Cosolvent and Dynamic Effects in Binding Pocket Search by Docking Simulations

Péter Bernát SZABÓ

dr. J. J. Nogueira, promotor
Prof. dr. J. Harvey, co-promotor
dr. F. S. Zariquiey, mentor

Thesis presented in partial fulfillment of the requirements for the degree of Master of Science in Theoretical Chemistry and Computational Modelling.

June 2021

Published in-house, Péter Bernát Szabó, Celestijnenlaan 200F box 2404, B-3001 Leuven (Belgium)

Preface

I am enormously grateful to my promoters and mentors, Juanjo, Quico and Prof. Harvey, for their continuous support throughout this somewhat unconventional, completely online thesis project. I would also like to thank my supervisors, Prof. Escudero and Prof. Loreau, for their time spent reviewing this work. I am thankful to my family, for providing the emotional and financial support necessary to complete my MSc in a foreign country. Finally, I would like to thank Dóri for making my life happier day after day.

Szabó P. Bernát

Abstract

In the last decades, the focus of many researchers in the field of drug discovery and development has shifted from experimental techniques to computational methods. Perhaps the most important applications of such methods lie in finding or designing small molecule drugs, interacting with a biologically relevant protein. Traditionally, the bulk of these approaches considered a single, frozen conformation of the target protein, into which they tried to insert the candidate ligands in millions of different poses, in order to find the one that exhibits the most favourable binding interaction. However, with more and more results indicating the importance of conformational changes in the protein during the ligand uptake process, the single, rigid protein conformation considered in these calculations can be inadequate. To remedy this problem, it was suggested that an ensemble of protein conformations could be utilised during the search for binders, which could improve the description of protein dynamics. In the present work, a specialised molecular dynamics simulation protocol is evaluated for the generation of this ensemble of protein structures. In these simulations, the protein is solvated in a mixture of water and some hydrophobic cosolvent. The introduced cosolvent molecules are expected to interact with the protein in a similar fashion as a true binder would, and therefore might improve the sampling of protein conformations which are favourable for ligand binding, but would be rarely visited in a simulation with pure water as the solvent.

After devising a suitable framework for the preparation, execution and analysis of the molecular dynamics simulations and the subsequent docking calculations, conclusions are drawn in connection with the applicability of the former to generate a suitable protein conformational ensemble. First and foremost, previously unreported binding sites of the target protein are discovered in the conformations obtained from these simulations. Utilising the multitude of obtained protein structures, the binding sites are characterised as either stable or transient. Stable pockets are shown to be accessible in multiple simulations with various solvents, while the necessity of a specific cosolvent to open some transient pockets is demonstrated. Furthermore, the most important residues for the opening mechanism of a selected transient site are highlighted. A clear connection between the hydrophobicity of the cosolvent employed during the molecular dynamics simulation and that of the binding pockets opened during the given trajectory is established. Based on the presented results, the devised method for generating the protein conformational ensemble is expected to perform well in other drug discovery projects as well. Given its relatively low computational costs, simple but effective working principle, and potential to be automatised, this approach shows great promise to be a useful addition to the toolbox of computational drug discovery.

List of Abbreviations

2D two dimensional. 16

3D three dimensional. 2, 3, 12, 16, 34, 63, 64

CMD cosolvent based molecular dynamics. 9, 12, 13, 26, 32, 43, 45, 46, 49–52, 55–57, 60–62, 64–67

COVID-19 coronavirus disease 2019. 2, 8, 64

CPU central processing unit. 19

CV collective variable. 7

DBI Davies–Bouldin index, clustering quality metric. 15, 27–29

FDA United States Food and Drug Administration. 2, 8, 10, 35, 65

GPU graphical processing unit. 19, 20

LJ Lennard–Jones (potential). 20, 21, 64

MD molecular dynamics. xi, 3, 5–9, 12–16, 19–21, 23–27, 29–33, 37, 38, 41, 43–50, 53, 54, 58–61, 64–68

NMR nuclear magnetic resonance. 3

NSP nonstructural protein. 19

PDB Protein Data Bank. 3, 20, 34, 47, 57–60

PES potential energy surface. 4–6

pSF pseudo-F statistic, clustering quality metric. 15, 27–29

RdRp RNA-dependent RNA polymerase. 2, 8, 19, 34, 64, 67, 68

RMSD root-mean-square deviation. xi, 10, 14, 15, 21, 23–33, 49, 64, 65

SARS-CoV severe acute respiratory syndrome coronavirus. 2, 19

SARS-CoV-2 severe acute respiratory syndrome coronavirus 2. 2, 8, 17, 19, 34, 64, 67, 68

VS virtual screening. 4, 8–12, 20, 49, 60, 61, 63, 67

List of Symbols

$\mathbf{r}$	position of an atom	[Å]
γ	Energy type parameter of the Lennard–Jones potential.	[kcal/mol]
ε	parameter of the density based clustering algorithm, distance defining cluster neighbourhood	[Å]
k	parameter of the density based clustering algorithm, minimum points near cluster seed	[−]
k_B	Boltzmann constant	[J K^{-1}]
N	number of atoms in the protein	[−]
R	ratio of common ligands in two or more ligand sets	[−]
$R_{\min}$	Distance type parameter of the Lennard–Jones potential.	[Å]
r_{ij}	Distance between atoms i and j.	[Å]
S	pocket druggability score	[−]
S_1, S_2	sets of binding ligands	[−]
S_C	pocket charge score	[−]
S_H	pocket hydrophobicity score	[−]
S_P	pocket polarity score	[−]
S_V	pocket volume score	[−]
T	absolute temperature	[K]
V_{ij}	Lennard–Jones potential between atoms i and j.	[kcal/mol]

Contents

Abstract iii

List of Abbreviations v

List of Symbols vii

Contents ix

List of Figures xi

List of Tables xiii

1 Introduction 1
- 1.1 Biological relevance . . . 1
  - 1.1.1 Importance of protein inhibition . . . 1
  - 1.1.2 Targeting the proteins of SARS-CoV-2 . . . 2
- 1.2 Importance of protein dynamics . . . 3
  - 1.2.1 Computational methods of drug discovery . . . 3
  - 1.2.2 Protein dynamics . . . 4
  - 1.2.3 Molecular dynamics for protein structure generation . . . 5
- 1.3 Objectives . . . 8

2 Methods 9
- 2.1 Virtual screening . . . 9
- 2.2 Computational docking . . . 11
- 2.3 Cosolvent molecular dynamics . . . 12
- 2.4 Clustering of molecular dynamics trajectories . . . 13
- 2.5 Pocket detection algorithms . . . 15
- 2.6 Ligand similarity . . . 16

3 Results and discussion 19
- 3.1 Computational details . . . 19
  - 3.1.1 The target protein . . . 19
  - 3.1.2 Molecular dynamics trajectories . . . 20
  - 3.1.3 Trajectory clustering . . . 21
  - 3.1.4 Docking calculations . . . 21
  - 3.1.5 Ligand similarity calculations . . . 22

3.1.6 Pocket description . . . 22
3.2 Analysis of the molecular dynamics trajectories . . . 23
3.2.1 Equilibration . . . 23
3.2.2 Residue mobility . . . 25
3.2.3 Conclusions . . . 26
3.3 Selecting representative protein conformations . . . 27
3.3.1 Tuning the parameters of the clustering algorithm . . . 27
3.3.2 Evaluating the chosen representatives . . . 31
3.3.3 Conclusions . . . 32
3.4 Preliminary analysis of the docking calculations . . . 33
3.4.1 Testing the performance of the docking protocol . . . 34
3.4.2 Binding energy distribution . . . 35
3.4.3 Necessity of multiple protein conformations . . . 37
3.4.4 Conclusions . . . 39
3.5 Criteria for selecting the best binding ligands . . . 39
3.5.1 Definition of the alternative selection method . . . 39
3.5.2 Evaluation of the alternative selection criteria . . . 40
3.5.3 Conclusions . . . 44
3.6 Identification of binding sites . . . 45
3.6.1 Selection of the most promising binders . . . 45
3.6.2 Definition of the binding sites . . . 46
3.6.3 Analysis of the pocket ligand populations . . . 49
3.6.4 Similarity of the binders of a given pocket . . . 50
3.6.5 Conclusions . . . 54
3.7 Description of the protein structure around the binding sites . . . 55
3.7.1 Hydrophobicity of the binding sites . . . 55
3.7.2 Conformational changes around the binding sites . . . 56
3.7.3 Conclusions . . . 61

4 Conclusions **63**
4.1 Project relevance . . . 63
4.2 Summary of the most important results . . . 64
4.3 Final remarks . . . 67
4.4 Outlook . . . 68

Bibliography **69**

List of Figures

1.1 Illustration of the sampling problem of molecular dynamics. 6

3.1 Equilibration of the protein structures during the MD simulations. 24
3.2 Average RMSDs for each residue of the protein. 26
3.3 Clustering descriptors for the water trajectory with only the alpha carbons considered. 29
3.4 Clustering descriptors for the water trajectory with all heavy atoms considered. 30
3.5 Clustering descriptors for the benzene trajectory with only the alpha carbons considered. 31
3.6 Distribution of RMSD distances between the obtained cluster representatives. . 33
3.7 Ligand binding energy distribution of the conformations obtained from different solvents. 36
3.8 Cumulative ratio of ligands discovered by each conformation and each trajectory. 38
3.9 Ratio of ligands common to both the binding energy and energy gap criteria. . 42
3.10 Comparison of the binding energy distribution of the ligands selected by their binding energy and the ligands selected by both their binding energy and energy gap. 43
3.11 Ratio of ligands selected for all three MD trajectories. 44
3.12 Comparison of the number of ligands selected for the three MD trajectories and the crystallised protein structure. 47
3.13 Discovered binding sites of the protein. 48
3.14 Average inrapocket and interpocket ligand similarity scores, between binders selected by the binding energy or energy gap criteria. 53
3.15 Average intrapocket ligand similarity for protein confomrations obtained from the same solvent trajectory. 54
3.16 The pocket score and volume of pocket seven across the different protein structures obtained from MD trajectories. 58
3.17 Conformational changes of the protein around pocket seven. 59
3.18 Conformational differences between the active sites of the apo and holo crystallised protein structures. 60
3.19 The volume and hydrophobicity score of the active site across the different protein confromations obtained from MD trajectories. 61

List of Tables

3.1 Comparison of the results of AutoDock Vina to the binding poses reported in the literature. . . . 35
3.2 Binding energy and energy gap thresholds utilised to obtain about one hundred ligands for each trajectory. . . . 41
3.3 The interacting residues of the binding pockets discovered through ensemble docking. . . . 47
3.4 The number of ligands, selected by the binding energy criterion, binding to each protein binding site. . . . 50
3.5 The number of ligands, selected by the energy gap criterion, binding to each protein binding site. . . . 51
3.6 Ratio of hydrophobic residues near each pocket, along with the cosolvent in which the pocket has the largest number of binders. . . . 56

Chapter 1

Introduction

The purpose of this chapter is to give a short introduction into the field of computational drug design and to place the presented study into wider context. First, the biological and medical relevance of the project is discussed, after which the relevant aspects of the state of the art of structure-based drug design are summarised. Finally, the motivation for our efforts as well as our objectives are stated.

1.1 Biological relevance

Proteins are ubiquitous building blocks of living organisms, playing a critical role in the reproduction, metabolism, regulation and growth of cells. Even viruses, the simplest forms of organic systems capable of reproduction, rely on proteins produced by the host cell for their functioning. Understanding and manipulating the way proteins interact with their surrounding is therefore of utmost interest from both a biological and a medical point of view. Currently, our most important method to manipulate the function of proteins is through the administering of drugs, used in the treatment of the widest range of illnesses, as illustrated in the following section.

1.1.1 Importance of protein inhibition

The practical importance of drugs targeting proteins is well illustrated by the ever growing interest in identifying new binders for a variety of proteins. In References 1–6, the authors set out to discover small molecule protein inhibitors or activators, in the hopes of treating a number of different sicknesses ranging from cancers through bacterial diseases to the African

sleeping sickness. In Reference 7, six further successful projects to find small molecule binders of proteins are described by Cosconati *et. al.* The vast majority of small molecule and biological drugs approved by the United States Food and Drug Administration (FDA) are already aimed at proteins, with 78 % of them having clear protein molecular targets [8]. More recently, small molecule inhibitors that affect protein–protein interactions have become widely recognised as a promising new family of drugs [9]. Since there are exponentially more protein–protein interactions than proteins, these compounds have vastly more potential targets than conventional drugs aimed at individual proteins. To summarise, the universality and effectiveness of drugs interacting with proteins have become clear a long time ago. Therefore it is not surprising that scientists turned to them once again, when faced with the new and immediate challenges of the coronavirus disease 2019 (COVID-19) pandemic.

1.1.2 Targeting the proteins of SARS-CoV-2

The COVID-19 pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), continues to claim thousands of lives every day more than a year after its outbreak [10]. However, our knowledge about it and consequently our tools against it are vastly more potent than they were a year before [11]. When it comes to combating this virus, both the drugs already available to us and our attempts to find new ones are in general no different than those outlined in Section 1.1.1 and references therein. Drugs targeting the proteins vital to the reproduction of SARS-CoV-2 have been our most important tools against it, aside from vaccines which can only be used as preventative measures. Remdesivir, one of the most widely used antiviral drugs against SARS-CoV-2 around the world [12], targets the RNA-dependent RNA polymerase (RdRp) protein of the virus [13]. Furthermore, given the urgency of developing an effective treatment, most attempts to find new inhibitor substances were in fact drug repurposing studies, targeting the virus's RdRp [14–18] or other important proteins [19–22]. The RdRp protein is an especially promising drug target as it is responsible for the replication of the viral RNA inside the host cell [23], it is highly similar to the RdRp of SARS-CoV [24], which already has a number of verified inhibitors [25], and its high-quality three dimensional (3D) structure has been available from as early as April 2020 [26]. In large part due to the urgent nature of the COVID-19 pandemic most of the above cited research projects relied heavily, or even exclusively, on computational techniques for the discovery of the potential inhibitors, due to the cost and time efficiency of such methods. The development and current state of these approaches is discussed in the next section.

1.2 Importance of protein dynamics

1.2.1 Computational methods of drug discovery

Discovering new drugs, perhaps understandably, has always been an area of great interest throughout human history. During centuries of practice and innovation it has evolved from a somewhat arbitrary, empirical experimentation with natural substances into a precise science with an ever expanding industry behind it [27]. With the advancements in organic synthesis leading to more and more organic molecules being synthesised, the search space for potential drugs grew rapidly throughout the twentieth century. To tackle the skyrocketing number of compounds to be tested, a highly automatable, *in vitro* selection of promising candidates was developed. The resulting method of high-throughput screening lead to an unprecedented number of compounds being evaluated routinely, with around ten thousand substances tested in a week by a single company being no unusual feat by the end of the century [28]. By this time, with the exponential increase in computing power and development of innovative algorithms, computational methods were also being employed in the drug discovery process. Their potential became most obvious with the advent of structure-based drug design, where potential drugs are created or found based on the 3D structure of the protein target [27, 29]. Until recently, such target structures were only obtainable reliably through costly and cumbersome experimental methods such as X-ray crystallography [30] or nuclear magnetic resonance (NMR) spectroscopy [31], with 96 % of the structures in the Protein Data Bank (PDB) in 2020 determined with one of these methods [32]. Consequently, in its early stages structure-based drug discovery was severely limited by the scarcity of suitably high-quality experimental structures for the desired targets [33]. Nowadays, 3D target protein structures are much more readily available due to the gradual improvement of existing methods, the appearance of new experimental methods such as cryo-electron microscopy [34], and perhaps most importantly the development of recent computational techniques such as homology modeling [35]. Taking advantage of the quickly growing body of available genomic data, computational tools capable of predicting protein structures from mere amino acid sequence information have also been developed. The Ensembler program [36] uses traditional computational techniques like comparative modeling and implicit solvent simulations, while AlphaFold [37] utilises machine learning to provide 3D structures which are suitable starting points for more costly simulations such as molecular dynamics (MD). By employing one (or a combination) of the above techniques, high-quality structures are available for a larger number of protein targets than ever before.

The current challenge to computational chemists is therefore how to best utilise the available structural information. The computational methods developed for structure-based drug design fall into two main categories: *de novo* design methods construct new, tailored ligands, while

docking methods select ligands complimentary to the target from the existing compound space [38]. Of the docking methods, virtual screening (VS) has emerged as a particularly successful technique as illustrated for example in References 7 and 39. This procedure can be thought of as a computational extension to high-throughput screening: a large number of compounds are docked to the target protein structure *in silico*, after which the ligands producing the most desirable fits can be selected. For more technical details on virtual screening and docking calculations the reader is referred to Sections 2.1 and 2.2, and the references therein.

1.2.2 Protein dynamics

Traditionally, VS campaigns have been carried out utilising a single, experimentally determined protein structure, often in the crystallised form [33, 38]. The main reasons for this were the difficulties of obtaining accurate protein structures, as outlined in Section 1.2.1. Recently, the deficiencies of using only a single, crystallised protein structure has been recognised [33, 38, 40–42]. Firstly, the structure of the crystallised protein often differs significantly from the conformations that the protein adopts *in vivo*. Secondly, even if the crystal structure is representative of the conformation most often visited in solution, a single structure cannot account for the dynamics of protein motion. The theoretical model that first rationalised the importance of protein motion was the induced-fit model of ligand docking [43, 44], which claims that in the process of ligand uptake the structure of the protein is changed, as opposed to the previously accepted lock and key model where a rigid protein is assumed [45]. More recent is the model of conformational selection [46–48] which views the target protein as a dynamic object, adopting a wide range of conformations even in the absence of ligands. In the context of this theory, a ligand near the corresponding binding site of the protein influences the potential energy surface (PES) of the system through the formation of a protein–ligand complex, stabilising some conformations of the target which would otherwise be rarely visited. Effectively, this also leads to the changing of the protein conformation upon ligand binding, via a mechanism that is completely neglected if only a single protein structure is considered. The need to take protein flexibility and motion into account became even clearer with the discovery of several "cryptic" or "hidden" pocket structures [49–51]. The characteristic property of these pockets is that they only appear in the presence of the appropriate ligand while their existence is not obvious from the equilibrium structure of the protein. The exact mechanism of their formation is not yet clear, although some combination of induced-fit and conformational selection has been hypothesised [50]. The discovery and theoretical description of such pockets are hindered by the fact that their opening often requires large scale rearrangements of the protein structure, events that are traditionally hard to predict with computational techniques [52].

With the importance of protein dynamics gaining wider recognition, new, more elaborate methods are appearing which aim to account for this phenomenon. On the one hand some of the modern computational docking programs, such as AutoDock Vina [53], can treat a selected number of protein residues as flexible at the cost of increased calculation times. This method is well suited to study a previously known, specific binding site of the protein. However it cannot account for larger structural changes of the protein and is limited to a handful of flexible residues due to its computational requirements. On the other hand, the family of ensemble docking techniques utilises traditional (rigid protein) docking calculations in combination with an ensemble of protein conformations to account for the flexibility of the target [33, 38]. The problem of multiple thermally accessible protein conformations can obviously be tackled by using several protein structures, moreover with careful selection of the structures of the ensemble one can hope to describe larger scale conformational changes and discover new cryptic pockets [48, 51, 52]. The main challenge for these methods is the generation of the protein structure ensemble. The use of several experimentally obtained, crystallised structures is widely accepted [38, 54, 55]. With this approach, the flexibility of the protein can be accounted for by selecting structures that were crystallised with different ligands binding to the protein. Unfortunately, this method is still limited by the difficulty of obtaining experimental protein structures. To remedy this, computational methods have been utilised during the generation of the structure ensemble as well. The range of approaches is quite wide, including, but not limited to, systematic local conformational space searches [56], neural networks [57] and MD [33, 58]. MD is an especially promising avenue, after all it has been designed for the very purpose of efficiently sampling the realistic conformational space of proteins.

1.2.3 Molecular dynamics for protein structure generation

After decades of steadily improving algorithms and force fields, MD simulations are accurate enough to be useful in the exploration of the PESs of proteins [51]. These methods work by integrating Newton's equations of motion under the forces resulting from precalculated force fields, creating a trajectory from the motions of proteins and other molecules. Aside from potential inaccuracies introduced by the employed force fields, the largest obstacle of MD calculations is the extremely slow convergence of the calculated trajectories [33]. Given the complicated structure of macromolecule PESs, and the nonergodic and nonequilibrium nature of protein dynamics, second-long MD simulations can sometimes be required for an equilibrated sampling [59]. However, even with highly specialised code and computers the longest timescales within reach are in the range of milliseconds [60].

This problem of sampling in connection with MD is illustrated on Figure 1.1. Here, the PES of a given protein is represented with the black curve, which has many local minima separated

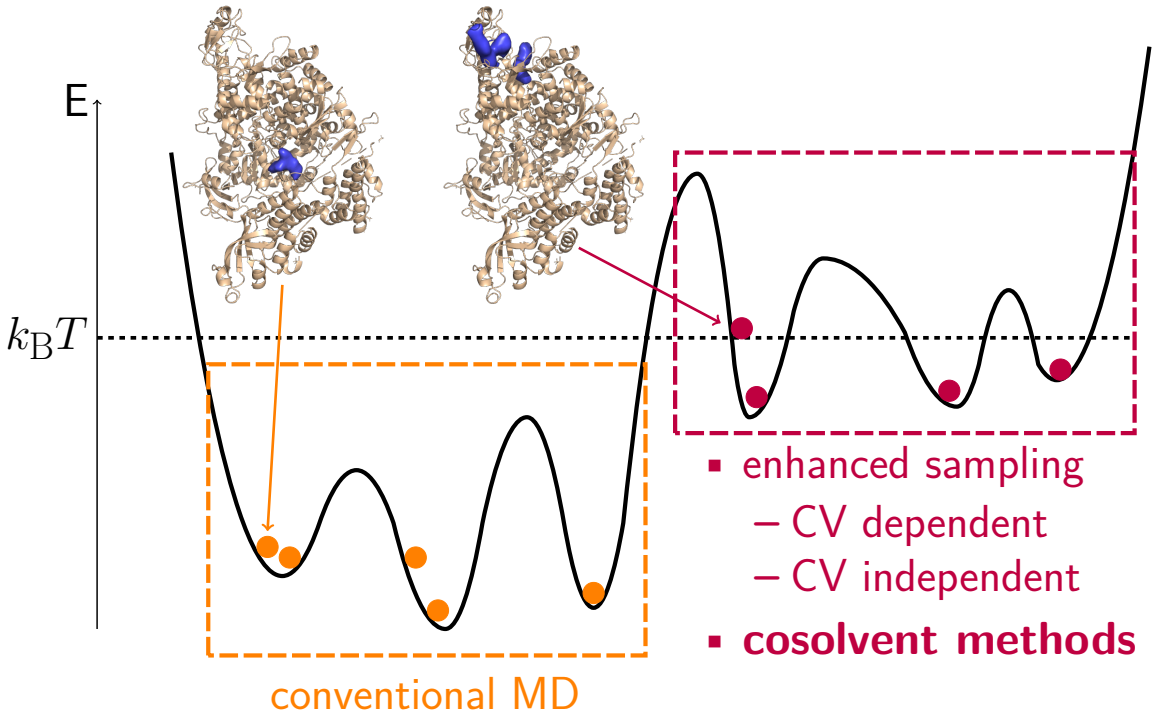

Figure 1.1: Illustration of the sampling problem of molecular dynamics. The curve represents the potential energy surface of the protein in question, the dots upon which represent some conformations of the protein. The conformations in orange can be visited by conventional molecular dynamics, while for the sampling of the conformations in purple more involved methods are required.

by large energy barriers. Some of the conformations which can be visited by conventional MD are shown by orange dots. The leftmost of these conformations features an open binding pocket at the active site of the protein, as indicated by the blue volume in the protein image corresponding to the conformation. Since this pocket is at the active site, it is unsurprising that conformations in which it is open lie very close to local minima of the PES and therefore they can be straightforwardly sampled using conventional MD. On the contrary, conformations represented by the purple dots are found somewhat higher in energy and are behind an energy barrier much larger than the thermal energy [k_BT, where k_B is the Boltzmann constant (in units of joule per kelvin), while T is the absolute temperature (in kelvins)]. As a consequence, these conformations are much more rarely (if ever) visited by traditional MD simulations. Unfortunately these rarely visited conformations, which might not even correspond to local minima of the PES, can harbor cryptic or transient pockets that could be targeted with drugs [51]. One such conformation is represented by the leftmost purple dot and the corresponding protein image, with the blue volume representing an open cryptic pocket distant to the active site.

In order to be able to sample these rarely visited conformations, a number of modified MD techniques have been developed. The first group of these is the enhanced sampling methods,

where some unphysical bias is introduced into the simulation in order to encourage the sampling of otherwise unlikely conformations. This group can be further subdivided into two categories based on whether they require a so called collective variable (CV) or not. A CV stores some information about the system and its desired conformations, which can then be utilised to apply the appropriate bias guiding the simulation towards these conformations. Some of the most notable CV dependent enhanced sampling methods in the context of cryptic pocket discovery are umbrella sampling [61], steered MD [62], and metadynamics [63]. The principal drawback of such methods is that the construction of an appropriate CV requires some *a priori* knowledge about the system in question, which is often not available, especially if the goal is to discover new binding sites for a target protein. Attempts have been made to remedy this problem, such as the JEDI framework [64]. This algorithm guides the MD simulation using a generalised measure of how likely a certain conformation is to harbor a typical small molecule drug or with other words, its druggability. Novel algorithms like this are certainly promising, but unfortunately they currently still suffer from some early stage imperfections such as high computational cost and a tendency to produce degenerate scores for a number of conformations resulting in simulations getting stuck in some undesirable state [51]. The other subclass of enhanced sampling, the CV independent methods avoid CVs and the problems associated with them completely. Perhaps the most popular of these methods is the temperature replica exchange MD [65], where several replicas of the system are simulated at different temperatures, regularly attempting to swap the temperatures between replicas based on some Metropolis criterion. Unfortunately, this method was deemed ineffective for opening cryptic pockets in a number of targets [50, 51]. Innovative approaches are being explored for CV independent methods as well, like the SWISH algorithm [50], which modulates the hydrophobicity of the protein residues instead of the temperature. While encouraging results have already been achieved with this method [66], the unfolding of proteins due to their artificially introduced hydrophilicity can pose serious obstacles [51].

A completely separate approach for the sampling of rarely visited conformations harboring cryptic pockets is that of the cosolvent methods. The main idea behind these frameworks is to replace the traditional water solvent in MD simulations with a mixture of water and some other cosolvent. The oftentimes hydrophobic or amphipatic cosolvent probes can then interact with the protein and occasionally induce conformational changes or stabilise some conformations where a cryptic pocket is open. Cosolvent methods have been successfully used to identify cryptic sites in a number of targets [48, 51, 67, 68]. Additionally, they do not require any preexisting knowledge about the system and have virtually no extra computational costs compared to traditional MD simulations. However they can also be prone to protein unfolding [67], and the discovery of cryptic pockets can require longer MD simulations compared to some enhanced sampling techniques [50, 51].

1.3 Objectives

The primary aim of the presented work is to aid the development of new computational frameworks with which the effectiveness of future VS campaigns can be improved. As discussed above, the current most important challenge for these projects is recognised to be the description of protein flexibility and dynamics. The necessity and efficiency of using an ensemble of protein structures in the context of VS are evaluated. The ensemble of protein structures is obtained via cosolvent MD techniques, the adeptness of which is thoroughly discussed. The viability of performing VS campaigns against several protein structures with thousands of ligands is examined. Attempts are made to establish standard methods for the detection and definition of cryptic pockets in arbitrary protein targets, from cosolvent MD and computational docking data. Recognising the severity of the COVID-19 pandemic, the benchmark calculations are carried out on the RdRp protein of SARS-CoV-2, and a set of FDA approved small molecule drugs, in the hopes of contributing to the development of effective treatments against this virus.

Chapter 2

Methods

In this chapter the computational methods employed during the present project are introduced more thoroughly and some technical details of their usage is discussed. In particular, the techniques of VS, computational docking, cosolvent based molecular dynamics (CMD), the clustering of MD trajectories, pocket detection and similarity calculations between small molecules are discussed. For the documentation of the settings and options of the utilised programs employed to obtain the presented results, see Section 3.1.

2.1 Virtual screening

As described in Section 1.2.1, VS is one of the most popular and successful computational methods in drug discovery. Its underlying principle is the same as of experimental high-throughput screening: searching through a large number of compounds in an efficient and parallel manner, to select the most promising drug candidates. The conceptual difference between the two approaches lies in the method used to perform the testing of individual compounds: while experimental techniques are employed in high-throughput screening, VS is performed solely with computational methods. In particular, computational docking calculations are utilised to evaluate the binding affinity of the ligand to the target protein [69]. The details of these docking calculations are discussed in Section 2.2.

Before the docking calculations can be performed, a crucial decision has to be made at the beginning of every VS calculation, concerning the definition of the set of ligand candidates to screen. To make a correct decision, the biological function of the target protein, the goal of the VS campaign and many other practicalities have to be taken into account. The number of considered ligands typically ranges from few hundreds to a couple thousands, with the most

severe limiting factor being the available computational resources. To reduce the number of ligands for which explicit docking calculations have to be performed, large ligand databases can be enriched in promising ligands by performing some crude pre-filtering [69]. To aid the definition of ligand sets, public databases containing millions of small molecules have been compiled, such as ZINC [70] or PubChem [71]. These databases often provide predefined subsets of ligands created with a specific goal in mind, for example the subset of FDA approved drugs for drug repurposing projects.

On top of assembling the ligands to be searched, the atomistic structure used to represent the target protein also has to be chosen or created. The most commonly employed techniques to arrive at such a structure are detailed in Section 1.2.1. After the ligands and the protein has been prepared, the VS campaign can proceed with the docking calculations for each ligand. This is the most computationally demanding step of the procedure as it often requires thousands of explicit docking calculations each of which constitutes a complicated global optimisation problem and therefore can take up to a couple of minutes to compute. Fortunately, the calculations for a given ligand are completely independent from every other ligand which enables us to perform these calculations in a massively parallel manner. Devising an execution scheme for the thousands of independent docking calculations which maximally utilises the available computational resources is a non-trivial problem from the field of computer science, especially if the complicated infrastructure of a modern supercomputer is involved. To make this challenge more approachable for computational chemists, VS frameworks have been developed [72], with which the scheduling of the docking calculations can be automatised. These tools often include features to manage large ligand libraries, extract the results of docking calculations, and other "bookkeeping" functionality. While the use of these tools is certainly justified in the largest projects, for smaller scale campaigns such as the one presented here, a simple set of scripts can be sufficient to manage the docking calculations. After the results of the dockings are available for all the ligands being considered, they can be used to select the most promising candidates.

The criteria chosen to perform the selection of the best ligands depends heavily on the goals of the VS campaign and to a lesser extent on the type of results provided by the docking program. A widespread choice is the use of a binding energy threshold (used e.g. in the projects described in Reference 7), as it is straightforward to apply and almost all docking programs provide an estimate of this value among their results. Occasionally applied further criteria are related to the distance root-mean-square deviation (RMSD) between the best poses found for a given ligand [7, 73, 74], and are aiming to measure the consistency of the predicted poses. Other possible criteria include manual visual inspection of the docked poses, and docking to decoy conformations or targets. The threshold for the best ligand selection is usually chosen to yield no more than a hundred promising candidates, which are usually further evaluated *in vitro* [7].

2.2 Computational docking

Given the structure of a macromolecular target and a small molecule ligand, the goal of a computational docking calculation is to find the spatial poses of the ligand for which the binding energy (docking affinity) to the target is optimal. The optimised poses and binding energy estimates can be utilised in a number of ways, the most important of which are most likely VS campaigns and fragment based drug discovery projects [75]. The rapidly increasing number of available protein structures along with the exploding computational capabilities allowing for more and more accurate simulations resulted in docking calculations becoming one of the most useful methods in predicting and designing small molecule binders. The continued interest in these algorithms yielded a large variety of both open-source and commercial docking programs in the last decades [76]. One of the better known of these is AutoDock, first released in 1990 [77], continuously improved since then, and freely available since 2007. In the present study, AutoDock Vina will be utilised, which is the newest release of the AutoDock developers, aiming to be a black-box docking tool [53]. This choice was motivated by the fact that its effective use requires little *a priori* knowledge about the target protein and that it has been shown to perform well in standardised tests compared to other docking programs [78].

The search for the best ligand poses during a docking calculation represents a complex global optimisation problem. The function to be optimised (also called the scoring function) associates an estimation of the ligand binding energy to each pose of the ligand relative to the target structure. Choosing the correct form for this function is crucial in the process of writing docking programs: a delicate compromise has to be found between accuracy and speed of evaluation. Most often, this is achieved by employing some empirical function containing a limited number of terms, although machine learning and deep learning approaches are also becoming more and more popular [79]. AutoDock Vina employs a traditional, simplistic scoring function containing only five terms dependent on interatomic distances, and one term dependent on the number of rotational degrees of freedom in the ligand [53]. The global optimisation of these scoring functions performed during computational docking is particularly challenging due to two main reasons. The first of these is the extraordinarily large size of the search space. To limit the number of possible conformations to something tractable with the current computational resources, the protein structure is usually considered rigid. Nonetheless, the remaining degrees of freedom of ligand translation, rotation, and in the case of most modern programs (including Vina) the internal rotations of the ligand, lead to a large number of potential poses that increases rapidly with the size of the ligand. It is worth noting that on top of intraligand rotations, AutoDock Vina is able to treat some important protein residues as flexible, which could further increase the complexity of the search space. However, the aim of the present study is to attempt to account for protein flexibility solely through utilising multiple protein

structures, therefore these structures were considered completely rigid throughout the docking calculations. The second problem often encountered during the optimisation is that the scoring function contains an excessive number of local minima for almost all protein ligand pairs, which can pose serious problems for all global optimisation algorithms [80]. It is because of these reasons that finding the absolute best docked pose for a given ligand remains challenging and why the results of docking calculations should be carefully verified, even with millions of scoring function evaluations and sophisticated optimisation algorithms.

After the docking calculation is performed, most programs present a handful of the best docked ligand poses as results. By returning more than a single pose, the programs allow the user to manually select some desirable conformations even if those were not ranked among the best, based on their scoring function values. This is especially useful as the binding energy estimation calculated by these programs is not particularly accurate, for example compared to experimental values a standard error of 2.85 kcal/mol is reported for AutoDock Vina on the CASF-2013 test set [78]. To improve on this, one can take advantage of the set of ligand poses to devise empirical corrections to the calculated binding energy. One commonly used idea is to correct for the entropic terms missing from most scoring functions, by accounting for the spatial distribution of the docked poses [73, 74]. In the case of AutoDock Vina, the obtained best poses are first clustered and only the cluster representative poses are returned [53], which makes the above mentioned, pose distribution based corrections less feasible.

2.3 Cosolvent molecular dynamics

The prerequisite of any docking calculation and consequently VS project, is at least one high-quality 3D target protein structure. Moreover, as one of the objectives of the present study is to investigate the effects of protein dynamics to docking calculations through ensemble docking, not just a single but a handful of such structures are required. As discussed in Section 1.2.3, CMD is a promising technique for the generation of appropriate protein conformations. The success of CMD to deliver protein conformations more suitable for docking in comparison to traditional MD [40, 67] can be rationalised as follows. On the one hand, traditional MD simulations with water as the solvent can only rely on thermal fluctuations to induce favorable conformational changes in the protein, as water molecules do not resemble traditional small molecule drugs and therefore cannot provoke induced-fit type changes in the protein [48]. On the other hand, the organic, apolar probes introduced in CMD simulations can interact with the protein in a similar manner as a binding ligand would, therefore it can prompt conformational changes based on the induced-fit model of ligand docking. In other words, the cosolvent probes are expected to mimic the behavior of the binding ligand in some aspects, opening pockets

that would normally only open in the presence of the appropriate binder. Hot-spot mapping, a related application of CMD simulations, utilises the fact that the cosolvent probes interact with the protein residues in a similar way that a binding drug would [81–83]. These methods identify regions of the protein surface where a small molecule drug is likely to bind based on the spatial distribution of the cosolvent probes. By choosing probes that show some of the usual characteristics of small molecule drugs, these molecules are stabilised around sites of the protein with which the drugs themselves would strongly interact. Consequently, by identifying regions where the local cosolvent concentration is higher than the bulk value, one can pinpoint protein residues that can potentially be targeted with binding ligands. Moreover, the interaction energy of ligands binding to the discovered site have also been estimated by more careful statistical analysis of the probe distribution [84]. Since in the present project the use of CMD simulations is restricted to generating a range of protein conformations to be used in docking calculations, such statistical analysis of the spatial distribution of the probes is not performed.

When preparing CMD simulations, the most important decisions to be made are those concerning the type and concentration of the cosolvent. In general, organic, at least partially apolar molecules containing moieties characteristic of small molecule drugs are preferred, for the reasons outlined above. The most common choices are benzene or phenol for their large apolar surfaces, and ethanol or isopropanol for their amphipathic character [67]. The concentration of the cosolvent is commonly chosen to be between 5–20 v/v % [40, 48, 67, 68], with 10 v/v% of phenol cosolvent found to give the best results for the targets of Reference 67. When higher concentrations are used, one must be careful to avoid the unfolding of the protein, while with more hydrophobic cosolvents such as benzene the clustering of the probes can also cause issues. To remedy this latter issue, modified force field parameters have been devised for the most problematic cosolvents that help reduce clustering by introducing an artificial repulsion term between the probes [85–87].

2.4 Clustering of molecular dynamics trajectories

Employing the above described CMD simulations, one is able to generate a large number of protein structures that can be suitable inputs to the subsequent docking calculations. However, the number of protein conformations generated in this way is actually too large: in their initial form, MD trajectories are commonly composed of millions of snapshots. In contrast, we have encountered up to minute long calculation times for the docking of a single ligand to a single protein structure. To run docking calculations for all frames of a trajectory is thus completely unfeasible. It is therefore unavoidable to filter the large number of protein structures obtained from a MD trajectory, before ensemble docking calculations can be performed with

them. The most straightforward approach to perform this filtering is simply to consider a small subset of the frames taken from the trajectory at regular intervals. This is especially adequate because snapshots lying close to each other usually depict very similar conformations and are consequently redundant for the purposes of ensemble docking. The millions of frames coming from a raw trajectory can therefore usually be reduced to a couple thousands without fear of losing any valuable information about the protein dynamics. Unfortunately, this number is still far too large for docking calculations, therefore some further, more refined selection is necessary.

A commonly used approach to obtain only a handful of representative protein structures to be used for docking purposes is that of trajectory clustering [1, 2]. Clustering is a general technique to reduce the size of data sets while minimising the loss of information. Its working principle is to retain only some representative data points, while discarding many others that contain little or no new information. To achieve this, the most similar data points are grouped into clusters, and subsequently a single data point is selected from each cluster to represent the whole group. In the context of MD trajectory clustering, the similarity of the conformations depicted in the various frames is used to define the clusters and to select representative snapshots. A frequently utilised metric of similarity is the RMSD value [88], calculated between conformations A and B as:

$$\mathrm{RMSD}^{A,B} = \sqrt{\frac{1}{N}\sum_i |\mathbf{r_i^A} - \mathbf{r_i^B}|^\mathbf{2}} \ , \tag{2.1}$$

where N is the number of atoms in the structure, while $\mathbf{r_i^A}$ and $\mathbf{r_i^B}$ are the coordinates of the ith atom of the structure in conformations A and B respectively.

Given RMSD distance data between all conformations of a trajectory, a number of different algorithms can be used to perform the definition of the clusters. These methods can be divided into top-down and bottom-up approaches [89]. The starting point of the top-down techniques is a single cluster, containing all snapshots. These algorithms then proceed by splitting this cluster into multiple smaller ones based on some similarity based criterion, until the desired number of clusters is reached. These methods tend to be very fast but mostly only produce similarly sized clusters. A further disadvantage of them is that the number of clusters to be created needs to be supplied as input, which is impractical when the number of relevant conformations is not known in advance. Conversely, at the beginning of bottom-up algorithms many smaller, initial clusters are considered and the program proceeds by systematically merging these based on their similarity. These methods can be more computationally intensive but can provide more diverse clusters in terms of shape and size [89].

A popular bottom-up algorithm used upon MD trajectories is density based clustering [90]. As this clustering algorithm is the default method implemented in `cpptraj` [91], the main

trajectory analysis tool used in the research group, and is generally recommended by the authors of the program, it was used to perform the trajectory clustering in the present work. A practical advantage of this algorithm is that it does not require the desired number of clusters to be determined in advance. Instead, it has two adjustable parameters with which the criteria for the defining and merging of the clusters are controlled. The first parameter k defines the minimum number of data points that have to be in the neighbourhood of a given data point for it to be considered the "seed" of an initial cluster. Increasing this parameter usually results in fewer final clusters, as the number of initial clusters is decreased. The second parameter ε is a RMSD value defining the neighbourhood of data points. It is used both during the definition of the initial clusters and also during the subsequent merging of them. Increasing this parameter usually decreases the number of final clusters, while also decreasing the number of frames not included in any cluster (noise frames), since it relaxes the criteria for the merging of two clusters. In order to perform an adequate clustering with the density based clustering algorithm, these two parameters have to be well tuned for the given trajectories. This tuning is a delicate matter for which no universally applicable procedure exists. The authors of the `cpptraj` program recommend simple trial and error approaches [92]. Different metrics characterising the quality of the achieved clustering can be very useful in this process, as they can indicate the optimal values of the two parameters. In the present work two of the better known clustering metrics will be used during the tuning of the density based clustering parameters. These are the pseudo-F statistic (pSF) and the Davies–Bouldin index (DBI), the general idea behind both of which is to compare the intra- and intercluster variances [89], with a small intracluster and large intercluster variance indicating good quality clustering. Although the general principle is the same between the two, the actual formulation is quite different, resulting in the fact that the two descriptors can behave quite differently. Moreover, while for the DBI a low value is desirable, in the case of the pSF a high value signals optimal clustering.

2.5 Pocket detection algorithms

The problem of selecting a few appropriate protein conformations from MD trajectories can also be tackled with so called pocket detection programs [93–95]. These algorithms take as input a single or an array of protein structures, that they use to make predictions about what regions of the protein are likely to bind typical small molecule ligands. The search for binding regions is based on the assumption that these are usually concave areas on the protein surface formed by hydrophobic residues. Since these algorithms do not consider the ligands explicitly, they circumvent the problem of dealing with the excessive number of possible ligand poses and consequently their computational requirements are much tamer than those of explicit

docking calculations. Their economic computational costs make it possible to apply them to thousands of frames of a MD trajectory. Therefore, while they are unable to predict the actual binding energy of specific ligands, they are perfectly suitable to select the most promising snapshots from a long trajectory. To facilitate the selection of those snapshots which harbor conformations best suited for docking, some of the pocket detection programs output a number of pocket descriptor metrics for each considered frame. In the present work, these pocket descriptors will be utilised to evaluate the effects of different cosolvents on the protein structure around binding sites.

The pocket detection algorithm of choice in this study is `fpocket` [93], a popular open-source program offering a wide range of pocket descriptors. It uses Voronoi tessellation and so called alpha spheres to find concave regions on the protein surface. The alpha spheres are spheres defined by exactly four protein atoms touching their surfaces. The regions of space where a large number of alpha spheres with appropriate radii can be constructed are considered binding pockets. A useful byproduct of this pocket definition, that is also the distinguishing feature of `fpocket`, is that protein residues can also be assigned to the discovered pockets, by considering the atoms touching the alpha spheres of the given pocket. This enables the calculation of many additional pocket descriptors that are based on for example the hydrophobicity of the surrounding residues.

2.6 Ligand similarity

The concept of molecule similarity is an inherently human one, and as such it can be difficult to translate to a language that computers understand. In the context of small molecule drug discovery, two ligands are usually considered similar if they have some common functional groups, share the same general structure or contain the same heteroatoms at matching positions. To capture this idea computationally, a variety of molecular fingerprints have been developed. A fingerprint is a string of bits with a fixed size associated with a given molecule, where each bit is set to zero or one based on specific properties of that molecule, following some predefined rules. The properties on which the value of the fingerprint depends are usually those that are associated with the concept of similarity as mentioned above. Depending on the application, these properties are extracted from the two dimensional (2D) or 3D structure of the molecule in question. The utilisation of the obtained fingerprints in similarity calculations is a key step as it replaces the complicated and ill-defined problem of molecular similarity with the much simpler problem of similarity of strings of bits. For the evaluation of the latter a number of expressions can be used, the most popular of which is the so called Tanimoto score [96]. With the above

technique a similarity score between zero and one can be obtained for any pair of molecules, with a higher score indicating a larger degree of similarity between the two structures.

In the presented work, ligand similarity calculations will be performed to determine whether the different discovered pockets of SARS-CoV-2 RdRp bind radically different ligands and whether there are some similarities between the ligands binding to the same pocket. To this end, the default fingerprint of the open-source cheminformatics toolbox RDKit [97] calculated from the SMILES representations of the ligands is used, along with the Tanimoto similarity score.

Chapter 3

Results and discussion

This chapter is dedicated to the presentation and discussion of the results obtained during the project. First, the technical details of the performed computations are documented for reproducibility. Afterwards, the main findings are presented, separated into logical units, each of them followed by a short paragraph dedicated to summarising the most important points discussed in that unit. Throughout the chapter the results are also explained and interpreted, so that the final conclusions can be drawn from them in Chapter 4.

3.1 Computational details

All calculations presented here were carried out on a single workstation with an Intel Core i7-9800X 3.8 GHz central processing unit (CPU), two Nvidia GeForce RTX 2080 8 GB graphical processing units (GPUs) and 32 GB of memory. For the inspection and manipulation of docked ligand poses, protein conformations and MD trajectories the open-source PyMOL molecular visualiser was used [98]. Other tasks were carried out with the help of shell scripts, Python and Julia scripts and notebooks on top of the programs cited below.

3.1.1 The target protein

The SARS-CoV-2 RdRp protein complex was chosen as the target of our investigations (see Section 1.1.2). In its active form it is composed of three domains: nonstructural proteins (NSPs) 7, 8 and 12 of SARS-CoV-2. Its active site is located in a deep groove and is highly similar to that of the analogous protein of the SARS-CoV [25]. Its simulation ready structure was obtained from the website of D. E. Shaw Research [99], where extensive MD simulations

have already been carried out for it. In Reference 14, the authors note that two zinc ions are necessary for the structural integrity of the protein. These ions were however not found in the structures and trajectories downloaded from D. E. Shaw Research. After numerous failed attempts at stabilising these zinc ions in their bound positions with restraining potentials and gradual heating, their inclusion was rejected in favor of the original D. E. Shaw structure. Additionally, two crystal structures determined with cryo-electron microscopy were downloaded from the website of the PDB: the apo structure 6M71 [100] and the holo structure 7B3B [101]. Out of these two crystallised structures, only the apo structure was utilised for docking calculations, to emulate drug discovery VS campaigns where the holo structure of the target is not available. The holo structure was used only to visualise the conformational changes around the active site occurring during ligand binding (see Section 3.7.2).

3.1.2 Molecular dynamics trajectories

The MD calculations were carried out with the Amber 18 program package [102], according to the following protocol. Three types of solvent boxes were prepared for the simulations: a simple water one and two with either benzene or phenol as cosolvent. The protein structures were solvated in octahedral solvent boxes containing the appropriate mixture of water and cosolvent molecules. A distance of at least 12 Å was left between the protein and all sides of the solvent box. The charge of the system was neutralised with sodium ions. During the simulations, periodic boundary conditions and a 12 Å cutoff for the Lennard–Jones interactions were used. The solvated systems were first minimised for 1000 gradient descent steps followed by an other 1000 conjugate gradient steps. Next, the heating of the systems to 300 K were performed during a 1 ns simulation with the Langevin thermostat in the NVT ensemble. Finally, 200 ns production simulations were carried out at 300 K and 1 bar pressure using the Langevin thermostat and Berendsen barostat in the NPT ensemble. Three replicas were run for the production calculation for each solvent. The last two simulations for each solvent were started from a random equilibrated frame of the first simulation for that solvent, with the velocities of all particles randomised according to the Boltzmann-distribution. The production calculations were run with GPU acceleration, using the `pmemd` program of Amber 18.

During the preparation of the solvent boxes containing cosolvents the `packmol` program was utilised [103]. The concentration of the cosolvents were set to 10 v/v % in both cases. In the case of the benzene cosolvent severe clustering of the cosolvent molecules was observed during the MD simulations when the default force field parameters were used. To circumvent this issue, scripts included in the `ParmEd` distribution [86] were utilised to introduce Lennard–Jones (LJ) potentials between the carbon atoms of different benzene molecules. The exact form of

this artificial potential between carbon atoms i and j is:

$$V_{ij} = \gamma \left[\left(\frac{R_{\min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min}}{r_{ij}} \right)^{6} \right] . \tag{3.1}$$

Here, V_{ij} is the introduced LJ potential, γ is the parameter determining the minimum value of the potential, $R_{\min}$ is the parameter controlling the position of the minimum, while r_{ij} is the distance between carbon atoms i and j. The default parameter values of $\gamma = 0.00036$ kcal/mol and $R_{\min} = 7.12719$ Å were utilised. After this modification was made, no clustering of the benzene molecules was observed during the simulations.

3.1.3 Trajectory clustering

For the clustering of the MD trajectories the `cpptraj` program [91] of Amber 18 was utilised. A density based clustering algorithm (chosen with the `dbscan` keyword of `cpptraj`) was employed, with the parameters k (unitless) and ε (in ångströms) set to 4 and 1.1 Å respectively (see Section 3.3 for the discussion of this choice). For each type of solvent the equilibrated part of the trajectories of the three replica simulations were concatenated and the clustering was carried out separately for each solvent. Before clustering, the structures in every frame were aligned to each other by their alpha carbon atoms. The clustering was performed using the RMSD values of the alpha carbon atoms as the distance metric between the conformations. A total of 19 cluster representatives were obtained with 13 coming from the trajectory with water as the solvent while the benzene and phenol cosolvent trajectories yielded 3 cluster representatives each.

3.1.4 Docking calculations

The set of FDA approved drugs were downloaded from the ZINC database [70] in the `mol2` format. This set is a popular choice for drug repurposing studies [16–18] and with approximately 2000 thousand contained ligands, it was feasible to perform docking calculations for all protein conformation, ligand pairs. From this set, 1957 ligand structures were converted to the `pdb` format, necessary for docking with AutoDock Vina, with the `openbabel` program [104]. The 19 cluster representative protein structures along with the holo crystal structure were aligned to each other by the RMSD distances between their alpha carbons. The protein and ligand structures were prepared for docking, relying on the scripts included in the AutoDockTools4 distribution [105]. The same docking region was used for all docking calculations, which encompassed the whole protein structure and was generated by AutoDockTools4. To carry

out the docking calculations, AutoDock Vina was run with the default command line options, except for the `exhaustiveness` option which was increased to 24, as is suggested by the authors for large docking regions and the `num_modes` option which was set to twenty to obtain the twenty best poses for each ligand. The parallel execution of the docking calculations were managed with in-house scripts.

3.1.5 Ligand similarity calculations

Based on the results of the docking calculations the best ligands binding to each discovered pocket were selected. Afterwards, similarity calculations were performed between all selected ligands. These calculations employed the RDKit program package [97], using its default RDKit small molecule fingerprint and the Tanimoto similarity score [96].

3.1.6 Pocket description

The binding sites of the protein, discovered through computational docking calculations, were analysed with the `mdpocket` program [106], part of the `fpocket` distribution. To this end, the 19 cluster representative protein structures were aligned to each other by their alpha carbons and concatenated to create a mock trajectory readable by `mdpocket`. The regions of space which the discovered binding pockets occupy were selected manually, by inspecting the poses of the ligands binding to the pocket in question. Based on the suggestions of the `mdpocket` authors, large regions were selected for each pocket, encompassing all or almost all docked ligand poses. With the protein structures concatenated and the binding regions selected, `mdpocket` was run with the `-S` option, instructing the program to score pockets by their druggability. Among its results `mdpocket` provides a number of pocket descriptors calculated for each frame of the supplied trajectory. From these descriptors, the pocket volume and various pocket druggability scores are utilised in the present study. To qualitatively evaluate the general druggability of a given pocket, a simple composite druggability score is defined here, that can be calculated from the descriptors provided by `mdpocket` as:

$$S = S_H + S_V + S_P + S_C \,. \tag{3.2}$$

Here, S_H, S_V, S_P and S_C are the hydrophobicity, volume, polarity and charge scores calculated by `mdpocket` respectively. It is emphasised that the definition of S is not suggested by the `mdpocket` developers and it is not intended as an absolute metric of pocket quality, but rather as a qualitative means to compare the druggability of the same pocket in different

protein conformations. For more information about the calculation and meaning of the pocket descriptors utilised for S, see the `mdpocket` documentation or Reference 106.

3.2 Analysis of the molecular dynamics trajectories

In this section, the trajectories obtained from the MD simulations are examined from two angles. First, the equilibration process of the protein during the first part of the production simulation is discussed. This phenomenon occurs because even after the simulated system is minimised and heated, it is possible that the protein is in a highly repulsive or simply unrealistic conformation. As a consequence, in the first part of any production MD simulation, the protein structure goes through more drastic changes as it adopts an equilibrated conformation. Moreover, since the simulations are carried out in the NPT ensemble, the volume of the simulation box is allowed to change to allow for the simulation of constant pressure. This can lead to rapid changes in the system volume before the equilibrated simulation pressure is reached, which entail sudden changes in the density of the system as well. It is therefore important that only the equilibrated part of a MD trajectory is considered for further analysis, as many of the changes happening in the early stages of the simulation are just artifacts of the computational method, and therefore no conclusion should be drawn from them. The pace of the equilibration process can usually be followed by plotting the RMSD distance of the protein from the starting conformation, against the time elapsed in the simulation. Finally, with the equilibrated section of the trajectories defined, the conformational changes occurring in the protein are examined. By plotting the average RMSD for each residue of the protein one can hope to identify regions with higher mobility, which could be of further interest, if the dynamics of protein motion are expected to be of high importance.

3.2.1 Equilibration

To examine the equilibration process of the protein during the MD trajectories, the RMSD values are plotted against the time elapsed in the simulations. On Figure 3.1, one can observe these plots for the first replica of each solvent. It is reassuring, that the protein structures seem to be well equilibrated after 50 ns of simulation time in all three cases. The equilibration appears to be happening slightly faster in the benzene and especially in the phenol cosolvent trajectory than in water. Admittedly, the determination of the point from which the structure is considered equilibrated is rather arbitrary, therefore no far reaching conclusions should be drawn from this difference in speed of equilibration. The RMSD curves for the other two replicas of each cosolvent, which were started from already equilibrated frames, are slightly

different. The rapidly rising first section of the curve corresponding to the equilibration process is much shorter or completely missing in those cases, as can be expected from the fact that those simulations were started from equilibrated system snapshots.

The equilibrated RMSD values plotted on Figure 3.1 are somewhat higher for the two cosolvent trajectories than in water. This could indicate that the cosolvent probes have stabilised some conformations that are not often visited with water as solvent and that are farther from the original protein conformation than those appearing frequently in water based simulations. The amplitude of the oscillations around the equilibrium RMSD value are also noticeably larger in the cosolvent simulations. It is therefore reasonable to expect that these trajectories sample a wider variety of conformations, increasing the chances to discover binding sites that are hidden during the water solvent simulation.

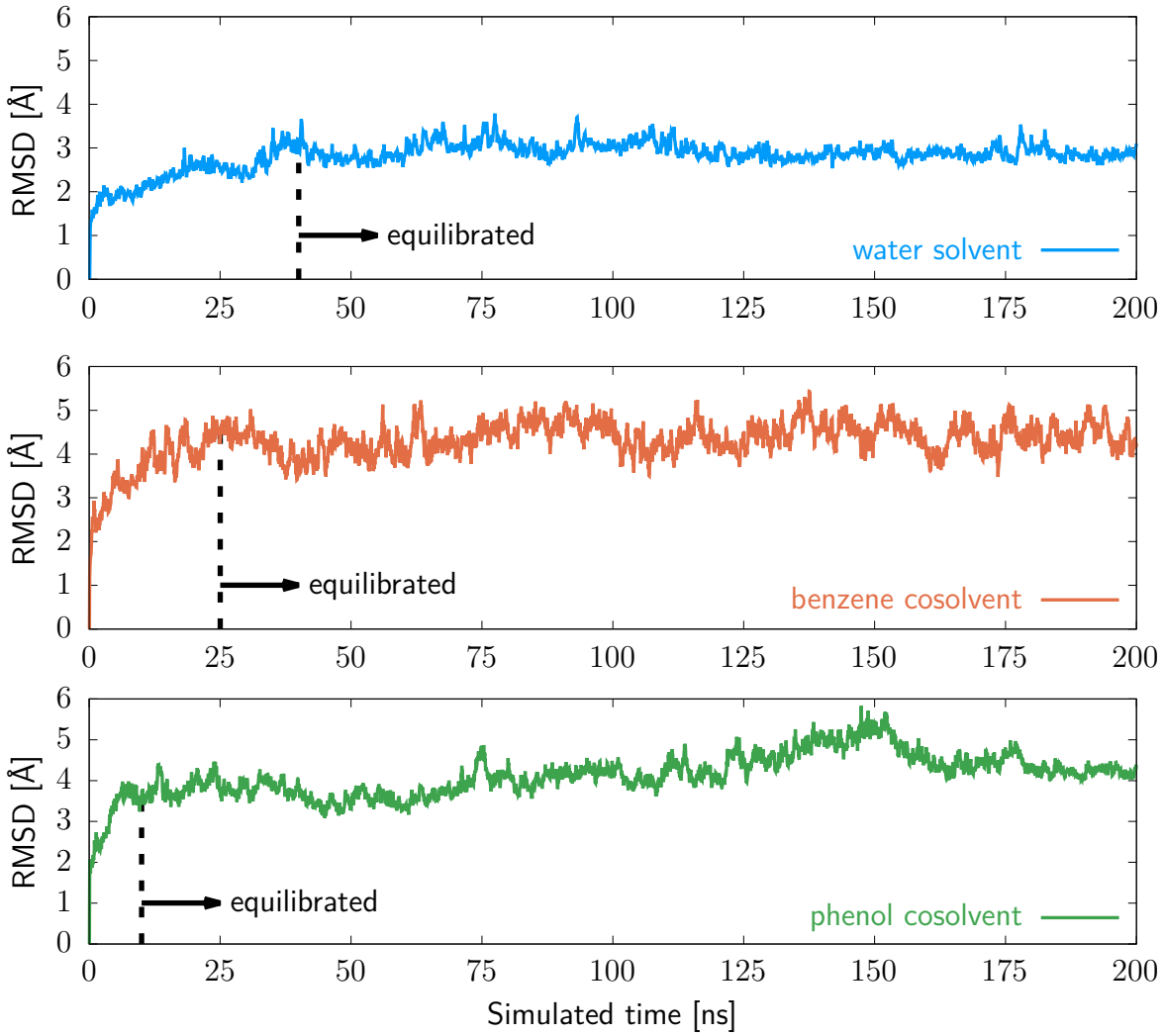

Figure 3.1: The evolution of the RMSD distance of the protein from the starting conformation during the MD trajectories. The first replica for each solvent is plotted.

3.2.2 Residue mobility

To investigate the mobility of different regions of the protein, the average residue-wise RMSD values are plotted for the equilibrated segment of the first production replicas of all solvents on Figure 3.2. To make the plots more readable, the moving average of residue mobility, calculated with a symmetric window with a width of nine residues, is plotted. The numbering of the residues on this figure and in the accompanying description simply describes the order in which Amber18 considers them, therefore these numbers cannot be straightforwardly compared with other numberings published in the literature. At first glance the three plotted curves appear fairly similar. The common characteristics include higher mobility of the terminal residues and lower average RMSD values for residues between 200 and 800. The section between residues 500 and 800 is roughly where the active site of the protein is located. Since the conformation of the residues around the active site is crucial for the activity of the protein, it is not surprising that these regions are more stable in the absence of substrates, than some other, less critical areas. The low mobility of the active site is confirmed again in Section 3.7.2, where its structure is more explicitly considered.

Upon more careful inspection, some differences that are worth mentioning can be identified between the three mobility curves. Most notable are the unusually high RMSD values around residues 870-880 observed in the trajectory with phenol as the cosolvent. These residues correspond to the terminal regions of the NSP12 and NSP7 proteins of the RdRp complex. Their higher mobility in itself is therefore not surprising. The fact that the RMSD values are outstandingly high only in the phenol trajectory could indicate some occasional interaction of these residues with the phenol probes. This observation signals that these regions could be of utmost interest in the subsequent docking calculations. Unfortunately, no binding pocket was found near this terminal site during the explicit docking calculations (see Section 3.6). Regardless, the fact that the phenol probes are able to induce significant changes in the protein structure around these regions is worth mentioning. A further difference between the three curves is the larger variance of the per residue RMSD values observed when the cosolvent MD trajectories are considered. Firstly, this manifest itself in the ever so slightly larger standard deviations shown for these trajectories. This difference shows most clearly around residues 1100-1192, where the average mobility is similar for all solvents, but much larger standard deviation areas can be seen for the cosolvent trajectories. Secondly, the differences between the average RMSD values between neighboring residues are also in general larger for the two cosolvent trajectories. This results in much deeper valleys and higher, more pronounced spikes in the bottom two curves of Figure 3.2. This latter observation can lead one to assume that the effects of the cosolvent probes are quite local in nature: they can significantly change the conformation of the handful of residues they are directly interacting with but leave the larger scale structure of the protein more or less intact.

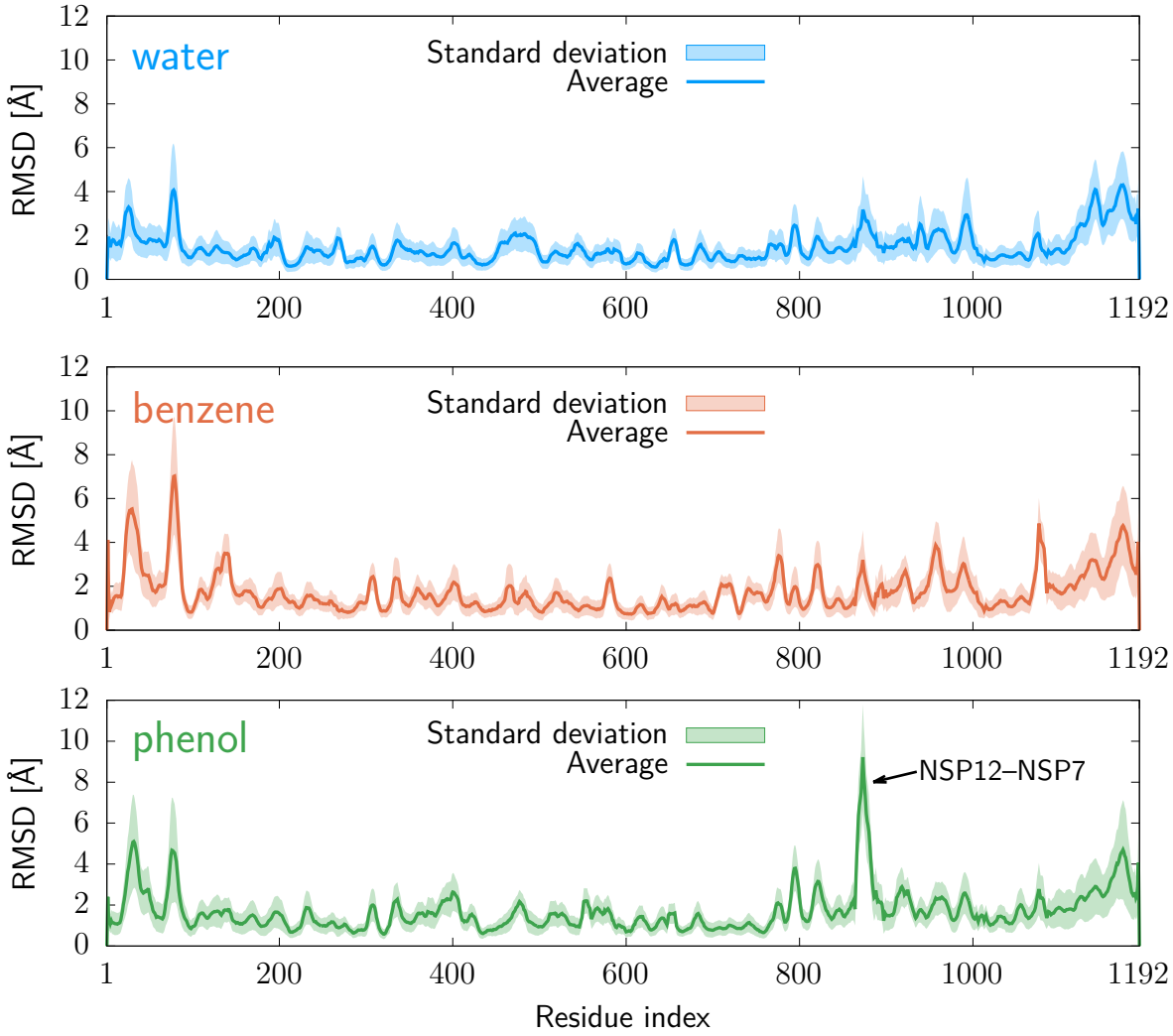

Figure 3.2: Average RMSD values for the protein residues in the three different solvent trajectories. The transparently coloured region around the average values represents the standard deviation throughout the simulation. The first replica for each solvent is plotted. The label "NSP7–NSP12" on the graph for the phenol cosolvent refers to the boundary region of the two nonstructural proteins.

3.2.3 Conclusions

The RMSD deviation values of the protein structure during the MD trajectories have been analysed. From the time evolution of this value during the simulations, the equilibrated section of the trajectories have been identified. The average and standard deviation of the per residue RMSD values have also been investigated. Based on this data, certain terminal regions of the protein have been identified as highly mobile, especially if phenol is utilised as cosolvent. On the contrary, the region around the active site of the protein has been classified as more rigid and stable in all MD trajectories. All in all, the available data indicates that the MD simulations sampled realistic protein structures with the CMD trajectories perhaps visiting a larger variety of conformations. With these trajectories in hand, the selection of a handful of representative

protein conformations can now be performed, through the technique of trajectory clustering.

3.3 Selecting representative protein conformations

As mentioned in Section 3.1.3, the `dbscan` algorithm of `cpptraj` is used to perform the clustering of the trajectories. The clustering is carried out separately for the three solvents, with the three replicas of each of them concatenated and treated as a single trajectory. In order to carry out a successful clustering of the trajectories, first the k and ε parameters of the density based clustering algorithm have to be tuned. On top of performing this tuning of the parameters, the effects of considering only the alpha carbon atoms for the RMSD calculations instead of all heavy atoms of the protein are also evaluated. Finally, possible redundancies in the set of representative protein structures are investigated.

3.3.1 Tuning the parameters of the clustering algorithm

The tuning of the parameters of the `dbscan` algorithm is performed by systematically varying the values for these parameters to see which combination yields the most optimal clustering. Additionally, clustering based only on the RMSDs of alpha carbon atoms is compared with all heavy atom RMSD clustering. The parameter k is varied between four and seven, as the authors of `cpptraj` write in the User's Manual [107], that values lower than four do not usually result in any improvements, while the value of of seven is already clearly inferior. The other parameter ε is varied from 0.95 Å to 1.3 Å in the case of alpha carbon clustering, and from 1.3 Å to 1.6 Å when all heavy atoms are considered. The computation of the RMSD distance between all frames of a trajectory is much more demanding if on top of the alpha carbons all other heavy atoms are considered as well. To speed up these computations, the technique of sieving is utilised: only every other frame is considered explicitly during the clustering, the remaining frames are simply added to the cluster with the cluster representative most similar to them. To measure the quality of the clustering four metrics are utilised. Two of these have already been discussed, namely the DBI and the pSF. Since both of these scores are heavily influenced by the number of obtained clusters [89], the comparison of their absolute values between different MD trajectories has limited meaning. Instead, the trends arising in these metrics through the systematic variation of the clustering parameters can be interpreted to optimise these parameters. At this point it is useful to reiterate, that low values of DBI and high values of pSF are desirable. The other two descriptors utilised to describe the quality of the clustering are the number of noise frames (frames not included in any cluster), and the number of clusters defined by the algorithm. The number of noise frames should clearly be

kept low to avoid missing any important conformations, only because it is visited very rarely and is therefore considered an outlier by the algorithm. Finally, while a high number of clusters is desirable as it can result in a wider variety of protein conformations, the computational limitations of performing explicit docking calculations to each representative conformation with thousands of ligands should be kept in mind.

On Figure 3.3, the descriptors of the water trajectory clustering can be seen, for the case when only the alpha carbons are considered during the RMSD distance calculations. Considering the top two plots at first, it can be observed that the DBI and pSF values are zero if ε is greater than or equal to 1.2 Å. The reason for this is that above this ε value all frames of the trajectory are grouped into a single cluster, for which these descriptors cannot provide a meaningful value. Since a single cluster is clearly not ideal, these large ε values do not need to be considered during the search for the optimal parameters. Focusing instead on parameter k, the most significant differences between the different values for this can be discovered on the pSF plot. Here, the curves with k= 4 or 6, reaching their peak at ε=1.1, are clearly superior to the other two. The curve corresponding to $k = 7$ is somewhat of an outlier on this graph, with its peak pSF at ε=1.15 Å instead of 1.1 Å. On the DBI plot, the variation of k has much more limited effects. In fact, all curves are more or less constant if ε is smaller than or equal to 1.1 Å, at which point the DBI values drop rapidly and become zero at 1.2 Å. Considering that ε=1.1 Å is the point at which the DBI values start decreasing, it is reasonable to assume that it is at this point that some significant changes are occurring in the way the clusters are defined. Together with the fact that ε=1.1 Å provides clusterings with the best pSF values, this observation makes this value of ε a promising candidate to be the optimal choice. By looking at the bottom two plots of Figure 3.3, the number of noise frames and number of clusters can be observed. On these plots one can find further advantages of the ε=1.1 Å choice. These are that the number of noise frames start their rapid increase only at slightly smaller epsilon values, and that a reasonable number of clusters, thirteen, are obtained for this value.

In comparison, similar descriptors can be seen on Figure 3.4 for the clusterings when all heavy atoms are considered during the RMSD calculations. The trends shown on this figure look very similar to those observed when only the alpha carbons are considered. The ε values are "shifted" by about 0.35 Å, which can be explained if higher mobility is assumed for the non-backbone heavy atoms of the protein in comparison to the alpha carbons. A further difference is that now the curve corresponding to the k value of four produces the best pSF values by a large margin instead of being only slightly better in the case of alpha carbon clustering. The number of noise frames is even more favorable in this case, however the ratio of noise frames is already negligible for reasonable values of k and ε, when only the alpha carbons are considered. On the contrary, the number of defined clusters is far too low if all heavy atoms are considered, with only two clusters at ε=1.45 Å and k=4. To summarise, no clear advantage of the heavy

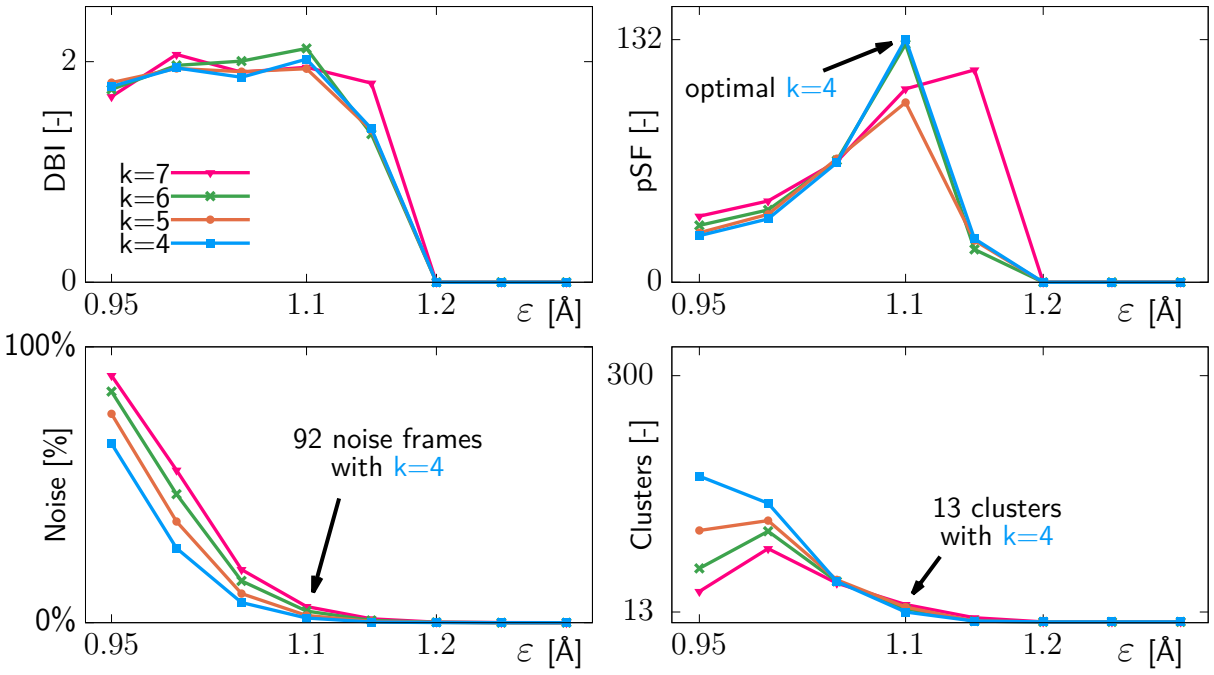

Figure 3.3: Plots of the clustering descriptors utilised for the tuning of the `dbscan` parameters. The descriptors were obtained by clustering the MD trajectory with water as the solvent and considering only the alpha carbon atoms for the RMSD calculations. The ε parameter in units of ångströms are shown on the horizontal axes in all cases, while on the vertical axes the various unitless descriptors are shown. From the top left in clockwise direction: the Davies–Bouldin index, the pseudo-F statistic, the number of clusters and the number of noise frames are shown.

atom clustering is found, and the significantly increased computational costs of considering much more atoms for the RMSD calculations make this type of clustering an inferior option compared to considering only the alpha carbons.

Next, the performance of the clustering algorithm is examined on the benzene cosolvent trajectory. The usual descriptors are plotted on Figure 3.5, with only the alpha carbons considered for clustering. Contrary to the water trajectory, in this case the number of clusters does not decrease to a single one at higher ε values but instead is saturated at three. As a consequence, the DBI and pSF values on the top graphs do not vanish for these values of ε. Instead, a sudden shift can be observed between ε=1.0 Å and 1.1 Å for both metrics, while for values higher or lower than these, the curves are more or less constant. The facts that this shift is occurring near $\varepsilon = 1.1$ Å, and that this value is already in the more favorable interval for both metric curves, highlight the attractiveness of choosing 1.1 Å as the value of the ε parameter. The pSF value at $\varepsilon = 1.1$ Å of the curve associated with $k = 4$ is again one of the best along with $k = 5$. The bottom two plots reveal no surprises: the number of noise frames is negligible with the parameter values being seriously considered, while the number of clusters stagnates around the reasonable value of three and increases sharply only for ε values less than 1.0 Å.

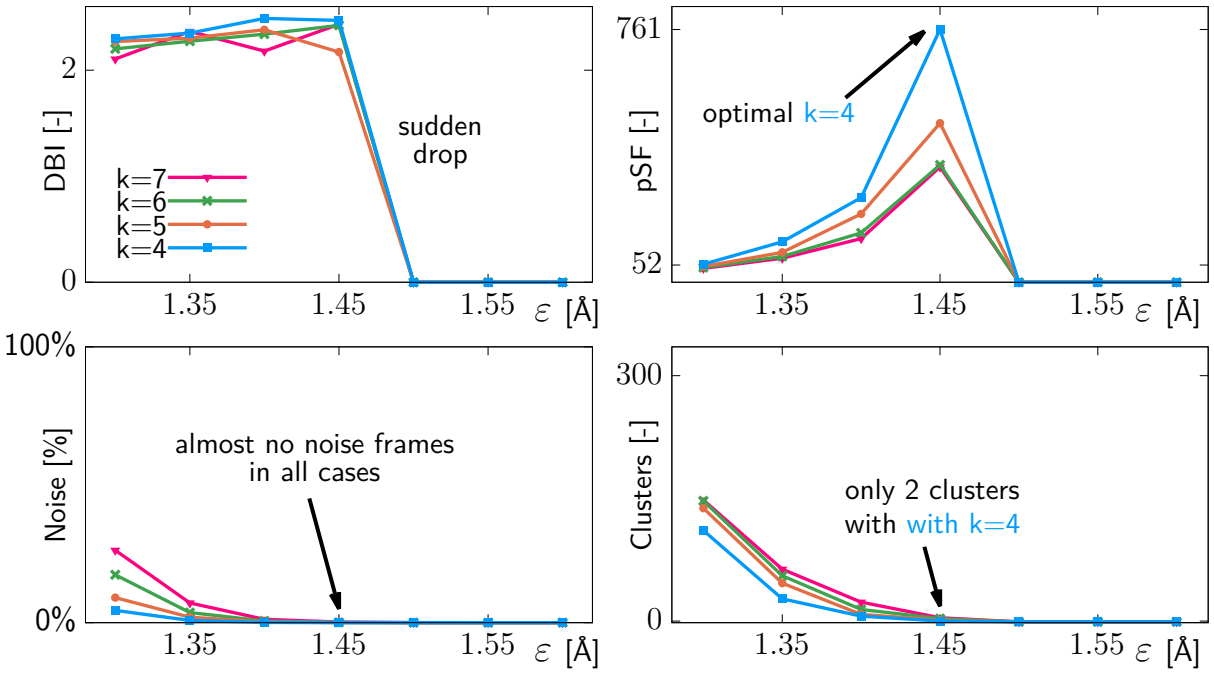

Figure 3.4: Plots of the clustering descriptors utilised for the tuning of the `dbscan` parameters. Descriptors were obtained by clustering the MD trajectory with water as the solvent and considering all heavy atoms for the RMSD calculations. The ε parameter in units of ångströms are shown on the horizontal axes in all cases, while on the vertical axes the various unitless descriptors are shown. From the top left in clockwise direction: the Davies–Bouldin index, the pseudo-F statistic, the number of clusters and the number of noise frames are shown.

Similar plots have been prepared for the phenol trajectory as well as for the heavy atom clustering of both cosolvents. However, since these are very similar to the graphs shown here already, they do not provide any new information but rather only reaffirm the conclusions that can be reached by considering the already presented figures. Therefore these plots are omitted here for brevity.

From the above analysis it can be seen that $\varepsilon = 1.1$ Å and k=4 are reasonable choices for the `dbscan` parameters, along with considering only the alpha carbon atoms during the RMSD calculations. For all three trajectories the quality of clustering achieved with these parameters are among the best ones investigated. The clustering performed in this manner selects 19 total protein conformations of which 13 are from the water trajectory while 3 are selected from the benzene and phenol trajectories each. Similar numbers of protein structures have already been utilised for ensemble docking calculations [1, 2], moreover the docking of the approximately 2000 selected ligands to all 19 snapshots is still well within reach with the computational resources available to us. To further investigate the suitability of the selected parameters, the potential redundancies in the resulting set of 19 representative protein structures were examined.

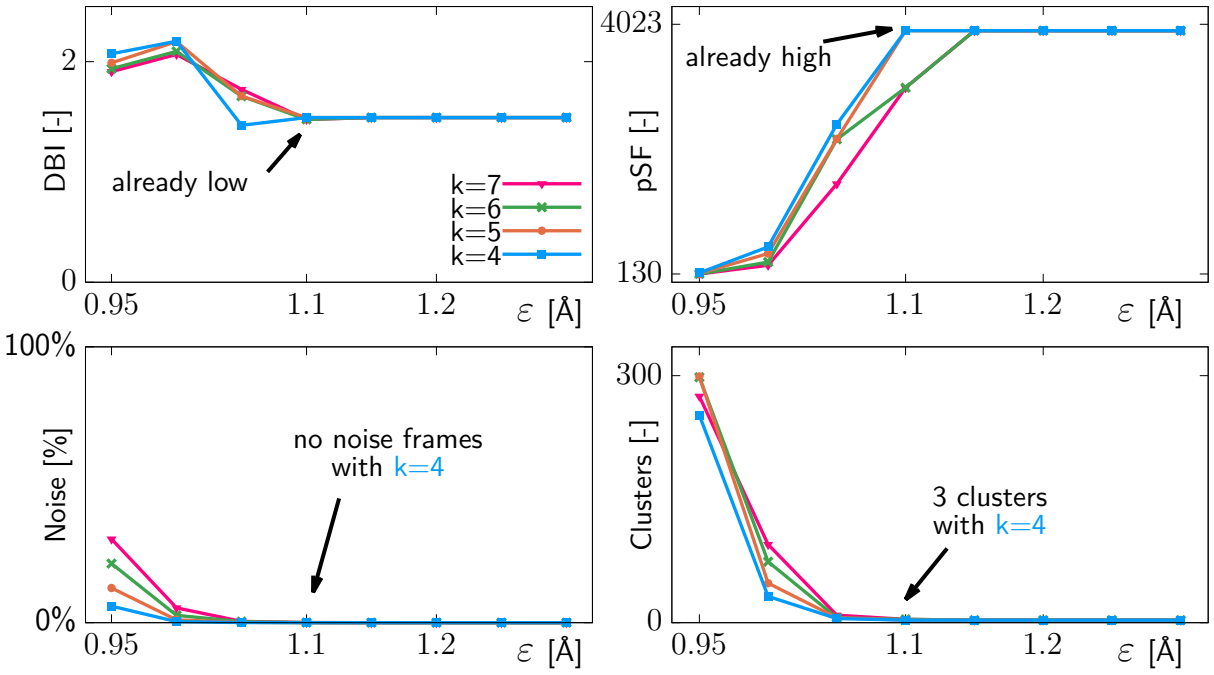

Figure 3.5: Plots of the clustering descriptors utilised for the tuning of the `dbscan` parameters. Descriptors were obtained by clustering the MD trajectory with benzene as the cosolvent and considering only the alpha carbon atoms for the RMSD calculations. The ε parameter in units of ångströms are shown on the horizontal axes in all cases, while on the vertical axes the various unitless descriptors are shown. From the top left in clockwise direction: the Davies–Bouldin index, the pseudo-F statistic, the number of clusters and the number of noise frames are shown.

3.3.2 Evaluating the chosen representatives

Since the clustering of the trajectories obtained with different solvents is carried out independently from each other, it is possible that some cluster representatives coming from different trajectories are quite similar to each other. This redundancy would clearly not be optimal as it increases the computational requirements of the ensemble docking calculations without providing much additional information. However, it is expected that the trajectories calculated with different cosolvents visit considerably different protein conformations and therefore significant redundancies between conformations coming from different solvents would be surprising. Nonetheless, this potential redundancy is worth investigating as its presence could indicate that the cosolvent simulations are not performing as expected. To this end, another clustering is performed utilising the parameters selected in the previous section, but with all trajectories considered at once. This clustering yields 18 cluster representative structures which is only marginally less than the 19 obtained with the original clustering scheme. The fact that the clustering algorithm cannot merge many clusters coming from different solvent trajectories, thus returns a similar number of clusters as when the trajectories are considered individually, signals that these trajectories indeed visit markedly different conformations.

To further confirm the assumption that conformations coming from trajectories with different solvents are more dissimilar to each other than conformations coming from the same trajectory, the RMSD distances between all cluster representatives are calculated. More specifically, the 19 representative protein conformations obtained in the previous section are taken, and RMSD values between all possible pairs formed from them are calculated, considering only their alpha carbons. By looking at the distribution of these RMSD values for conformation pairs obtained from the same or from different MD trajectories, one can compare the intra- and intertrajectory similarity of protein conformations. On Figure 3.6, one can observe this data, grouped by the solvent pairs from which the protein conformations are obtained. The various curves on this plot represent the frequencies with which some RMSD value is found among the distances calculated between cluster representatives of the two given solvents. For example the blue curve represents the distribution of the RMSD distances between cluster representatives coming from the water solvent trajectory, while the yellow dashed curve represents this distribution measure between the representatives of the water and benzene trajectories. The most noticeable feature of this graph is that the intratrajectory distances are noticeably smaller than the intertrajectory ones, with the solid curves being to the left of the dashed ones. There is only a single outlier benzene conformation, which is quite dissimilar to all other cluster representatives coming from this trajectory. To summarise, the data represented on this graph validates our assumption that the cosolvent trajectories visit conformations that are distinct from those visited by the traditional MD simulation with water as the solvent.

Since the extent of redundancy present in the selected set of 19 protein conformations is confirmed to be minimal and the cosolvent trajectories are shown to visit new conformations, the obtained set of structures is accepted for ensemble docking calculations to be performed with them.

3.3.3 Conclusions

It has been shown that the results of clustering a MD trajectory depend heavily on the employed parameters of the clustering algorithm. In the first part of this section, the effect of changing these parameters have been investigated on a variety of clustering quality metrics. The results have shown that there is no single choice of parameters that results in the objectively best clustering. However, after systematic experimentation with a range of parameter values, optimal or close to optimal settings have been selected. Moreover, the necessity of considering all heavy atoms for the RMSD calculations have been investigated and it was found to provide no significant advantages, to validate its increased computational costs. Finally, the set of cluster representative protein structures, obtained with the selected parameters, were analysed. The conformations coming from CMD trajectories have been shown to be markedly different from

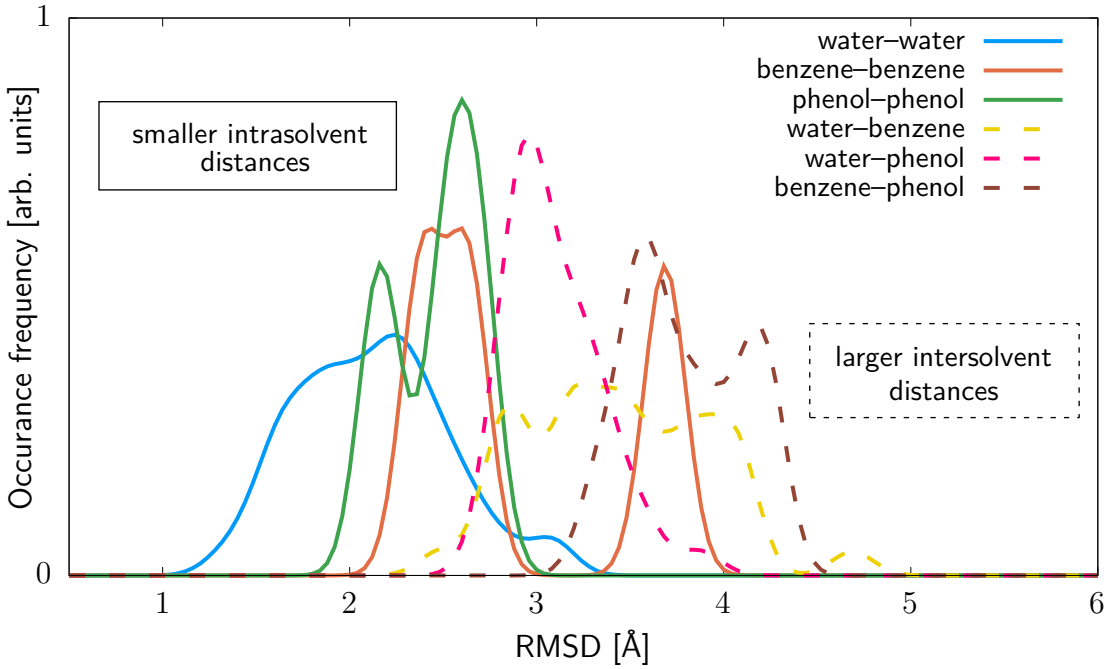

Figure 3.6: The distribution of the RMSD deviation distances between the selected protein conformation cluster representatives. The curves represent the frequencies with which a given RMSD value is found among all distances calculated between conformations coming from the trajectories denoted by their solvent in the legend.

the conformations coming from conventional MD trajectories with water as solvent. The set of cluster representatives were also checked for potential redundancies but nothing egregious was found.

3.4 Preliminary analysis of the docking calculations

In this section, some preliminary analysis of the docking calculations are presented. The findings presented here are utilised in later sections where the effect of protein dynamics and cosolvent simulations on the docking calculations are discussed. First, the accuracy of the docking calculations performed with AutoDock Vina are assessed by considering the crystal structure of the target and a handful of known binders. Then, the distribution of the calculated binding energies are examined and compared between protein conformations obtained from trajectories with different solvents. The data gathered here will be useful to determine the selection criteria for the best binding ligands, employed in Section 3.6. Finally, the importance of the various protein conformations are assessed by investigating whether any unique ligands bind to them.

3.4.1 Testing the performance of the docking protocol

In this section, the accuracy of the results provided by AutuDock Vina, with the employed settings described in Section 3.1.4, is evaluated. To this end, a number of confirmed binders for the RdRp of SARS-CoV-2 are obtained from Reference 14, along with the most important interacting residues for each binder. To obtain these results, the authors utilised the experimentally determined apo protein structure with PDB code 6M71. To compare the performance of AutoDock Vina, we have performed docking calculations with it, also considering this crystallized protein structure alone. However, there are still significant differences in the preparation of the protein structure and the execution of the docking calculations between their work and the one presented here. Most notably, they carried out the minimisation of the protein structure beforehand and treated key residues of the protein as flexible during the docking calculations. During both the structure preparation and docking calculations commercial software were utilised, meaning that we were unable to replicate the exact process. Nonetheless, it is expected that the published binders and their poses can be at least partially reproduced using the docking procedure adopted in this work.

The results obtained with AutoDock Vina and from Reference 14 are compared in Table 3.1 for seven ligands. This subset of the binders presented in the original article, comprises the ligands for which a 3D structure could be obtained from the ZINC database. During our docking calculations for these known binders, the settings described in Section 3.1.4 are utilised. The data shown in the last two columns of Table 3.1 were gathered by visually inspecting the twenty docked poses returned by AutoDock Vina, and comparing their position to the interacting residues, given in the second column of the same table. In the output of Vina, the top twenty poses are ranked based on their estimated binding energy. In the third column of the table, this rank is shown for the first pose returned by Vina, which interacts with the residues given in the second column. In the fourth column, the total number of poses which interact with the specified residues is shown. It can be seen, that there are only two ligands, out of the seven considered, for which AutoDock Vina does not find the pose reported in the literature. For these two ligands, a binding site near the active site of the protein is clearly favored by Vina, over the respective sites of the ligands reported in Reference 14. Since the referenced study defined its docking region to encompass only the active site it is possible although unlikely, that the binding site found by Vina is indeed more favourable for these ligands, and it is simply not considered in Reference 14. More likely is that for the binding of these two ligands some rearrangements of the active site residues are necessary, which was accounted for in the work of Ahmed *et al.* by employing flexible residue docking, but is ignored in the present work with the single rigid protein structure considered. For the other five ligands, the pose predicted in the literature is found by Vina as well, and in three cases it is among the top five returned

Table 3.1: Comparison of the results of AutoDock Vina to the binding poses reported in the literature. The data in the first two columns of the table is reproduced from Tables 1 and 2 of Reference 14, originally published by Ahmad *et al.*. The third and fourth columns show the rank of the best pose and the total number of poses returned by AutoDock Vina that interact with the residues shown in the second column.

ligand	reported interacting residues	rank of first matching pose	number of matching poses
ornipressin	ASP760, THR591, ASP865, GLN815, SER814, CYS813, GLU811, TYR619	1	8 (40 %)
benzquercin	ARG553, LYS798	–	0 (0%)
cisatracurium	ARG553, ARG555, GLU811	2	1 (5 %)
ditercalinium	ASP623, ASP760	9	1 (5 %)
examorelin	ASN691, HIS810	15	3 (15 %)
nacartocin	LYS798	4	3 (15 %)
pegamotecan	LYS621	–	0 (0 %)

poses. Moreover, in the case of three ligands, the predicted interacting residues are reproduced by more than one pose, indicating the stability of these results.

All in all, the results presented in this section show, that even though a much less sophisticated docking procedure is utilised in the present work than that of Reference 14, it is yielding sensible results for this set of known binders. Since the main focus of this study is not on the accuracy of the docking calculations in itself, but rather on the investigation of the effects of cosolvents and protein dynamics on the docking results, the proposed docking procedure is deemed suitable and is therefore utilised throughout the rest of our work.

3.4.2 Binding energy distribution

After the adeptness of the docking calculation protocol has been verified, the distribution of the binding energies obtained from docking the approximately 2000 FDA approved drugs to the representative protein conformations is examined. The best binding energy for each protein structure-ligand pair is extracted from the output of AutoDock Vina. The binding energies are grouped based on the solvent utilised to obtain the protein conformation used during the docking calculation. The distributions of binding energies obtained in this manner are plotted on Figure 3.7. Perhaps the most noticeable feature of this plot is that the distribution for the water solvent is much wider around the -8 kcal/mol mark, then it narrows much more rapidly

at lower binding energy values, in comparison to the distributions for the cosolvents. On the other hand, the cosolvent distributions extend to much lower binding energies, with the best binding energy for the phenol conformations being -12.4 kcal/mol. Accordingly, the number of unique ligands with favorable binding energies are larger for the cosolvent trajectories as can be observed from the coloured numbers to the left of each distribution on the figure. The presence of such general and clear trends in a large number of binding energies calculated for a variety of ligands, most likely indicate some significant differences in the protein conformations utilised during the docking calculations. For example, the above described trend of lower binding energies for the cosolvent structures could be explained, if one assumes that one or more binding regions are more accessible to ligands in the conformations obtained from the cosolvent trajectories, than in those obtained from the water solvent trajectory. Finally, it is noted that if the binding energy is utilised to select the best binders, the threshold of -9.9 kcal/mol seems like an appropriate threshold, as it selects binders in a relatively balanced way from all trajectories while also not selecting so many total ligands that would make further analysis cumbersome.

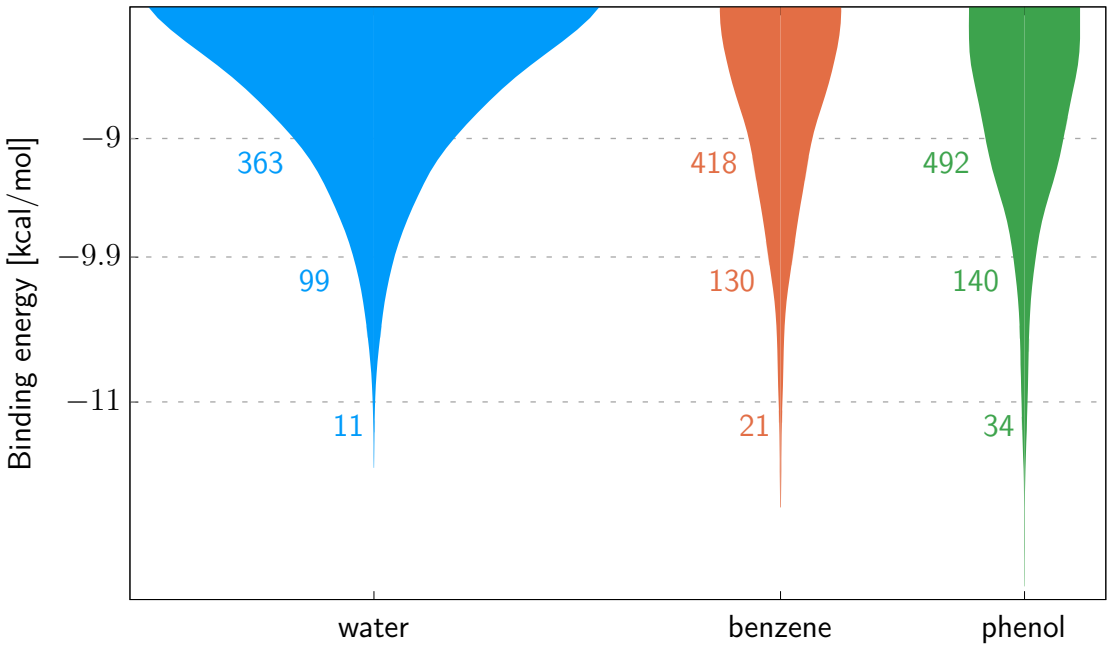

Figure 3.7: The distribution of the ligand binding energies as calculated by AutoDock Vina. The data is grouped according to the solvent utilised to obtain the protein structure the ligand is being docked to. The numbers to the left of the distributions indicate the number of unique ligands which have a binding energy lower than the given value for at least one of the conformations obtained from that solvent. Note that the distributions were normalized separately for each solvent and therefore the widths of the curves obtained for different trajectories cannot be directly compared.

3.4.3 Necessity of multiple protein conformations

Since the binding energy of a given ligand is very strongly dependent on the conformation of the target protein, it is expected that some ligands bind strongly only to a specific protein conformation. Moreover, in Section 3.3.2, it has been demonstrated that the different solvent trajectories sample markedly dissimilar protein conformations. Consequently, it can be expected that some strong binders can only be found when the docking is performed to a specific cluster representative conformation or the cluster representatives of a trajectory with a specific solvent. If this is the case, then one should observe that docking to one or even to just a few conformations is not sufficient to find all the best binding ligands for a given protein. To test this hypothesis, the selection criterion for choosing the best binders is set to -9.9 kcal/mol, as it selects sufficiently many ligands for each trajectory. The number of ligands selected by this criterion for each trajectory can be seen on Figure 3.7. The binders selected in this way are subsequently grouped according to which cluster representative protein structures they bind to. Note that a ligand can be in multiple groups simultaneously, if its binding energy is lower than -9.9 kcal/mol with more than one protein conformation. Next, the cluster representatives are ordered in descending order according to the number of ligands they bind. Finally, the representative with the highest number of binding ligands is selected, and its ligands are removed from the lists of all other conformations, which happen to also bind that ligand. This last step is repeated until all conformations and all binding ligands are accounted for. From the data obtained in this manner, cumulative ligand binding plots are created and displayed on Figure 3.8. These diagrams visualise the number of protein conformations which are necessary to find a certain percentage of all the best binding ligands, discovered through any of the cluster representatives. On the left pane of Figure 3.8, the ligands discovered by different solvent trajectories are separated. The cluster representatives are ordered on the horizontal axis in decreasing importance (from many to few new ligands discovered). It can be observed that with a single protein conformation, only about 75 % of all binders of the cosolvent trajectories and less than 50 % of the binders of the water solvent trajectory would have been discovered, even if the conformation with the most binders would have been utilised. Moreover, all conformations coming from cosolvent trajectories have ligands that only bind to them, and not to other conformations of the given trajectory. In the case of the 13 representative conformations obtained from the water solvent trajectory, only a single one could have been neglected to still find all of the 99 best binders of that trajectory. These observations confirm, that even the cluster representatives of a single MD trajectory harbor binding site conformations, that are different from the other conformations of that trajectory in a way that affects the binding of ligands. These results therefore validate the increased computational costs of docking to an ensemble of protein structures, as they show that significantly more binders can be discovered

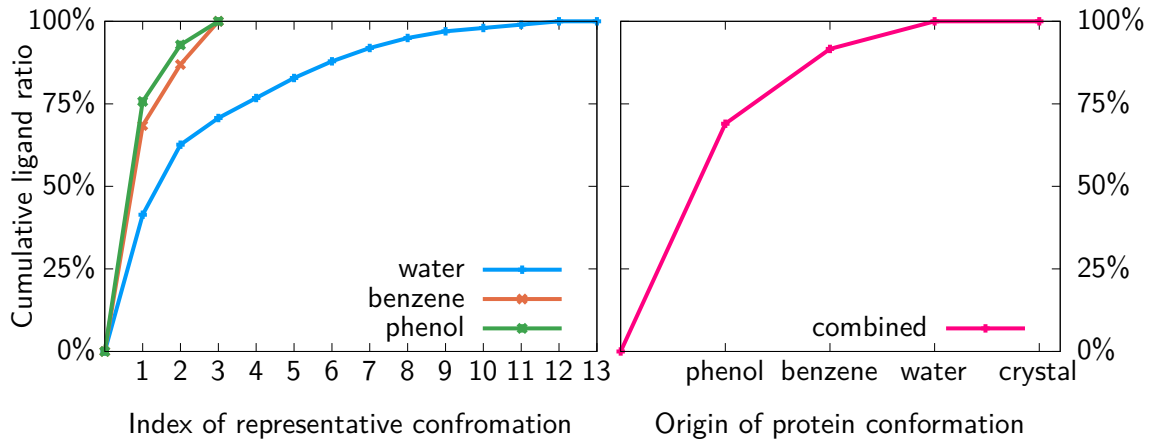

Figure 3.8: Cumulative ratio of the ligands that would have been found if only a subset of the available protein conformations was considered for docking. On the left pane, the protein conformations obtained from different solvent trajectories are treated separately, with cumulative ligand ratios calculated for the cluster representatives of each trajectory. On the right pane, all ligands discovered by any trajectory or the holo crystal structure are merged together, and the cumulative ligand ratios are calculated for all conformations of the trajectories or the crystal structure. The cluster representatives or trajectories on the horizontal axes are ordered in decreasing importance, from most to least new ligands discovered.

by utilising multiple protein conformations than in the case of a single considered structure. On the right pane of Figure 3.8, a similar curve is shown but instead of considering the ligands of the MD trajectories separately, they are merged together. In this case, the horizontal axis displays the various solvents that the conformations were obtained from, while on the vertical axis the cumulative ratio of ligands discovered by the cluster representatives of the trajectory with the given solvent is shown. It can be observed, that the conformations of the phenol cosolvent trajectory have the highest number of unique binding ligands, accounting for about 75 % of all ligands. However, to discover the last 25 % of the best binders, the other two trajectories are also necessary. The last data point of this plot corresponds to the binders discovered by docking to the apo crystal structure of the protein. It is clear that no new ligand is found by docking to this structure, that would not have been already found by one of the conformations obtained from the MD trajectories.

To conclude, it has been shown that almost all of the protein conformations obtained from the MD trajectories are necessary to be considered in order to find this set of binding ligands. In other words, it is barely possible to reduce the number of protein conformations without sacrificing on the completeness of the found binders. This conclusion validates the hypothesis proposed at the beginning of this section, and justifies the computationally expensive use of the ensemble protein conformations during the docking calculations.

3.4.4 Conclusions

First and foremost, results proving the adeptness of the adopted docking protocol have been shown. AutoDock Vina was able to find the reported poses for a large portion of the best binders taken from the literature. Subsequently, the binding energy distribution obtained by docking to the ensemble of protein structures were examined. It was observed that the cosolvent trajectories and especially the conformations obtained from the phenol cosolvent simulation, yield more ligands with highly favorable binding energy. Finally, the importance of the various protein conformations in discovering binders has been assessed. It was found, that almost all protein structures in the ensemble have unique binding ligands that do not interact strongly with any of the other protein conformations.

3.5 Criteria for selecting the best binding ligands

The inaccuracies of the binding energy estimates calculated by docking programs have already been mentioned in Section 2.2. Additionally, it is noted that this binding energy is naturally dependent on the size of the ligand being considered. Since ligands with a high number of atoms can benefit from more interactions between their atoms and those of the protein, this metric makes the selection of ligands that contain only a small number of very favorably interacting moieties difficult. These imperfections of the binding energy estimate make it a somewhat unreliable criterion for the selection of the best binding ligands. In the same section, a possible approach to remedy this issue has been introduced: the spatial distribution of the best poses calculated for each ligand can be utilised for the selection of the binders, either in itself or in some combination with the binding energy estimate of the docking program. However, it has also been discussed, that AutoDock Vina automatically performs a clustering of the found docked ligand poses, only returning the requested number of cluster representatives as a result. This clustering removes much of the information about the spatial distribution of the binding poses from the original docking results, greatly hindering the application of the above mentioned improvements to the ligand selection criteria. To remedy this issue, in this section, an alternative approach to enhance the selection of the best binders is devised and evaluated.

3.5.1 Definition of the alternative selection method

The key points during the design of this alternative ligand selection criterion are to avoid being dependent on the size of the ligand as much as possible, and to consequently enable

the selection of ligands harboring particularly effective chemical moieties. Furthermore, relying on information about the spatial distribution of the best ligand poses is avoided, so that the criterion can be used to filter the results of docking programs that do not provide such information, such as AutoDock Vina, as well. Our first attempts towards devising such a criterion revolved around scaling the binding energy estimate by the number of heavy atoms in the considered ligand. Unfortunately, this approach turned out to favor small ligands too much, while the inclusion of some parameter to scale this bias down, seemed too arbitrary. Consequently, the idea of scaling the reported binding energies was discarded, and the procedure presented here is based solely on the binding energies themselves, as reported by the docking program for the best poses of the given ligand. More specifically, it evaluates the energy gap between the two best poses found for the ligand. The idea behind this approach is that a truly good docked pose should be something out of the ordinary compared to the myriad of other suboptimal poses. Given the scarcity of such extraordinary minima of the scoring function, it is expected that only a small fraction of the stochastic optimisation runs, performed during a docking calculation, will find the corresponding ligand poses. In practice, this phenomenon could manifest itself in the form of a single docking pose with a highly desirable binding energy estimate, found among the many other mediocre poses in the results of a docking calculation. A large gap between the binding energies of the two best docked poses for a given ligand could indicate such a situation and consequently signal a true binder of the target protein. Therefore, it is proposed to categorise ligands as binders (non-binders) based on whether the gap between their two best binding energies is larger (smaller) than a given threshold. It is expected that this energy gap does not significantly depend on the size of the ligand and is therefore a less biased criterion than the binding energy itself. Furthermore, it is possible that this criterion functions best in combination with a simple binding energy threshold, where it would serve to eliminate some large ligands that are only categorised as binders due to the bias of the latter criterion towards larger ligands.

3.5.2 Evaluation of the alternative selection criteria

As the first step of evaluating this criterion, its selected ligands are compared to those selected based on their binding energy. If very few or no common ligands are found between the two criteria, it could indicate that this alternative selection method does not perform as expected, because even though the binding energy estimate provided by the docking program is an imperfect one, in general it will select reasonable ligands. To carry out this comparison, two sets of selection thresholds were determined for the binding energy and energy gap criteria each. The thresholds in these sets were constructed separately for the protein conformations coming from different trajectories, such that they select approximately one hundred or twenty ligands

Table 3.2: The energy thresholds determined for the binding energy and energy gap criteria, along with the number of ligands selected for the various protein trajectories.

solvent	desired ligands	E_{bind} [kcal/mol]	ligands selected by E_{bind}	E_{gap} [kcal/mol]	ligands selected by E_{gap}
water	20	-10.7	20	1.6	21
	100	-9.9	99	1.2	96
benzene	20	-11.0	21	1.4	16
	100	-10.0	106	1.0	88
phenol	20	-11.3	18	1.3	19
	100	-10.2	97	0.9	89
crystal	20	-9.5	20	0.8	24
	100	-8.8	110	0.6	90

for each trajectory. The values of the thresholds and the exact number of ligands selected by them for each trajectory is summarised in Table 3.2. Here, E_{bind} and E_{gap} denote the binding energy and energy gap thresholds respectively. It is noteworthy, that while the binding energy threshold which selects about one hundred ligands is stricter for the cosolvent trajectories than for the water one, in the case of the energy gap criteria the opposite of this relation is true. For example, to select about one hundred binders for the conformations of the water trajectory a binding energy threshold of -9.9 kcal/mol is sufficient, while for the phenol cosolvent trajectory the stricter -10.2 kcal/mol criterion has to be utilised. On the contrary, with the energy gap criterion, a higher threshold of 1.2 kcal/mol selects about one hundred binders for the water solvent conformations, but a looser value of 0.9 kcal/mol is sufficient to achieve the same for the phenol cosolvent trajectory. Furthermore, both thresholds have to be set to significantly looser values if one wishes to select the same number of ligands for the crystal structure. This last observation could indicate that in general, binding to the crystallised protein conformation is weaker, and therefore utilising MD to simulate *in vivo* conditions is beneficial for finding the best ligands. With the best binders selected, the ratio of ligands selected by both criteria can now be defined to be:

$$R = \frac{|S_1 \cap S_2|}{\max\left(|S_1|, |S_2|\right)}\,, \tag{3.3}$$

where S_1 and S_2 are the sets of ligands selected by one of the criteria.

On Figure 3.9, the ratio of ligands selected by both criteria are plotted separately for each trajectory and the crystal structure. From this figure, one can observe that about 15-20 % of the best one hundred ligands are common to both criteria. For the cosolvent trajectories and the crystal structure this ratio is somewhat higher, and for the water trajectory it is slightly lower. It is also clear from the graph, that the ratio of common ligands decreases when going from the top hundred to the top twenty ligands selected. This trend is the most pronounced for

the conformations obtained from the benzene trajectory, where no common ligands are found in the top twenty binders. Considering all of this, one can conclude that even though the two criteria select ligands in an entirely different manner, at least part of the ligands they favor are common to both of them. This gives us some confidence that the energy gap based selection criterion can indeed be suitable to augment simple binding energy based selection methods.

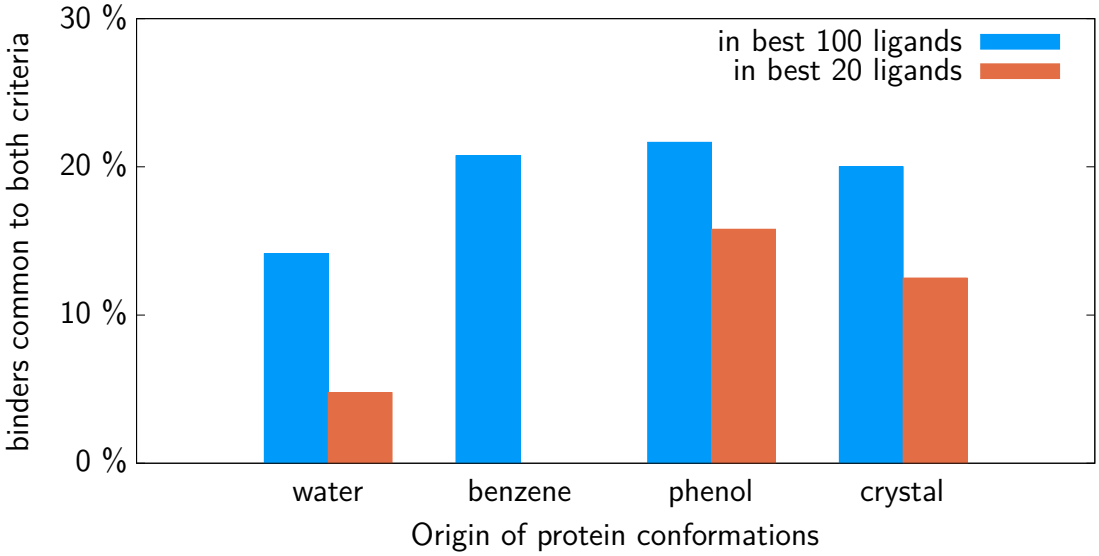

Figure 3.9: The ratio of ligands selected both based on their binding energy and the energy gap between their two best poses. Ligands binding to conformations coming from different solvent trajectories (or to the crystal structure) are treated separately. For the exact definition of the ratio plotted here, see Equation 3.3.

After observing that a substantial ratio of the selected ligands are common for the two selection criteria, the question naturally arises, whether these common ligands are especially suitable binders. This would validate the use of the composite ligand selection method involving the binding energy and energy gap criteria, described earlier in this section. Furthermore, it would be beneficial to confirm that the ligands selected based on the energy gap between their two best poses exhibit a favorable binding energy as well, thus further increasing the confidence in the adeptness of this ligand selection method. To investigate these properties of the energy gap criterion, the normalised distribution of the binding energies are plotted for the ligands present in the top hundred ligands of both selection methods. This distribution is then compared to a similar normalised distribution of the binding energies of the best hundred ligands of the binding energy criterion. Figure 3.10 collects these distributions separately for the ligands obtained for the three MD trajectories and for the crystal structure of the protein. Here, the more lightly coloured left half of each violin plot represents the normalised distribution of the binding energy of the ligands selected solely on the basis of their binding energy. On the right sides of the same violin plots with a darker color, the same distribution can be seen for the

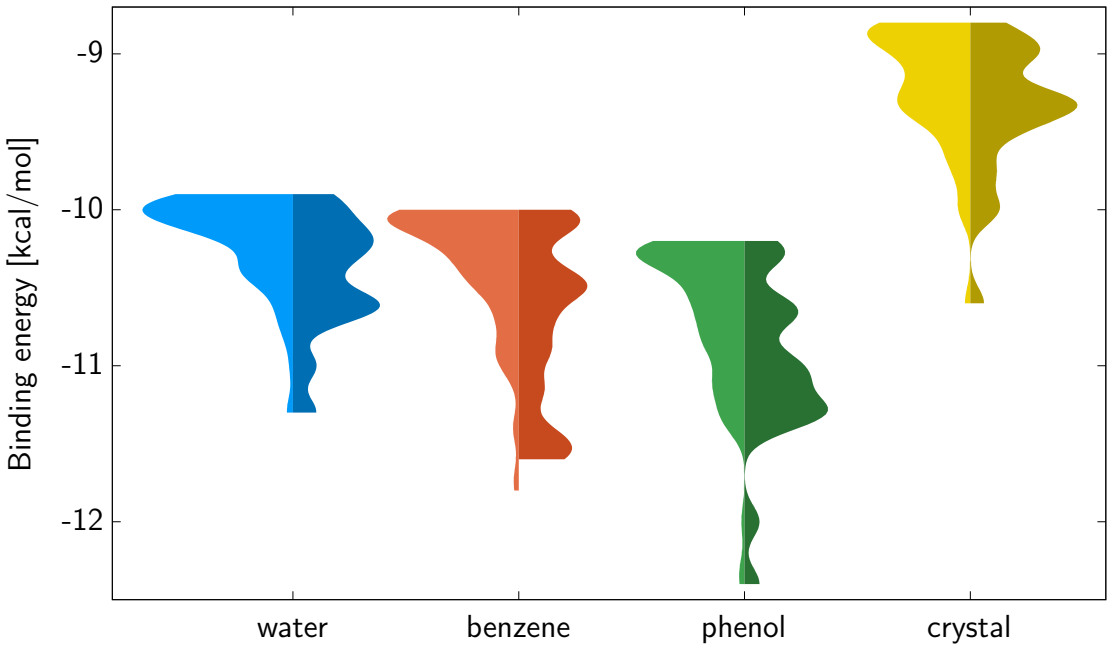

Figure 3.10: Comparison of the binding energy distributions for the ligands selected based on solely their binding energy and for the ligands selected both based on their binding energy and energy gap between their two best poses. The distributions are plotted separately for the ligands of the three trajectories and the crystal structure. The left half of each violin plot represents the distribution obtained for the ligands selected by their binding energy, while the right half denotes the same distribution for the ligands selected by both their binding energy and energy gap.

ligands which satisfy both the binding energy and energy gap criteria. The figure clearly shows that the binding energies of those ligands which are common to both selections, are lower (more favorable), than those of ligands selected solely by their binding energy. This trend is most pronounced for the ligands selected for the phenol cosolvent trajectories, where the bulk of the distribution is present just below the binding energy selection threshold on the left side of the violin plot and it is shifted almost a full kcal/mol lower on the right side. It is therefore confirmed that ligands satisfying both criteria tend to have a lower binding energy than ligands which satisfy only the binding energy criterion.

Lastly, the sensitivity of the two selection criteria to the type of solvent employed during the MD simulations are compared. To this end, the ratio of the ligands that are selected as good binders for at least one protein conformation per MD trajectory are plotted on Figure 3.11. The best one hundred (or twenty) ligands are collected separately for the conformations obtained from a trajectory simulated with a given solvent and their common ligand ratios are calculated analogously to Equation 3.3. If this ratio is high for a selection method, it indicates that that method selects a similar set of binders for all trajectories and is therefore not sensitive to the differences in protein conformation present between structures obtained from different CMD

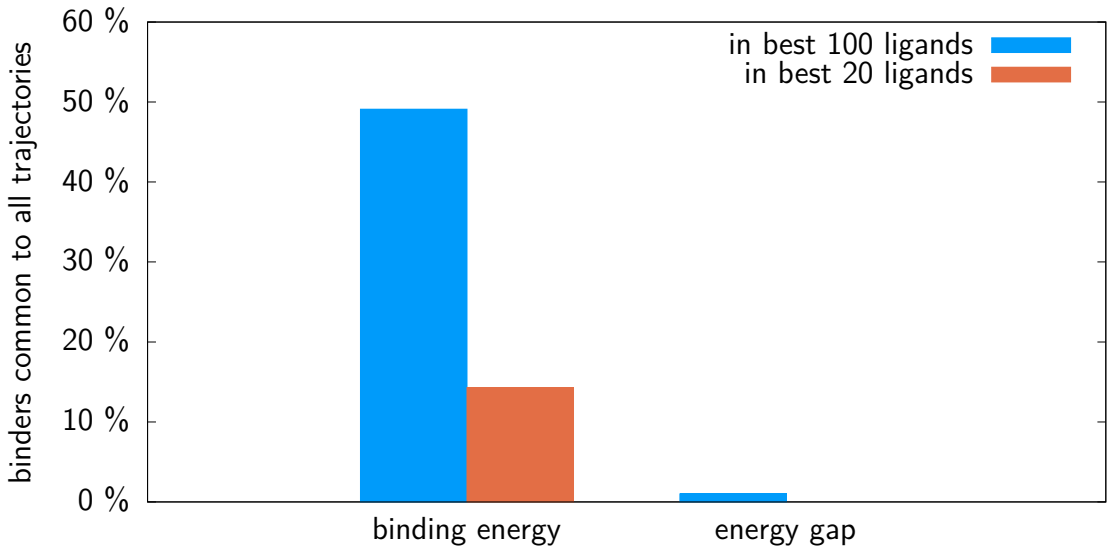

Figure 3.11: Ratio of ligands selected for all three MD trajectories. The top one hundred or twenty ligands were selected for each trajectory either with the binding energy or energy gap criterion, utilising the threshold values displayed in Table 3.2. The ratio of ligands that are common to all three trajectories were calculated analogously to Equation 3.3.

trajectories. Looking at the columns corresponding to the binding energy selection criteria on Figure 3.11, one can observe that about 50 % of the best one hundred ligands are common to all three trajectories. Therefore, while this ratio reduces considerably if only the best twenty binders are considered for each trajectory, it can be concluded that the binding energy selection criterion is not extremely sensitive to the various conformations obtained from trajectories with different solvents. On the contrary, when the results obtained with the energy gap criterion are considered, one finds a negligible common ligand ratio between the three MD trajectories. Only 1 % of the best one hundred ligands are considered good binders for all three trajectories, and no common ligands are found among the best twenty. Consequently, the energy gap criterion is deemed very sensitive to the conformational changes encountered between protein structures obtained from MD trajectories with different solvents.

3.5.3 Conclusions

In this section, an alternative ligand selection criterion has been devised, that is based on the energy gap between the two best poses for each ligand. It was shown, that this new selection method can be a promising alternative to the simple binding energy based selection, if information about the spatial distribution of the docked poses is not available. First, it was established that the two criteria select a significant number of common ligands, indicating that the energy gap between the two best poses of a ligand indeed correlates with its estimated

binding energy. Then, the ligands selected by the combined use of the two criteria has been shown to have significantly lower binding energies than the ligands selected by the same binding energy criterion in itself. Based on this observation, it can be expected that the two criteria can work together beneficially, with the energy gap criteria serving to eliminate the bias of the binding energy estimation towards larger ligands. Finally, the sensitivity of the two criteria to the conformational changes of the protein structure were compared. On the one hand, it was found that the binding energy criterion is relatively unaffected by the differences in conformation caused by considering different solvents. On the other hand, the energy gap criterion was found extremely sensitive to the same conformational differences, indicating that it could be especially effective in distinguishing ligands that only bind to specific protein conformations. These first results are encouraging enough to validate further experimentation with the energy gap based ligand selection method, discussed in Section 3.6.

3.6 Identification of binding sites

In the previous sections, the design and evaluation of the computational methodology employed in the present work has been discussed: the analysis of the MD trajectories, the tuning of the protein structure ensemble to be utilised for docking, the testing of the docking program with known binders of the target, statistical analysis of the docking results and criteria for selecting the best binding ligands have all been touched upon. With this preparatory work out of the way, the scene is now set for one of the main goals of the project: the evaluation of the effects of protein dynamics and CMD simulations to the formation of binding sites. To this end, first the ligands that are considered to be the most promising binders of the target are selected. Next, the areas of the protein to which these ligands bind are identified by visual inspection. Finally, the similarity of the binders of a given pocket is investigated and compared between various MD trajectories.

3.6.1 Selection of the most promising binders

The selection of the best binding ligands is performed according to the two selection criteria discussed in Section 3.5. In both cases, the energy threshold is determined in a way to select approximately one hundred ligands for the water solvent MD trajectory. For this purpose, a binding energy threshold of -9.9 kcal/mol, and an energy gap threshold of 1.2 kcal/mol are the most optimal. On Figure 3.12, the number of ligands selected for the three MD trajectories and the crystal structure are compared. From the diagram on the left it can be seen, that if the binding energy is considered as the selection criterion, ligands are selected quite evenly

for the three MD trajectories, while much less binders are found for the crystallised protein structure. On the right side of the same figure, similar data is shown for the case when the energy gap ligand selection criterion is utilised. The distribution of the binders selected with this approach is somewhat different than what can be seen on the left side of the same figure. In particular, the organic cosolvents yield significantly less binders than the water trajectory, with the phenol trajectory producing only one fourth as many. This could at least partly be explained by the fact that this ligand selection criterion is much more sensitive to the conformation of the protein, as it has been shown in Section 3.5.2. Due to this sensitivity one can assume that this criterion selects much more unique ligands for each cluster representative protein conformation even if they are obtained from the same MD trajectory. Since for each conformation a largely unique set of binders is selected and since the CMD trajectories only yielded three representative protein conformations each, while the water solvent trajectory yielded thirteen, it is not surprising that the water trajectory produces more unique ligands. This bias does not appear when the binding energy is used as the ligand selection criterion, because that criterion seems less sensitive to the changes in protein conformation. Therefore the fewer number protein structures in the case of the organic cosolvents are already sufficient to identify all important binders, while the additional ten conformations considered for the water trajectory do not result in many new unique binders. The results for the crystallised protein structure do not change qualitatively between the two selection criteria, with only very few identified binders in both cases.

3.6.2 Definition of the binding sites

With the set of promising binders selected, their binding regions on the surface of the protein can now be defined. Upon visual inspection of the most favorable docked poses of the best binders, nine major binding sites of the protein are identified. The locations of these sites are illustrated on Figure 3.13, while the interacting residues of each site are summarised in Table 3.3. The data collected in this table is obtained by visually inspecting the docked poses for the ligands binding to a pocket, and trying to select the protein residues that seemed closest to many of these poses. It is noted that binding site number four identified here, corresponds well to the active site of the protein, as reported in Reference 14, therefore the interacting residues shown for this pocket in Table 3.3 are taken from that work. After the binding sites of the protein have been identified, the number of ligands which bind to these sites are counted, separately for the MD trajectories coming from different cosolvents, and for the apo crystal structure.

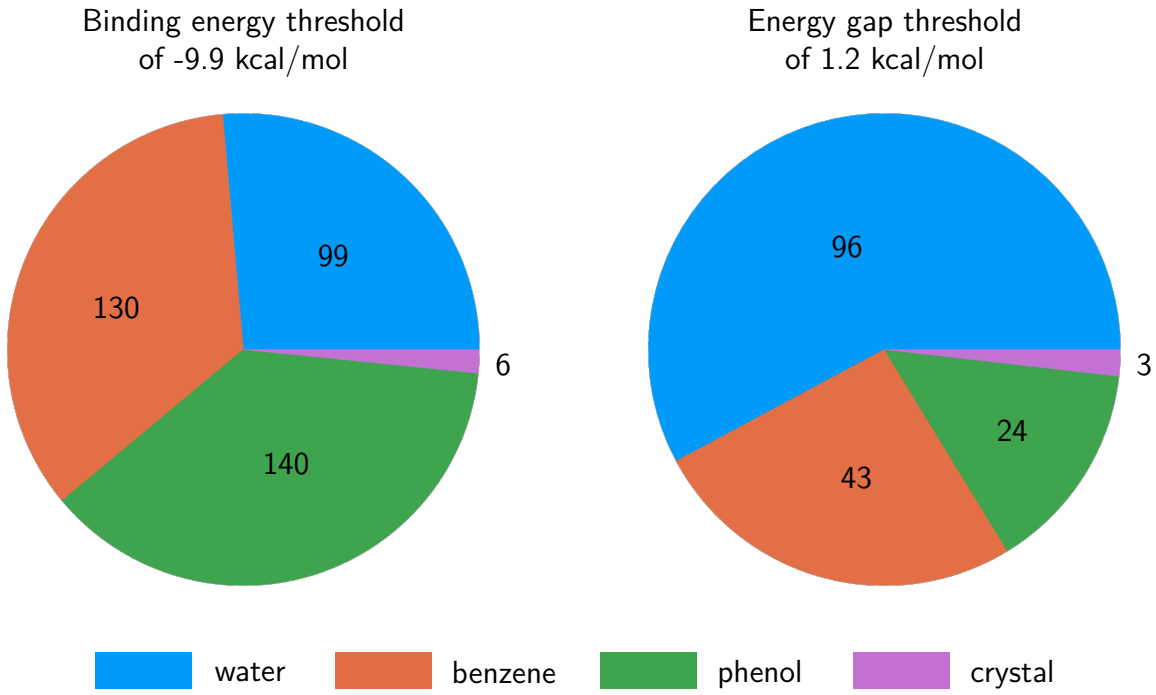

Figure 3.12: Comparison of the number of ligands selected for the three MD trajectories and the crystallised protein structure. On the left side, the ligand binding energy of maximum -9.9 kcal/mol, while on the right side an energy gap between the two best poses of the ligand of at least 1.2 kcal/mol is considered as the selection criterion.

Table 3.3: The interacting residues of the binding pockets discovered through ensemble docking. The interacting residues of the active site, which corresponds to the fourth pocket in the numbering of this work, are taken from Section 1.1 of Reference 14 by Ahmad *et al*. The chain identifiers and residue numbers are consistent with the numbering of the structure with PDB identifier 6M71 [100].

pocket nr.	interacting residues
1	**chain A:** LEU172, TYR265, THR319, PRO323, THR394, PHE396, LEU460
2	**chain A:** ASP36, TYR38, ILE66, SER68, ASP208
3	**chain A:** LEU49, ASP711, ASP714, GLN773
4	**chain A:** ASP618, CYS622, ASN691, ASN695, MET755, ILE757, LEU758, SER759, ASP760, ASP761, ALA762, VAL763, GLU811, PHE812, CYS813, SER814
5	**chain A:** ASN447, **chain B:** PRO133, ASP134, TRP182, PRO183, **chain B:** LYS27
6	**chain A:** ASN414, ASN416, ASP418, VAL844, **chain C:** ILE68, **chain D:** ARG111
7	**chain B:** TYR138, THR145, TRP154, GLU155, LEU169
8	**chain A:** GLU254, ASP269, LYS272, ARG285
9	**chain A:** GLN292, THR293, LEU302, ASP303, ARG305, LEU470

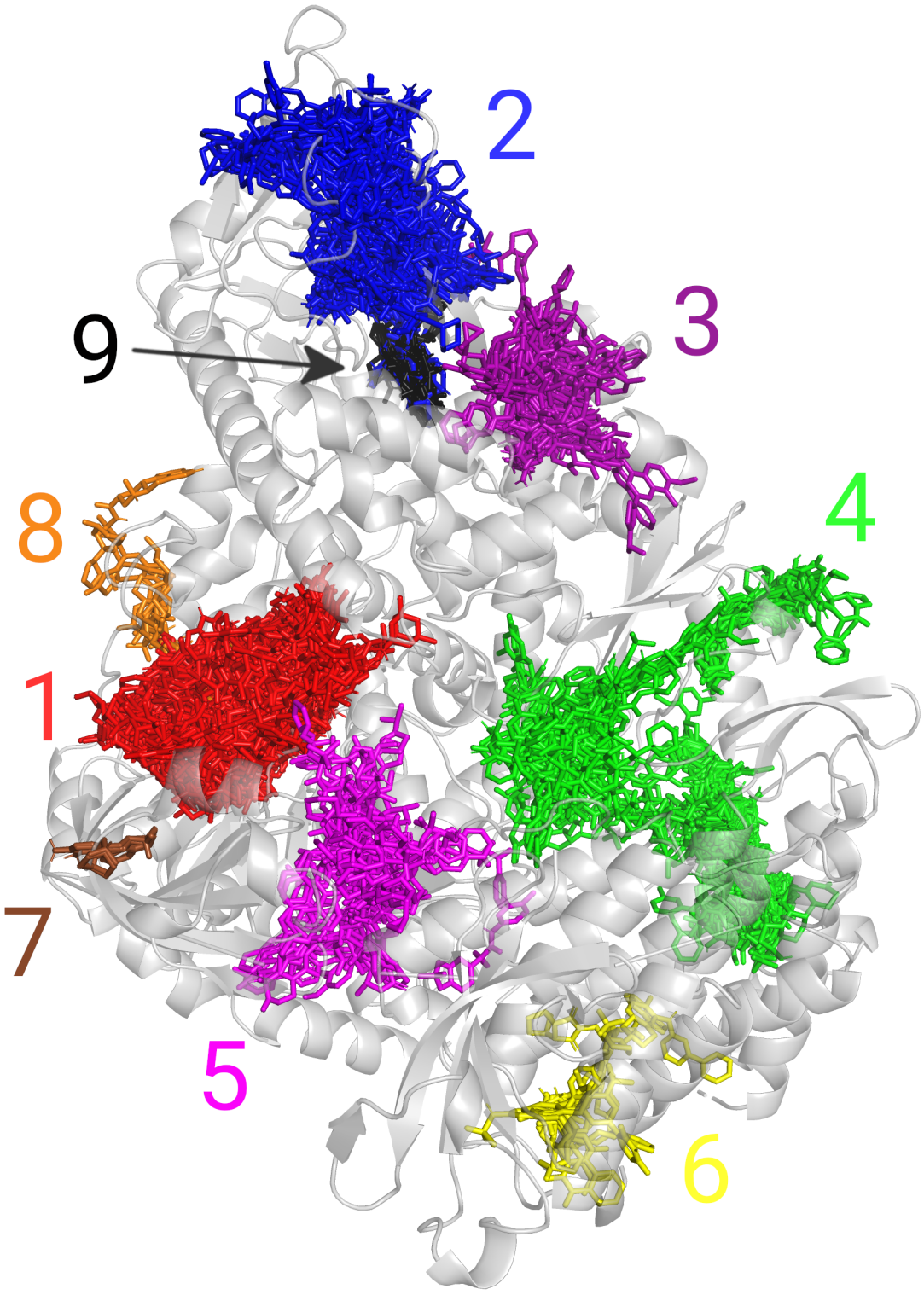

Figure 3.13: Binding sites on the protein, as identified by the visual inspection of the poses of the best binding ligands. The protein structure is the first cluster representative structure obtained from the water solvent MD trajectory. The coloured molecule clusters are the superimposed binders for all cluster representative protein structures as selected by either one of the binding energy criteria. Note that only those ligands are shown which bind to one of the identified pockets, an additional 2 % of the binders, which bind to other regions of the protein are hidden.

3.6.3 Analysis of the pocket ligand populations

The number of binders in each pocket, selected by the binding energy and the energy gap criteria are shown in Tables 3.4 and 3.5 respectively. Overall, the nine identified pockets are responsible for 97-99 % of the best binders for all trajectories, which clearly highlights the importance of these regions of the protein surface for potential subsequent targeted VS campaigns. It is most interesting that some of the identified pockets only bind ligands if specific cosolvents or even a specific ligand selection method is employed. This observation again shows, that the cosolvents significantly influence the protein structure, so much so that it can lead to the opening or the closing of certain pockets. Regarding the populations of the individual pockets, on the one hand, it is reassuring to see that multiple binders are found, with all three solvents considered, for the active site of the protein, which corresponds to the fourth pocket in the numbering of the present work. On the other hand, no binders are found for this site if the crystallised, apo protein structure is employed in the docking calculations. This fact highlights the importance of protein flexibility during ligand uptake, and makes it abundantly clear why docking to a single, crystallised, apo protein structure can be woefully inadequate to utilise the full power of docking calculations and VS. Returning to the binders discovered by the conformations originating from MD trajectories, the most highly populated pocket is in fact not the active site, instead it is either the first or the second pocket, depending on the cosolvent and binder selection criterion utilised. Among these, the first identified pocket is especially promising, as ligands bound here tend to be very well buried in the protein (see Figure 3.13). The fact that the first and second pockets are populated by a relatively large number of ligands, regardless of the cosolvent utilised to obtain the protein conformations, signals that they are quite stable and are most likely open in the conformations most often visited by the protein. These pockets therefore definitely cannot be characterised as transient or cryptic. On the other hand, the sixth and seventh pockets for example, are only populated with a few ligands when protein structures obtained from the benzene CMD trajectories are considered (except for the single ligand found by the energy gap criterion for the water solvent trajectory). It seems likely that the benzene probes were able to induce some sort of conformational change in these regions of the protein. These latter two pockets are therefore more likely to be cryptic pockets, open only in rarely visited protein conformations or in the presence of binders. These rarely appearing pockets that are seemingly only open during the benzene CMD trajectory, could partly explain the larger RMSD distance between the protein conformations obtained from this trajectory and all other conformations, found in Section 3.3.2. For the third pocket, similar results are found as for the sixth and seventh, with the exception that some ligands are also selected for the conformations obtained from the phenol CMD trajectory. The fact that on top of the benzene cosolvent molecules, the more polar phenol probes were also able to

open this pocket to a certain degree, could be connected to the higher ratio of hydrophilic residues around this binding site, as discussed in Section 3.7.1. Finally, the fifth, eighth and ninth pockets all seem to be more open in the conformations coming from the water solvent trajectory, with the latter two only having ligands if the energy gap ligand selection criterion is considered. As it will be seen in Section 3.7.1, the residues around these pockets also tend to be more hydrophilic than in the case of some other sites. Remarkably, these pockets are also not present in the crystallised apo structure, as can be seen by the lack of binders in the last column of Tables 3.4–3.5. It is therefore very likely that some reorganisation of the neighbouring residues, accounted for during the conventional MD calculations with water as the solvent, are necessary for their opening.

Table 3.4: The distribution of the binders, as selected by the binding energy criterion, across the various binding sites of the protein. Ligand counts are reported separately for each cosolvent utilised to obtain the protein conformation to which the ligand binds and the apo crystal structure. For the definition of the pocket numbering see Figure 3.13 and Table 3.3.

pocket	type of protein structure			
	water	benzene	phenol	crystal structure
1	56	8	99	2
2	9	67	22	1
3	0	11	4	3
4	4	29	9	0
5	29	2	5	0
6	0	6	0	0
7	0	4	0	0
other	1	3	1	0
total	99	130	140	6

3.6.4 Similarity of the binders of a given pocket

The strength of the binding interaction during docking calculations is determined by a scoring function which considers how "compatible" the structure of the ligand is with the conformation of the binding site residues. Therefore it is an interesting question, whether the effect of the different protein conformations obtained from different CMD trajectories can be detected in the (dis-)similarity of the ligands binding to these conformations. Observing the similarity of the ligands docked to a given binding site but considering different cluster representative protein structures, one could learn about how preserved the given pocket is, across the various conformations the protein tends to adopt. A further point of concern is whether the binders of an identified pocket exhibit some structural similarity between each other, that could be

Table 3.5: The distribution of the binders, as selected by the energy gap criterion, across the various binding sites of the protein. Ligand counts are reported separately for each cosolvent utilised to obtain the protein conformation to which the ligand binds and the apo crystal structure. For the definition of the pocket numbering see Figure 3.13 and Table 3.3.

pocket	type of protein structure			
	water	benzene	phenol	crystal structure
1	31	3	9	1
2	25	13	11	0
3	0	3	2	1
4	3	10	2	0
5	19	0	0	0
6	1	7	0	1
7	0	5	0	0
8	4	2	0	0
9	8	0	0	0
other	5	0	0	0
total	96	43	24	3

exploited by future drug design projects. To investigate the questions posed above, similarity coefficients are calculated between the best binding ligands of all protein conformations. The technical details of the calculation of these coefficients are given in Section 2.6, the only fact reiterated here is that the value of a similarity score can range between zero and one, with a score closer to one indicating a higher degree of similarity.

To tackle the question of intrapocket ligand similarity, average similarity scores are calculated between all ligands binding to a specific pocket (irrespective of the cosolvent utilised to obtain the protein structures). As a comparison, average interpocket scores are also calculated: the average similarity score of the ligands of a given pocket with the ligands of all other pockets are taken. On Figure 3.14, these average similarity scores can be compared with each other for all nine discovered pockets. The plot on the left pane shows these scores calculated for the ligands selected by the binding energy criterion, while the right plot shows the same for the ligands of the energy gap criterion. Considering the ligands of the binding energy criterion on the left, one can observe that the intrapocket similarity score is in most cases slightly higher than the interpocket one for all pockets. This difference is most noticeable for the fifth and especially the third and seventh pockets. The fact that these binding sites were all characterised as transient in Section 3.6, appearing in only a subset of the CMD trajectories, can help explain this observation. Since these pockets only open in very specific circumstances, most likely with the help of certain types of cosolvent probes, there are only a handful of protein conformations considered in which they are accessible. With fewer protein structures considered, the conformation of the residues around these sites vary less than in the case of other, more

easily accessible pockets which are open in more protein structures. This clearly defined protein conformation in turn results in the fact that only ligands with a well defined structure can bind to these pockets. Pocket six does not fit into this explanation, because even though it was characterised as a cryptic pocket in Section 3.6, its intrapocket similarity is practically equal to its interpocket one. This could perhaps suggest multiple opening mechanisms for this pocket, where the resulting different pocket conformations bind slightly different ligands. For the other, more stable pockets, the two average similarity scores are quite close to each other. This seems reasonable, considering the higher number of protein structures within which these pockets are open, resulting in a larger variety of local residue conformations, leading to more diverse binders.

Looking at the right pane of Figure 3.14, where the same average similarity scores are plotted for the ligands selected by the energy gap criterion, one finds a slightly different picture. For most pockets, the intrapocket similarity is quite similar to and in one case it is smaller than the interpocket score. This trend is only broken in the case of pocket seven, where the intrapocket score is much higher, as can be expected based on the reasoning presented above. In general however, the similarity scores plotted on the right pane of Figure 3.14 are significantly lower than those plotted on the left graph. To understand these lower similarity scores especially those calculated between ligands binding to the same pocket, it is useful to remember the findings of Section 3.5 about the energy gap ligand selection criterion. As demonstrated by Figure 3.11 of that section, this selection method is much more sensitive to the conformational differences of the binding site residues, especially to those that are present between the protein structures obtained from different CMD trajectories. This can result in very different sets of ligands selected for the same pocket if conformations originating from different cosolvent trajectories are considered. Moreover, the intrapocket similarity scores plotted on Figure 3.14 are calculated between all ligands binding to the same pocket, no matter which cosolvent the protein conformation utilised during docking is obtained from. These two facts can explain the lower similarity scores seen on the right pane of the figure: the sensitive binder criterion selects dissimilar ligands for conformations obtained from different solvent trajectories which reduce the average intrapocket similarity. In the case of pocket seven, the intrapocket similarity score is much higher as this pocket only harbors ligands when the conformations of the benzene cosolvent trajectory is considered, from which only three cluster representative protein structures were obtained. For this low number of protein conformations, all coming from the same cosolvent trajectory, even the energy gap criterion selects largely similar binders for this pocket.

Finally, the effect of cosolvents on the similarity of ligands binding to a given pocket is also examined. On Figure 3.15 the average intrapocket similarity scores can be seen for all pockets. These averages were calculated by considering only those ligands which bind to

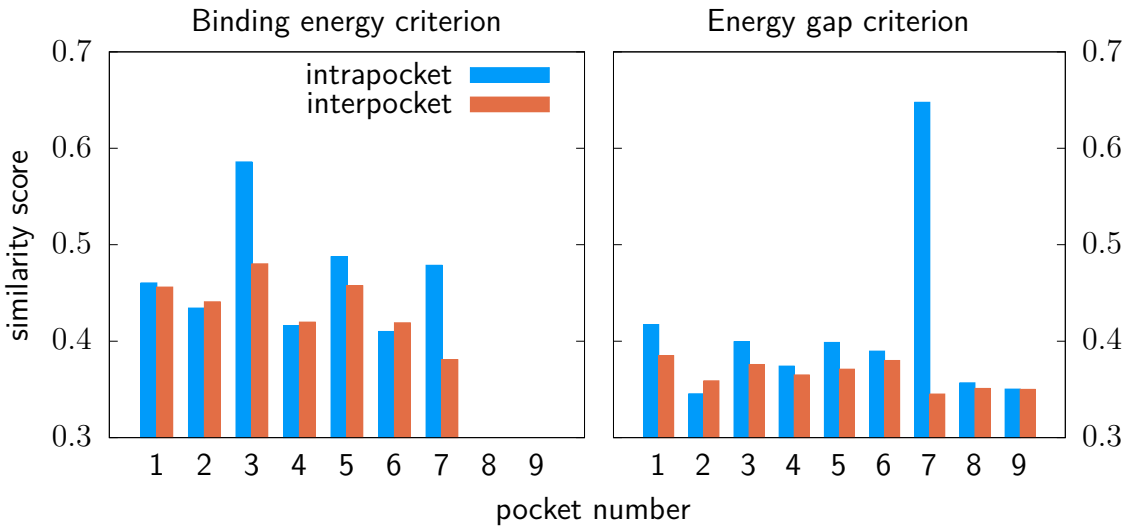

Figure 3.14: Average similarity scores between the ligands binding to a given pocket and other ligands binding to that pocket (intrapocket) or all other ligands binding to different pockets (interpocket). On the left pane the binders were selected with the binding energy criterion, while on the right pane the energy gap selection method was utilised. Ligands binding to any of the cluster representative protein conformations were considered together. No binders are found with the binding energy criterion to pockets eight and nine, therefore the corresponding columns are empty on the plot on the left.

protein conformations obtained from a given solvent trajectory. On the left pane of this figure, the ligands selected by the binding energy criterion can be seen, while on the right pane the binders according to the energy gap criterion are shown. One can indeed observe slightly higher intrapocket similarities on these plots, than on Figure 3.14 where the ligands of every trajectory are considered together. The strength of this effect varies heavily with the pocket being considered. It is most pronounced in the case of pocket three, where the ligands found for the conformations obtained from the benzene cosolvent trajectory are especially similar to each other regardless of the employed binder selection criterion. As for the similarity of the ligands selected by the energy gap criterion, a slight increase can be observed when only the ligands of a single MD trajectory are considered compared to the case when the ligands of all trajectories are included (see Figure 3.14 for the latter).

Considering everything mentioned above, one can conclude that the binding energy ligand selection criterion indeed tends to select at least somewhat similar ligands for a given pocket, even if ligands selected for different protein conformations are considered. On the other hand, the energy gap criterion is more sensitive to the subtle conformational changes of the protein residues around the binding sites, and therefore selects more dissimilar ligands for the different protein conformations. Moreover, the type of the pocket can also significantly influence the similarity of the ligands binding to it. On the one hand, stable binding sites that can adopt a wide range of conformations while still being accessible to ligands can have a larger variety of

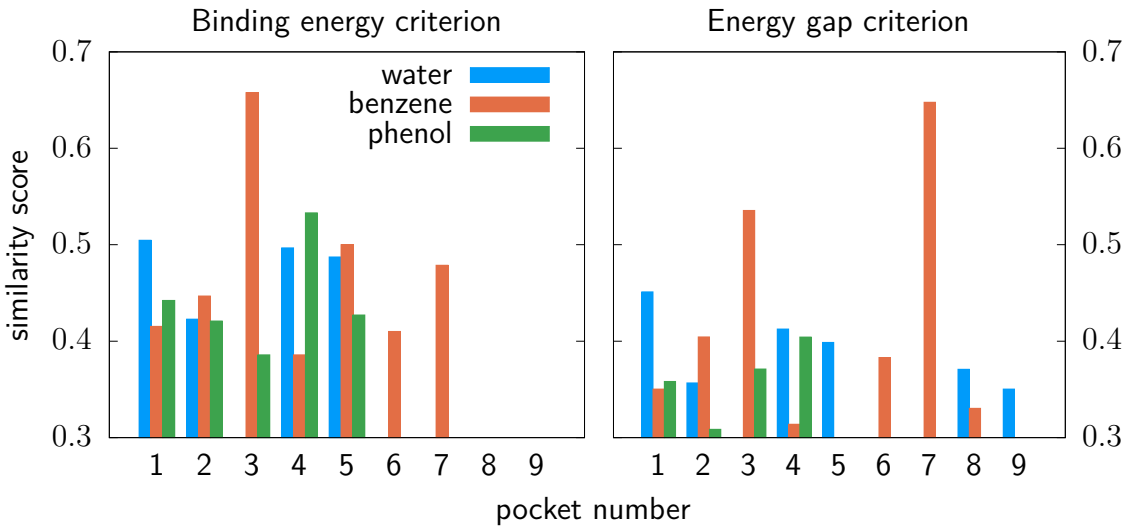

Figure 3.15: Average intrapocket ligand similarity scores separated based on the solvent utilised to obtain the protein conformation the ligand binds to. On the left pane, the ligands selected by the binding energy selection criterion are considered, while on the right pane the ligands chosen by the energy gap threshold are shown. Ligand similarity scores are averaged over all ligand pairs that bind to the same pocket, in protein conformations obtained from the given solvent MD trajectory.

binders. On the other hand, transient, cryptic binding sites that are only open when the protein adopts some specific conformation, are much more selective when the binding of ligands is considered, leading to a small, highly similar set of binders.

3.6.5 Conclusions

During the course of this section, the most important binding regions of the target protein have been identified. It has been shown that some of these binding sites are only accessible to ligands if protein conformations from a specific MD trajectory are considered. From data regarding the number of trajectories in which the pocket is open and the number of ligands binding to the said pocket, the binding sites were characterised as either stable or transient. In particular, pockets three, five, six, seven, eight and nine were identified as likely transient, cryptic pockets while pockets one, two, and four seem to be more stable binding sites. Finally, the similarities of the ligands binding to the same pocket were investigated. It was revealed that stable pockets tend to bind a larger variety of ligands, while cryptic pockets bind only a small set of more similar ligands.

3.7 Description of the protein structure around the binding sites

As a final analysis, the protein structure around the binding sites is investigated. First, the hydrophobicity of the residues around the different pockets is compared. Next, more detailed structural analysis is carried out for two selected pockets. The first of these is the active site of the protein, because it represents a known, stable pocket. The other described pocket is the seventh one, as the results discussed in Section 3.6 clearly point to it being a typical cryptic pocket. If the discovered binding sites of the protein are to be further investigated, similar analysis should be carried out for the other pockets as well. However, the analysis for the remaining pockets is not performed here, as this would have exceeded the scope of the present Master's thesis.

3.7.1 Hydrophobicity of the binding sites

The primary goal of this section is to assess whether the polarity of the cosolvents correlate with the hydrophilicity of the binding pockets that are open during the CMD trajectory calculated with that cosolvent. To perform this analysis, the residues nearest to the nine binding sites identified in the previous section have to be selected. This is done by first selecting the protein conformation for which the highest number of binders are found for the pocket in question. Then, those residues are selected which have atoms not farther than 3 Å from one of the atoms of one of the binders. This distance threshold yields a sufficient number of residues (at least four) for all pockets. Then, the selected protein residues are categorised as either hydrophilic or hydrophobic based on the results presented in Reference 108. The ratio of hydrophobic residues for each pocket, along with the cosolvent with which the most populated protein conformation is obtained for that pocket, is shown in Table 3.6. Looking at this table, it is apparent that the pockets which were most open in the protein conformations coming from the water solvent trajectory all have a low hydrophobic residue ratio. In particular it is seen here, as well as in Section 3.6.3, that pockets eight and nine bind ligands almost exclusively when protein conformations coming from the water trajectory are considered. This phenomenon is now partly explained by the high ratio of hydrophilic residues around these pockets. On the contrary, the most apolar cosolvent, benzene, opens the most hydrophobic pockets best. More specifically, pockets six and seven, which were only discovered with the benzene cosolvent protein conformations, both have a hydrophobic residue ratio of at least 50 %. Pockets two and three are also most populated with the benzene cosolvent protein conformations, even though both of them are decidedly hydrophilic in character. However, it is worth noting that

Table 3.6: The ratio of hydrophobic residues near each pocket, along with the cosolvent with which the protein conformation which produced the most binders for that pocket was obtained. The assignment of residues to each pocket is discussed in the main text.

pocket	total no. of residues	hydrophobic residues [%]	cosolvent
1	34	47	phenol
2	35	34	benzene
3	14	21	benzene
4	19	58	benzene
5	26	31	water
6	9	55	benzene
7	6	50	benzene
8	4	0	water
9	9	33	water

these pockets are also well populated when protein structures from trajectories with water and especially with phenol as a cosolvent are considered. Based on the above results, a link between the polarity of the solvent, and the hydrophilicity of the pockets opened by said cosolvent can be confidently established. This highlights once again the usefulness of employing CMD simulations for the generation of protein conformations in ensemble docking. With a highly polar solvent such as water, the more hydrophobic pockets of the protein, such as the sixth or seventh, are less open when compared to conformations obtained from apolar cosolvent trajectories such as the benzene or the phenol one. Therefore, by utilising the cosolvent probes' ability to open hydrophobic transient pockets, one can find binding ligands to targets that would have otherwise been considered undruggable.

3.7.2 Conformational changes around the binding sites

The two pockets (number four and seven) that are selected for further analysis, are first characterised using the pocket descriptors of the `mdpocket` program. Then, the main differences between the various conformations obtained for these two pockets are visualised. Based on these results, the conformational stability of the residues around the pocket and the mechanism of pocket opening for these pockets are discussed.

A typical transient pocket

This section is dedicated to the analysis of pocket seven, which was identified as a cryptic pocket in the previous sections. On Figure 3.16, the pocket score (see Equation 3.2 for its definition) and pocket volume, as calculated by `mdpocket` are shown for each protein structure.

Looking at the top plot of this figure, where the pocket scores are shown, one can observe that most of the protein structures coming from the water solvent trajectory do not provide favorable pocket conformations. On the contrary, conformations coming from the benzene trajectory all harbor pockets with appreciable score, with two out of the three conformations yielding the top two scores from any conformation. The phenol trajectory is closer to the water one in terms of pocket druggability scores calculated for pocket seven. When looking at the volumes of the pocket plotted on the bottom pane, the trend observed for the pocket scores is even clearer. The benzene trajectory is able to provide the pockets with the largest volume, in one case reaching more than 400 $Å^3$. These observations are to some extent in line with the fact that ligands were only able to bind to the conformations coming from the benzene trajectory, during the computational docking calculations. However, `mdpocket` calculates non-zero volumes and pocket scores for some protein conformations coming from the water solvent trajectory, indicating the presence of an open pocket in those conformations. The fact that the docking calculations found no ligands for the water solvent trajectory, not even for these conformations, indicates that the descriptors provided by `mdpocket` are not perfectly reliable and cannot be considered a substitute to the more accurate explicit docking calculations.

To better understand the varying degree of openness of pocket seven between the different conformations, the structure of the protein around this pocket is visualised. On Figure 3.17, two structures coming from either the water or the benzene trajectories are compared. The residue identifiers in this figure refer to the residues of chain B in the protein structure with PDB identifier 6M71. The conformation shown in magenta is the thirteenth conformation of the water trajectory, harboring a practically closed pocket (see Figure 3.16). In green, the second representative conformation of the benzene trajectory is shown (structure 15 in Figure 3.16), which offers the pocket with the highest score and a significant volume. The largest differences between the two structures are highlighted with grey arrows on Figure 3.17. The displacement of residue THR145 to the left, and residues TRP154 and GLU155 downwards, lead to a significantly larger cavity in the conformation obtained from the benzene CMD trajectory, into which ligands can bind. The fact that this conformational change occurs much more frequently with benzene (or to a lesser extent phenol) as the cosolvent, suggests that the (partially) apolar probes facilitate the opening of this pocket quite efficiently. The high ratio of hydrophobic residues around this pocket shown in Table 3.6 also supports this theory.

A well known, stable pocket

Finally, the structure of the active site is also analysed, to compare the characteristics of a cryptic binding site to a well established, stable one. First, to investigate the nature of conformational changes that can be expected to happen during ligand uptake by the active site,

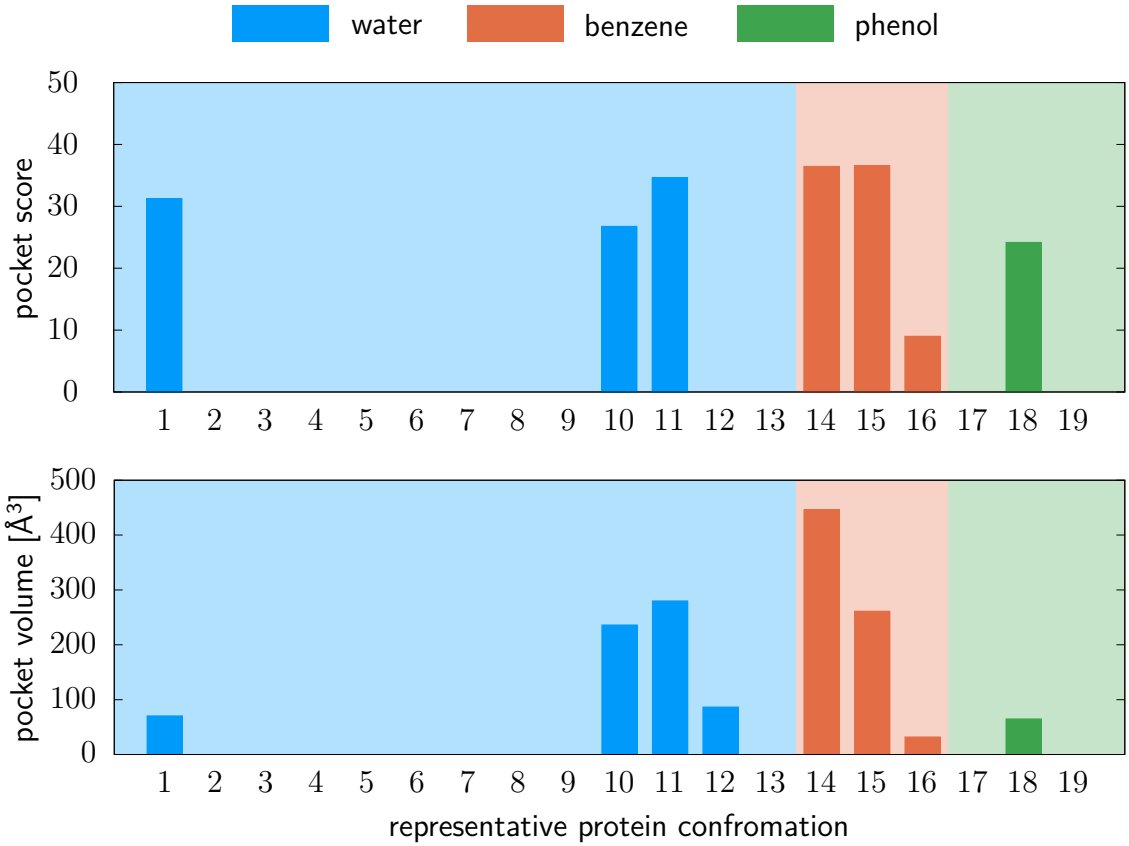

Figure 3.16: The pocket score (top) and volume (bottom) of pocket seven across the different protein structures. Both the pocket score and volume of the pocket is calculated from the pocket descriptors provided by `mdpocket` (see Section 3.1.6). The cosolvent trajectory, from which the given protein structure is taken, is indicated by the color of the columns and the background.

the experimentally obtained apo and holo crystallised structures of the protein are superimposed and visualised. On Figure 3.18, the holo crystallised structure is shown in green, while the apo structure is represented in magenta. The residue numbering in this figure refers to chain A of the protein structure with PDB identifier 6M71. As it can be seen, the conformation of the protein around the active site is largely unaffected by ligand binding: slight differences can only be observed in the orientations of some amino acid side chains. Based on this observation, it might be expected that no fundamental changes are occurring in the protein structure around the active site throughout the various MD simulations.

To investigate the presence of such conformational changes, the volume of the pocket corresponding to the active site is calculated for all representative protein conformations coming from MD trajectories, with the `mdpocket` program. These results are shown on the top pane of Figure 3.19. In comparison to the results found for pocket seven, one can observe a slightly more homogeneous pocket volume distribution across the various protein structures. However, the fluctuation between protein conformations is still large with the volume of the

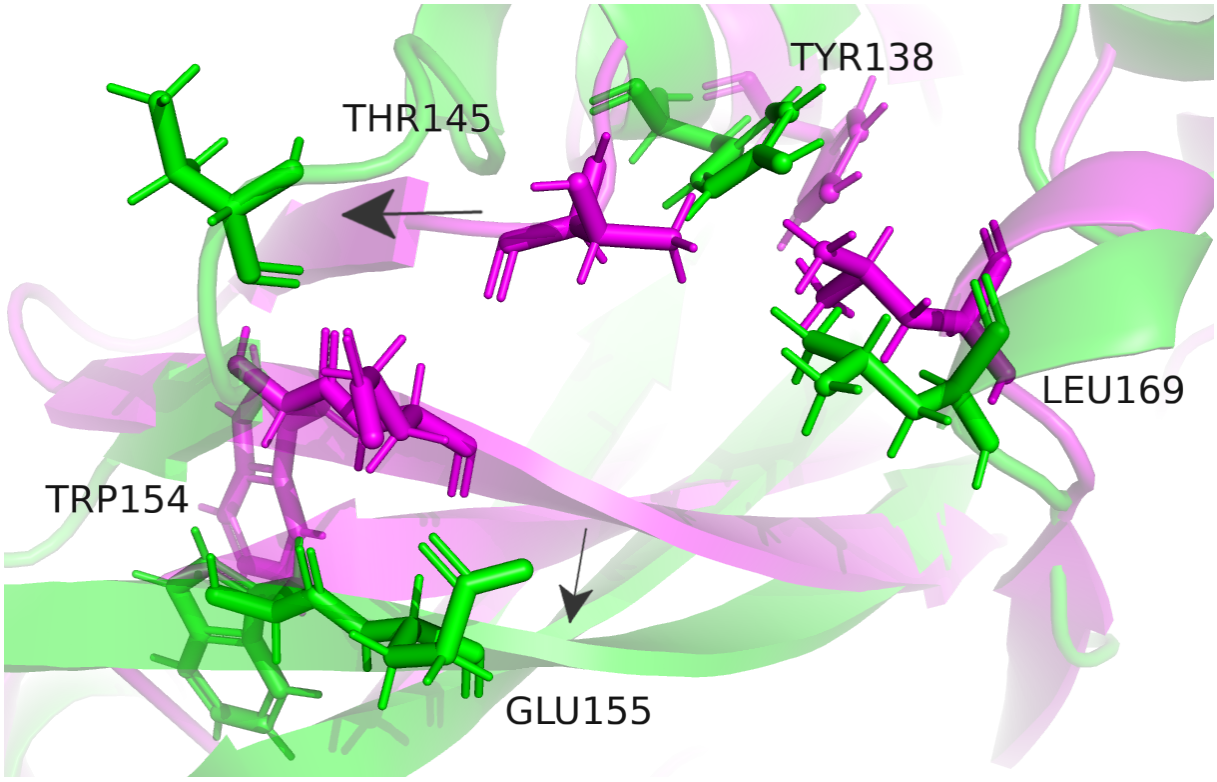

Figure 3.17: Conformational changes of the protein around pocket seven. In magenta, the thirteenth protein conformation coming from the water solvent MD trajectory is shown, while in green the second representative of the benzene trajectory can be seen. The residue numbering corresponds to chain B of the protein structure with PDB identifier 6M71. The most notable differences between the two conformations are highlighted with grey arrows.

active site pocket ranging between 800 and 1700 Å^3. Comparing the two binding sites further, this pocket is found to have a much larger volume than the transient pocket seven in all frames. The main similarity between these pockets is that their largest respective volumes are calculated when conformations coming from one of the cosolvent trajectories are considered. That being said, the difference in pocket volume between the water solvent and benzene cosolvent trajectory frames is far less significant in the case of the active site. As it will be seen from the next section, this more even pocket volume does not mean that the conformation around the pocket is completely rigid and that the druggability of the binding site also remains constant throughout the MD trajectories.

To further investigate the conformational changes happening around the active site, the pocket score as calculated by `mdpocket` is also examined. In this descriptor, there are still noticeable differences when conformations obtained from different cosolvent trajectories are considered. Upon inspecting the different types of contributions included in the pocket score individually (see Eq. 3.2), the hydrophobicity score is found to have the largest variations between the different cosolvents. Thus the average and standard deviation of the hydrophobicity score is

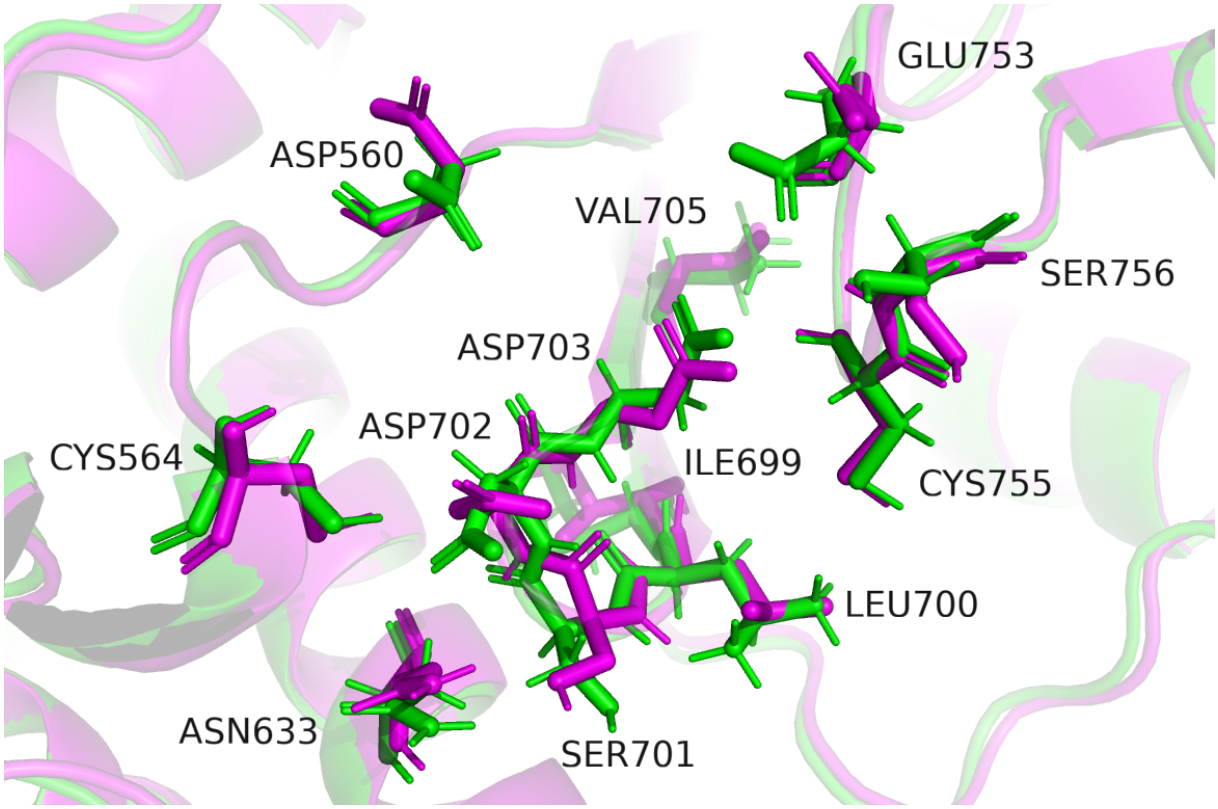

Figure 3.18: Conformational differences between the active sites of the apo and holo crystallised protein structures. In magenta the crystallised apo structure is shown (PDB: 6M71), while the holo structure is represented in green (PDB: 7B3B). The residue numbering corresponds to chain A of the 6M71 protein structure.

shown for each cosolvent on the bottom pane of Figure 3.19. The average hydrophobicity score can be observed to be about twice as much for the protein conformations obtained from the two cosolvent trajectories, than for those obtained from the water solvent one. This higher score found for the cosolvent trajectories corresponds well to the higher number of binding ligands found for the benzene or phenol cosolvent conformations, during the computational docking calculations (see Section 3.6.3). The change in the hydrophobicity score also indicates that there are still some non-negligible conformational changes around the active site occurring during the MD trajectories, even for this more stable region of the protein. In particular, the apolar cosolvent probes most likely manage to stabilise some conformations where hydrophobic residues are more accessible, leading to a higher average hydrophobicity scores for the pocket. This increased hydrophobicity once again validates the use of CMD trajectories for the generation of protein structure in a VS campaign.

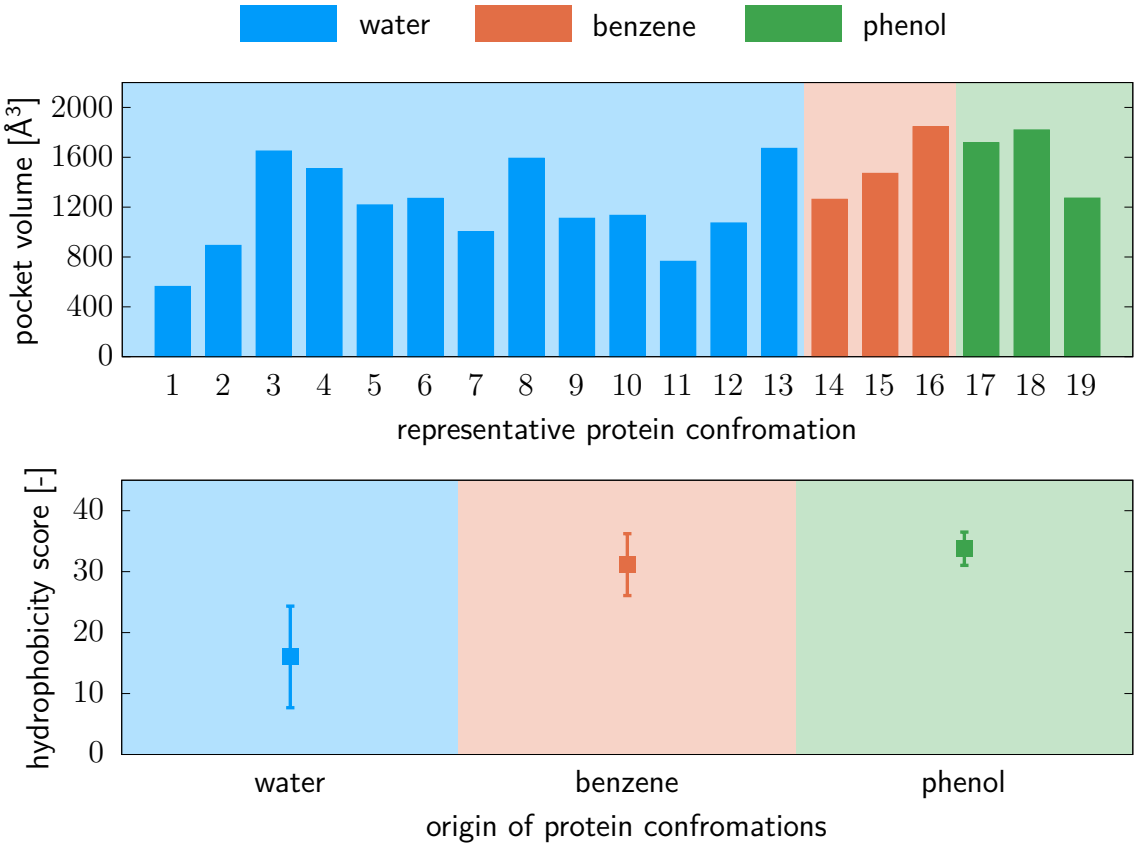

Figure 3.19: The volume (top) and average hydrophobicity score (bottom) of pocket four across the different protein conformations. The error bars on the bottom plot represent the standard deviations of the hydrophobicity score calculated for all cluster representatives of the given trajectory. Both the pocket score and volume of the pocket is calculated from the pocket descriptors provided by `mdpocket` (see Section 3.1.6). The cosolvent trajectory, from which the given protein structure is taken, is indicated by the color of the columns or markers and the background.

3.7.3 Conclusions

First, it has been shown that the polarity of the solvent significantly influences which pockets are likely to adopt open conformations during a MD trajectory. This observation encourages the utilisation of CMD to generate protein structures for docking calculations and VS campaigns, as decidedly hydrophobic pockets are more likely to be accessible in the structures generated in this way, than in crystallised structures or structures obtained from conventional MD trajectories. Next, the conformational changes around two binding sites were examined more closely: pocket four corresponding to the active site, and pocket seven shown to be a transient pocket. These pockets represent opposite ends of the pocket stability spectrum: while the active site is a known stable pocket, the residues around pocket seven are mobile and this binding site is rarely open. For the active site, the use of cosolvent simulations have been shown to increase its hydrophobicity score, indicating that docking calculations are expected to be more successful

when conformations from these trajectories are considered. This result highlights the ability of the cosolvent probes to induce meaningful conformational changes in the protein, not only in the region of transient pockets but for well known binding sites as well. Moreover, in the case of pocket seven, the cosolvent probes were shown to be absolutely necessary to induce the conformational changes of the protein, required for the opening of this cryptic pocket. In both cases, considering protein dynamics, and CMD trajectories in particular, during the preparation of the protein conformational ensemble increased the druggability of the pocket in question. It is therefore our expectation that these techniques could be advantageously applied during drug discovery projects.

Chapter 4

Conclusions

In this final chapter, the relevance of the presented work is assessed, its most important results are summarised, conclusions are drawn, and an outlook on the future of the project is given. First, the topic of the thesis is placed into wider context, emphasising its connections to recent developments in the fields of drug discovery and computational docking. Then, its most important results are collected and reviewed. Based on these results, the applicability of the presented methods in drug discovery projects are evaluated. Finally, directions for potential future efforts in connection with the presented work are discussed.

4.1 Project relevance

Computational approaches are widely utilised in the field of drug design and discovery. Their lower costs, increasing accuracy, and ability to provide quicker results make them a more and more attractive alternative to traditional experimental methods. In drug discovery projects, the technique of VS has proved to be especially useful, contributing to the discovery of hundreds of small molecule drug candidates already. The key to its success is that it is capable of efficiently evaluating thousands of ligands with only a 3D structure of the target protein as input, by performing parallel docking calculations to the protein structure with each considered ligand. Owing to its numerous successes, considerable efforts are expended to achieve further improvements in its accuracy and usability. Recently, the fact that it ignores the dynamics of the target protein by considering only a single rigid structure, has been identified as one of the main weaknesses of the traditional VS method. With newer models of ligand binding (such as that of conformational selection) attributing a central role to the motions of the protein during this process, it has become clear that assuming a rigid protein structure during docking calculations can severely undermine the accuracy of the results.

To remedy this problem, several approaches that are attempting to account for the dynamics of the protein structure are being evaluated. A promising family of such approaches, the ensemble docking methods, are utilising a handful of rigid protein structures during the docking calculations, to describe the potential conformational changes of the protein. The main challenge for these techniques is the construction of a suitable protein structure ensemble. Since the experimental determination of even a single 3D protein structure requires time, trained researchers and expensive equipment, computational methods are being considered for the generation of such structure ensembles. In the present work, the ability of CMD simulations is evaluated to perform this task. This approach is promising, as it has already been successfully applied to discover new transient binding sites in some targets and to sample druggable conformations of a range of proteins.

A further point of relevance for the project is connected to the COVID-19 pandemic. With little or no effective treatments available during the first months of the epidemic, the SARS-CoV-2 have already claimed millions of lives around the globe. This dire situation quickly prompted researchers of various backgrounds to embark on a search for drugs that can be utilised in the fight against the virus. The urgency of the pandemic means that shortening the time required to develop new treatments has never been more important. Due to their speed and efficiency, computational techniques have become the method of choice for many drug discovery projects in this race for a cure. These efforts are supported by the extraordinarily large number of high quality 3D structures of the most important SARS-CoV-2 proteins. The presented work joins the line of research projects that utilise computational methods to investigate the RdRp protein, key for the reproduction of this virus, in the hopes of revealing information that can contribute to the development of effective treatments.

4.2 Summary of the most important results

The first goal of the project was to prepare and execute the CMD simulations. Out of these initial steps, the preparation of the solvent boxes containing the appropriate mixture of water and cosolvent molecules proved to be somewhat challenging. In particular, the simulations with benzene as the cosolvent were problematic, as at first they exhibited severe clustering of the cosolvent molecules. This phenomenon is not desirable as it significantly reduces the interactions between the cosolvent probes and the protein. To overcome this problem, the introduction of an artificial LJ potential between the carbon atoms of different benzene molecules was most helpful. After the calculation of the production MD trajectories were finished, some preliminary analysis was performed on them. By plotting the evolution of the RMSD distance of the protein from its starting structure throughout the trajectories, it was

confirmed that the CMD simulations equilibrate just as fast or even slightly faster than the traditional MD calculation. This observation gave us some confidence that the cosolvent probes do not induce unrealistic changes in the protein and the frames of these trajectories sample relevant conformations.

With the various MD trajectories in hand, the next objective was to select a handful of snapshots from them, which can serve as the protein conformational ensemble for the subsequent ensemble docking calculations. To this end, clustering of the trajectories with the density based clustering algorithm was utilised. After careful tuning of the clustering parameters, this approach gave satisfying results, with 19 representative protein conformations selected in total. The cluster representative frames selected from different cosolvent trajectories proved to harbor protein conformations with meaningful differences, as evidenced by their large RMSD distances from each other. Considering only the alpha carbons of the protein during the trajectory clustering turned out to significantly reduce the computational costs of the clustering without sacrificing its accuracy.

As the next step of the present work, the ensemble docking calculations were performed utilising the 19 representative protein conformations and a set of approximately two thousand FDA approved drug molecules. The parallel execution of the independent calculations for all ligands were managed by simple in-house scripts. Based on the results obtained, first an alternative binder selection method was devised and evaluated. This method selects the binders according to the binding energy gap between their two best docked poses. It is intended to help in selecting the ligands with the best interacting chemical moieties, regardless of the size of the ligand. It was found to select a reasonable ratio of common ligands with the simple binding energy criterion, with the ligands selected by both methods showcasing the most favorable binding energies of all the considered ligands. Subsequently, utilising either of the two selection criteria, the best binders of the protein were selected. By inspecting the poses of these ligands, nine important binding sites of the protein were identified, that harbored the best pose for over 98 % of the best binders. By analysing the populations of the binding pockets across different CMD trajectories they were separated into two groups, corresponding to stable and transient binding sites. This categorisation of the pockets was reinforced through the calculation of the similarities of the ligands binding to them as well. It was found that stable pockets are open in many or all frames of the various MD trajectories, and they bind a relatively large number of ligands that can be quite dissimilar to each other. On the contrary, transient or cryptic binding sites were found to be open only in a limited set of protein conformations, often originating from the same CMD trajectory calculated with an apolar cosolvent. Moreover, they only bind a limited number of ligands that are highly similar to each other. These differences were explained as follows. On the one hand, stable pockets do not require special protein conformations to be open and consequently the surrounding residues can adopt a range of

local conformations while still allowing for ligands to dock. This results in a more diverse set of binders, binding to protein structures obtained from various MD trajectories. On the other hand, cryptic pockets require rarely occurring conformational changes to happen in the protein in order to open. This means that there are much fewer representative conformations in the ensemble that harbor an open cryptic pocket than those harboring open stable pockets. Furthermore, these conformations usually originate from a single CMD trajectory as typically only a specific cosolvent is able to induce the necessary conformational changes in the protein. Since the protein conformations coming from a single MD trajectory tend to be more similar to each other, the fact that only a single cosolvent is able to open the transient pockets also means that the binders will be similar to each other. To summarise, by considering a handful of protein structures for the ensemble docking calculations, obtained through various CMD simulations, not only the binding ligands and binding sites could be identified, but additional information about the nature of the binders and pockets could also be obtained.

Finally, in order to understand the effects of cosolvents better, the protein conformations around two selected binding sites were compared between representative structures originating from different MD trajectories. One of these selected pockets was the active site of the protein, which was characterised as a stable pocket during the previous analysis. The other pocket was distal to the active site, and was previously described as transient, due to number and similarity of the ligands it binds. By comparing the conformational changes the cosolvent probes induce around such pockets with different characteristics, one can hope to better understand how they interact with the protein and to identify situations where their usage is advantageous. Considering first the transient pocket, it was found that only the apolar cosolvent probes, especially benzene, were able to induce conformational changes which lead to the opening of the binding site. Superimposing an open and a closed pocket conformation revealed that a significant rearrangement of the primarily apolar residues around the pocket is necessary for its opening. It was pointed out that the apolar nature of this pocket makes hydrophobic cosolvents like benzene especially useful for sampling protein conformations in which it is open. Looking now at the active site, it was revealed that even though this region of the protein seems conserved between the apo and holo crystallised structures, the MD simulations sample some protein conformations containing important changes around the active site. These changes were most noticeable in the hydrophobicity descriptors of the pocket in question, which increased sharply when conformations obtained from CMD trajectories were considered. This was considered as evidence that even in well known, stable pockets, cosolvent probes can induce conformational changes that increase the druggability of such pockets, thus leading to protein conformations better suited for computational docking calculations. All in all, it was found that the effects of cosolvents are significant on either cryptic or well known pockets, with more druggable protein conformations visited during CMD simulations than either the

crystallised protein structure or those visited during traditional MD trajectories with water as the solvent.

4.3 Final remarks

During the course of this project, a thorough investigation of the effects of cosolvent simulations and protein dynamics on docking calculations have been carried out. This was achieved through performing a small scale VS campaign for the RdRp protein of SARS-CoV-2 and analysing its results from many different angles. Perhaps the most important conclusion of this work is that such VS projects are most effective when a wide range of analysis techniques are utilised in combination. The fact that no single metric can provide the ultimate best description of complex phenomena such as those encountered in a VS campaign is of course not surprising. However, the necessity of investigating such problems with various approaches and avoiding relying on a single description should be emphasised as often as possible. Ample evidence of this wisdom was seen during the course of this work, starting from the clustering of the MD trajectories, through the selection methods employed to obtain the best binding ligands, to the evaluation of the druggability of certain protein conformations.

Among the more concrete conclusions of the project are those regarding the applicability of MD and CMD simulations to enhance the quality of computational docking calculations. It was clearly demonstrated that even short, traditional MD trajectories, simulated in a simple water solvent, can be of use in the discovery of new binding sites and ligands of a protein. By utilising a handful of protein structures selected from such trajectory in an ensemble docking calculation, one can at least partially account for the dynamics of the protein, known to be crucial in many ligand uptake processes. With the help of the so obtained protein conformations, significantly lower binding energies were obtained for a number of ligands and numerous previously unreported binding sites of the RdRp protein were discovered. Moreover, it was shown that employing additional CMD simulations can further improve the completeness of the docking results. They are especially suitable to discover the more hydrophobic, transient binding sites of the target, as the employed apolar cosolvent probes can accelerate the conformational changes necessary for the opening of such pockets. To summarise, it was found that (cosolvent) MD simulations can be successfully used to enhance the accuracy of VS campaigns. They require relatively little computational resources as quite short simulations are sufficient for these purposes. With their help, new binding sites and binding ligands can be identified. The resulting ensemble of protein structure can be tremendously useful in understanding the nature of the discovered pockets or even their mechanism of opening. It is therefore expected that these techniques can be advantageously utilised in many future VS campaigns.

4.4 Outlook

When the discovery of new inhibitors of the SARS-CoV-2 RdRp protein is considered, further investigations of the newly discovered binding sites is of utmost importance. More specifically, one or two of the most promising binding pockets could be selected and their binders could be more accurately evaluated. For this purpose protein–ligand MD simulations could be carried out, which would not only provide a more accurate binding energy for the selected ligands but could also help in understanding the mechanism of ligand uptake for the pocket in question, especially if enhanced sampling techniques are also utilised.

In connection to the development of new ensemble docking methods, that can efficiently account for the dynamics of proteins, the evaluation of other enhanced sampling methods for the generation of the conformational ensemble would be most beneficial. Some of these methods, discussed in Section 1.2.3, could be evaluated for this purpose utilising the same protein and drug candidates. Afterwards the resulting binders and discovered binding pockets could be compared with those obtained in the present work to select the most efficient method for the protein structure generation. Alternatively, the protocol devised in this work could be applied to new proteins in order to assess the consistency of the presently obtained results.

Bibliography

[1] J. D. Durrant, M. D. Urbaniak, M. A. J. Ferguson, and J. A. McCammon, *Journal of Medicinal Chemistry* **53**, 5025 (2010).

[2] J. D. Durrant, R. Cao, A. A. Gorfe, W. Zhu, J. Li, A. Sankovsky, E. Oldfield, and J. A. McCammon, *Chemical Biology & Drug Design* **78**, 323 (2011).

[3] X. Li, X. Zhang, Y. Lin, X. Xu, L. Li, and J. Yang, *Chemistry & Biodiversity* **16**, e1900170 (2019).

[4] C. Li, L. Xu, D. W. Wolan, I. A. Wilson, and A. J. Olson, *Journal of Medicinal Chemistry* **47**, 6681 (2004).

[5] S. Cosconati, J. A. Hong, E. Novellino, K. S. Carroll, D. S. Goodsell, and A. J. Olson, *Journal of Medicinal Chemistry* **51**, 6627 (2008).

[6] E. Mullarky, N. C. Lucki, R. Beheshti Zavareh R, J. L. Anglin, A. P. Gomes, B. N. Nicolay, J. C. Y. Wong, S. Christen, H. Takahashi, P. K. Singh, J. Blenis, J. D. Warren, S.-M. Fendt, J. M. Asara, G. M. DeNicola, C. A. Lyssiotis, L. L. Lairson, and L. C. Cantley, *Proceedings of the National Academy of Sciences of the United States of America* **113**, 1778 (2016).

[7] S. Cosconati, S. Forli, A. L. Perryman, R. Harris, D. S. Goodsell, and A. J. Olson, *Expert Opinion on Drug Discovery* **5**, 597 (2010).

[8] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, *Nature Reviews. Drug Discovery* **5**, 993 (2006).

[9] X. Morelli, R. Bourgeas, and P. Roche, *Current Opinion in Chemical Biology* **15**, 475 (2011).

[10] E. O.-O. Max Roser, Hannah Ritchie and J. Hasell, *Our World in Data* (2020), `https://ourworldindata.org/coronavirus` [Online; accessed on 2 May 2021].

[11] H. Rathi, V. Burman, S. K. Datta, S. V. Rana, A. A. Mirza, S. Saha, and R. Kumar, *Indian Journal of Clinical Biochemistry* **36**, 3 (2021).

[12] Reuters Staff, `https://www.reuters.com/article/healthcoronavirus-gilead-remdesivir-idUSL1N2HD1UX`, [Online; accessed on 14 May 2021].

[13] J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R. W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T. F. Patterson, R. Paredes, D. A. Sweeney, W. R. Short, G. Touloumi, D. C. Lye, N. Ohmagari, M.-d. Oh, G. M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M. G. Kortepeter, R. L. Atmar, C. B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J. D. Neaton, T. H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H. C. Lane, *The New England Journal of Medicine* **383**, 1813 (2020).

[14] J. Ahmad, S. Ikram, F. Ahmad, I. U. Rehman, and M. Mushtaq, *Heliyon* **6**, e04502 (2020).

[15] S. Ao, D. Han, L. Sun, Y. Wu, S. Liu, and Y. Huang, *Frontiers in Genetics* **11**, 581668 (2020).

[16] Z. Ruan, C. Liu, Y. Guo, Z. He, X. Huang, X. Jia, and T. Yang, *Journal of Medical Virology* **93**, 389 (2021).

[17] M. Kandeel, Y. Kitade, and A. Almubarak, *PeerJ (San Francisco, CA)* **8**, e10480 (2020).

[18] R. Cozac, N. Medzhidov, and S. Yuki, Predicting inhibitors for SARS-CoV-2 RNA-dependent RNA polymerase using machine learning and virtual screening, `arxiv.org`, 2020.

[19] S. Koulgi, V. Jani, M. Uppuladinne, U. Sonavane, A. K. Nath, H. Darbari, and R. Joshi, *Journal of Biomolecular Structure & Dynamics* , 1 (2020).

[20] S. Guo, H. Xie, Y. Lei, B. Liu, L. Zhang, Y. Xu, and Z. Zuo, *Bioorganic Chemistry* **110**, 104767 (2021).

[21] M. U. Mirza and M. Froeyen, *Journal Of Pharmaceutical Analysis* **10**, 320 (2020).

[22] P. Delre, F. Caporuscio, M. Saviano, and G. F. Mangiatordi, *Frontiers in Chemistry* **8**, 594009 (2020).

[23] X. Xu, Y. Liu, S. Weiss, E. Arnold, S. G. Sarafianos, and J. Ding, *Nucleic Acids Research* **31**, 7117 (2003).

[24] P. Procacci, M. Macchiagodena, M. Pagliai, G. Guarnieri, and F. Iannone, *Chemical Communications (Cambridge, England)* **56**, 8854 (2020).

[25] J. S. Morse, T. Lalonde, S. Xu, and W. R. Liu, *Chembiochem : a European Journal of Chemical Biology* **21**, 730 (2020).

[26] M. Bucci, *Nature Chemical Biology* **16**, 712 (2020).

[27] K. Strømgaard, P. Krogsgaard-Larsen, and U. Madsen, editors, *Textbook of drug design and discovery*, CRC Press, Taylor & Francis Group, Boca Raton ; London ; New York, fifth edition. edition, 2017.

[28] D. A. Pereira and J. A. Williams, *British Journal of Pharmacology* **152**, 53 (2007).

[29] I. D. Kuntz, *Science (American Association for the Advancement of Science)* **257**, 1078 (1992).

[30] A. Ilari and C. Savino, *Protein Structure Determination by X-Ray Crystallography*, pages 63–87, Humana Press, Totowa, NJ, 2008.

[31] J. M. Würz, S. Kazemi, E. Schmidt, A. Bagaria, and P. Güntert, *Archives of Biochemistry and Biophysics* **628**, 24 (2017), Nuclear Magnetic Resonance.

[32] J. T. Seffernick and S. Lindert, *The Journal of Chemical Physics* **153**, 240901 (2020).

[33] R. E. Amaro, J. Baudry, J. Chodera, O. Demir, J. A. McCammon, Y. Miao, and J. C. Smith, *Biophysical Journal* **114**, 2271 (2018).

[34] E. Nwanochie and V. N. Uversky, *International Journal of Molecular Sciences* **20** (2019).

[35] T. E. Lewis, I. Sillitoe, A. Andreeva, T. L. Blundell, D. W. A. Buchan, C. Chothia, D. Cozzetto, J. M. Dana, I. Filippis, J. Gough, D. T. Jones, L. A. Kelley, G. J. Kleywegt, F. Minneci, J. Mistry, A. G. Murzin, B. Ochoa-Montaño, M. E. Oates, M. Punta, O. J. L. Rackham, J. Stahlhacke, M. J. Sternberg, S. Velankar, and C. Orengo, *Nucleic Acids Research* **43**, D382 (2015).

[36] D. L. Parton, P. B. Grinaway, S. M. Hanson, K. A. Beauchamp, and J. D. Chodera, *PLoS Computational Biology* **12**, e1004728 (2016).

[37] M. AlQuraishi, *Bioinformatics* **35**, 4862 (2019).

[38] K. R.M.A, K. I.D, and O. C.M, *Journal of Molecular Biology* **266**, 424 (1997).

[39] W. Patrick Walters, M. T. Stahl, and M. A. Murcko, *Drug Discovery Today* **3**, 160 (1998).

[40] A. Kalenkiewicz, B. J. Grant, and C.-Y. Yang, *Biology (Basel, Switzerland)* **4**, 344 (2015).

[41] S. J. Teague, *Nature Reviews. Drug Discovery* **2**, 527 (2003).

[42] H. A. Carlson and J. A. McCammon, *Molecular Pharmacology* **57**, 213 (2000).

[43] D. E. Koshland, *Proceedings of the National Academy of Sciences - PNAS* **44**, 98 (1958).

[44] W. L. Jorgensen, *Science (American Association for the Advancement of Science)* **254**, 954 (1991).

[45] C. Talbot, *School Science Review* **93**, 9 (2011).

[46] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, *Science (American Association for the Advancement of Science)* **254**, 1598 (1991).

[47] C.-J. Tsai, S. Kumar, B. Ma, and R. Nussinov, *Protein Science* **8**, 1181 (1999).

[48] S. Uehara and S. Tanaka, *Journal of Chemical Information and Modeling* **57**, 742 (2017).

[49] J. M. Ostrem, U. Peters, M. L. Sos, J. A. Wells, and K. M. Shokat, *Nature (London)* **503**, 548 (2013).

[50] V. Oleinikovas, G. Saladino, B. P. Cossins, and F. L. Gervasio, *Journal of the American Chemical Society* **138**, 14257 (2016).

[51] A. Kuzmanic, G. R. Bowman, J. Juarez-Jimenez, J. Michel, and F. L. Gervasio, *Accounts of Chemical Research* **53**, 654 (2020).

[52] P. Cimermancic, P. Weinkam, T. J. Rettenmaier, L. Bichmann, D. A. Keedy, R. A. Woldeyes, D. Schneidman-Duhovny, O. N. Demerdash, J. C. Mitchell, J. A. Wells, J. S. Fraser, and A. Sali, *Journal of Molecular Biology* **428**, 709 (2016).

[53] O. Trott and A. J. Olson, *Journal of Computational Chemistry* **31**, 455 (2010).

[54] S. Huang and X. Zou, *Proteins, Structure, Function, and Bioinformatics* **66**, 399 (2007).

[55] C. Strecker and B. Meyer, *Journal of Chemical Information and Modeling* **58**, 1121 (2018).

[56] P. Bradley, K. M. S. Misura, and D. Baker, *Science (American Association for the Advancement of Science)* **309**, 1868 (2005).

[57] A. Wang, Y. Zhang, H. Chu, C. Liao, Z. Zhang, and G. Li, *Journal of Chemical Information and Modeling* **60**, 2939 (2020).

[58] E. S. D. Bolstad and A. C. Anderson, *Proteins: Structure, Function, and Bioinformatics* **73**, 566 (2008).

[59] X. Hu, L. Hong, M. Dean Smith, T. Neusius, X. Cheng, and J. C. Smith, *Nature Physics* **12**, 171 (2016).

[60] D. Shaw, R. Dror, J. Salmon, J. Grossman, K. Mackenzie, J. Bank, C. Young, M. Deneroff, B. Batson, K. Bowers, et al., Millisecond-scale molecular dynamics simulations on Anton. acm, in *IEEE Conference on Supercomputing (SC09)*, 2009.

[61] G. Torrie and J. Valleau, *Journal of Computational Physics* **23**, 187 (1977).

[62] J. R. Gullingsrud, R. Braun, and K. Schulten, *Journal of Computational Physics* **151**, 190 (1999).

[63] A. Laio and F. L. Gervasio, *Reports on Progress in Physics* **71**, 126601 (2008).

[64] R. Cuchillo, K. Pinto-Gil, and J. Michel, *Journal of Chemical Theory and Computation* **11**, 1292 (2015).

[65] Y. Sugita and Y. Okamoto, *Chemical Physics Letters* **314**, 141 (1999).

[66] F. Comitani and F. L. Gervasio, *Journal of Chemical Theory and Computation* **14**, 3321 (2018).

[67] D. Schmidt, M. Boehm, C. L. McClendon, R. Torella, and H. Gohlke, *Journal of Chemical Theory and Computation* **15**, 3331 (2019).

[68] J. P. Arcon, L. A. Defelipe, E. D. Lopez, O. Burastero, C. P. Modenutti, X. Barril, M. A. Marti, and A. G. Turjanski, *Journal of Chemical Information and Modeling* **59**, 3572 (2019).

[69] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, *The AAPS Journal* **14**, 133 (2012).

[70] T. Sterling and J. J. Irwin, *Journal of Chemical Information and Modeling* **55**, 2324 (2015).

[71] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, *Nucleic Acids Research* **49**, D1388 (2020).

[72] S. Forli, R. Huey, M. E. Pique, M. F. Sanner, D. S. Goodsell, and A. J. Olson, *Nature Protocols* **11**, 905 (2016).

[73] R. J. Rosenfeld, D. S. Goodsell, R. A. Musah, G. M. Morris, D. B. Goodin, and A. J. Olson, *Journal of Computer-aided Molecular Design* **17**, 525 (2003).

[74] G. Varela-Salinas, C. A. García-Pérez, R. Peláez, and A. J. Rodríguez, Visual Clustering Approach for Docking Results from Vina and AutoDock, in *Hybrid Artificial Intelligent Systems*, volume 10334 of *Lecture Notes in Computer Science*, pages 342–353, Cham, 2017, Springer International Publishing.

[75] E. Yuriev, M. Agostino, and P. A. Ramsland, *Journal of Molecular Recognition* **24**, 149 (2011).

[76] E. Yuriev, J. Holien, and P. A. Ramsland, *Journal of Molecular Recognition* **28**, 581 (2015).

[77] D. S. Goodsell and A. J. Olson, *Proteins* **8**, 195 (1990).

[78] T. Gaillard, *Journal of Chemical Information and Modeling* **58**, 1697 (2018).

[79] C. Shen, J. Ding, Z. Wang, D. Cao, X. Ding, and T. Hou, *Wiley Interdisciplinary Reviews. Computational Molecular Science* **10** (2020).

[80] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for Optimization*, The MIT Press, 2019.

[81] F. S. Zariquiey, J. V. de Souza, and A. K. Bronowska, *Scientific Reports* **9**, 19118 (2019).

[82] P. Ghanakota and H. A. Carlson, *The Journal of Physical Chemistry. B* **120**, 8685 (2016).

[83] C. H. Ngan, T. Bohnuud, S. E. Mottarella, D. Beglov, E. A. Villar, D. R. Hall, D. Kozakov, and S. Vajda, *Nucleic Acids Research* **40**, W271 (2012).

[84] A. Bakan, N. Nevins, A. S. Lakdawala, and I. Bahar, *Journal of Chemical Theory and Computation* **8**, 2435 (2012).

[85] L. Zuzic, J. K. Marzinek, J. Warwicker, and P. J. Bond, *Journal of Chemical Theory and Computation* **16**, 5948 (2020).

[86] M. R. Shirts, M. R. Shirts, C. Klein, C. Klein, J. M. Swails, J. M. Swails, J. Yin, J. Yin, M. K. Gilson, M. K. Gilson, D. L. Mobley, D. L. Mobley, D. A. Case, D. A. Case, E. D. Zhong, and E. D. Zhong, *Journal of Computer-aided Molecular Design* **31**, 147 (2017).

[87] A. Soni, Uncovering Cryptic Pockets in Biologically Relevant Proteins: An Improved Computational Methodology, Master's thesis, Indian Institute of Science Education and Research, Pune, Maharashtra, India, 2017.

[88] H. Kim, C. Jang, D. K. Yadav, and M. Kim, *Journal of Cheminformatics* **9**, 1 (2017).

[89] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, *Journal of Chemical Theory and Computation* **3**, 2312 (2007).

[90] J. Sander, *Density-Based Clustering*, pages 270–273, Springer US, Boston, MA, 2010.

[91] D. R. Roe and T. E. Cheatham, *Journal of Chemical Theory and Computation* **9**, 3084 (2013).

[92] `https://amberhub.chpc.utah.edu/cluster/`, [Online; accessed on 14 May 2021].

[93] V. Le Guilloux, P. Schmidtke, and P. Tuffery, *BMC Bioinformatics* **10**, 168 (2009).

[94] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey, *Bioinformatics* **28**, 2074 (2012).

[95] I. R. Craig, C. Pfleger, H. Gohlke, J. W. Essex, and K. Spiegel, *Journal of Chemical Information and Modeling* **51**, 2666 (2011).

[96] D. Bajusz, A. Rácz, and K. Héberger, *Journal of Cheminformatics* **7**, 1 (2015).

[97] RDKit: Open-source cheminformatics, `http://www.rdkit.org`, [Online; accessed 15 May 2021].

[98] L. Schrödinger and W. DeLano, Pymol.

[99] D. E. Shaw Research, Molecular Dynamics Simulations Related to SARS-CoV-2, `https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/`, 2020, D. E. Shaw Research Technical Data.

[100] Y. Gao, L. Yan, Y. Huang, F. Liu, Y. Zhao, L. Cao, T. Wang, Q. Sun, Z. Ming, L. Zhang, J. Ge, L. Zheng, Y. Zhang, H. Wang, Y. Zhu, C. Zhu, T. Hu, T. Hua, B. Zhang, X. Yang, J. Li, H. Yang, Z. Liu, W. Xu, L. W. Guddat, Q. Wang, Z. Lou, and Z. Rao, *Science (American Association for the Advancement of Science)* **368**, 779 (2020).

[101] G. Kokic, H. S. Hillen, D. Tegunov, C. Dienemann, F. Seitz, J. Schmitzova, L. Farnung, A. Siewert, C. Höbartner, and P. Cramer, *Nature Communications* **12**, 279 (2021).

[102] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman, University of California, San Francisco, 2018.

[103] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, *Journal of Computational Chemistry* **30**, 2157 (2009).

[104] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, *Journal of Cheminformatics* **3**, 1 (2011).

[105] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, *Journal of Computational Chemistry* **16**, 2785 (2009).

[106] P. Schmidtke, A. Bidon-Chanal, F. J. Luque, and X. Barril, *Bioinformatics* **27**, 3276 (2011).

[107] Amber developers, `https://amberhub.chpc.utah.edu/`, [Online; accessed on 30 May 2021].

[108] M. Moret and G. Zebende, *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **75**, 011920 (2007).