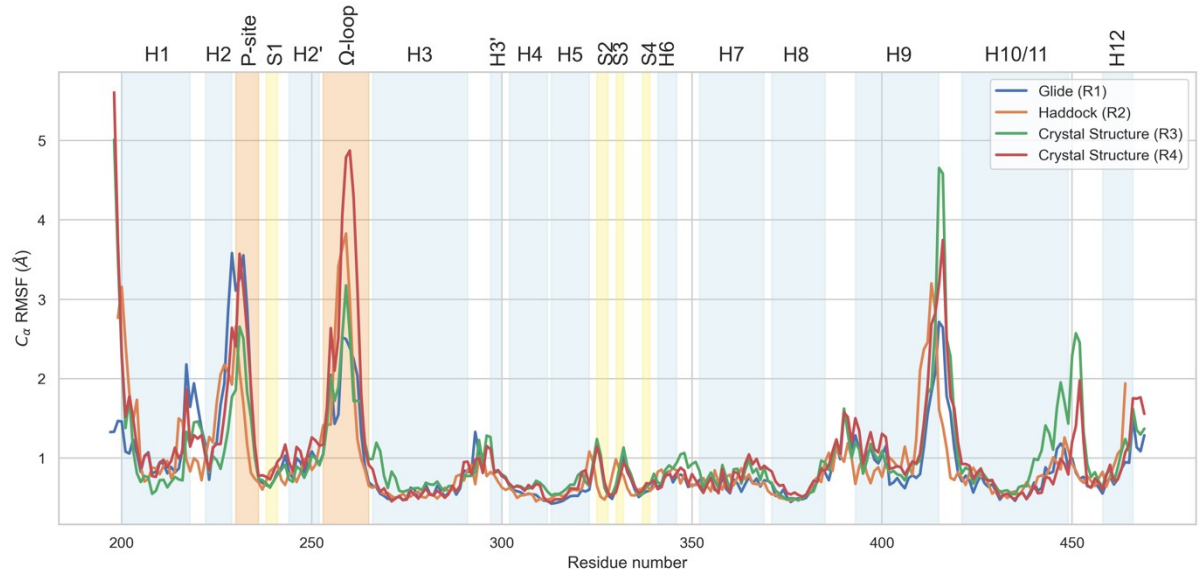


Figure S63. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

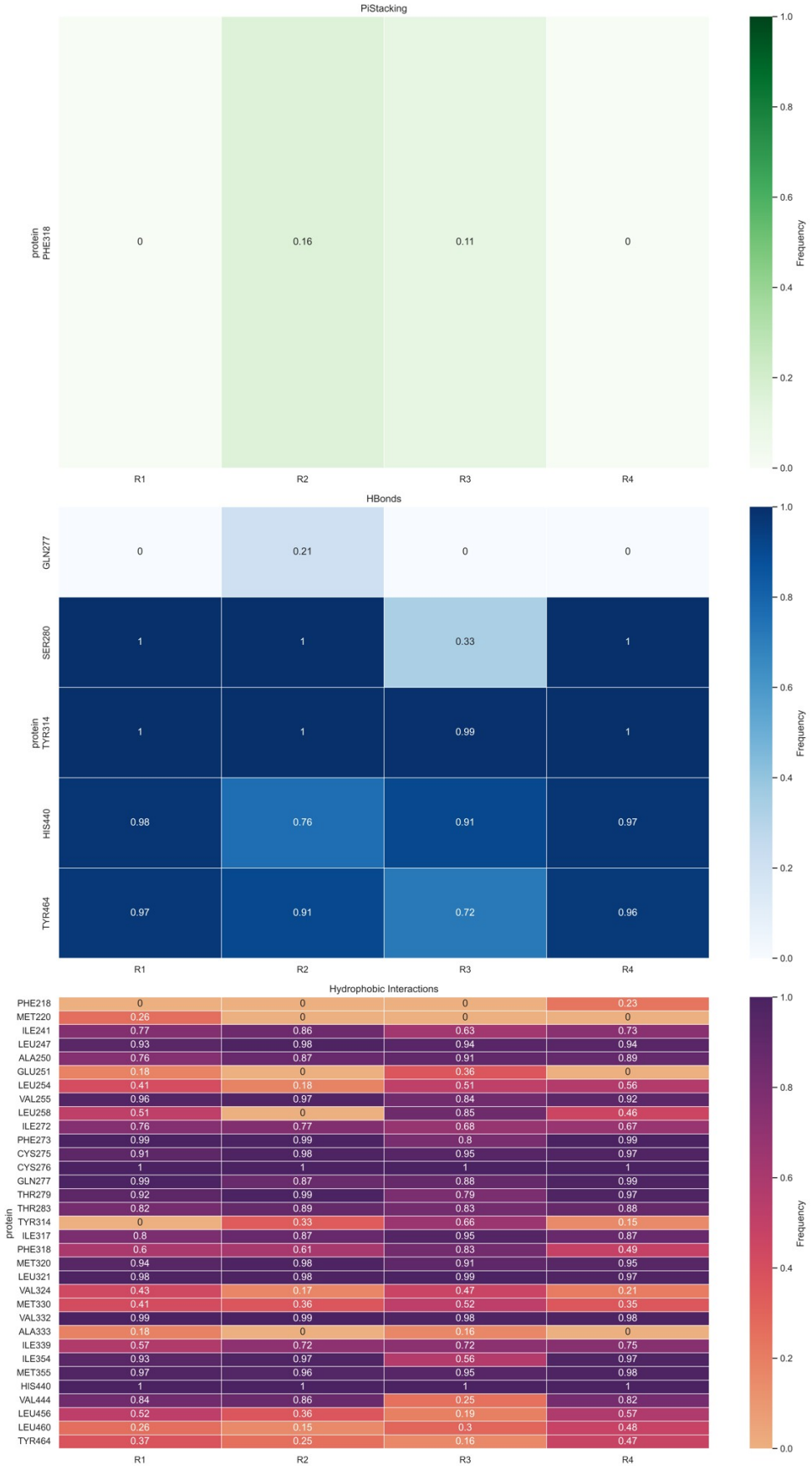


Figure S64. Heatmap of GW7647-protein interactions.

Figure S65. Detailed overview of H-bonds formed between GW7647 and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

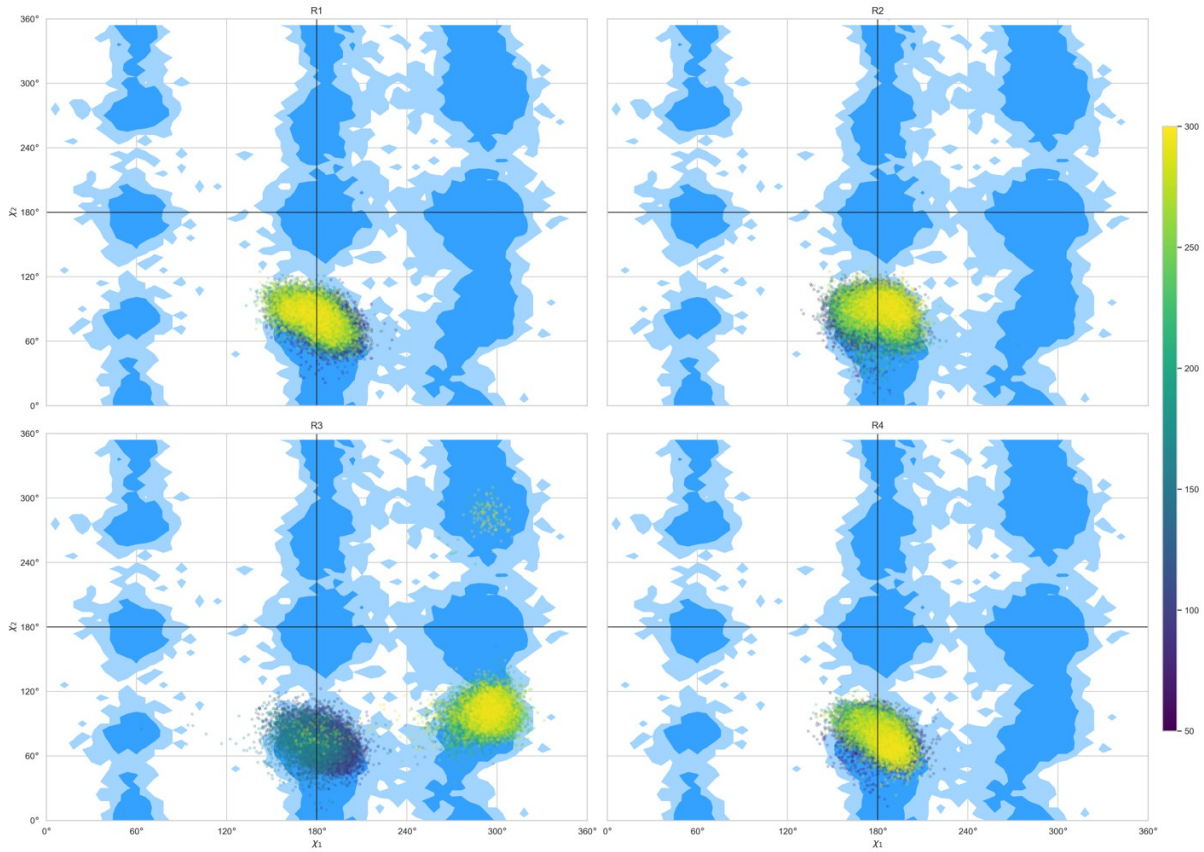


Figure S66. Time evolution of the Chi$_1$ and Chi$_2$ angles of F273. A data point represents the Chi$_1$ - Chi$_2$ angle combination in one particular MD frame. The colour coding indicates simulation time.

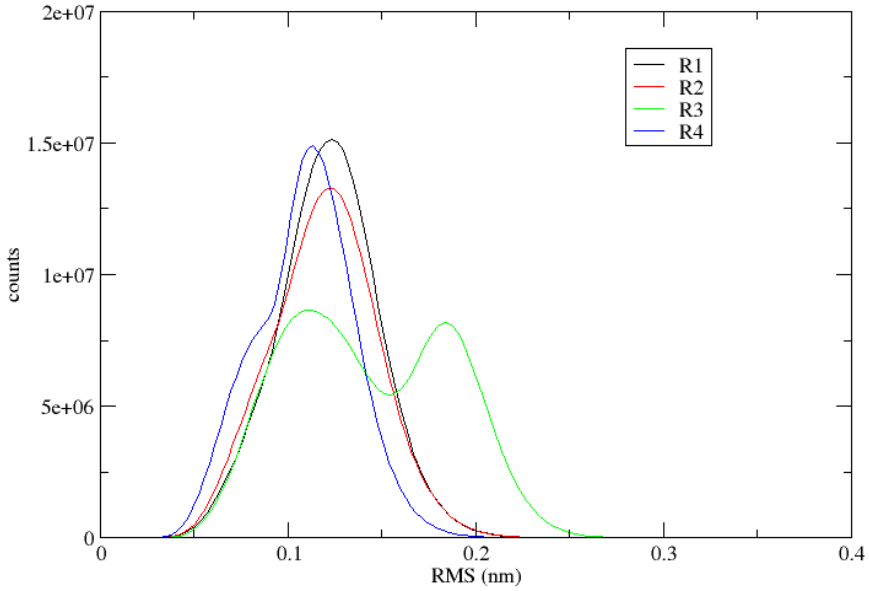


Figure S67. RMSD distribution of the GW6747 ligand in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.13nm was selected to cluster the respective trajectories into representative groups (see Table S7).

Table S7. Overview of clusters from simulations of GW7647 and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

GW7647			
Pocket	Replica	# clusters	Percentage (%)
CENTER	R1	3	(1) **93.0** (2) **4.2** (3) **2.5**
	R2	3	(1) **94.0** (2) **4.4** (3) **1.4**
	R3	4	(1) **69.3** (2) **26.2** (3) **2.1** (4) **1.3**
	R4	2	(1) **97.2** (2) **2.2**

8.2 Ciprofibrate

8.2.1 Ciprofibrate bound to Arm I

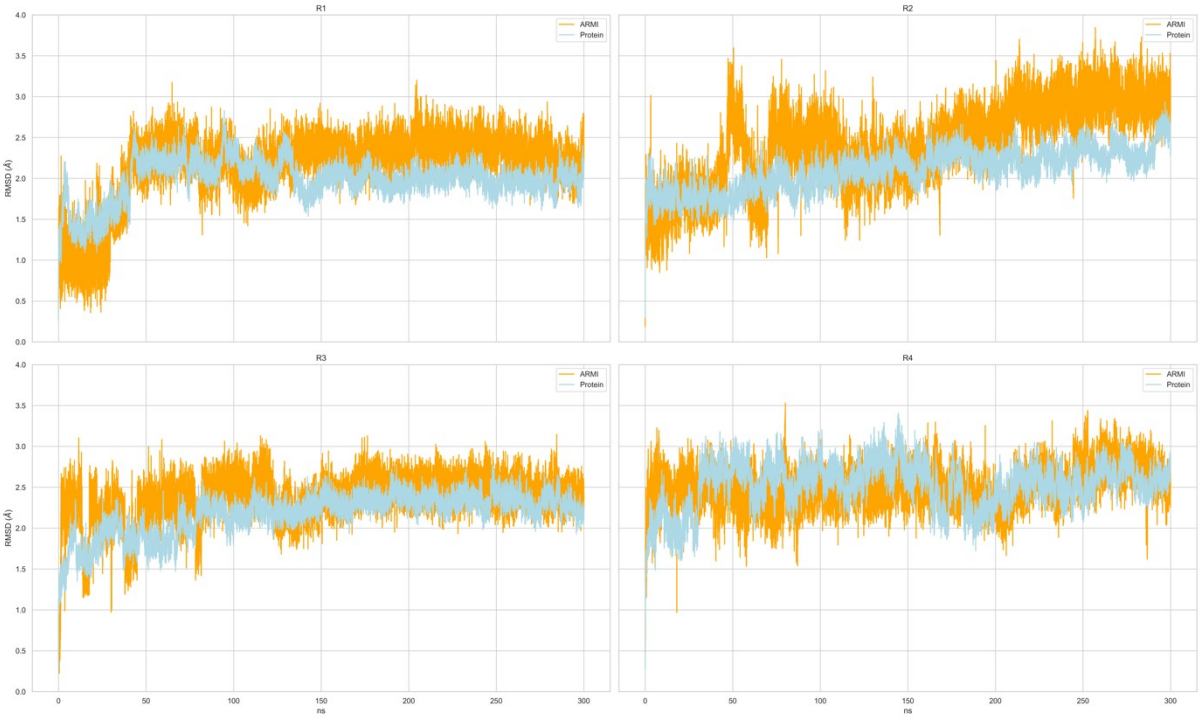


Figure S68. RMSD calculation of the protein backbone (light blue) and ciprofibrate bound to Arm I (orange). The starting structure of the MD simulations is the Glide top docking pose (R1), the Haddock top docking pose (R2) and the crystal structure with PDB code 6KB3 (R3 and R4, each with different initial velocities). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

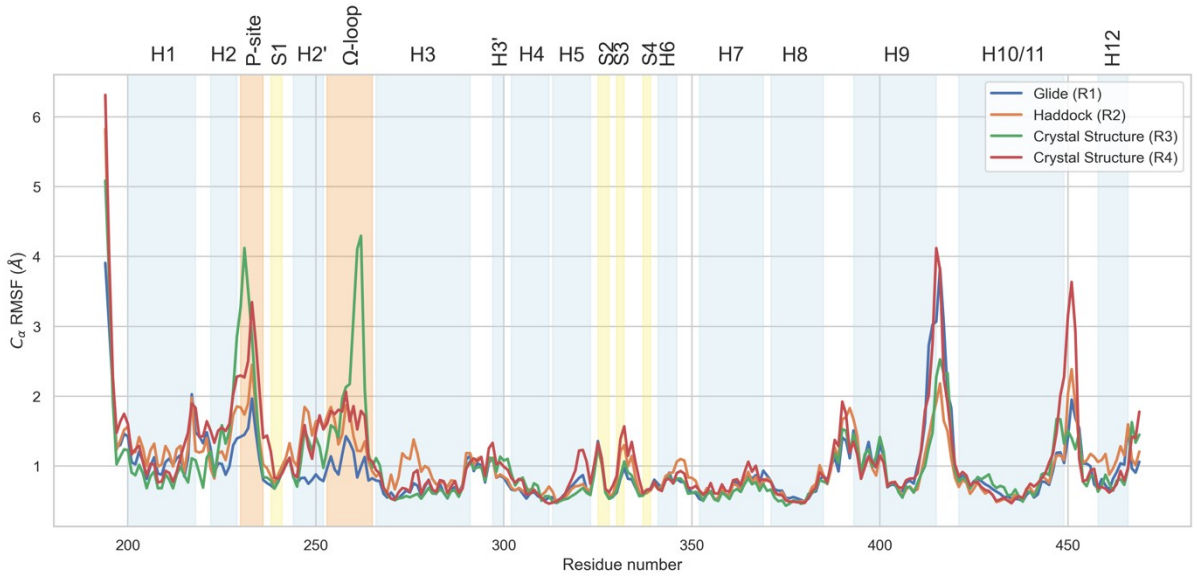


Figure S69. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

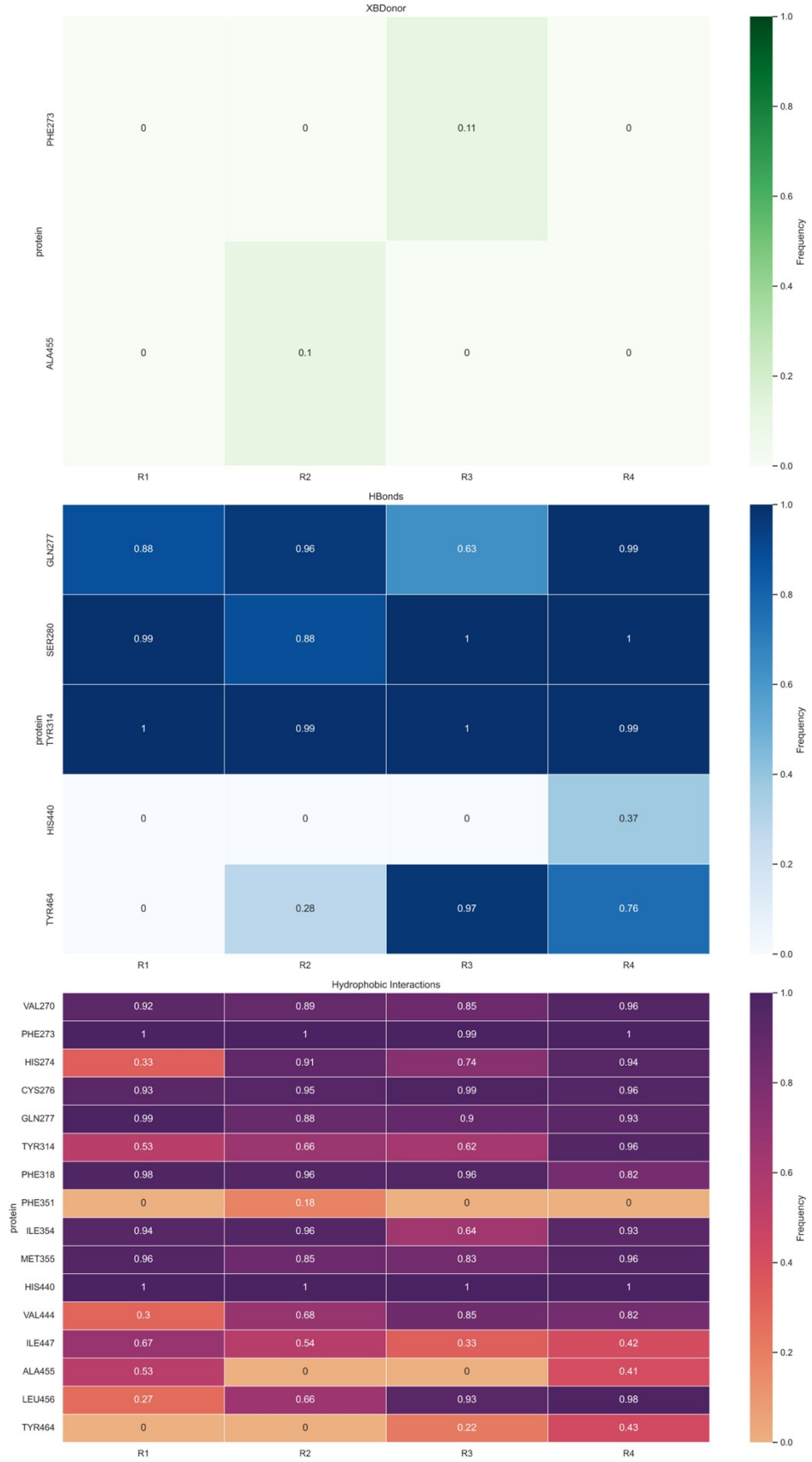


Figure S70. Heatmap of ciprofibrate-protein interactions

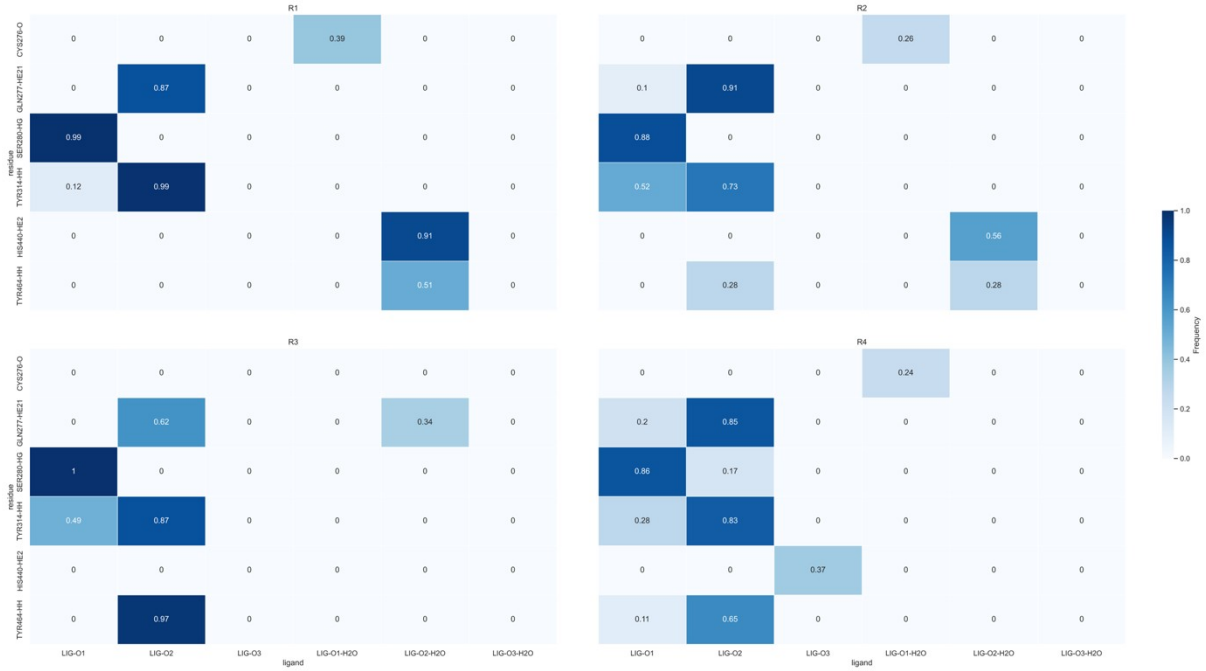


Figure S71. Detailed overview of H-bonds formed between ciprofibrate and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

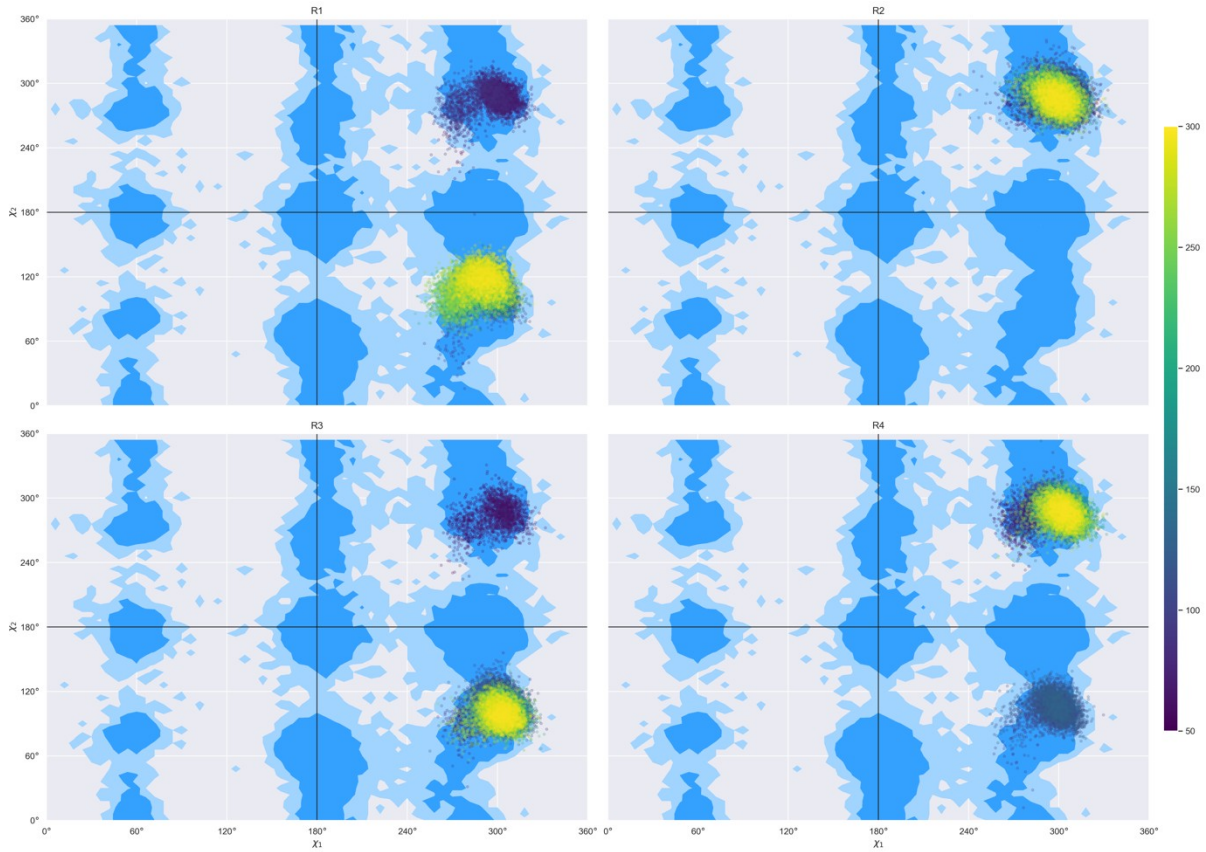


Figure S72. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

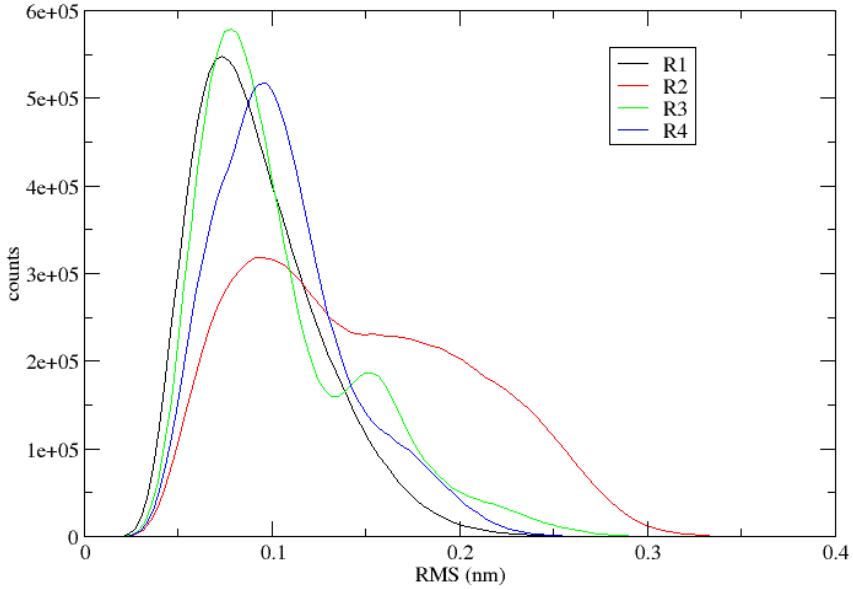


Figure S73. RMSD distribution of ciprofibrate ligand in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.125nm was selected to cluster the respective trajectories into representative groups (see Table S8).

Table S8. Overview of clusters from simulations of ciprofibrate and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

CIPROFIBRATE (1 molecule)			
Pocket	Replica	# clusters	Percentage (%)
A	R1	1	(1) **98.2**
R	R2	4	(1) **71.5** (2) **22.7** (3) **2.0** (4) **1.0**
M	R3	4	(1) **91.6** (2) **3.8** (3) **3.1** (4) **1.0**
I	R4	2	(1) **93.8** (2) **5.3**

8.2.2 Ciprofibrate bound to Arm I and Arm X

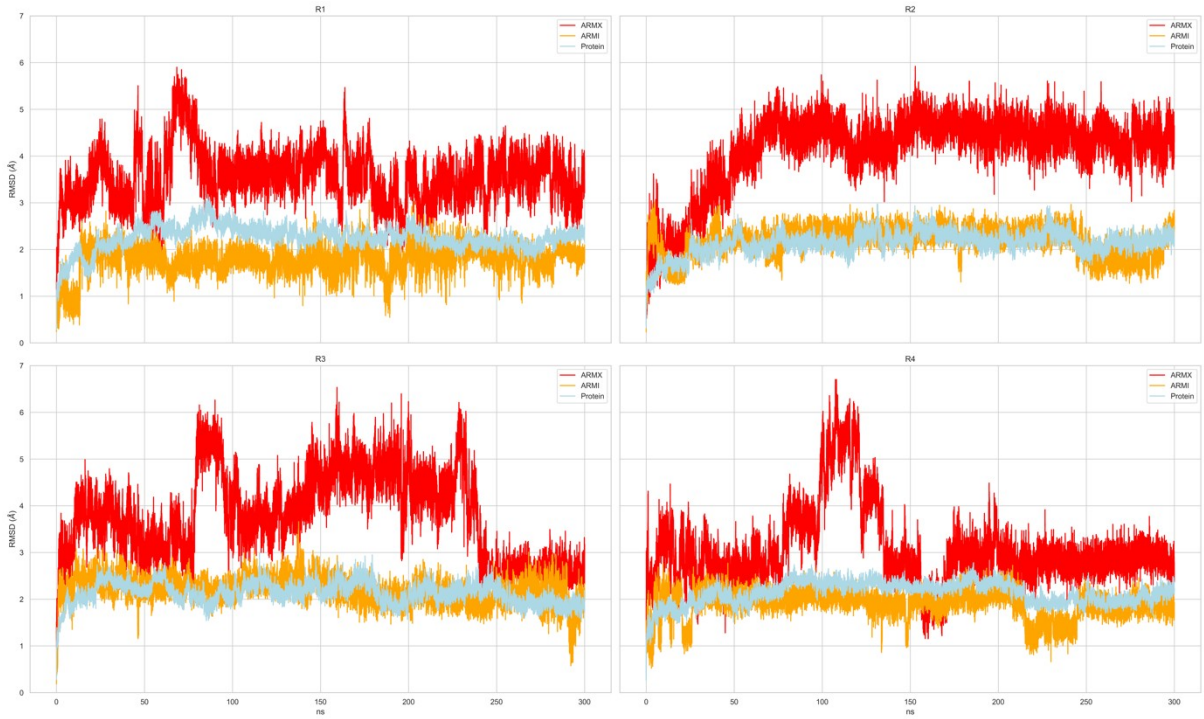


Figure S74. RMSD calculation of the protein backbone (light blue), ciprofibrate bound to Arm I (orange) and ciprofibrate bound to Arm X (red). The starting structure of the MD simulations is the Glide top docking pose (R1), the Haddock top docking pose (R2) and the crystal structure with PDB code 6KB3 (R3 and R4, each with different initial velocities). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

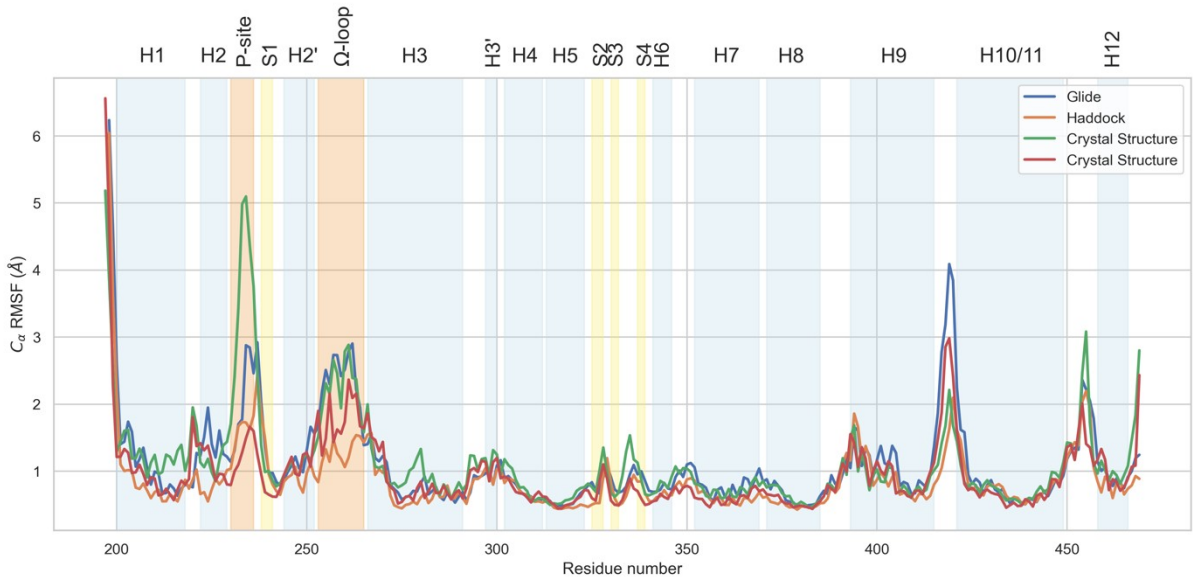


Figure S75. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

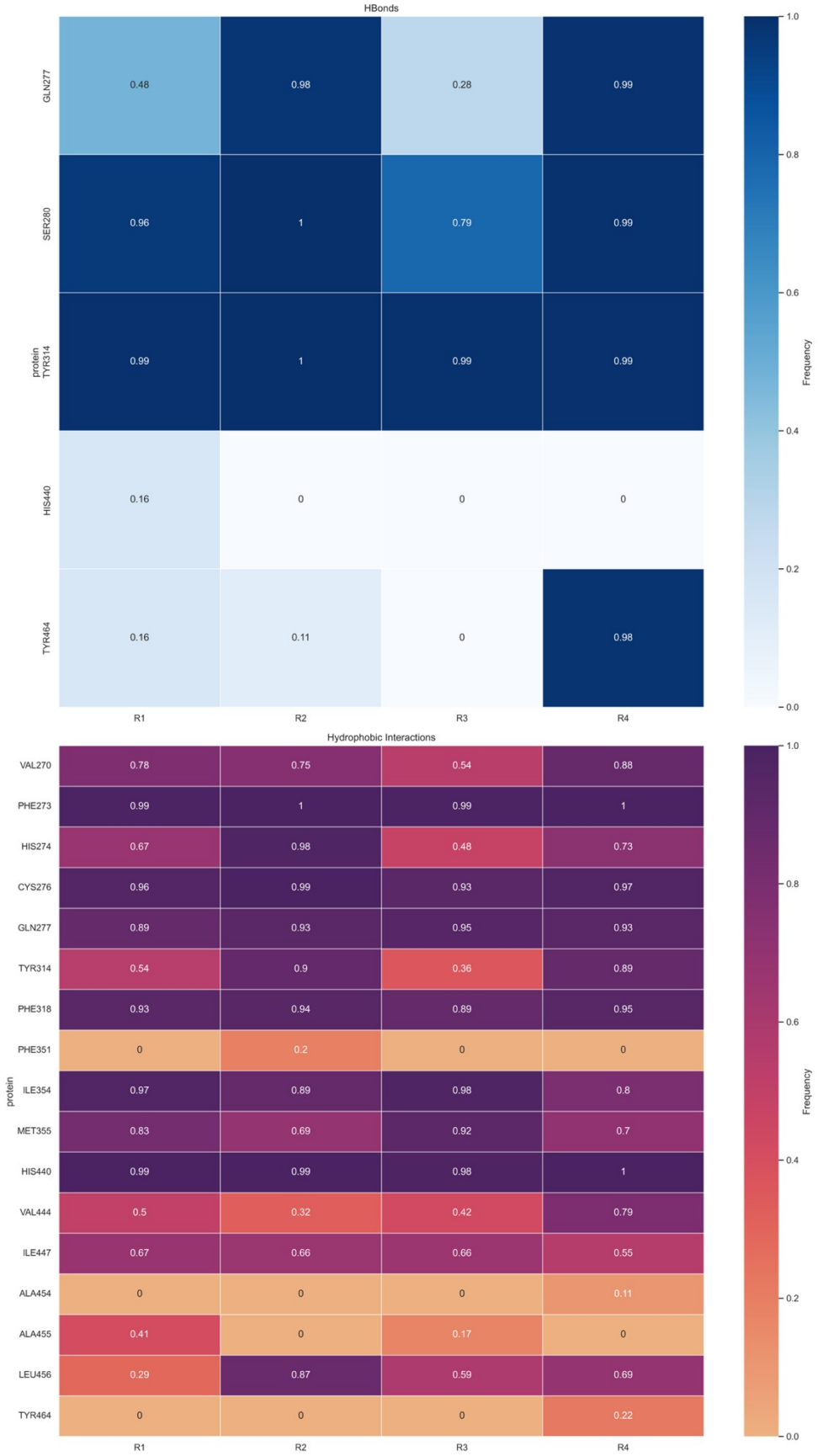


Figure S76. Heatmap of ciprofibrate-protein interactions (ciprofibrate bound to Arm I).

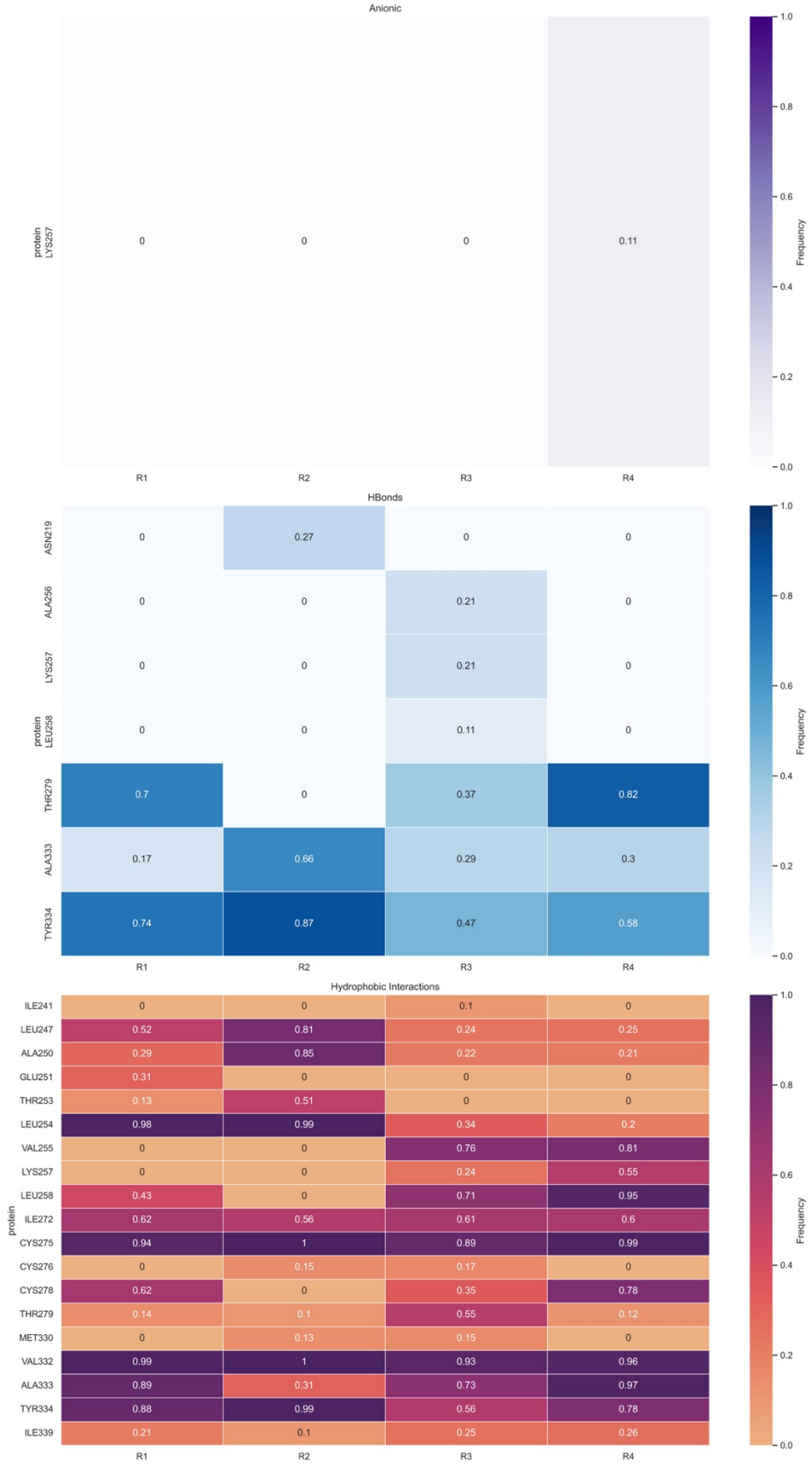


Figure S77. Heatmap of ciprofibrate-protein interactions (ciprofibrate bound to Arm X).

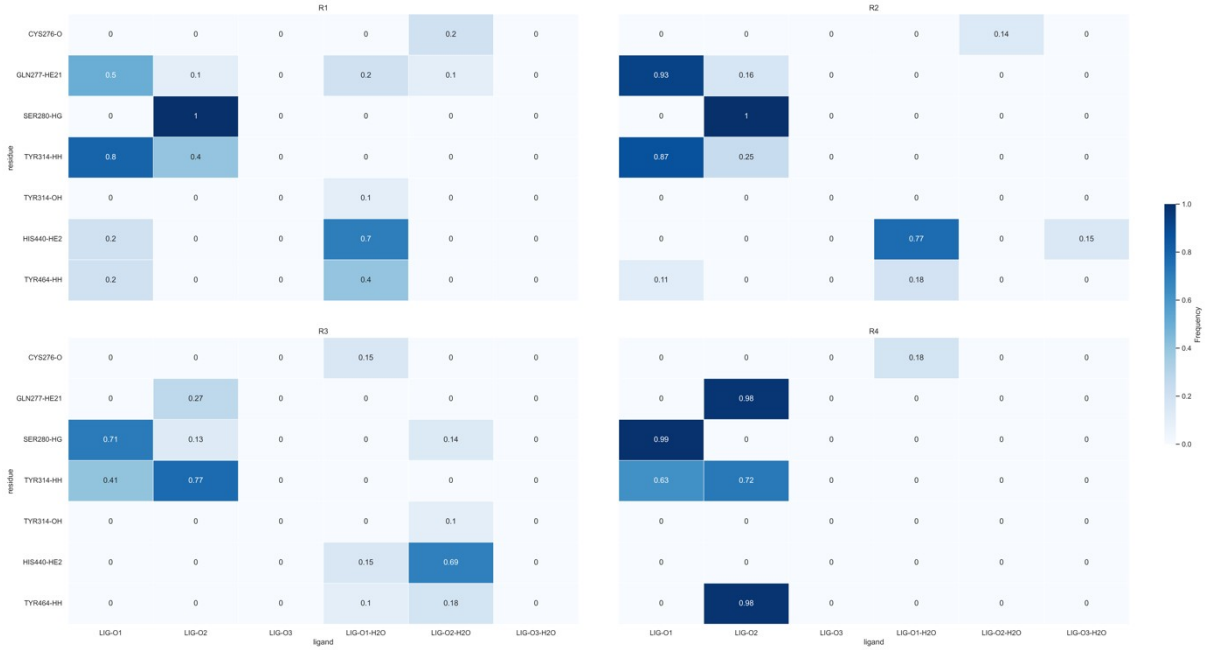


Figure S78. Detailed overview of H-bonds formed between ciprofibrate (bound to Arm I) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

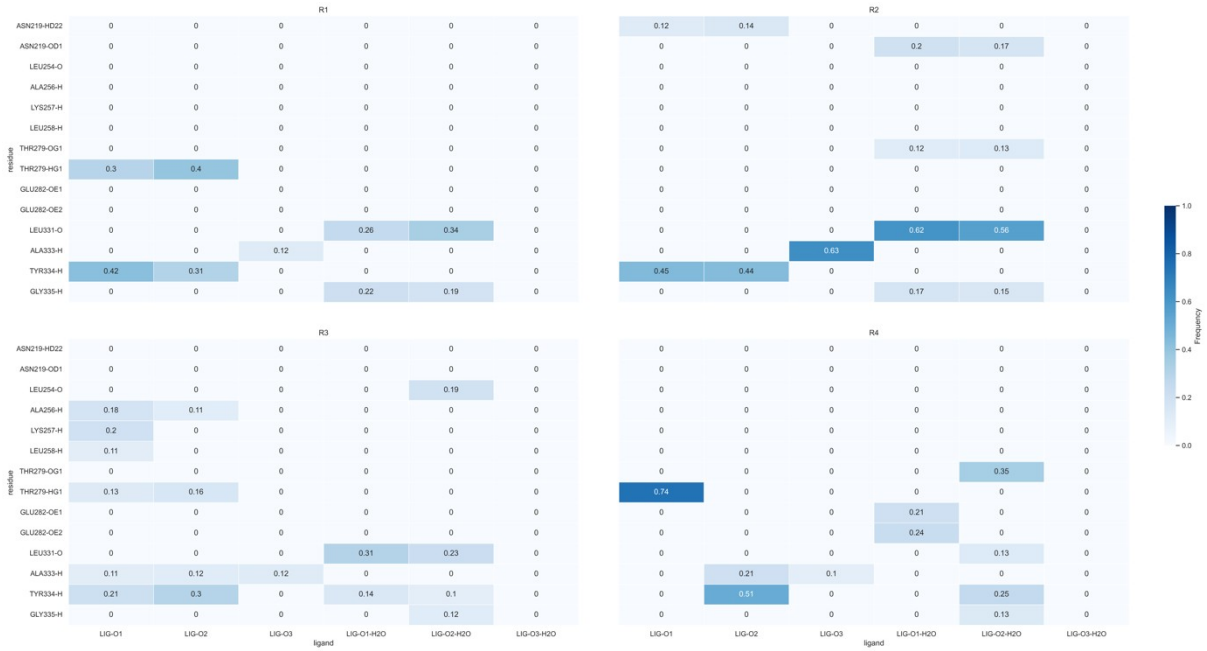


Figure S79. Detailed overview of H-bonds formed between ciprofibrate (bound to Arm X) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

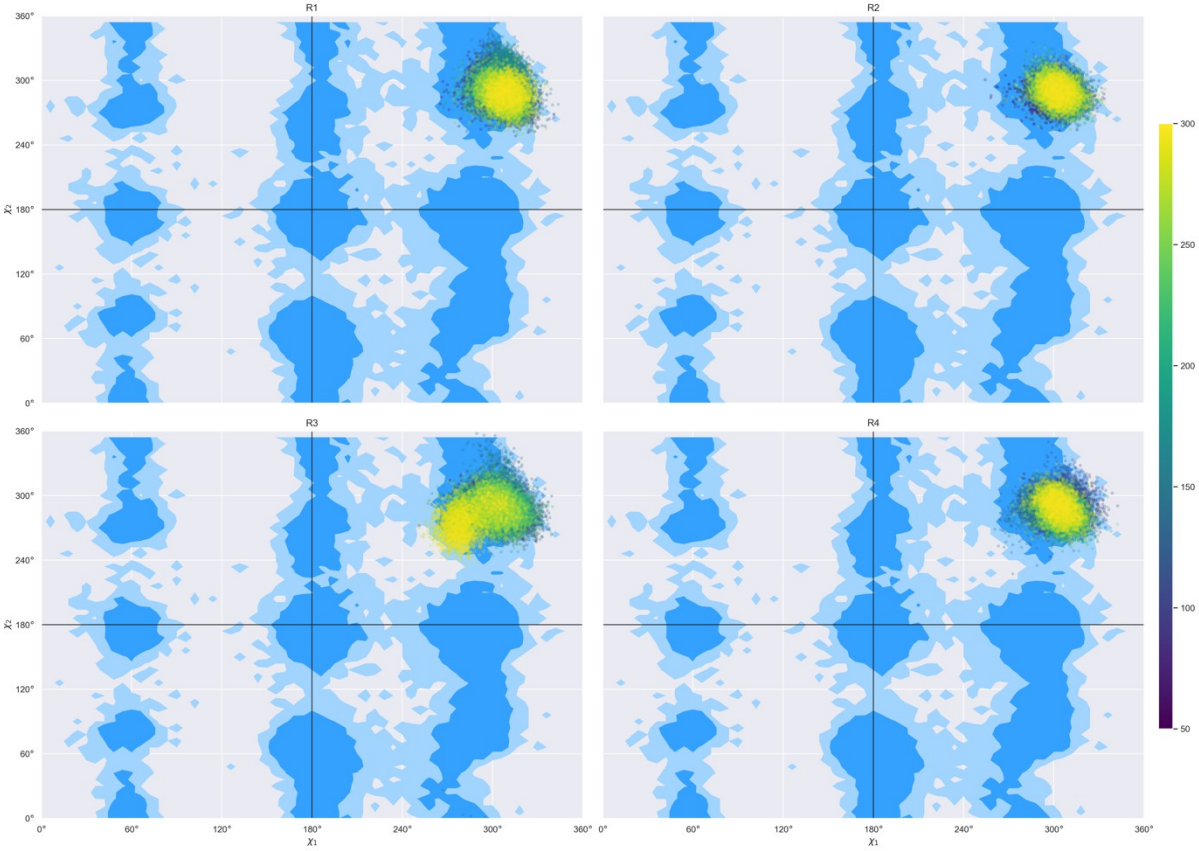


Figure S80. Time evolution of the Chi$_1$ and Chi$_2$ angles of F273. A data point represents the Chi$_1$ - Chi$_2$ angle combination in one particular MD frame. The colour coding indicates simulation time.

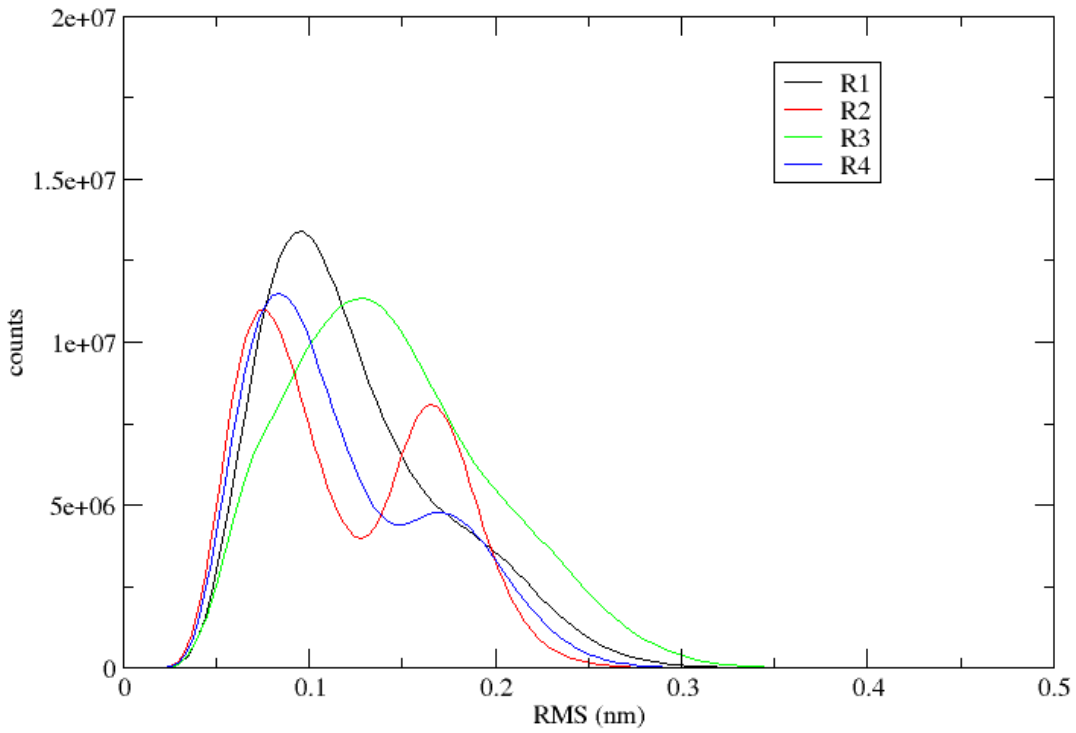


Figure S81. RMSD distribution of ciprofibrate (bound to Arm I) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.125nm was selected to cluster the respective trajectories into representative groups (see Table S9).

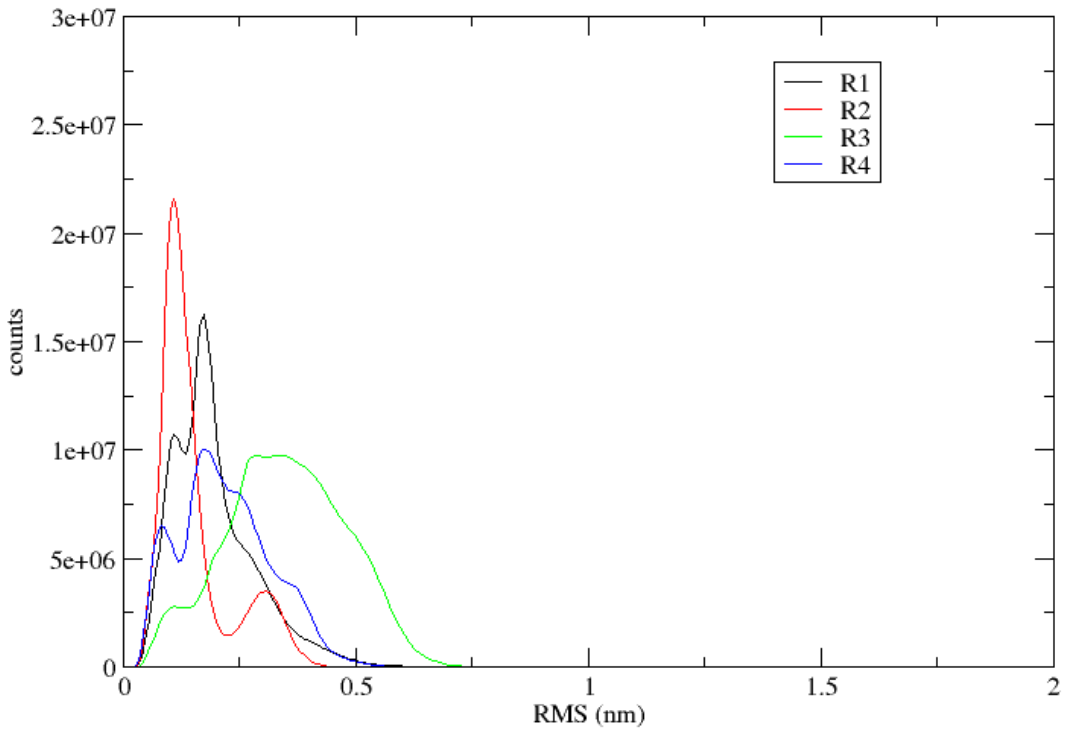


Figure S82. RMSD distribution of ciprofibrate (bound to Arm X) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.14 nm was selected to cluster the respective trajectories into representative groups (see Table S9).

Table S9. Overview of clusters from simulations of ciprofibrate and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

CIPROFIBRATE (2 molecules)			
Pocket	Replica	# clusters	Percentage (%)
ARMI	R1	4	(1) **85.9** (2) **6.5** (3) **4.3** (4) **1.7**
	R2	3	(1) **82.7** (2) **8.4** (3) **8.0**
	R3	5	(1) **70.7** (2) **17.4** (3) **4.2** (4) **3.4** (5) **1.5**
	R4	3	(1) **85.3** (2) **12.3** (3) **1.1**
ARMX	R1	10	(1) **62.3** (2) **13.7** (3) **3.7** (4) **3.6** (5) **2.4** (6) **1.6** (7) **1.4** (8) **1.4** (9) **1.2** (10) **1.0**
	R2	3	(1) **85.8** (2) **9.7** (3) **1.9**
	R3	16	(1) **15.3** (2) **12.9** (3) **9.4** (4) **8.5** (5) **8.3** (6) **6.4** (7) **5.0** (8) **4.5** (9) **2.9** (10) **1.9** (11) **1.7** (12) **1.4** (13) **1.3** (14) **1.1** (15) **1.1** (16) **1.0**
	R4	11	(1) **53.8** (2) **11.3** (3) **5.3** (4) **4.2** (5) **4.0** (6) **3.5** (7) **3.5** (8) **2.4** (9) **2.0** (10) **1.2** (11) **1.0**

8.3 Gemfibrozil

8.3.1 Gemfibrozil bound to the Arm I pocket

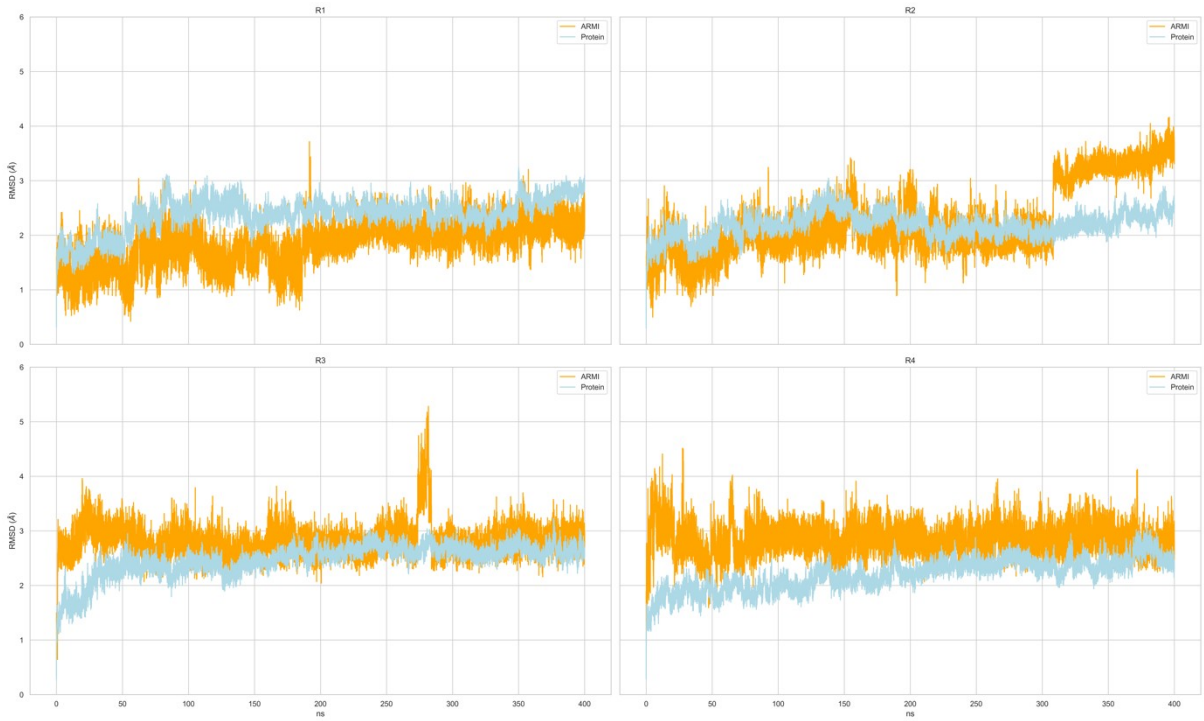


Figure S83. RMSD calculation of the protein backbone (light blue) and gemfibrozil bound to Arm I (orange). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

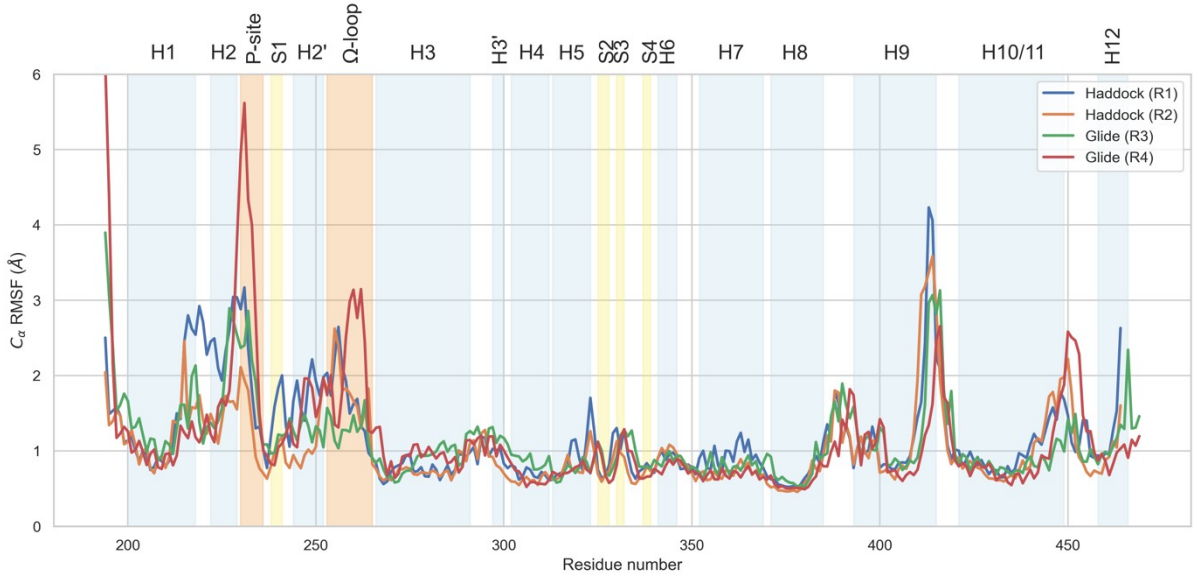


Figure S84. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

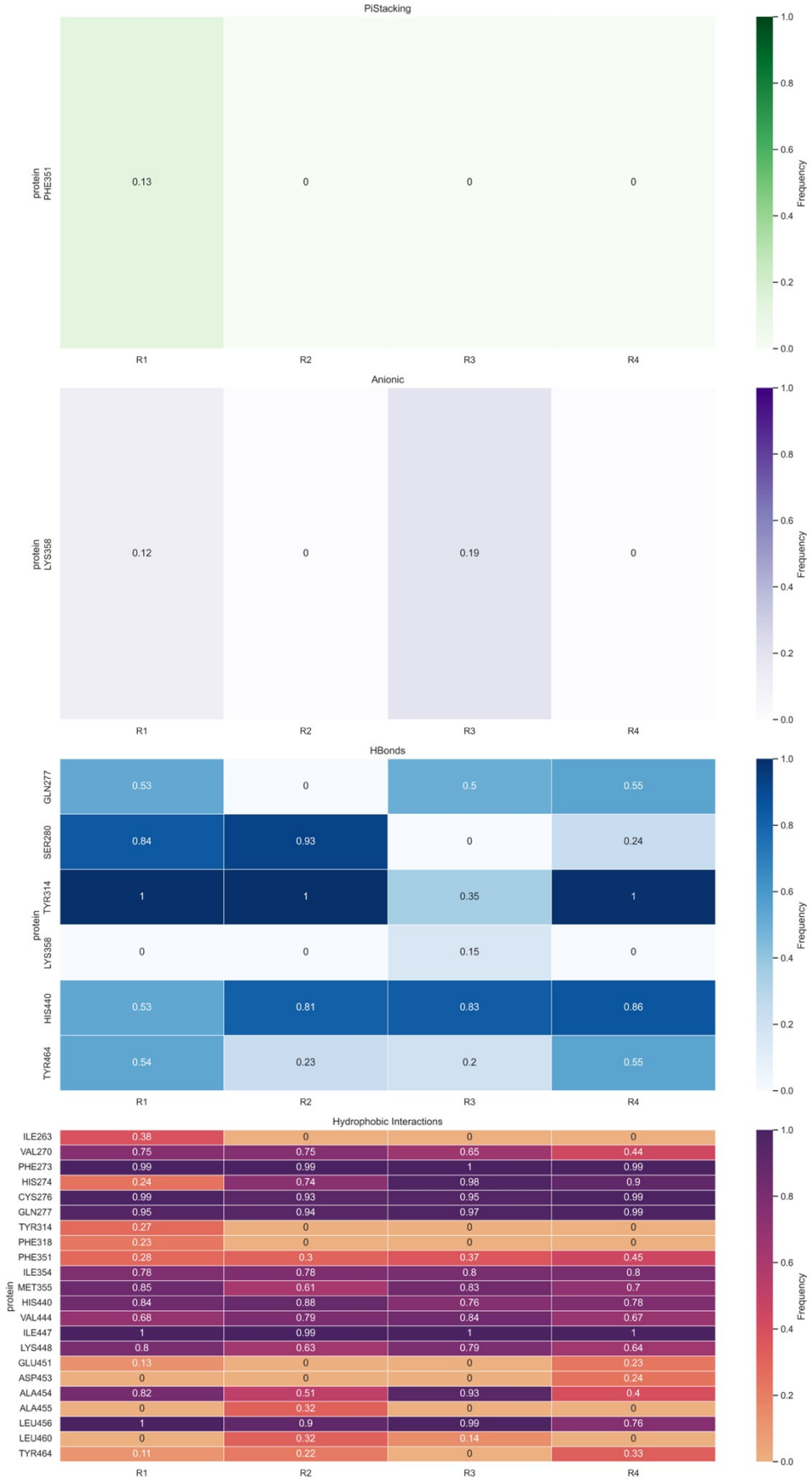


Figure S85. Heatmap of gemfibrozil-protein interactions (gemfibrozil bound to Arm I).

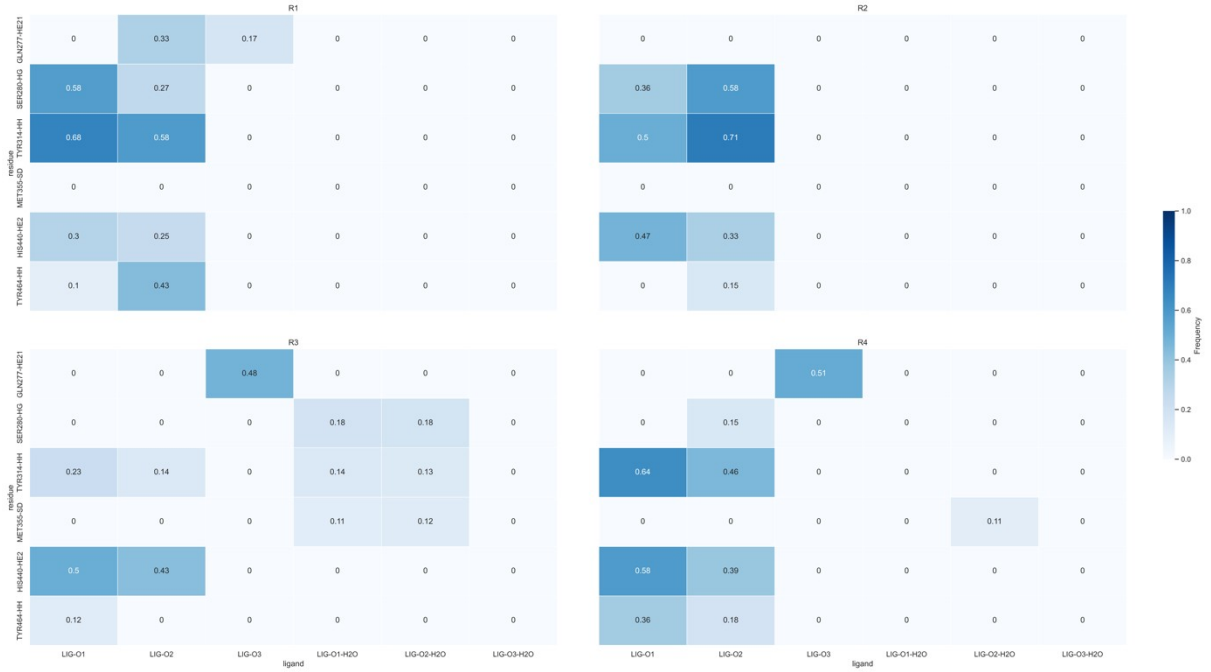


Figure S86. Detailed overview of H-bonds formed between gemfibrozil (bound to Arm I) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

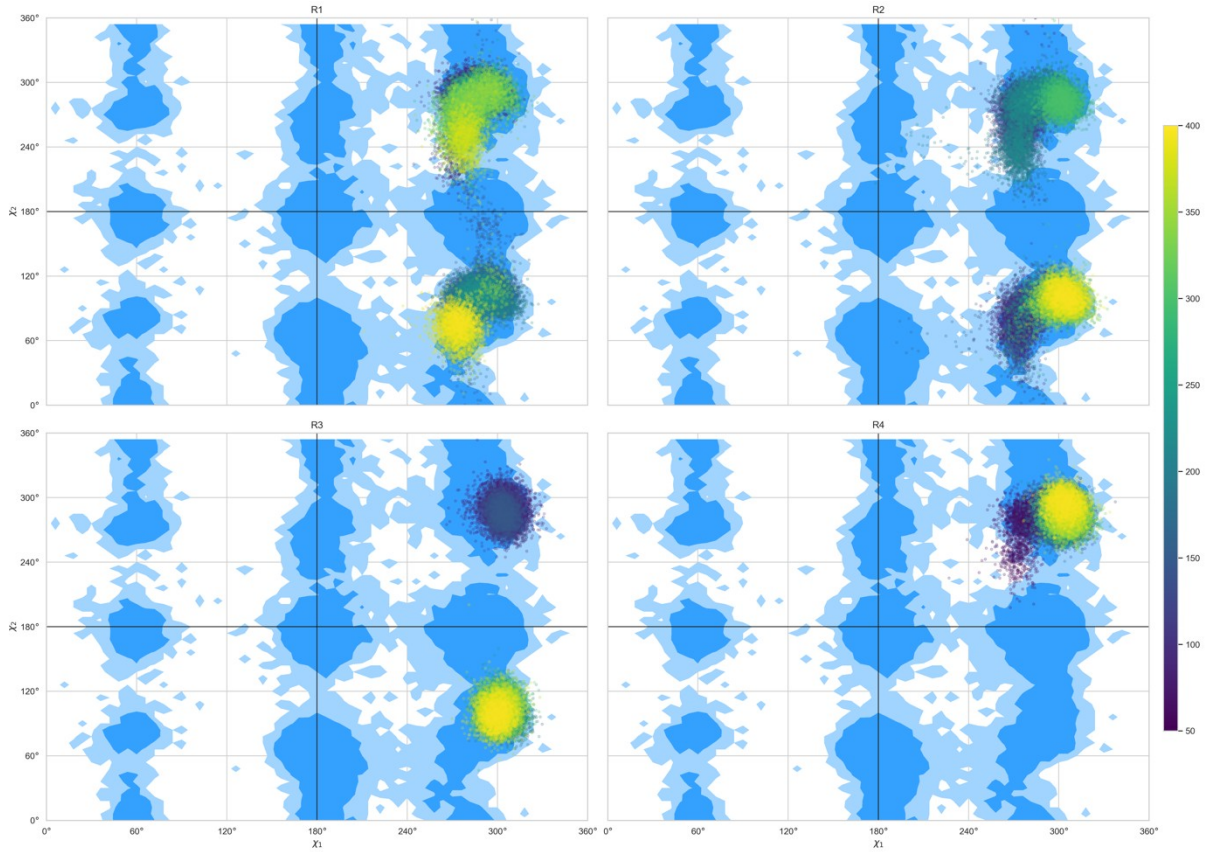


Figure S87. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

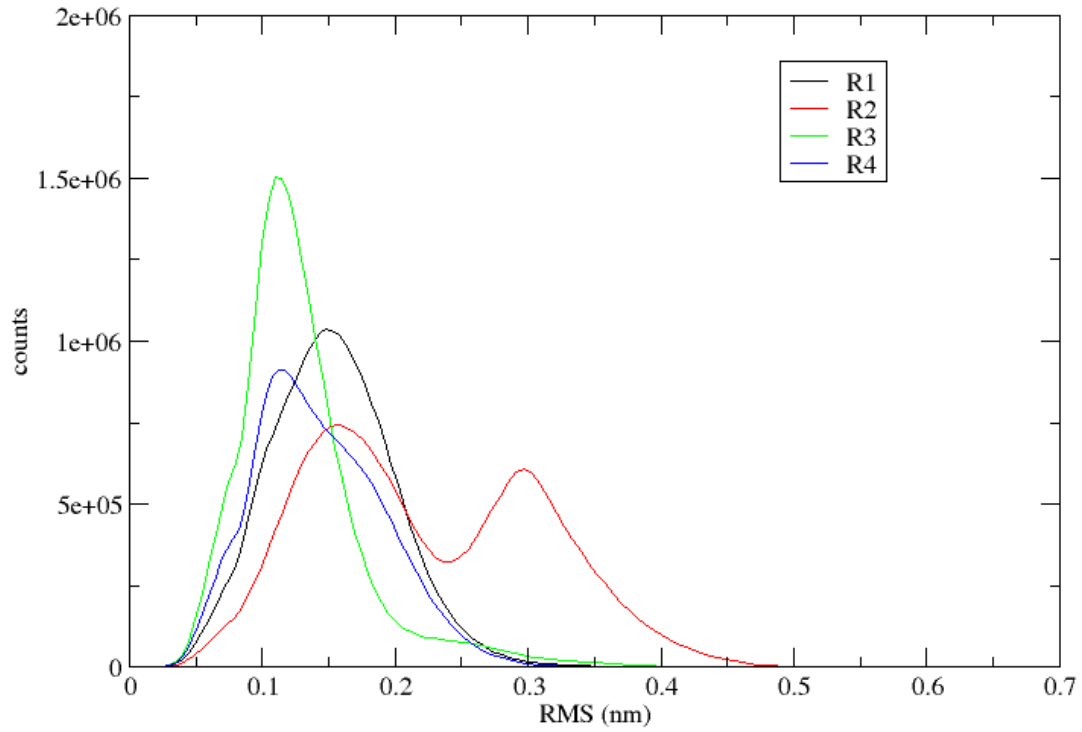


Figure S88. RMSD distribution of gemfibrozil (bound to Arm I) in the four replica MD simulations. Here, frames from 50-400ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S10).

Table S10. Overview of clusters from simulations of gemfibrozil and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

GEMFIBROZIL (1 molecule)			
Pocket	Replica	# clusters	Percentage (%)
A	R1	1	(1) **99.1**
R	R2	3	(1) **71.4** (2) **25.5** (3) **1.3**
M	R3	2	(1) **98.1** (2) **1.7**
I	R4	1	(1) **99.5**

8.3.2 Gemfibrozil bound to the Center/Arm II/Arm III pocket

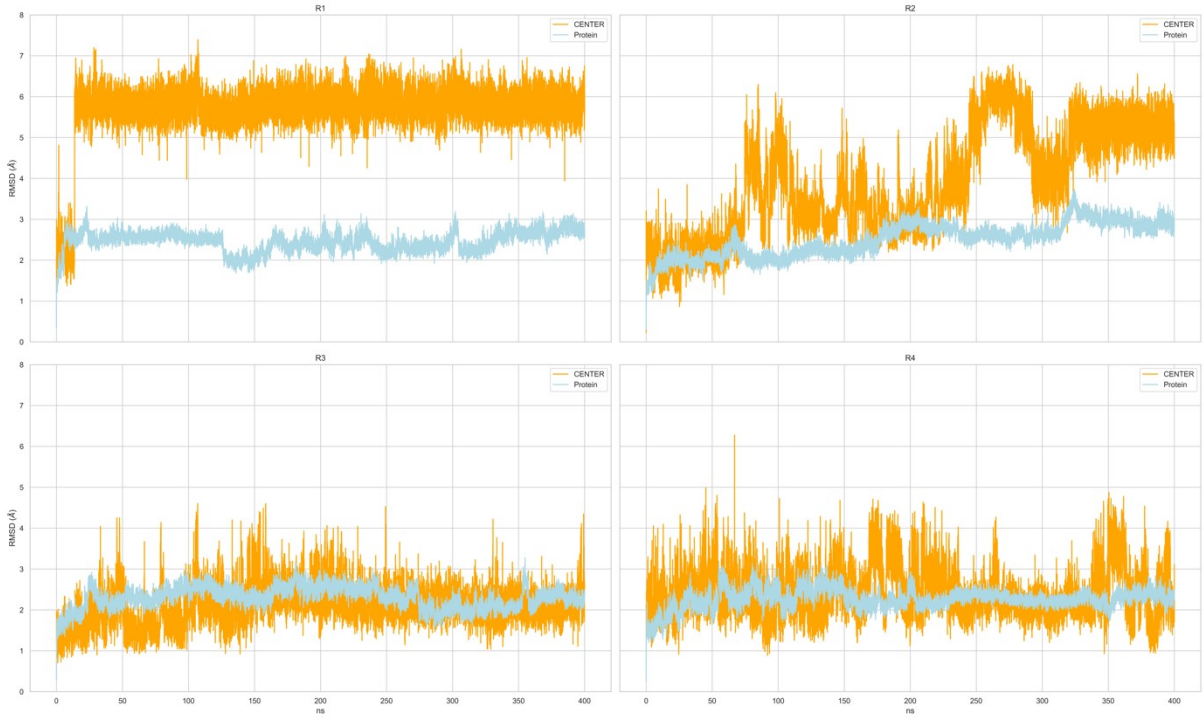


Figure S89. RMSD calculation of the protein backbone (light blue) and gemfibrozil bound to the Center/Arm II/Arm III pocket (orange). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

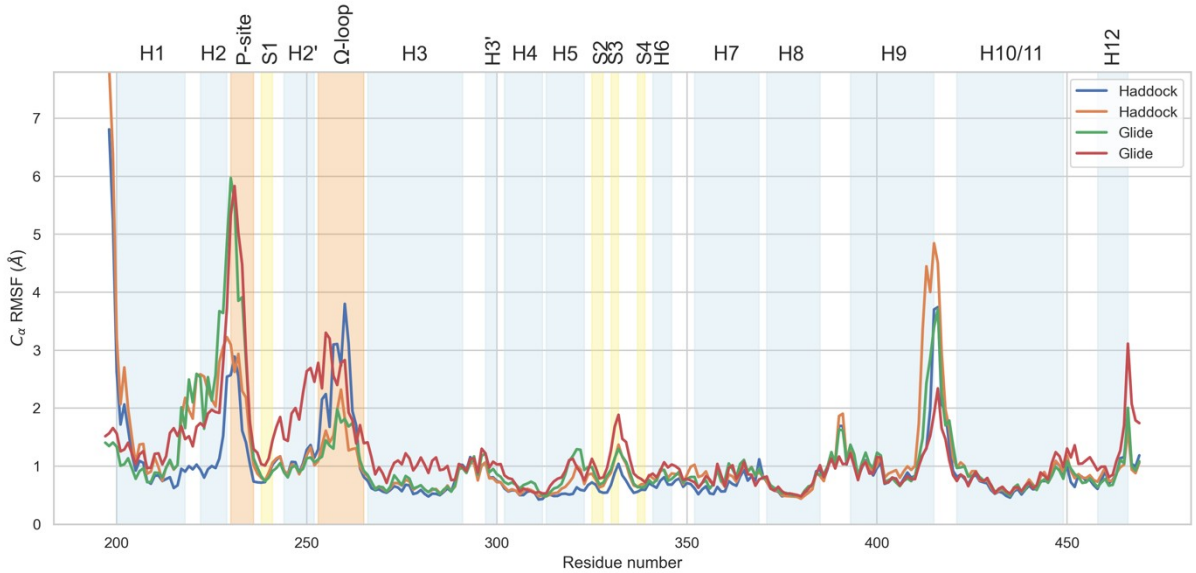


Figure S90. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

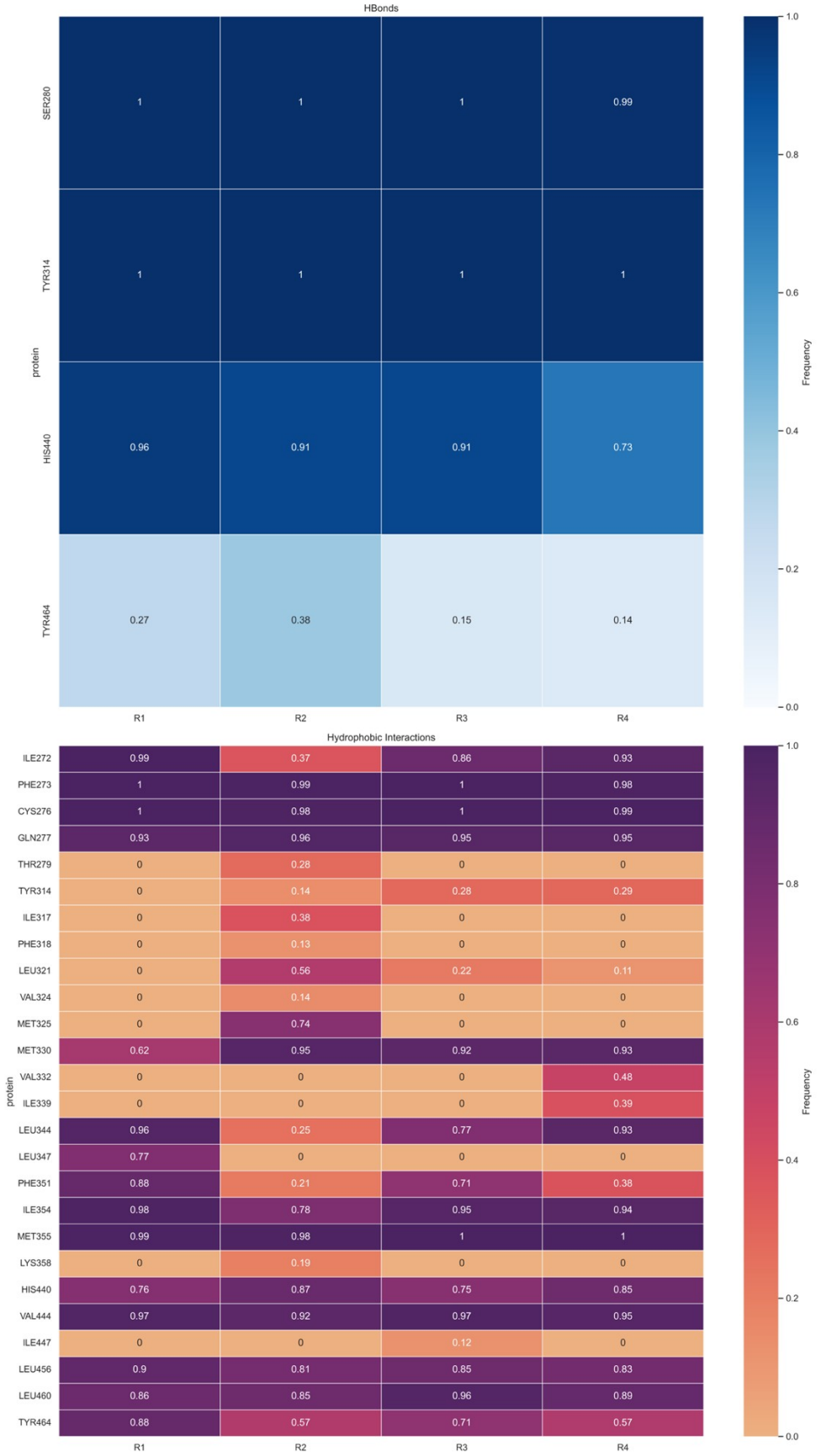


Figure S91. Heatmap of gemfibrozil-protein interactions (gemfibrozil bound to the Center/Arm II/Arm III pocket).

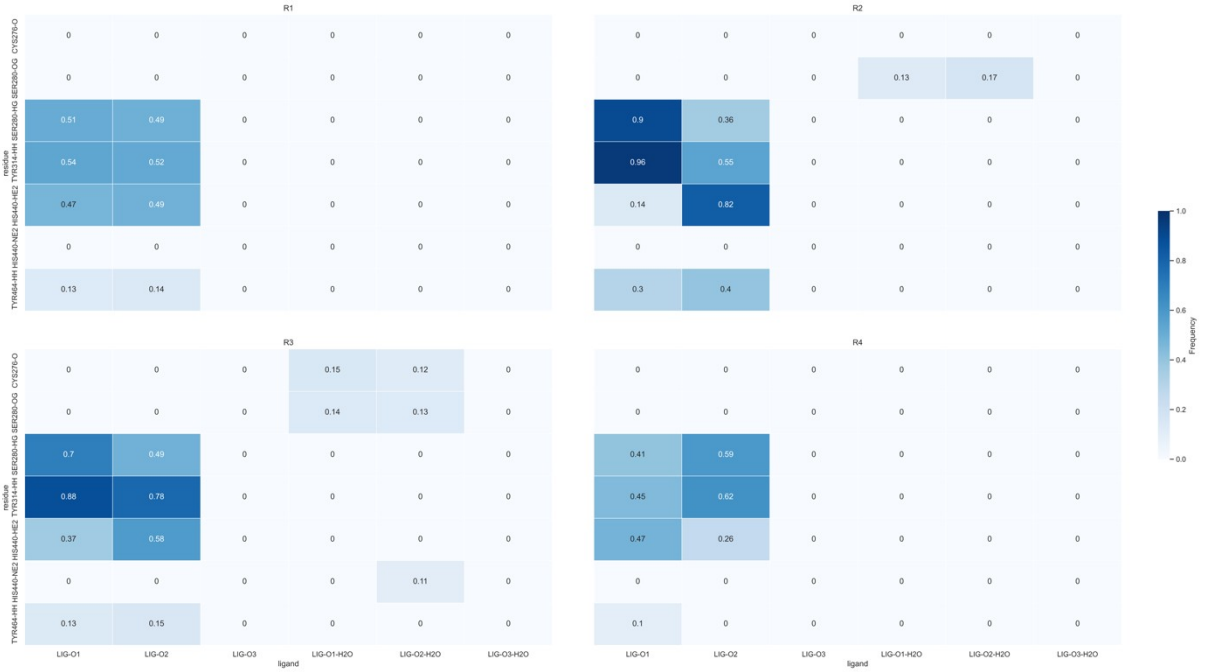


Figure S92. Detailed overview of H-bonds formed between gemfibrozil (bound to the Center/Arm II/Arm III pocket) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

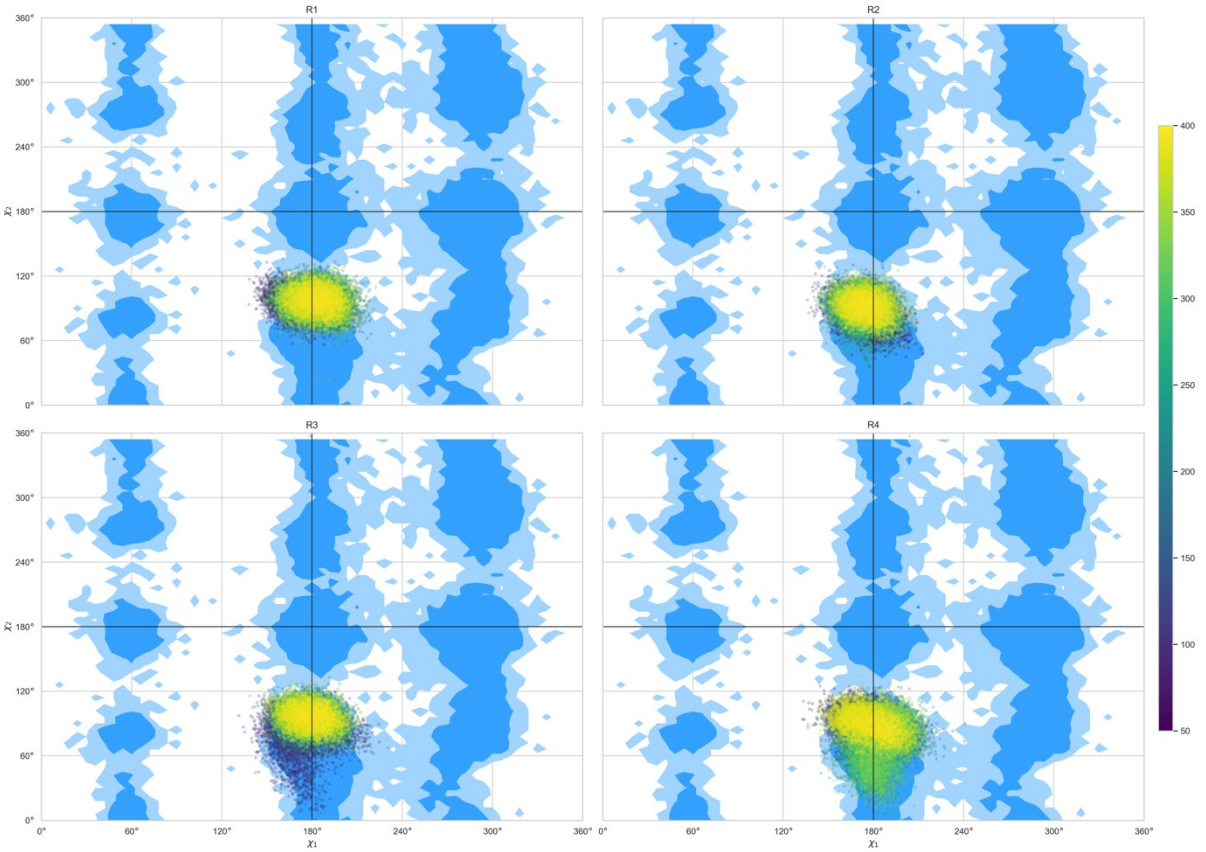


Figure S93. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

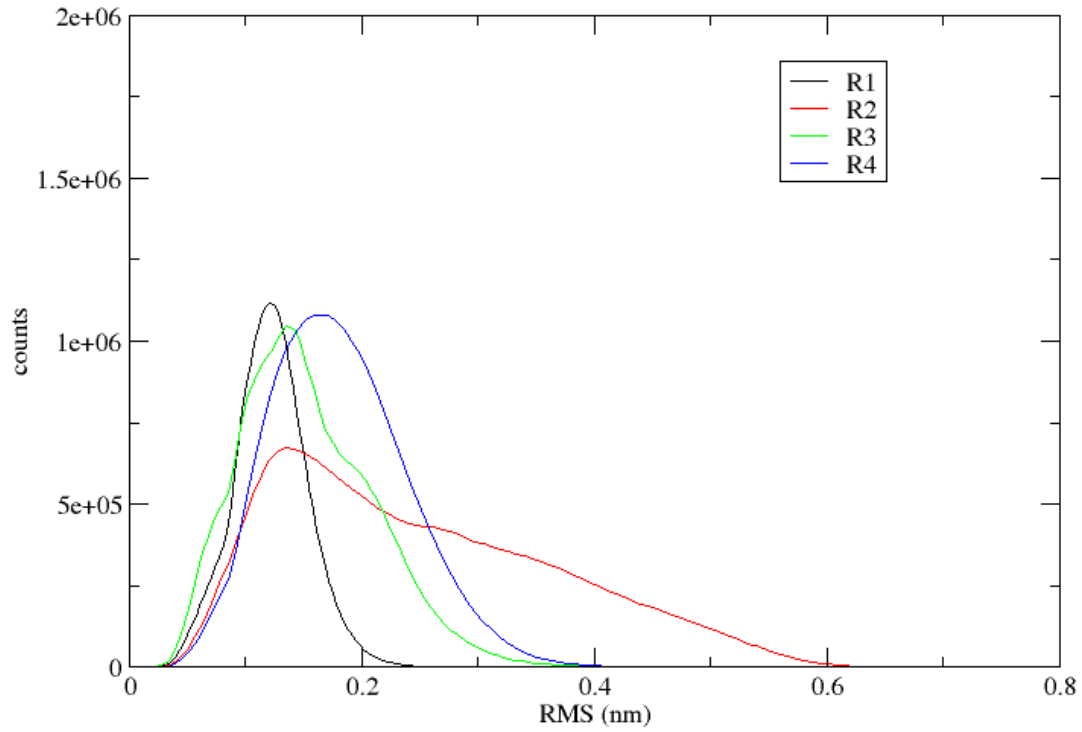


Figure S94. RMSD distribution of gemfibrozil (bound to the Center/Arm II/Arm III pocket) in the four replica MD simulations. Here, frames from 50-400ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.15 nm was selected to cluster the respective trajectories into representative groups (see Table S11).

Table S11. Overview of clusters from simulations of gemfibrozil (bound to the Center/Arm II/Arm III pocket) and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

GEMFIBROZIL (1 molecule)			
Pocket	Replica	# clusters	Percentage (%)
CENTER	R1	2	(1) **98.1** (2) **1.0**
	R2	6	(1) **44.4** (2) **32.5** (3) **7.7** (4) **6.8** (5) **4.9** (6) **1.5**
	R3	4	(1) **79.7** (2) **13.8** (3) **2.1** (4) 2.0
	R4	6	(1) **65.0** (2) **20.1** (3) **4.8** (4) **3.4** (5) **2.4** (6) **1.3**

8.3.3 Gemfibrozil bound to the Arm I and Arm X

Figure S95. RMSD calculation of the protein backbone (light blue), gemfibrozil bound to Arm I (orange) and gemfibrozil bound to Arm X (red). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

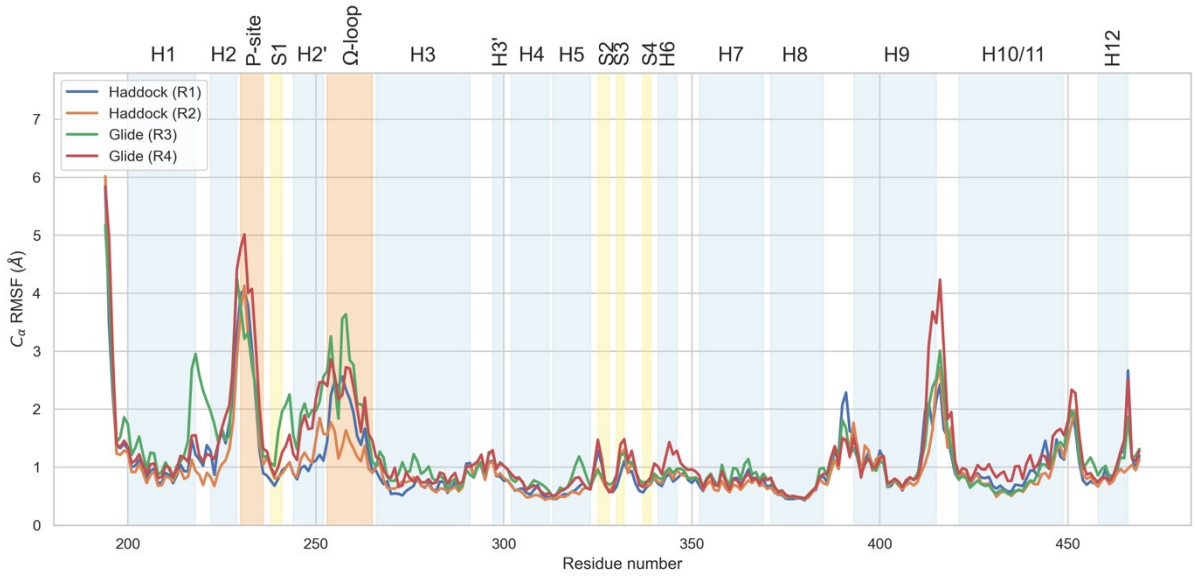


Figure S96. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

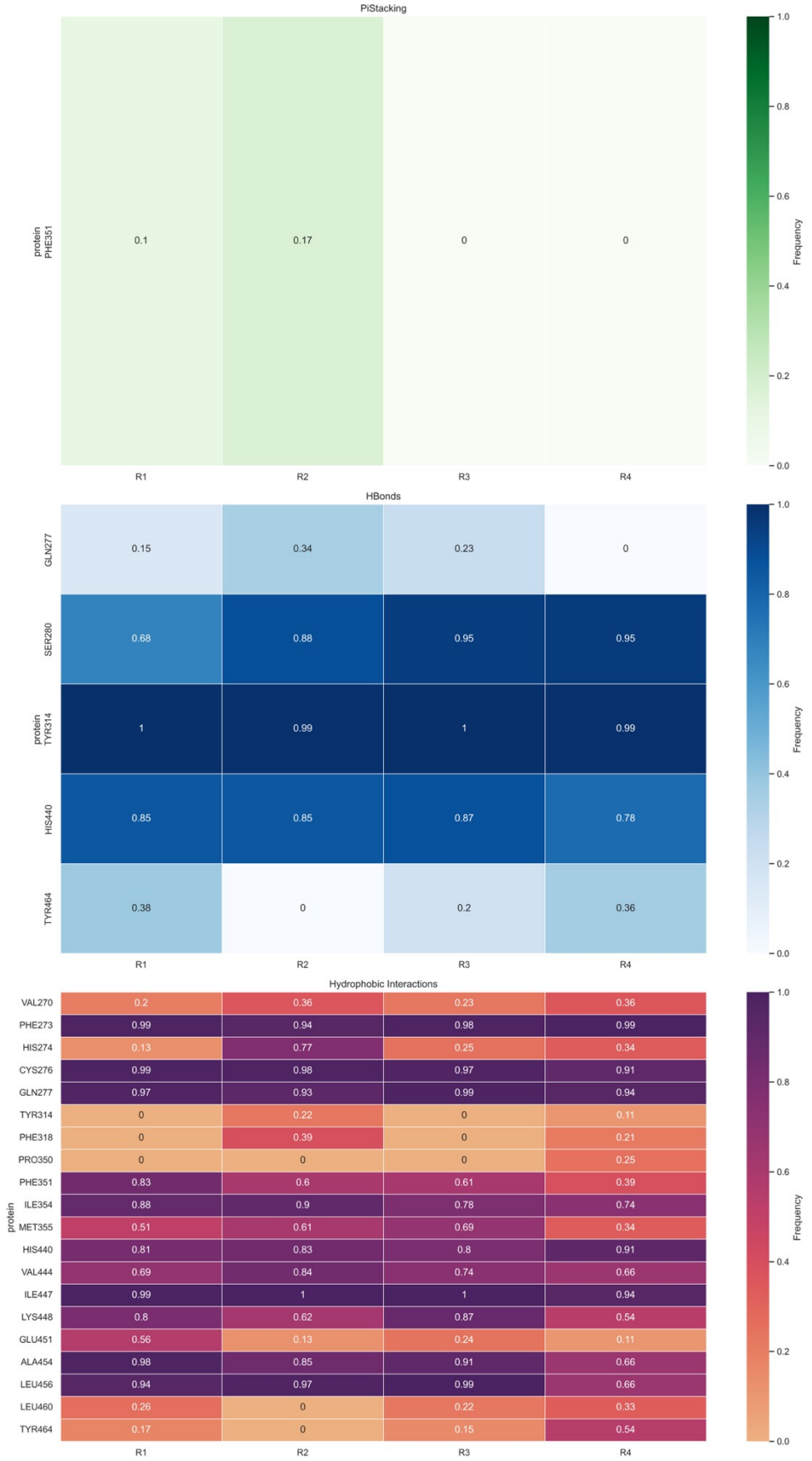


Figure S97. Heatmap of gemfibrozil-protein interactions (gemfibrozil bound to Arm I).

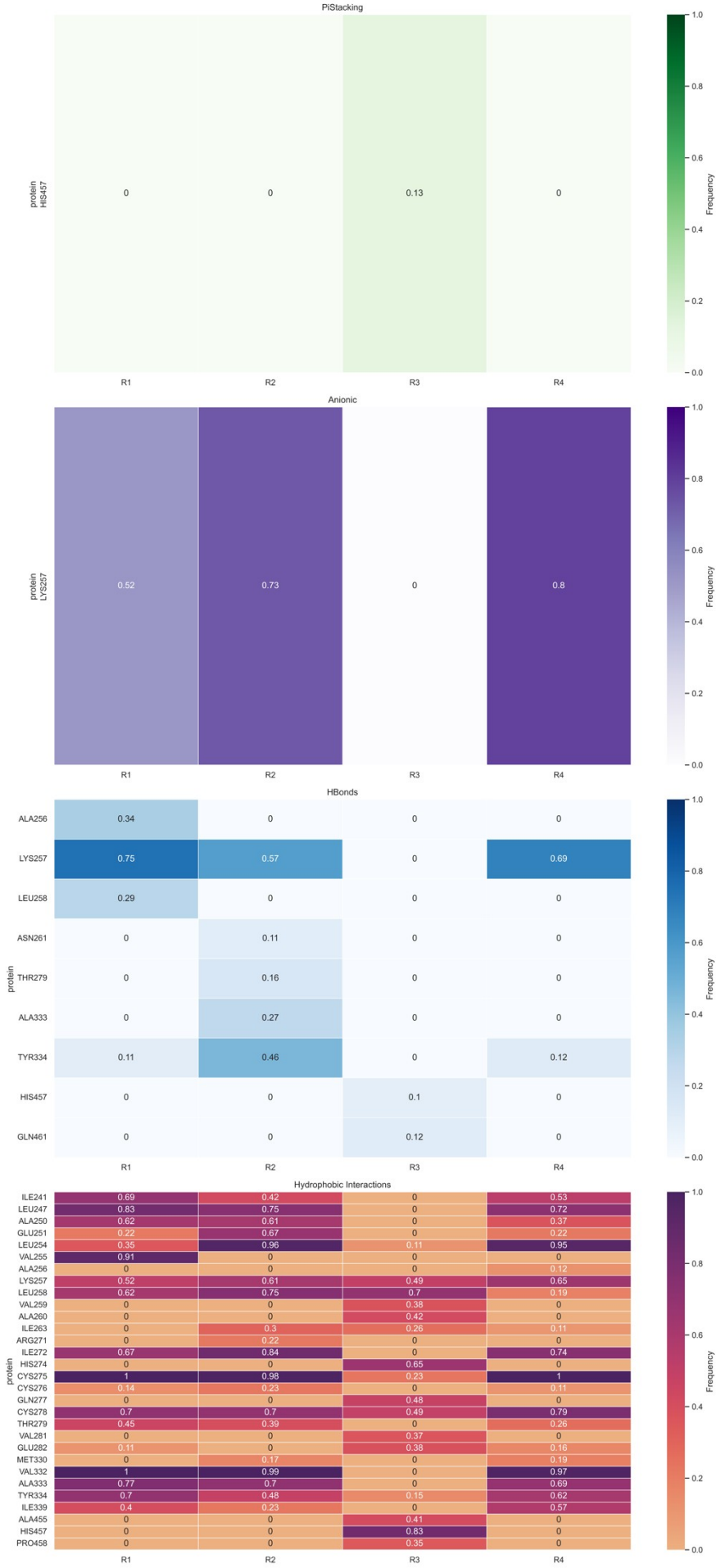

PiStacking
protein HIS457
0 0 0.13 0
R1 R2 R3 R4
Frequency
Anionic
protein LYS257
0.52 0.73 0 0.8
R1 R2 R3 R4
Frequency
HBonds
protein
ALA256 0.34 0 0 0
LYS257 0.75 0.57 0 0.69
LEU258 0.29 0 0 0
ASN261 0 0.11 0 0
THR279 0 0.16 0 0
ALA333 0 0.27 0 0
TYR334 0.11 0.46 0 0.12
HIS457 0 0 0.1 0
GLN461 0 0 0.12 0
R1 R2 R3 R4
Frequency
Hydrophobic Interactions
protein
ILE241 0.69 0.42 0 0.53
LEU247 0.83 0.75 0 0.72
ALA250 0.62 0.61 0 0.37
GLU251 0.22 0.67 0 0.22
LEU254 0.35 0.96 0.11 0.95
VAL255 0.91 0 0 0
ALA256 0 0 0 0.12
LYS257 0.52 0.61 0.49 0.65
LEU258 0.62 0.75 0.7 0.19
VAL259 0 0 0.38 0
ALA260 0 0 0.42 0
ILE263 0 0.3 0.26 0.11
ARG271 0 0.22 0 0
ILE272 0.67 0.84 0 0.74
HIS274 0 0 0.65 0
CYS275 1 0.98 0.23 1
CYS276 0.14 0.23 0 0.11
GLN277 0 0 0.48 0
CYS278 0.7 0.7 0.49 0.79
THR279 0.45 0.39 0 0.26
VAL281 0 0 0.37 0
GLU282 0.11 0 0.38 0.16
MET330 0 0.17 0 0.19
VAL332 1 0.99 0 0.97
ALA333 0.77 0.7 0 0.69
TYR334 0.7 0.48 0.15 0.62
ILE339 0.4 0.23 0 0.57
ALA455 0 0 0.41 0
HIS457 0 0 0.83 0
PRO458 0 0 0.35 0
R1 R2 R3 R4
Frequency

Figure S98. Heatmap of gemfibrozil-protein interactions (gemfibrozil bound to Arm X).

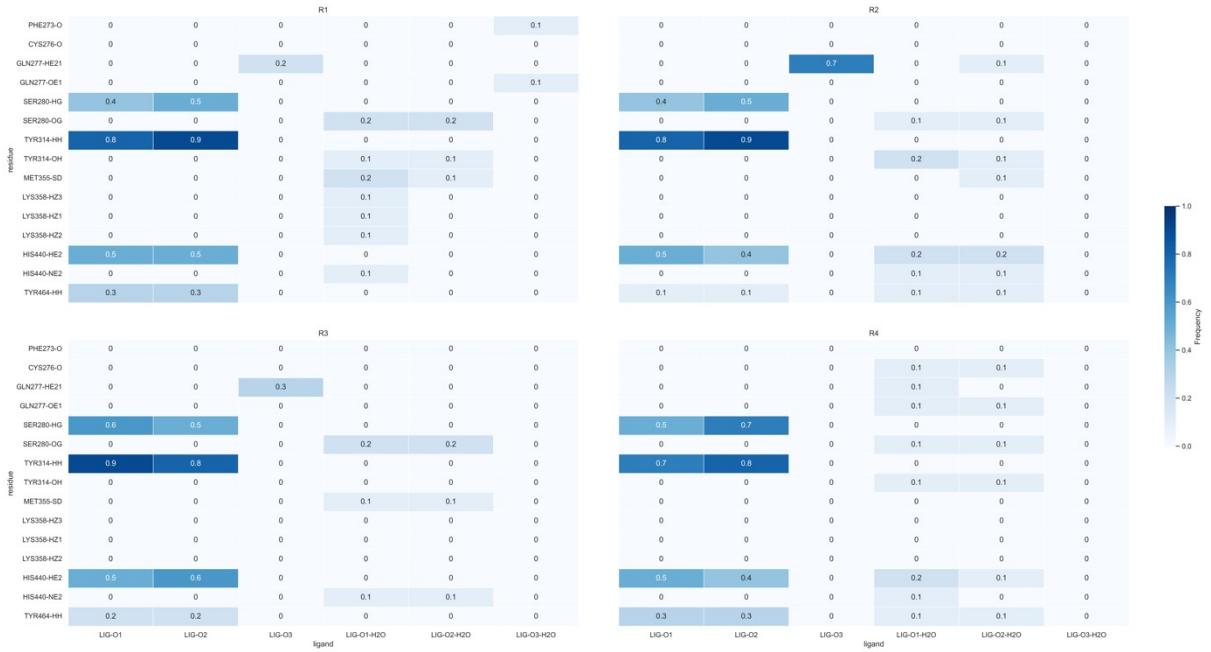


Figure S99. Detailed overview of H-bonds formed between gemfibrozil (bound to Arm I) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

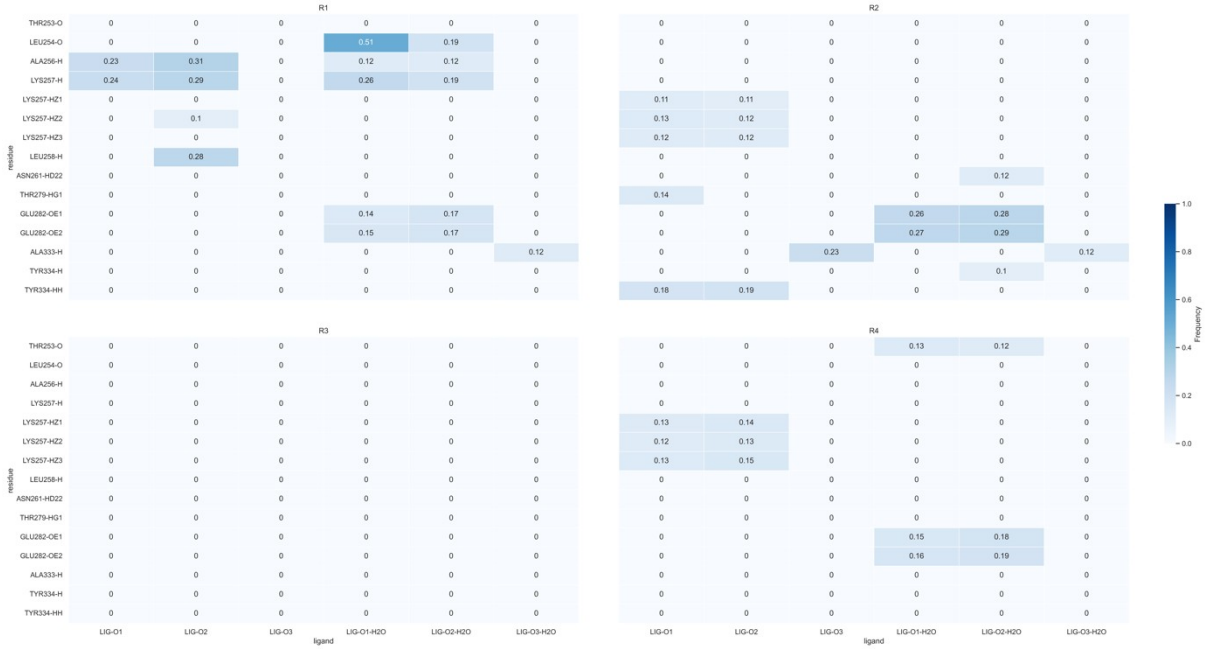


Figure S100. Detailed overview of H-bonds formed between gemfibrozil (bound to Arm X) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

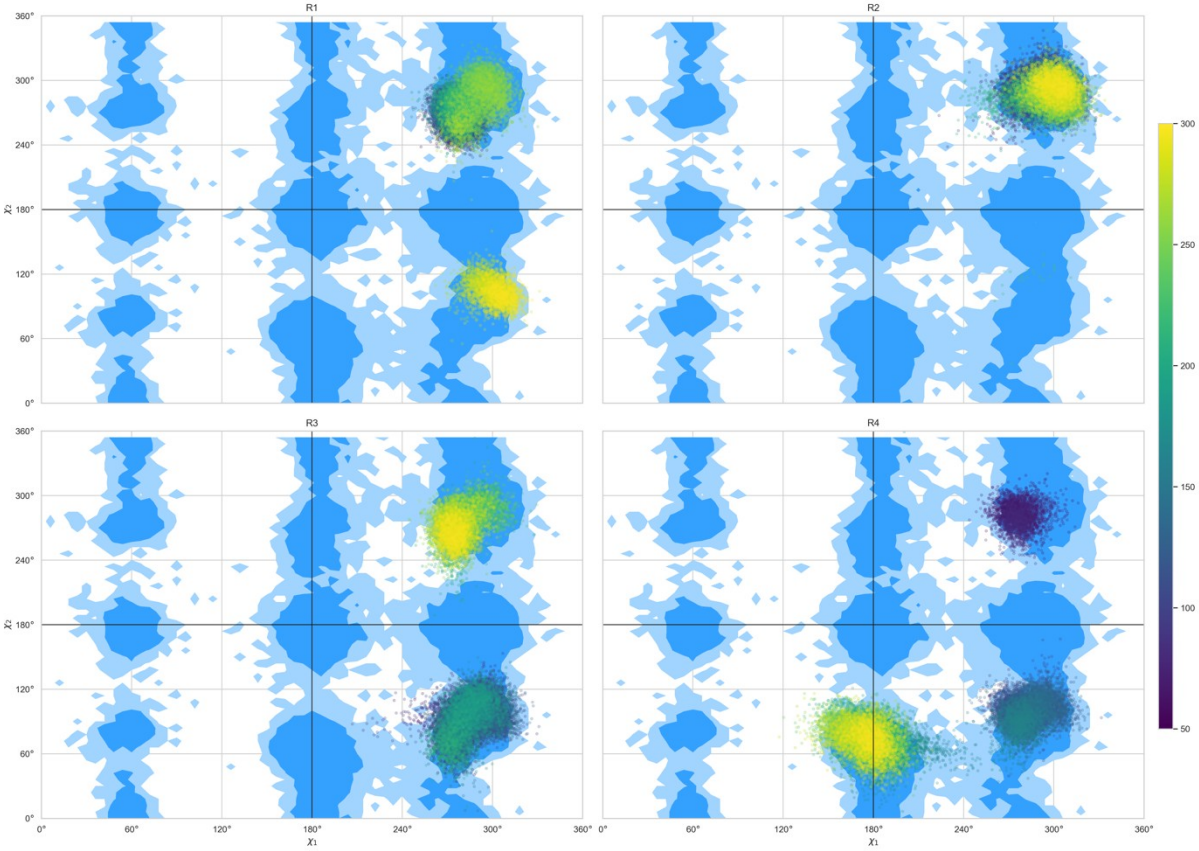


Figure S101. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

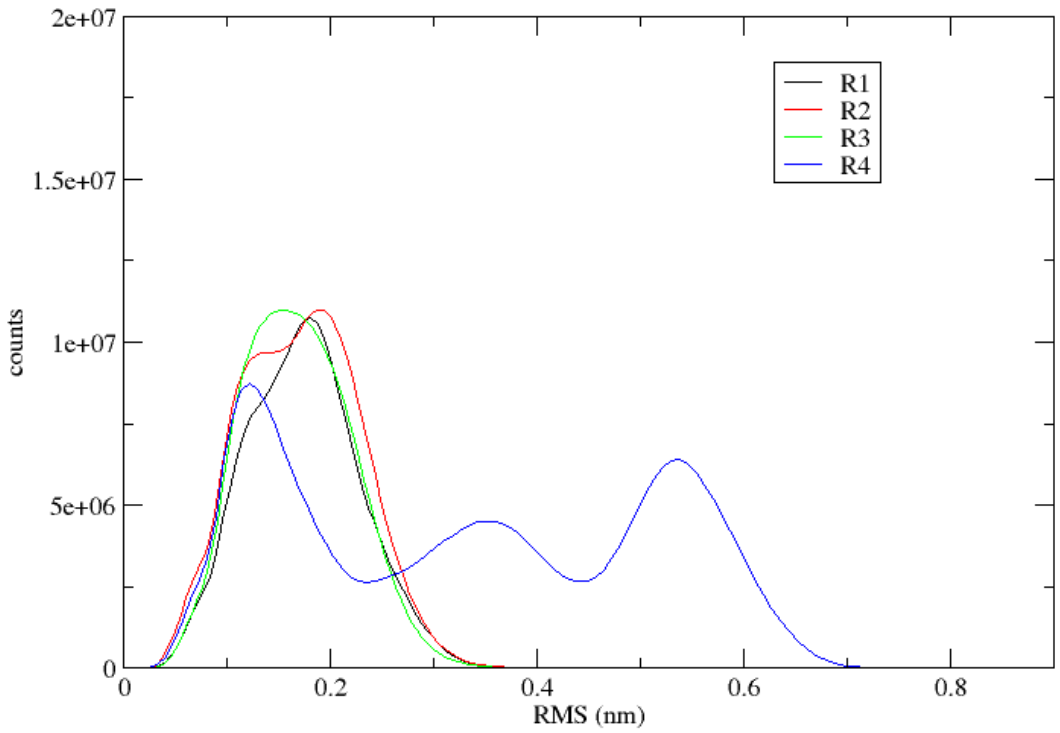


Figure S102. RMSD distribution of gemfibrozil (bound to Arm I) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S12).

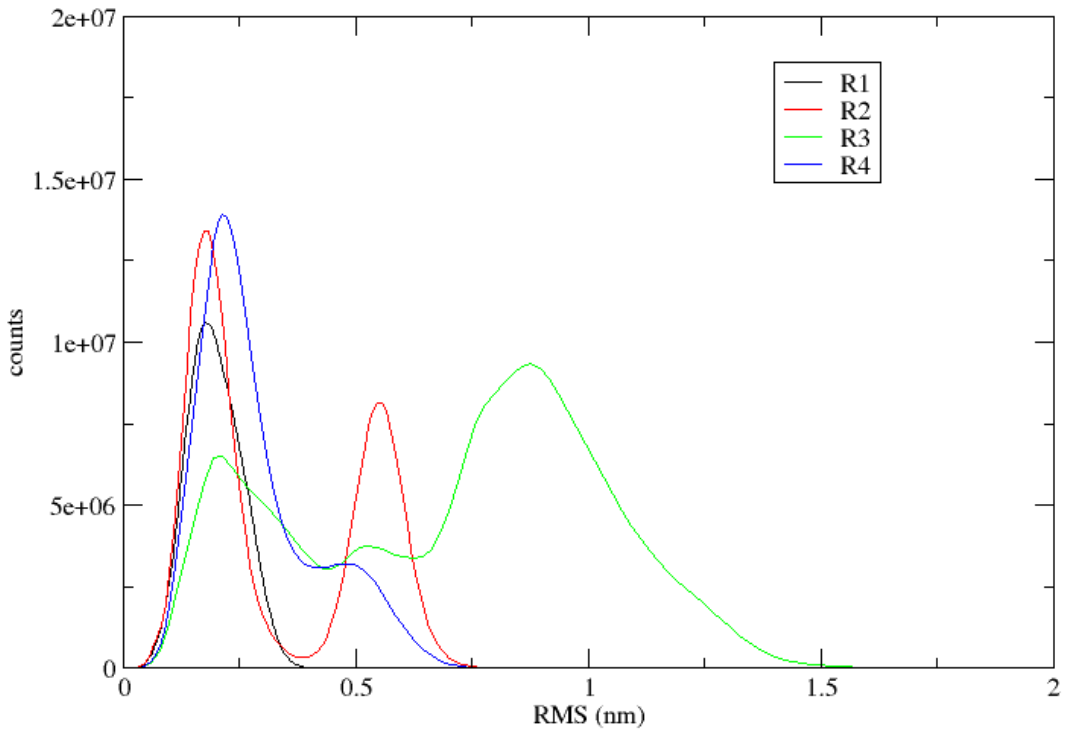


Figure S103. RMSD distribution of gemfibrozil (bound to Arm X) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.35 nm was selected to cluster the respective trajectories into representative groups (see Table S12).

Table S12. Overview of clusters from simulations of gemfibrozil and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

GEMFIBROZIL (2 molecules)			
Pocket	Replica	# clusters	Percentage (%)
ARMI	R1		
	R2	3	(1) **96.8** (2) **1.6** (3) **1.1**
	R3	1	(1) **98.2**
	R4	4	(1) **47.2** (2) **37.2** (3) **12.7** (4) **1.3**
ARMX	R1	1	(1) **100**
	R2	2	(1) **73.2** (2) **26.6**
	R3	7	(1) **37.0** (2) **28.6** (3) **13.0** (4) **10.7** (5) **2.6** (6) **2.4** (7) **2.2**
	R4	3	(1) **92.0** (2) **4.9** (3) **3.1**

8.3.4 Gemfibrozil bound to the Center/Arm II/Arm III and Arm X pocket

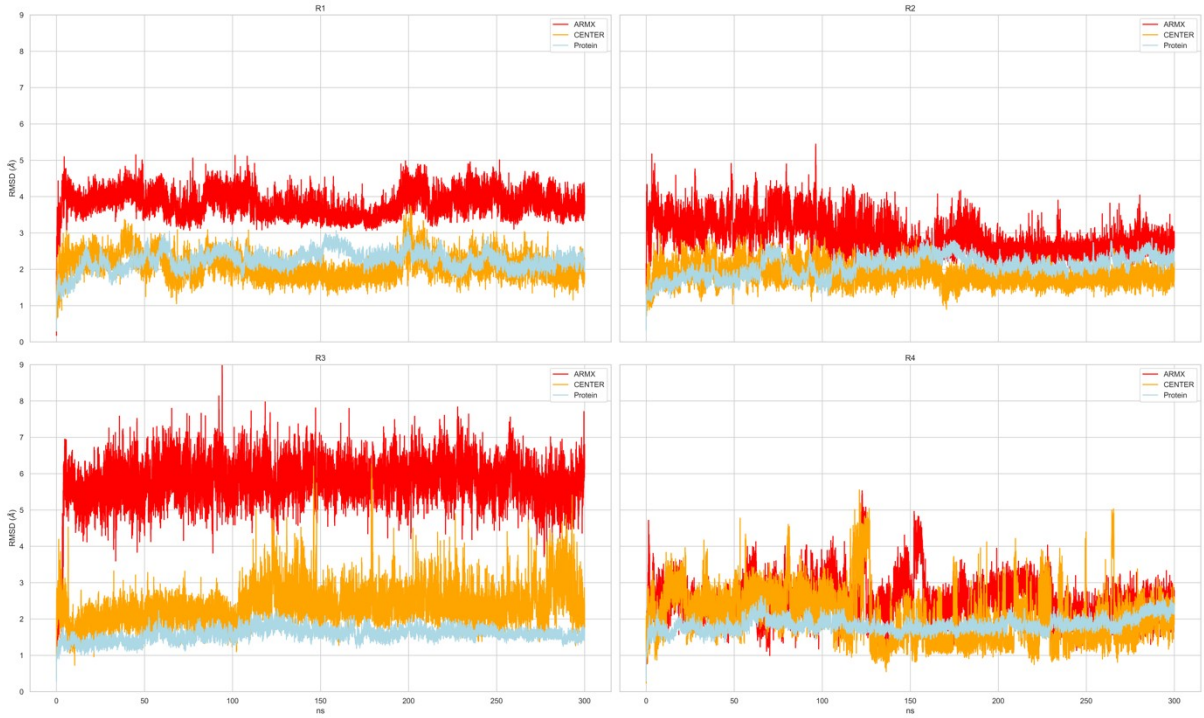


Figure S104. RMSD calculation of the protein backbone (light blue), gemfibrozil bound to Center/Arm II/Arm III pocket (orange) and gemfibrozil bound to Arm X (red). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

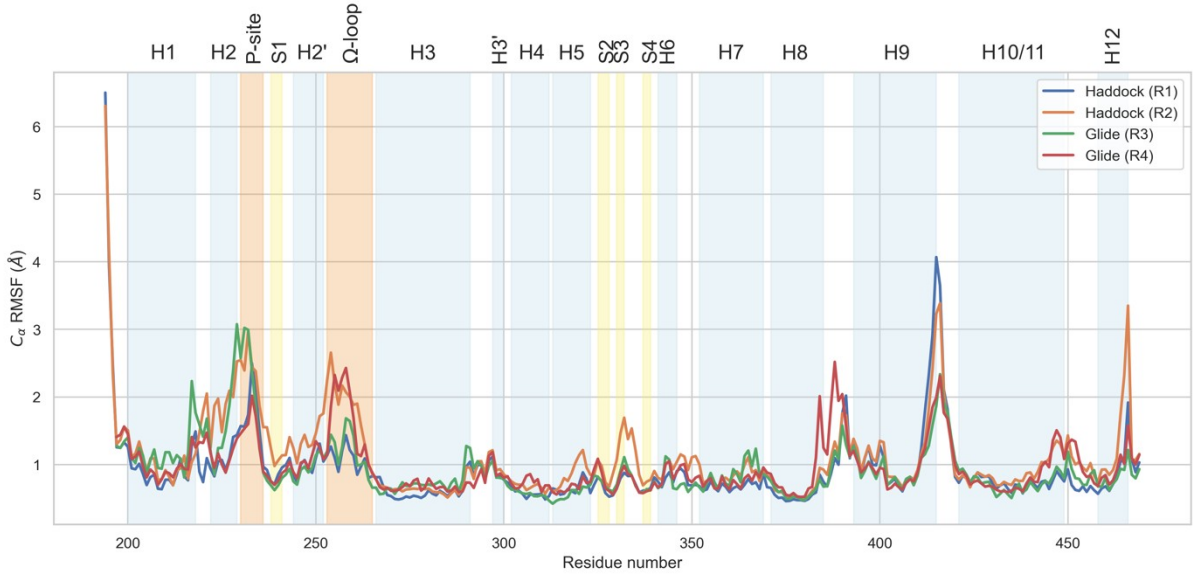


Figure S105. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

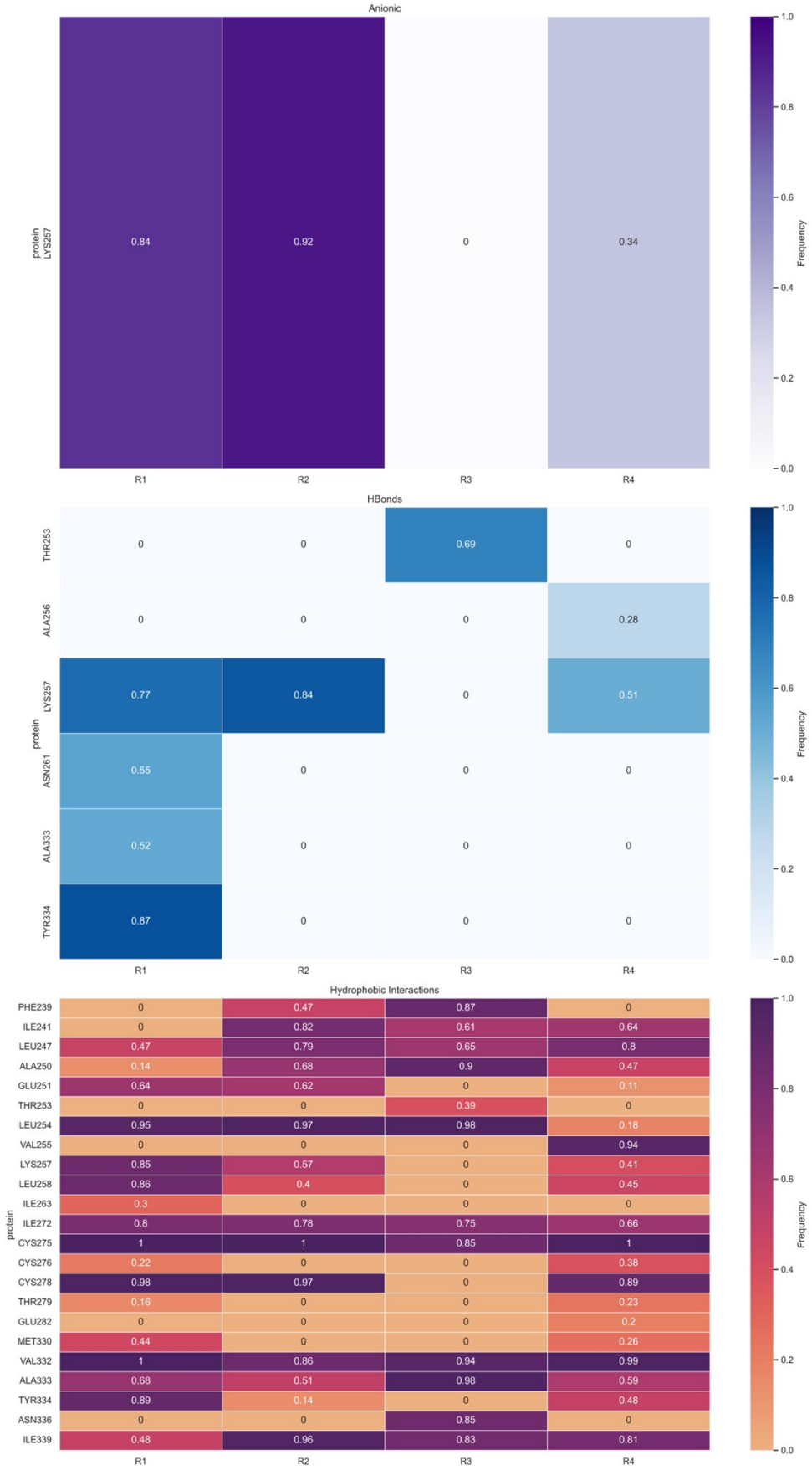


Figure S106. Heatmap of gemfibrozil-protein interactions (gemfibrozil bound to Arm X).

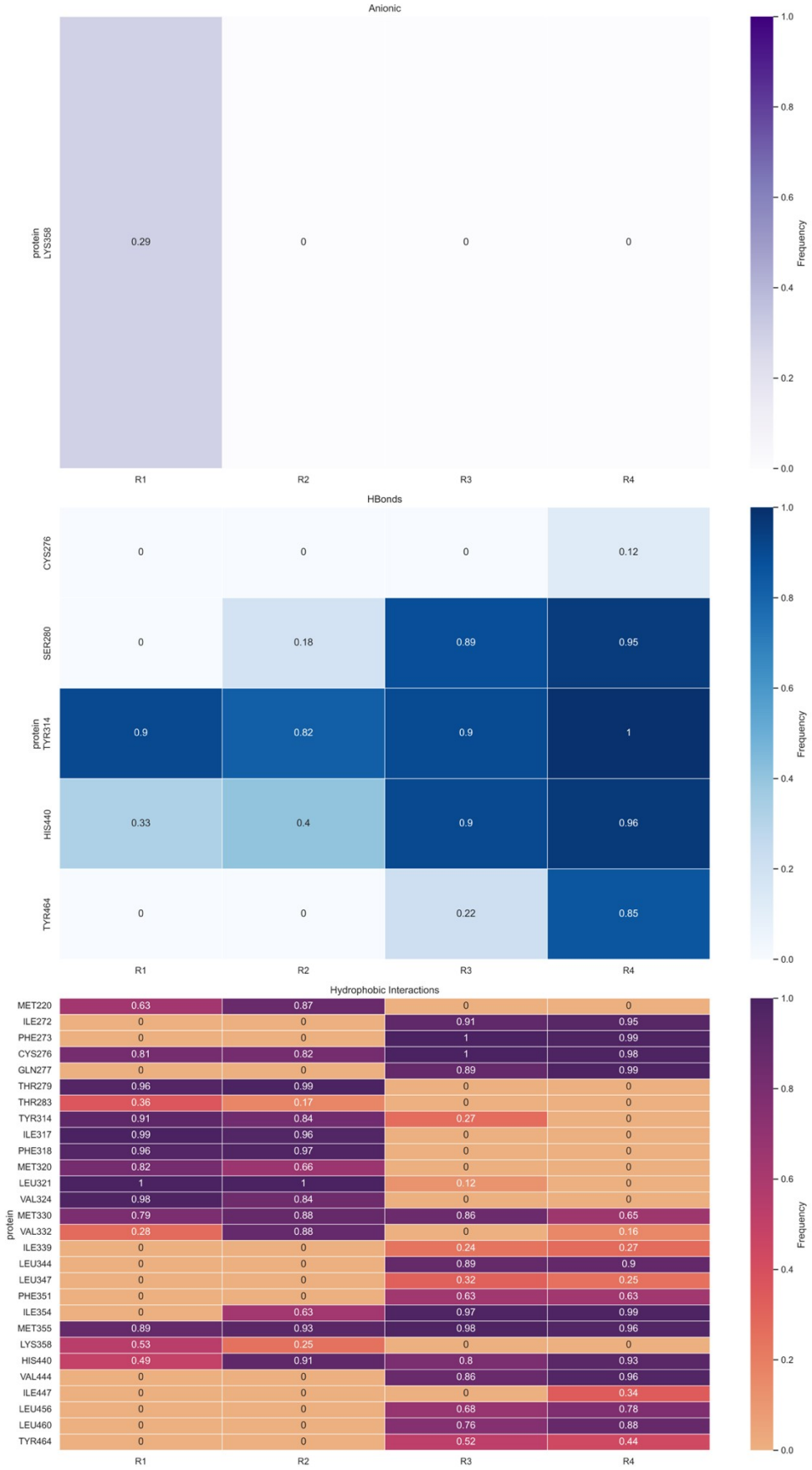


Figure S107. Heatmap of gemfibrozil-protein interactions (gemfibrozil bound to the Center/Arm II/Arm III pocket).

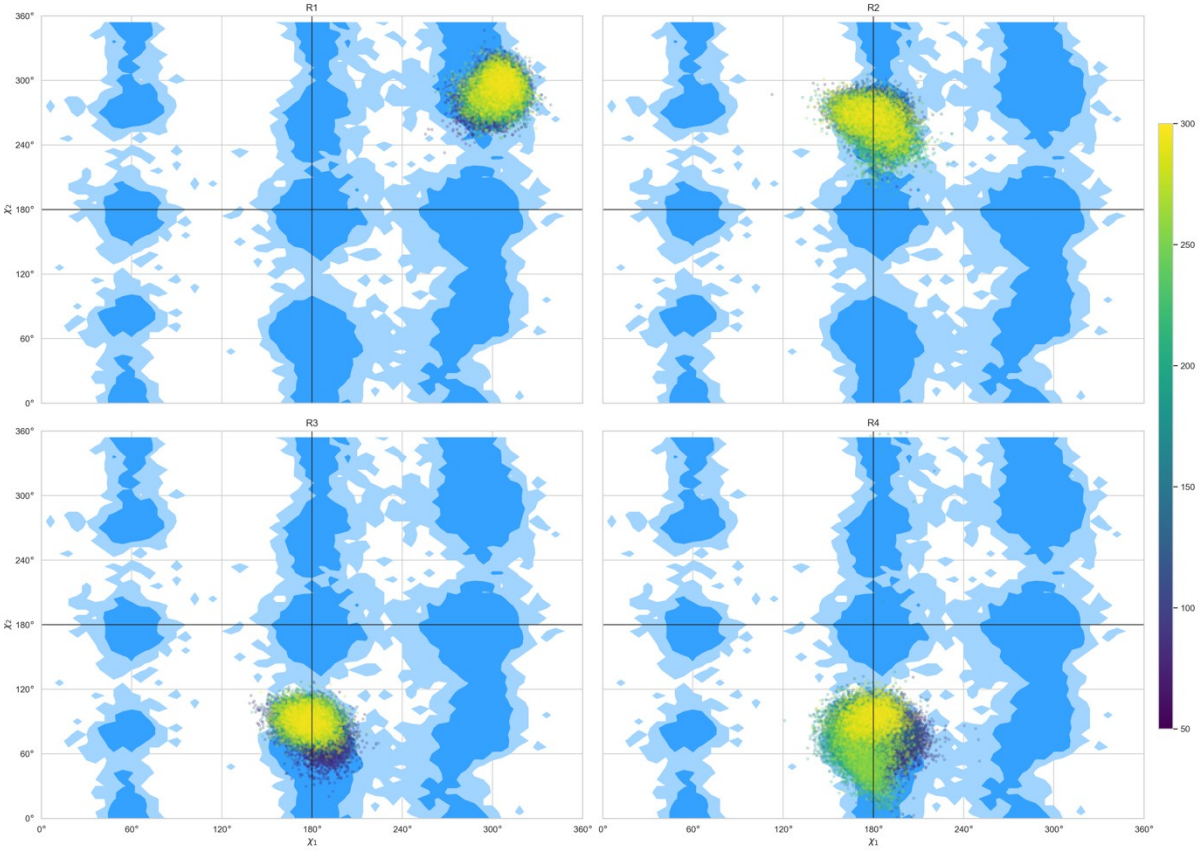


Figure S108. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

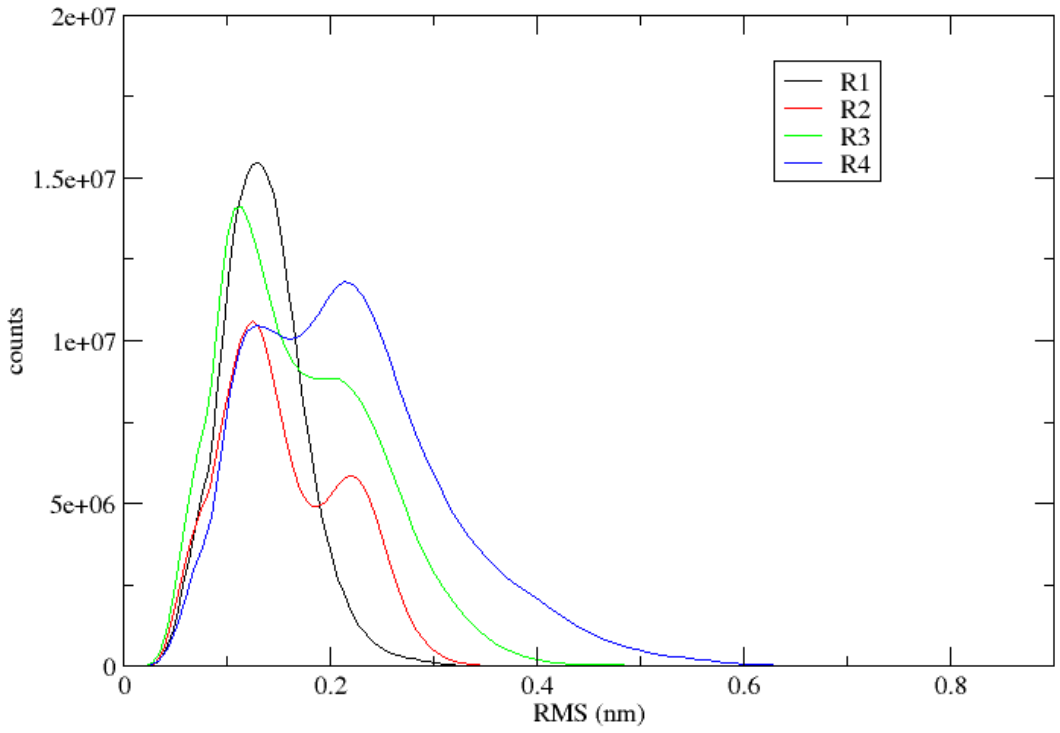


Figure S109. RMSD distribution gemfibrozil (bound to the Center/Arm II/Arm III pocket) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.15 nm was selected to cluster the respective trajectories into representative groups (see Table S13).

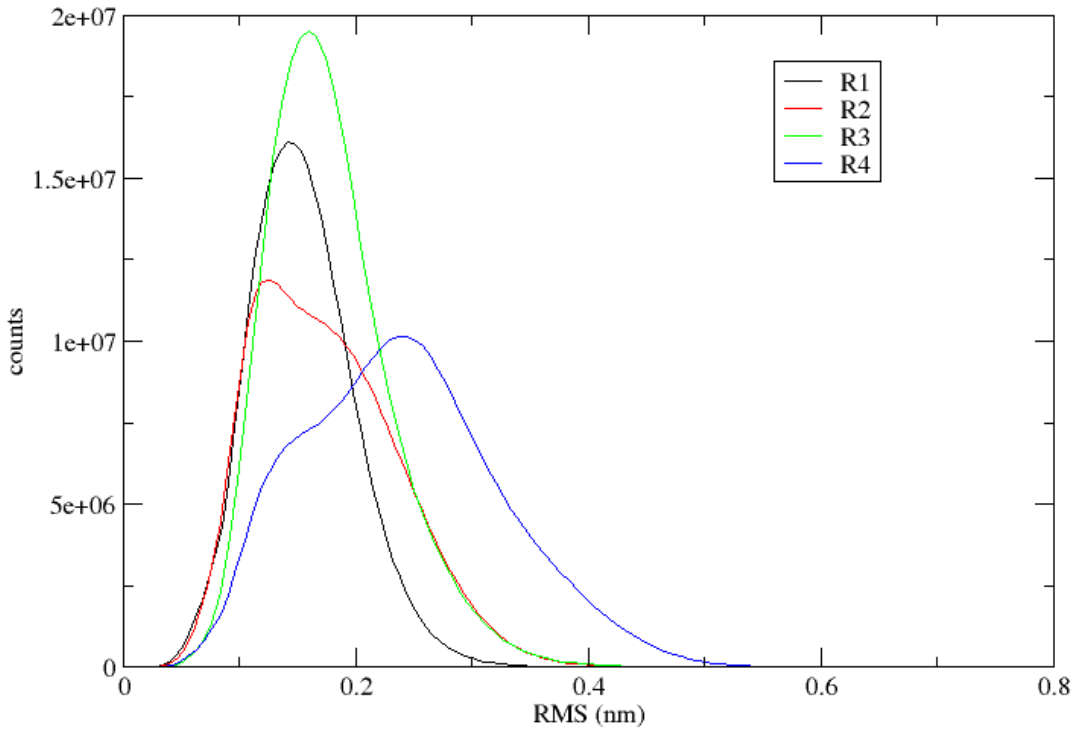


Figure S110. RMSD distribution of gemfibrozil (bound to Arm X) ligand in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S13).

Table S13. Overview of clusters from simulations of gemfibrozil and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

GEMFIBROZIL (2 molecules)			
Pocket	Replica	# clusters	Percentage (%)
CENTER	R1	3	(1) **93.3** (2) **2.6** (3) **1.9**
	R2	4	(1) **78.0** (2) **18.2** (3) **1.7** (4) **1.2**
	R3	4	(1) **71.0** (2) **21.8** (3) **4.3** (4) **1.4**
	R4	6	(1) **45.1** (2) **20.6** (3) **16.0** (4) **12.5** (5) **2.5** (6) **1.0**
ARMX	R1	1	(1) **98.7**
	R2	3	(1) **93.0** (2) **4.0** (3) **1.6**
	R3	3	(1) **92.4** (2) **3.6** (3) **2.6**
	R4	6	(1) **67.6** (2) **11.7** (3) **10.4** (4) **3.3** (5) **2.6** (6) **2.2**

8.4 Cinnamic Acid

8.4.1 Cinnamic acid bound to the Arm I pocket

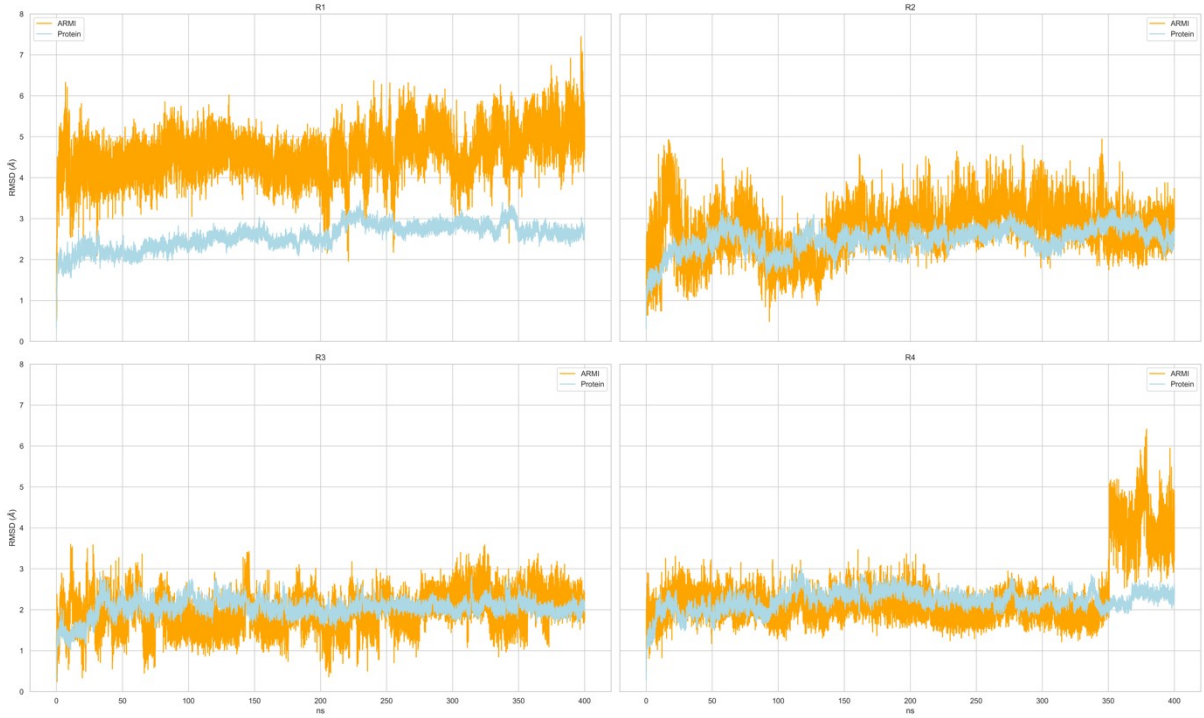


Figure S111. RMSD calculation of the protein backbone (light blue) and cinnamic acid bound to Arm I (orange). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

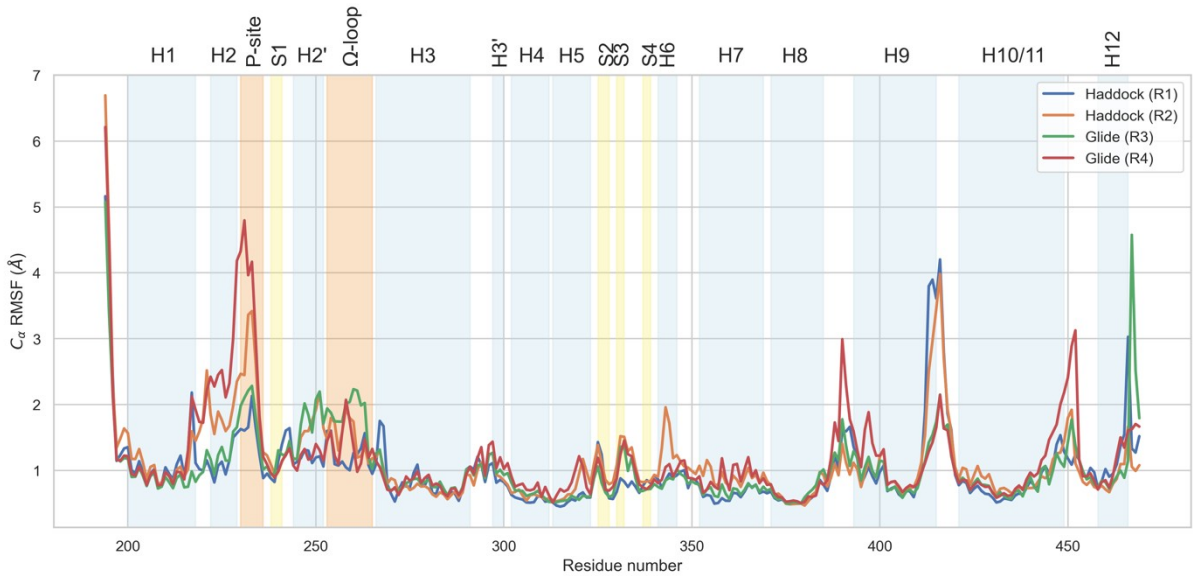


Figure S112. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

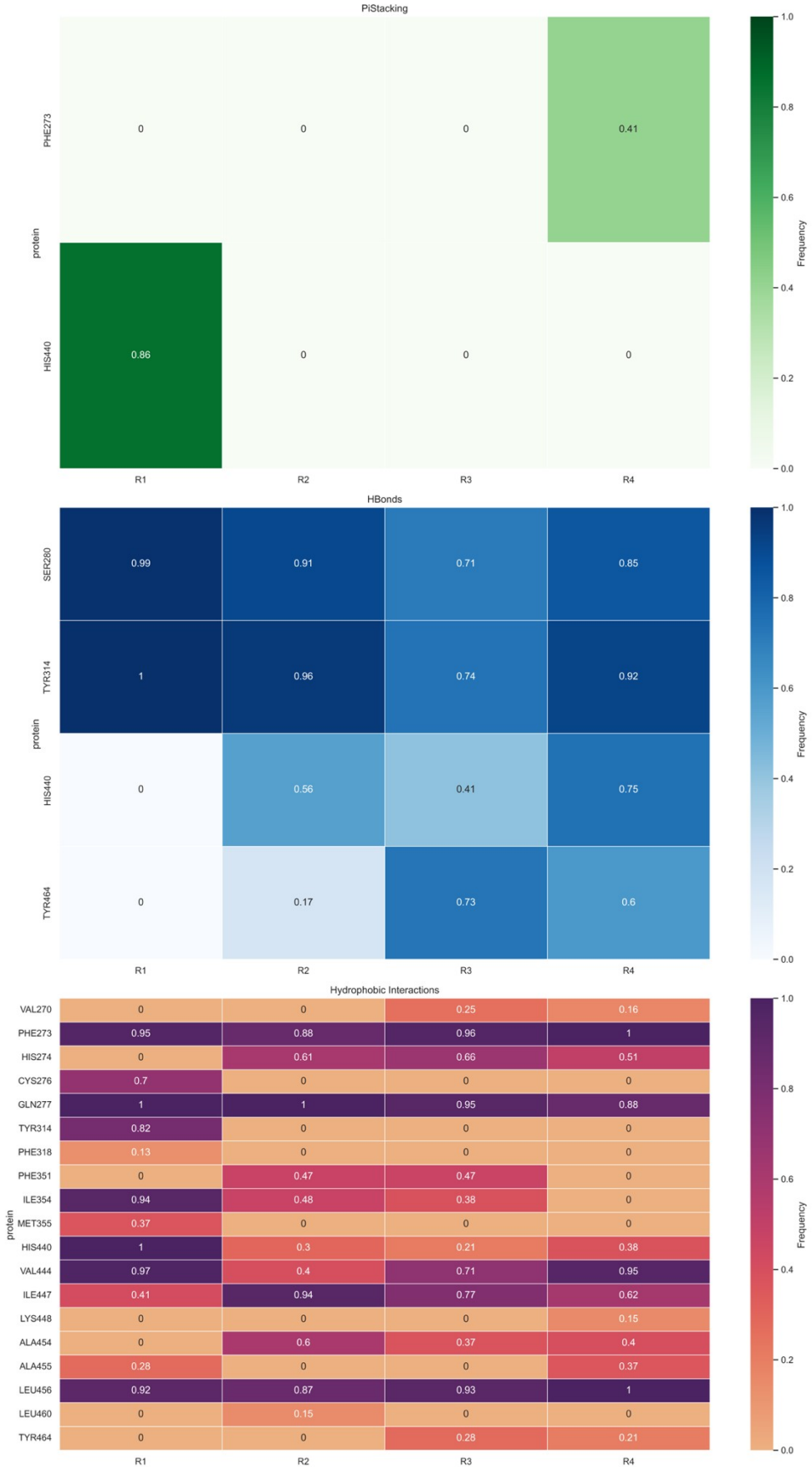


Figure S113. Heatmap of cinnamic acid-protein interactions (cinnamic acid bound to Arm I).

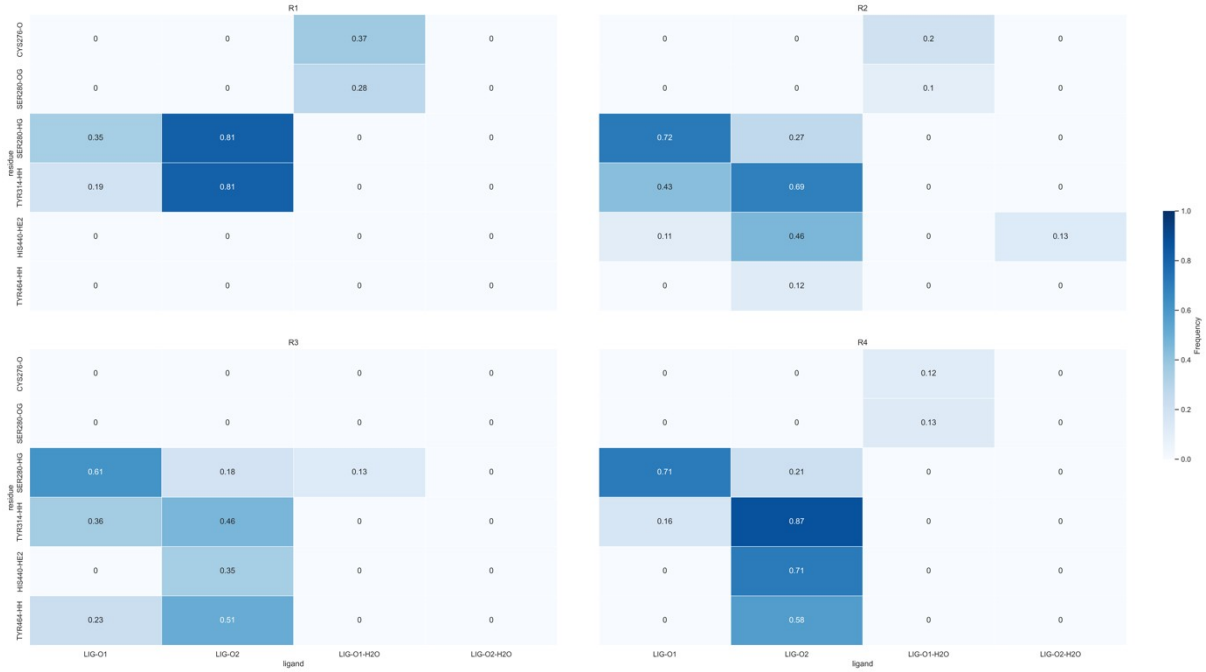


Figure S114. Detailed overview of H-bonds formed between cinnamic acid (bound to Arm I) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

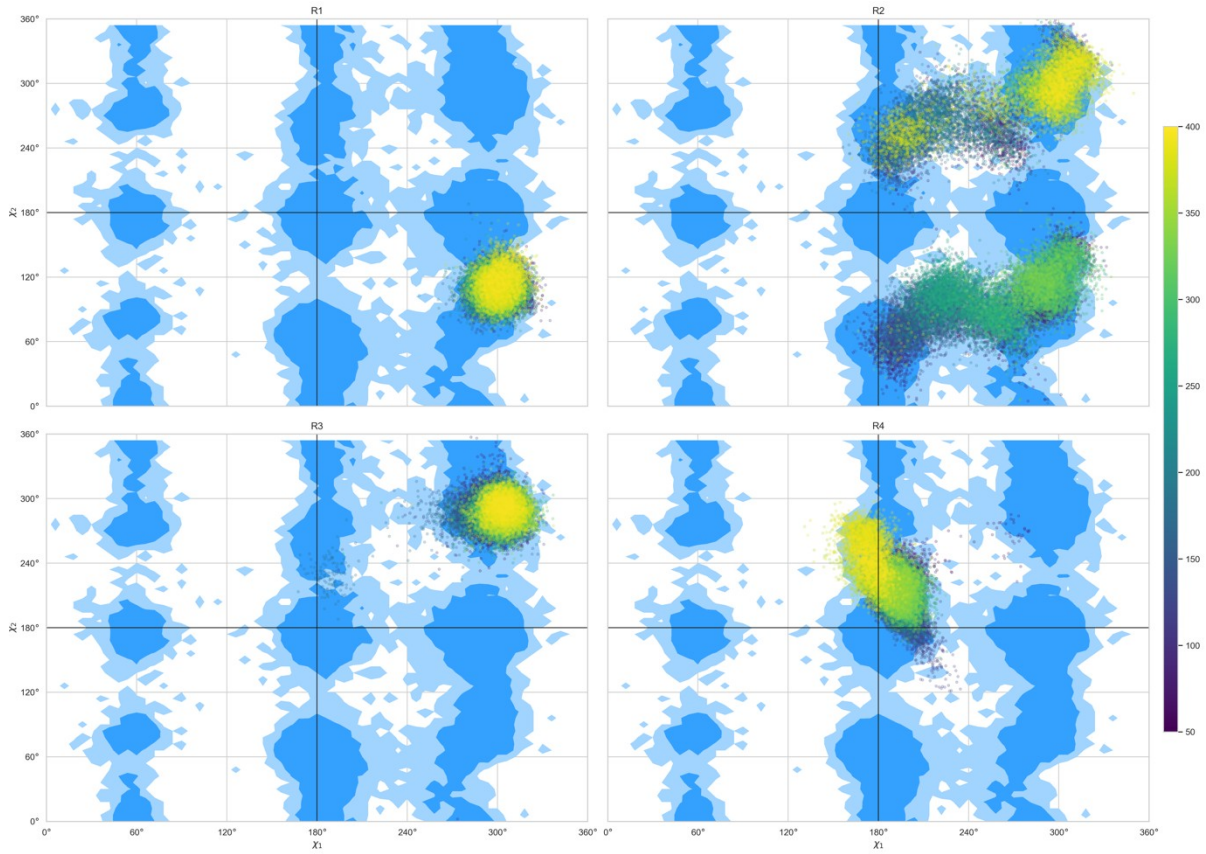


Figure S115. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

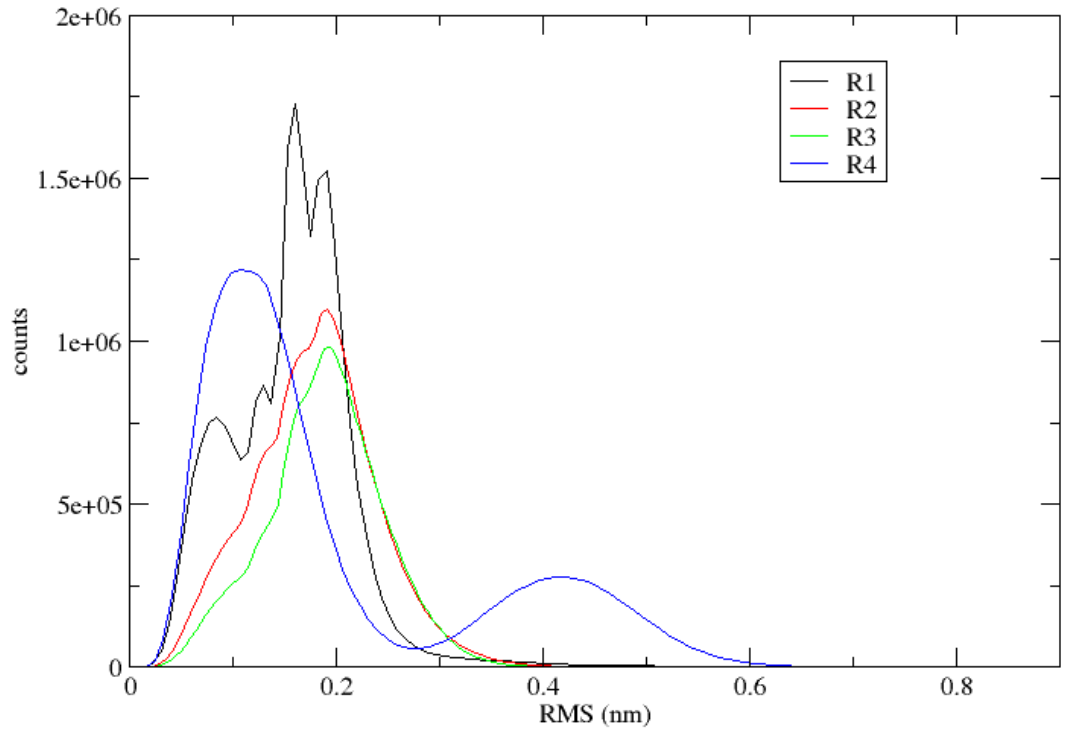


Figure S116. RMSD distribution of cinnamic acid (bound to Arm I) in the four replica MD simulations. Here, frames from 50-400ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S14).

Table S14. Overview of clusters from simulations of cinnamic acid and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

CINNAMIC ACID (1 molecule)			
Pocket	Replica	# clusters	Percentage (%)
A	R1	2	(1) **97.0** (2) **1.7**
R	R2	3	(1) **90.0** (2) **6.1** (3) **2.9**
M	R3	4	(1) **86.3** (2) **8.7** (3) **3.3** (4) **1.0**
I	R4	2	(1) **85.6** (2) **13.5**

8.4.2 Cinnamic acid bound to the Center/Arm II/Arm III pocket

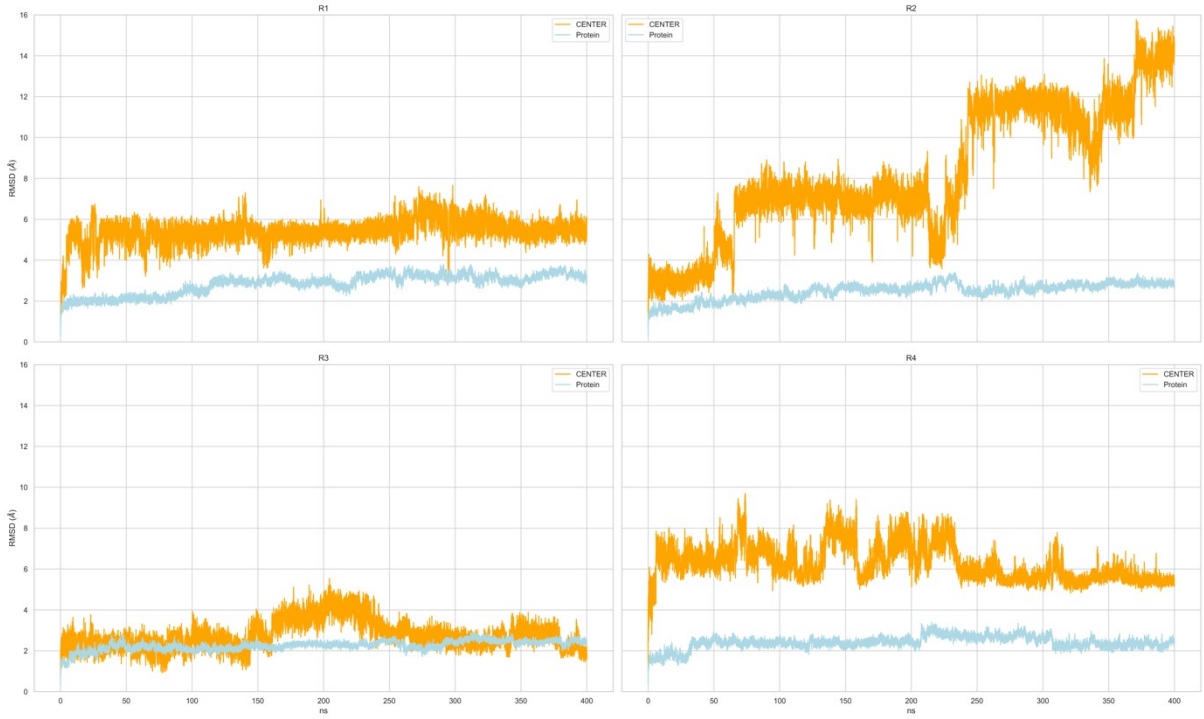


Figure S117. RMSD calculation of the protein backbone (light blue) and cinnamic acid bound to Center pocket (orange). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

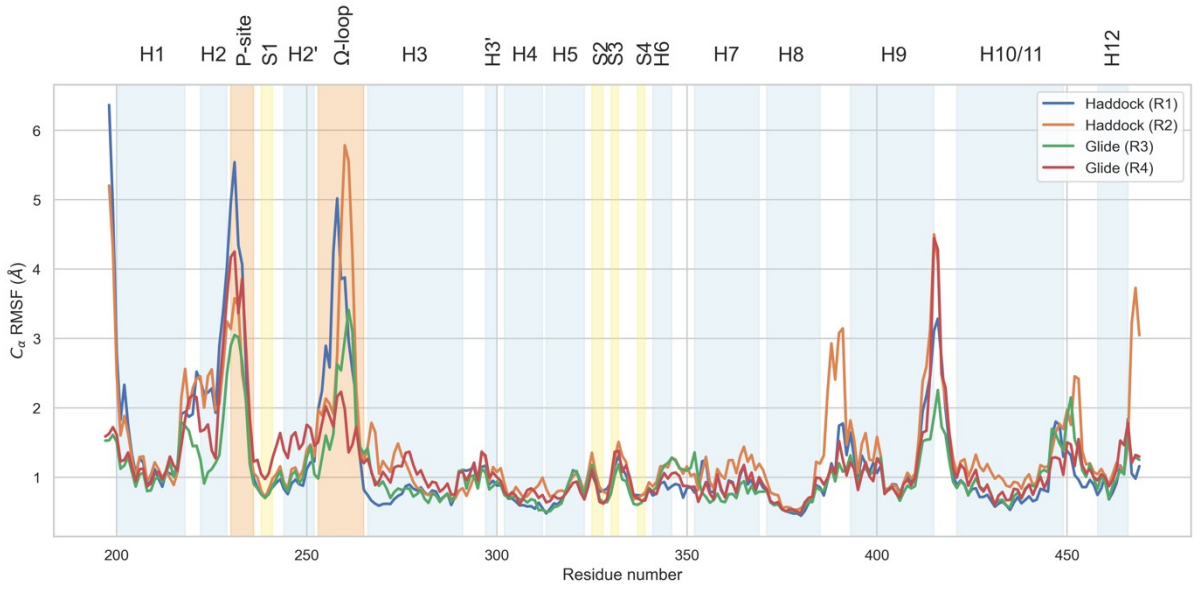


Figure S118. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

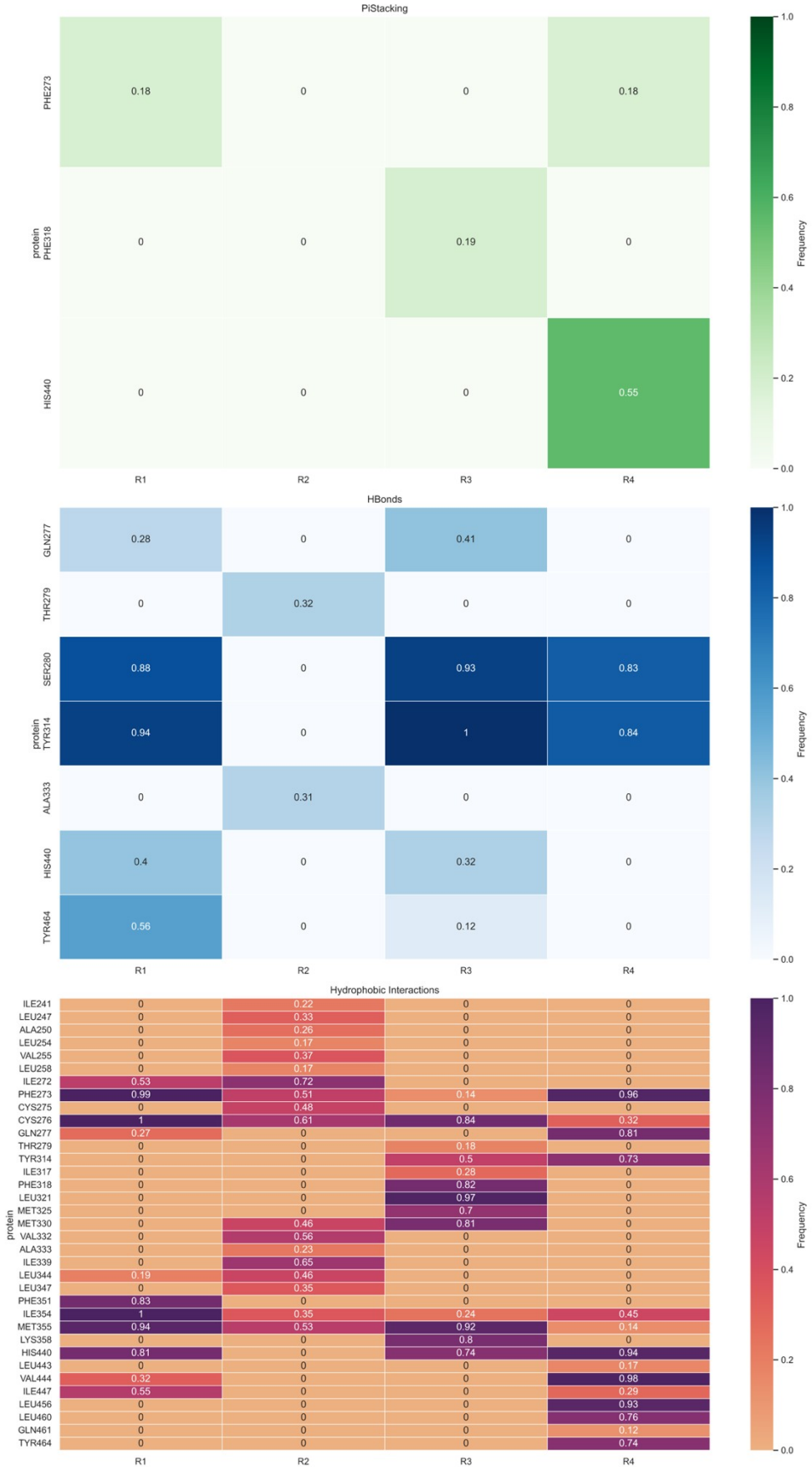


Figure S119. Heatmap of cinnamic acid-protein interactions (cinnamic acid bound to Center/Arm II/Arm III pocket).

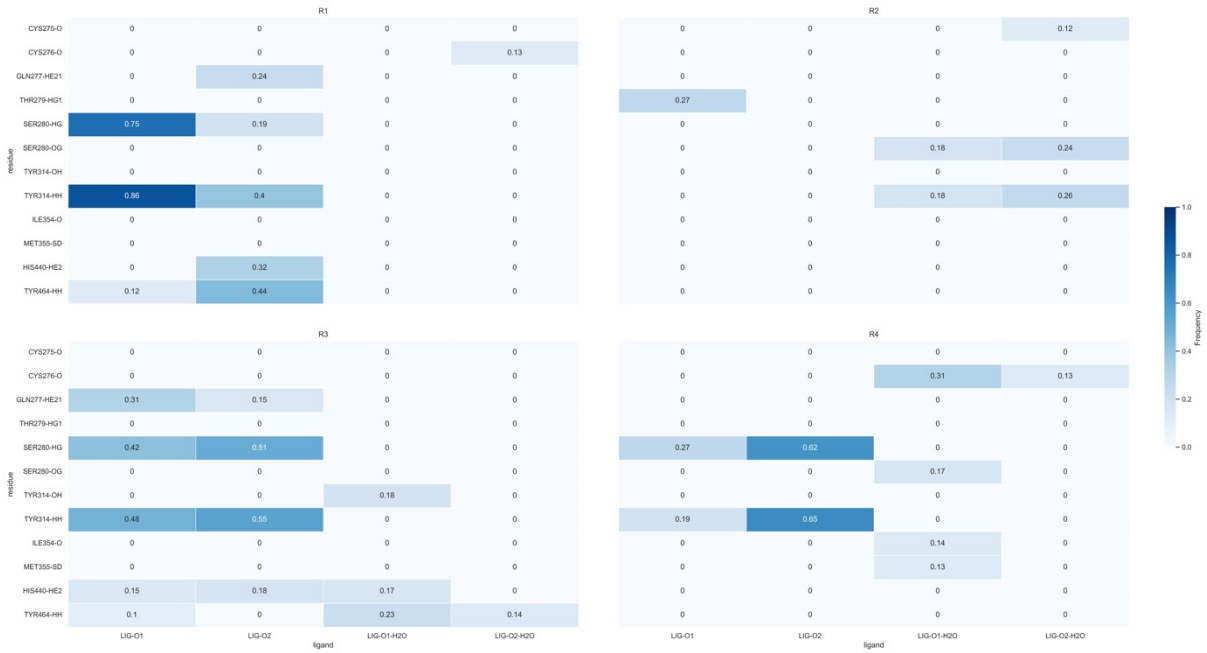


Figure S120. Detailed overview of H-bonds formed between cinnamic acid (bound to Center/Arm II/Arm III) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

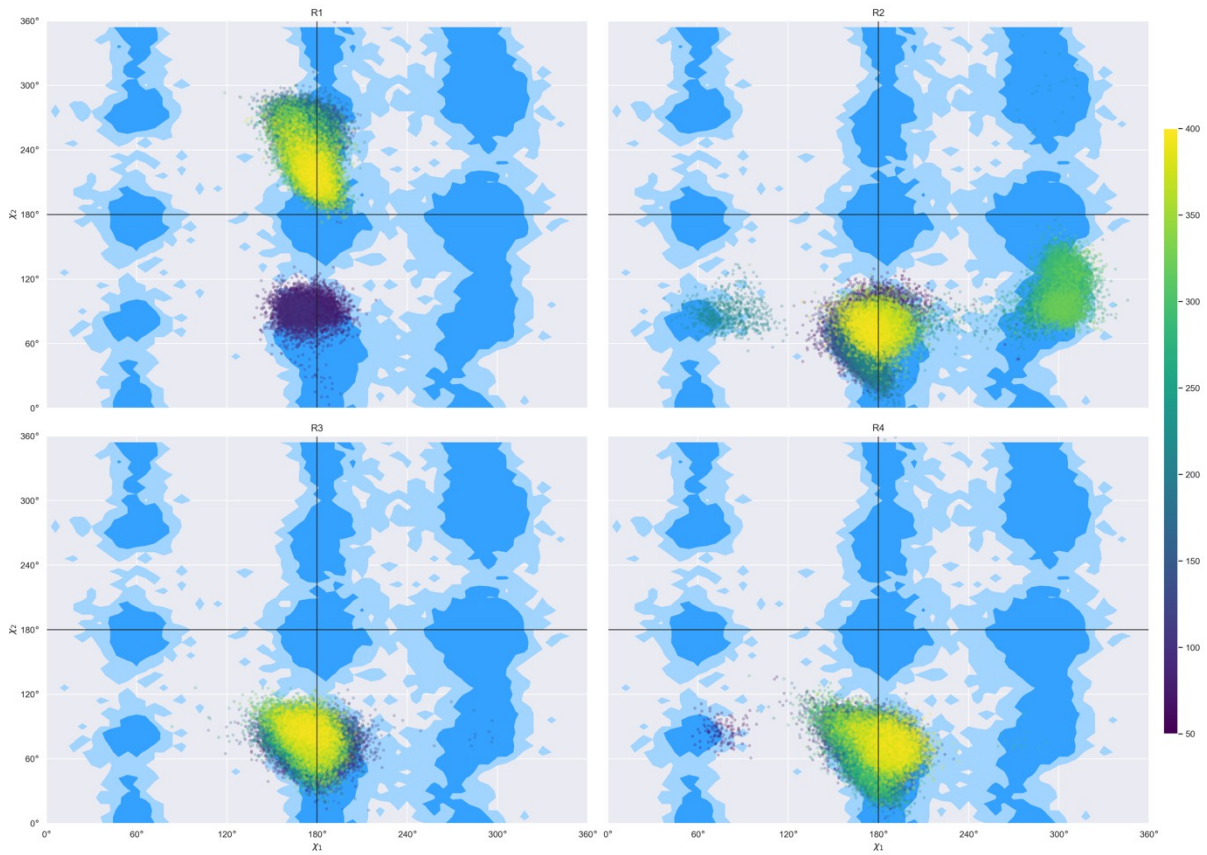


Figure S121. Time evolution of the Chi_1 and Chi_2 angles of F273. A data point represents the Chi_1 - Chi_2 angle combination in one particular MD frame. The colour coding indicates simulation time.

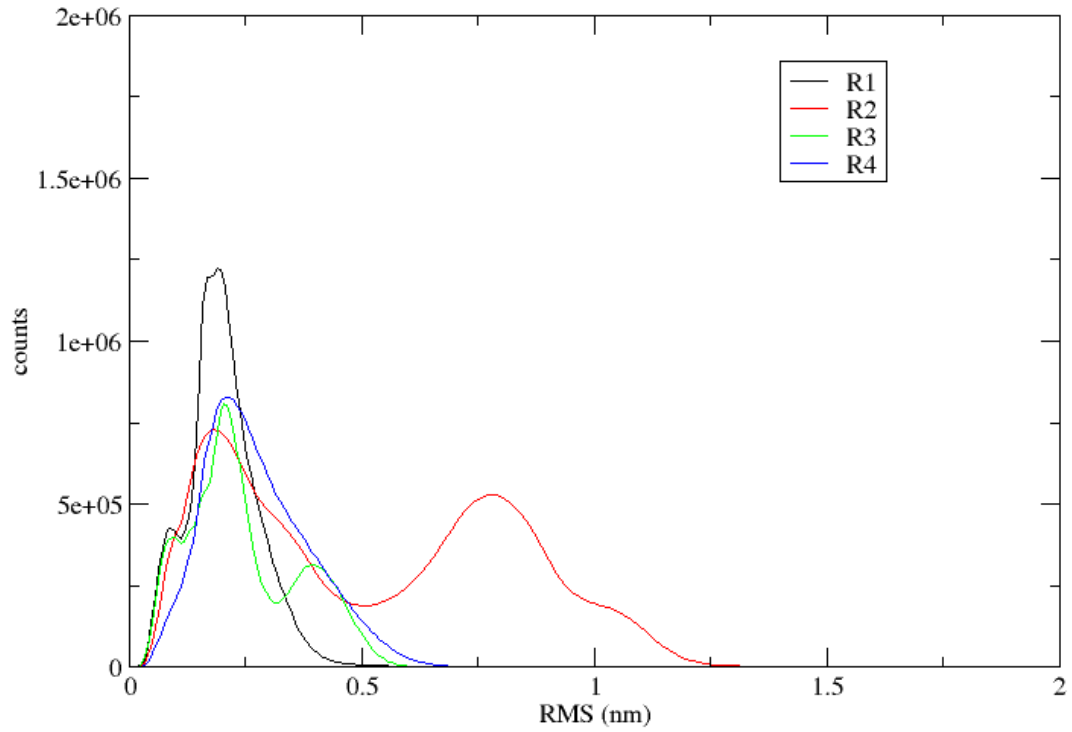


Figure S122. RMSD distribution of cinnamic acid (bound to Center/Arm II/Arm III) in the four replica MD simulations. Here, frames from 50-400ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.25 nm was selected to cluster the respective trajectories into representative groups (see Table S15).

Table S15. Overview of clusters from simulations of cinnamic acid (bound to Center/Arm II/Arm III) and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

CINNAMIC ACID (1 molecule)			
Pocket	Replica	# clusters	Percentage (%)
CENTER	R1	2	(1) **97.6** (2) **1.2**
	R2	7	(1) **44.4** (2) **34.4** (3) **7.7** (4) **4.0** (5) **3.4** (6) **2.6** (7) **1.0**
	R3	3	(1) **82.7** (2) **15.6** (3) **1.4**
	R4	3	(1) **75.5** (2) **19.0** (3) **4.4**

8.4.3 Cinnamic acid bound to the Arm I and Arm X

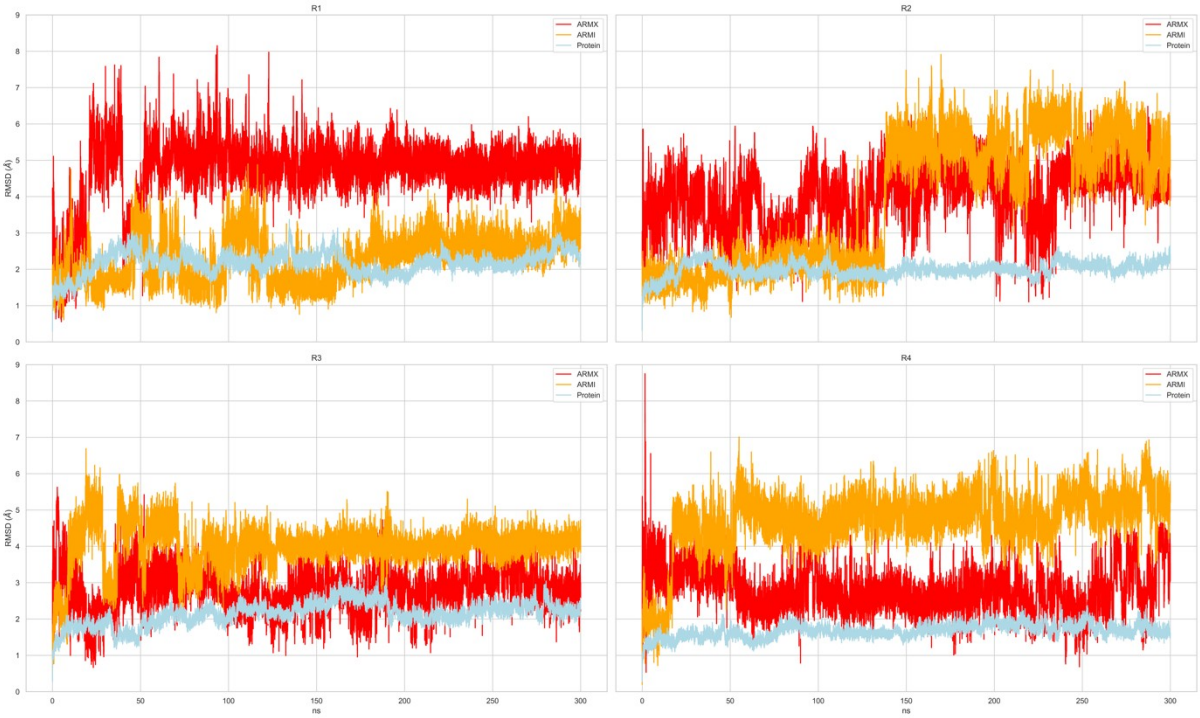


Figure S123. RMSD calculation of the protein backbone (light blue), cinnamic acid bound to Arm I (orange) and cinnamic acid bound to Arm X (red). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

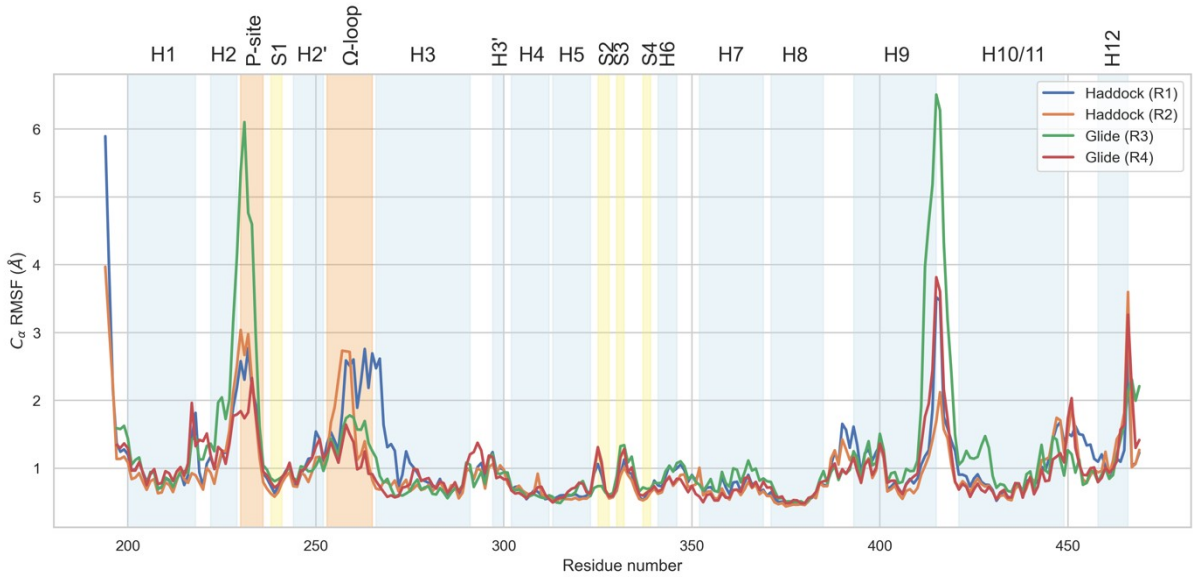


Figure S124. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

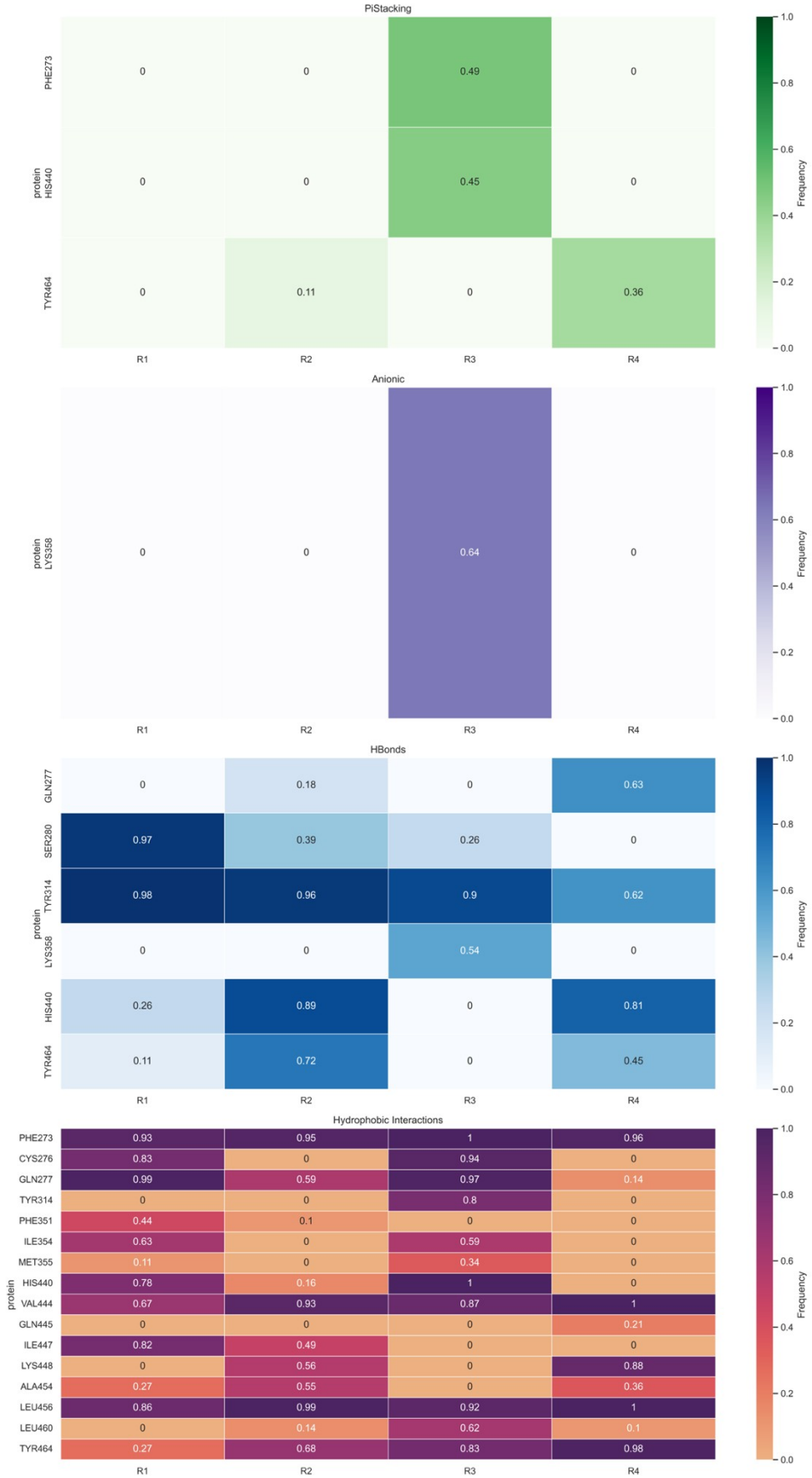


Figure S125. Heatmap of cinnamic acid-protein interactions (cinnamic acid bound to Arm I).

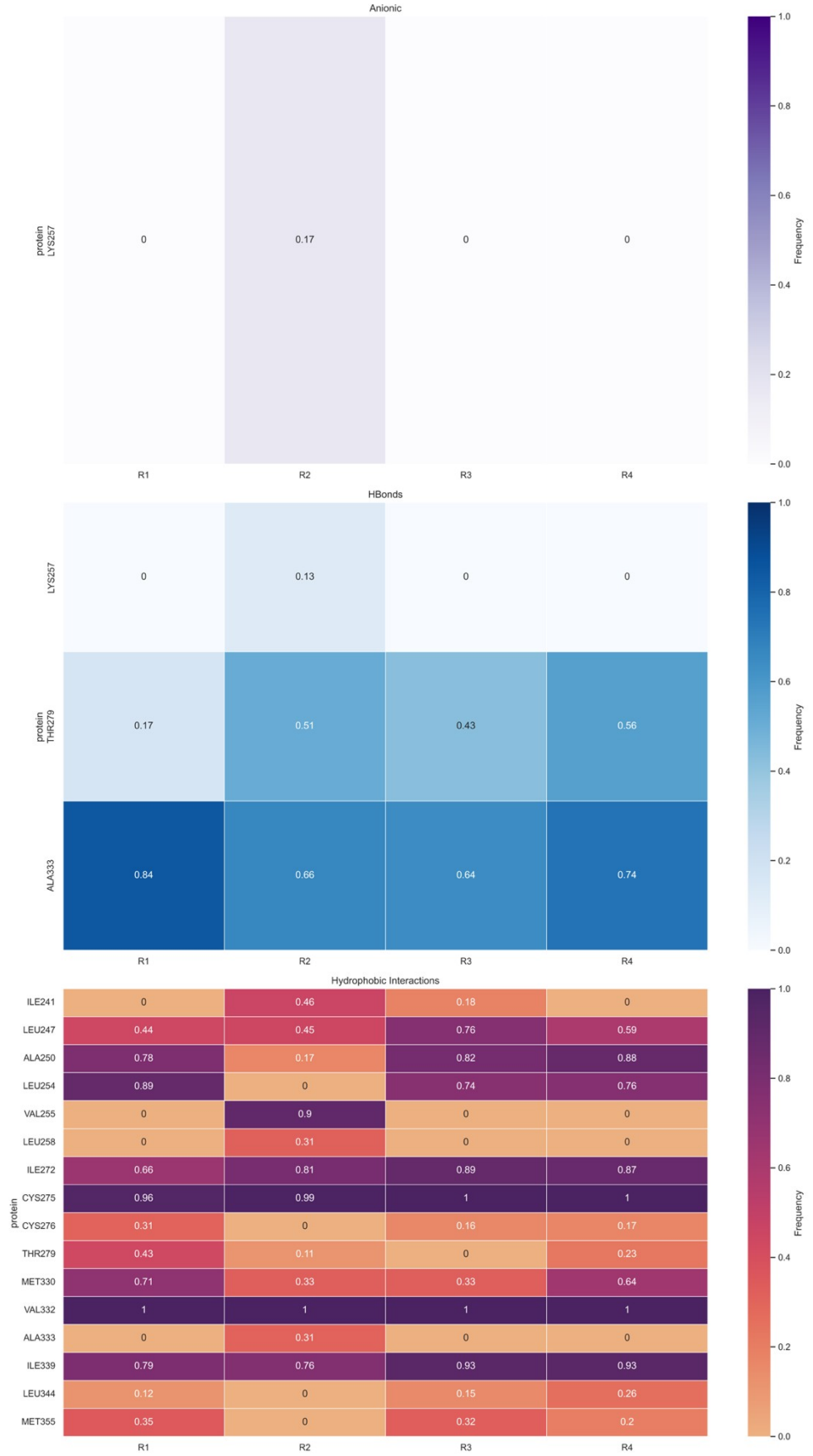


Figure S126. Heatmap of cinnamic acid-protein interactions (cinnamic acid bound to Arm X).

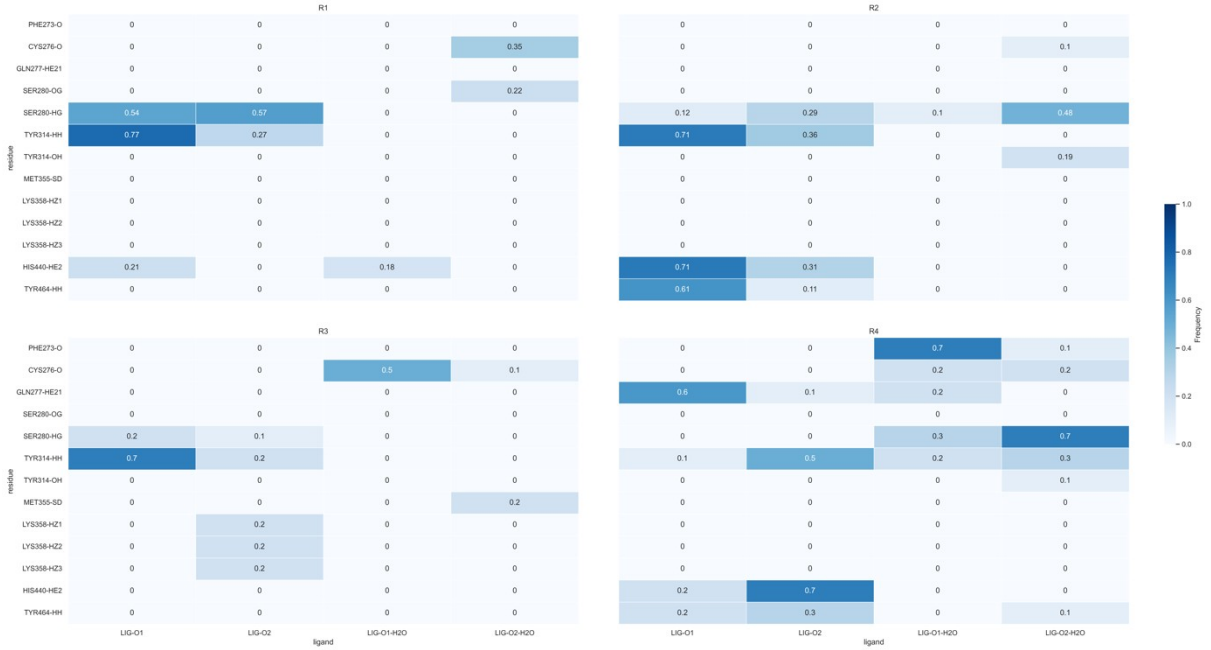


Figure S127. Detailed overview of H-bonds formed between cinnamic acid (bound to Arm I) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

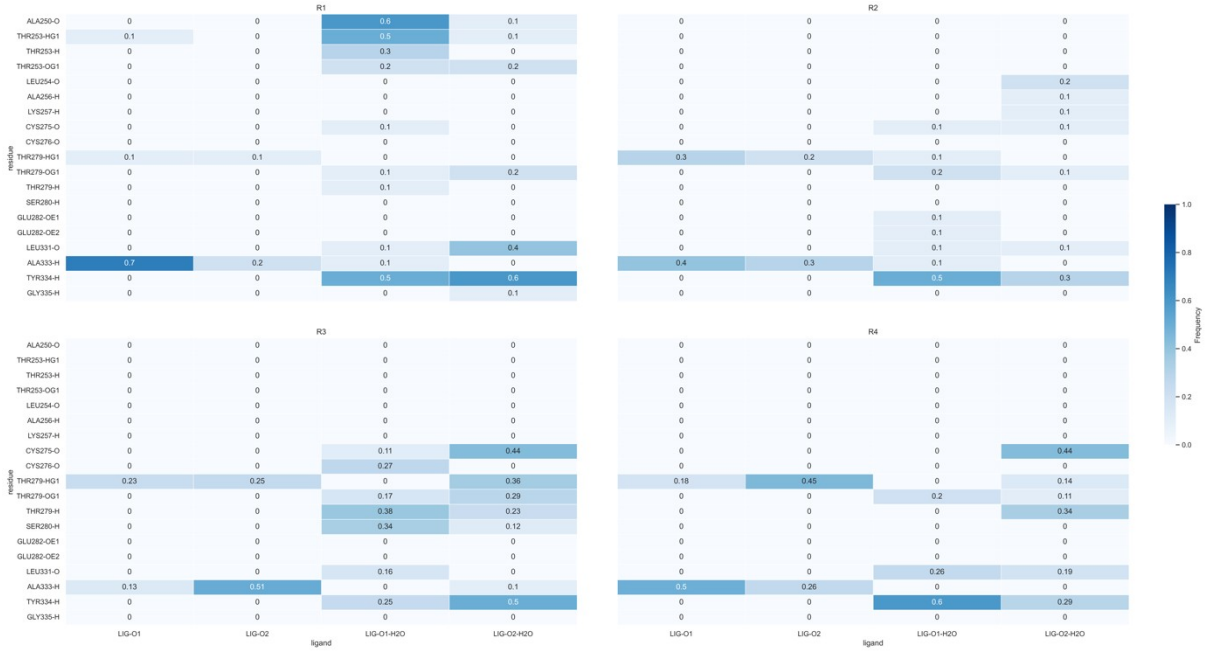


Figure S128. Detailed overview of H-bonds formed between cinnamic acid (bound to Arm X) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

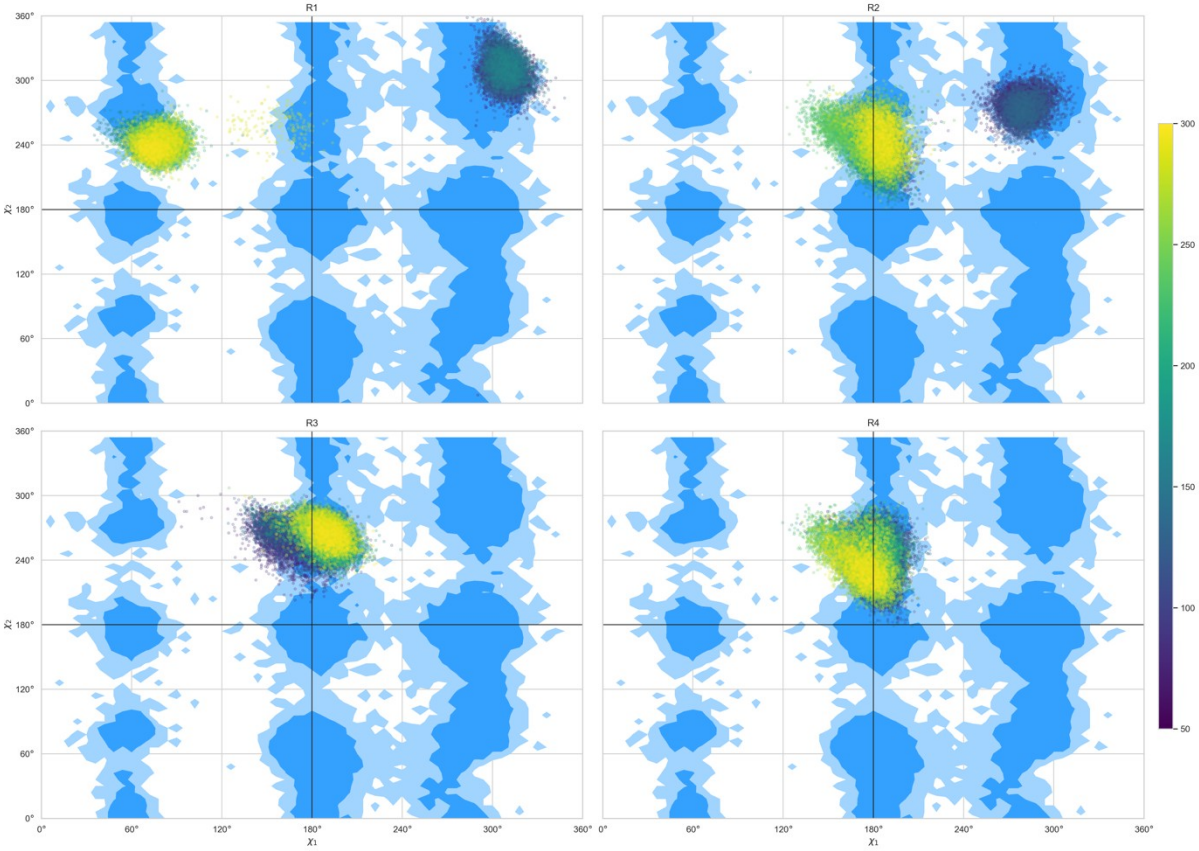


Figure S129. Time evolution of the Chi$_1$ and Chi$_2$ angles of F273. A data point represents the Chi$_1$ - Chi$_2$ angle combination in one particular MD frame. The colour coding indicates simulation time.

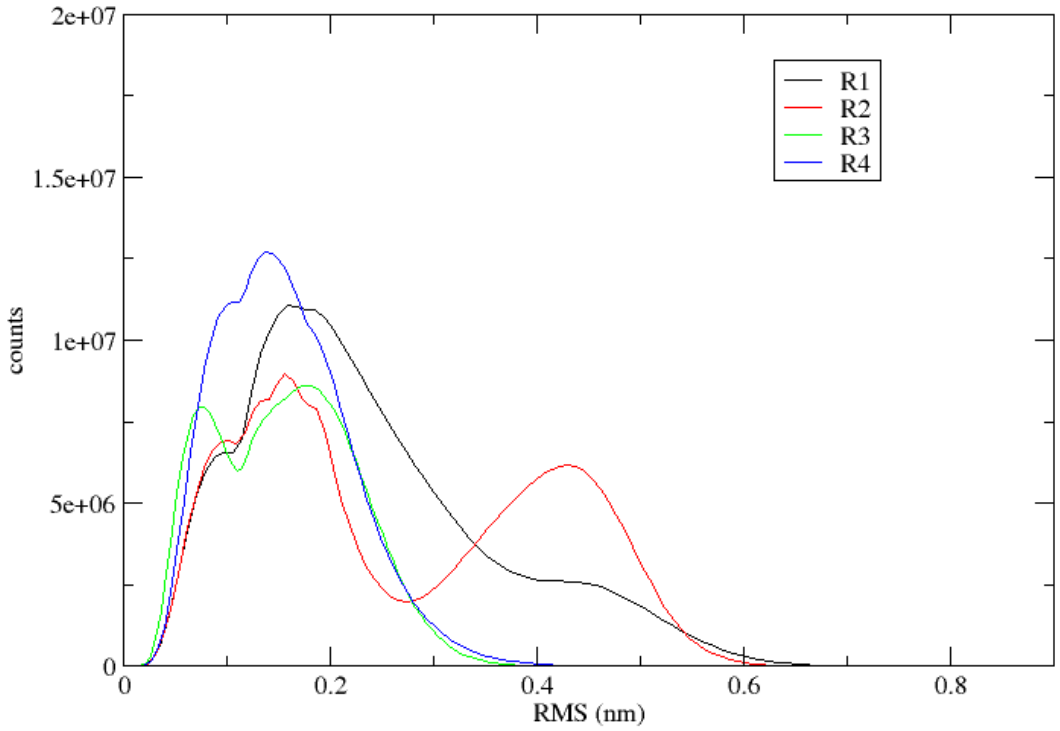


Figure S130. RMSD distribution of cinnamic acid (bound to Arm I) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S16).

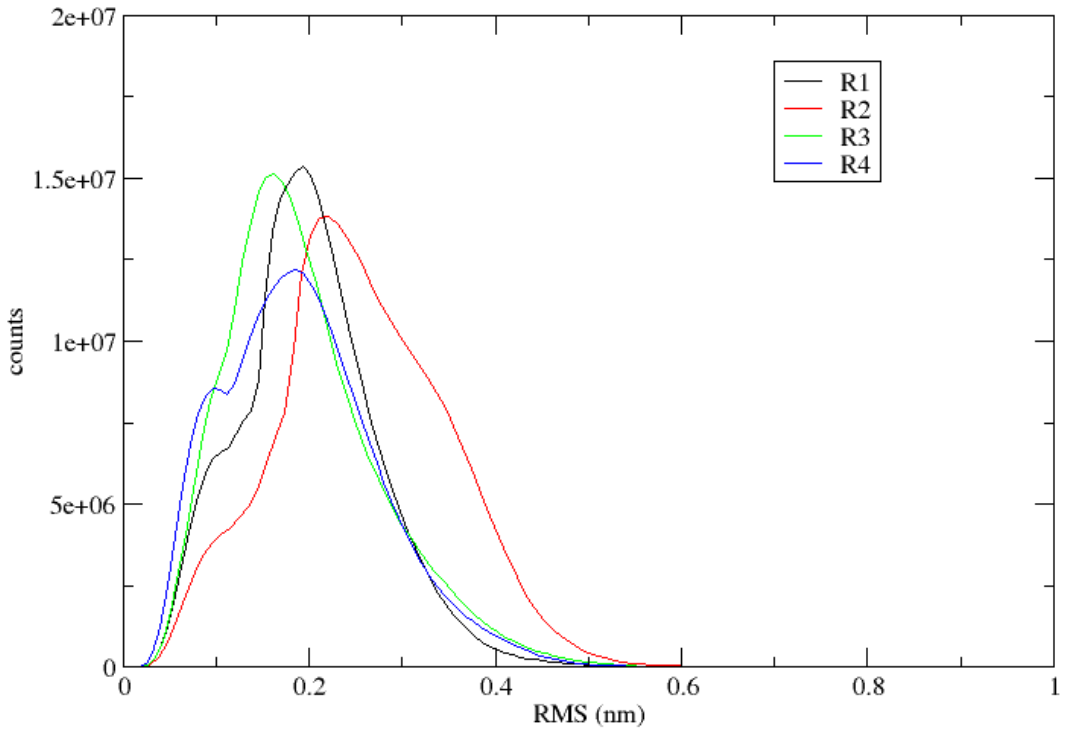


Figure S131. RMSD distribution of cinnamic acid (bound to Arm X) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S16).

Table S16. Overview of clusters from simulations of cinnamic acid and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

CINNAMIC ACID (2 molecules)			
Pocket	Replica	# clusters	Percentage (%)
A R M I	R1 R2 R3 R4	5 2 2 3	(1) **74.2** (2) **16.8** (3) **4.6** (4) **2.3** (5) **1.3** (1) **63.9** (2) **34.1** (1) **96.0** (2) 2.4 (1) **93.7** (2) **3.3** (3) **1.4**
A R M X	R1 R2 R3 R4	4 7 3 3	(1) **82.4** (2) **7.8** (3) **6.0** (4) **1.7** (1) **54.9** (2) **24.5** (3) **7.1** (4) **7.0** (5) **2.7** (6) **1.4** (7) **1.1** (1) **84.1** (2) **9.7** (3) **4.6** (1) **84.4** (2) **9.8** (3) **4.0**

8.4.4 Cinnamic acid bound to the Center/Arm II/Arm III and Arm X pocket

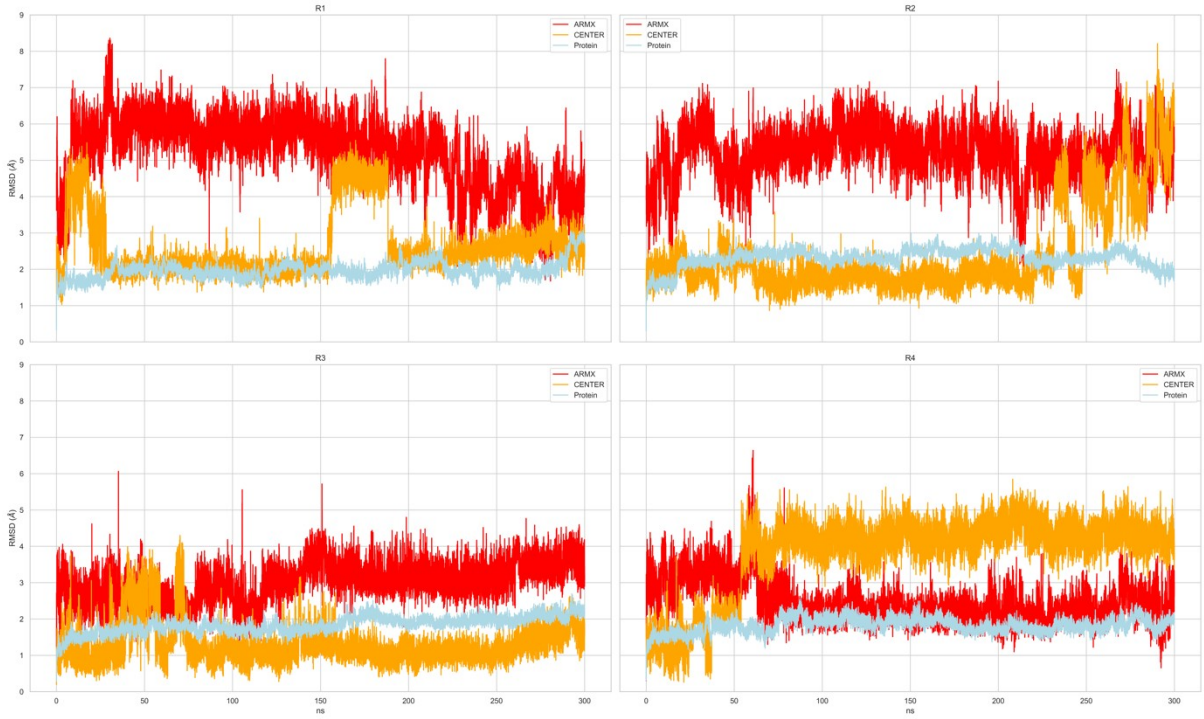


Figure S132. RMSD calculation of the protein backbone (light blue), cinnamic acid bound to Center/Arm II/Arm III pocket (orange) and cinnamic acid bound to Arm X (red). The starting structure of the MD simulations is the Haddock top docking pose (R1 and R2) and the Glide top docking pose (R3 and R4). For each protein-ligand complex separate velocities were generated according to a Maxwell-Boltzmann distribution at a temperature of 300K.

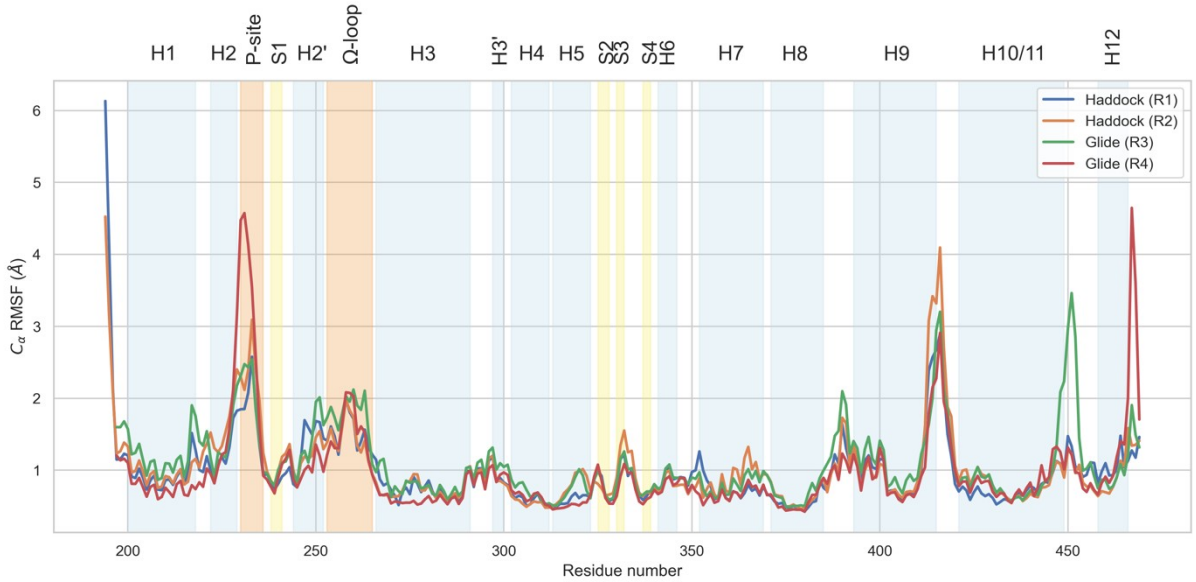


Figure S133. RMSF calculations of the C_α backbone atoms. Secondary structure elements were named according to reference [58].

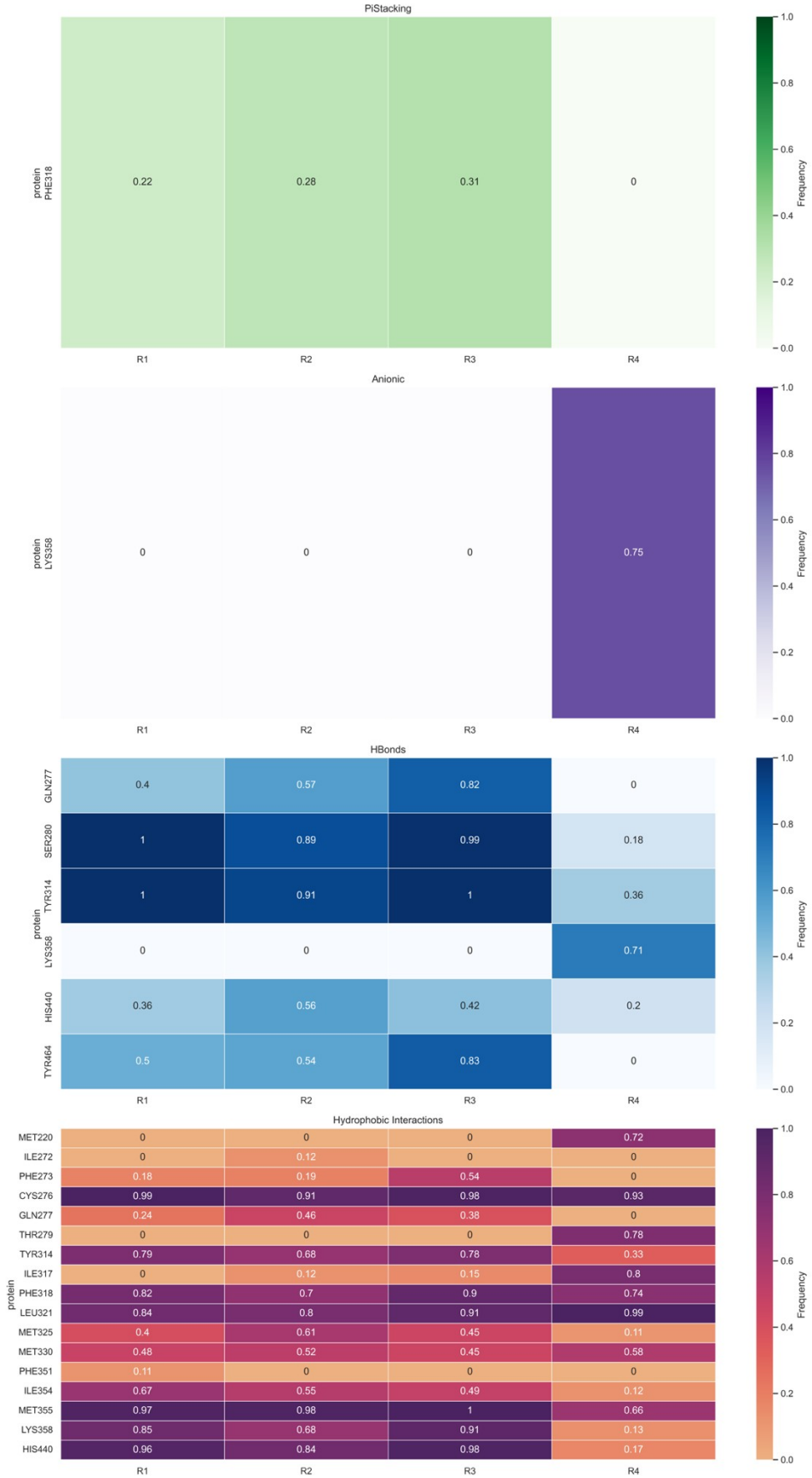


Figure S134. Heatmap of cinnamic acid-protein interactions (cinnamic acid bound to Center/Arm II/Arm III pocket)

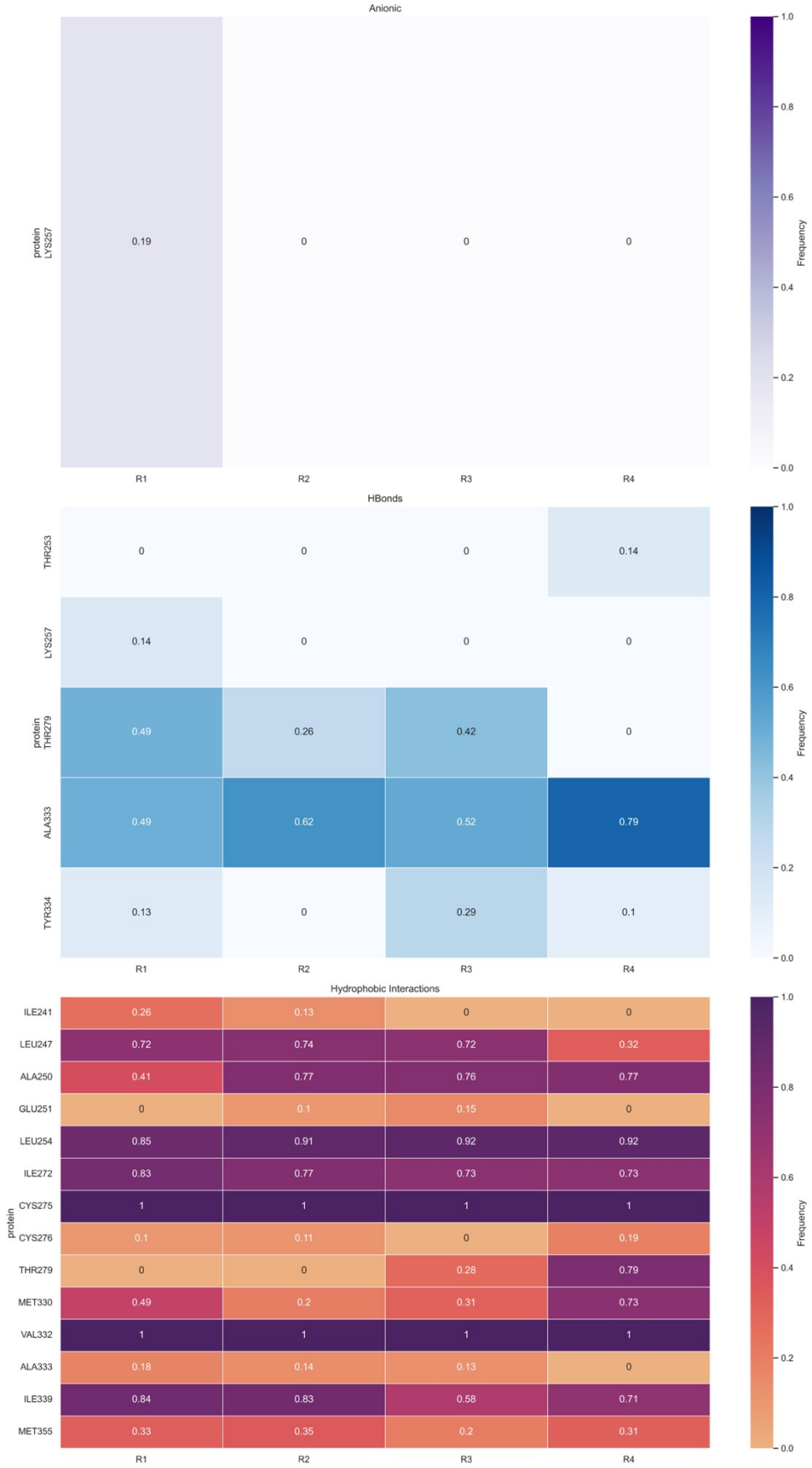


Figure S135. Heatmap of cinnamic acid-protein interactions (cinnamic acid bound to Arm X)

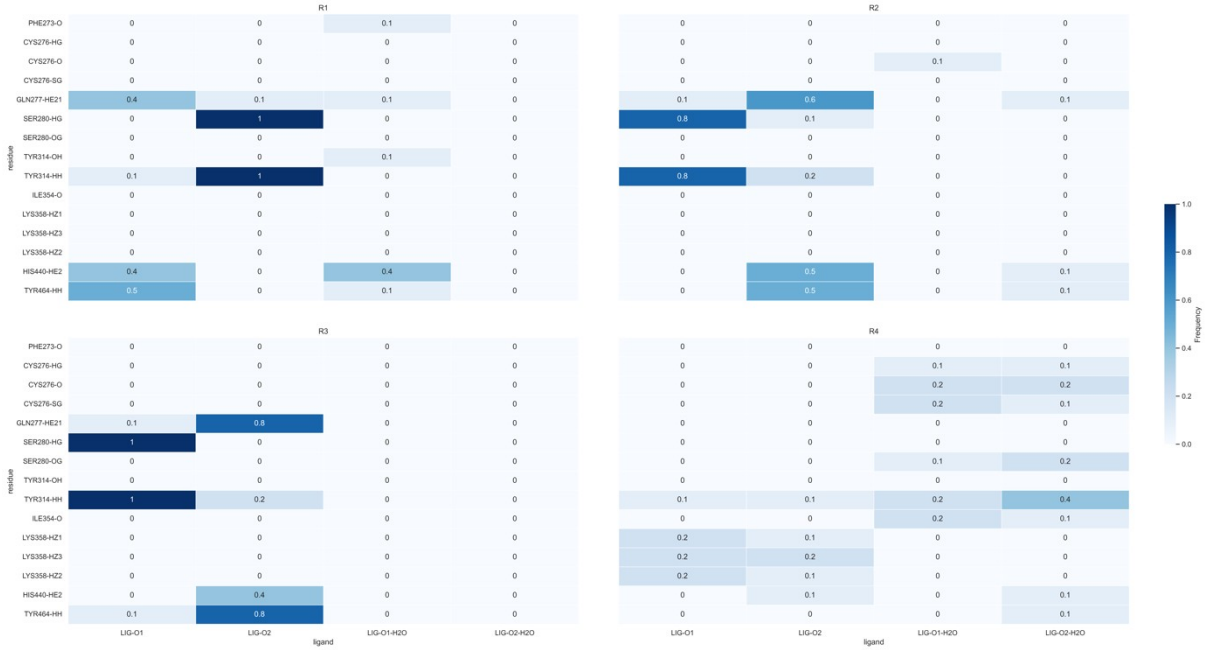


Figure S136. Detailed overview of H-bonds formed between cinnamic acid (bound to Center/Arm II/Arm III) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

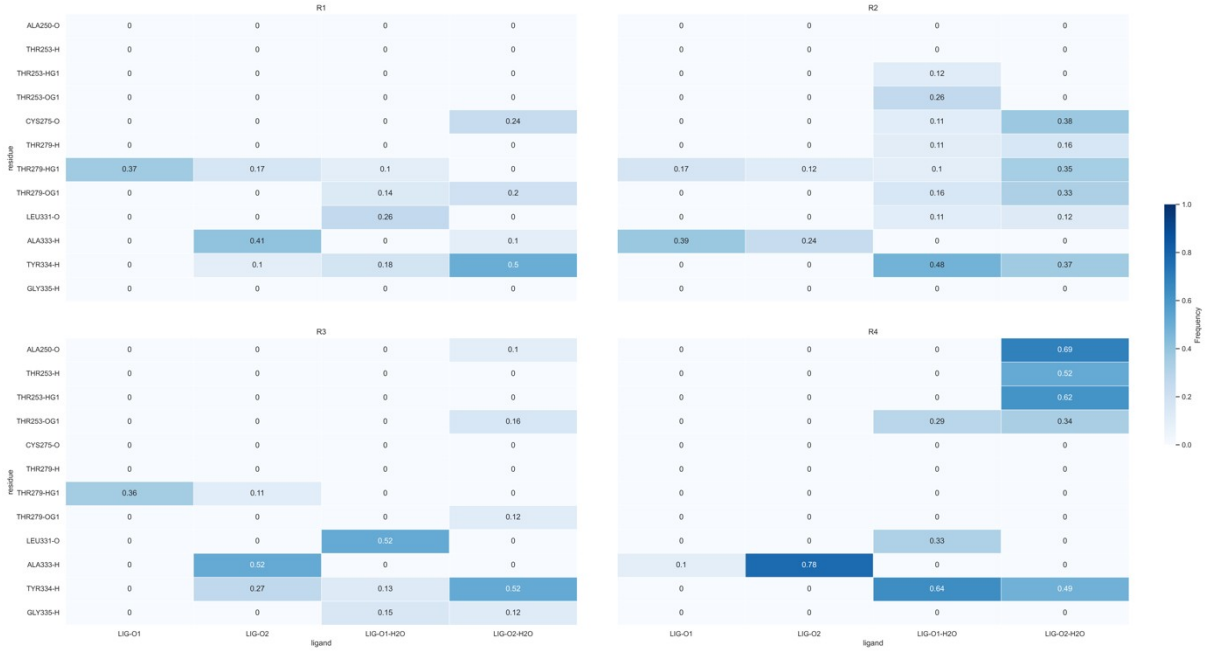


Figure S137. Detailed overview of H-bonds formed between cinnamic acid (bound to Arm X) and PPARα residues. Only interactions with frequency ≥ 0.1 are shown.

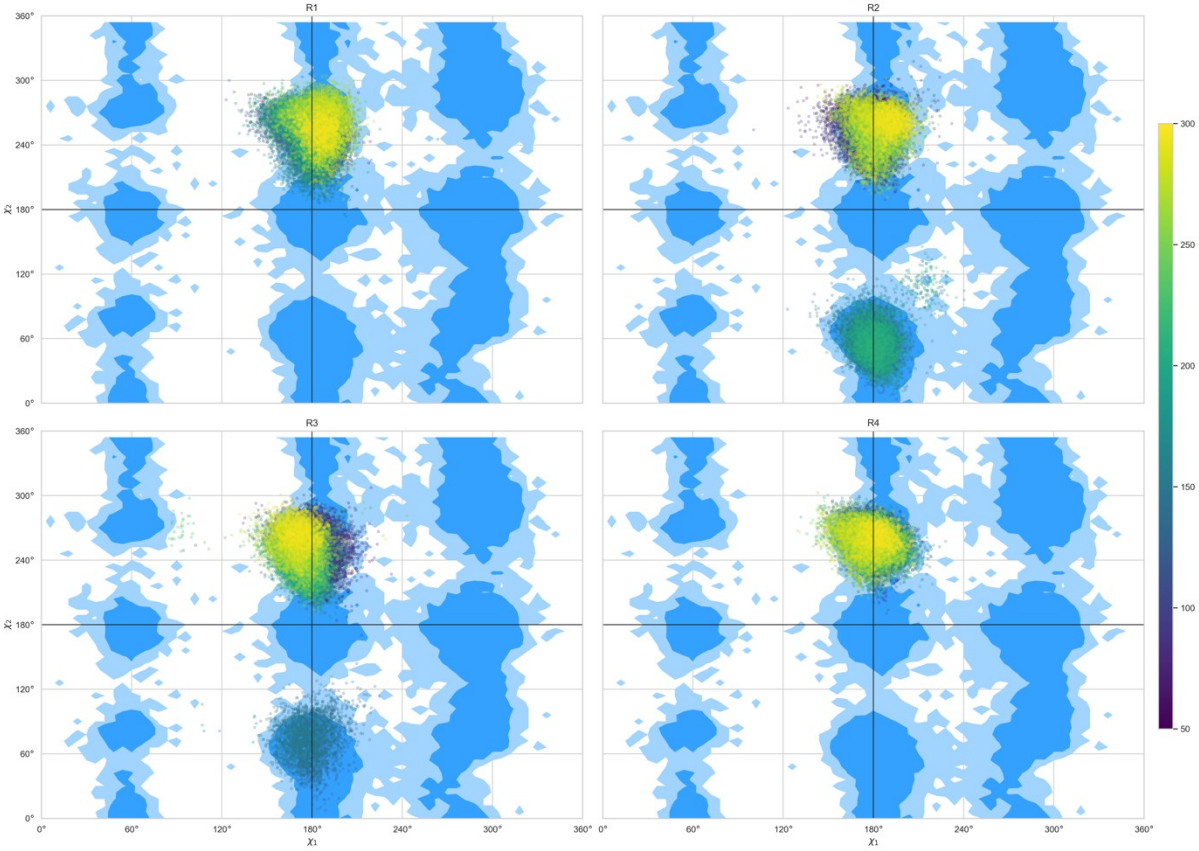


Figure S138. Time evolution of the Chi$_1$ and Chi$_2$ angles of F273. A data point represents the Chi$_1$ - Chi$_2$ angle combination in one particular MD frame. The colour coding indicates simulation time.

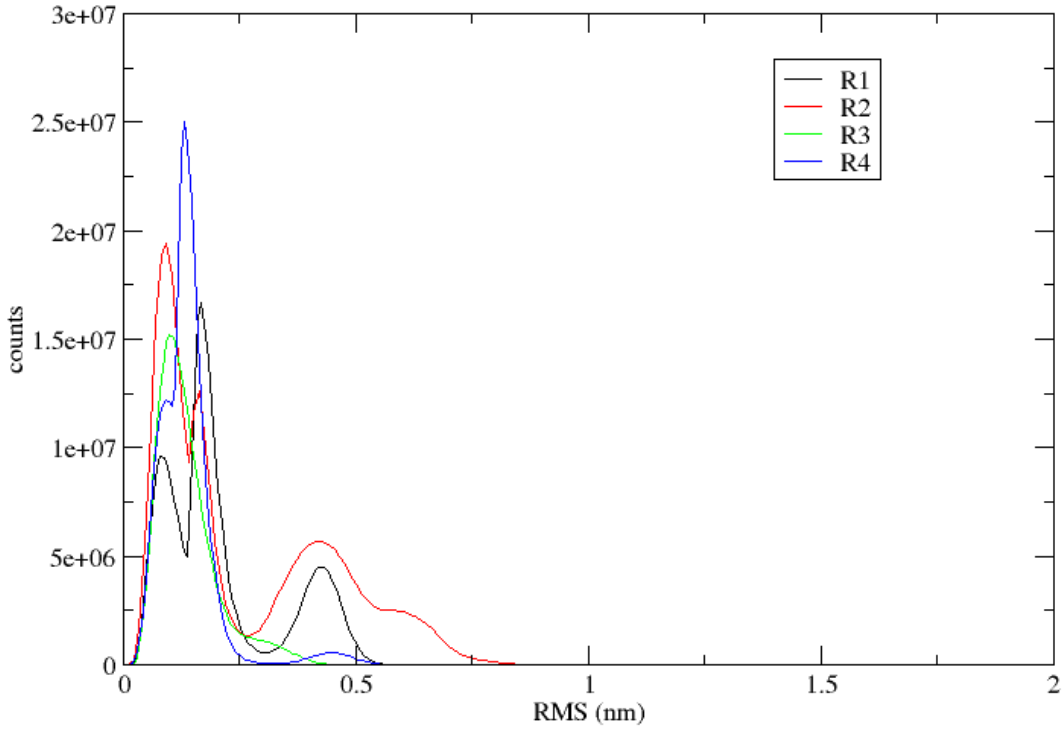


Figure S139. RMSD distribution of cinnamic acid (bound to Center/Arm II/Arm III) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.25 nm was selected to cluster the respective trajectories into representative groups (see Table S17).

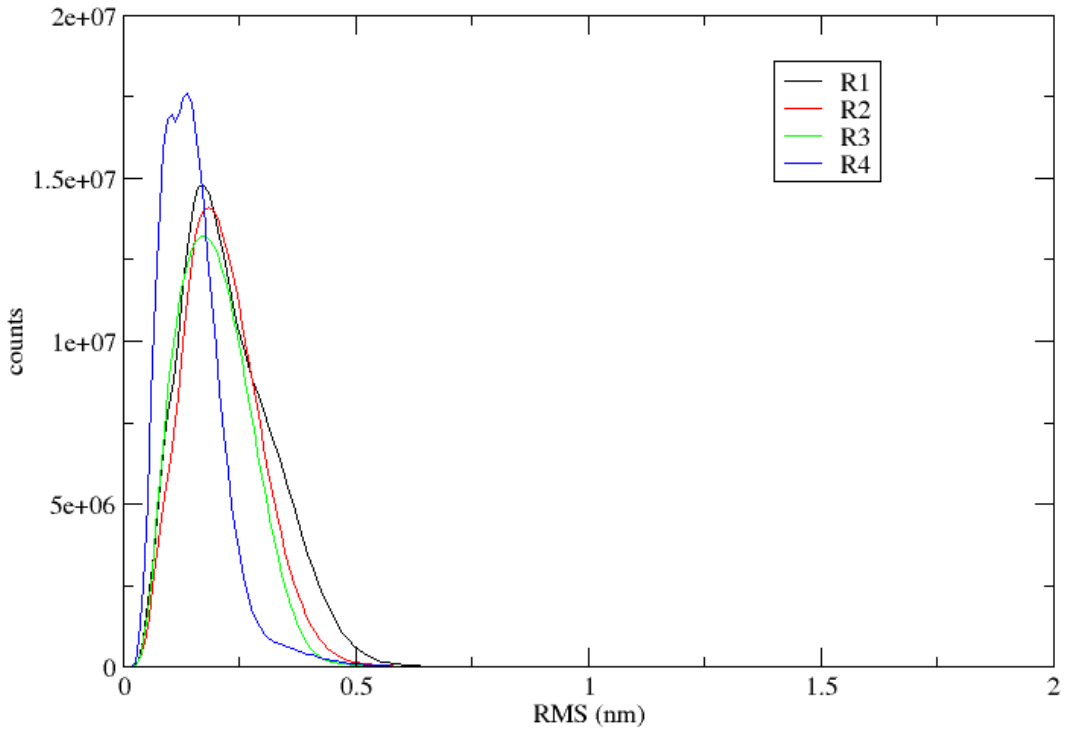


Figure S140. RMSD distribution of cinnamic acid (bound to Arm X) in the four replica MD simulations. Here, frames from 50-300ns were used in order to generate the respective RMSD distance matrices. Based on that distribution, a value of 0.2 nm was selected to cluster the respective trajectories into representative groups (see Table S17).

Table S17. Overview of clusters from simulations of cinnamic acid and PPARα. Clusters are included if they contain more than 1% of MD frames from the part of the trajectory used.

CINNAMIC ACID (2 molecules)			
Pocket	Replica	# clusters	Percentage (%)
CENTER	R1	2	(1) **87.0** (2) **12.8**
	R2	4	(1) **76.6** (2) **14.0** (3) **6.5** (4) **1.0**
	R3	1	(1) **99.7**
	R4	2	(1) **98.4** (2) **1.6**
ARMX	R1	4	(1) **74.2** (2) **12.4** (3) **9.5** (4) **1.6**
	R2	5	(1) **78.7** (2) **9.0** (3) **5.4** (4) **3.2** (5) **1.4**
	R3	3	(1) **86.0** (2) **7.1** (3) **5.4**
	R4	2	(1) **96.5** (2) **1.8**

9 References

1. Mueller, N.P.F., P. Carloni, and M. Alfonso-Prieto, *Molecular determinants of acrylamide neurotoxicity through covalent docking.* Front Pharmacol, 2023. **14**: p. 1125871.
2. Nin-Hill, A., et al., *Photopharmacology of Ion Channels through the Light of the Computational Microscope.* Int J Mol Sci, 2021. **22**(21).
3. Swaen, G.M., et al., *Mortality study update of acrylamide workers.* Occup Environ Med, 2007. **64**(6): p. 396-401.
4. Pennisi, M., et al., *Neurotoxicity of acrylamide in exposed workers.* Int J Environ Res Public Health, 2013. **10**(9): p. 3843-54.
5. Busova, M., et al., *Risk of exposure to acrylamide.* Cent Eur J Public Health, 2020. **28**: p. S43-S46.
6. Mottram, D.S., B.L. Wedzicha, and A.T. Dodson, *Acrylamide is formed in the Maillard reaction.* Nature, 2002. **419**(6906): p. 448-9.
7. Schouten, M.A., S. Tappi, and S. Romani, *Acrylamide in coffee: formation and possible mitigation strategies - a review.* Crit Rev Food Sci Nutr, 2020. **60**(22): p. 3807-3821.
8. Guenther, H., et al., *Acrylamide in coffee: review of progress in analysis, formation and level reduction.* Food Addit Contam, 2007. **24 Suppl 1**: p. 60-70.
9. Reynolds, T., *Acrylamide and cancer: tunnel leak in Sweden prompted studies.* J Natl Cancer Inst, 2002. **94**(12): p. 876-8.
10. Semla, M., et al., *Acrylamide: a common food toxin related to physiological functions and health.* Physiol Res, 2017. **66**(2): p. 205-217.
11. Kumar, J., S. Das, and S.L. Teoh, *Dietary Acrylamide and the Risks of Developing Cancer: Facts to Ponder.* Front Nutr, 2018. **5**: p. 14.
12. EU Commission. *Commission Regulation (EU) 2017/2158 of 20 November 2017 establishing mitigation measures and benchmark levels for the reduction of the presence of acrylamide in food*. 2017 [cited 2023; Available from: http://data.europa.eu/eli/reg/2017/2158/oj.
13. LoPachin, R.M. and T. Gavin, *Molecular mechanism of acrylamide neurotoxicity: lessons learned from organic chemistry.* Environ Health Perspect, 2012. **120**(12): p. 1650-7.
14. Erkekoglu, P. and T. Baydar, *Acrylamide neurotoxicity.* Nutr Neurosci, 2014. **17** (2): p. 49-57.
15. Li, J., et al., *Acrylamide induces locomotor defects and degeneration of dopamine neurons in Caenorhabditis elegans.* J Appl Toxicol, 2016. **36**(1): p. 60-7.
16. Murray, S.M., B.M. Waddell, and C.W. Wu, *Neuron-specific toxicity of chronic acrylamide exposure in C. elegans.* Neurotoxicol Teratol, 2020. **77**: p. 106848.
17. Faria, M., et al., *Acrylamide acute neurotoxicity in adult zebrafish.* Sci Rep, 2018. **8**(1): p. 7918.
18. Raldua, D., et al., *Targeting redox metabolism: the perfect storm induced by acrylamide poisoning in the brain.* Sci Rep, 2020. **10**(1): p. 312.
19. Koutsidis, G., et al., *Investigations on the effect of amino acids on acrylamide, pyrazines, and Michael addition products in model systems.* J Agric Food Chem, 2009. **57**(19): p. 9011-5.
20. Dill, K.A. and J.L. MacCallum, *The protein-folding problem, 50 years on.* Science, 2012. **338**(6110): p. 1042-6.

21. Dill, K.A., et al., *The protein folding problem.* Annu Rev Biophys, 2008. **37**: p. 289-316.
22. Dobson, C.M., *Principles of protein folding, misfolding and aggregation.* Semin Cell Dev Biol, 2004. **15**(1): p. 3-16.
23. Onuchic, J.N. and P.G. Wolynes, *Theory of protein folding.* Curr Opin Struct Biol, 2004. **14**(1): p. 70-5.
24. Pereira, J., et al., *High-accuracy protein structure prediction in CASP14.* Proteins, 2021. **89**(12): p. 1687-1699.
25. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.
26. Thornton, J.M., R.A. Laskowski, and N. Borkakoti, *AlphaFold heralds a data-driven revolution in biology and medicine.* Nat Med, 2021. **27**(10): p. 1666-1669.
27. Samoggia, A. and B. Riedel, *Consumers' Perceptions of Coffee Health Benefits and Motives for Coffee Consumption and Purchasing.* Nutrients, 2019. **11**(3).
28. Pourshahidi, L.K., et al., *A Comprehensive Overview of the Risks and Benefits of Coffee Consumption.* Compr Rev Food Sci Food Saf, 2016. **15**(4): p. 671-684.
29. Alonso-Salces, R.M., et al., *Botanical and geographical characterization of green coffee (Coffea arabica and Coffea canephora): chemometric evaluation of phenolic and methylxanthine contents.* J Agric Food Chem, 2009. **57**(10): p. 4224-35.
30. Sunarharum, W.B., D.J. Williams, and H.E. Smyth, *Complexity of coffee flavor: A compositional and sensory perspective.* Food Research International, 2014. **62**: p. 315-325.
31. Janissen, B. and T. Huynh, *Chemical composition and value-adding applications of coffee industry by-products: A review.* Resources, Conservation and Recycling, 2018. **128**: p. 110-117.
32. Sova, M., *Antioxidant and antimicrobial activities of cinnamic acid derivatives.* Mini Rev Med Chem, 2012. **12**(8): p. 749-67.
33. Ruwizhi, N. and B.A. Aderibigbe, *Cinnamic Acid Derivatives and Their Biological Efficacy.* Int J Mol Sci, 2020. **21**(16).
34. Naveed, M., et al., *Chlorogenic acid (CGA): A pharmacological review and call for further research.* Biomed Pharmacother, 2018. **97**: p. 67-74.
35. Tajik, N., et al., *The potential effects of chlorogenic acid, the main phenolic components in coffee, on health: a comprehensive review of the literature.* Eur J Nutr, 2017. **56**(7): p. 2215-2244.
36. He Fengsheng, et al., *Neurological and Electroneuromyographic Assessment of the Adverse Effects of Acrylamide on Occupationally Exposed Workers.* Scandinavian Journal of Work, Environment & Health, 1989. **15**: p. 125-129.
37. Awoonor-Williams, E. and C.N. Rowley, *Evaluation of Methods for the Calculation of the pKa of Cysteine Residues in Proteins.* J Chem Theory Comput, 2016. **12**(9): p. 4662-73.
38. Roseli, R.B., A.B. Keto, and E.H. Krenske, *Mechanistic aspects of thiol additions to Michael acceptors: Insights from computations.* WIREs Computational Molecular Science, 2022. **e1636**.
39. Boettcher, M.I., et al., *Mercapturic acids of acrylamide and glycidamide as biomarkers of the internal exposure to acrylamide in the general population.* Mutat Res, 2005. **580**(1-2): p. 167-76.
40. Catalgol, B., G. Ozhan, and B. Alpertunga, *Acrylamide-induced oxidative stress in human erythrocytes.* Hum Exp Toxicol, 2009. **28**(10): p. 611-7.

41. Moon, J.K., H.S. Yoo, and T. Shibamoto, *Role of Roasting Conditions in the Level of Chlorogenic Acid Content in Coffee Beans: Correlation with Coffee Acidity.* J Agric Food Chem, 2009. **57**(12): p. 5365–5369.
42. Lu, H., et al., *Chlorogenic acid: A comprehensive review of the dietary sources, processing effects, bioavailability, beneficial properties, mechanisms of action, and future directions.* Compr Rev Food Sci Food Saf, 2020. **19**(6): p. 3130-3158.
43. Monteiro, M., et al., *Chlorogenic acid compounds from coffee are differentially absorbed and metabolized in humans.* J Nutr, 2007. **137**(10): p. 2196-201.
44. Liang, N. and D.D. Kitts, *Role of Chlorogenic Acids in Controlling Oxidative and Inflammatory Stress Conditions.* Nutrients, 2015. **8**(1).
45. Singh, A.K., et al., *Evaluation of antidiabetic activity of dietary phenolic compound chlorogenic acid in streptozotocin induced diabetic rats: Molecular docking, molecular dynamics, in silico toxicity, in vitro and in vivo studies.* Comput Biol Med, 2021. **134**: p. 104462.
46. Jin, U.H., et al., *A phenolic compound, 5-caffeoylquinic acid (chlorogenic acid), is a new type and strong matrix metalloproteinase-9 inhibitor: isolation and identification from methanol extract of Euonymus alatus.* Life Sci, 2005. **77**(22): p. 2760-9.
47. Lee, K., et al., *Chlorogenic acid ameliorates brain damage and edema by inhibiting matrix metalloproteinase-2 and 9 in a rat model of focal cerebral ischemia.* Eur J Pharmacol, 2012. **689**(1-3): p. 89-95.
48. Oboh, G., et al., *Comparative study on the inhibitory effect of caffeic and chlorogenic acids on key enzymes linked to Alzheimer's disease and some pro-oxidant induced oxidative stress in rats' brain-in vitro.* Neurochem Res, 2013. **38** (2): p. 413-9.
49. Mollica, A., et al., *Microwave-assisted extraction, HPLC analysis, and inhibitory effects on carbonic anhydrase I, II, VA, and VII isoforms of 14 blueberry Italian cultivars.* J Enzyme Inhib Med Chem, 2016. **31**(sup4): p. 1-6.
50. Burns, K.A. and J.P. Vanden Heuvel, *Modulation of PPAR activity via phosphorylation.* Biochim Biophys Acta, 2007. **1771**(8): p. 952-60.
51. Papageorgiou, L., et al., *Conserved functional motifs of the nuclear receptor superfamily as potential pharmacological targets.* International Journal of Epigenetics, 2021. **1**(2).
52. Pyper, S.R., et al., *PPARalpha: energy combustion, hypolipidemia, inflammation and cancer.* Nucl Recept Signal, 2010. **8**: p. 1-21.
53. Warden, A., et al., *Localization of PPAR isotypes in the adult mouse and human brain.* Scientific Reports, 2016. **6**.
54. Wojtowicz, S., et al., *The Novel Role of PPAR Alpha in the Brain: Promising Target in Therapy of Alzheimer's Disease and Other Neurodegenerative Disorders.* Neurochemical Research, 2020. **45**(5): p. 972-988.
55. Chandra, S., et al., *Cinnamic acid activates PPARα to stimulate Lysosomal biogenesis and lower Amyloid plaque pathology in an Alzheimer's disease mouse model.* Neurobiol Dis, 2019. **124**: p. 379-395.
56. Prorok, T., et al., *Cinnamic acid protects the nigrostriatum in a mouse model of Parkinson's disease via peroxisome proliferator-activated receptor α.* Neurochem Res, 2019. **44**(4): p. 751-762.
57. Kamata, S., et al., *PPARalpha Ligand-Binding Domain Structures with Endogenous Fatty Acids and Fibrates.* iScience, 2020. **23**(11): p. 101727.
58. Zoete, V., A. Grosdidier, and O. Michielin, *Peroxisome proliferator-activated receptor structures: ligand specificity, molecular switch and interactions with regulators.* Biochim Biophys Acta, 2007. **1771**(8): p. 915-25.

59. Moreno-Santos, I., et al., *Computational and biological evaluation of N-octadecyl-N'-propylsulfamide, a selective PPARalpha agonist structurally related to N-acylethanolamines.* PLoS One, 2014. **9**(3): p. e92195.
60. Xu, H.E., et al., *Molecular recognition of fatty acids by peroxisome proliferator-activated receptors.* Mol Cell, 1999. **3**(3): p. 397-403.
61. Xu, H.E., et al., *Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPARalpha.* Nature, 2002. **415**(6873): p. 813-7.
62. Michalik, L., et al., *Combined simulation and mutagenesis analyses reveal the involvement of key residues for peroxisome proliferator-activated receptor alpha helix 12 dynamic behavior.* J Biol Chem, 2007. **282**(13): p. 9666-9677.
63. Capelli, D., et al., *Structural basis for PPAR partial or full activation revealed by a novel ligand binding mode.* Sci Rep, 2016. **6**: p. 34792.
64. Bruning, J.B., et al., *Partial agonists activate PPARgamma using a helix 12 independent mechanism.* Structure, 2007. **15**(10): p. 1258-71.
65. Bernardes, A., et al., *Molecular mechanism of peroxisome proliferator-activated receptor alpha activation by WY14643: a new mode of ligand recognition and receptor stabilization.* J Mol Biol, 2013. **425**(16): p. 2878-93.
66. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
67. Xu, J., M. McPartlon, and J. Li, *Improved protein structure prediction by deep learning irrespective of co-evolution information.* Nat Mach Intell, 2021. **3**: p. 601-609.
68. AlQuraishi, M., *End-to-End Differentiable Learning of Protein Structure.* Cell Syst, 2019. **8**(4): p. 292-301.e3.
69. Chowdhury, R., et al., *Single-sequence protein structure prediction using a language model and deep learning.* Nat Biotechnol, 2022. **40**(11): p. 1617-1623.
70. Lee, D., et al., *Deep learning methods for 3D structural proteome and interactome modeling.* Curr Opin Struct Biol, 2022. **73**.
71. Majumder, P., *Computational Methods Used in Prediction of Protein Structure*, in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*. 2019. p. 119-133.
72. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
73. Vyas, V.K., et al., *Homology modeling a fast tool for drug discovery: current perspectives.* Indian J Pharm Sci, 2012. **74**(1): p. 1-17.
74. Sutcliffe, M.J., et al., *Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures.* Protein Eng, 1987. **1**(5): p. 377-84.
75. Levitt, M., *Accurate modeling of protein conformation by automatic segment matching.* J Mol Biol, 1992. **226**(2): p. 507-33.
76. Eswar, N., et al., *Comparative protein structure modeling using Modeller.* Curr Protoc Bioinformatics, 2006. **Chapter 5**: p. Unit-5 6.
77. Xiang, Z., *Advances in homology protein structure modeling.* Curr Protein Pept Sci, 2006. **7**(3): p. 217-27.
78. van Gelder, C.W., et al., *A molecular dynamics approach for the generation of complete protein structures from limited coordinate data.* Proteins, 1994. **18**(2): p. 174-85.
79. Rohl, C.A., et al., *Protein structure prediction using Rosetta.* Methods Enzymol, 2004. **383**: p. 66-93.

80. Olivella, M., et al., *Relation between sequence and structure in membrane proteins.* Bioinformatics, 2013. **29**(13): p. 1589-92.
81. Piccoli, S., et al., *Genome-wide Membrane Protein Structure Prediction.* Curr Genomics, 2013. **14**(5): p. 324-9.
82. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-6.
83. Studer, G., et al., *QMEANDisCo-distance constraints applied on model quality estimation.* Bioinformatics, 2020. **36**(6): p. 1765-1771.
84. Studer, G., M. Biasini, and T. Schwede, *Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane).* Bioinformatics, 2014. **30**(17): p. i505-11.
85. David S. Goodsell, A.J.O., *Automated Docking of Substrates to Proteins by Simulated Annealing.* PROTEINS Structure, Function, and Genetics, 1990. **8**: p. 195-202.
86. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.
87. de Vries, S.J., M. van Dijk, and A.M. Bonvin, *The HADDOCK web server for data-driven biomolecular docking.* Nat Protoc, 2010. **5**(5): p. 883-97.
88. Combs, S.A., et al., *Small-molecule ligand docking into comparative models with Rosetta.* Nat Protoc, 2013. **8**(7): p. 1277-98.
89. Friesner, R.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* J Med Chem, 2004. **4 7**(7): p. 1739-49.
90. Rueda, M., et al., *A consensus view of protein dynamics.* Proc Natl Acad Sci U S A, 2007. **104**(3): p. 796-801.
91. Hollingsworth, S.A. and R.O. Dror, *Molecular Dynamics Simulation for All.* Neuron, 2018. **99**(6): p. 1129-1143.
92. Raghavan, B., et al., *Drug Design in the Exascale Era: A Perspective from Massively Parallel QM/MM Simulations.* Chem Rxiv, 2023.
93. Kmiecik, S., et al., *Coarse-Grained Protein Models and Their Applications.* Chem Rev, 2016. **116**(14): p. 7898-936.
94. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhanced sampling techniques in molecular dynamics simulations of biological systems.* Biochim Biophys Acta, 2015. **1850**(5): p. 872-877.
95. Cornell, W.D., et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.* Journal of the American Chemical Society, 2002. **117**(19): p. 5179-5197.
96. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins, 2010. **78**(8): p. 1950-8.
97. He, X., et al., *A fast and high-quality charge model for the next generation general AMBER force field.* J Chem Phys, 2020. **153**(11).
98. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
99. Chen, A.A. and R.V. Pappu, *Parameters of monovalent ions in the AMBER-99 forcefield: assessment of inaccuracies and proposed improvements.* J Phys Chem B, 2007. **111**(41): p. 11884-7.
100. Daura, X., et al., *Peptide Folding: When Simulation Meets Experiment.* Angewandte Chemie International Edition, 1999. **38**(1-2): p. 236-240.
101. Rodrigues, J.P., et al., *Clustering biomolecular complexes by residue contacts similarity.* Proteins, 2012. **80**(7): p. 1810-7.

102. Genheden, S. and U. Ryde, *The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities.* Expert Opin Drug Discov, 2015. **10**(5): p. 449-61.
103. Srinivasan, J., et al., *Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices.* Journal of the American Chemical Society, 1998. **120**(37): p. 9401-9409.
104. Tan, C., Y.H. Tan, and R. Luo, *Implicit nonpolar solvent models.* J Phys Chem B, 2007. **111**(42): p. 12263-74.
105. Valdes-Tresanco, M.S., et al., *gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS.* J Chem Theory Comput, 2021. **17**(10): p. 6281-6291.
106. Kim, S., et al., *PubChem 2019 update: improved access to chemical data.* Nucleic Acids Res, 2019. **47**(D1): p. D1102-D1109.
107. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
108. Burley, S.K., et al., *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 Res, 2021. **49**(D1): p. D437-D451.
109. Bienert, S., et al., *The SWISS-MODEL Repository - new features and functionality.* Nucleic Acids Res, 2017. **45**(D1): p. D313-D319.
110. Waterhouse, A., et al., *SWISS-MODEL: homology modelling of protein structures and complexes* Nucleic Acids Res, 2018. **46**(W1): p. W296-W303.
111. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
112. Steinegger, M., et al., *HH-suite3 for fast remote homology detection and deep protein annotation.* BMC Bioinformatics, 2019. **20**(1): p. 473.
113. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(1): p. 12-21.
114. Williams, C.J., et al., *MolProbity: More and better reference data for improved all-atom structure validation.* Protein Sci, 2018. **27**(1): p. 293-315.
115. Foloppe, N., et al., *Structure, dynamics and electrostatics of the active site of glutaredoxin 3 from Escherichia coli: comparison with functionally related proteins.* J Mol Biol, 2001. **310**(2): p. 449-70.
116. Lutolf, M.P., et al., *Systematic modulation of Michael-type reactivity of thiols through the use of charged amino acids.* Bioconjug Chem, 2001. **12**(6): p. 1051-6.
117. LoPachin, R.M., et al., *Structure-toxicity analysis of type-2 alkenes: in vitro neurotoxicity.* Toxicol Sci, 2007. **95**(1): p. 136-46.
118. Nair, D.P., et al., *The Thiol-Michael Addition Click Reaction: A Powerful and Widely Used Tool in Materials Chemistry.* Chemistry of Materials, 2013. **26**(1): p. 724-744.
119. van Zundert, G.C.P., et al., *The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.* J Mol Biol, 2016. **428**(4): p. 720-725.
120. HADDOCK developer team. *HADDOCK covalent docking tutorial.* 2018 [cited 2023 30th May]; Available from: https://www.bonvinlab.org/education/biomolecular-simulations-2018/HADDOCK_tutorial/.
121. Kurkcuoglu, Z., et al., *Performance of HADDOCK and a simple contact-based protein-ligand binding affinity predictor in the D3R Grand Challenge 2.* J Comput Aided Mol Des, 2018. **32**(1): p. 175-185.

122. Koukos, P.I., L.C. Xue, and A.M.J.J. Bonvin, *Protein-ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3.* J Comput Aided Mol Des, 2019. **33**(1): p. 83-91.
123. Karaca, E., et al., *Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server.* Mol Cell Proteomics, 2010. **9**(8): p. 1784-94.
124. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8.
125. Bouysset, C. and S. Fiorucci, *ProLIF: a library to encode molecular interactions as fingerprints.* J Cheminform, 2021. **13**(1): p. 72.
126. Mazmanian, K., et al., *Preferred Hydrogen-Bonding Partners of Cysteine: Implications for Regulating Cys Functions.* J Phys Chem B, 2016. **120**(39): p. 10288-10296.
127. Ha, H.-J., et al., *Fluorescence turn-on probe for biothiols: intramolecular hydrogen bonding effect on the Michael reaction.* Tetrahedron, 2011. **67**(40): p. 7759-7762.
128. Weber, B.W., et al., *The mechanism of the amidases: mutating the glutamate adjacent to the catalytic triad inactivates the enzyme due to substrate mispositioning.* J Biol Chem, 2013. **288**(40): p. 28514-23.
129. Noort, D., A. Fidder, and A.G. Hulst, *Modification of human serum albumin by acrylamide at cysteine-34: a basis for a rapid biomonitoring procedure.* Arch Toxicol, 2003. **77**(9): p. 543-5.
130. Tong, G.C., W.K. Cornwell, and G.E. Means, *Reactions of acrylamide with glutathione and serum albumin.* Toxicol Lett, 2004. **147**(2): p. 127-31.
131. Ferguson, S.A., et al., *Preweaning behaviors, developmental landmarks, and acrylamide and glycidamide levels after pre- and postnatal acrylamide treatment in rats.* Neurotoxicol Teratol, 2010. **32**(3): p. 373-82.
132. Dixit, R., H. Mukhtar, and P.K. Seth, *In vitro inhibition of alcohol dehydrogenase by acrylamide: interaction with enzyme-SH groups.* Toxicol Lett, 1981. **7**(6): p. 487-92.
133. Hansch, C., et al., *A quantitative structure-activity relationship and molecular graphics analysis of hydrophobic effects in the interactions of inhibitors with alcohol dehydrogenase.* J Med Chem, 1986. **29**(5): p. 615-20.
134. Dobryszycki, P., et al., *Effect of acrylamide on aldolase structure. I. Induction of intermediate states.* Biochim Biophys Acta, 1999. **1431**(2): p. 338-50.
135. Matsuoka, M., H. Matsumura, and H. Igisu, *Creatine kinase activities in brain and blood:possible neurotoxic indicator of acrylamide intoxication.* Occup Environ Med, 1996. **53**(7): p. 468-71.
136. Lin, L., et al., *Determination of the catalytic site of creatine kinase by site-directed mutagenesis.* Biochim Biophys Acta, 1994. **1206**(1): p. 97-104.
137. Meng, F.G., H.W. Zhou, and H.M. Zhou, *Effects of acrylamide on creatine kinase from rabbit muscle.* Int J Biochem Cell Biol, 2001. **33**(11): p. 1064-70.
138. Sheng, Q., et al., *Effects of acrylamide on the activity and structure of human brain creatine kinase.* Int J Mol Sci, 2009. **10**(10): p. 4210-4222.
139. LoPachin, R.M. and D.S. Barber, *Synaptic cysteine sulfhydryl groups as targets of electrophilic neurotoxicants.* Toxicol Sci, 2006. **94**(2): p. 240-55.
140. Howland, R.D., et al., *The etiology of toxic peripheral neuropathies: In vitro effects of acrylamide and 2,5-hexanedione on brain enolase and other glycolytic enzymes.* Brain Research, 1980. **202**(1): p. 131-142.
141. Hogervorst, J.G., et al., *The carcinogenicity of dietary acrylamide intake: a comparative discussion of epidemiological and experimental animal research.* Crit Rev Toxicol, 2010. **40**(6): p. 485-512.

142. Aliau, S., et al., *Cysteine 530 of the human estrogen receptor alpha is the main covalent attachment site of 11beta-(aziridinylalkoxyphenyl)estradiols.* Biochemistry, 1999. **38**(45): p. 14752-62.
143. Martyniuk, C.J., et al., *Molecular mechanism of glyceraldehyde-3-phosphate dehydrogenase inactivation by alpha,beta-unsaturated carbonyl derivatives.* Chem Res Toxicol, 2011. **24**(12): p. 2302-11.
144. Basile, A., et al., *Proteomic approach for the analysis of acrylamide-hemoglobin adducts. Perspectives for biological monitoring.* J Chromatogr A, 2008. **1215**(1-2): p. 74-81.
145. Feng, C.H. and C.Y. Lu, *Modification of major plasma proteins by acrylamide and glycidamide: Preliminary screening by nano liquid chromatography with tandem mass spectrometry.* Anal Chim Acta, 2011. **684**(1-2): p. 80-6.
146. Friedman, M.A., et al., *Inhibition of rat testicular nuclear kinesins (krp2; KIFC5A) by acrylamide as a basis for establishing a genotoxicity threshold.* J Agric Food Chem, 2008. **56**(15): p. 6024-30.
147. Sciandrello, G., et al., *Acrylamide catalytically inhibits topoisomerase II in V79 cells.* Toxicol In Vitro, 2010. **24**(3): p. 830-4.
148. Fasano, M., et al., *The extraordinary ligand binding properties of human serum albumin.* IUBMB Life, 2005. **57**(12): p. 787-96.
149. Carter, D.C. and J.X. Ho, *Structure of serum albumin.* Adv Protein Chem, 1994. **45**: p. 153-203.
150. Pedersen, A.O. and J. Jacobsen, *Reactivity of the thiol group in human and bovine albumin at pH 3--9, as measured by exchange with 2,2'-dithiodipyridine.* Eur J Biochem, 1980. **106**(1): p. 291-5.
151. Sampath, V., X.J. Zhao, and W.S. Caughey, *Anesthetic-like interactions of nitric oxide with albumin and hemeproteins. A mechanism for control of protein function.* J Biol Chem, 2001. **276**(17): p. 13635-43.
152. Wang, P.F., et al., *Exploring the role of the active site cysteine in human muscle creatine kinase.* Biochemistry, 2006. **45**(38): p. 11464-72.
153. Lü, Z.R., et al., *The effects of acrylamide on brain creatine kinase: inhibition kinetics and computational docking simulation.* Int J Biol Macromol, 2009. **44**(2): p. 128-32.
154. Alberts, G.L., J.F. Pregenzer, and W.B. Im, *Contributions of cysteine 114 of the human D3 dopamine receptor to ligand binding and sensitivity to external oxidizing agents.* Br J Pharmacol, 2009. **125**(4): p. 705-10.
155. Kooistra, A.J., et al., *GPCRdb in 2021: integrating GPCR sequence, structure and function.* Nucleic Acids Res, 2021. **49**(D1): p. D335-D343.
156. Michino, M., et al., *What can crystal structures of aminergic receptors tell us about designing subtype-selective ligands?* Pharmacol Rev, 2015. **67**(1): p. 198-213.
157. Giros, B. and M.G. Caron, *Molecular characterization of the dopamine transporter.* Trends Pharmacol Sci, 1993. **14**(2): p. 43-9.
158. Park, S.U., et al., *Peroxynitrite inactivates the human dopamine transporter by modification of cysteine 342: potential mechanism of neurotoxicity in dopamine neurons.* J Neurosci, 2002. **22**(11): p. 4399-405.
159. Chen, N. and M.E. Reith, *Structure and function of the dopamine transporter.* Eur J Pharmacol, 2000. **405**(1-3): p. 329-39.
160. Pidathala, S., et al., *Structural basis of norepinephrine recognition and transport inhibition in neurotransmitter transporters.* Nat Commun, 2021. **12**(1): p. 2199.
161. Coleman, J.A., et al., *Serotonin transporter-ibogaine complexes illuminate mechanisms of inhibition and transport.* Nature, 2019. **569**(7754): p. 141-145.

162. McHugh, P.C. and D.A. Buckley, *The structure and function of the dopamine transporter and its role in CNS diseases.* Vitam Horm, 2015. **98**: p. 339-69.
163. Jayaramayya, K., et al., *Unraveling correlative roles of dopamine transporter (DAT) and Parkin in Parkinson's disease (PD) - A road to discovery?* Brain Res Bull, 2020. **157**: p. 169-179.
164. Whitehead, R.E., et al., *Reaction of oxidized dopamine with endogenous cysteine residues in the human dopamine transporter.* J Neurochem, 2001. **76**(4): p. 1242-51.
165. Whiteheart, S.W., et al., *N-ethylmaleimide-sensitive fusion protein: a trimeric ATPase whose hydrolysis of ATP is required for membrane fusion.* J Cell Biol, 1994. **126**(4): p. 945-54.
166. Tanii, H. and K. Hashimoto, *Effect of acrylamide and related compounds on glycolytic enzymes of rat brain.* Toxicol Lett, 1985. **26**(1): p. 79-84.
167. Soukri, A., et al., *Role of the histidine 176 residue in glyceraldehyde-3-phosphate dehydrogenase as probed by site-directed mutagenesis.* Biochemistry, 1989. **28**(6): p. 2586-92.
168. Ahmed, M.H., M.S. Ghatge, and M.K. Safo, *Hemoglobin: Structure, Function and Allostery.* Subcell Biochem, 2020. **94**: p. 345-382.
169. Barber, D.S., et al., *Metabolism, toxicokinetics and hemoglobin adduct formation in rats following subacute and subchronic acrylamide dosing.* Neurotoxicology, 2001. **22**(3): p. 341-53.
170. Perutz, M.F., *Regulation of oxygen affinity of hemoglobin: influence of structure of the globin on the heme iron.* Annu Rev Biochem, 1979. **48**: p. 327-86.
171. Hwang, P.K. and J. Greer, *Interaction between hemoglobin subunits in the hemoglobin . haptoglobin complex.* Journal of Biological Chemistry, 1980. **255**(7): p. 3038-3041.
172. Hoyle, J., et al., *Localization of human and mouse N-ethylmaleimide-sensitive factor (NSF) gene: a two-domain member of the AAA family that is involved in membrane fusion.* Mamm Genome, 1996. **7**(11): p. 850-2.
173. May, A.P., S.W. Whiteheart, and W.I. Weis, *Unraveling the mechanism of the vesicle transport ATPase NSF, the N-ethylmaleimide-sensitive factor.* J Biol Chem, 2001. **276**(25): p. 21991-4.
174. Matsushita, K., et al., *Nitric oxide regulates exocytosis by S-nitrosylation of N-ethylmaleimide-sensitive factor.* Cell, 2003. **115**(2): p. 139-50.
175. Barber, D.S. and R.M. LoPachin, *Proteomic analysis of acrylamide-protein adduct formation in rat brain synaptosomes.* Toxicol Appl Pharmacol, 2004. **201** (2): p. 120-36.
176. Zhao, C., J.T. Slevin, and S.W. Whiteheart, *Cellular functions of NSF: not just SNAPs and SNAREs.* FEBS Lett, 2007. **581**(11): p. 2140-9.
177. Zhao, M., et al., *Mechanistic insights into the recycling machine of the SNARE complex.* Nature, 2015. **518**(7537): p. 61-7.
178. Cidon, S. and T.S. Sihar, *Characterization of a H+-ATPase in Rat Brain Synaptic Vesicle.* The Journal of Biological Chemistry, 1989. **264**(14): p. 8281-8288.
179. Feng, Y. and M. Forgac, *A Novel Mechanism for Regulation of Vacular Acidification.* The Journal of Biological Chemistry, 1992. **267**(28): p. 19769-19772.
180. Hunt, I.E. and D. Sanders, *The Kinetics of N-Ethylmaleimide Inhibition of a Vacuolar H+-ATPase and Determination of Nucleotide Dissociation Constants.* Plant Physiol, 1996. **110**(1): p. 97-103.

181. Feng, Y. and M. Forgac, *Inhibition of vacuolar H(+)-ATPase by disulfide bond formation between cysteine 254 and cysteine 532 in subunit A.* Journal of Biological Chemistry, 1994. **269**(18): p. 13224-13230.
182. Dennehy, M.K., et al., *Cytosolic and nuclear protein targets of thiol-reactive electrophiles.* Chem Res Toxicol, 2006. **19**(1): p. 20-9.
183. King, J.L. and T.H. Jukes, *Non-Darwinian evolution.* Science, 1969. **164**(3881): p. 788-98.
184. Calakos, N. and R.H. Scheller, *Synaptic vesicle biogenesis, docking, and fusion: a molecular description.* Physiol Rev, 1996. **76**(1): p. 1-29.
185. Anandakrishnan, R., B. Aguilar, and A.V. Onufriev, *H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W537-41.
186. Rostkowski, M., et al., *Graphical analysis of pH-dependent properties of proteins predicted using PROPKA.* BMC Struct Biol, 2011. **11**: p. 6.
187. Song, Y., J. Mao, and M.R. Gunner, *MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling.* J Comput Chem, 2009. **30**(14): p. 2231-47.
188. Pahari, S., L. Sun, and E. Alexov, *PKAD: a database of experimentally measured pKa values of ionizable groups in proteins.* Database (Oxford), 2019. **2019**.
189. Roos, G., N. Foloppe, and J. Messens, *Understanding the pK(a) of redox cysteines: the key role of hydrogen bonding.* Antioxid Redox Signal, 2013. **18**(1): p. 94-127.
190. Soylu, I. and S.M. Marino, *Cy-preds: An algorithm and a web service for the analysis and prediction of cysteine reactivity.* Proteins, 2016. **84**(2): p. 278-91.
191. Soylu, I. and S.M. Marino, *Cpipe: a comprehensive computational platform for sequence and structure-based analyses of Cysteine residues.* Bioinformatics, 2017. **33**(15): p. 2395-2396.
192. Li, S., et al., *pCysMod: Prediction of Multiple Cysteine Modifications Based on Deep Learning Framework.* Front Cell Dev Biol, 2021. **9**: p. 617366.
193. Mondal, D. and A. Warshel, *Exploring the Mechanism of Covalent Inhibition: Simulating the Binding Free Energy of alpha-Ketoamide Inhibitors of the Main Protease of SARS-CoV-2.* Biochemistry, 2020. **59**(48): p. 4601-4608.
194. Luo, Y.L., *Mechanism-Based and Computational-Driven Covalent Drug Design.* J Chem Inf Model, 2021. **61**(11): p. 5307-5311.
195. Mihalovits, L.M., G.G. Ferenczy, and G.M. Keserű, *The role of quantum chemistry in covalent inhibitor design.* International Journal of Quantum Chemistry, 2021. **122**(8).
196. Ahdritz, G., et al., *OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization.* bioRxiv, 2022.
197. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
198. Dominguez, C., R. Boelens, and A.M. Bonvin, *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.* J Am Chem Soc, 2003. **125**(7): p. 1731-7.
199. Scarpino, A., G.G. Ferenczy, and G.M. Keseru, *Comparative Evaluation of Covalent Docking Tools.* J Chem Inf Model, 2018. **58**(7): p. 1441-1458.
200. Borisek, J., et al., *Development of N-(Functionalized benzoyl)-homocycloleucyl-glycinonitriles as Potent Cathepsin K Inhibitors.* J Med Chem, 2015. **58**(17): p. 6928-37.

201. Zhu, K., et al., *Docking covalent inhibitors: a parameter free approach to pose prediction and scoring.* J Chem Inf Model, 2014. **54**(7): p. 1932-40.
202. Warshel, A. and M. Levitt, *Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme.* J Mol Biol, 1976. **103**(2): p. 227-49.
203. Warshel, A., *Computer modeling of chemical reations in enzymes and solutions* . 1991: Wiley.
204. Kim, S., et al., *PubChem 2023 update.* Nucleic Acids Res, 2023. **51**(D1): p. D1373-D1380.
205. Bienfait, B. and P. Ertl, *JSME: a free molecule editor in JavaScript.* J Cheminform, 2013. **5**: p. 24.
206. Hawkins, P.C., et al., *Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database.* J Chem Inf Model, 2010. **50**(4): p. 572-84.
207. Farid, R., et al., *New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies.* Bioorg Med Chem, 2006. **14**(9): p. 3160-73.
208. Sherman, W., H.S. Beard, and R. Farid, *Use of an induced fit receptor structure in virtual screening.* Chem Biol Drug Des, 2006. **67**(1): p. 83-4.
209. Sherman, W., et al., *Novel procedure for modeling ligand/receptor induced fit effects.* J Med Chem, 2006. **49**(2): p. 534-53.
210. Schrödinger, L., New York, NY, 2023, *LigPrep*. 2021.
211. Shelley, J.C., et al., *Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules.* J Comput Aided Mol Des, 2007. **21**(12): p. 681-91.
212. Greenwood, J.R., et al., *Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution.* J Comput Aided Mol Des, 2010. **24**(6-7): p. 591-604.
213. Sastry, G.M., et al., *Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments.* J Comput Aided Mol Des, 2013. **27** (3): p. 221-34.
214. Hawkins, P.C., A.G. Skillman, and A. Nicholls, *Comparison of shape-matching and docking as virtual screening tools.* J Med Chem, 2007. **50**(1): p. 74-82.
215. Malde, A.K., et al., *An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0.* J Chem Theory Comput, 2011. **7**(12): p. 4026-37.
216. Vangone, A., et al., *Large-scale prediction of binding affinity in protein-small ligand complexes: the PRODIGY-LIG web server.* Bioinformatics, 2019. **35**(9): p. 1585-1587.
217. Berendsen, H.J.C., D. van der Spoel, and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation.* Computer Physics Communications, 1995. **91**(1-3): p. 43-56.
218. Sousa da Silva, A.W. and W.F. Vranken, *ACPYPE - AnteChamber PYthon Parser interfacE.* BMC Res Notes, 2012. **5**: p. 367.
219. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath.* The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
220. Parrinello, M. and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method.* Journal of Applied Physics, 1981. **52**(12): p. 7182-7190.
221. Michaud-Agrawal, N., et al., *MDAnalysis: a toolkit for the analysis of molecular dynamics simulations.* J Comput Chem, 2011. **32**(10): p. 2319-27.

222. Mysinger, M.M., et al., *Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking.* J Med Chem, 2012. **55**(14): p. 6582-94.
223. Daina, A., O. Michielin, and V. Zoete, *SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules.* Sci Rep, 2017. **7**: p. 42717.
224. Lou, L.L. and J.C. Martin, *Selected Thoughts on Hydrophobicity in Drug Design.* Molecules, 2021. **26**(4).
225. Guterres, H. and W. Im, *Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations.* J Chem Inf Model, 2020. **60**(4): p. 2189-2198.
226. *Reliability and reproducibility checklist for molecular dynamics simulations.* Commun Biol, 2023. **6**(1): p. 268.
227. Basith, S., et al., *A Molecular Dynamics Approach to Explore the Intramolecular Signal Transduction of PPAR-alpha.* Int J Mol Sci, 2019. **20**(7).
228. Sullivan, H.J., et al., *To Probe Full and Partial Activation of Human Peroxisome Proliferator-Activated Receptors by Pan-Agonist Chiglitazar Using Molecular Dynamics Simulations.* PPAR Res, 2020. **2020**: p. 5314187.
229. Berrabah, W., et al., *Control of nuclear receptor activities in metabolism by post-translational modifications.* FEBS Lett, 2011. **585**(11): p. 1640-50.
230. Roy, A., et al., *Identification and characterization of PPARalpha ligands in the hippocampus.* Nat Chem Biol, 2016. **12**(12): p. 1075-1083.
231. Sergeeva, O.A., et al., *OLHA (N(alpha)-oleoylhistamine) modulates activity of mouse brain histaminergic neurons.* Neuropharmacology, 2022. **215**: p. 109167.
232. Liu, M., et al., *Molecular recognition of agonist and antagonist for peroxisome proliferator-activated receptor-alpha studied by molecular dynamics simulations.* Int J Mol Sci, 2014. **15**(5): p. 8743-52.
233. Laio, A. and M. Parrinello, *Escaping free-energy minima.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12562-6.
234. Baler, K., et al., *Electrostatic unfolding and interactions of albumin driven by pH changes: a molecular dynamics study.* J Phys Chem B, 2014. **118**(4): p. 921-30.
235. Aquino Neto, S., et al., *The kinetic behavior of dehydrogenase enzymes in solution and immobilized onto nanostructured carbon platforms.* Process Biochemistry, 2011. **46**(12): p. 2347-2352.
236. Eagles, P.A. and M. Iqbal, *A comparative study of aldolase from human muscle and liver.* Biochem J, 1973. **133**(3): p. 429-39.
237. Bernt, E., W. Gruber, and G. Szasz, *Creatine kinase in serum: 1. Determination of optimum reaction conditions.* Clinical Chemistry, 1976. **22**(5): p. 650-656.
238. Kapolka, N.J., et al., *Proton-gated coincidence detection is a common feature of GPCR signaling.* Proc Natl Acad Sci U S A, 2021. **118**(28).
239. Berfield, J.L., L.C. Wang, and M.E. Reith, *Which form of dopamine is the substrate for the human dopamine transporter: the cationic or the uncharged species?* J Biol Chem, 1999. **274**(8): p. 4876-82.
240. Shimizu, A., F. Suzuki, and K. Kato, *Characterization of alpha alpha, beta beta, gamma gamma and alpha gamma human enolase isozymes, and preparation of hybrid enolases (alpha gamma, beta gamma and alpha beta) from homodimeric forms.* Biochim Biophys Acta, 1983. **748**(2): p. 278-84.
241. de Oliveira, V.M., et al., *pH and the Breast Cancer Recurrent Mutation D538G Affect the Process of Activation of Estrogen Receptor alpha.* Biochemistry, 2022. **61**(6): p. 455-463.

242. Heinz, F. and B. Freimuller, *Glyceraldehyde-3-phosphate dehydrogenase from human tissues.* Methods Enzymol, 1982. **89 (Pt D)**: p. 301-5.
243. Imai, K. and T. Yonetani, *PH dependence of the Adair constants of human hemoglobin. Nonuniform contribution of successive oxygen bindings to the alkaline Bohr effect.* Journal of Biological Chemistry, 1975. **250**(6): p. 2227-2231.
244. Riera, F. and A. Álvarez, *Influence of temperature and pH on the antigen-binding capacity of immunoglobulin G in cheese whey derived from hyper-immune milk.* International Dairy Journal, 2014. **37**(2): p. 111-116.
245. Verhey, K.J., et al., *Light chain-dependent regulation of Kinesin's interaction with microtubules.* J Cell Biol, 1998. **143**(4): p. 1053-66.
246. Selby, C., *Sex hormone binding globulin: origin, function and clinical significance.* Ann Clin Biochem, 1990. **27 (Pt 6)**: p. 532-41.
247. Gardiner, L.P., et al., *The N-terminal domain of human topoisomerase IIalpha is a DNA-dependent ATPase.* Biochemistry, 1998. **37**(48): p. 16997-7004.
248. Osheroff, N., E. Shelton, and D. Brutlag, *DNA topoisomerase II from Drosophila melanogaster. Relaxation of supercoiled DNA.* 1983. **258**(15): p. 9536-43.
249. Kakinuma, Y., Y. Oshumi, and Y. Anraku, *Properties of H+-translocating adenosine triphosphatase in vacuolar membranes of Saccharomyces cerevisiae.* 1981. **256**(21): p. 10859-63.
250. RCSB PDB Ligand Summary. *Propionamide*. 2023 [cited 2023; Available from: https://www.rcsb.org/ligand/ROP.
251. Raj, H., et al., *Engineering methylaspartate ammonia lyase for the asymmetric synthesis of unnatural amino acids.* Nat Chem, 2012. **4**(6): p. 478-84.
252. Groftehauge, M.K., et al., *Crystal Structure of a Hidden Protein, YcaC, a Putative Cysteine Hydrolase from Pseudomonas aeruginosa, with and without an Acrylamide Adduct.* Int J Mol Sci, 2015. **16**(7): p. 15971-84.
253. Eklund, H., et al., *Three-dimensional structure of horse liver alcohol dehydrogenase at 2-4 A resolution.* J Mol Biol, 1976. **102**(1): p. 27-59.
254. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
255. Ishii, T. and K. Uchida, *Induction of reversible cysteine-targeted protein oxidation by an endogenous electrophile 15-deoxy-delta12,14-prostaglandin J2.* Chem Res Toxicol, 2004. **17**(10): p. 1313-22.
256. Liu, H. and K. May, *Disulfide bond structures of IgG molecules: structural variations, chemical modifications and possible impacts to stability and biological function.* MAbs, 2012. **4**(1): p. 17-23.
257. Sickles, D.W., et al., *Direct effect of the neurotoxicant acrylamide on kinesin-based microtubule motility.* Journal of Neuroscience Research, 1996. **46**(1): p. 7-17.
258. Miki, H., et al., *All kinesin superfamily protein, KIF, genes in mouse and human.* Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7004-11.
259. Talapatra, S.K., B. Harker, and J.P. Welburn, *The C-terminal region of the motor protein MCAK controls its structure and activity through a conformational switch.* Elife, 2015. **4**.
260. Förster, T., et al., *2-Sulfonylpyrimidines Target the Kinesin HSET via Cysteine Alkylation.* European Journal of Organic Chemistry, 2019. **2019**(31-32): p. 5486-5496.
261. Park, H.W., et al., *Structural basis of small molecule ATPase inhibition of a human mitotic kinesin motor protein.* Sci Rep, 2017. **7**(1): p. 15121.

262. Wong, A.S., et al., *Rabbit sex hormone-binding globulin: expression in the liver and testis during postnatal development and structural characterization by truncated proteins.* Int J Androl, 2001. **24**(3): p. 165-74.

Anhang 3 zur Promotionsordnung

aus dem Institut für
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Frau Dr. Mercedes Alfonso-Prieto

2. Prof. Dr. Patricia Hidalgo

Tag der mündlichen Prüfung: 22.05.2024