UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Combined ligand and structure-based computational drug discovery of innovative Aquaporin modulators

Sara Gabriela Ferraz Ferreira

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Doutor Bruno Lourenço da Silva Victor

2020

“Suckin’ at something is the first step to being sorta good at something.”
- Jake the Dog

Acknowledgments

Right, acknowledgments, how does one start them?

After conducting an exhaustive search in this area of study, the main conclusion was that the first person to thank is the supervisor. Having this in mind, I would like to acknowledge my stupendous supervisor, Dr. Bruno Victor. I would have to write a whole another thesis in order to refer every single thing that I'm thankful for, but since my brain isn't capable of handling that kind of challenge, at least for now, let's keep it simple. Thank you for the opportunity to work on such a challenging project, thank you for your patience every time I had a question that could be incorporated in a "science for dummies" kind of book, thank you for allowing me to have a voice, even when it meant disagreeing with you, thank you for letting me express myself and leave part of myself in everything I did, which was often associated with using a whole rainbow of colors in my figures (very girly, I know), and finally, thank you for telling me that I wrote like Saramago and that my sentences were waaaaay to long (this one was made incredibly long in your honor). This past year was a lot, in every possible way (Covid and stuff), but I'm forever thankful for the tremendous support you were able to give me, even if most of it was through Zoom calls.

Okay, the hardest part is done. Now, most people like to thank those who surrounded them throughout the past year, so let's do that. I could refer to every single person that encompasses our research group, but the truth is that it has been growing exponentially in the past few months, so let's make this a bit general. First of all, I have to thank the bosses, for every word of support, for every time they were curious about my project, for every time they made fun of my funky colors, and for allowing me to work in a fantastic workspace with my own desk and shelves that I couldn't reach. Second, I have to thank the younger ones for always making me feel that I was part of the group, for all the lunch hours, and for not judging me, at least not a lot, for not playing Pokémon with them.

And now what? Family? Friends? Well, let's do them together then. First, my mom, Paula, my completely out of this world mom. Thank you for everything (let's be honest, if it wasn't for you, I wouldn't even exist), thank you for sharing, again, a house with me after I moved out five years ago (Covid and stuff, again), thank you for always making my favourite dishes when I was feeling down, thank you for always feeding me candy while saying that the brain also needs to eat, and lastly thank you for the financial support (let's not be hypocrites and say that it doesn't matter). Then, my sister, Rafaela, a pretty amazing human being. Thank you for always talking me down in my worst days, thank you for understanding my anxiety, thank you for the company in the last months, and finally, thank you for always making me see the good side of things, even though I always failed in this task. Third, my father, Joge, thank you for listening to me every night, thank you for always asking how things were going, for always trying to understand what I was doing in front of a screen for so many hours, and once again for the financial support. Fourth, my dog, Poppy, thank you for always being the perfect excuse when I wanted to take a break, being it to pet you or to take you to the street. Lastly, my dear friends. Although a lot of people surrounded me in the last year, there are four names that I cannot ignore: Miguel, Raquel, Rita, and Francisco. For the first three, thank you for always being there, since I can remember actually, if it wasn't for you guys, I wouldn't be here, and you know it. Thank you for the late-night talks, every message asking, "Coffee tonight?" even though I only drink tea, thank you for always lifting my head, and lastly for making me believe in myself (I swear no one believes in me like you three). For the last one, even though you entered this adventure almost at the end, thank you for making sure that I wasn't taking this too far and losing myself along the way, thank you for letting me scream when I was frustrated, and for smiling alongside me every time I felt proud of this work.

And now, the overall universe. Thank you for allowing me to have a fantastic year, that even though was extremely tiring, never felt like "work" to me.

Resumo

As Aquaporinas (AQPs) são uma família de 13 proteínas membranares, cuja principal função é promover o transporte passivo de água, através da célula, em resposta a gradientes osmóticos gerados pelo transporte ativo de solutos [1–3]. As AQPs podem ser divididas em dois grupos, de acordo com os materiais transportados: aquelas que transportam exclusivamente água, que é o caso do nosso objeto de estudo APQ - subtipo 1; e aquelas que para além de água podem transportar pequenos solutos, como por exemplo glicerol [4] - denominadas aquagliceroporinas. Estas proteínas, que possuem a capacidade de atravessar a membrana plasmática, apresentam uma estrutura na forma de um homotetrámero. Cada um dos monómeros inclui um vestíbulo extracelular e citoplasmático, que é conectado através de uma região central, anfipática, designada por poro, num arranjo em forma de barril [5]. Em contraste com a maioria dos canais iónicos, nas AQPs a região do poro não reside no eixo de simetria das quatro subunidades; em vez disso, cada um dos quatro monómeros contém o seu próprio poro [6]. Duas regiões do poro são essenciais para a função das AQPs: a região aromática-arginina, ou filtro de seletividade ar/R, e a região asparagina-prolina-alanina (Asn-Pro-Ala), ou motivo NPA [7]. Tanto o filtro de seletividade ar/R como o motivo NPA participam da exclusão de protões, uma vez que as cargas positivas dos seus resíduos ajudam a repelir a entrada destas partículas no poro da AQP.

As AQPs são amplamente expressas em todo o reino animal e vegetal, estando localizadas na membrana plasmática e nos compartimentos citoplasmáticos, particularmente em tipos de células envolvidas no transporte de fluidos. Foi já comprovado que as AQPs desempenham papéis cruciais no desenvolvimento biológico de tumores, estando incluídas na sua histologia, proliferação, migração, angiogénese, e edema associado [8]. Esta correlação é devida ao facto de as AQPs serem sobre expressas nestes tecidos, em comparação com tecidos ditos normais. Consequentemente, as AQPs podem funcionar como potenciais alvos diagnósticos e terapêuticos no tratamento do cancro, uma vez que a sua inibição em células endoteliais e tumorais pode limitar o crescimento e a disseminação do tumor [2,4]. Infelizmente, a taxa de sucesso para a identificação de pequenas moléculas moduladoras de AQPs é muito baixa em comparação com outras proteínas membranares, e os poucos moduladores farmacológicos identificados carecem de especificidade ou acabam por revelar uma alta toxicidade [6]. Uma possível explicação para esta baixa "drogabilidade" é o reduzido tamanho do monómero funcional das AQPs e o pequeno diâmetro do seu poro. No entanto, o contínuo crescimento do conhecimento acerca da relação estrutura-função das AQPs, particularmente da geometria, a nível atómico, dos resíduos hidrofóbicos e hidrofílicos específicos da região dos poros, torna as AQPs um alvo terapêutico promissor para ser futuramente usado em campanhas de descoberta de novos fármacos [5].

Na sua vertente experimental, as campanhas de descoberta de fármacos são reconhecidas por serem muito demoradas, arriscadas, e caras [9]. Felizmente, com os constantes avanços na tecnologia e o aumento da compreensão dos princípios fundamentais das interações proteína-ligando, as etapas iniciais das campanhas de descoberta de fármacos estão agora mais focadas em previsões computacionais, nomeadamente abordagens baseadas no ligando (*Ligand-based*) e na estrutura (*Structure-based*).

A abordagem baseada no ligando é um método fundamentado no conhecimento e informação, acerca de compostos ativos e inativos, relativamente a um alvo específico [10]. Este método indireto de desenvolvimento de novos fármacos baseia-se na compreensão da ligação de uma determinada molécula a um alvo de interesse. Essa molécula pode então ser usadas para derivar um modelo farmacóforo que determina as características estruturais mínimas indispensáveis que uma outra molécula deve reter para ser capaz de se ligar ao alvo em estudo [11].

O conceito baseado na estrutura é fundamentado na perceção da estrutura tridimensional do alvo biológico, obtida, por exemplo, através de métodos de cristalografia de raios-X [12]. Este método direto

de desenvolvimento de fármacos é análogo ao *screening* de alto rendimento (HTS), onde o conhecimento da estrutura alvo e do ligando é imperativo [13]. Esta abordagem depende de uma filtragem automática de várias moléculas, na procura daquelas que evoquem a resposta biológica desejada.

Uma das principais razões pelas quais os fármacos acabam por falhar nos ensaios clínicos é porque o efeito pretendido, no alvo desejado, não é produzido [14]. Desta forma, a identificação apropriada do alvo é uma etapa crucial no desenvolvimento de novos fármacos. Um bom alvo precisa de satisfazer vários critérios: ser eficaz; seguro; atender às necessidades clínicas e comerciais; e, acima de tudo, ser "drogável". Um alvo "drogável" é acessível ao fármaco em teste e, após a sua ligação ao mesmo, evoca a resposta biológica desejada que é consistente tanto *in vitro* como *in vivo* [9]. Posteriormente ao processo de validação do alvo, os ensaios de *screening* de compostos são desenvolvidos durante a fase de identificação de *hits* [9]. Um composto pode ser descrito como um *hit* se mostrar a atividade desejada através de um *screening* do composto em questão, e se a sua atividade puder ser confirmada após um novo teste. Desta forma, os desafios significativos deste campo incluem a identificação de *hits* promissores e o desenvolvimento de *leads* de alta qualidade para um maior desenvolvimento em potenciais candidatos clínicos [10]. Infelizmente, ainda surgem vários desafios nesta área de estudo, incluindo os limites definidos pelas estruturas proteicas disponíveis, o reconhecimento de ligandos de referência para a proteína alvo, a identificação de resultados promissores, e o subsequente desenvolvimento de compostos de alta qualidade, para melhorias adicionais, em candidatos clínicos.

Dito isto, de forma a obter resultados satisfatórios e a superar estas limitações, novas e inovadoras abordagens e métodos computacionais devem ser desenvolvidos. Consequentemente, o trabalho apresentado nesta tese foi focado no desenvolvimento de um novo plano de trabalho computacional assente em vários métodos distintos que combinam abordagens inovadoras, baseadas no ligando e na estrutura, para identificar novos moduladores da AQP1.

Na abordagem baseada no ligando, a identificação de novos moduladores da AQP1 foi realizada com base nas características químicas mais importantes descritas a partir de moduladores de AQPs já identificados. Essas mesmas características já foram previamente usadas em campanhas de *screening* virtual na procura de novos potenciais moduladores terapêuticos, mas com uma eficiência limitada. Para fortalecer a nossa triagem, aplicámos essa mesma abordagem a várias bases de dados, disponíveis comercialmente, para garantir a construção de uma base de dados global de compostos, com a mais ampla diversidade química possível. Juntamente com abordagens de *screening* virtual, uma combinação com *Docking* molecular permitiu a identificação de moduladores alternativos para a AQP1. Nesta abordagem, o foco foi na região dos poros da proteína, que é, até ao momento, a região mais importante e caracterizada como "drogável". No entanto, e como abordagem alternativa, também foi feita a identificação de novas regiões sensíveis a influenciar a função da AQP1. Usando uma abordagem multiconformacional, com base em simulações de Dinâmica Molecular (MD) da AQP1, totalmente incorporada na bicamada fosfolipídica da membrana, tentámos identificar novas regiões-alvo na superfície da proteína.

O uso de conformações adicionais de MD da proteína AQP1 é o que distingue o nosso trabalho de estudos previamente descritos nesta área de investigação, e o que nos dá uma vantagem na identificação de novos moduladores. Esta abordagem multi-conformacional foi seguida com o objetivo de encontrar as conformações mais prováveis para uso neste tipo de estudos, superando assim as limitações impostas por um sistema de proteína membranar com diferentes monómeros.

Posteriormente, usámos estas regiões em novas campanhas de *screening* virtual, baseadas na estrutura, semelhantes ao que tem sido descrito anteriormente para a região dos poros. Após uma classificação de eficiência de ligação baseada na energia e caracterização estrutural obtidas a partir do *screening* baseado na estrutura, foi criada uma pequena lista dos moduladores mais promissores, para a AQP1. Esta lista foi então usada em novos ensaios de afinidade de ligação através de uma abordagem combinada de simulações de Dinâmica Molecular e cálculos MM/PBSA. Esta abordagem permitiu atribuir uma pontuação, e posterior classificação, mais precisas dos compostos mais eficientes que se ligam

à AQP1, o que tornou possível a criação de uma pequena lista final de compostos cujo efeito sobre a função desta proteína será avaliado.

Para garantir a validade dos resultados obtidos, todas as previsões computacionais realizadas terão, no futuro, a sua contrapartida experimental.

Palavras Chave: aquaporinas, tratamento do cancro, *screening* virtual, desenvolvimento de fármacos baseado no ligando, desenvolvimento de fármacos baseado na estrutura.

Abstract

Aquaporins (AQPs) are a family of 13 small integral membrane proteins whose primary function is to promote the passive transport of water across the cell in response to osmotic gradients generated by solutes' active transport [1–3]. These small membrane-spanning proteins assemble as homotetramers. Each Monomer comprises an extracellular and cytoplasmic vestibule connected by a central amphipathic pore region in a barrel-like arrangement [5]. AQPs are widely expressed throughout the animal and plant kingdom, being localized in the plasma membrane and the cytoplasmic compartments, particularly in cell types involved in fluid transport. AQPs have been proven to play crucial roles in tumor biology, including histological tumor grade, proliferation, migration, angiogenesis, and tumor-associated edema [8]. This correlation is due to AQPs' over-expression when compared to normal tissues.

Taking these facts into account, AQPs are seen as new potential diagnostic and therapeutic targets in anticancer treatment since their inhibition in endothelial and tumor cells might limit tumor growth and spread [2,4]. Unfortunately, the hit rate for identifying small-molecule AQP modulators appears to be very low when compared to other membrane proteins, mainly due to the lack of specificity or high toxicity [6]. One possible explanation for these proteins' low druggability is the small size of the functional AQP Monomer and its small pore diameter. Nonetheless, the continuous growth of structure-function knowledge on AQPs, particularly the atomic-level geometry of specific hydrophobic and hydrophilic residues in the pore region, makes AQPs a promising therapeutic target to be used in future Structure-based drug discovery campaigns [5].

In its experimental aspect, drug discovery campaigns are recognized to be time-consuming, risky, and costly [9]. Fortunately, with the constant advances in technology and the increased understanding of the fundamental principles of protein-ligand interactions, the early stages of drug discovery campaigns are now focused on the use of computational predictions, namely Ligand and Structure-based drug design approaches. Several challenges still arise from this field, including the boundaries defined by the available protein structures, the recognition of reference ligands for the target protein, the identification of promising hits, and the subsequent development of high-quality leads for further improvement into clinical candidates. Therefore, to achieve successful results and overcome these limitations, new and innovative computational approaches and methods must be developed. Consequently, the work presented in this thesis was focused on developing a new computational workflow based on several distinct methods that combine innovative Ligand and Structure-based approaches to identify new AQP1 modulators.

The multiconfigurational use of different AQP conformations obtained from Molecular Dynamics simulations distinguishes our work from previously described studies focused on this protein target, giving us a fundamental advantage in identifying new AQP1 modulators. This approach was pursued in order to find the most representative conformations of the protein target that could be used as the reference for subsequent Structure-based virtual Molecular Docking screening campaigns. By using different Molecular Docking approaches, together with various commercially available compound Databases, we have identified a list of compounds with promising activity for AQP1 function modulation. Lastly, to narrow down and sort this list, we have used MM/PBSA methods to determine these compounds' binding free energy at the previously determined binding regions at the AQP1 surface. In the end, the most promising hits identified under this thesis work will be purchased and experimentally tested as new AQP1 modulators.

Keywords: aquaporins, anticancer treatment, virtual screening, Structure-based drug discovery, Ligand-based drug discovery.

Table of Contents

Acknowledgments II
Resumo III
Abstract VI
Table of Contents VII
List of Tables IX
List of Figures X
List of Abbreviations XV
1. Introduction 1
1.1. Aquaporins 1
1.2. Computational Approach 6
2. Aim of this work 7
3. Theory 8
3.1. Molecular Modeling: Description and Relevance 8
Molecular Mechanics 9
Potential Energy Function 9
Force Field 12
Binding pocket predictions 13
DogSiteScorer 13
AutoLigand 13
SeeSAR 14
Molecular Docking 14
Molecular Dynamics 16
Automated force field Topology Builder 18
Periodic Boundary Conditions 18
Treatment of nonbonded interactions 19
Temperature and Pressure control 20
Energy Minimization 21
MM/PBSA 22
3.2. Analysis 24
RMSD 24
4. Methodological Workflow 25
4.1. Identification of the binding sites of AQP1 25
4.1.1. Creating and preparing a small compound ligand Database for the identification of AQP1 binding regions 25

4.1.2. Preparing the protein and ligand files for the Docking screening with ADT4 and VINA 25
4.1.3. Preparing the Docking calculations for ADT4 26
4.1.4. Preparing the Docking calculations for VINA 27
4.1.5. Sampling the conformational space of AQP1 using MD simulations 28
4.1.6. Binding pocket predictions 29
4.1.7. Compound Database enrichment using a Ligand-based approach 29
4.1.8. Quantitative evaluation of Ligand-Protein Interactions 30
4.2. Screening of commercially available Databases with VINA 30
4.3. MM/PBSA calculations of the most promising identified AQP1 modulators 32
5. Results and Discussion 34
5.1. Screening with a 35 compound Database 34
5.2. Screening with a 35 compound Database using different AQP1 structures 37
5.3. Binding pocket predictions 38
5.4. Screening with a 7204 compound Database using different AQP1 structures 40
5.5. Quantitative evaluation of Ligand-Protein Interactions 42
5.6. Molecular Docking screening to the regions and conformers of interest using different commercially available Compound Databases 50
5.7. MM/PBSA calculations 56
6. Concluding Remarks 63
7. Future Work 64
References 65
Appendix 77

List of Tables

Table 5.1. Scoring function values calculated with VINA, for the top 20 compounds, for the B_NPA binding site 51

Table 5.2. Scoring function values calculated with VINA, for the top 20 compounds, for the C_ar/R binding site 51

Table 5.3. Scoring function values calculated with VINA, for the top 20 compounds, for the C_NPA binding site 52

Table 5.4. Binding Energy values calculated with g_mmpbsa, for the top 20 compounds, for the B_NPA binding site 58

Table 5.5. Binding Energy values calculated with g_mmpbsa, for the top 20 compounds, for the C_ar/R binding site 58

Table 5.6. Binding Energy values calculated with g_mmpbsa, for the top 20 compounds, for the C_NPA binding site 59

Table 5.7. Binding Energy, Van der Wall Energy, Electrostatic Energy, Polar Solvation Energy, and SASA Energy values, for the top 5 compounds, based on the MM/PBSA ranking, for the B_NPA binding site 60

Table 5.8. Binding Energy, Van der Wall Energy, Electrostatic Energy, Polar Solvation Energy, and SASA Energy values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_ar/R binding site 60

Table 5.9. Binding Energy, Van der Wall Energy, Electrostatic Energy, Polar Solvation Energy, and SASA Energy values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_NPA binding site 60

Table A.1. Molweight, cLogP, cLogS, Hydrogen Acceptor, Hydrogen Donors, Rotatable Bonds, and Aromatic Rings values, for the top 5 compounds, based on the MM/PBSA ranking, for the B_NPA binding site 85

Table A.2. Molweight, cLogP, cLogS, Hydrogen Acceptor, Hydrogen Donors, Rotatable Bonds, and Aromatic Rings values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_ar/R binding site 85

Table A.3. Molweight, cLogP, cLogS, Hydrogen Acceptor, Hydrogen Donors, Rotatable Bonds, and Aromatic Rings values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_NPA binding site 85

List of Figures

Figure 1.1. Crystallographic structure of AQP1 embedded in a POPC membrane bilayer 1

Figure 1.2. Topology of an AQP protein embedded in a POPC membrane bilayer 2

Figure 1.3. Crystallographic structure of AQP1 highlighting the four Monomers.....3

Figure 1.4. Monomer A of the Crystallographic structure of AQP1 highlighting the ar/R selectivity filter and the NPA motif.....4

Figure 1.5. Top and bottom view of the Crystallographic structure of AQP1 highlighting the four pores and each of the four Monomers.....5

Figure 3.1. Representation of the Lennard-Jones potential.....11

Figure 3.2. Representation of the five contributions to a Molecular Mechanics Force Field 12

Figure 3.3. Schematic representation of the MD algorithm 17

Figure 3.4. Representation of PBC in two dimensions 19

Figure 3.5. Representation of the local and global minimum in an energy minimization problem.21

Figure 3.6. Graphical representation of the steepest descent method 22

Figure 4.1. Crystallographic structure of AQP1 highlighting the “top” grid (extracellular side) and the used parameters for ADT4 calculations. 26

Figure 4.2. Crystallographic structure of AQP1 highlighting the “bottom” grid (cytoplasmic side) and the used parameters for ADT4 calculations. 27

Figure 4.3. Crystallographic structure of AQP1 highlighting the “top” grid (extracellular side) and the used parameters for VINA calculations.....28

Figure 4.4. Crystallographic structure of AQP1 highlighting the “bottom” grid (cytoplasmic side) and the used parameters for VINA calculations. 28

Figure 4.5. Monomer B of the MD2 structure of AQP1 highlighting the “top” grid (extracellular side) and the used parameters for VINA calculations.....31

Figure 4.6. Monomer C of the MD2 structure of AQP1 highlighting the “top” grid (extracellular side) and the used parameters for VINA calculations.....31

Figure 4.7. Monomer C of the MD2 structure of AQP1 highlighting the “bottom” grid (cytoplasmic side) and the used parameters for VINA calculations.....32

Figure 5.1. Crystallographic structure of AQP1 after undergoing Molecular Docking using the ADT4 software with 35 ligands.....35

Figure 5.2. Crystallographic structure of AQP1 after undergoing Molecular Docking using the VINA software with 35 ligands.....35

Figure 5.3. Representation of the RMSD variation through simulation time (50 ns) of the five replicate MD simulations concerning the Crystallographic structure conformation of AQP1.....36

Figure 5.4. Monomer D of the MD1 structure of AQP1, highlighting the ar/R selectivity filter....36

Figure 5.6. Top view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Molecular Docking using the VINA software with 35 ligands.....38

Figure 5.6. Bottom view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Molecular Docking using the VINA software with 35 ligands.....38

Figure 5.7. Top view of the Crystallographic structure of AQP1 after undergoing three different automated pocket detection approaches.....39

Figure 5.8. Bottom view of the Crystallographic structure of AQP1 after undergoing three different automated pocket detection approaches.....40

Figure 5.9. Top view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Molecular Docking using the VINA software with 7204 ligands.....40

Figure 5.10. Bottom view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Molecular Docking using the VINA software with 7204 ligands.....42

Figure 5.11. Top view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing the “pairwisedistances.py” Python script.....43

Figure 5.12. Identification of the population of interactions between the compound dataset and the extracellular side of the Crystallographic structure of AQP1.....43

Figure 5.13. Identification of the population of interactions between the compound dataset and the extracellular side of the MD1 structure of AQP1. 44

Figure 5.14. Identification of the population of interactions between the compound dataset and the extracellular side of the MD2 structure of AQP1. 44

Figure 5.15. Identification of the population of interactions between the compound dataset and the extracellular side of the MD3 structure of AQP1. 45

Figure 5.16. Bottom view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing the “pairwisedistances.py” Python script.....46

Figure 5.17. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the Crystallographic structure of AQP1. 46

Figure 5.18. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the MD1 structure of AQP1.....47

Figure 5.19. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the MD2 structure of AQP1.....47

Figure 5.20. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the MD3 structure of AQP1.....48

Figure 5.21. Monomer B of the MD2 structure of AQP1 highlighting the B_NPA binding site and the NPA motif 49

Figure 5.22. Monomer C of the MD2 structure of AQP1 highlighting the C_ar/R binding site and the ar/R selectivity filter.....49

Figure 5.23. Monomer C of the MD2 structure of AQP1 highlighting the C_NPA binding site and the NPA motif 50

Figure 5.24. Identification of the population of interactions between the compound dataset and the B_NPA region. 52

Figure 5.25. Monomer B of the MD2 structure of AQP1, highlighting the B_NPA binding pocket 53

Figure 5.26. Identification of the population of interactions between the compound dataset and the C_ar/R region. 53

Figure 5.27. Monomer C of the MD2 structure of AQP1, highlighting the C_ar/R binding pocket 54

Figure 5.28. Identification of the population of interactions between the compound dataset and the C_NPA region. 54

Figure 5.29. Monomer C of the MD2 structure of AQP1, highlighting the C_NPA binding pocket 55

Figure 5.30. Representation of the RMSD variation through simulation time (30 ns) of five compounds bound at the B_NPA binding site.....57

Figure 5.31. Identification of the population of interactions between the CHEMBL4090944 compound and the B_NPA binding site.....62

Figure 5.32. Monomer B of the MD2 structure of AQP1, highlighting the CHEMBL4090944 compound, the interacting amino acids, and the existing H-bonds. 62

Figure 5.33. Identification of the population of interactions between the CHEMBL255641 compound and the C_ar/R binding site. 63

Figure 5.34. Monomer C of the MD2 structure of AQP1, highlighting the CHEMBL255641 compound, the interacting amino acids, and the existing H-bonds 63

Figure 5.35. Identification of the population of interactions between the CHEMBL500565 compound and the C_NPA binding site.....64

Figure 5.36. Monomer C of the MD2 structure of AQP1, highlighting the CHEMBL500565 compound, the interacting amino acids, and the existing H-bonds 64

Figure A.1. Top view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Automated Pocket Detection using DogSiteScorer 77

Figure A.2. Bottom view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Automated Pocket Detection using DogSiteScorer 77

Figure A.3. Top view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Automated Pocket Detection using AutoLigand 78

Figure A.4. Bottom view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Automated Pocket Detection using AutoLigand 78

Figure A.5. Top view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Automated Pocket Detection using SeeSar 79

Figure A.6. Bottom view of the Crystallographic structure of AQP1 and the three MD snapshots after undergoing Automated Pocket Detection using SeeSar 79

Figure A.7. Representation of the RMSD variation through simulation time (10 ns) of twelve compounds bound at the B_NPA binding site of AQP1. 80

Figure A.8. Representation of the RMSD variation through simulation time (50 ns) of two compounds bound at the B_NPA binding site of AQP1. 80

Figure A.9. Representation of the RMSD variation through simulation time (100 ns) of one compound bound at the B_NPA binding site of AQP1. 81

Figure A.10. Representation of the RMSD variation through simulation time (10 ns) of thirteen compounds bound at the C_ar/R binding site of AQP1. 81

Figure A.11. Representation of the RMSD variation through simulation time (30 ns) of three compounds bound at the C_ar/R binding site of AQP1. 82

Figure A.12. Representation of the RMSD variation through simulation time (50 ns) of four compounds bound at the C_ar/R binding site of AQP1. 82

Figure A.13. Representation of the RMSD variation through simulation time (10 ns) of eleven compounds bound at the C_NPA binding site of AQP1. 83

Figure A.14. Representation of the RMSD variation through simulation time (30 ns) of seven compounds bound at the C_NPA binding site of AQP1. 83

Figure A.15. Representation of the RMSD variation through simulation time (50 ns) of one compound bound at the C_NPA binding site of AQP1. 84

Figure A.16. Representation of the RMSD variation through simulation time (100 ns) of one compound bound at the C_NPA binding site of AQP1. 84

Figure A.17. Top 20 compounds for the B_NPA binding site 86

Figure A.18. Top 20 compounds for the C_ar/R binding site 87

Figure A.19. Top 20 compounds for the C_NPA binding site 88

List of Abbreviations

AQP	Aquaporin
ADT4	AutoDock 4
ar/R	Aromatic-arginine region
ATB	Automated Force Field Topology Builder
B_NPA	Extracellular side, including the NPA motif and the region above it, of Monomer B of the MD2 structure of AQP1
C_ar/R	Extracellular side, including the ar/R selectivity filter and the region above it, of Chain C of the MD2 structure of AQP1
C_NPA	Cytoplasmic side, including the NPA motif and the region below it, of Chain C of the MD2 structure of AQP1
DoG	Difference of Gaussian
EJIBCE	Encontro de Jovens Investigadores em Biologia Computacional Estrutural
Ftrees	Feature Trees
H-bond	Hydrogen Bond
HTS	High-throughput screening
LIE	Linear interaction energy
MM/PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
NPA	Asparagine-proline-alanine region
NPT	Constant number of particles, pressure, and temperature
PBC	Periodic boundary conditions
pdb	Protein Data Bank
pdbqt	Protein Data Bank Partial Charge Atom Type
PEF	Potential energy function
PME	Particle mesh Ewald
POPC	Palmitoyl-oleoyl-phosphatidyl-choline
QM	Quantum Mechanics
RF	Reaction Field
RMSD	Root-mean-square deviation
SASA	Solvent-accessible surface area
SPC	Simple Point Charge
VINA	AutoDock Vina

Amino Acids:

Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartate
Cys	Cysteine
Glu	Glutamate
Gln	Glutamine
Gly	Glycine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
Phe	Phenylalanine
Pro	Proline
Ser	Serine
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine

Chapter 1

1. Introduction

1.1. Aquaporins

Aquaporins (AQPs) are a highly conserved group of mechanosensitive transmembrane proteins, of predominantly alpha-helical composition [15], whose functions enable the passive transport of water and small solutes through cell membranes in response to osmotic and hydrostatic pressure or concentration gradients [5,16].

To date, there are thirteen types of aquaporins identified in mammals, denoted AQP0-AQP12 [17], being widely expressed in the plasma membrane and in the cytoplasmic compartments, particularly in cell types involved in fluid transport (Figure 1.1) [5,18,19]. AQPs can be divided into two groups, according to the transported materials: those that are exclusively selective for water (AQP0, 1, 2, 4, 5, 6, 8, 11, and 12) [20], and those that additionally to water can transport small solutes, such as glycerol [4], and several other permeants such as urea [21], ammonia, carbon dioxide [22], hydrogen peroxide, metalloids, nitric oxide, and even anions [23,24] - called aquaglyceroporins (AQP3, 7, 9, and 10) [20]. However, AQPs are rather strict about excluding charged substrates, disabling the passage of cations, protons, inorganic cations, and ammonium [25], due to natural electrostatic energy barriers that prevent the transit of charged molecules and consequently maintaining the cell membrane's electrochemical potential [15,26].

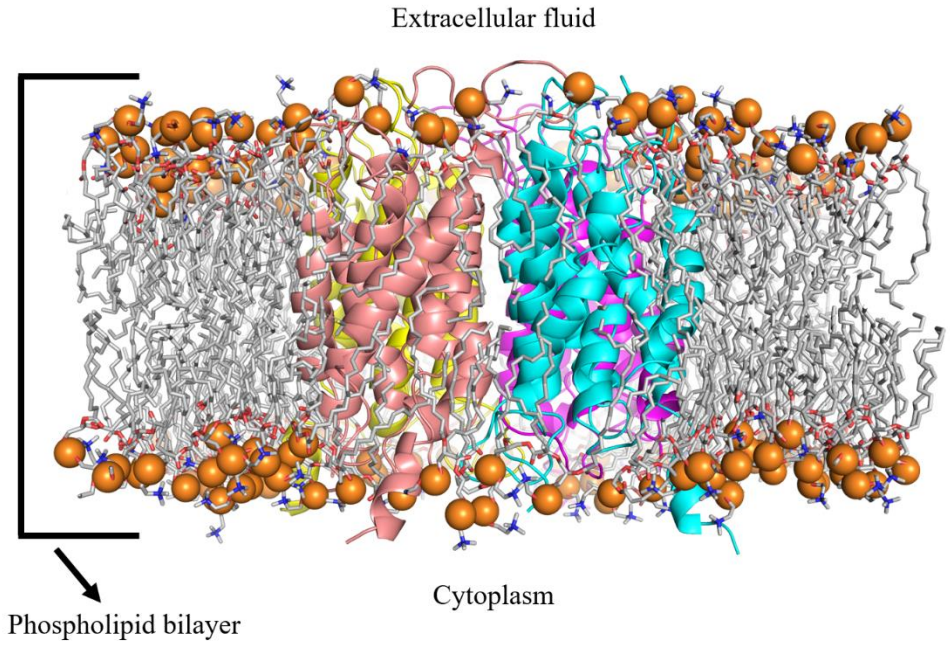

Figure 1.1. Crystallographic structure of AQP1 embedded in a palmitoyl-oleoyl-phosphatidyl-choline (POPC) membrane bilayer. The POPCs are represented as grey sticks, the phosphorus molecules are represented as orange spheres, and the different Monomers are represented as cartoon, in four different colors, one per Monomer.

These small membrane-spanning proteins assemble as homotetramers. Each of the four Monomers weights approximately 30kDa, and possesses six transmembrane alpha-helical domains - six helices (I-VI) connected by five loops (loops A-E), with both carboxylic and amino terminals on the cytoplasmic side (Figure 1.2) [27]. Each Monomer spans in the plasma membrane with a barrel-like structure, divided into an extracellular and a cytoplasmic vestibule, connected by a central amphipathic pore region (Figure 1.3 and 1.5) [5,28]. In contrast to most ion channels, the pore region does not reside at the four subunits' symmetry axis; instead, each of the four existing Monomers has its own permeable pathway, containing a separate pore (Figure 1.3 and 1.5) [6].

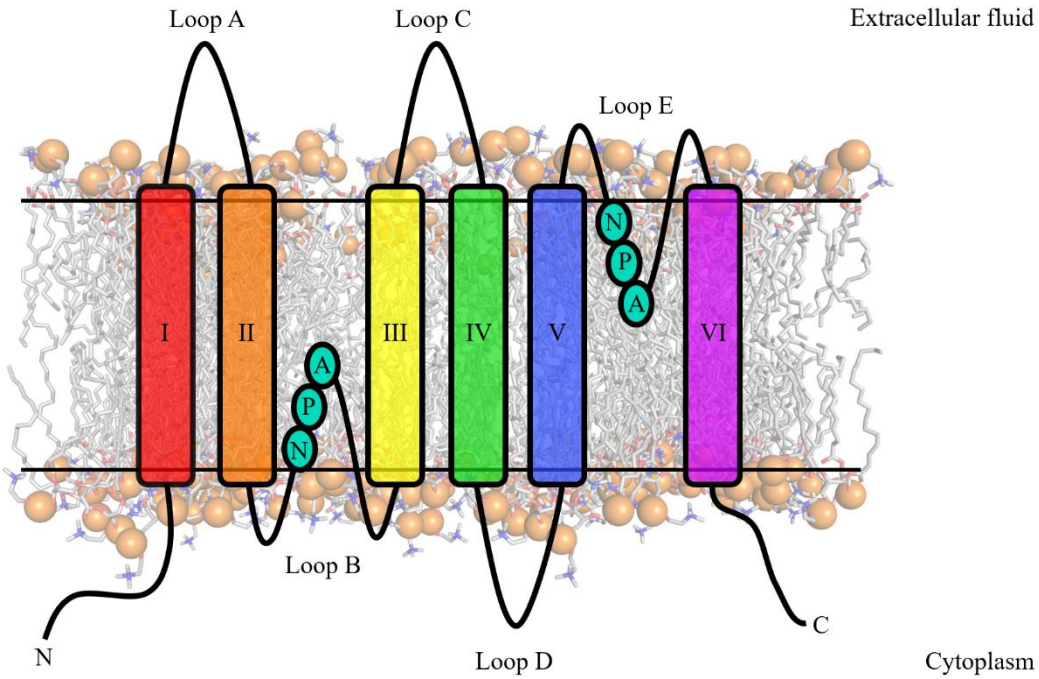

Figure 1.2. Topology of an AQP protein embedded in a POPC membrane bilayer, highlighting the six transmembrane alpha-helical domains (helices I-IV and loops A-E) and the conserved NPA motifs in loops B and E.

It has been reported that AQPs are structurally related and have highly similar consensus regions, particularly in the pore-forming domains [27]. In this way, an analogous transport mechanism can be presumed. The hydrophobic domain is created by loops B and E (Figure 1.2) and is involved in substrate size restriction and specificity. The pathway through the domain that constitutes the central pore is coated with conserved hydrophobic residues, which originates a single hydrogen-bonded chain of water molecules, granting a rapid transport of these particles [27,29].

Two pore regions are critical for AQPs function: the aromatic-arginine region, or ar/R selectivity filter, and the asparagine-proline-alanine (Asn-Pro-Ala) region, or NPA motif [7]. The first one is located close to the extracellular side of the pore, displaying a highly conserved Arg residue, in loop E (Figure 1.2), in an aromatic neighboring [25,28,30], formed by the following amino acids, in AQP1: Phe 56; His 180; and Arg 195 (Figure 1.4). The ar/R region diameter defines selectivity by size exclusion, allowing the aquaporin to let through or block, in a selective way, the passage of different molecules [28], acting as a selectivity filter by blocking particles larger than water [7].

The NPA motif contains two highly conserved hydrophobic regions (Figure 1.4), with the following amino acids, in AQP1: Asn 192; Pro 193; and Ala 194 on one side; and Asn 76; Pro 77; and Ala 78 on the other. This motif is located in the center of the channel, where the two half helices meet, on the two hydrophobic loops, one from the cytoplasmic region (loop B), and the other from the extracellular medium (loop E) (Figure 1.2), that contain inward-facing Asn polar side chains, with the Asn residues residing at the positive helix ends [25]. In this way, the helices' positive partial dipole moments are focused on NPA's Asn residues in the pore center [26]. Together with the backbone alpha-carbonyl groups, the NPA motifs act as hydrogen-bond donors and acceptors that regulate water transport through the AQP1 pore [5,29,31]. The NPA motif surrounds the central pore-like region that contains additional protein density, creating a distinctive hourglass shape, making the water channel narrow in the middle and wider at each end [32].

Due to its positive charge, both the ar/R region and the NPA motif are involved in proton repulsion [25,33,34]. The ar/R selectivity filter forms a joint filter, together with the electrostatic field originated from the positive ends of the half helix dipoles in the NPA region, against the passage of inorganic cations [5,25,35–37]. In this way, both the ar/R selectivity filter and the NPA motif participate in proton exclusion since these residues' positive charges help to repel protons from entering the AQP's pore.

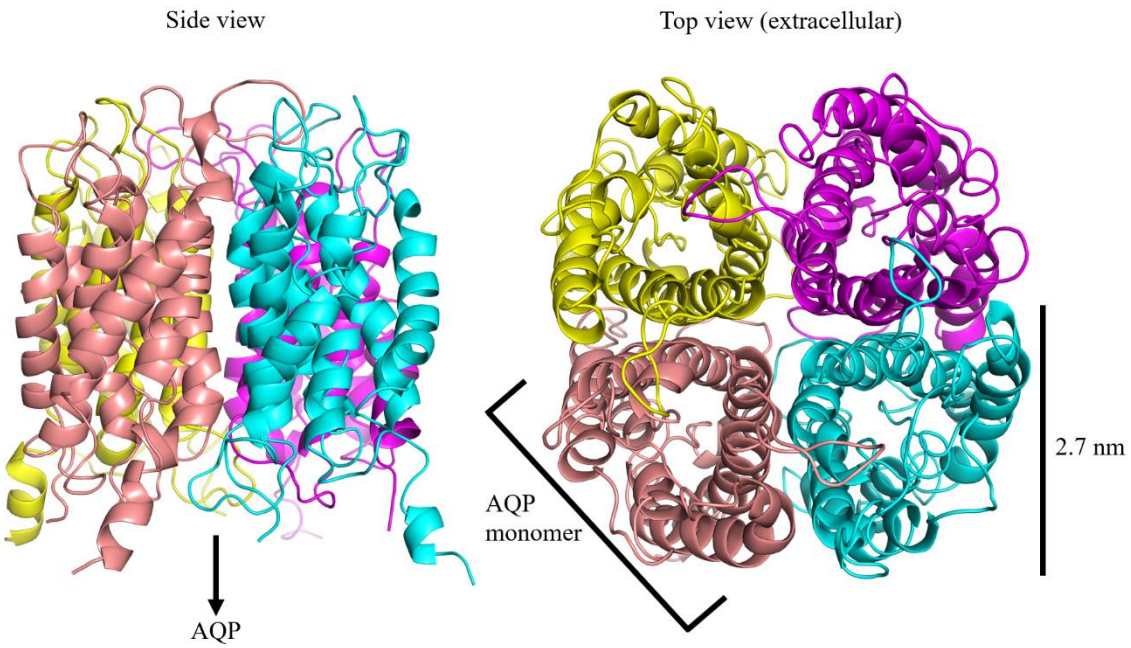

Figure 1.3. Crystallographic structure of AQP1, represented as cartoon, with the four Monomers in four different colors, one per Monomer. On the left: side view highlighting the barrel-like structure of each Monomer. On the right: top view highlighting the four pores, one in each Monomer.

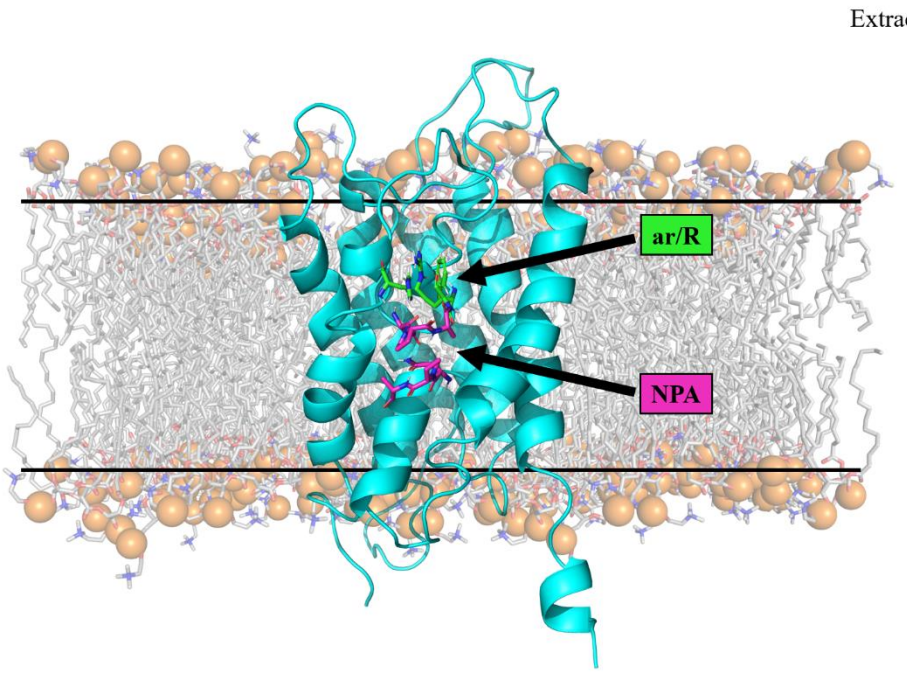

Figure 1.4. Monomer A of the Crystallographic structure of AQP1, represented as cartoon, highlighting the ar/R selectivity filter (in green) and the NPA motif (in pink).

Genotype and phenotype studies have established AQPs' role in obesity, brain swelling, glaucoma, epilepsy, refractory edema, cancer, neuroinflammation, and pain [3,38]. In cell migration, AQPs' presence has been associated with local invasion, tumor angiogenesis, and metastasis [38–40]. AQPs have also been proven to be overexpressed in different tumors and correlated with malignancy and cancer progression [2,41]. Studies have shown that these proteins are closely associated with cancer's biological functions, being highly expressed in more than twenty human cancer cell types, suggesting that AQPs play a crucial role in tumor growth [3,4]. Therefore, these proteins can work as potential diagnostic and therapeutic targets in anticancer treatment since their modulation might limit tumor growth and spread [2,4].

However, several disadvantages could be identified, making AQPs problematic drug targets. To this day, specific AQPs' inhibitors suitable for clinical use have not yet been identified, probably due to the small size and pore diameter of the functional AQP Monomer, preventing small molecules' access [42,43]. Additionally, AQPs are straightforward passive pores, meaning they lack sophisticated gating and transport mechanisms that can be targeted by small molecules [5]. Regulation of AQPs is allowed by short-term regulation, also known as gating, which is often achieved by directly affecting the protein's channel conformation [44]. Various conditions such as phosphorylation, pH, pressure, temperature, and solute gradients play an essential role in the gating process of AQPs [25,44,45]. However, there is still a considerable amount of controversial information due to the limited knowledge about these proteins. Most recent available data [5] also suggests that an inhibitor's binding should occur deep in the narrow AQP pore to physically prevent water conduction since mutations in the extracellular and cytoplasmic domains of AQPs generally have little effect on water permeability, making the AQPs even less druggable.

On the other side, there are considerable advantages to use AQPs as future drug targets. From studies using knockout mice [46], there is compelling evidence that AQPs may be potential drug targets, not only for new diuretic reagents but also for novel therapies for brain edema, inflammatory disease,

glaucoma, obesity, and cancer [40,47]. Multiple phenotype abnormalities were found in the null mice, such as: increased protection for cytotoxic brain edema, induced by water and cerebral ischemic injury, when using AQP4 knockout mice [48]; when using AQP7 knockout mice, these displayed markedly greater fat mass, when compared to wild-type mice, revealing adult-onset obesity [49–51]; in AQP1-deficient mice, significantly slower tumor growth and improved survival was achieved, when compared to wild-type mice, in a tumor-bearing model with melanoma and carcinoma cells implanted subcutaneously [52,53]. Additionally, some studies also indicate that the pharmacological modulation of water and solute transport targeting AQPs may provide novel opportunities for therapeutic interventions in a wide range of human disorders [47].

Current knowledge of the pore structure-function relationships should allow the development of "designer drugs" that can specifically target AQPs [2]. Furthermore, the continuous growth of knowledge concerning the structure-function pair, particularly the geometry, on the atomic level, of specific hydrophobic and hydrophilic residues in the pore region, makes AQPs an encouraging therapeutic target to be used in future drug discovery campaigns [5].

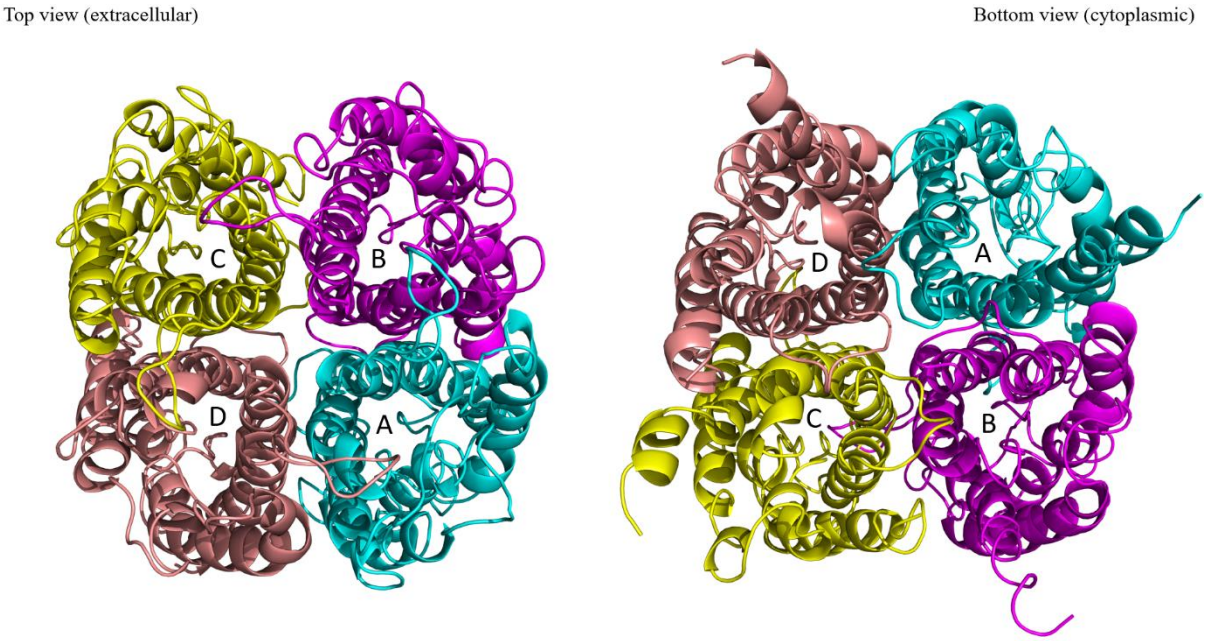

Figure 1.5. Top (left) and bottom (right) view of the Crystallographic structure of AQP1, represented as cartoon, highlighting the four pores and each of the four Monomers, colored in four different colors, one per Monomer (Monomer A-Cyan; Monomer B-Magenta; Monomer C-Yellow; Monomer D-Salmon).

1.2. Computational Approach

Drug discovery campaigns, including novel drug discovery and development, are recognized to be time-consuming, risky, and costly [9]. A new medicine will take a median of 10 to 15 years, and more than US$1 billion, from the birth of an original idea to the launch of a finished product to a pharmacy shelf [14]. The initial research, often developed in academia, generates data that enables developing a hypothesis, which declares that a protein's inhibition, or activation, will lead to a therapeutic effect in a specific disease state [9]. In the last years, with the constant advances in technology, drug discovery has moved towards more rational and advanced strategies, supported by our increasing understanding of the fundamental principles of protein-ligand interactions. Ligand and Structure-based drug design approaches summon the most compelling concepts in modern chemistry and biology, linking medicinal chemistry with structural biology [10].

The Structure-based concept relies on the perception of the biological target's three-dimensional structure, obtained, for example, through x-ray crystallography methods [12]. This direct drug design is analogous to high-throughput screening (HTS), where knowledge of both target and ligand structure is imperative [13]. This brute force concept relies on an automatic approach to screen several molecules while searching for those that evoke the desired biological response. This method has two significant advantages: it requires minimal compound design or prior knowledge, and technologies required to do this type of screening are becoming more efficient day by day [13]. Structure-based approaches have gained more and more attention from the scientific community, especially after the 2013 Nobel Prize in chemistry awarded to Martin Karplus, Michael Levitt, and Arieh Warshel [54].

Ligand-based approaches are powerful methods based on the information of known active and inactive compounds [10]. This indirect drug design method relies on understanding the molecules' binding to a target of interest. These molecules can then be used to derive a pharmacophore model that determines the minimal indispensable structural characteristics that another molecule should retain to be able to bind to the designated target [11]. It is worth noting that these knowledge-driven approaches are often explored together, creating new opportunities for innovation in medicinal chemistry so as to maximize the accomplishment of drug discovery projects [10].

However, although traditional high-throughput screening often generates multiple hit compounds, some of which are, in theory, qualified to be modified into a lead and later a novel therapeutic, the overall hit rate is often meager, limiting the usage of HTS in research programs capable of screening large compound libraries [13]. One of the main reasons why drugs fail in trials is because the desired target's intended effect is not produced [14]. In this way, appropriate target identification is a crucial step in developing new drugs. A good target needs to meet several criteria: be efficacious, safe, meet clinical and commercial needs, and, above all, be 'druggable'. A 'druggable' target is accessible to the presumed drug molecule and, upon binding, evokes a desired biological response that is consistent both in vitro and in vivo [9]. Following the target validation process, compound screening assays are developed during the hit identification phase [9]. A compound can be described as a 'hit' if it shows the desired activity in a compound screen and if it's activity can be confirmed upon retesting. A way to identify hit molecules is by applying HTS, which involves screening entire compound libraries directly against the desired drug target [55], and where both Structure and Ligand-based approaches can be applied.

In this way, this field's significant challenges include identifying promising hits and developing high-quality leads for further development into clinical candidates [10]. To achieve successful results, new and innovative approaches and methods must also be developed.

Chapter 2

2. Aim of this work

In this thesis, a combined Ligand and Structure-based computational drug discovery approach was pursued to identify new and innovative AQP1 modulators that in the future could be used as new therapeutics for the treatment of brain edema, inflammatory disease, glaucoma, obesity, and cancer.

In the Ligand-based approach, the identification of new AQP1 modulators was performed based on the most important chemical characteristics described from previously reported AQP modulators. These characteristics were used in previous virtual screening campaigns to search for new potential therapeutic modulators with some limited efficiency. To strengthen our screening, we have applied this same approach to multiple commercially available Databases to guarantee the construction of a global compound Database with the broadest possible chemical diversity that could be used in the subsequent Structure-based screening approach. To further optimize and curate our compound Database, we calculated and identified the most important chemical descriptors for a drug discovery campaign and used them to reduce the number of potential hits.

Later on, the previously prepared compound Database was used in a Structure-based approach. Using a combined Molecular Dynamics and Molecular Docking approach, it was possible to identify new possible alternative AQP1 modulators. In this approach, the focus was on the pore region of the protein, which is, at the moment, the most important and characterized druggable region. However, this approach additionally allowed the identification of new regions liable to influence the function of AQP1. We tried to identify and characterize these new target regions on the protein surface and used them in a new Structure-based virtual screening campaign, similar to what was done for the pore region.

After an energy-based binding efficiency ranking and structural characterization obtained from the Structure-based screening, a short list of the most promising AQP modulators was obtained. This list was then used in new binding affinity assays using a combined approach of Molecular Dynamics simulations and MM/PBSA calculations. This approach allowed a more accurate scoring and ranking of the most efficient compounds binding to AQP1, which granted the creation of a final short list of compounds whose effect on this protein function will be experimentally evaluated.

Chapter 3

3. Theory

3.1. Molecular Modeling: Description and Relevance

Molecular Modeling is the term applied to various theoretical models that provide simplified descriptions of molecular systems, mimicking some aspects of the molecules' behavior [56]. More straightforwardly, it encompasses all methods, theoretical and computational, used to try to simulate particles based on several physical and mathematical models. The more elemental calculations can be performed by hand, in a very feasible way, by using mechanical models. However, nowadays, Molecular Modeling is undoubtedly associated with computer modeling since computers are mandatory to perform any reasonably sized simulations [57]. This fact implies that computers have, without any doubt, allowed an extension of the existent models to the study of a more considerable number of systems. These methods are applied in computational chemistry, drug design, and computational biology to study molecular systems, which can be as simple as small chemical structures, or as complex as large biological molecules. Computational Molecular Modeling has, in this way, been gaining more and more significance in the studies of biology, biochemistry, and biophysics since it can complement experimental work.

Empirical methods are generally the fastest, cheapest, and most straightforward option when studying biological systems. However, some exceptions are systems that do not benefit from such an approach due to their inherent complexity, methodological limitations, time scales involved, or equipment costs [58]. Along these lines, Molecular Modeling relevance is supported by the fact that nowadays, only a minimal number of methodologies addressing molecular systems are experimentally accessible. Computational methods, such as Molecular Dynamics (MD), simulate the biomolecular process and determine properties that are estimated as too short or too small to be reachable through empirical evaluation [59]. These simulations are achieved through statistical mechanics and the classical Newton laws of motion [59–61]. Even though it is a computational technique, in order to be reliable and accurately described, the system requires experimental data to be based on. The inadequacy of empirical results and inaccurate data will make it impossible to describe the "rules" that the system must abide by [59].

In Molecular Mechanics, these previously referred "rules" are known as the Force Field. A biomolecular Force Field is made of potential energy terms that describe bonded (bond-stretching, bond-angle bending, improper and proper dihedrals) and nonbonded (Van der Waals and electrostatic) interactions between atoms [62]. Molecular Modeling methods are a useful and powerful tool to study molecular systems and should be seen as a way of complementing empirical methods [58]. Nonetheless, experimental measurement should always be favored instead of a prediction by simulation, unless very difficult to obtain. Even though more powerful machines are further developed every year, allowing the simulations to achieve the millisecond's timescale [63], unfortunately, computing power is still somewhat limiting for the simulations. Despite this fact, if we have energy, movement, and time, through the classical laws of motion, a trajectory of a system of any size and in any time scale, if not too computationally demanding, can be theoretically simulated, for example, with an MD approach [59,60].

Molecular Mechanics

Quantum Mechanics (QM) is one of the most used methods in Molecular Modeling. This method meticulously treats the molecular system by providing an accurate and precise description of several phenomena that classical theories cannot explain at the atomic and subatomic levels [64]. However, many of the problems one would like to address have too many particles. Therefore, as more detail exists in the system of interest or the higher number of degrees of freedom used to describe the system, more computing power is needed to accomplish this task [65,66]. Hence, these studies are limited to smaller systems, and larger ones require fewer degrees of freedom [59]. Due to these limitations, some methods have been developed, using simplifications, to treat molecules' conformational behavior.

MD is one of such nonquantum methods based on an empirical Force Field that uses Newtonian, or classical mechanics, to describe the models' physical evidence. This consists of seeing a molecule as a gathering of particles, held together by the elastic forces defined in terms of a Potential Energy Function (PEF) of internal coordinates, such as bond lengths, angles, and torsion angles [56,67]. In this way, Molecular Mechanics does not consider the electronic motions. The system is described as a set of classical nuclei, where electrons are implicitly treated (atoms are treated as the smallest entity), diverging from the more detailed and complex QM methods, where the electrons are explicitly treated (modelling protons, neutrons, and electrons). One can remove the electronic motions and adopt the nuclear positions to calculate the system's energy using the Born-Oppenheimer approximation. With this adjustment, the system is described by a Molecular Mechanics Force Field using a PEF, with parameters from an empirical Force Field [56,59].

Potential Energy Function

A description of the system using a PEF is practiced in a Molecular Mechanics procedure [68]. This function computes the molecular potential energy considering four types of bonded interactions (bond length stretch, bending angles, and torsions of proper and improper dihedral) and two nonbonded interactions (van der Waals forces and electrostatic) [69]. PEF describes the atoms and their interactions defined by their nuclei positions while ignoring electron movements [68].

The potential energy is dependent on the atom coordinates (r^N), meaning it is a function of the positions r of N particles, usually atoms:

$$V(r^N) = V(r^N)_{bonded} + V(r^N)_{non-bonded} \tag{3.1}$$

$$V(r^N)_{bonded} = V(r^N)_{bonds} + V(r^N)_{angles} + V(r^N)_{dihedral} \tag{3.2}$$

$$V(r^N)_{non-bonded} = V(r^N)_{vdW} + V(r^N)_{elect} \tag{3.3}$$

$$V(r^N) = V(r^N)_{bonds} + V(r^N)_{angles} + V(r^N)_{proper} + V(r^N)_{improper} + V(r^N)_{vdW} + V(r^N)_{elect} \tag{3.4}$$

The most straightforward way to compute the movements between an atom pair, regarding bond stretching interaction with bond b, is expressed by a harmonic potential based on Hooke's Law [56]:

$$V(l) = \frac{k_b}{2}(b - b_0)^2 \tag{3.5}$$

where k_b is the force constant between an atom pair, b is the distance between them, and b_0 is their optimal distance, or the bond length with the minimum energy.

Similarly, angle bending interaction (three-bodies) also relies on a harmonic potential:

$$V(\theta) = \frac{k_b}{2}(\theta - \theta_0)^2 \tag{3.6}$$

where k_b is the force constant associated with angle bend, θ is the angle between three atoms, and θ_0 is their optimal angle. These molecular movements (stretching/compressing and bending) require significant amounts of energy compared to the torsional or nonbonded terms. Due to this fact, these two terms, and how they interact with each other, have a massive impact on the structure [59].

There are two types of torsional angles (four-body): proper and improper dihedrals. The proper dihedral is computed through a torsional angle formed by four sequential atoms, while four non-sequential atoms define improper torsional angles. The proper torsional potential is expressed as:

$$V(\omega) = \sum_{n=0}^{N} k_\varphi[1 + cos(n\varphi - \delta)] \tag{3.7}$$

where φ is the value of the proper dihedral angle, n is the multiplicity which stands for the number of minima, and maxima, of energy in one complete rotation, and δ is the reference maximum, and it can be 0 or π.

The improper torsional potential can be expressed as:

$$V(\omega) = \frac{k_\xi}{2}(\xi - \xi_0)^2 \tag{3.8}$$

where k represents the force constant associated with angle torsion, ξ is the torsion angle formed between the atoms' pair that are 3 bonds apart, and ξ_0 is their optimal angle.

The nonbonded interactions do not rely on any specific bond interactions, though they are usually dependent on the distance's inverse power [59]. The nonbonded terms can be separated into two sub-components: the van der Waals and electrostatic interactions. Van der Waals interactions consider two opposite forces between two interacting atoms, and this component can be described as a Lennard-Jones interaction:

$$V(r) = \sum_{i=1}^{N}\sum_{j>1}^{N} \frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}} \tag{3.9}$$

where r_{ij} is the distance between two atoms (i and j), and C_{ij}^{12} and C_{ij}^{6} are interaction parameters: C_{ij}^{12} is related to the repulsion between the two atoms, while C_{ij}^{6} is related to their attraction (Figure 3.1).

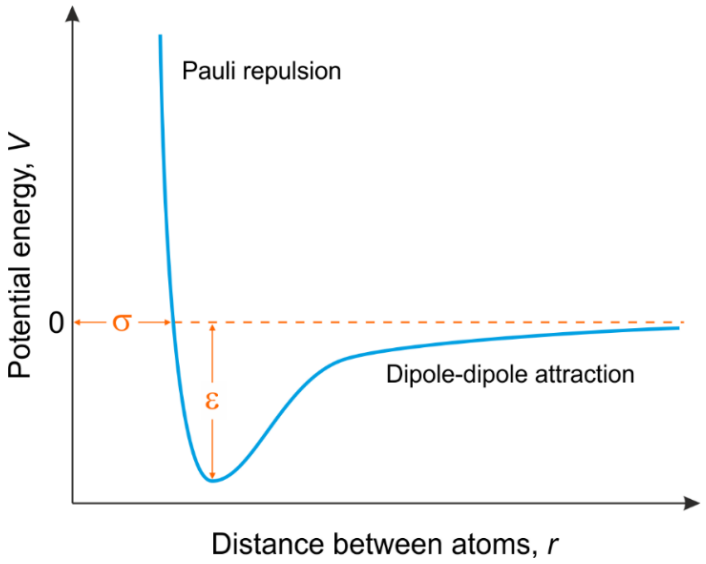

Figure 3.1. Representation of the Lennard-Jones potential, C^6 and C^{12}. ε is the well's depth and a measure of how strongly the two particles attract each other, and σ is the distance at which the intermolecular potential energy between the two particles is zero. Adapted from [70].

When the distance between the two atoms starts to increase, the energy diminishes, passing through the collision distance σ (when the potential interaction energy equals zero), until it reaches a minimum ε (Figure 3.1). Usually, in simulations, a cut-off radius is used so that atom pairs separated by a greater distance than the user-defined cut-off will have a null van der Waals interaction energy. Thus, we are saving computational time without compromising the results of the simulation.

An electrostatic interaction can reside between two different molecules or between two regions of the same molecule if both possess partial or net atomic charges. These charge distributions exist due to the electronegativity's variation of the atoms in their constitution [59]. An electrostatic interaction can be calculated as a sum of pairwise interactions of point charges, with Coulomb's law:

$$V(r) = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \tag{3.10}$$

where qi and qj are the partial atomic charges for atom i and j, respectively, N is the total number of atoms in the system, ε_0 is the permittivity in vacuum, ε_r is the relative dielectric constant of the medium, and rij is the distance between the atom pairs.

All the previously referred terms can now be added to form the potential energy function of a system:

$$
\begin{aligned}
V(r^N) = & \sum_{bonds} \frac{k_b}{2}(b - b_0)^2 \\
& + \sum_{angles} \frac{k_b}{2}(\theta - \theta_0)^2 \\
& + \sum_{proper} k_\varphi[1 + cos(n\varphi - \delta)] \\
& + \sum_{improper} \frac{k_\xi}{2}(\xi - \xi_0)^2 \\
& + \sum_{i=1}^{N_A}\sum_{j=1}^{N_B}\left[\frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} + \left(\sum_{i=1}^{N}\sum_{j>1}^{N}\frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}}\right)\right]
\end{aligned}
\tag{3.11}
$$

Force Field

A computational method used to estimate the forces between atoms, within and between molecules is designated as Force Field, which refers to the practical form and parameters used to determine the potential energy of a system of atoms, in a Molecular Mechanics approach [71].

A Force Field encompasses the components and parameters used to describe a molecular system and will depend on the four main features, previously referred to: bonds, angles, torsions, and nonbonded interactions (Figure 3.2). Each of these pieces will supply an energy term to the potential energy function, allowing an interpretation of the entire system [59]. In this way, a force field is built up from two distinct components: the equations, or potential functions, and corresponding derivatives, the forces; and the used parameters.

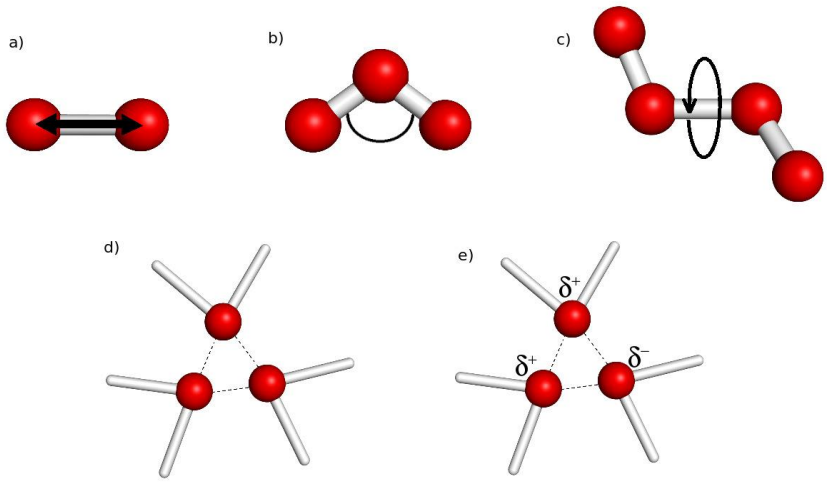

Figure 3.2. Representation of the five contributions to a Molecular Mechanics Force Field: (a) bond stretching, (b) angle bending, (c) dihedral angle interactions, (d) nonbonded van der Waals forces, and (e) nonbonded electrostatic interactions. Adapted from [58].

The parameterization is accomplished over an agreement with experimental values and theoretical calculations. Force fields are empirical, which means they intend to reproduce a given collection of experimental data. Therefore, different Force Fields must be used accordingly to their defined purpose and specified parametrization [56,72]. In this way, for different types of atomic models, there are specific Force Fields.

An example of a Force Field is GROMOS [62,73–77], which is a performance and biomolecular focused Force Field, parameterized to reproduce thermodynamic properties of small polar molecules and solvation energies of amino acids [62]. GROMOS is a united-atom force field, meaning that all nonpolar hydrogens are collapsed into the nearest atom, which becomes slightly bigger and reflects their charges.

Binding pocket predictions

Rating the attractiveness of a drug target is one of the foremost challenges within the early stages of the development of "designer drugs". This challenge arises due to the massive amount of available structural data, and thus the automated collection of target information gains importance. In this way, protein's active sites' prediction is crucial for large-scale protein function prediction, classification, and druggability estimates [78].

DogSiteScorer

DoGSiteScorer [79,80] provides all the necessary tools to detect potential binding pockets of a protein of interest [78]. The first step is predicting potential pockets on the protein's surface, exclusively supported by the protein's heavy atom coordinates. This is achieved by using a grid that covers the whole protein, where points are labeled as free or occupied based on their overlap with protein atoms. Afterwards, a filter based on a difference of Gaussian (DoG) is applied to the grid. This filter identifies areas on the protein's surface where the positioning of a sphere-like object is encouraging, based on the grid points' labeling. Lastly, based on a density threshold, these areas are clustered to potential subpockets, and the subpocktes are then merged into pockets. This method differs from existing ones since DoGSiteScorer splits predicted pockets into subpockets, conceding a refined description of the active sites' topology [78–80].

AutoLigand

AutoLigand [81], which is included within the AutoDockTool [82] package, was developed to identify ligand binding sites within receptor proteins. AutoLigand uses a compelling approach to browse promptly for high-affinity binding pockets and then reports their optimal characteristics, such as volume, shape, and best atom types [83]. Using AutoGrid, a program that pre-calculates grid maps of interaction energies for diverse atom types with a macromolecule, the user conceives affinity maps. These maps are created by browsing each atom type, carbon, oxygen, or hydrogen, through the grid and recording its affinity at that distinct point. Formerly, a simple "flood-fill" of the user's input volume is performed, starting at each point of the grid, and saving the ten best non-overlapping fills. Next, AutoLigand lets the fill volumes to a shape and location with the best binding energy and lowest volume. There are many possible binding sites with small volumes, but with larger fill volumes the results end up collapsing to a single best fill region, which usually is characterized as an active site [81].

SeeSAR

SeeSAR [84] is a software tool for interactive compound evolution and visual compound prioritization. To maximize the likelihood of success, Structure-based design work supports optimization of multi-parameter, having the relevant parameters at hand in consolidation with real-time visual computer assistance, in a three-dimensional structure, instead of affinity alone [85]. SeeSaR's "Binding Site mode" computes proposals for accessible empty pockets, allowing us allows us to identify binding pockets and also how many and which amino acids form those individual pockets. This functionality is based on a heuristic model that applies a DoG on a 3D grid to assess the likelihood of a specific area allowing the existence of a pocket shape. Hydrogen bonds, the lipophilic character, the solvent accessible surface (SAS), and the distances between pairs of functional group atoms of the potential pocket are also considered. As stated by the authors, "the pocket detection algorithms are part of the DoGSiteScorer that emerged from a collaboration with Merck and Hamburg University" [84].

Molecular Docking

Within the Molecular Mechanics framework, Molecular Docking is the most commonly used method to identify both the binding sites of a protein of interest and also the best pose of small compounds at its surface region. This is done with as many degrees of freedom as possible and with a realistic energy evaluation [86]. Due to these features, Molecular Docking is one of the most popular methods in Structure-based drug design since portraying the binding behaviour has a crucial role in rational drug design [87–89].

Molecular Docking is a bioinformatic modeling approach that comprises the interaction of two or more molecules to give a final stable pose [69]. Molecular Docking conceives numerous ligand poses that are then ranked and grouped using the software's scoring function [69]. In this way, Docking simulations can predict, based on the system's total energy, optimized docked conformers. Molecular Docking can be exemplified as a "lock-and-key" kind of problem. One wants to find the correct orientation of the key, or in this case, the ligand, regarding the lock, or the protein [90]. In this way, it is described as an optimization dilemma, representing the optimal orientation of a ligand that binds to a protein of interest. During the Docking process, the ligand accommodates its position, relative to the protein, to produce a "best-fit" conformation. Docking protocols are a combination of search algorithms, exploring all the degrees of freedom of the guest-host system, and scoring functions, allowing the system to be sampled efficiently, ensuring the detection of the most critical binding modes [87,91]. As a significant advantage, this method grants the analysis of a large area of the system's conformational space [92]. However, it has two major problems: the global search problem [93] and how to score the obtained solutions [94].

The search space, in theory, consists of all possible orientations and conformations of both the ligand and the protein [69]. However, it is impossible to explore the search space's plenitude in practice since this would involve enumerating all possible distortions of each molecule and all possible translational and rotational orientations of the ligand [95,96]. The first developed algorithms only considered translation and rotation and treated the ligand and receptor as rigid structures to overcome the previously referred problem. With the computational evolution throughout the years and with the aid of new algorithms, it is now possible to treat the ligand with full flexibility and the host as a rigid structure [97] or even more ambitious both structures as flexible [98]. However, in this last case, some problems arise.

In this thesis, since we are screening large Databases, we need to be extra careful in terms of computational demands and overall computational time. If we add flexibility to the protein's side chains,

the number of degrees of freedom that the Docking software has to calculate becomes enormous. Many degrees of freedom imply that we have to generate a lot of poses to better sample the conformational space, which is very time consuming. So, in our case, we decided to treat the protein as a rigid structure.

As a result, these programs generate a considerable number of promising ligand poses, which are then classified using a scoring function. The scoring function takes a specific pose and returns a number indicating the likelihood of that pose representing a favourable binding interaction [99]. However, we have to reach a compromise between using a scoring function that is not too computationally demanding and, at the same time, that assures the desired differentiation between solutions. Nevertheless, we should not forget that the final result is still a rigid complex. In this thesis's work, we have used MD simulations to generate the most populated conformations of the protein's conformational space. Like this, we tried to counterbalance the question of conformational variability or the flexibility problem.

Most scoring functions are physics-based Molecular Mechanics Force Fields that estimate the pose's energy within the binding site:

$$\Delta G_{bind=}\Delta G_{solvent} + \Delta G_{conf} + \Delta G_{int} + \Delta G_{rot} + \Delta G_{t/t} + \Delta G_{vib} \tag{3.12}$$

where different factors consist of solvent effects, conformational changes, free energy due to protein-ligand interactions, internal rotations, association energy of ligand and receptor to create one complex, and free energy owing to modifications in vibrational modes, respectively [100]. A low (negative) energy value indicates a stable system, and therefore a possible binding interaction. Docking usually uses a scoring function supported by approximating the system's standard chemical potentials, with Coulomb energies and 6-12 van der Waals interactions. Two software's are mostly used to achieve the previously referred scoring functions: AutoDock 4 (ADT4) [82]; and AutoDock Vina (VINA) [101]. ADT4 requires the pre-calculation of grids for the different atom-types found in the target protein while, on the other hand, VINA does not require choosing atom types and pre-calculating grid maps [101]. Instead, it calculates the grids internally and instantly and does not store them on the disk, making this software significantly faster than ADT4. Furthermore, the search algorithm implemented in VINA is faster when compared to ADT4. As previously referred, we need to be careful in terms of computational time, and for this reason, VINA was the major Molecular Docking software used in this thesis. VINA incorporates a general form for the conformation dependency within the scoring function:

$$c = \sum_{i<j} f_{t_i t_j}(r_{ij}) \tag{3.13}$$

where each atom i is attributed to a type t_i and a symmetric group of interaction functions $f_{t_i t_j}$, considering an interatomic distance, r_{ij}.

The intermolecular and intramolecular contributions can obtain this value as follows:

$$c = c_{inter} + c_{intra} \tag{3.14}$$

The scoring function of VINA was inspired by the X-score function [102]. As stated by the authors, the scoring function's nature is "more of a machine learning than a physics-based function" [101], which differentiates VINA from ADT4.

The interaction functions $f_{t_i t_j}$ are defined considering the surface distance, d_{ij}:

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j} \tag{3.15}$$

where R_t represents the van der Waals radius of atom type t.

$$f_{t_i t_j}(r_{ij}) \equiv h_{t_i t_j}(d_{ij}) \tag{3.16}$$

In this scoring function, $h_{t_i t_j}$ is the weighted sum of steric interactions, hydrophobic interactions between the hydrophobic atoms and hydrogen bonds, when applied [101].

Molecular Dynamics

Once the model and the Force Field have been chosen, a method to search the configuration space must be selected [60]. One of the most used methods to perform this search is MD.

MD simulations study particles' movements and interactions, using a molecular Force Field, through the integration of their coordinates with a descriptor of the potential energy [59]. In this way, MD analyses the physical trends of atoms and molecules. The atoms and molecules interact for a user-fixed duration, giving an overall picture of the system's evolution. Newton's motion equations are used to obtain a sequence of positions and velocities of the particles throughout the simulation time, also known as a trajectory.

We start with the acceleration of each particle at instant t by using Newton's second law of motion:

$$\frac{d\vec{r_i}(t)}{dt^2} = \frac{\vec{F_i}(t)}{m_i} \tag{3.17}$$

where $\vec{r_i}$ and m_i are the coordinates and mass of atom i, respectively, at instant t, and $\vec{F_i}$ is the force acting on that specific atom.

The acceleration vector can then be integrated to obtain velocities and positions at instant $t + \Delta t$. Since this problem does not have an analytic solution, the equations need to be solved using a numerical method [68]. The Verlet algorithm [103] allows this type of calculation. The Leapfrog algorithm is one of its variants and was used in this work. In this method, the position at time t and the velocities at time $t + \frac{\Delta t}{2}$ are used to calculate new positions and velocities as follows:

$$\vec{r_i}(t + \Delta t) = \vec{r_i}(t) + \vec{v_i}\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{3.18}$$

$$\vec{v_i}\left(t + \frac{\Delta t}{2}\right) = \vec{v_i}\left(t - \frac{\Delta t}{2}\right) + \frac{\vec{F_i}(t)\Delta t}{m_i} \tag{3.19}$$

where Δt is the already mentioned time step. If this value is too small, the computational time will increase if it is too large, the system's dynamic behaviour will be wrongly described. The number of integration steps (a step between t and $t + \Delta t$) is a user-defined parameter.

One way to increase the time step is by using algorithms that are able to constrain some degrees of freedom of the system that are not coupled with molecular motions of interest. This means that the bond distances are forced to maintain their original values [68]. For example, we can use LINCS [104] to constrain bond lengths to predetermined reference values. Another example of this kind of algorithm is SETTLE [105], which is applied to the simulation of water molecules, for example, the simple point charge, SPC [106], that we have used in the MD simulations in our work.

Having this in mind, the general MD algorithm consists of the following [58] (Figure 3.3):

1. In the first step, the initial conditions of the system are inputted: the potential interaction function V as a function of atom positions; the positions r of all atoms; and the velocities v of all particles;
2. Next, the forces are computed. The force on any atom is calculated as a result of bonded interactions (bond stretching, angle bending, and proper and improper dihedral interaction terms) and nonbonded interactions (van der Waals forces and electrostatic);
3. The atoms movement is simulated by solving Newton's equations of motion, and therefore the configuration is updated;
4. Steps 2 and 3 are recurrent until an output step is requisite. As a result, the positions, velocities, energies, temperature, and pressure are written.

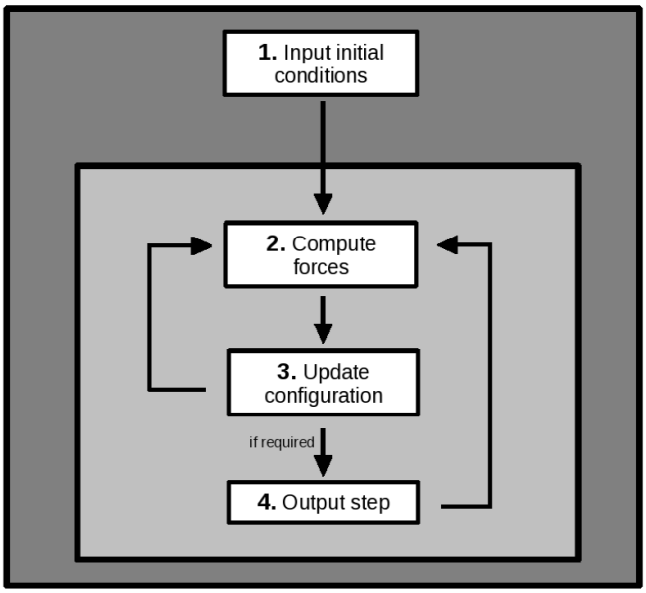

Figure 3.3. Schematic representation of the MD algorithm. Adapted from [107].

Automated force field Topology Builder

Force Fields, like GROMOS [62,73–77], provide a set of parameters for a core set of molecules such as amino acids, nucleotides, simple sugars, common lipids, and common solvents. These parameters have explicitly been parametrized to reproduce a given set of properties of this core set of molecules. Consequently, each new heteromolecule, such as a substrate, inhibitor, cofactor, or drug molecule, must be parametrized individually [108].

Several programs and Web servers exist to facilitate the generation of Force Field descriptions for novel molecules, such as the ligands we have used in this thesis, allowing, therefore, their use in MD simulations. The Automated force field Topology Builder (ATB) [108–110] is a Web server that provides topologies and parameters for a wide range of molecules to be applied in molecular simulations and computational drug design, compatible with the GROMOS family of force fields.

The ATB has three primary functions [108]:

1. To store the information of molecules that have already been parametrized;
2. To store pre-equilibrated systems to apply as starting configurations in MD simulations;
3. To develop descriptions of novel molecules compatible with the GROMOS family of force fields.

Results from a series of QM calculations are combined with a knowledge-based approach to ensure compatibility with the GROMOS family of force fields, designated as a multistep process [108].

Periodic Boundary Conditions

In most cases, the models used in a biomolecular simulation try to approximate the effect of an infinite amount of solvent molecules [68], as they intend to simulate an infinite dilution or an endless membrane [56]. However, to avoid large computational times, these models are restricted to a relatively small number of molecules. In this way, we can predict specific properties in a simulation where the system is "stuck" in a box of finite size.

Periodic Boundary Conditions (PBC) enable performing a simulation using a small number of particles that experience forces as if they were in a bulk fluid. A periodic system is characterized by the system of interest being simulated in the center [59]. When trajectories are simulated, a molecule may get closer to the box's borders and eventually span across them (Figure 3.4). When this happens, and an atom leaves the central compartment on one side, it enters it with an identical velocity at the opposite side, at the translated image position, allowing consistency in the number of atoms [60]. Furthermore, the box's size is relevant since an atom must not simultaneously interact with another one and its image. Consequently, the cut-off radius R_{cut} (Figure 3.4) must not exceed half of the shortest box vector.

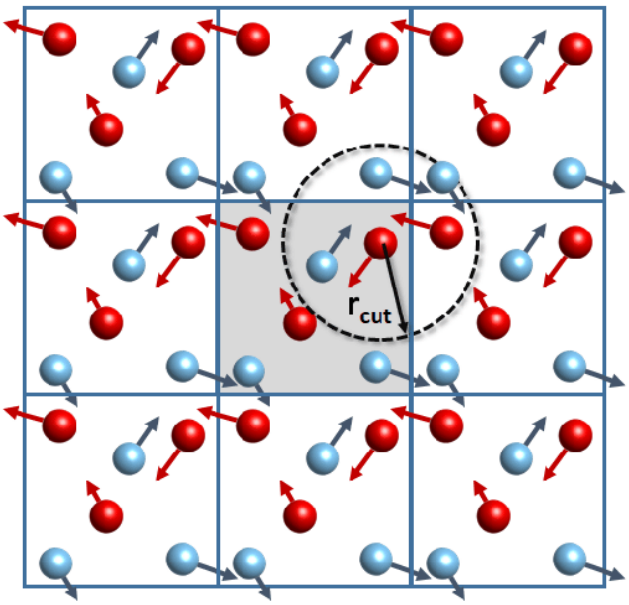

Figure 3.4. Representation of PBC in two dimensions. Eight identical images surround the central box. Also represented is the cut-off radius R_{cut}, highlighting that the atom does not interact with other atoms outside the dashed line. Adapted from [111].

Treatment of nonbonded interactions

As previously mentioned, with the aid of PBC, one can model an endless system. However, this creates a problem when dealing with long-range interactions, as we cannot calculate them all since the number of particles is virtually immeasurable [68].

The final term in the PEF, which covers both types of nonbonded interactions (van der Waals and electrostatic), runs over all atom pairs in the system. Its computation time is proportional to N^2, the square of the total number of atoms in the system, while the computation time of the other parts of the potential energy function is proportional to N [60]. In this way, computational efficiency can be improved by reducing the nonbonded interactions' summation. The most straightforward method to solve this problem is by applying a cut-off, which means that the interactions are truncated above a certain user-defined distance.

For short-range interactions, this cut-off might be applied to a single cut-off scheme together with the Verlet list update method. In this method, the user defines a cut-off value. The interactions of neighbours of each atom within the cut-off are calculated. For each particle, a Verlet list is constructed listing all other particles within the potential cut-off distance, which is updated at regular intervals. The restrictions are then placed in two ways: the use of a cut-off; and the frequency of updates [60].

Since the neighbouring search is one of the limiting steps in a Molecular Dynamics simulation, this saves a significant amount of computational time [68]. Lennard-Jones/van der Waals interactions are then solved with this approach since they decay to zero at relatively short interatomic distances [112] (Figure 3.1).

However, we cannot forget the electrostatic contributions beyond the cut-off. The treatment of long-range electrostatics mainly relies on two methods: Reaction Field (RF) [113]; and particle mesh Ewald (PME) [114]. A more in-depth description of these methods can be found in [115] or [72].

The RF approach is a method in which the region beyond the cut-off is treated as homogeneous. This homogeneous medium has a dielectric constant that attenuates all electrostatic interactions within the cut-off region [68].

PME [114,116,117] is a fast algorithm to calculate Ewald summations [116]. In this method, the system is positioned in a box, surrounded by an endless number of identical copies of itself. Although the problem does not cease to exist, it is moved to the infinite scale in an artificial periodicity. This can be applied to MD simulations since MD boxes are also periodic, when using PBC, and inhomogeneous. Nonetheless, there are also some PME inherent problems since it establishes an artificial periodicity and obligates a neutralization of the full system.

Given that both approaches have their limitations, and neither of these alternatives is theoretically unquestionable, one should be pragmatic when choosing a method to model lipid bilayers.

Temperature and Pressure control

When simulating biological systems, it is also required to control external components, such as temperature and pressure, in order to simulate the system accurately. Several methods were developed for the treatment of such parameters. All MD steps in this work were performed in the NPT ensemble, which means that the number of particles, pressure, and temperature were the three thermodynamic properties kept constant during the simulations [107].

To keep the thermodynamic temperature and pressure constants, we need to model an exterior bath that can give or extract kinetic energy from the system and adequately resize the simulation box. On a general note, the used algorithms adjust the temperature by tweaking the velocities applied in each atom and the pressure by altering the atoms' positions and therefore accommodating the system's volume [59].

Temperature

In this thesis, the velocity rescaling thermostat (v-rescale) [118], originated from the Berendsen thermostat [119], was used to keep the system under constant temperature. The temperature of a system is given by the average kinetic energy over time [107] by coupling it to an external bath, which either increments or removes energy through random velocity reassignment, from the Maxwell-Boltzmann distribution [56,59,107]:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}(T_{bath} - T(t)) \tag{3.20}$$

where τ_T is a coupling parameter whose magnitude determines how tightly the bath and the system are coupled together, T_{bath} is the desired temperature of the bath, and $T(t)$ is the temperature at any given time t.

Pressure

Analogously to the temperature, the Parrinello-Rahman method [119] keeps the system under constant pressure by coupling it to an external pressure bath.

The used method for pressure adjustment is similar to the one used for the temperature, with a rate of change of pressure given by:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_{bath} - P(t)) \tag{3.21}$$

where τ_T is the coupling constant, P_{bath} is the desired pressure of the bath, and $P(t)$ is the pressure at any given time t.

For both temperature and pressure, a high τ_T value means that the system will have more time to adjust. Hence, the system has a smaller temperature/pressure coupling, which means that it needs more time until the established value is achieved. A smaller τ_T translates into a higher-temperature/pressure coupling, and therefore the convergence to the desired value is faster. Nevertheless, systems with a low τ_T may not adjust fast enough, leading to undesired effects. Meanwhile, systems with a high τ_T will achieve the desired temperature/pressure slowly but more steadily [59].

Energy Minimization

When a starting configuration is at a great distance from equilibrium, the forces may be excessive, and the MD simulation may fail since the system can quickly diverge to nonphysical conformations [107]. A solution to this problem is to slightly change the system coordinates to bring it to an energy minimum, which guarantees its stability. The potential energy function has a deepest point, a global minimum, and a vast number of local minima (Figure 3.5). There are also saddle points between the local minima, which are the mountain passes through which the system can migrate from one local minimum to another [120] (Figure 3.5).

The steepest descent method [121] performs a downhill search on the phase space until it hits the closest energy minimum, which might not necessarily be the global energy minimum [59]. This iterative method is conceptually intuitive: a new point is calculated, being accepted if its energy is smaller than the current one. As the new positions get closer to a local minimum, displacement values will be lower [59] (Figure 3.6). The search will stop if a certain user-specified number of steps has been performed or if the convergence criteria are met.

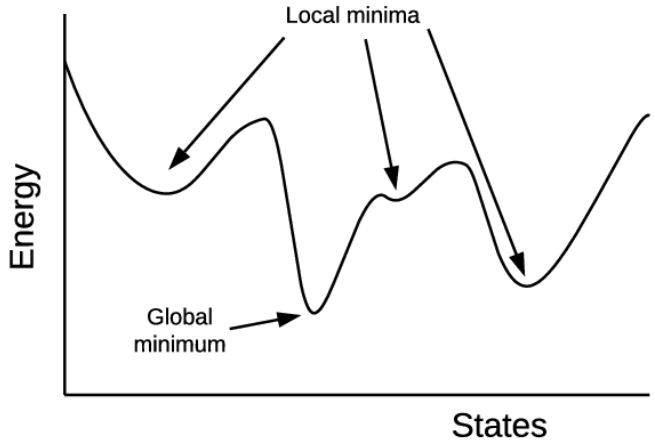

Figure 3.5. Representation of the local and global minimum in an energy minimization problem. Adapted from [122].

Other minimization methods complement the system's surface search after the initial use of the steepest descent technique. Although no method guarantees the global minimum determination in any reasonable amount of time, given a starting configuration, it is possible to find the nearest local minimum through energy minimization methods [120].

When the system is steady, the MD production may start.

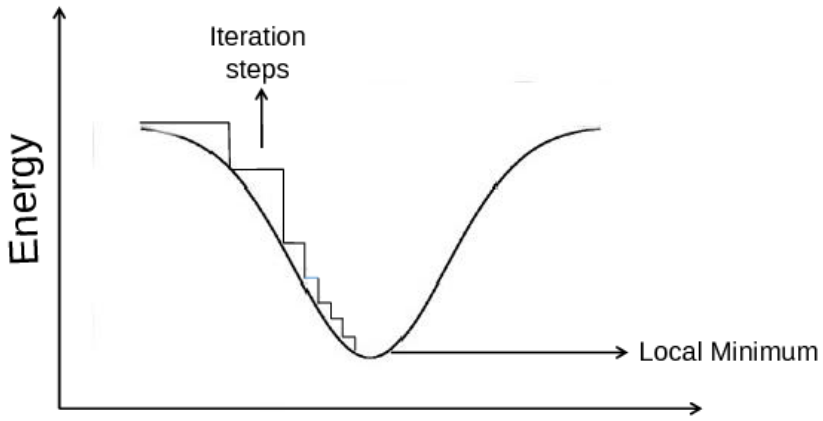

Figure 3.6. Graphical representation of the steepest descent method iterating over different energy values until it converges to a local minimum. Adapted from [59].

MM/PBSA

Free energy calculations are part of the repertory of computational biochemical research, encompassing the prediction of binding affinities of small compounds as drug-candidates and evaluating the relative stability of bio-macromolecular structures [123].

Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) combines an explicit Molecular Mechanical model for a solute with a continuum method for the solvation free energy [124]. The leverage of including Molecular Mechanical energies is substantial, allowing torsional energies and van der Waals interactions to be explicitly included [125], which distinguishes it from methods that use just continuum models to calculate $\Delta G_{binding}$ [126]. The MM/PBSA is one of the most commonly used methods to compute interaction energies. Combined with MD simulations, MM/PBSA can also incorporate conformational fluctuations and entropic contributions to the binding energy [123,127,128], and is a good compromise in accuracy and computational demand between empirical scoring and strict alchemical or "synthetic" perturbation methods [128,129]. Therefore, this approach can be used to rescore complexes obtained from Molecular Docking calculations to better distinguish active and inactive compounds.

In this thesis work, we have used the g_mmpbsa tool to perform such calculations. Although it is not an official tool from GROMACS, it has been widely used in determining binding free energy calculations from a set of compounds in the binding site of a protein target. In this approach, three energetic terms are combined to account for the change in binding free energies. The first one corresponds to a change in the potential energy in vacuum. It includes bonded terms such as bonds, angles, and torsion energies, and also nonbonded terms such as van der Waals and electrostatic interactions. The second term reports the desolvation of the different species. It is quantified by the sum of two energy

terms, the polar and nonpolar solvation energies using an implicit solvation model [130–133]. The third term accounts for the configurational entropy associated with the gas phase's complex formation [127].

In general terms, the binding free energy of the protein with the ligand can be expressed as [124,130]:

$$\Delta G_{binding} = G_{complex} - (G_{protein} + G_{ligand}) \tag{3.22}$$

where $G_{complex}$ is the total free energy of the protein-ligand complex and $G_{protein}$ and G_{ligand} are, respectively, total free energies of the isolated protein and the ligand in a solvent.

Furthermore, the free energy for each entity can be given by [134]:

$$G_x = \langle E_{MM} \rangle - TS + \langle G_{solvation} \rangle \tag{3.23}$$

where x is the protein or ligand, or protein-ligand complex, $\langle E_{MM} \rangle$ is the standard Molecular Mechanic's potential energy in vacuum, TS is the entropic contribution to the free energy in vacuum where T and S stand for temperature and entropy, respectively, and $\langle G_{solvation} \rangle$ is the free energy of solvation. The vacuum potential energy, E_{MM}, includes the energy of both bonded as well as nonbonded interactions, and it's calculated based on the Molecular Mechanics Force Field parameters.

Note that the conformation of the protein and the ligand in the bound and unbound forms is assumed to be identical in the single trajectory approach. Thus, ΔE_{bonded} is always zero.

The free energy of solvation is the energy required to transfer a solute from vacuum into a solvent and is calculated using an implicit solvent model [130,133,135]:

$$G_{solvation} = G_{polar} + G_{nonpolar} \tag{3.24}$$

G_{polar} and $G_{nonpolar}$ are, respectively, the electrostatic and nonelectrostatic contributions. The electrostatic term, G_{polar}, is estimated by solving the Poisson−Boltzmann equation, [135,136], while the nonelectrostatic term, $G_{nonpolar}$, includes the forces between solute and solvent (repulsive and attractive ones), that are generated by cavity formation, and van der Waals interactions:

$$G_{nonpolar} = G_{cavity} + G_{vdW} \tag{3.25}$$

G_{cavity} stands for the solute's "work" to create a cavity in the solvent and depends on the solute's shape and geometry, and G_{vdW} is the van der Waals energy between solvent and solute. These terms can be estimated using various models, such as: SASA-Only Nonpolar Model; SAV-Only Nonpolar Model; SASA−SAV Nonpolar Model; and SASA−SAV−WCA Nonpolar Model.

The SASA-Only Nonpolar Model is one we have used in our calculations, and also the most commonly used of nonpolar models. The solvent-accessible surface area (SASA) model is based on the assumption that the SASA is roughly dependent on the $G_{nonpolar}$ term and can thus be calculated as follows:

$$G_{nonpolar} = \gamma A + b \tag{3.26}$$

where γ is a coefficient associated with the physical phenomenon of surface tension of the solvent, A is the solvent-accessible surface area (SASA), and b is the fitting parameter.

Using the g_mmpbsa tool, the binding energy is decomposed on a per residue basis. Initially, the energy components E_{MM}, G_{polar}, and $G_{nonpolar}$ are calculated within the bound, as well as the unbound form, and subsequently their contribution to the binding energy ΔR_x^{BE} of residue x is calculated:

$$\Delta R_x^{BE} = \sum_{i=0}^{n} \left(A_i^{bound} - A_i^{free} \right) \tag{3.27}$$

A_i^{bound} and A_i^{free} are the energy of the ith atom from x residue in the bound and unbound forms, respectively, and n is the total number of atoms in the residue. The summed energy contribution is equal to the binding energy, $\Delta G_{binding} = \sum_{x=0}^{m} \Delta R_x^{BE}$, where m is the total number of residues in either protein-protein or protein-ligand complexes.

3.2. Analysis

RMSD

The root-mean-square deviation (RMSD) is a simple and widely used method to evaluate and quantify the similarity between two structures [137,138]. This is achieved by measuring the average distance between the atoms by superimposing their atomic coordinates:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i^2} \tag{3.28}$$

where N is the total number of atoms, and δ_i is the distance between atom i and the same atom in a reference structure. In the first instance of an RMSD calculation, both structures under comparison are fitted by superimposing them through rotations and translations until they hit a minimum value [56]. Furthermore, the RMSD calculations can be made for all atoms or a particular group of them.

Chapter 4

4. Methodological Workflow

4.1. Identification of the binding sites of AQP1

4.1.1. Creating and preparing a small compound ligand Database for the identification of AQP1 binding regions

This thesis's first goal was to identify the most probable binding pockets of AQP1, which could then be used for the Structure-based virtual screening part of the workplan. To perform such task, an MD approach was used together with a small compound Database composed of 35 ligands previously reported in the available literature as function modulators of different AQPs. Using different Docking methodological approaches, two software's were used for this purpose: ADT4 [82]; and VINA [101]. To perform such calculations, we needed, in the first place, to build, prepare, and curate the compound Database. To achieve this goal, we first generated the 3D conformers for all the compounds. Ideally, this should have been done using QM methodologies [64]. Unfortunately, and as already mentioned, a QM approach is computationally demanding, making it almost impossible to apply to big compound Databases. Therefore, several tools relying on Molecular Mechanics were developed. Software's like molconvert from ChemAxon, Balloon, Unicon, Babel, among others, are just a limited number of applications that can perform this task in a much shorter time, with lower computational demands and with reasonable accuracy [139].

We started this process by organizing the 35 compounds into two categories: the ones that had a compound identifier from a commercially available Database, like ChEMBL; and those whose only chemical information available was based on reported publications. A search based on the first group's identifier was performed, followed by the download of the structure file, which in the majority of the cases was in sdf file format. Afterwards, we used ChemAxon's Molconvert software [140] to generate a representative conformer of each compound to the mol2 format. For the second group, we manually built every compound with ChemAxon's MarvinSkecth [141] (a chemically intelligent desktop toolkit that can be used to draw chemical structures), and saved it in InchI file format. Similarly to the first group of compounds, we then used Molconvert to perform the final conformer generation, saving the result again in the mol2 file format.

4.1.2. Preparing the protein and ligand files for the Docking screening with ADT4 and VINA

To perform ADT4 and VINA calculations, protein and ligand files must be converted to pdbqt (Protein Data Bank, partial charge (Q), atom type (T)) file format. This type of file is based on the Protein Data Bank (pdb) file format, with additional information regarding the partial atomic charges, and AutoDocks' atom types [142].

In order to prepare the pdbqt files for the protein receptor and the ligands, we used the python scripts "prepare_receptor4.py" and "prepare_ligand4.py" from AutoDockTools [82], respectively. The "prepare_receptor4.py" script takes a .pdb file from the Crystallographic structure of AQP1 (1IH5)

[143], while the “prepare_ligand4.py” script takes, as an input, the .mol2 files previously generated for each ligand.

4.1.3. Preparing the Docking calculations for ADT4

ADT4 consists of two major programs: Autogrid, which pre-calculates Ligand-based grids for the various atom-types, and Autodock, which performs the ligand's Docking using the formerly grids focused on the binding region of interest of our protein [144]. In order to generate the grid files, the first step is to create the grid parameter file (gpf) for AutoGrid4. This was achieved by defining a grid box on AutoDockTools, that covered the desired binding site. We focused our calculations on two different regions: the “top” region that covered the extracellular side (Figure 4.1) and the “bottom” region that covered the cytoplasmic side (Figure 4.2). This choice was based on the fact that AQP1 is a transmembrane protein, and therefore is embedded in a membrane bilayer. Thus, the areas accessible for compounds’ binding are the extracellular and cytoplasmic sides since the rest of the protein is surrounded by the membrane.

Regarding the setup parameters, for both sides of the protein, we used 126 points in the x-dimension, 126 points in the y-dimension, and 60 points in the z-dimension, as the total grid points per map. The coordinates used to identify the center of the grid box were the following: 50.97 for the x center; and 50.936 for the y center. The z coordinated defined for the extracellular region was 60.526, while for the cytoplasmatic site was 27.087 (Figures 4.1 and 4.2). Lastly, the grid spacing was set to 0.375 angstroms.

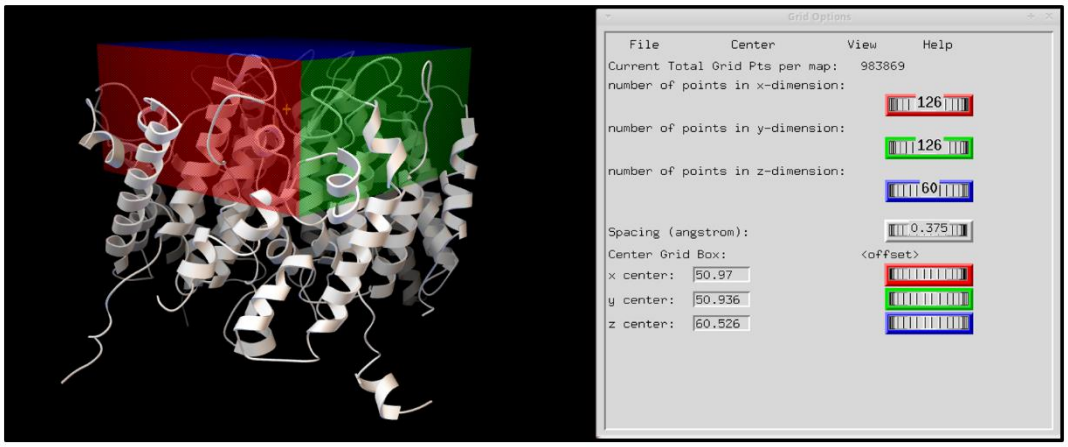

Figure 4.1. On the left: side view of the Crystallographic structure of AQP1, highlighting the “top” grid (extracellular side). On the right: parameters of the “top” grid defined on AutoDockTools for ADT4 calculations.

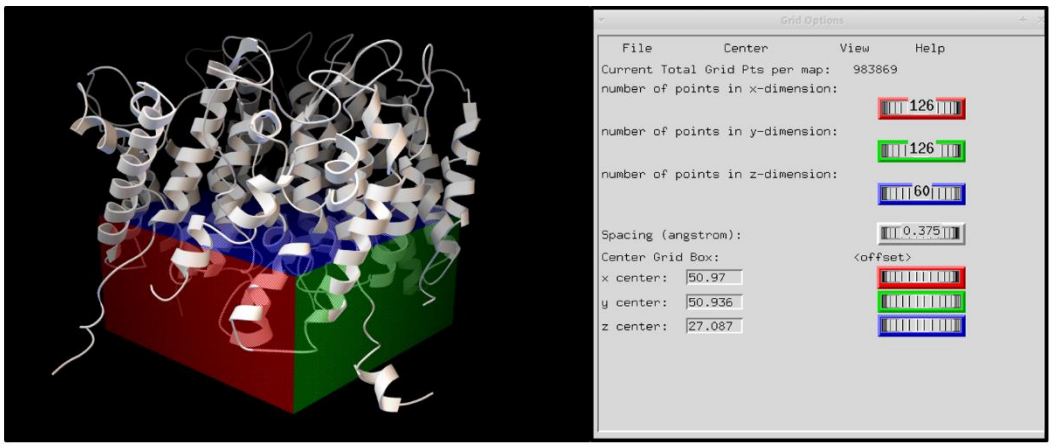

Figure 4.2. On the left: side view of the Crystallographic structure of AQP1, highlighting the "bottom" grid (cytoplasmic side). On the right: parameters of the "bottom" grid defined on AutoDockTools for ADT4 calculations.

4.1.4. Preparing the Docking calculations for VINA

As previously mentioned, ADT4 requires the pre-calculation of grids for the different atom-types found in the target ligand. On the other hand, VINA does not require choosing any atom types and pre-calculating grid maps [101]. Instead, it calculates the grids internally and instantly and does not store them on the disk, making this software significantly faster than ADT4. Furthermore, the search algorithm implemented in VINA is faster when compared to ADT4. Using VINA, the first step is the creation of a configuration (config.txt) file. Analogously to what was previously described for the ADT4 calculation, this was achieved by defining a grid box on AutoDockTools, that covered the same two regions of search previously mentioned: the extracellular (Figure 4.3) and the cytoplasmic (Figure 4.4). The total grid points per map for both sides of the protein were: 56 points in the x-dimension, 56 points in the y-dimension, and 18 points in the z-dimension. For the center of the grid box, the coordinates were as follows: 50.97 for the x center; and 50.936 for the y center. Once again, the coordinate center's definition for the z axis of the two target regions was different. For the extracellular side, it was set to 60.526, while for the cytoplasmic side, it was set to 27.087 (Figures 4.3 and 4.4). Lastly, the grid spacing was set to 1.000 angstroms. Although the used grid spacing is different from those defined for ADT4, they guarantee that the same search region was defined for the two programs.

The previously described Docking process preparation can be found with a high level of detail in [139] - "In silico end-to-end protein-ligand interaction characterization pipeline: the case of SARS-CoV-2", Nícia Rosário-Ferreira, Salete J. Baptista, Carlos A.V. Barreto, Filipe Rodrigues, Tomás Silva, Sara Ferreira, Rita Melo, Bruno L. Victor, Irina S. Moreira, Miguel Machuqueiro", section "Basic Protocol 4 - ADT4 and VINA - two tools for Structure-based virtual screenings". This section is part of a paper that we have recently submitted for publication and should be used as a guideline for this section. In this publication, the protein target was the Main Protease of the SARS-CoV-2 virus.

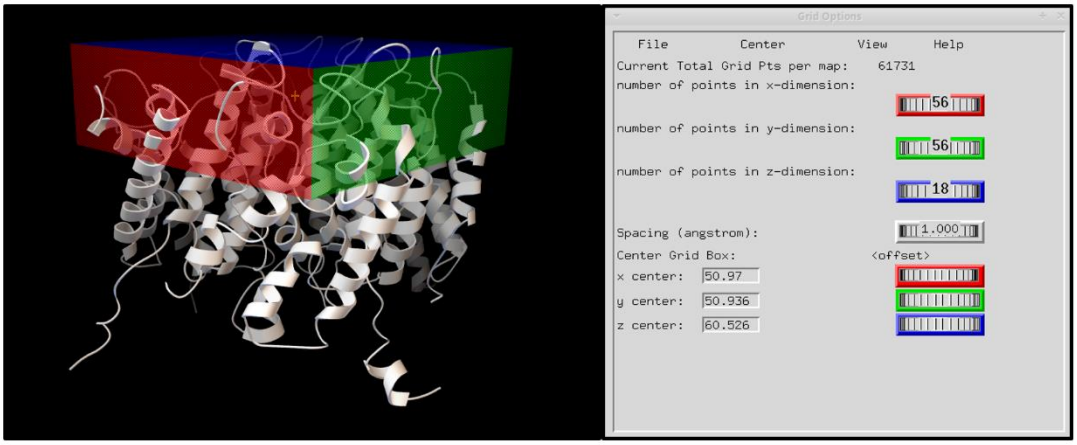

Figure 4.3. On the left: side view of the Crystallographic structure of AQP1, highlighting the "top" grid (extracellular side). On the right: parameters of the "top" grid defined on AutoDockTools for VINA calculations.

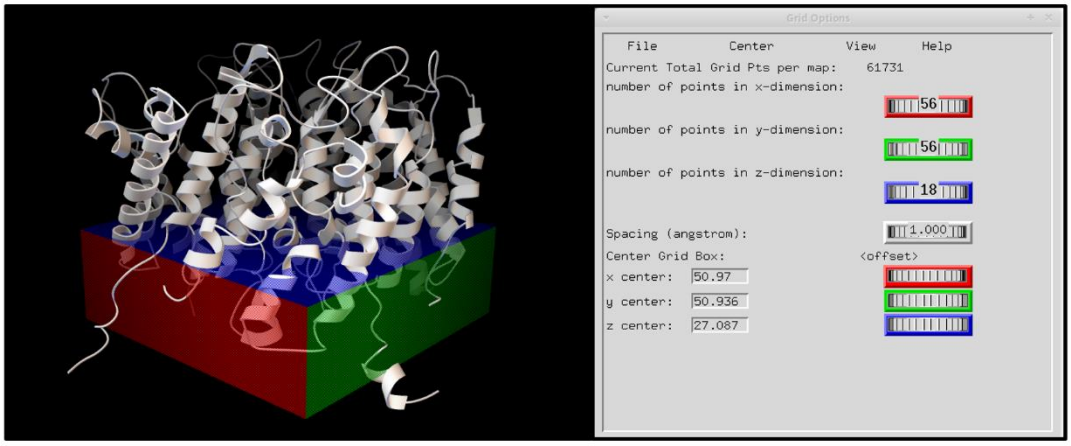

Figure 4.4. On the left: side view of the Crystallographic structure of AQP1, highlighting the "bottom" grid (cytoplasmic side). On the right: parameters of the "bottom" grid defined on AutoDockTools for VINA calculations.

4.1.5. Sampling the conformational space of AQP1 using MD simulations

After this initial approach, we quickly realized that using only the Crystallographic structure of AQP1 wouldn't give us the expected results since the low energy binding solutions obtained from the different ligands weren't focused on the pore or access pore regions, as described in the literature (see Results and Discussion section). For this reason, we decided to use MD simulations to generate different conformations of AQP1 to test with the same Docking procedure as described for the Crystallographic structure.

AQP1 was simulated fully embedded in membrane bilayer composed of 300 POPC lipids, which had already been equilibrated in our lab using long MD simulations. These simulations were performed using the GROMOS 54A7 [77] force field and the GROMACS 2018.6 package at constant pressure and temperature. We used the v-rescale thermostat [118] at 298.15 K, with a relaxation time of 0.1 ps to control the temperature. Regarding the pressure control, we used the Parrinelo-Rahman barostat

[119] for a semi-isotropic coupling at 1 bar, with an isothermal compressibility of 4.5e-5 bar^{-1}, along with a relaxation time of 2.0 ps [145]. For all simulations, the integrator time step was 2 fs and, to constrain the solute bonds, a P-LINCS [146] constraint algorithm was used. Van der Waals interactions were calculated until 1 nm, while long-range electrostatic interactions were calculated with PME [150], with a grid spacing of 12 Å and a 4^{th} order of interpolation.

Before the MD production run was performed, we performed a step of energy minimization to ensure a stable starting system. Initially, we performed 2000 steps of steepest descent minimization with all heavy atoms restrained, followed by another 2000 steps of minimization with the same algorithm, without any restraining. After the minimization procedure, and to allow proper repacking of the lipids around the protein, we performed 5 ns of MD simulation (at constant temperature and pressure) with all the protein atoms harmonically restrained (the force constant used was 1000 kJ mol^{-1} nm^{-2}). We then performed additional 5 ns of MD simulation, with position restraints (the force constant used was 750 kJ mol^{-1} nm^{-2}) applied only to the protein's heavy-atoms. Five replicate MD simulations for full production were then run for a total time per replicate of 50 ns. After evaluating the system stability (using RMSD), we have clustered all the conformations extracted from the equilibrated system and identified the most populated clusters. A representative snapshot of the conformation from the top 3 most populated clusters was extracted and used in the subsequent studies.

4.1.6. Binding pocket predictions

Additionally, to the use of Molecular Docking methodologies to identify possible new binding site regions at the AQP1 surface, we have also used other computational approaches to validate and combine the obtained data. To accomplish this task, we used three different automated pocket detection approaches, available in three different software's: DogSiteScorer [78–80], a web server for automatic binding pocket prediction, analysis, and druggability assessment; AutoDockTools AutoLigand [81], a ligand site finder and drug development tool based on Autodock; and SeeSar [84], a software tool from BiosolveIT for interactive, visual compound prioritization as well as compound evolution. This evaluation process was applied, as previously described, to the four selected structures/conformations of AQP1: the Crystallographic structure of AQP1 and the three MD snapshots.

4.1.7. Compound Database enrichment using a Ligand-based approach

As previously described, our initial compound dataset was only composed of 35 ligands reported to have some activity towards different AQP subtypes. In order to have a significant statistical approach to identify the most probable regions where the Docking of compounds at the AQP1 protein surface occurs, we needed to perform a higher number of Docking assays with a higher number of chemically diverse compounds. To achieve that, we needed, in the first place, to enrich our library of compounds. Therefore, we used Feature Trees (FTrees) software by BioSolveIT [147], a highly efficient software tool for pharmacophore-style ligand similarity searching. This software is able to capture a molecule's overall topology and its pharmacophore properties, which are then used as similarity descriptors on the search in compound Databases for new similar compounds. We applied this software to our initial Database of 35 compounds against four chemical compound Databases: ChEMBL, DrugBank, NCI, and ZINC. As a result, we obtained an initial list of 11808 compounds, which after filtering (to detect and remove compound duplicates), was composed of 7204 compounds. This new dataset was then used to repeat a new screening, this time solely with VINA software, using the previously described Docking

protocol and settings. This decision was made due to the fact that VINA is much faster than ADT4, without compromising the accuracy of the results. Once again, the Docking screening was done at the two sides of the protein: the extracellular side (Figure 4.3); and the cytoplasmic side (Figure 4.4).

4.1.8. Quantitative evaluation of Ligand-Protein Interactions

To determine, more rigorously, which binding regions of the protein were the most adequate to focus our subsequent screening studies, we decided to identify the most populated regions of AQP1 where the compounds from our new dataset docked. We have focused our analysis on the overall tetrameric AQP1 structure and also on the individual Monomers (A, B, C, and D) composing the functional structure. This was achieved using the "pairwisedistances.py" Python script on Pymol, which, given any two selections, calculates and returns the pairwise distances between all atoms that fall within a user-defined cut-off [148]. Using a 4.5 Å cut-off, we again focused the analysis on the two sides of the protein: the extracellular side and the cytoplasmic side. Finally, to analyse the obtained results, we have used two complementary approaches. In the first one, a visual approach with PyMOL was developed using a color code analysis directly related to the number of interactions between the protein and the compounds. The used spectrum was the "blue_white_red" one, where blue indicates a low number of interactions, white indicates an intermediate number of interactions, and red indicates a high number of interactions. In the second approach, we followed on with a more quantitative perspective, where every interaction between the amino acids of each Monomer (A, B, C, and D) and the different compounds was counted and plotted in a bar graph.

4.2. Screening of commercially available Databases with VINA

After identifying the desired binding sites, we performed a Molecular Docking screening with four different compound Databases: the one containing the 7204 ligands identified from the previously mentioned Ftrees approach; and three commercially available ones (DrugBank, NCI, and ChEMBL).

This part of the work became one of the most time-consuming steps of the thesis here presented due to the fact that although the previously referred Databases are available, they need a lot of pre-preparation and curation before they can be used in screening approaches. As previously mentioned in the section "Preparing the protein and ligand files for the screening with ADT4 and VINA", these calculations require ligand files in the pdbqt format, which, from a .sdf file, can be obtained through a series of conversions. At first glance, this didn't sound like a problem to us, so we simply used some of the tools that exist for this end (molconvert from ChemAxon, Ballon, Unicon, Babel, among others). Unfortunately, some of the selected tools ended up not working since some of the compounds' generated 3D conformations were inconsistent with their correct chemical 3D chemical arrangement. We only realized that something was wrong because some of these problematic compounds showed up in our top Docking scores. When we started to analyze these solutions carefully, we immediately identified these problems. To correct them, we then tested multiple combinations of the referred conformer generation/conversion tools. The solution was achieved using babel to go from the InChi code to the Smiles code (2D conversion), then Molconvert to go from the Smiles code to the .mol2 file (3D conformer generation), and finally the python "script prepare_ligand4.py" to go from the .mol2 file to the .pdbqt file. This successive conversion scheme was chosen because some of the used software's do not allow a direct conversion between the desired file formats.

At this stage, we decided to focus these calculations only on one Monomer of AQP1 from one of the four previously evaluated conformations of the tetrameric protein. We decided to select only the

Monomers since this thesis goal was to identify modulators of the AQP1 function, which was already referred to occur in the pore region of each of the four Monomers. This selection procedure was made based on the Docking results and Ligand-Protein Interactions (see Results and Discussion section). In the end, the final screening campaigns were focused on three different regions of AQP1: two were focused on two different regions on the extracellular site (close to the ar/R selectivity filter, and the other closer to the NPA region), and one on the cytoplasmatic site (including the NPA motif and the region below it). Two different monomers (B and C) of a single conformation of AQP1 (referred to previously as MD2) were used in these studies. Monomer B showed the most suitable structure for the focus of the first binding region (from now on referred to as B_NPA) (Figure 5.21). In contrast, monomer C was displayed as the most advisable region for the other two binding regions, which from now on will be referred to as C_ar/R (Figure 5.22) and C_NPA (Figure 5.23) (see Results and Discussion section).

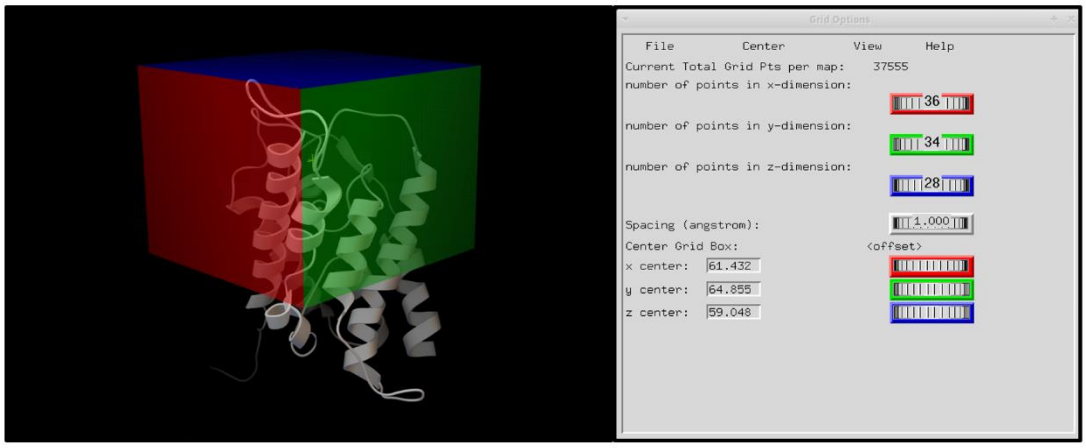

Figure 4.5. On the left: side view of the Monomer B of the MD2 structure of AQP1, highlighting the "top" grid (extracellular side). On the right: parameters of the "top" grid defined on AutoDockTools for VINA calculations.

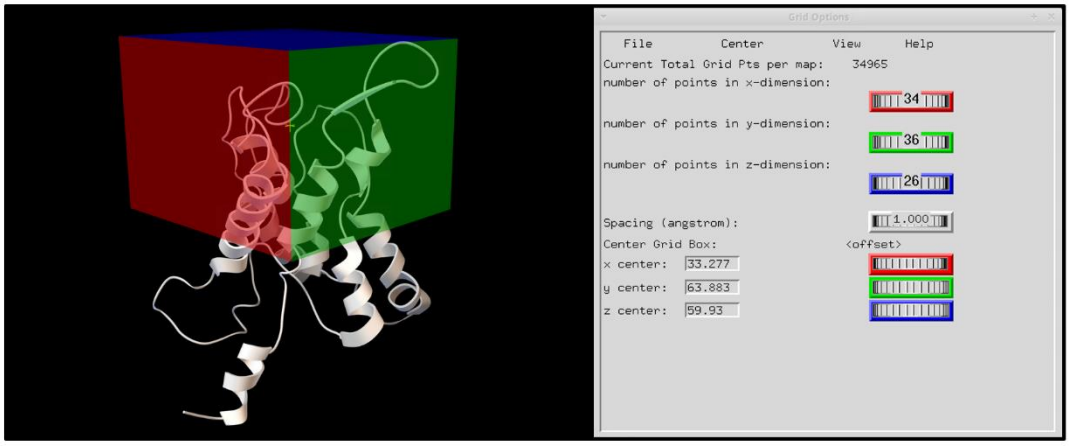

Figure 4.6. On the left: side view of the Monomer C of the MD2 structure of AQP1, highlighting the "top" grid (extracellular side). On the right: parameters of the "top" grid defined on AutoDockTools for VINA calculations.

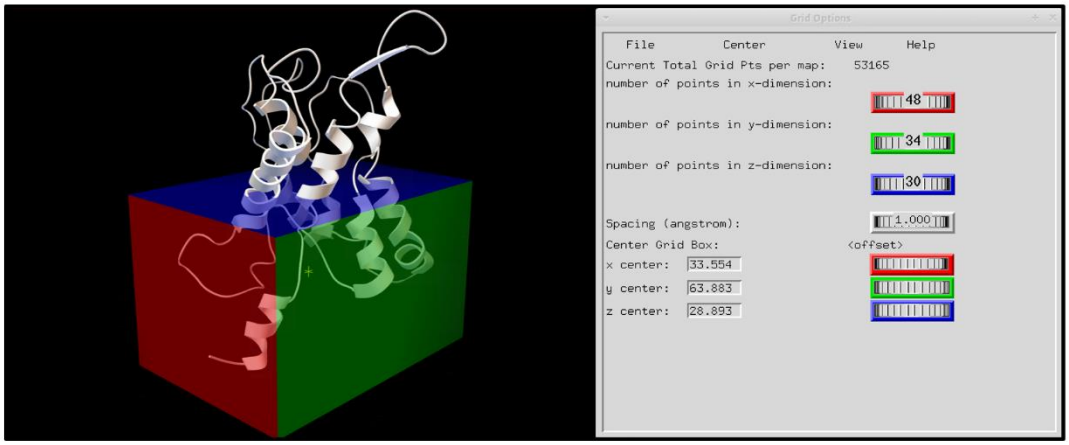

Figure 4.7. On the left: side view of the Monomer C of the MD2 structure of AQP1, highlighting the "bottom" grid (cytoplasmic side). On the right: parameters of the "bottom" grid defined on AutoDockTools for VINA calculations.

In the B_NPA region setup, the Docking screening calculations used a grid with 36 points in the x-dimension, 34 points in the y-dimension, and 28 points in the z-dimension. The grid's center was set to: 61.432 for the x center; 64.855 for the y center; and 59.048 for the z center (Figure 4.5). For the C_ar/R region, the simulation grid was defined by 34 points in the x-dimension, 36 points in the y-dimension, and 26 points in the z-dimension. The center of this grid was set to the following 3D coordinates: 33.277 for x axis, 63.883 for the y axis, and 59.93 for the z axis (Figure 4.6). Lastly, for the C_NPA region, we used 48 points in the x-dimension, 34 points in the y-dimension, and 30 points in the z-dimension. For the center of the grid box, we defined 33.554 for the x axis, 63.883 for the y axis, and 28.893 for the z axis (Figure 4.7). For all three regions, the Spacing was set to 1.000 angstroms.

The results obtained from the Molecular Docking screening were then sorted according to the VINA scoring function. At this stage, we have also performed the filtering of the evaluated compounds according to Lipinski's rule of five filters [149]. The filters in this rule state that, in general, an orally active drug has no more than one violation of the following criteria:

1. No more than 5 hydrogen bond donors (the total number of nitrogen-hydrogen and oxygen-hydrogen bonds);
2. No more than 10 hydrogen bond acceptors (all nitrogen or oxygen atoms);
3. A molecular mass less than 500 daltons;
4. An octanol-water partition coefficient [150] (log P) that does not exceed 5.

Note that all numbers are multiples of five, which is where the origin of the rule's name resides. To perform such task, we have used the software DataWarrior [151], ending up by selecting the top 20 compounds for each of the binding sites previously defined (a total of 60 compounds).

4.3. MM/PBSA calculations of the most promising identified AQP1 modulators

With the 60 compounds (20 for each binding site) previously identified, the following step was to perform short MD simulations. Firstly, this was done to evaluate the compounds' stability in the binding regions of the AQP1, and then to also use, from these results, an ensemble of complex conformations to calculate the complexes' binding energy using the MM/PBSA method. This approach is commonly

used to rescore Docking complexes, thereby identifying the most promising and stable AQP1 binding compounds.

To perform such simulations, we had, in the first place, to parameterize all 60 compounds according to the GROMOS 54A7 force field. To accomplish this task, we have used the ATB Web server [108,110]. This approach allows the generation of Force Field descriptors of novel molecules that are not included in the Force Field core set. After the compounds' parameterization, we downloaded all the GROMOS 54A7 force field files, adapting them to simulate all the complexes. All the MD simulations in this section followed the same parameters previously specified in the section "Sampling the conformational space of AQP1 using MD simulations". Initially, each of the 60 MD simulations was run for a total of 10 ns. After evaluating the system stability (using RMSD), we realized that in 24 of them, the compound did not show a stable conformational sampling at the binding site of the protein. In this way, we decided to extend the MD simulations of these 24 complexes until all complexes reached an equilibrated conformational state. Overall, we ran 36 simulations for a total of 10 ns, 15 for 30 ns, 7 for 50 ns, and 2 for 100 ns. In the end, we have selectively used part of MD trajectories where the complexes showed to be in their equilibrated conformational state (20 individual snapshots) to perform binding free energy calculations using the MM/PBSA method implemented in the g_mmpbsa application [127].

Chapter 5

5. Results and Discussion

5.1. Screening with a 35 compound Database

In the initial Docking screening (ADT4 and VINA) with the AQP1 Crystallographic structure and the 35 compound Database, the results showed divergent results. As can be seen in figure 5.1, where the results obtained with ADT4 on both the extracellular and cytoplasmatic regions are shown, the majority of the solutions are found at the interface between the different monomers of AQP1. Interestingly, a different Docking pattern was predicted with VINA at both regions of the protein. As shown in figure 5.2, when focusing our attention on the extracellular region results, most of the solutions are found close to the hydrophobic central pore of the protein (at the different monomers interface). This pattern is not observed at the cytoplasmatic site, where the Docking solutions are more spanned over all of the proteins' surface. From these results, two main conclusions can be taken. The first one is that both programs give different results. The second is that, independently of the target region, none of the programs can identify the protein regions already reported as being important for the development of function modulators. Most probably, what is happening in this situation is that the conformation obtained from crystallography is biased to a representation that is not the most adequate to be used in binding calculations. The loops connecting the transmembrane helices form one of the main problems with the Crystallographic structure of AQPs. The conformations of these loops, found in the Crystallographic structure, occupies the region of access to the pores of each monomer, making it impossible for the Docking programs to identify low energy Docking solutions focused on the most druggable regions of the protein (as reported in several publications [152–154]). Therefore, one possible solution to overcome this problem is to generate different conformations of the protein. Several methods, such as flexible Docking approaches and Monte Carlo simulations, allow performing such task. However, MD simulations have previously been described as one of the most efficient solutions to problems such as the ones we have faced with Aquaporins.

Taking these facts into account, we have performed five replicate MD simulations, starting from long pre-equilibrated simulations of AQP1 (R1 - R5). We have run these simulations for 50 ns. As can be seen in Figure 5.3, after 25 ns of simulations, AQP1 reached a stable conformational state with a converged RMSD value of approximately 0.65 nm (in respect to the Crystallographic structure). In order to identify the most populated and, consequently, more representative conformational states of AQP1 under the equilibrated part of these simulations, we have performed a clustering analysis of the last 15 ns of simulations. From this analysis, we have identified three different snapshots (MD1, MD2, MD3) representing the most populated conformations. As expected, most of the structural variations observed in these different conformations are focused on the loops connecting the transmembrane helices. Interestingly, some of the selected conformations show the regions of functional interest (ar/R and NPA) to be more exposed to the solvent, making them consequently more adequate to perform the simulation of the binding of different compounds. As can be seen, for example, in Figure 5.4, which shows a binding region that exposes the ar/R region in a manner that could not be observed in AQP1 Crystallographic structure. The use of this multiconfigurational approach is quite different from what has been previously followed by others, where the use of the Crystallographic structure is commonly used to identify new AQP modulators [152,154–156].

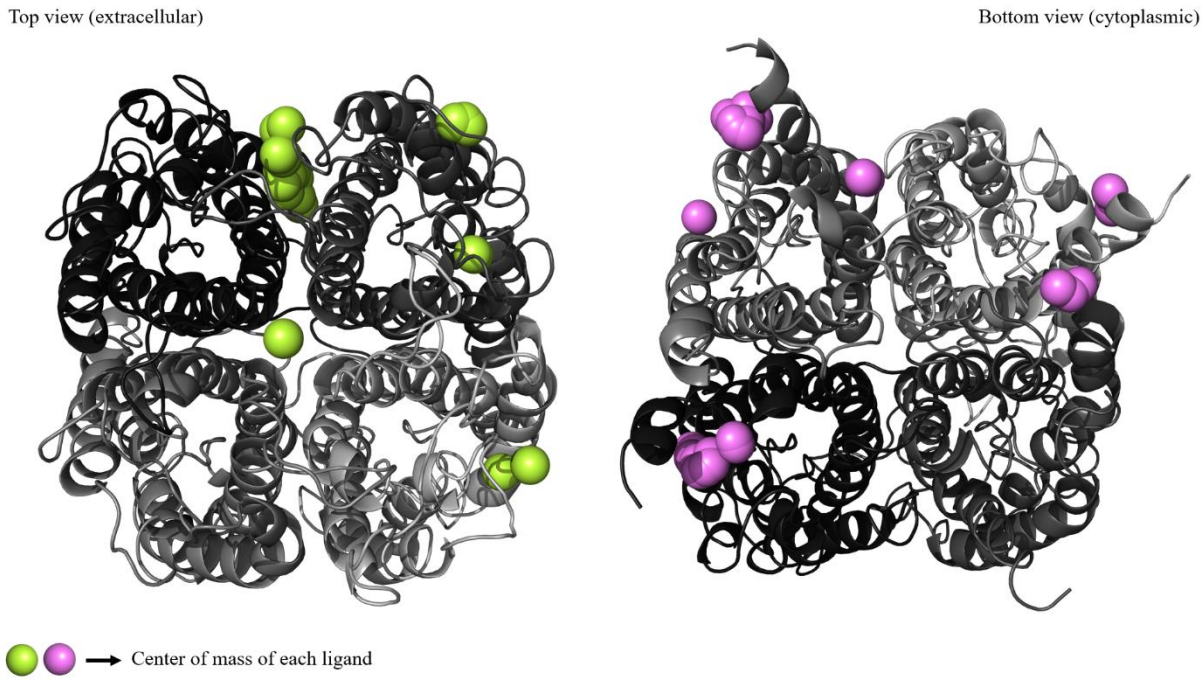

Figure 5.1. Crystallographic structure of AQP1 after undergoing Molecular Docking using the ADT4 software, with 35 ligands. The secondary structure of the protein is represented as cartoon, and the centers of mass of the docked ligands are represented as small green spheres for the top view (left) and as small pink spheres for the bottom view (right).

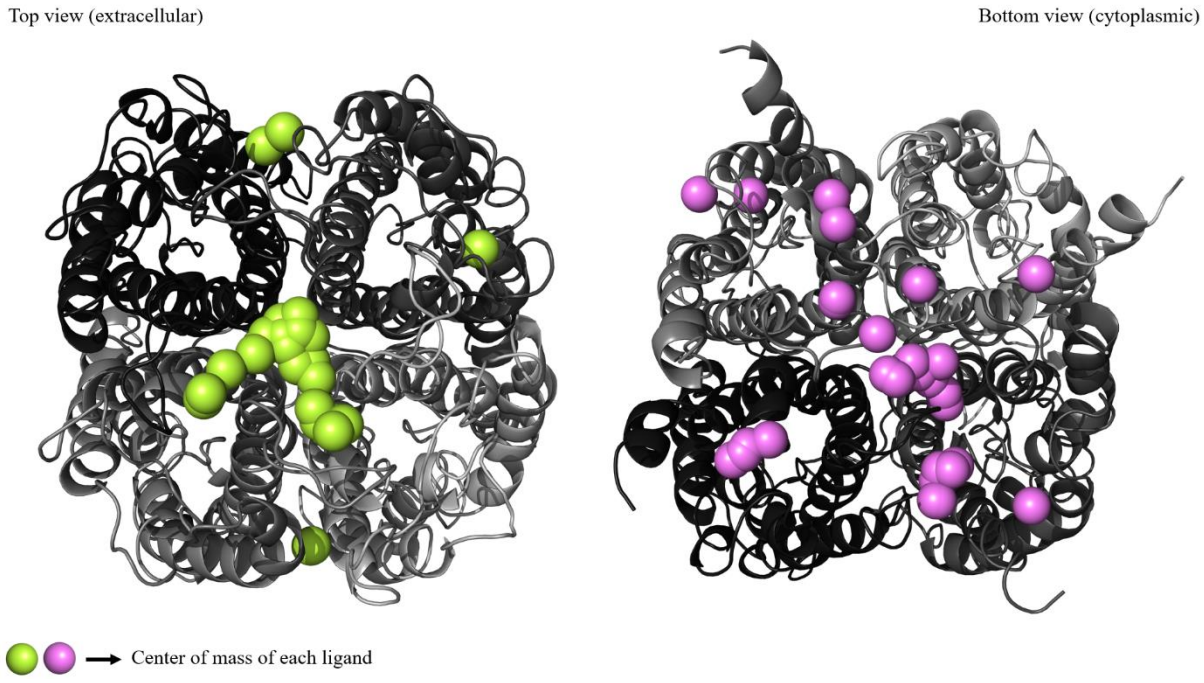

Figure 5.2. Crystallographic structure of AQP1 after undergoing Molecular Docking using the VINA software, with 35 ligands. The secondary structure of the protein is represented as cartoon, and the centers of mass of the docked ligands are represented as small green spheres for the top view (left) and as small pink spheres for the bottom view (right).

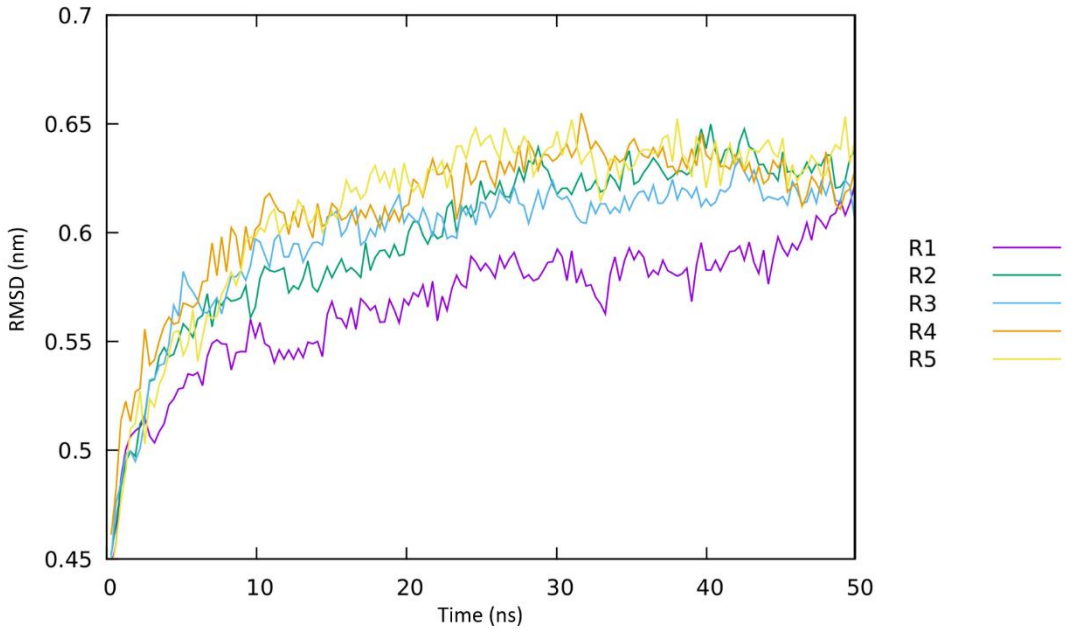

Figure 5.3. Representation of the RMSD (nm) variation through simulation time (50 ns) of the five replicate MD simulations concerning the Crystallographic structure conformation of AQP1.

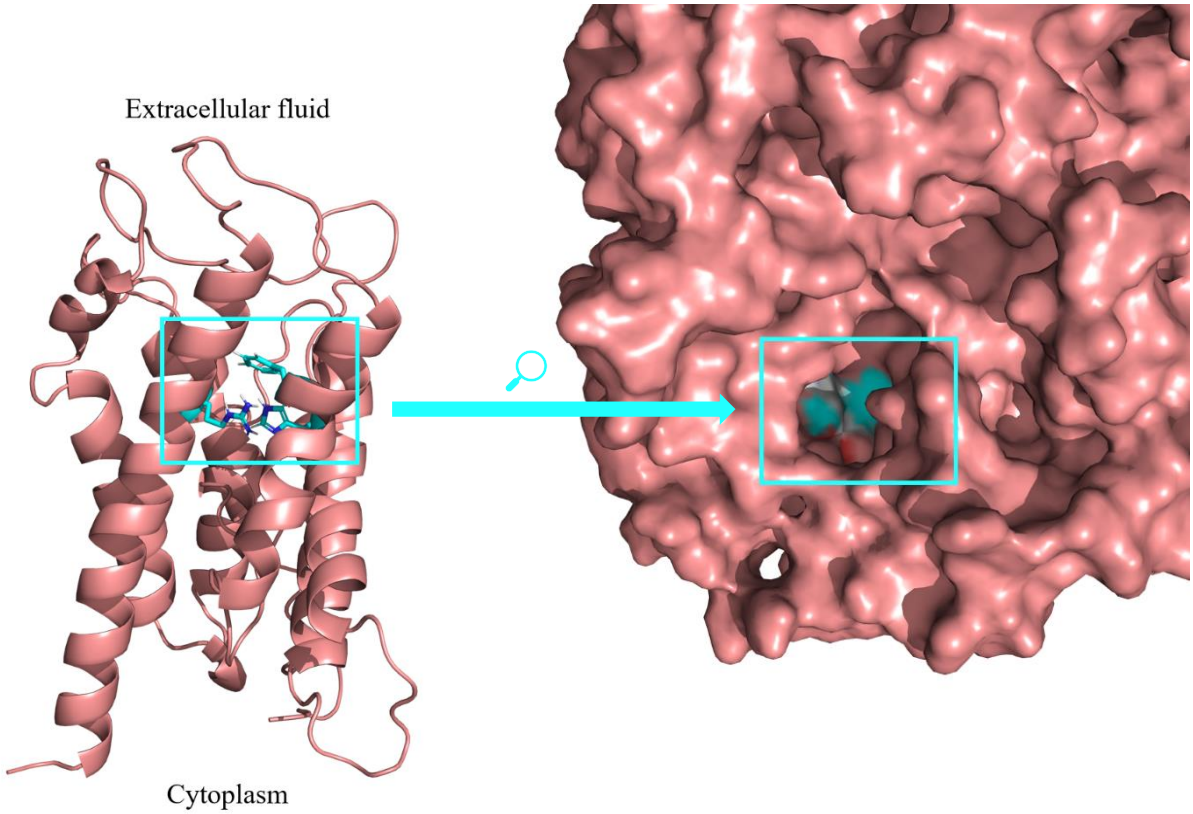

Figure 5.4. Monomer D of the MD1 structure of AQP1, highlighting the ar/R selectivity filter (blue square). On the right: cartoon representation. On the left: surface representation.

5.2. Screening with a 35 compound Database using different AQP1 structures

From our initial tests with the 35 compound dataset, both ADT4 and VINA (Figure 5.1 and 5.2) behaved similarly regarding the binding of the Crystallographic structure of AQP1. As previously reported, independently of the use of ADT4 and VINA, most compounds bind with higher stability to regions of the protein that, according to different reports, are not functionally interesting or relevant for modulating the function of this type of proteins. Therefore, we have used MD simulations to generate more conformations of AQP1 and selected those that are at the same time more populated and also with the necessary characteristics (such as the exposure to the functional relevant sites such as ar/R) to be more adequate to use in the identification of new functional modulators. As already referred, we need to be careful in terms of computational time and, for this reason, VINA was the major Molecular Docking software used in this thesis. Therefore, all the following Docking results were obtained using the VINA software solely.

As shown in Figure 5.5 (focused on the "top" region), we can observe that the pattern of the Docking solutions observed with the Crystallographic structure of AQP1 is not observed in the different selected MD snapshots. While in the Crystallographic structure, the compounds are binding at the hydrophobic central pore region, in the different MD snapshots, the Docking solutions are seen more dispersed over the surface of the different monomers, with a pin-point focus on the access region to the ar/R region. Regarding the results obtained for the cytoplasmic faced surface of AQP1 (Figure 5.6), similar conclusions can be drawn, although in this case, the initial Docking pattern obtained with the Crystallographic structure is not so evident but still observed.

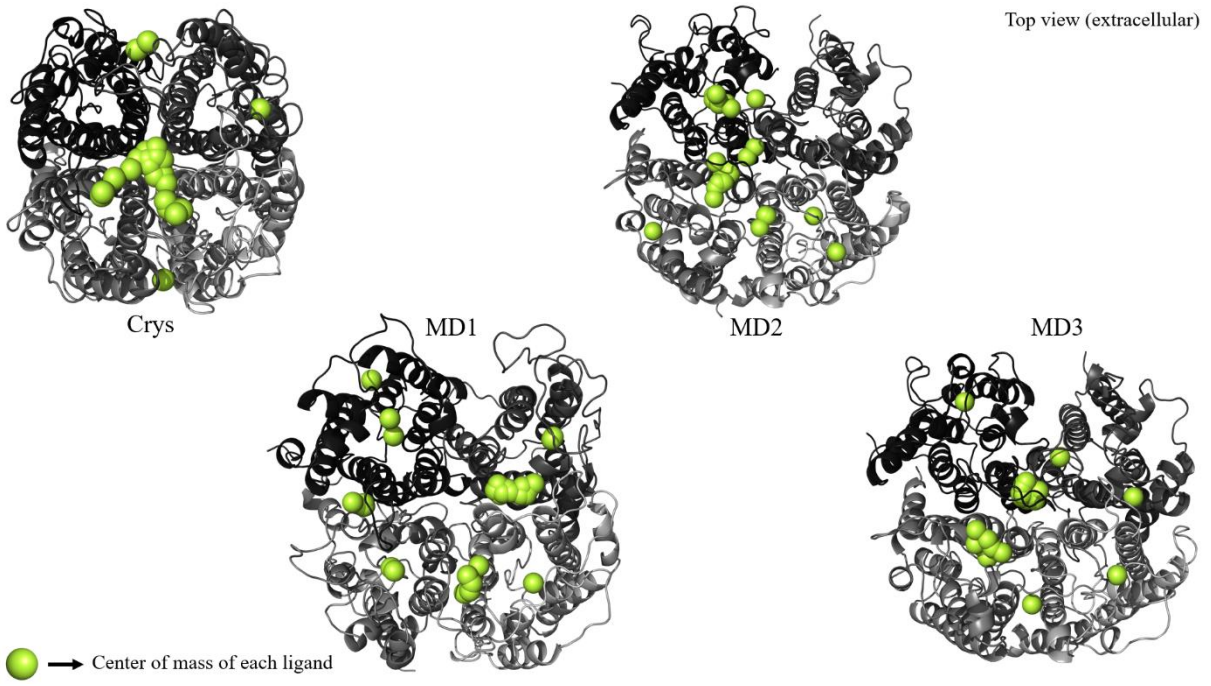

Figure 5.5. Top view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Molecular Docking using the VINA software with 35 ligands. The secondary structure of the protein is represented as cartoon, and the centers of mass of the docked ligands are represented as small green spheres.

In this thesis, we have focused our attention on both extracellular and cytoplasmatic regions of the AQP1 protein, as possible druggable regions. The extracellular faced region is the one that is more prone to possible inhibitors coming from the exterior side of the cells where AQPs are found. However, it has also been suggested that some compounds, due to their amphipathic chemical properties, could diffuse through the lipid bilayer and reach the cytoplasmic side of AQPs pores [157,158]. The probability of a compound acting upon this region is significantly low compared to the extracellular side since an additional diffusion step, penetrating the membrane to reach this region of the protein, must be successfully performed. Without discarding this region, the extracellular region will be our main focus since the binding process focused on this region seems, in our view, more simplified.

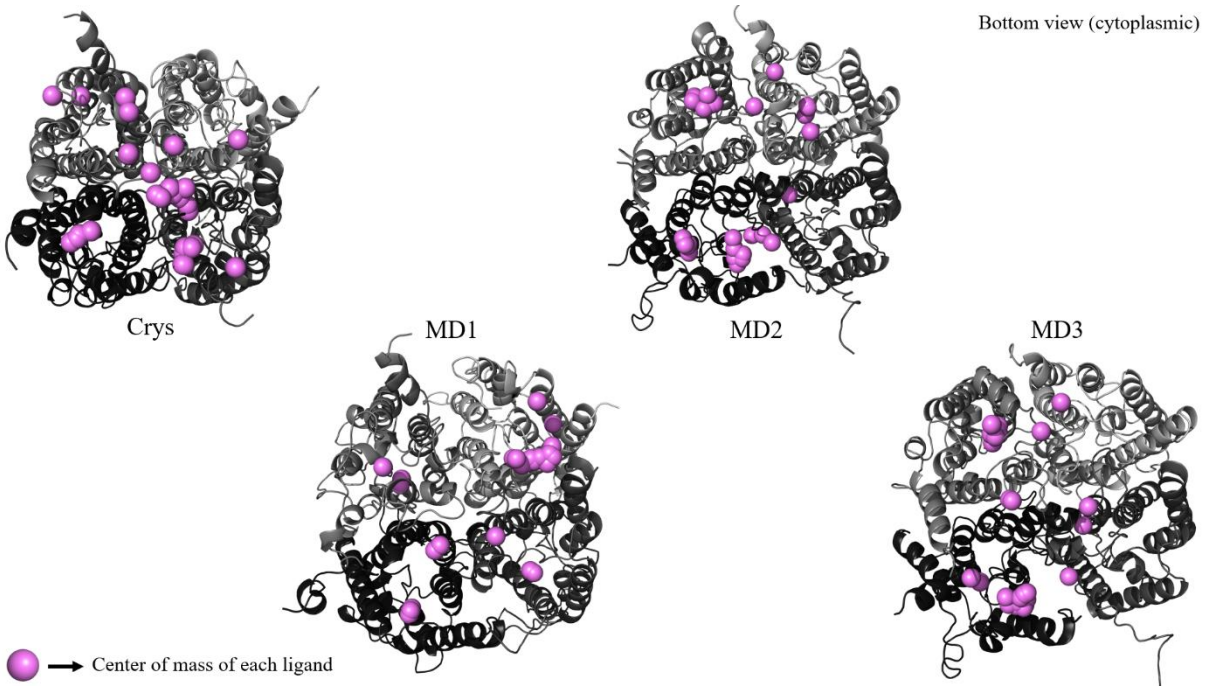

Figure 5.6. Bottom view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Molecular Docking using the VINA software with 35 ligands. The secondary structure of the protein is represented as cartoon, and the centers of mass of the docked ligands are represented as small pink spheres.

5.3. Binding pocket predictions

To complement the results we have obtained from the interactions obtained with the 35 compound Database and the different structures of AQP1, we have additionally used Automated Pocket Detection software's, which are able to detect potential binding pockets of a protein of interest. By applying DogSiteScorer, AutoLigand and Seesar, to both extracellular and cytoplasmatic surface regions (Figures 5.7 and 5.8, respectively), we can conclude that clear druggable binding regions at AQP1 Crystallographic surface could not be identified close to the central pore regions of each Monomer. The identified binding pockets are scattered all over the used structure, covering almost the majority of AQP1, including the protein's outer regions. However, when we submitted the MD snapshots to the previously mentioned binding site detection software's, distinct conclusions can be taken, similar to what we have observed in the previous section, where we used a small Docking screen approach to

identify binding site regions on AQP1. Although the solutions obtained with the different MD snapshots are not conclusive about the identification of promising druggable regions at the protein's surface, it is clear that the use of these structures *in silico* drug discovery campaigns is more suitable than solely using the Crystallographic structure. As can be seen in Figures A.1-A.6 (Appendix), the number of regions identified as druggable with the different software's is much higher in the MD snapshots when compared to the results obtained with Crystallographic structure. Overall, these conclusions justify our choice to focus our subsequent screening studies solely on the MD snapshots.

All the presented results and drawn conclusions, up to this part of the thesis, were gathered and presented, in the form of a scientific poster, at EJIBCE 2019 (Encontro de Jovens Investigadores em Biologia Computacional Estrutural), that took place at Faculdade de Ciências da Universidade de Lisboa, on the 20th of December of 2019 [159].

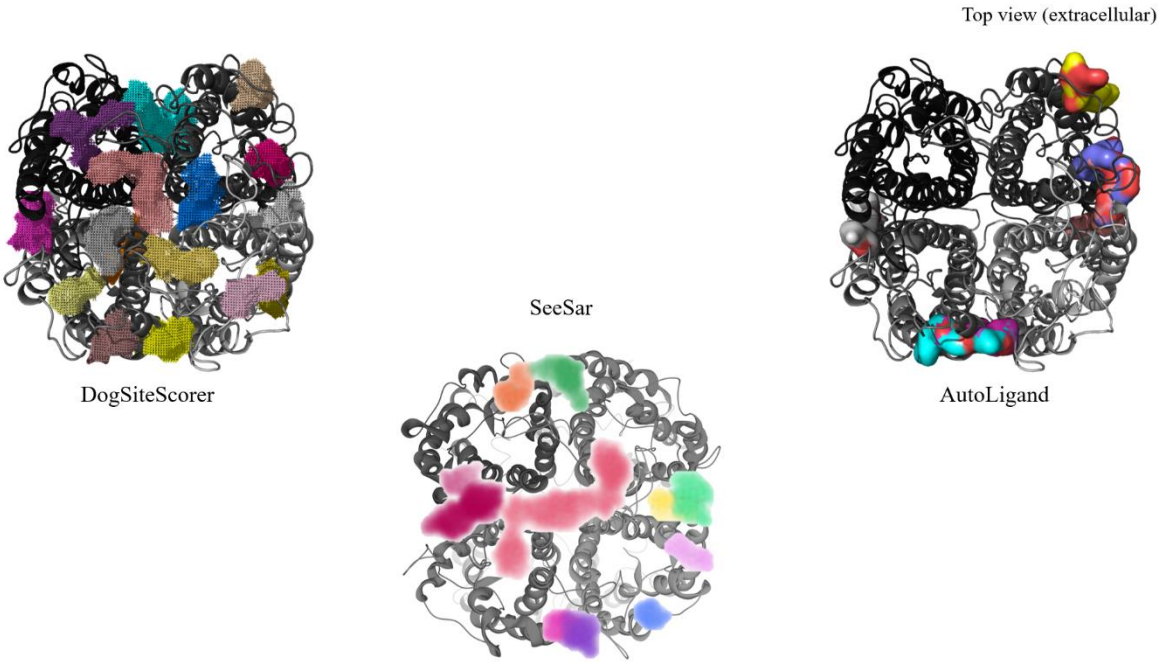

Figure 5.7. Top view of the Crystallographic structure of AQP1 after undergoing three different automated pocket detection approaches. On the left: the DogSiteScorer results (colored mesh). On the right: the AutoLigand results (colored surface). On the bottom: the SeeSar results (colored surface).

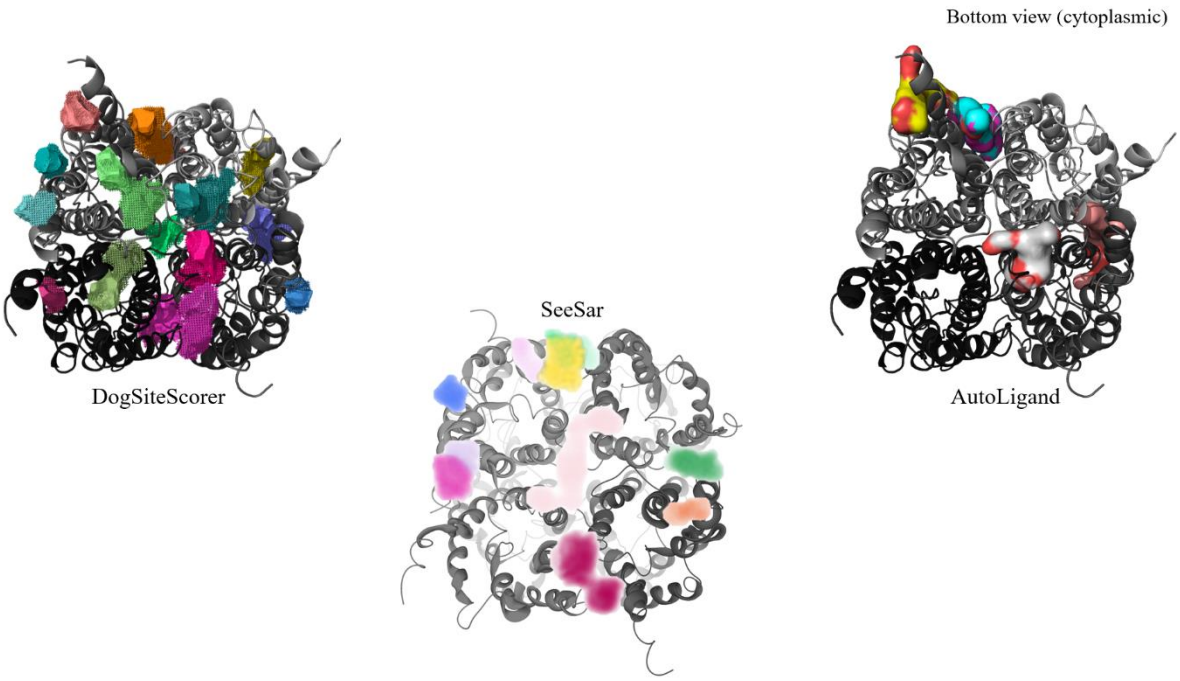

Figure 5.8. Bottom view of the Crystallographic structure of AQP1 after undergoing three different automated pocket detection approaches. On the left: the DogSiteScorer results (colored mesh). On the right: the AutoLigand results (colored surface). On the bottom: the SeeSar results (colored surface).

5.4. Screening with a 7204 compound Database using different AQP1 structures

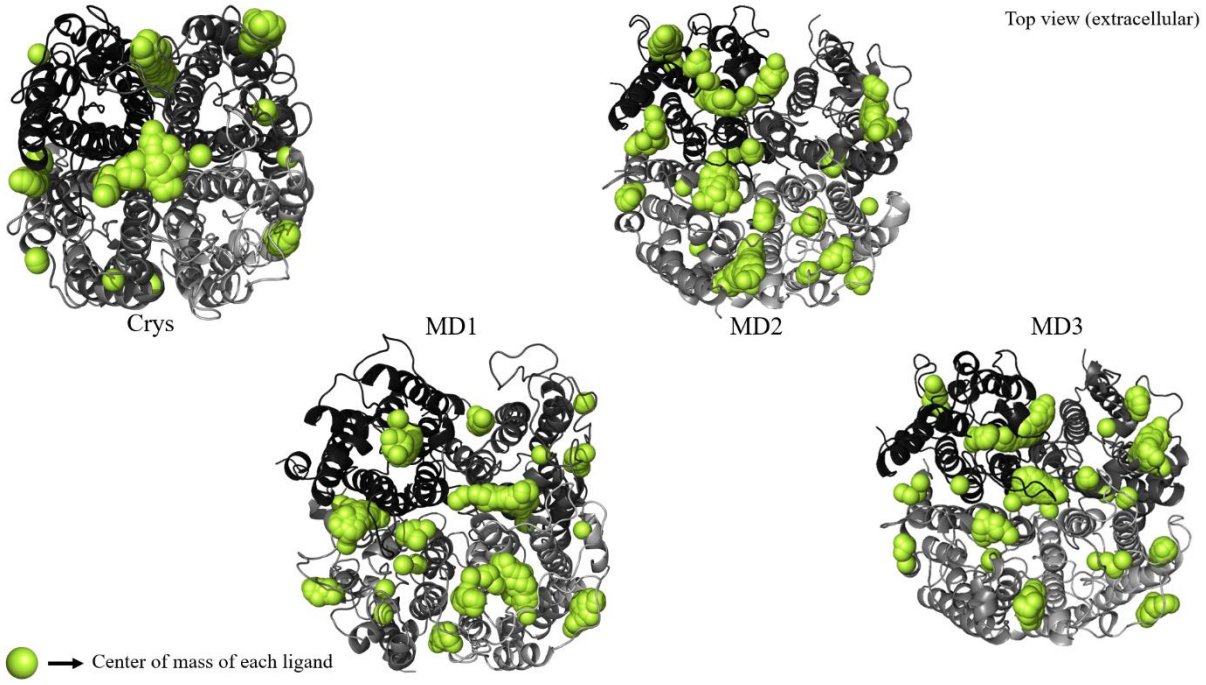

Figure 5.9. Top view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Molecular Docking using the VINA software with 7204 ligands. The secondary structure of the protein is represented as cartoon, and the centers of mass of the docked ligands are represented as small green spheres.

With the ultimate goal of selecting/identifying the most adequate structural representation of AQP1 to be used in our final screening campaigns, we decided to enrich our compound Database in order to see if the same type of Docking mapping conclusion was obtained with both the Crystallographic structure and MD snapshots of AQP1. As previously described, in this thesis, we have used FTrees to generate a compound database with 7204 compounds, showing similar chemical properties as the 35 described AQPs modulators.

As shown in Figure 5.9, similar conclusions can be drawn when using this larger compound dataset (compare with Figure 5.5) for the Crystallographic structure. As previously observed for the top region, most Docking solutions are focused on the central area of the AQP tetramer. However, this Docking pattern is not observed in the results obtained from the three MD snapshots. Using these different conformations of the AQP1 protein, we could identify a more scattered distribution of the Docking solutions, with pin-point clusters focused on the different Monomers' pore entrance regions. These more populated regions could not be identified in the previous screening with the 35 compound Database evidencing, therefore, the need to use enriched compound Databases for this type of screening approach.

It is essential to recall the reader that the Monomers' positions in all Docking results are the same as previously highlighted in Figure 1.5. When a top view (extracellular) is applied, Monomer A is in the bottom right position, Monomer B is in the top right position, Monomer C is in the top left position, and Monomer D is in the bottom left position. However, when a bottom view (cytoplasmic) is applied, Monomer A is in the top right position, Monomer B is in the bottom right position, Monomer C is in the bottom left position, and Monomer D is in the top right position.

Focusing on the MD1 structure, especially in Monomers A, C, and D, it is possible to identify clusters of Docking solutions at the pores' region. In the MD2 structure, a similar pattern is observed in Monomer C and D, while in the MD3 structure, Monomer D is the only one where we can identify a high density of Docking solutions close to the pore region.

Regarding the results obtained for the cytoplasmic region of AQP, the conclusions previously drawn (Figure 5.6) cannot be applied. As shown in Figure 5.10, the pattern previously described in all structures, where most ligands are localized on the protein's central area, is not observed. In this case, the ligands are scattered throughout the whole protein's bottom surface. In the Crystallographic structure, the obtained results are more focused on the protein's outer region. Nonetheless, in the MD1 structure, all the four Monomers (A - D) show aggregates of ligands focused on the central pore region. In the MD2 structure, Monomer A is the only one where we can observe an arrangement of the ligands on the pore region. In the MD3 structure, this pattern is observable in Monomer A and D. Overall, the difference between the observed pattern for the Crystallographic structure and the MD conformations is that in the latter, we can observe the formation of some aggregates of ligands localized in the central pore region of each of the four Monomers, displaying probable binding pockets.

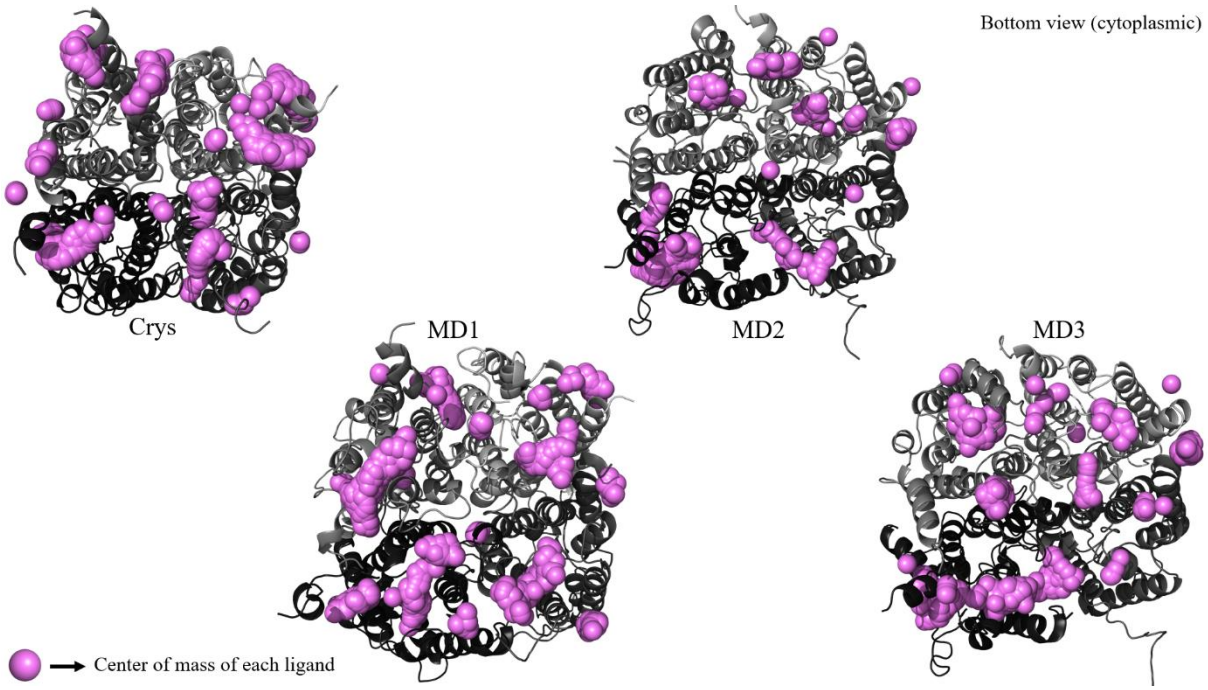

Figure 5.10. Bottom view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Molecular Docking using the VINA software with 7204 ligands. The secondary structure of the protein is represented as cartoon, and the centers of mass of the docked ligands are represented as small pink spheres.

5.5. Quantitative evaluation of Ligand-Protein Interactions

In order to quantitatively evaluate the regions at both extracellular and cytoplasmatic regions of AQP1, more prone to bind the 7204 ligands (section 5.4), we have measured the number of interactions these compounds made with the different regions of the protein. To perform such a task, we have used a pairwise distance criterium using the "pairwisedistances.py" Python script, on Pymol, with a maximum cut-off distance of 4.5 Å. In Figures 5.11 and 5.16, we have highlighted the results using a blue to red Pymol spectrum mapped on the secondary structures of AQP1. The regions colored in blue indicate a low number of interactions, while the areas colored in red indicate those with a high number of interactions. In Figures 5.12-5.15 and 5.17-5.20, these results were highlighted by plotting the number of interactions against all the amino acids of the AQP1. In order to discriminate between the different Monomers of the protein, we have colored the bars accordingly: blue bars represent Monomer A; the magenta bars represent Monomer B; the yellow bars represent Monomer C; and the salmon bars represent Monomer D. In the two referred approaches, the Pymol spectrum analysis, and the plotted interactions, represent the same data set, only presented in complementary ways.

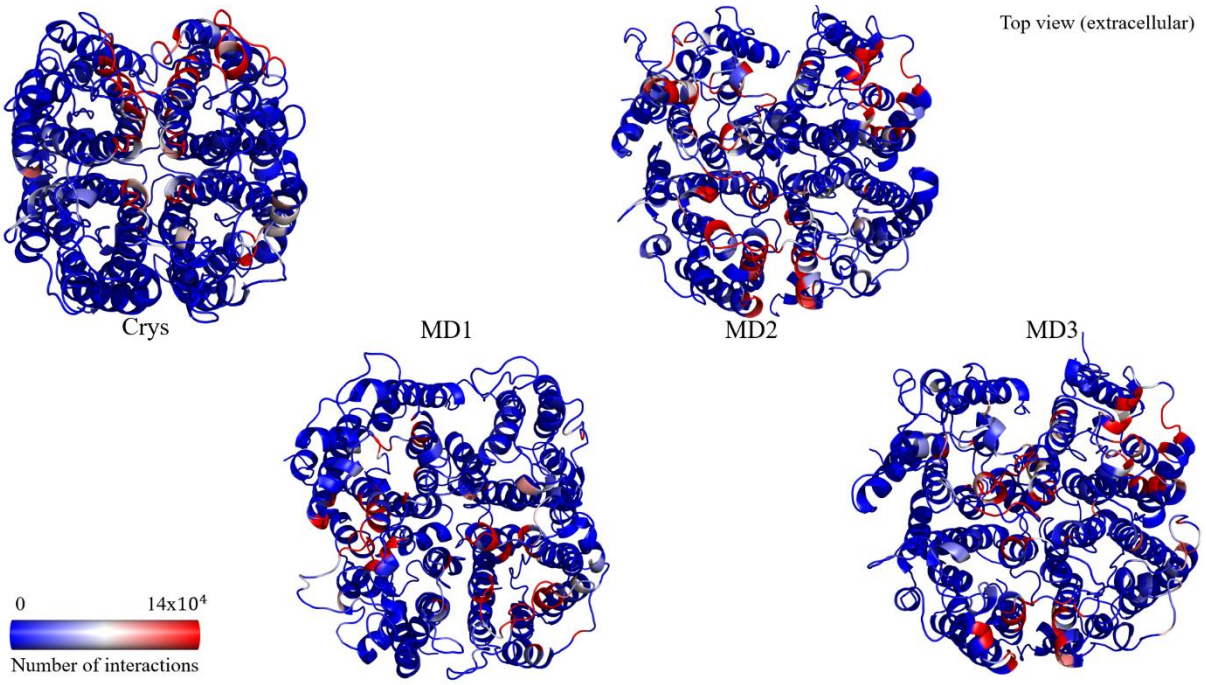

Figure 5.11. Top view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing the "pairwisedistances.py" Python script, with a maximum distance of 4.5 Å, highlighting the identified interactions with the blue_white_red Pymol spectrum.

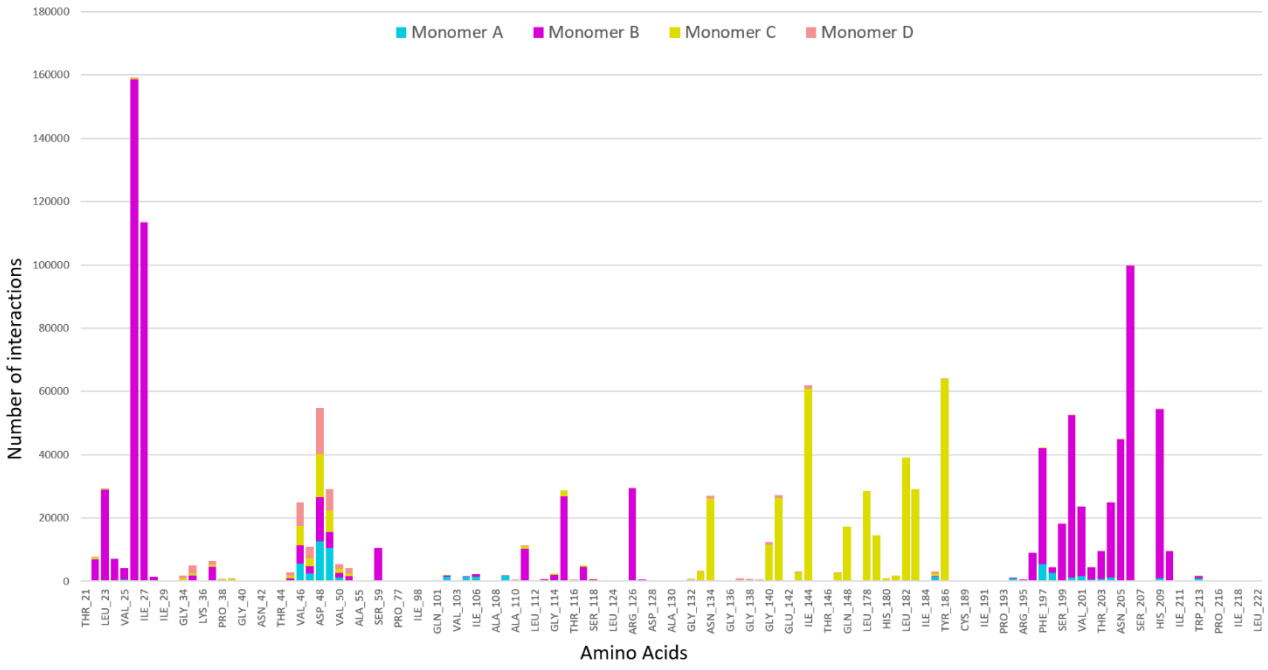

Figure 5.12. Identification of the population of interactions between the compound dataset and the extracellular side of the Crystallographic structure of AQP1.

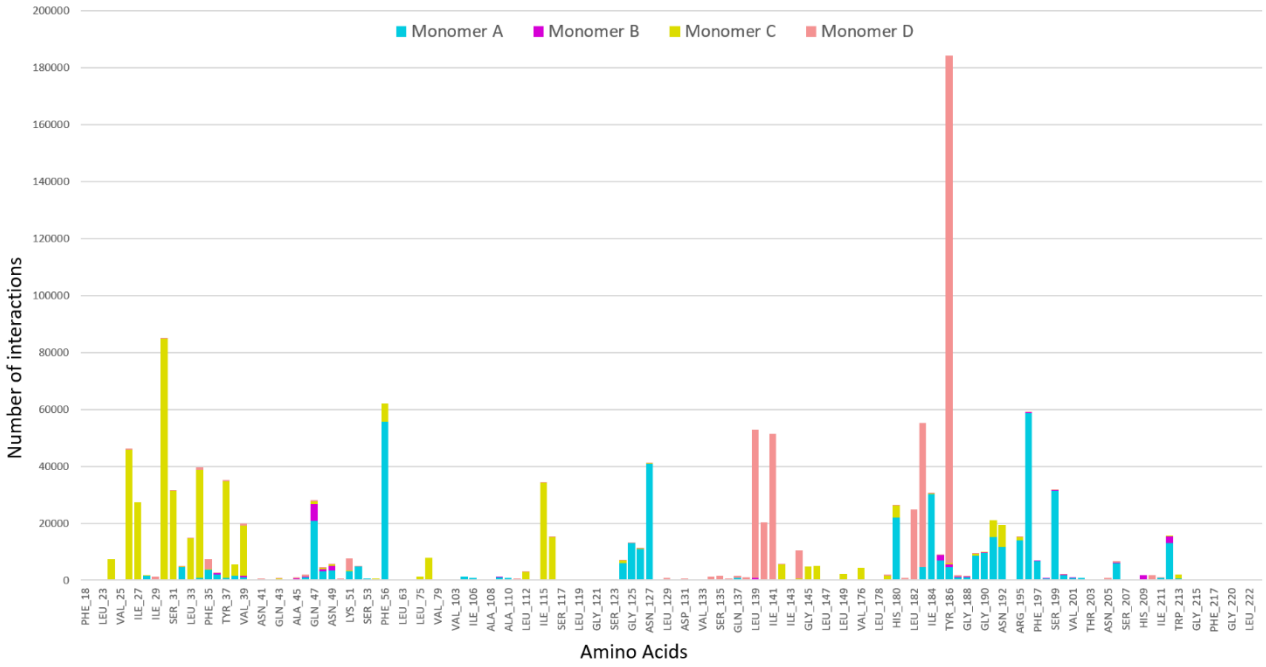

Figure 5.13. Identification of the population of interactions between the compound dataset and the extracellular side of the MD1 structure of AQP1.

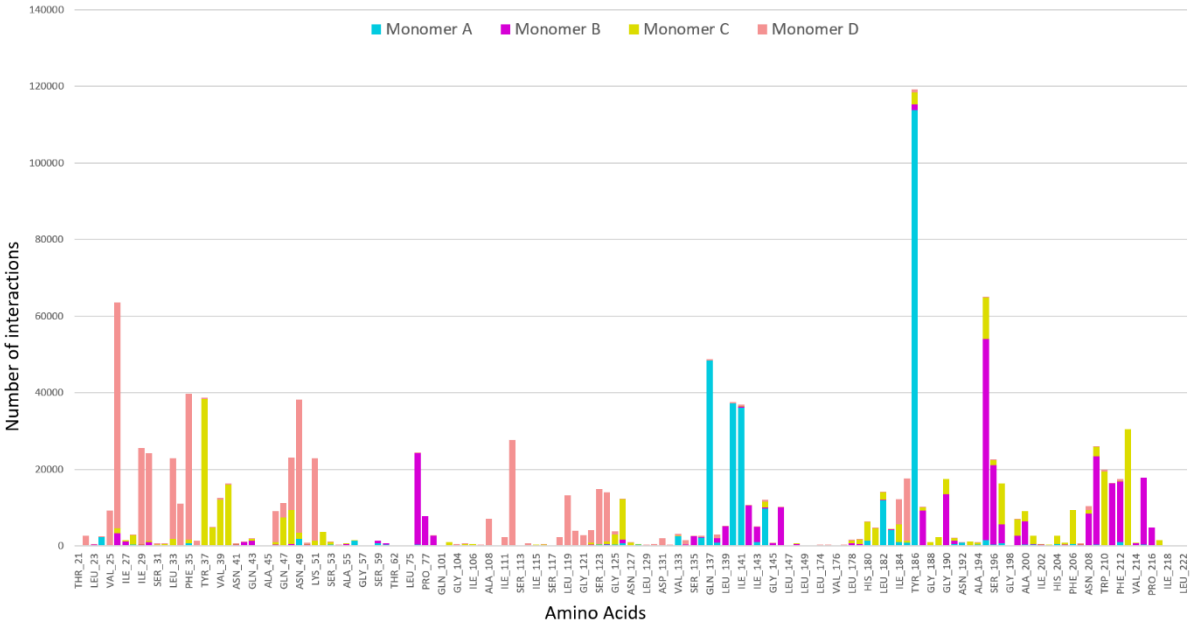

Figure 5.14. Identification of the population of interactions between the compound dataset and the extracellular side of the MD2 structure of AQP1.

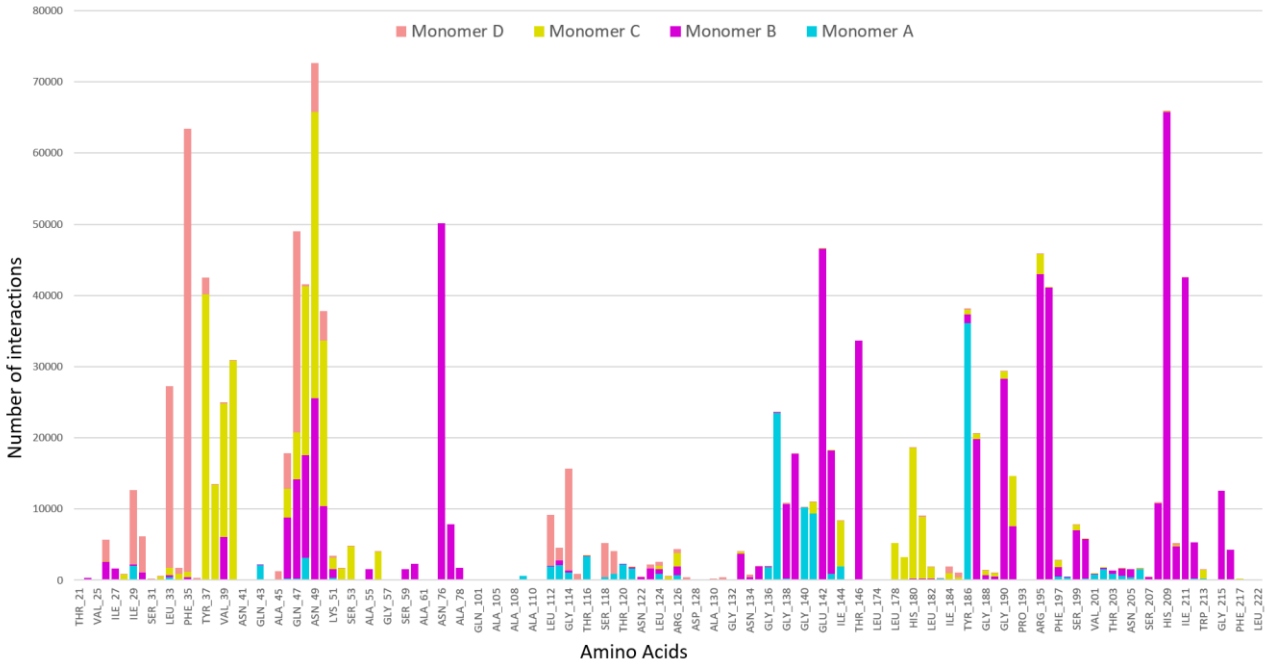

Figure 5.15. Identification of the population of interactions between the compound dataset and the extracellular side of the MD3 structure of AQP1.

The results obtained for the extracellular region are presented in Figures 5.11-5.15. For the Crystallographic structure, most interactions are located on both the protein's central region, namely at Monomer B. For the MD1 structure, Monomer A shows the highest number of interactions. In contrast, for the MD2 structure, most interactions are focused on Monomer B, C, and D in residues close to the access region of the pores. Regarding the MD3 structure, both Monomer B and C presented a higher number of interactions.

For the cytoplasmic region, the results are highlighted in Figures 5.16-5.20. For the Crystallographic structure, we could observe that the interactions occurred mainly on Monomer A, B, and D, while for the MD1 structure, Monomer D seems more prone to interactions. For the MD2 structure, a higher number of interactions are observed on both Monomer A and C, while for the MD3 structure, Monomer C and D are more populated.

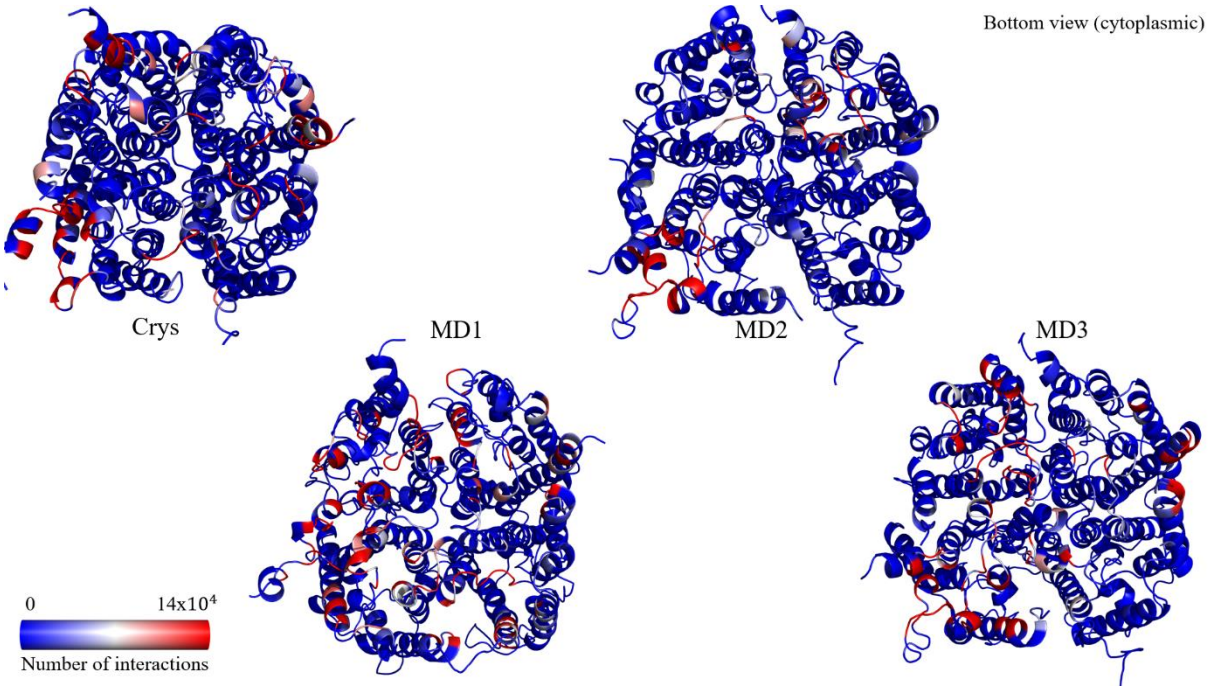

Figure 5.16. Bottom view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing the "pairwisedistances.py" Python script, with a maximum distance of 4.5 Å, highlighting the identified interactions with the blue_white_red Pymol spectrum.

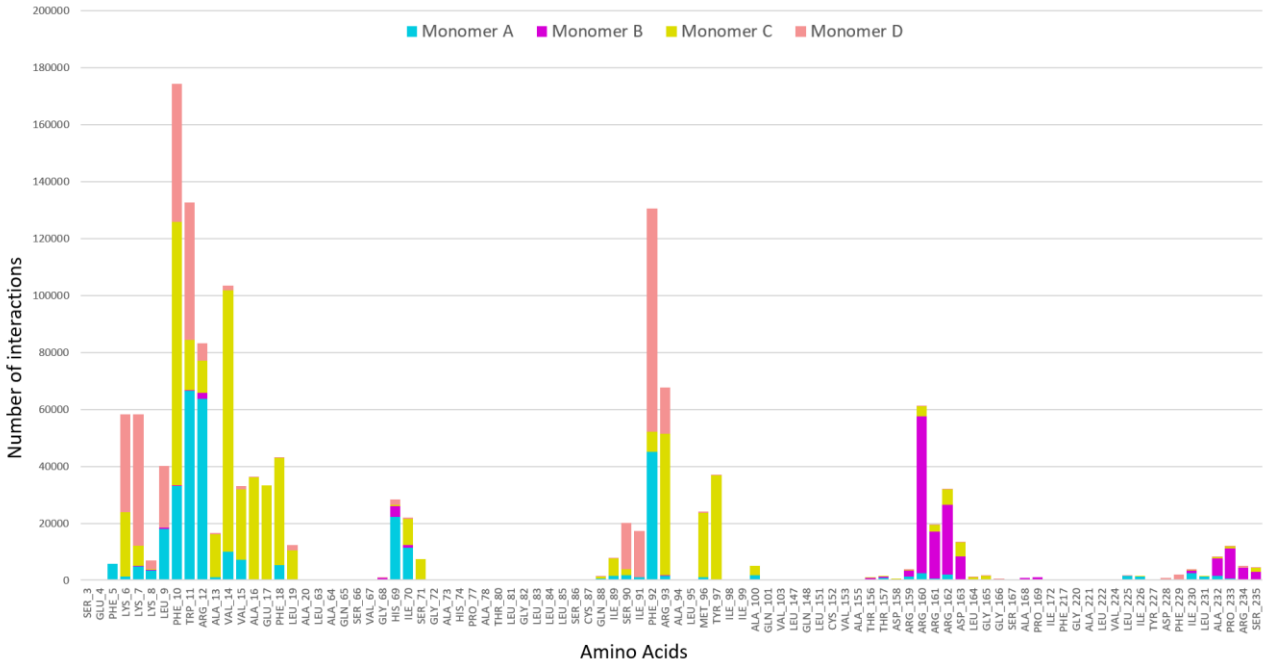

Figure 5.17. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the Crystallographic structure of AQP1.

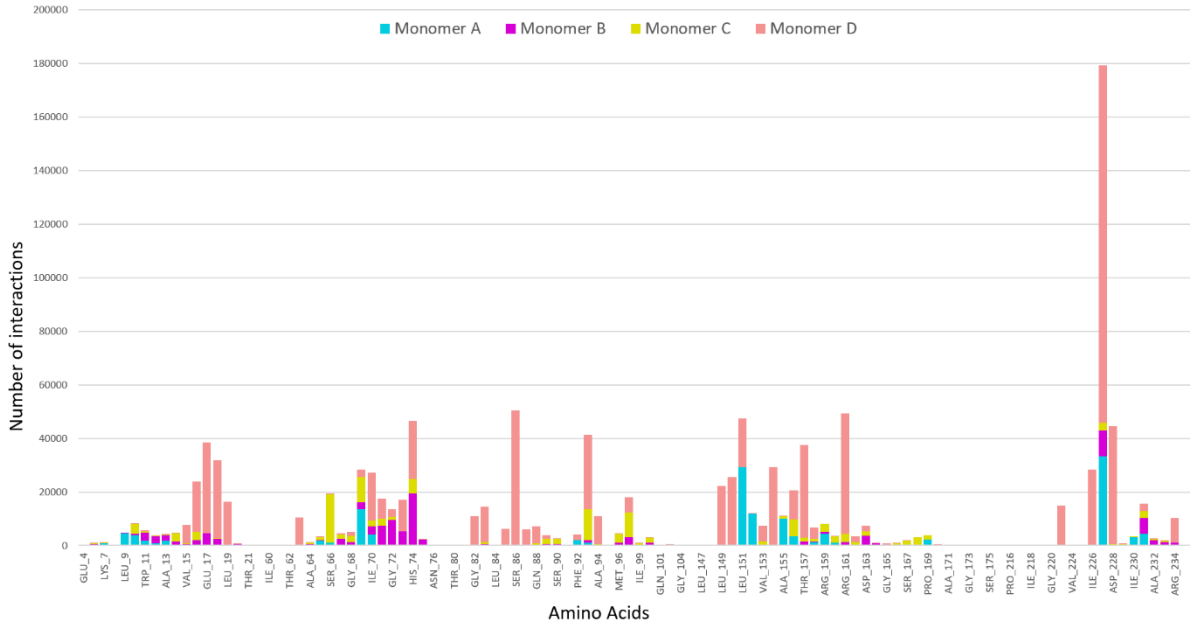

Figure 5.18. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the MD1 structure of AQP1.

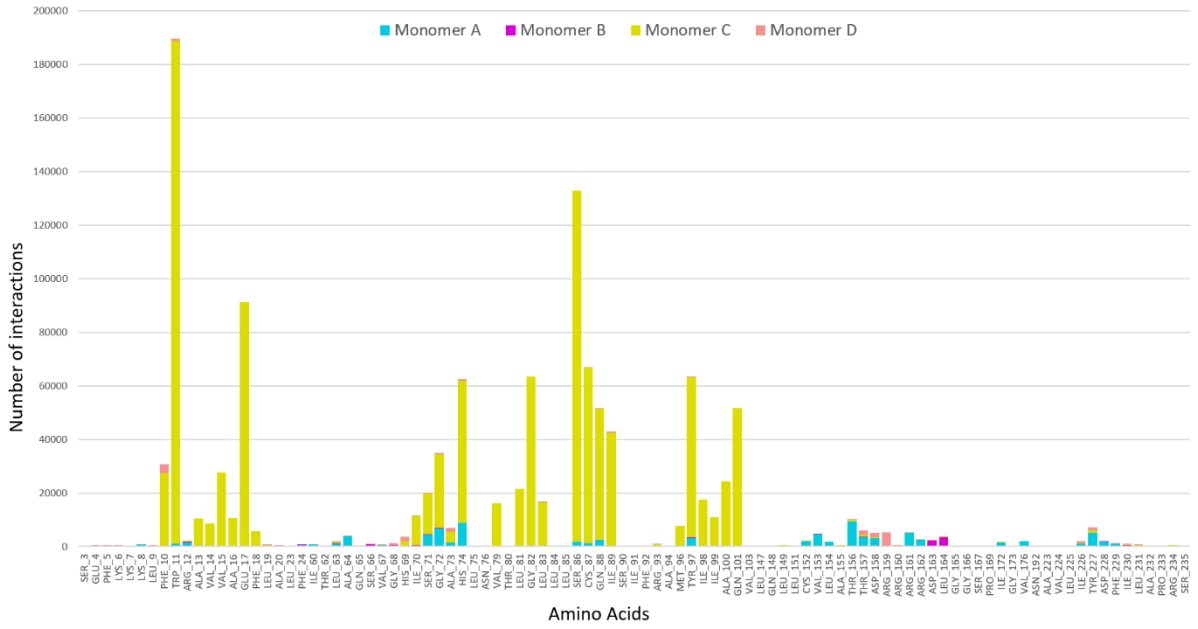

Figure 5.19. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the MD2 structure of AQP1.

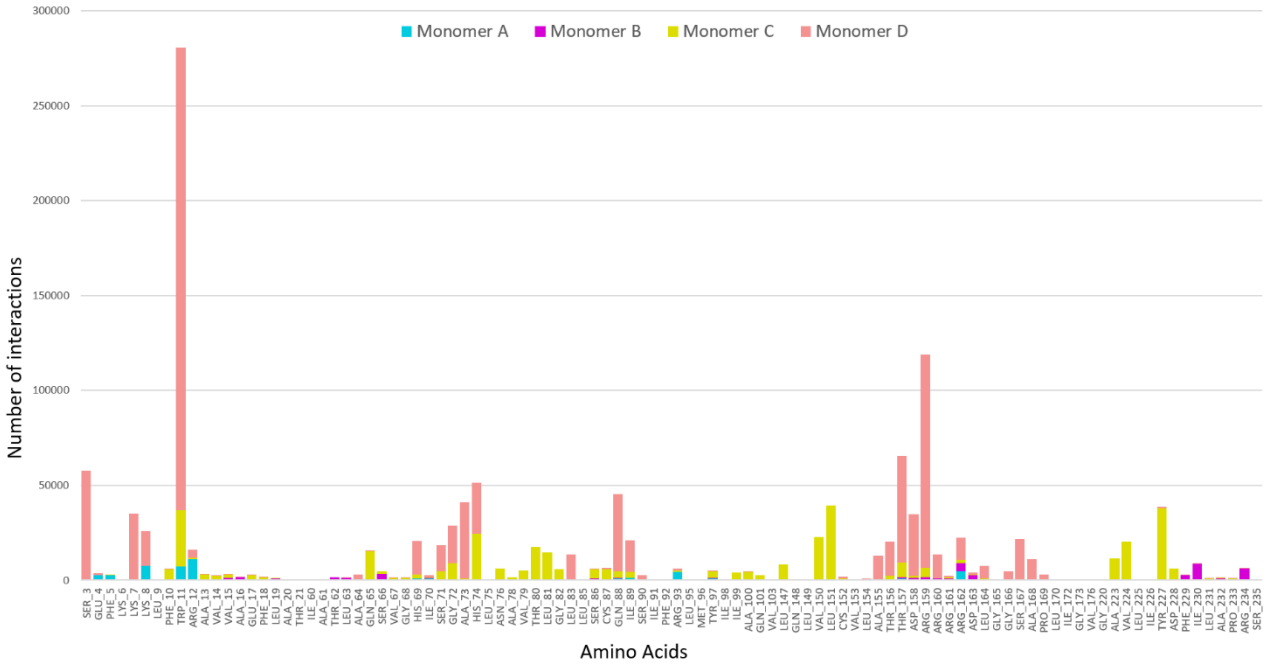

Figure 5.20. Identification of the population of interactions between the compound dataset and the cytoplasmic side of the MD3 structure of AQP1.

Considering all the previous results, we have identified three regions at the surface of this protein (B_NPA, C_ar/R, and C_NPA) as the most populated ones. The identified regions are located in two different Monomers of the MD2 confirmation, as this structure emerged as the most suitable one due to its overall higher number of interactions. The first region is located in Monomer B and is focused on the extracellular side, including the NPA motif and the region above (Figure 5.21). The other two regions are located in Monomer C. The first is also focused more on the extracellular side, including the ar/R selectivity filter and the region above (Figure 5.22). The second region is defined at the cytoplasmic site, including the NPA motif and the region below it (Figure 5.23).

It is crucial to refer that we started with a Docking approach that encompassed the whole AQP1 protein, with its four Monomers, on each side (extracellular and cytoplasmic), and that at this stage of the work, we have focused our studies on three specific regions of two Monomers. By using these regions, we are able to map all sites of interest and also take into account the conformational variability characteristics of these proteins, which has been shown to be crucial in the identification of new AQP1 modulators.

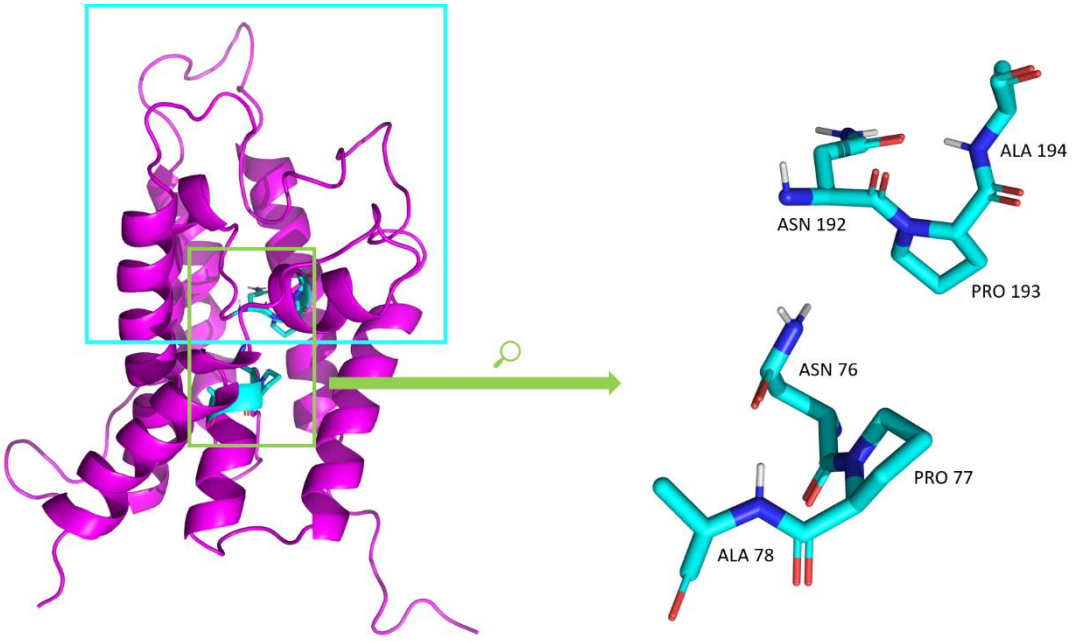

Figure 5.21. Monomer B of the MD2 structure of AQP1, highlighting the B_NPA binding site (blue square), and the NPA motif (green rectangle), with the constituent amino acids.

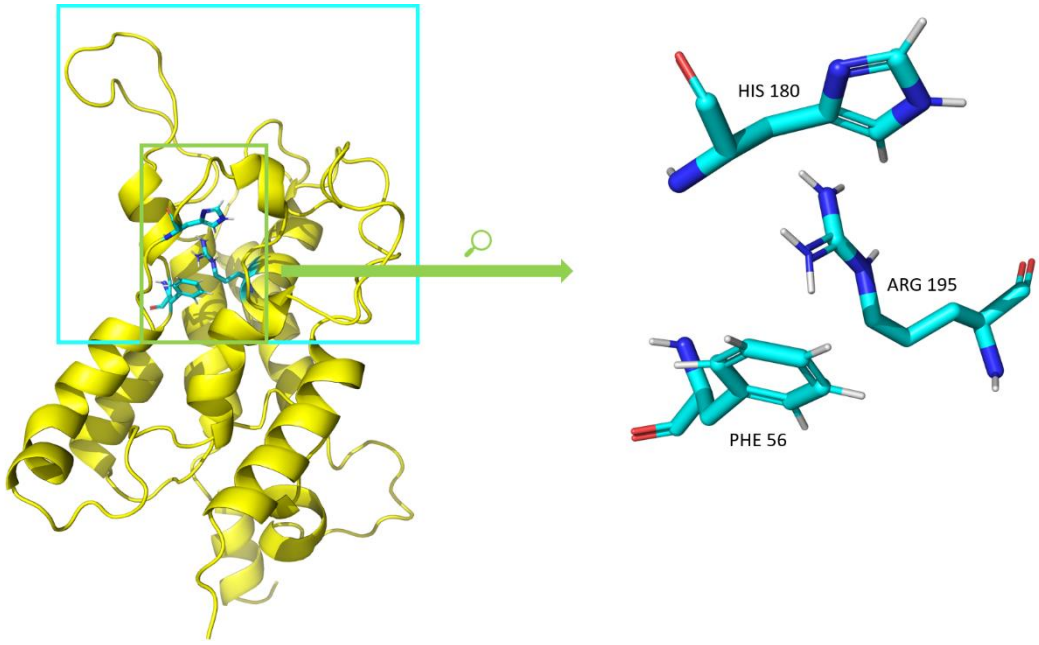

Figure 5.22. Monomer C of the MD2 structure of AQP1, highlighting the C_ar/R binding site (blue square), and the ar/R selectivity filter (green rectangle), with the constituent amino acids.

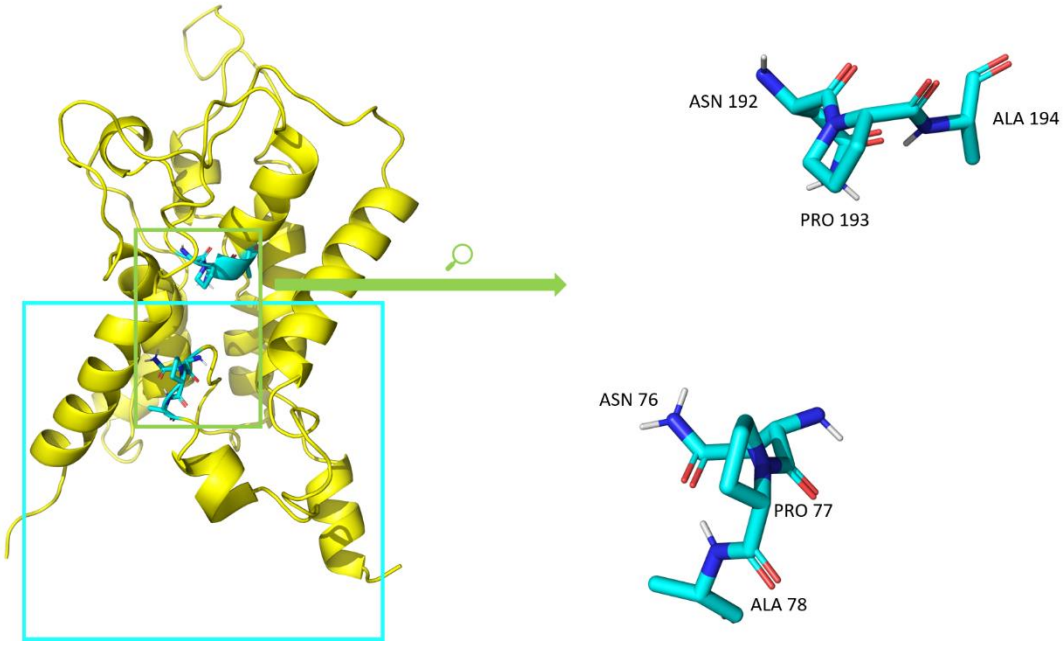

Figure 5.23. Monomer C of the MD2 structure of AQP1, highlighting the C_NPA binding site (blue square), and the NPA motif (green rectangle), with the constituent amino acids.

5.6. Molecular Docking screening to the regions and conformers of interest using different commercially available Compound Databases

After identifying the desired conformations of the AQP1 protein, and also the regions of the Monomers to focus on (B_NPA, C_ar/R, and C_NPA), we have performed the MD screenings using four different compound Databases: the one containing the 7204 ligands identified from the Ftrees approach and three commercially available ones (DrugBank, NCI, and ChEMBL).

As previously mentioned, to perform the different screenings, we have used VINA due to its speed and accuracy [139]. The results obtained from each dataset were sorted according to the VINA's scoring function. To further filter the compounds with undesired chemical properties, at this stage of the work, we have also applied Lipinski's rule of five filters. In the end, we ended up selecting the top 20 compounds for each of the binding sites: B_NPA (Table 5.1), C_ar/R (Table 5.2), and C_NPA (Table 5.3).

Overall, the best compounds identified in each binding region showed different ranges of binding affinities. For the B_NPA region, the obtained values ranged between -52.33 and -49.40 kJ/mol, for the C_ar/R region, the energies ranged between -45.64 and -43.54 kJ/mol, and finally, for the C_NPA region, the obtained values ranged between -53.59 and -50.24 kJ/mol. From these results, two conclusions can be immediately drawn: the compounds binding to the C_ar/R region show higher binding energy values when compared to the other two regions (B_NPA and C_NPA). This can mean that the compounds can bind tightly to the B_NPA and C_NPA than to the C_ar/R region. Another very interesting fact that we could observe was that the overwhelming majority of the selected compounds came from the ChEMBL Database. From the 60 selected compounds, only one was from a different Database (NCI 235813). We can understand this fact since the CHEMBL Database is the biggest Database used in this work with compounds with a good range of chemical diversity [160].

Table 5.1. Scoring function (Affinity (kJ/mol)) values calculated with VINA, for the top 20 compounds, for the B_NPA binding site.

B_NPA	
Ligand	Affinity (kJ/mol)
CHEMBL1806998	-52.33
CHEMBL1082826	-51.50
CHEMBL4090944	-51.50
CHEMBL2296092	-51.08
CHEMBL1083757	-50.66
CHEMBL2269285	-50.66
CHEMBL2322618	-50.66
CHEMBL246750	-50.66
CHEMBL3669378	-50.66
CHEMBL1714709	-50.24
CHEMBL3740177	-50.24
CHEMBL1374342	-49.82
CHEMBL246192	-49.82
CHEMBL3822940	-49.82
CHEMBL3933195	-49.82
CHEMBL3938690	-49.82
CHEMBL1383242	-49.40
CHEMBL1566748	-49.40
CHEMBL3809352	-49.40
CHEMBL4060285	-49.40

Table 5.2. Scoring function (Affinity (kJ/mol)) values calculated with VINA, for the top 20 compounds, for the C_ar/R binding site.

C_ar/R	
Ligand	Affinity (kJ/mol)
CHEMBL549017	-45.64
CHEMBL588059	-45.64
CHEMBL1765736	-44.80
CHEMBL3657567	-44.38
CHEMBL546902	-44.38
CHEMBL128477	-43.96
CHEMBL3288678	-43.96
CHEMBL3668659	-43.96
CHEMBL3758926	-43.96
CHEMBL3904599	-43.96
CHEMBL528300	-43.96
CHEMBL548389	-43.96
CHEMBL586861	-43.96
CHEMBL163764	-43.54
CHEMBL1765731	-43.54
CHEMBL1801152	-43.54
CHEMBL255641	-43.54
CHEMBL3770233	-43.54
CHEMBL528622	-43.54
CHEMBL586786	-43.54

Table 5.3. Scoring function (Affinity (kJ/mol)) values calculated with VINA, for the top 20 compounds, for the C_NPA binding site.

C_NPA	
Ligand	Affinity (kJ/mol)
CHEMBL1092882	-53.59
CHEMBL1092837	-52.75
CHEMBL1098578	-52.75
CHEMBL1089433	-52.33
CHEMBL3353215	-52.33
CHEMBL1277129	-51.50
CHEMBL3943580	-51.50
NCI235813	-51.50
CHEMBL464079	-51.08
CHEMBL1092838	-50.66
CHEMBL1644380	-50.66
CHEMBL2314047	-50.66
CHEMBL2374868	-50.66
CHEMBL3402066	-50.66
CHEMBL500565	-50.66
CHEMBL1277038	-50.24
CHEMBL1420518	-50.24
CHEMBL1974174	-50.24
CHEMBL3809352	-50.24
CHEMBL511940	-50.24

To structurally characterize, in detail, the interaction of all the 60 compounds with the binding sites of the protein, we have used the previously described pymol script applying a maximum cut-off distance of 4.5 Å. In Figures 5.24, 5.26, and 5.28, we plotted the number of Ligand-Protein interactions against the amino acids with the highest number of interactions. In Figures 5.25, 5.27, and 5.29, we have a structural analysis of the three binding pockets, B_NPA, C_ar/R, and C_NPA, respectively, highlighting the top 15 residues with the highest number of interactions.

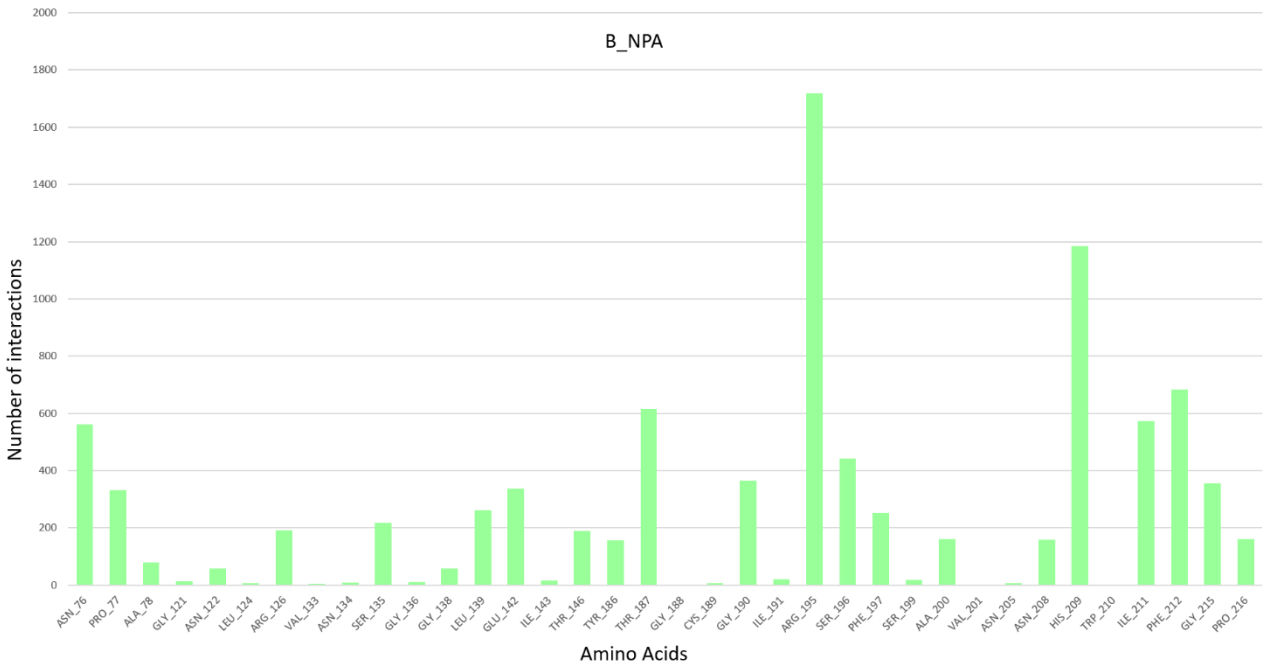

Figure 5.24. Identification of the population of interactions between the compound dataset and the B_NPA binding site.

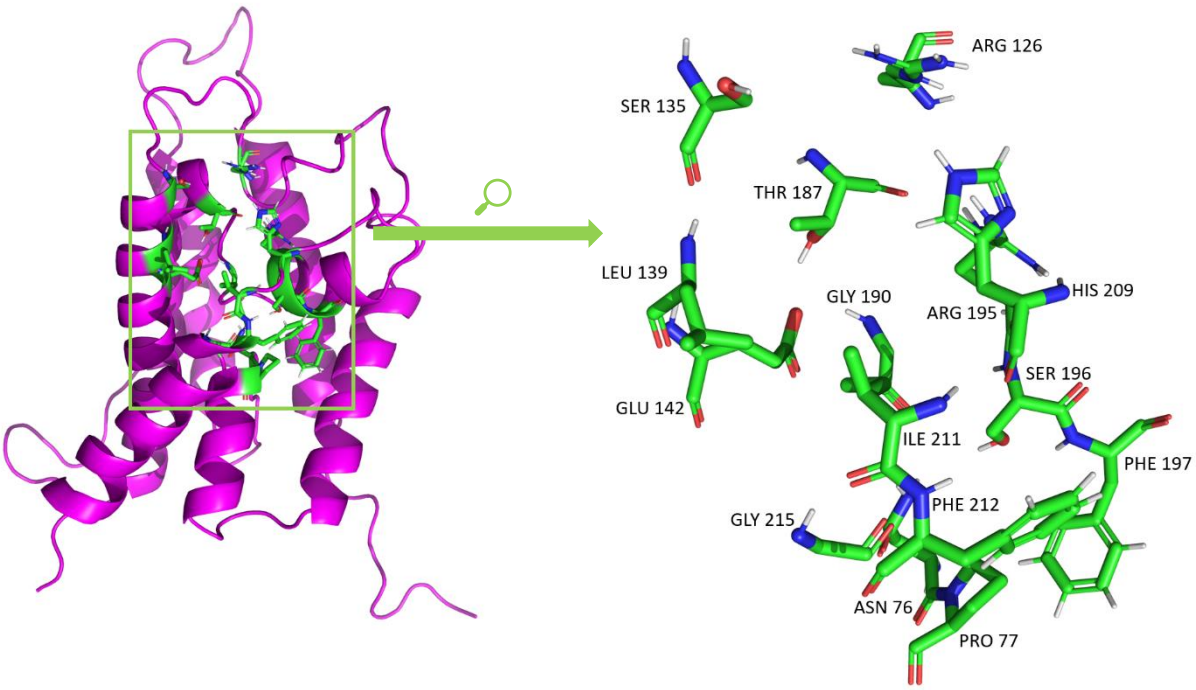

Figure 5.25. Monomer B of the MD2 structure of AQP1, highlighting the B_NPA binding pocket (green rectangle), with the constituent amino acids.

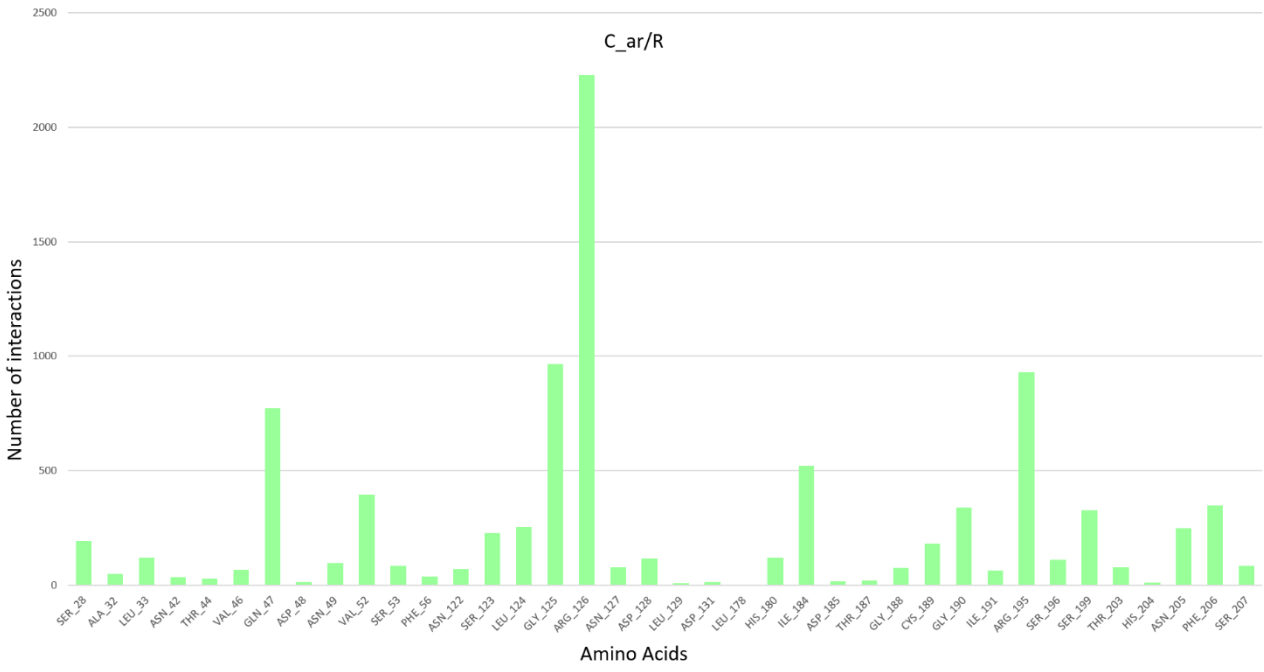

Figure 5.26. Identification of the population of interactions between the compound dataset and the C_ar/R binding site.

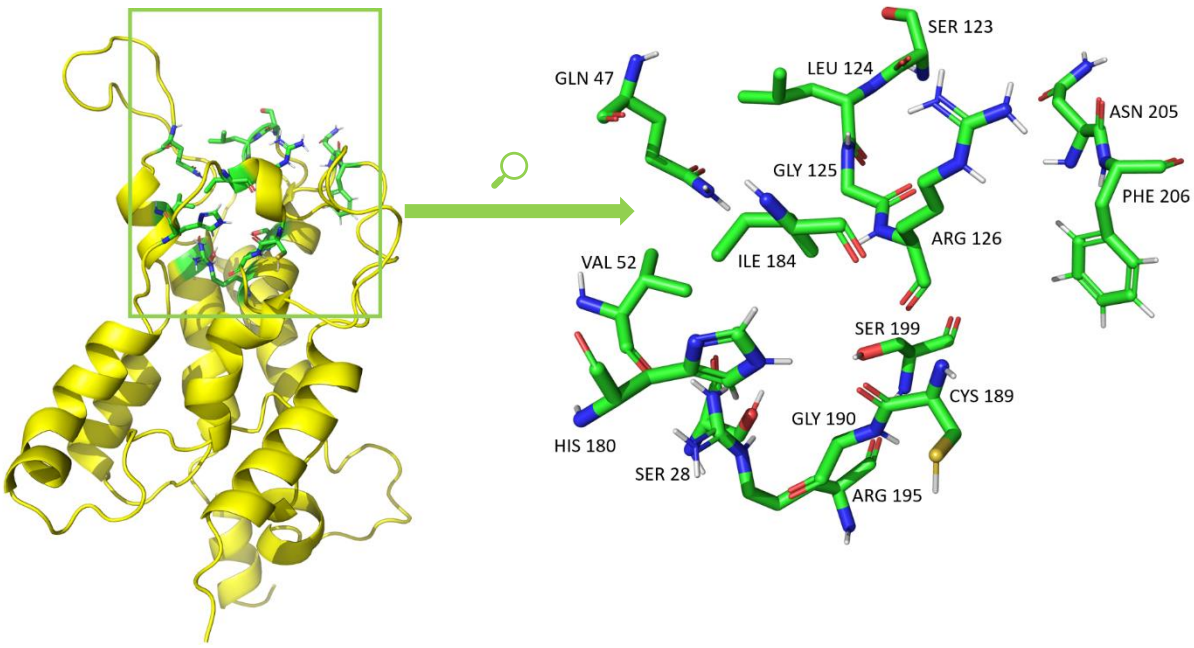

Figure 5.27. Monomer C of the MD2 structure of AQP1, highlighting the C_ar/R binding pocket (green rectangle), with the constituent amino acids.

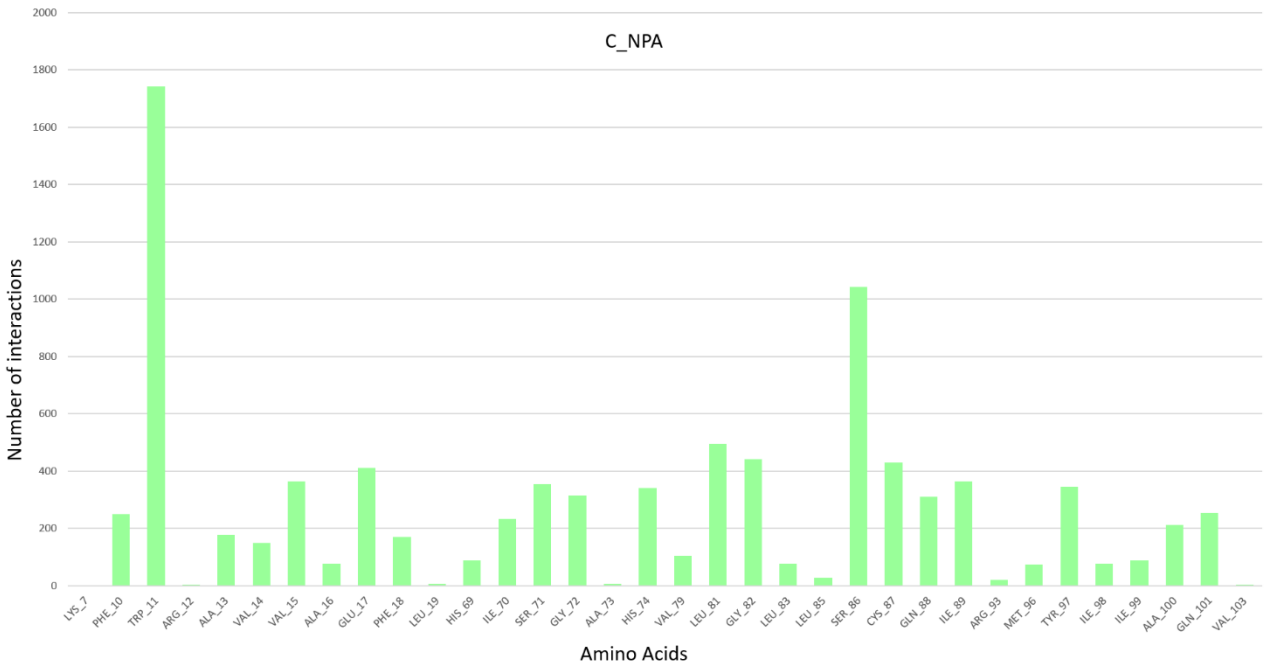

Figure 5.28. Identification of the population of interactions between the compound dataset and the C_NPA binding site.

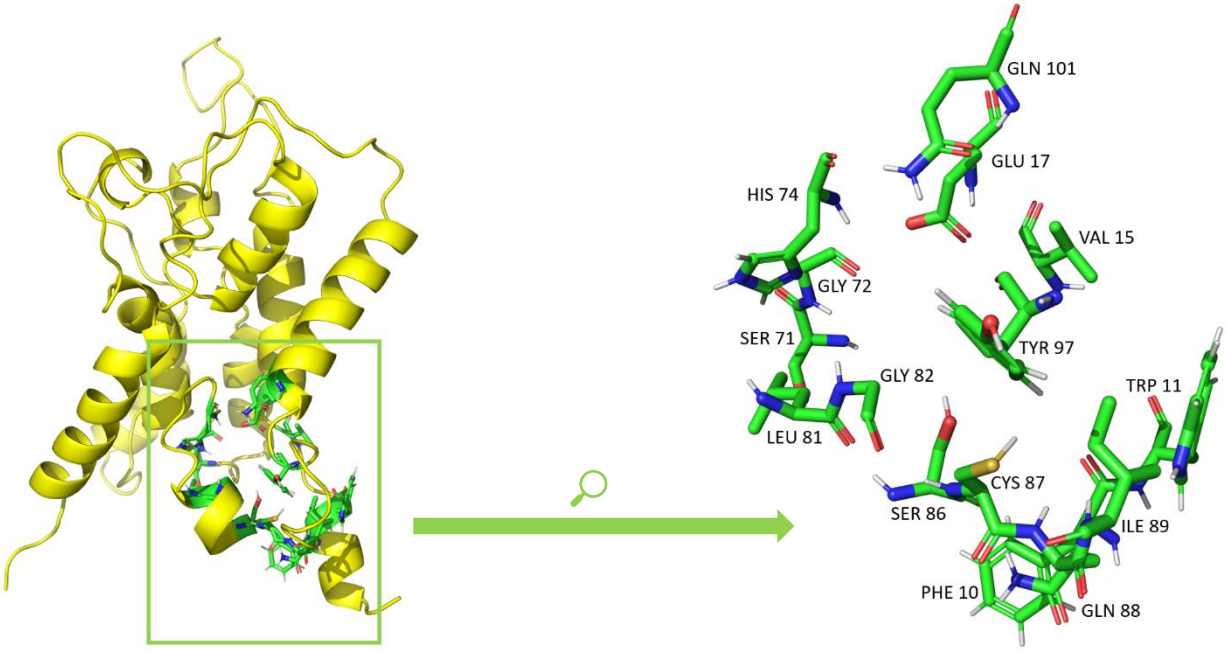

Figure 5.29. Monomer C of the MD2 structure of AQP1, highlighting the C_NPA binding pocket (green rectangle), with the constituent amino acids.

According to these results, we can easily observe that some amino acids are clear hits for each of the binding sites, evidencing their importance in the interaction with different compounds. For the B_NPA binding site, two amino acids stand out due to the high number of identified interactions: Arg 195 and His 209 (Figure 5.24). Additionally, residues such as Asn 76, Pro 77, Thr 187, Gly 190, Ser 196, Ile 211, Phe 212, and Gly 215 participate in the interaction of many of the selected compounds. It is important to mention that the Arg 195 residue is one of the ar/R selectivity filter components and that Asn 76 and Pro 77 residues are included in the NPA motif, as highlighted in Figure 5.21. The chemistry of the amino acids that compose the protein's binding pockets is critical to Ligand-Protein interactions, since the amino acids' side chains can bond with the ligand to hold an interaction that might promote a strong binding of the ligand to a specific pocket of the protein. Charged amino acids (Arg, Asp, Glu, His, and Lys) can form ionic bonds, polar amino acids (Asn, Cys, Gln, Ser, and Thr) are capable of forming hydrogen bonds and lastly, hydrophobic side chains (Ala, Gly, Ile, Leu, Met, Phe, Pro, Trp, Tyr, and Val) interact via weak van der Waals interactions. Based on the analysis of the results obtained for this binding region, we can conclude that although there are regions of this binding site more apolar, the most important regions that allow the establishment of ionic or hydrogen bonds, and that at the same time confer specificity to the binding, come from residues such as Arg 195 and His 209 (Figure 5.25).

Regarding the results obtained for the C_ar/R binding region (Figure 5.26), we can observe that Arg 126 easily stands out from all the other residues. Some other residues also show to be highly involved in the compound interactions with the selected compounds: Gln 47, Gly 125, Ile 184, and Arg 195. Arg 195 residue is one of the components of the ar/R selectivity filter and has also shown to be quite important in the binding of compounds since it was one of the most used residues in interactions in the B_NPA binding region. Due to its positively charged character, and hydrogen bond donor characteristics, this residue stands out as one of the most relevant interaction points atoms of the compounds with hydrogen bond acceptor and/or negative charge characteristics. Interestingly, the other two residues

that compose the selectivity filter (Phe 56 and His 180) do not seem to be as important to the compounds' interaction as Arg 195 (Figure 5.26 and Figure 5.27).

For the C_NPA binding site, two amino acids stand out: Trp 11 and Ser 86 (Figure 5.28). Some other residues also present high interaction values, like Glu 17, Leu 81, Gly 82, and Cys 87. None of these residues (see Figure 5.29) are members of either the NPA motif or the ar/R selectivity filter. However, the Cys 87 residue has been reported to be involved in the inhibition of water permeability through the AQP1 pore due to its sensitivity to mercury [161]. Acknowledging the two amino acids with the highest number of interactions, we can conclude that the majority of interactions observed in the C_NPA pocket are of hydrophobic nature. This is natural since this Trp residue is in the more hydrophobic region of the pore's cytoplasmatic side. Facing the region of more solvent exposure, we also find Ser 86, a residue with high capabilities of hydrogen bond donor interactions with compounds that bind to the protein from the cytoplasmic side (Figure 5.29).

5.7. MM/PBSA calculations

A quantitative knowledge of protein-ligand binding affinities is essential in understanding molecular recognition. Therefore, efficient and accurate binding free energy calculations are a central goal in computer-aided drug design. Molecular Docking methods use some simplistic but fast conformational search for complex solutions, but with a low degree of accuracy. These methods are commonly used for discriminating binders and nonbinders within large chemical Databases, but they typically lack accuracy in quantitatively ranking and predicting $\Delta G_{binding}$ values. Therefore, in order to further characterize the top binders of AQP1 in the different selected binding regions, we decided to use alternate end-point methods offering a good compromise of effectiveness and efficiency in $\Delta G_{binding}$. Frequently applied end-point methods make use of linear interaction energy (LIE) theory and the MM/PBSA approach. This latter approach can easily be performed using results from protein-ligand complex MD simulations. Therefore, the top 20 docked poses, for each binding site, were submitted to MD simulations with the first objective of accessing the stability, explicit solvation, conformational adaptation of the protein, and reliability of the docked poses at the atomic level [38].

In Figure 5.30, we can observe the RMSD variation through simulation time (30 ns) of five compounds docked at the B_NPA site (CHEMBL1383242, CHEMBL246750, CHEMBL3669378, CHEMBL3809352, and CHEMBL4060285). As can be seen, the majority of the compounds reached a stable overall RMSD value (Figure 5.30) after 5 ns of simulation. Except for compound CHEMBL246750, all the other compounds showed stable conformational sampling at the binding site of AQP1. Although the CHEMBL246750 compound changed its configuration after approximately 20 ns of simulation, it has reached a plateau, another stable configurational plateau. It is noteworthy to mention here that all the selected 60 compounds are structurally diverse regarding their structural flexibility (Tables A.1-A.3 and Figures A.17-A.19). Still, independently of these characteristics, the compounds were relatively stable at each one of the binding sites, allowing us to have a good conformational sampling to use in MM/PBSA binding energies calculations [38]. However, not all compounds showed to be stable at the AQP1 binding site after 30 ns of simulation. In these cases, the simulation time was extended until a stable conformational dynamic was achieved. Due to this fact, we had the need to ran MD simulations for four different time intervals: 36 of the 60 compounds ran for 10 ns, 15 of them ran for 30 ns, 7 ran for 50 ns, and finally, 2 ran for 100 ns (Tables 5.4-5.6).

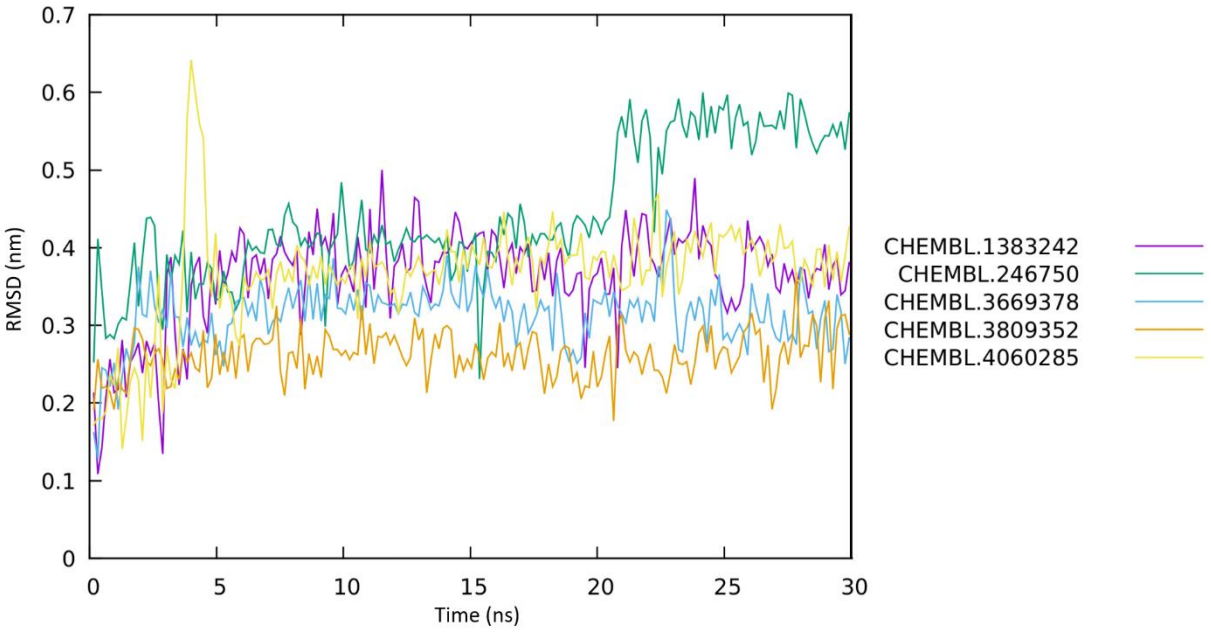

Figure 5.30. Representation of the RMSD (nm) variation through simulation time (30 ns) of five compounds bound at the B_NPA binding site.

Figures A.7-A.9 (Appendix) show the RMSD variation throughout three different simulation times (10 ns, 50 ns, and 100 ns, respectively) of the compounds docket at the B_NPA binding site. Figures A.10-A.12 represent the RMSD variation throughout three different simulation times (10 ns, 30 ns, and 50 ns, respectively) of the compounds docket at the C_ar/R binding site. Lastly, in Figures A.13-A.16, we can observe the RMSD variation throughout four different simulation times (10 ns, 30 ns, 50 ns, and 100 ns, respectively) of the compounds docket at the C_NPA binding site. As can be seen, all compounds, after the different used simulation times, reached a stable plateau, indicating, therefore, the sampling of their equilibration states. The presented results suggest that these ligands did not dissociate from the protein binding site, remaining bound at the protein binding site in a stable configuration throughout the simulation.

Once the RMSD value of the compounds reached a stable configuration, regardless of the simulation time, 20 snapshots of the simulated system were taken to use on the binding free energy calculations, using the MM/PBSA method. To perform such calculations, we have used the g_mmpbsa tool for GROMACS (Tables 5.4, 5.5, and 5.6).

If we analyze the required simulation time to achieve stable conformations in each one of the evaluated complexes, we can conclude that the simulation time (10 ns, 30ns, 50ns, or 100ns) is not correlated to a specific higher or lower Binding Energy. Complexes that presented the highest Binding Energy values were associated with both high and low simulation times, and the same can be observed for complexes with the lowest Binding Energy values.

Table 5.4. Binding Energy (kJ/mol) values calculated with g_mmpbsa, for the top 20 compounds, for the B_NPA binding site. Different simulation times are highlighted with different coloring (10 ns, 30 ns, 50 ns, and 100 ns).

B_NPA	
Ligand	Binding Energy (kJ/mol)
CHEMBL4090944	-167.93
CHEMBL3740177	-160.33
CHEMBL1083757	-151.89
CHEMBL1383242	-151.66
CHEMBL4060285	-147.96
CHEMBL2269285	-142.53
CHEMBL3938690	-135.34
CHEMBL246750	-134.44
CHEMBL3933195	-131.68
CHEMBL1082826	-131.17
CHEMBL2322618	-125.85
CHEMBL1806998	-125.00
CHEMBL1714709	-121.82
CHEMBL3669378	-117.82
CHEMBL1566748	-117.64
CHEMBL246192	-117.63
CHEMBL3809352	-114.82
CHEMBL1374342	-113.70
CHEMBL3822940	-110.76
CHEMBL2296092	-105.78

	10 ns
	30 ns
	50 ns
	100 ns

Table 5.5. Binding Energy (kJ/mol) values calculated with g_mmpbsa, for the top 20 compounds, for the C_ar/R binding site. Different simulation times are highlighted with different coloring (10 ns, 30 ns, and 50 ns).

C_ar/R	
Ligand	Binding Energy (kJ/mol)
CHEMBL255641	-140.57
CHEMBL3904599	-124.59
CHEMBL3668659	-117.07
CHEMBL128477	-116.29
CHEMBL1801152	-105.39
CHEMBL1765731	-104.48
CHEMBL3770233	-102.71
CHEMBL3758926	-96.14
CHEMBL549017	-90.36
CHEMBL586861	-83.76
CHEMBL3288678	-83.00
CHEMBL548389	-81.08
CHEMBL588059	-77.95
CHEMBL1765736	-76.93
CHEMBL586786	-74.62
CHEMBL163764	-70.71
CHEMBL546902	-67.26
CHEMBL528622	-60.94
CHEMBL528300	-54.70
CHEMBL3657567	-13.02

	10 ns
	30 ns
	50 ns

Table 5.6. Binding Energy (kJ/mol) values calculated with g_mmpbsa, for the top 20 compounds, for the C_NPA binding site. Different simulation times are highlighted with different coloring (10 ns, 30 ns, 50 ns, and 100 ns).

C_NPA	
Ligand	Binding Energy (kJ/mol)
CHEMBL500565	-158.40
CHEMBL2374868	-149.72
CHEMBL464079	-135.39
CHEMBL2314047	-132.34
CHEMBL1092882	-128.14
CHEMBL3943580	-126.49
CHEMBL3353215	-123.34
CHEMBL1092837	-122.13
CHEMBL1092838	-120.64
CHEMBL1089433	-119.54
CHEMBL511940	-107.32
CHEMBL1277038	-106.62
CHEMBL1098578	-103.81
CHEMBL1420518	-103.59
CHEMBL1277129	-100.45
CHEMBL3809352	-94.33
NCI235813	-88.70
CHEMBL1974174	-88.12
CHEMBL1644380	-75.05
CHEMBL3402066	34.28

	10 ns
	30 ns
	50 ns
	100 ns

By analyzing Tables 5.4 to 5.6, we can see that the top 20 compounds for each binding region had Binding Energy (kJ/mol) values ranging from: -167.932 to -105.784 kJ/mol for the B_NPA binding site; -140.567 to -13.018 kJ/mol for the C_ar/R binding site; and -158.401 to 34.275 kJ/mol for the C_NPA binding site. The compounds displaying the lowest free energy binding values for each binding site were: CHEMBL4090944 for the B_NPA binding site, with a free energy value of -167.932 kJ/mol; CHEMBL255641 for the C_ar/R binding site, with a free energy value of -140.567 kJ/mol; and lastly, CHEMBL500565 for the C_NPA binding site, with a free energy value of -158.401 kJ/mol. Interestingly, the range of binding energies observed for the top 20 results of the B_NPA is the lowest one, indicating, therefore, that all the solutions show good binding energies for this protein's region. The same cannot be said for the results obtained for the C_NPA region. As can be seen in table 5.6, although the top binding compound shows a high binding affinity to this region of the protein, the 20th top solution shows a positive value for the binding affinity, indicating an unfavored binding of this compound for this binding site.

We have also tried to see if a correlation between VINA's binding affinities and the MM/PBSA approach would be observed. As reported by others [162–164], the compounds' scoring and sorting according to their affinity to the different binding sites using Molecular Docking approaches and MM/PBSA methods is divergent. In our results, we could identify compounds that in the Molecular Docking results stood out due to their lower binding score (top results), while in the MM/PBSA results, they were placed on the bottom of the list as the worst binding compounds. Since the Molecular Docking scores and binding free energies of all top 20 poses are not correlated, we decided to focus our attention on the top scorer ligands identified from the MM/PBSA calculations. Binding free energy values, calculated with and MM/PBSA method, rely on a dynamic trajectory, with several snapshots of the ligand-protein complex. On the other hand, the Affinity values calculated with a Docking approach are focused on a stationary pose, which can be seen as a single snapshot. Having this in mind, MM/PBSA calculations provide us more reliable results regarding binding values, which can better mimic the actual ligand-

protein complex interaction. MM/PBSA has already been compared with other ligand-binding methods, and typically its accuracy is better when compared to a Docking approach [129].

Having this in mind, Tables 5.7-5.9 have a more in-depth description of the Binding Energy values, regarding the four components that can be extracted from the MM/PBSA calculations: van der Wall Energy; Electrostatic Energy; Polar Solvation Energy; and SASA Energy values. This description was made only for the top 5 compounds, per binding pocket: B_NPA (Table 5.7); C_ar/R (Table 5.8); and C_NPA (Table 5.9).

Table 5.7. Binding Energy, Van der Wall Energy, Electrostatic Energy, Polar Solvation Energy, and SASA Energy values, for the top 5 compounds, based on the MM/PBSA ranking, for the B_NPA binding site.

B_NPA					
Ligand	Binding energy (kJ/mol)	Van der Waals energy (kJ/mol)	Electrostatic energy (kJ/mol)	Polar solvation energy (kJ/mol)	SASA energy (kJ/mol)
CHEMBL4090944	-167.93	-251.70	-17.68	123.44	-22.00
CHEMBL3740177	-160.33	-239.74	-13.45	117.20	-24.34
CHEMBL1083757	-151.89	-215.92	-9.79	95.34	-21.52
CHEMBL1383242	-151.66	-227.81	-24.45	124.11	-23.52
CHEMBL4060285	-147.96	-225.14	-5.46	104.22	-21.58

Table 5.8. Binding Energy, Van der Wall Energy, Electrostatic Energy, Polar Solvation Energy, and SASA Energy values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_ar/R binding site.

C_ar/R					
Ligand	Binding energy (kJ/mol)	Van der Waals energy (kJ/mol)	Electrostatic energy (kJ/mol)	Polar solvation energy (kJ/mol)	SASA energy (kJ/mol)
CHEMBL255641	-140.57	-234.10	-38.16	153.11	-21.42
CHEMBL3904599	-124.59	-215.81	-27.84	141.10	-22.04
CHEMBL3668659	-117.07	-257.92	-22.06	189.95	-27.03
CHEMBL128477	-116.29	-211.44	-53.51	170.66	-22.00
CHEMBL1801152	-105.39	-212.82	-35.03	164.11	-21.65

Table 5.9. Binding Energy, Van der Wall Energy, Electrostatic Energy, Polar Solvation Energy, and SASA Energy values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_NPA binding site.

C_NPA					
Ligand	Binding energy (kJ/mol)	Van der Waals energy (kJ/mol)	Electrostatic energy (kJ/mol)	Polar solvation energy (kJ/mol)	SASA energy (kJ/mol)
CHEMBL500565	-158.40	-234.01	-43.92	141.11	-21.57
CHEMBL2374868	-149.72	-186.31	-6.72	61.34	-18.02
CHEMBL464079	-135.39	-210.73	-30.82	127.47	-21.32
CHEMBL2314047	-132.34	-204.26	-18.61	113.30	-22.76
CHEMBL1092882	-128.14	-208.75	-18.86	121.75	-22.28

Overall, the most significant contribution to the binding energy of all compounds listed in tables 5.7 to 5.9 comes from the van der Waals energy. Although some small disparity can be found in the different components of the MM/PBSA calculations, due to small bridges of favorable interactions, which ends up reflecting in a better-ranked compound, we can conclude that the 15 selected ligands (5 per binding site) have very similar Binding Energy values and similar values for the four different terms that compose the Binding value, having the same order of magnitude.

Lastly, in order to further structurally characterize the interaction of the top compound, per binding pocket, we have again used the pymol script previously described applying a maximum cut-off distance of 4.5 Å. In Figures 5.31, 5.33, and 5.35, we present the Ligand-Protein interactions plotted in a bar graph for each of the used binding sites, where the number of interactions is plotted against the amino acids interacting with the used compound. Furthermore, in Figures 5.32, 5.34, and 5.36, we have a structural analysis of the top compound for each binding site, B_NPA, C_ar/R, and C_NPA, respectively, highlighting the top 8 or 9 residues with the highest number of interactions with the top compounds.

CHEMBL4090944, the compound showing higher stability to the B_NPA binding site of AQP1, adopted a favored conformation where it was preferably interacting with five amino acids: Arg 195, Ser 199, Asn 208, His 209, and Ile 211 (Figure 5.31). As shown in Figure 5.32, Arg 195 and His 209 seem to be performing a stable pi-pi interaction with the two rings of the compound at the top of the binding site. Additionally, the long carbon chain occupying the deeper region of this binding site, do seem to perform some stable interactions with Asn 76 and Ser 199. CHEMBL4090944 establishes an H-bond with Ser 199 at a distance of 2.7 Å (Figure 5.32).

Regarding the top result of the C_ar/R binding region, CHEMBL255641 compound (Figure 5.33) showed a higher preference to interact with Gln 47, Gly 125, Arg 126, and Ser 199 amino acids. Similar to the CHEMBL4090944 compound, Arg 126 seems to be promoting stable interactions, but in this case of more ionic characteristics. Additionally, several nitrogen atoms in the compound perform hydrogen bonds with Gln 47 and Ser 199 and hydrophobic interactions with Gly 125. CHEMBL255641 additionally establishes several H-bonds with Leu 124, Arg 126, and Ser 199, at a distance of 2.5, 2.1, and 3.0 Å, respectively (Figure 5.34).

Lastly, for the C_NPA binding site, compound CHEMBL500565 establishes several preferred interactions with four amino acids found in this region of the protein: Ala 16, Glu 17, Ser 86, and Ala 100 (Figure 5.35). Some ionic interactions between Glu 17 and this compound are observed. Also, hydrogen bonds with Ser 86 and hydrophobic interactions with Ala 16 and Ala 100 seem to contribute to an overall stable binding configuration (Figure 5.36). CHEMBL500565 additionally establishes an H-bonds with Ala 100, at a distance of 1.9 Å, and two H-bond with Gln 101, at 1.9 and 2.2 Å (Figure 5.36).

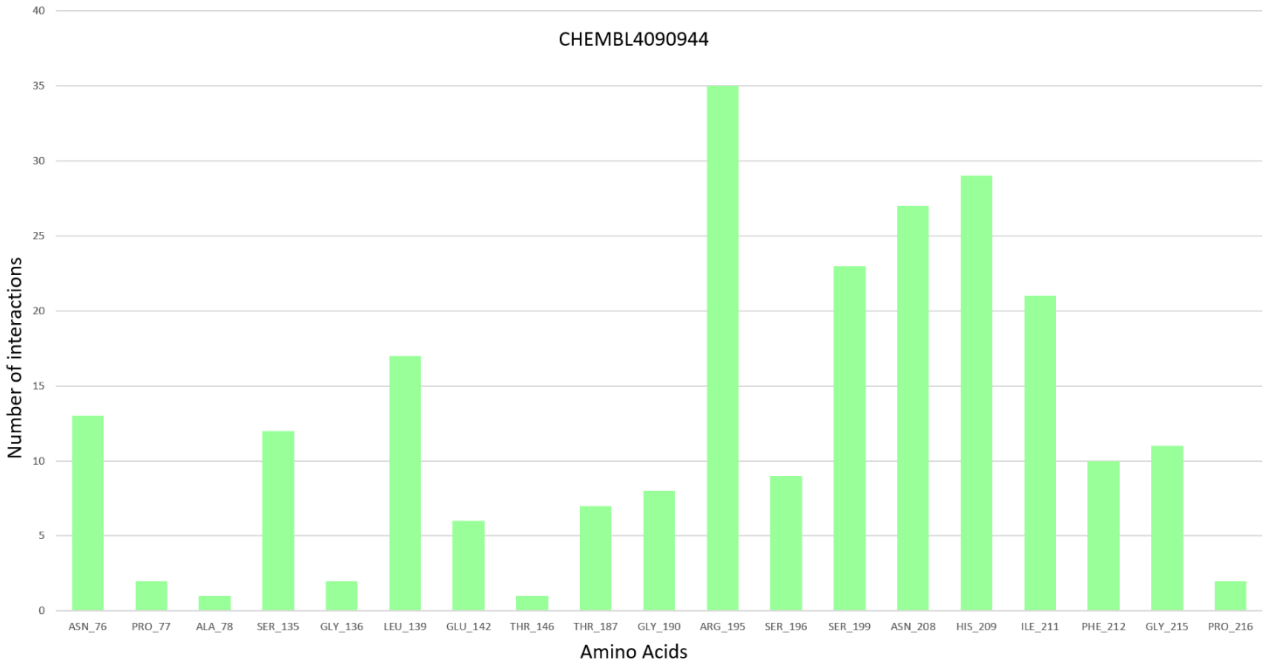

Figure 5.31. Identification of the population of interactions between the CHEMBL4090944 compound and the B_NPA binding site.

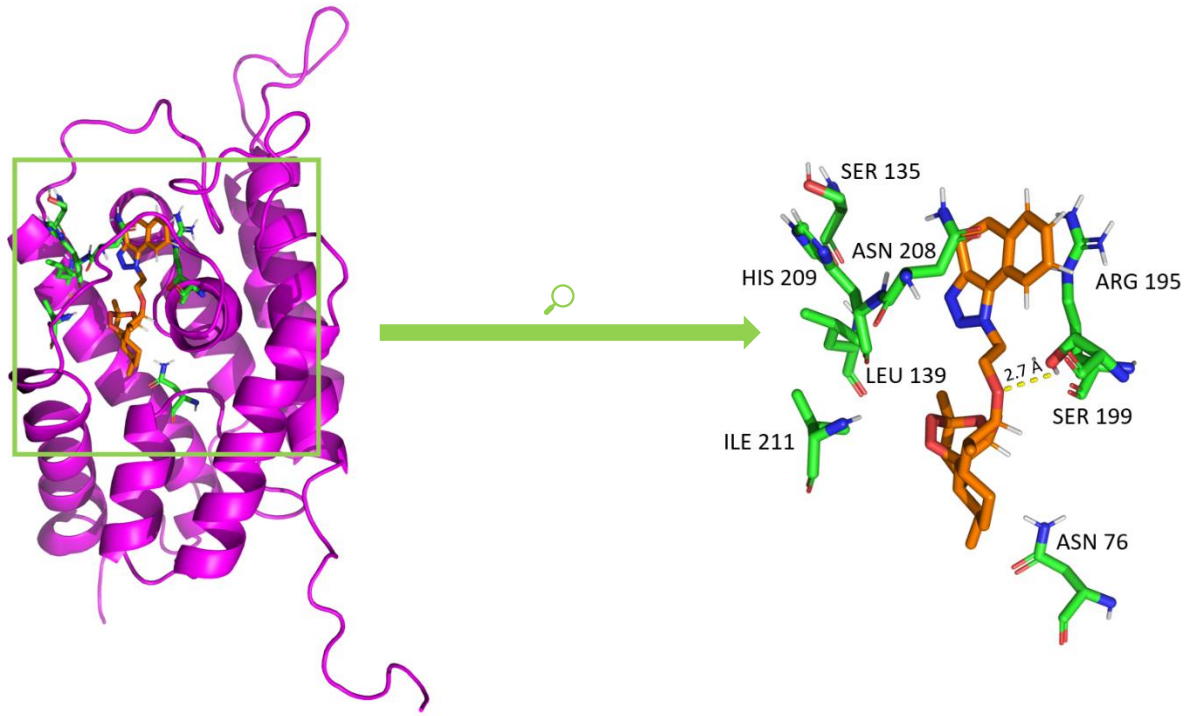

Figure 5.32. Monomer B of the MD2 structure of AQP1, highlighting the CHEMBL4090944 compound (in orange), the top 8 interacting amino acids (in green), and the existing H-bonds (yellow dashed line).

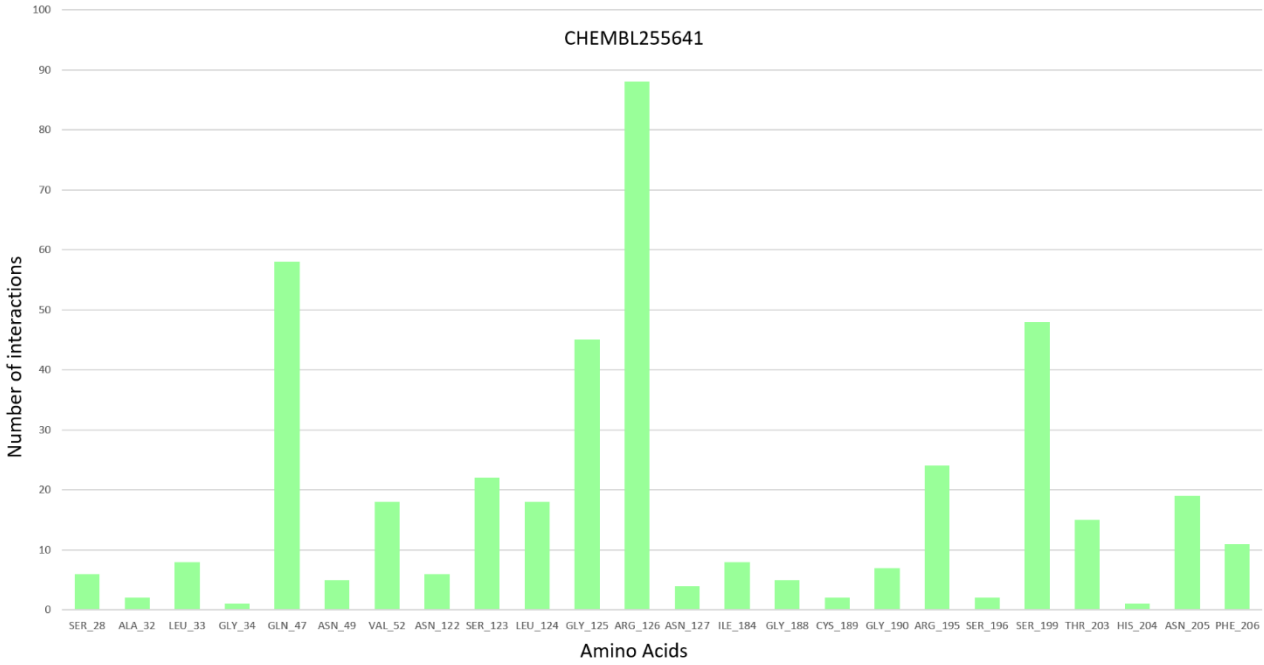

Figure 5.33. Identification of the population of interactions between the CHEMBL255641 compound and the C_ar/R binding site.

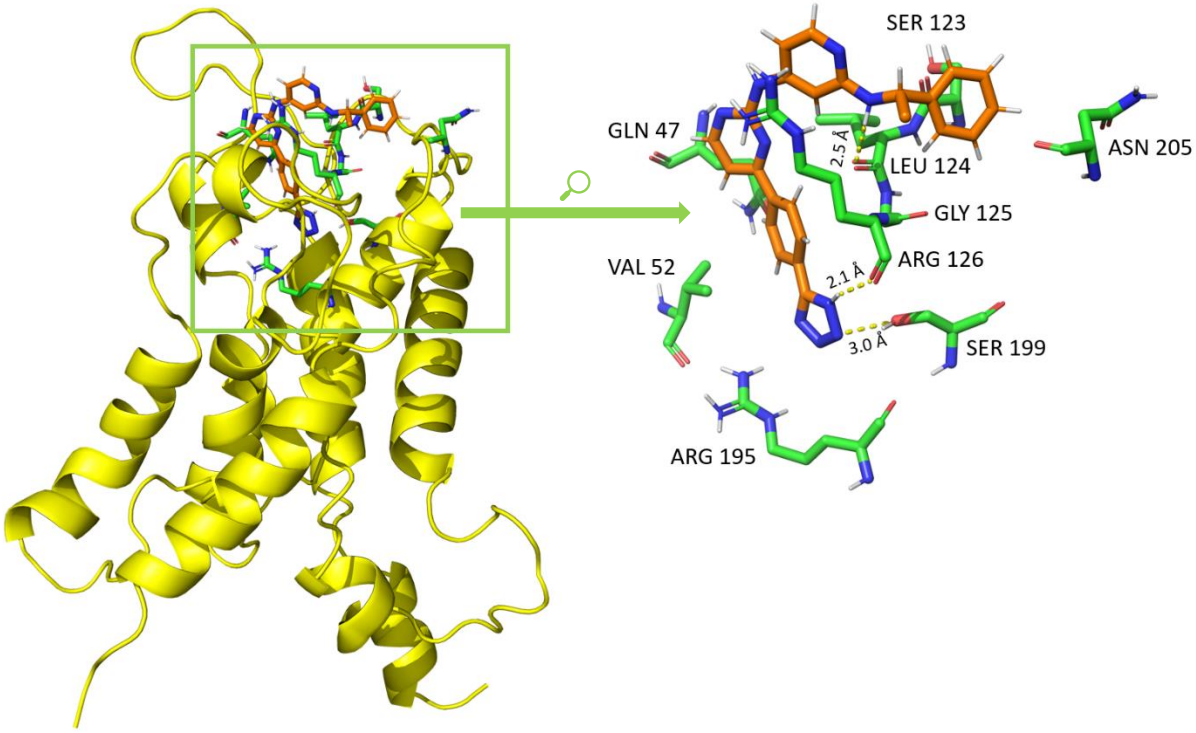

Figure 5.34. Monomer C of the MD2 structure of AQP1, highlighting the CHEMBL255641 compound (in orange), the top 9 interacting amino acids (in green), and the existing H-bonds (yellow dashed line).

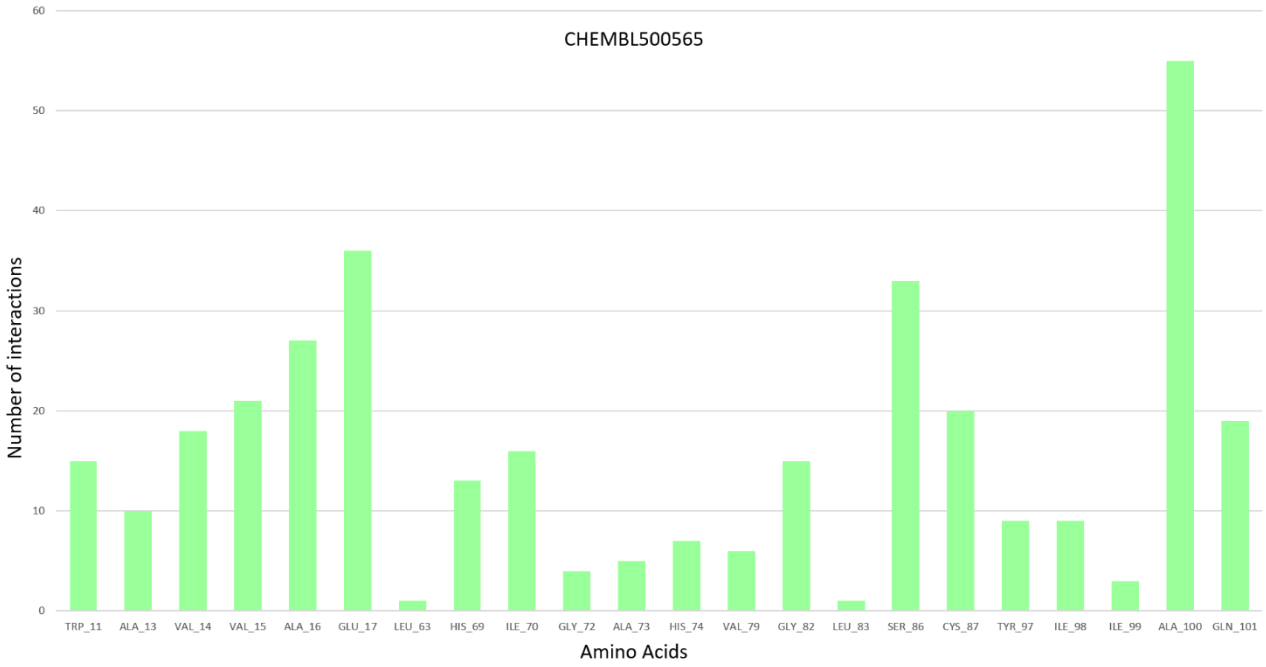

Figure 5.35. Identification of the population of interactions between the CHEMBL500565 compound and the C_NPA binding site.

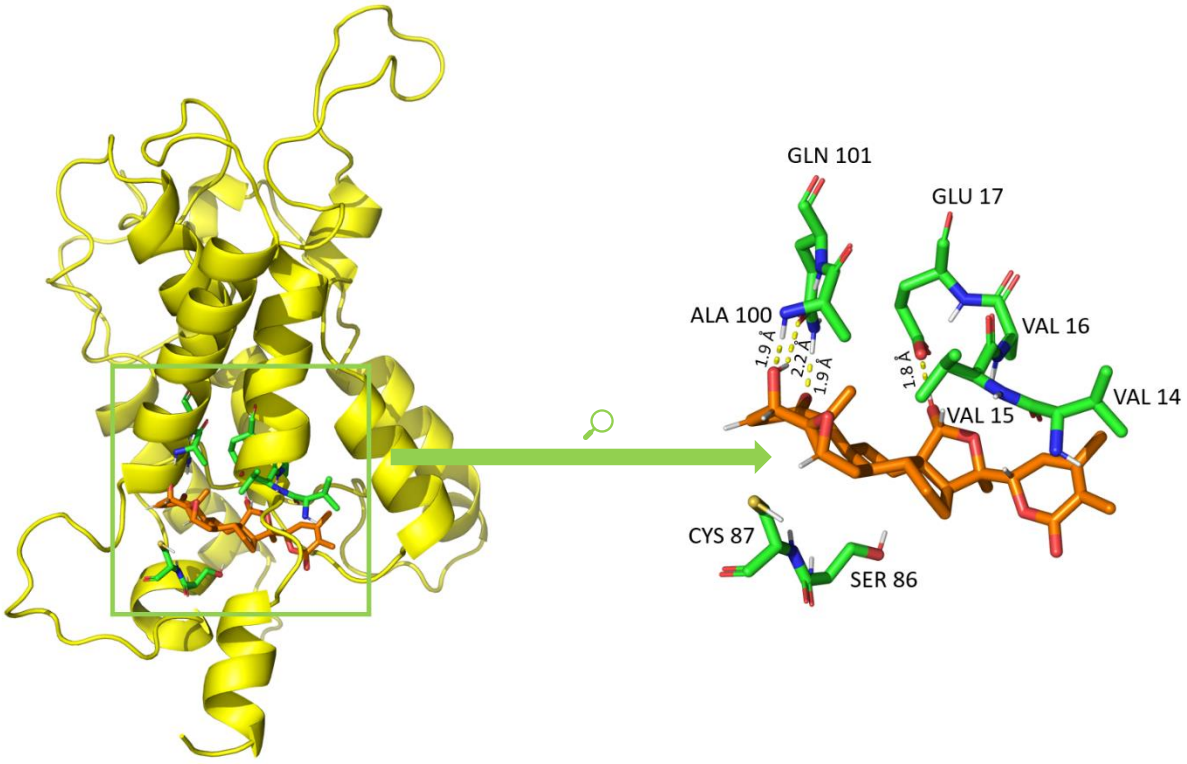

Figure 5.36. Monomer C of the MD2 structure of AQP1, highlighting the CHEMBL500565 compound (in orange), the top 8 interacting amino acids (in green), and the existing H-bonds (yellow dashed line).

As already referred to in this thesis, to physically prevent water conduction, an inhibitor's binding should occur deep in the narrow AQP1 pores. As previously described, especially in what concerns the results obtained at the B_NPA regions, compounds are able to interact deeper into the pore, which has been reported by others [5] as one major innovative factor correlated with compounds affinity and efficiency of inhibition.

Considering all the presented results, we can conclude that all of the compounds identified as top hits might be used as potential orally active drugs in future drug design campaigns. All compounds present good Binding Energy values to the AQP1 structure (Table 5.4-5.6). Additionally, they prove to have structural and chemical characteristics that respect the previously referred Lipinski's rule of five [149] and the Veber rule [165]. This later one states that orally active compounds should only have 10 or fewer rotatable bonds (Tables A.1-A.3), and satisfactory binding poses with interactions that promote the stabilization of the Ligand-Protein complex (Figures 5.32, 5.34, and 5.36).

Chapter 6

6. Concluding Remarks

The distinct AQPs' roles have already been well-established. Their involvement in the prognosis of various diseases states the urgency in the discovery of selective modulators or inhibitors as therapeutic agents. AQP1 has been proven to play a crucial role in tumor biology, including histological tumor grade, proliferation, migration, angiogenesis, or tumor-associated edema, namely due to its over-expression when compared to normal tissues [8]. Therefore, this protein can work as a potential diagnostic and therapeutic target in anticancer treatment since its inhibition in endothelial and tumor cells might limit tumor growth and spread [2,4]. Unfortunately, the hit rate for identifying small-molecule AQP modulators appears to be very low compared to other membrane proteins, and the few available modulators lack specificity or show high toxicity [6].

This thesis presented the performed steps of a new computational workflow based on several distinct methods that combine Ligand and Structure-based approaches to identify new AQP1 modulators. The use of a protein target multiconfigurational procedure, based on Molecular Dynamics simulations, distinguishes our approach from those previously followed by others on the quest for new and innovative AQP modulators. The need for a new approach arises from the fact that the available Crystallographic structures of this protein seem to be biased to configurational arrangements that are not suitable for Structure-based drug discovery campaigns. This is where the use of Molecular Dynamics simulations becomes useful. This methodology allows the sampling of the protein's configurational space, making it possible to identify the protein's conformations that could have never been obtained in Crystallographic structures. These conformations are of outmost importance since they highlight binding regions that could not be observed in the Crystallographic structures or that are in inadequate conformations to use in Structure-based drug discovery approaches. Working with a structure that isn't representative is a setback since we would be developing a modulator that is only able to bind to the protein in a precise and narrow moment.

In this work, after performing a clustering analysis of AQP1 conformations obtained from different MD replicate simulations, we identified three different conformations as good representatives for subsequent drug discovery screenings (MD1, MD2, MD3). One conformation (MD2) stood-out in these structures, where it was possible to identify three potential binding regions (B_NPA, C_ar/R, and C_NPA). A total of 20 hits, per binding site, with good binding affinity values and promising drug-like properties, were identified from a combination of pharmacophore and Docking based screening strategies. This list was then sorted using a combined approach of Molecular Dynamics simulations and MM/PBSA calculations, allowing the identification of three promising modulators of the activity of AQP1: CHEMBL4090944, CHEMBL255641, and CHEMBL500565.

Although further work is necessary, we can state that this thesis goal was accomplished since the proposed steps were successfully performed, and satisfactory conclusions were reached.

Chapter 7

7. Future Work

Following the presented work, the succeeding step will be purchasing and testing top hitters for each of the identified binding sites. Our collaborators at the Faculty of Pharmacy of the University of Lisbon, Professor Graça Soveral's research group, will perform this validation step. Subsequently, if any of the compounds show any promising modular activity of AQP1, further computational studies focused on their optimization will be developed. The further optimization steps will be focused on improving the compounds' physical chemistry properties or improving the compounds' affinity to the protein. In both approaches, to accomplish the desired objective, we will perform replacements on atoms, or group of atoms, always with a rational approach. As a final step, with the help of Organic or/and Medicinal Chemists, we will proceed to the synthesis of the derivative compounds identified in the previous step and once more, to assure the legitimacy of the obtained results, an experimental counterpart focused on functional validation of the compounds will be again performed.

References

[1] K. Takata, T. Matsuzaki, and Y. Tajika, “Aquaporins: Water channel proteins of the cell membrane,” *Prog. Histochem. Cytochem.*, vol. 39, no. 1, pp. 1–83, 2004, doi: 10.1016/j.proghi.2004.03.001.

[2] D. Ribatti, G. Ranieri, T. Annese, and B. Nico, “Aquaporins in cancer,” *Biochim. Biophys. Acta - Gen. Subj.*, vol. 1840, no. 5, pp. 1550–1553, 2014, doi: 10.1016/j.bbagen.2013.09.025.

[3] A. S. Verkman, “Aquaporins in clinical medicine,” *Annu. Rev. Med.*, vol. 63, no. 1, pp. 303–316, 2012, doi: 10.1146/annurev-med-043010-193843.

[4] J. Wang *et al.*, “Aquaporins as diagnostic and therapeutic targets in cancer: How far we are?,” *J. Transl. Med.*, vol. 13, no. 1, p. 96, 2015, doi: 10.1186/s12967-015-0439-7.

[5] A. S. Verkman, M. O. Anderson, and M. C. Papadopoulos, “Aquaporins: Important but elusive drug targets,” *Nat. Rev. Drug Discov.*, vol. 13, no. 4, pp. 259–277, 2014, doi: 10.1038/nrd4226.

[6] N. A. Castle, “Aquaporins as targets for drug discovery,” *Drug Discov. Today*, vol. 10, no. 7, pp. 485–493, 2005, doi: 10.1016/S1359-6446(05)03390-8.

[7] S. Kreida and S. Törnroth-Horsefield, “Structural insights into aquaporin selectivity and regulation,” *Curr. Opin. Struct. Biol.*, vol. 33, pp. 126–134, 2015, doi: 10.1016/j.sbi.2015.08.004.

[8] M. C. Papadopoulos and S. Saadoun, “Key roles of aquaporins in tumor biology,” *Biochim. Biophys. Acta - Biomembr.*, vol. 1848, no. 10, pp. 2576–2583, 2015, doi: 10.1016/j.bbamem.2014.09.001.

[9] J. P. Hughes, S. S. Rees, S. B. Kalindjian, and K. L. Philpott, “Principles of early drug discovery,” *Br. J. Pharmacol.*, vol. 162, no. 6, pp. 1239–1249, 2011, doi: 10.1111/j.1476-5381.2010.01127.x.

[10] R. V. C. Guido, G. Oliva, and A. D. Andricopulo, “Structure-and Ligand-based drug design approaches for neglected tropical diseases,” *Pure Appl. Chem.*, vol. 84, no. 9, pp. 1857–1866, 2012, doi: 10.1351/PAC-CON-11-11-07.

[11] S.-K. Lin, “Pharmacophore Perception, Development and Use in Drug Design. Edited by Osman F. Güner,” *Molecules*, vol. 5, no. 7, pp. 987–989, 2000, doi: 10.3390/50700987.

[12] H. Jhoti and A. R. Leach, *Structure-based drug discovery*. Springer Netherlands, 2007.

[13] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, “Computational methods in drug discovery,” *Pharmacol. Rev.*, vol. 66, no. 1, pp. 334–395, 2014, doi: 10.1124/pr.112.007336.

[14] N. Berdigaliyev and M. Aljofan, “An overview of drug discovery and development,” *Future Med. Chem.*, vol. 12, no. 10, pp. 939–947, 2020, doi: 10.4155/fmc-2019-0307.

[15] S. Höfinger *et al.*, “Structural features of aquaporin 4 supporting the formation of arrays and junctions in biomembranes,” *Biochim. Biophys. Acta - Biomembr.*, vol. 1818, no. 9, pp. 2234–2243, 2012, doi: 10.1016/j.bbamem.2012.04.009.

[16] K. Alleva, O. Chara, and G. Amodeo, “Aquaporins: Another piece in the osmotic puzzle,” *FEBS Lett.*, vol. 586, no. 19, pp. 2991–2999, 2012, doi: 10.1016/j.febslet.2012.06.013.

[17] G. P. Nagaraju *et al.*, “Aquaporins: Their role in gastrointestinal malignancies,” *Cancer Lett.*, vol. 373, no. 1, pp. 12–18, 2016, doi: 10.1016/j.canlet.2016.01.003.

[18] L. S. King and P. Agre, “Pathophysiology of the aquaporin water channels,” *Annu. Rev. Physiol.*, vol. 58, no. 1, pp. 619–648, 1996, doi: 10.1146/annurev.ph.58.030196.003155.

[19] A. S. Verkman, “More than just water channels: Unexpected cellular roles of aquaporins,” *J. Cell Sci.*, vol. 118, no. 15, pp. 3225–3232, 2005, doi: 10.1242/jcs.02519.

[20] S. Y. Jung, S. S. Kim, Y. Il Kim, S. H. Kim, and S. G. Yeo, “A review: Expression of aquaporins in otitis media,” *Int. J. Mol. Sci.*, vol. 18, no. 10, p. 2164, 2017, doi: 10.3390/ijms18102164.

[21] A. Rojek, J. Praetorius, J. Frøkiaer, S. Nielsen, and R. A. Fenton, “A current view of the mammalian aquaglyceroporins,” *Annu. Rev. Physiol.*, vol. 70, no. 1, pp. 301–327, 2008, doi: 10.1146/annurev.physiol.70.113006.100452.

[22] G. J. Cooper, Y. Zhou, P. Bouyer, I. I. Grichtchenko, and W. F. Boron, “Transport of volatile solutes through AQP1,” *J. Physiol.*, vol. 542, no. 1, pp. 17–29, 2002, doi: 10.1113/jphysiol.2002.023218.

[23] B. Wu and E. Beitz, “Aquaporins with selectivity for unconventional permeants,” *Cell. Mol. Life Sci.*, vol. 64, no. 18, pp. 2413–2421, 2007, doi: 10.1007/s00018-007-7163-2.

[24] M. Yasul, A. Hazama, T. H. Kwon, S. Nielsen, W. B. Guggino, and P. Agre, “Rapid gating and artion permeability of an intracellular aquaporin,” *Nature*, vol. 402, no. 6758, pp. 184–187, 1999, doi: 10.1038/46045.

[25] J. Rambow, B. Wu, D. Rönfeldt, and E. Beitz, “Aquaporins with anion/monocarboxylate permeability: Mechanisms, relevance for pathogen-host interactions,” *Front. Pharmacol.*, vol. 5, p. 199, 2014, doi: 10.3389/fphar.2014.00199.

[26] T. Gonen and T. Walz, “The structure of aquaporins,” *Q. Rev. Biophys.*, vol. 39, no. 4, pp. 361–396, 2006, doi: 10.1017/S0033583506004458.

[27] E. Kruse, N. Uehlein, and R. Kaldenhoff, “The aquaporins,” *Genome Biol.*, vol. 7, no. 2, p. 206, 2006, doi: 10.1186/gb-2006-7-2-206.

[28] H. Sui, B. G. Han, J. K. Lee, P. Walian, and B. K. Jap, “Structural basis of water-specific transport through the AQP1 water channel,” *Nature*, vol. 414, no. 6866, pp. 872–878, 2001, doi: 10.1038/414872a.

[29] D. Fu *et al.*, “Structure of a glycerol-conducting channel and the basis for its selectivity,” *Science*, vol. 290, no. 5491, pp. 481–486, 2000, doi: 10.1126/science.290.5491.481.

[30] N. Mitani-Ueno, N. Yamaji, F. J. Zhao, and J. F. Ma, “The aromatic/arginine selectivity filter of NIP aquaporins plays a critical role in substrate selectivity for silicon, boron, and arsenic,” *J. Exp. Bot.*, vol. 62, no. 12, pp. 4391–4398, 2011, doi: 10.1093/jxb/err158.

[31] N. Chakrabarti, E. Tajkhorshid, B. Roux, and R. Pomès, “Molecular Basis of Proton Blockage in Aquaporins,” *Structure*, vol. 12, no. 1, pp. 65–74, 2004, doi: 10.1016/j.str.2003.11.017.

[32] S. Gravelle, L. Joly, F. Detcheverry, C. Ybert, C. Cottin-Bizonne, and L. Bocquet, “Optimizing water permeability through the hourglass shape of aquaporins,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 41, pp. 16367–16372, 2013, doi: 10.1073/pnas.1306447110.

[33] E. Beitz, B. Wu, L. M. Holm, J. E. Schultz, and T. Zeuthen, “Point mutations in the aromatic/arginine region in aquaporin 1 allow passage of urea, glycerol, ammonia, and protons,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 2, pp. 269–274, 2006, doi: 10.1073/pnas.0507225103.

[34] H. Li *et al.*, “Enhancement of proton conductance by mutations of the selectivity filter of aquaporin-1,” *J. Mol. Biol.*, vol. 407, no. 4, pp. 607–620, 2011, doi: 10.1016/j.jmb.2011.01.036.

[35] B. Wu, C. Steinbronn, M. Alsterfjord, T. Zeuthen, and E. Beitz, “Concerted action of two cation filters in the aquaporin water channel,” *EMBO J.*, vol. 28, no. 15, pp. 2188–2194, 2009, doi: 10.1038/emboj.2009.182.

[36] D. Wree, B. Wu, T. Zeuthen, and E. Beitz, “Requirement for asparagine in the aquaporin NPA sequence signature motifs for cation exclusion,” *FEBS J.*, vol. 278, no. 5, pp. 740–748, 2011, doi: 10.1111/j.1742-4658.2010.07993.x.

[37] U. K. Eriksson, G. Fischer, R. Friemann, G. Enkavi, E. Tajkhorshid, and R. Neutze, “Subangstrom resolution x-ray structure details aquaporin-water interactions,” *Science*, vol. 340, no. 6138, pp. 1346–1349, 2013, doi: 10.1126/science.1234306.

[38] D. K. Yadav, S. Kumar, E. H. Choi, S. Chaudhary, and M. H. Kim, “Computational Modeling on Aquaporin-3 as Skin Cancer Target: A Virtual Screening Study,” *Front. Chem.*, vol. 8, p. 250, 2020, doi: 10.3389/fchem.2020.00250.

[39] A. S. Verkman, M. Hara-Chikuma, and M. C. Papadopoulos, “Aquaporins - New players in cancer biology,” *J. Mol. Med.*, vol. 86, no. 5, pp. 523–529, 2008, doi: 10.1007/s00109-008-0303-9.

[40] G. Soveral, S. Nielsen, and A. Casini, *Aquaporins in health and disease: New molecular targets for drug discovery*. CRC Press, 2016.

[41] H. Pan, C. C. Sun, C. Y. Zhou, and H. F. Huang, “Expression of aquaporin-1 in normal, hyperplasic, and carcinomatous endometria,” *Int. J. Gynecol. Obstet.*, vol. 101, no. 3, pp. 239–244, 2008, doi: 10.1016/j.ijgo.2007.12.006.

[42] A. Engel, T. Wspalz, and Y. Fujiyoshi, “The AQP structure and functional implications,” *Handbook of Experimental Pharmacology*. pp. 31–56, 2009, doi: 10.1007/978-3-540-79885-9_2.

[43] P. Kitchen *et al.*, “Water channel pore size determines exclusion properties but not solute selectivity,” *Sci. Rep.*, vol. 9, no. 1, p. 20369, 2019, doi: 10.1038/s41598-019-56814-z.

[44] A. Madeira, T. F. Moura, and G. Soveral, “Detecting aquaporin function and regulation,” *Front. Chem.*, vol. 4, p. 3, 2016, doi: 10.3389/fchem.2016.00003.

[45] S. Dajani, A. Saripalli, and N. Sharma-Walia, “Water transport proteins–aquaporins (AQPs) in cancer biology,” *Oncotarget*, vol. 9, no. 91, pp. 36392–36405, 2018, doi: 10.18632/oncotarget.26351.

[46] A. S. Verkman, B. Yang, Y. Song, G. T. Manley, and T. Ma, “Role of water channels in fluid transport studied by phenotype analysis of aquaporin knockout mice,” *Exp. Physiol.*, vol. 85, no. s1, pp. 233s-241s, 2000, doi: 10.1111/j.1469-445x.2000.tb00028.x.

[47] F. Wang, X. C. Feng, Y. M. Li, H. Yang, and T. H. Ma, “Aquaporins as potential drug targets,” *Acta Pharmacol. Sin.*, vol. 27, no. 4, pp. 395–401, 2006, doi: 10.1111/j.1745-7254.2006.00318.x.

[48] G. T. Manley *et al.*, “Aquaporin-4 deletion in mice reduces brain edema after acute water intoxication and ischemic stroke,” *Nat. Med.*, vol. 6, no. 2, pp. 159–163, 2000, doi: 10.1038/72256.

[49] N. Maeda *et al.*, “Adaptation to fasting by glycerol transport through aquaporin 7 in adipose tissue,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 51, pp. 17801–17806, 2004, doi: 10.1073/pnas.0406230101.

[50] T. Hibuse *et al.*, “Aquaporin 7 deficiency is associated with development of obesity through activation of adipose glycerol kinase,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 31, pp. 10993–10998, 2005, doi: 10.1073/pnas.0503291102.

[51] M. Hara-Chikuma *et al.*, “Progressive adipocyte hypertrophy in aquaporin-7-deficient mice: Adipocyte glycerol permeability as a novel regulator of fat accumulation,” *J. Biol. Chem.*, vol. 280, no. 16, pp. 15493–15496, 2005, doi: 10.1074/jbc.C500028200.

[52] S. Saadoun, M. C. Papadopoulos, M. Hara-Chikuma, and A. S. Verkman, “Impairment of angiogenesis and cell migration by targeted aquaporin-1 gene disruption,” *Nature*, vol. 434, no. 7034, pp. 786–792, 2005, doi: 10.1038/nature03460.

[53] J. Yong, Z. Xin, C. Kun, S. Rui, and Z. Tongcun, “Aquaporin-1 deletion in mice inhibits hepa1-6 hepatocellular carcinoma growth,” 2009, doi: 10.1109/ICBBE.2009.5162405.

[54] “The Nobel Prize in Chemistry 2013,” 2020. https://www.nobelprize.org/prizes/chemistry/2013/summary/.

[55] S. Fox *et al.*, “High-throughput screening: Update on practices and success,” *J. Biomol. Screen.*, vol. 11, no. 7, pp. 864–869, 2006, doi: 10.1177/1087057106292473.

[56] A. R. Leach, “Molecular Modeling: Principles and Applications 2nd Ed,” *Pearson Educ.*, 2001.

[57] B. L. da S. Victor, “Molecular Modeling of Redox Proteins involved in the Oxygen and Nitric Oxide Metabolism,” NOVA University of Lisbon, 2007.

[58] P. R. da S. Á. Magalhães, “Interaction between the neuropeptide kyotorphin and a lipid bilayer: a constant-ph molecular dynamics study,” University of Lisbon, Faculty of Sciences, 2009.

[59] T. F. da Silva, “Computational study of pH-dependent membrane insertion mechanism of pHLIP peptides,” University of Lisbon, Faculty of Sciences, 2017.

[60] W. F. van Gunsteren and H. J. C. Berendsen, “Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry,” *Angew. Chemie Int. Ed. English*, vol. 29, no. 9, pp. 992–1023, 1990, doi: 10.1002/anie.199009921.

[61] W. F. Van Gunsteren *et al.*, “Biomolecular modeling: Goals, problems, perspectives,” *Angew. Chemie - Int. Ed.*, vol. 45, no. 25, pp. 4064–4092, 2006, doi: 10.1002/anie.200502655.

[62] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, “A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1656–1676, 2004, doi: 10.1002/jcc.20090.

[63] K. Lindorff-Larsen, P. Maragakis, S. Piana, and D. E. Shaw, “Picosecond to Millisecond Structural Dynamics in Human Ubiquitin,” *J. Phys. Chem. B*, vol. 120, no. 33, pp. 8313–8320, 2016, doi: 10.1021/acs.jpcb.6b02024.

[64] R. Feynman, “Vol 3 Ch 01 - Quantum Behavior.pdf,” *The Feynman Lectures on Physics*. 1963.

[65] N. J. H. Dunn and W. G. Noid, “Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids,” *J. Chem. Phys.*, vol. 143, no. 24, p. 243148, 2015, doi: 10.1063/1.4937383.

[66] V. Zoete, M. Cuendet, U. F. Röhrig, A. Grosdidier, and O. Michielin, “Molecular modeling of proteins: From simulations to drug design applications,” in *Bioinformatics: A Swiss Perspective*, 2009, pp. 247–281.

[67] N. Vaidehi and A. Jain, “Internal coordinate molecular dynamics: A foundation for multiscale dynamics,” *J. Phys. Chem. B*, vol. 119, no. 4, pp. 1233–1242, 2015, doi: 10.1021/jp509136y.

[68] D. R. dos S. V. Viçosa, “Molecular modeling and simulation of pH effects in lipid bilayers,” University of Lisbon, Faculty of Sciences, 2015.

[69] A. M. Dar and S. Mir, “Molecular Docking: Approaches, Types, Applications and Basic Challenges,” *J. Anal. Bioanal. Tech.*, vol. 8, no. 2, 2017, doi: 10.4172/2155-9872.1000356.

[70] E. Generalic, “Lennard-Jones potential,” *Croatian-English Chemistry Dictionary & Glossary*, 2018. https://glossary.periodni.com/glossary.php?en=Lennard-Jones+potential.

[71] K. Vanommeslaeghe, O. Guvench, and A. D. MacKerell, “Molecular Mechanics,” *Curr. Pharm. Des.*, vol. 20, no. 20, pp. 3281–3292, 2014, doi: 10.2174/13816128113199990600.

[72] M. P. Allen, D. J. Tildesley, and J. R. Banavar, “Computer Simulation of Liquids,” *Phys. Today*, vol. 42, no. 3, pp. 105–106, 1989, doi: 10.1063/1.2810937.

[73] C. Oostenbrink, T. A. Soares, N. F. A. Van Der Vegt, and W. F. Van Gunsteren, “Validation of the 53A6 GROMOS force field,” *Eur. Biophys. J.*, vol. 34, no. 4, pp. 273–284, 2005, doi: 10.1007/s00249-004-0448-6.

[74] R. D. Lins and P. H. Hünenberger, “A new GROMOS force field for hexopyranose-based carbohydrates,” *J. Comput. Chem.*, vol. 26, no. 13, pp. 1400–1412, 2005, doi: 10.1002/jcc.20275.

[75] T. A. Soares, X. Daura, C. Oostenbrink, L. J. Smith, and W. F. van Gunsteren, “Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme,” *J. Biomol. NMR*, vol. 30, no. 4, pp. 407–422, 2004, doi: 10.1007/s10858-004-5430-1.

[76] T. A. Soares *et al.*, “An improved nucleic acid parameter set for the GROMOS force field,” *J. Comput. Chem.*, vol. 26, no. 7, pp. 725–737, 2005, doi: 10.1002/jcc.20193.

[77] N. Schmid *et al.*, “Definition and testing of the GROMOS force-field versions 54A7 and 54B7,” *Eur. Biophys. J.*, vol. 40, no. 7, pp. 843–856, 2011, doi: 10.1007/s00249-011-0700-9.

[78] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey, “Dogsitescorer: A web server for automatic binding site prediction, analysis and druggability assessment,” *Bioinformatics*, vol. 28, no. 15, pp. 2074–2075, 2012, doi: 10.1093/bioinformatics/bts310.

[79] A. Volkamer, A. Griewel, T. Grombacher, and M. Rarey, “Analyzing the topology of active sites: On the prediction of pockets and subpockets,” *J. Chem. Inf. Model.*, vol. 50, no. 11, pp. 2041–2052, 2010, doi: 10.1021/ci100241y.

[80] A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann, and M. Rarey, “Combining global and local measures for Structure-based druggability predictions,” *J. Chem. Inf. Model.*, vol. 52, no. 2, pp. 360–372, 2012, doi: 10.1021/ci200454v.

[81] R. Harris, A. J. Olson, and D. S. Goodsell, “Automated prediction of ligand-binding sites in proteins,” *Proteins Struct. Funct. Genet.*, vol. 70, no. 4, pp. 1506–1517, 2008, doi: 10.1002/prot.21645.

[82] G. M. Morris *et al.*, “Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity,” *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, 2009, doi: 10.1002/jcc.21256.

[83] R. Al Olaby, H. M. Azzazy, R. Harris, B. Chromy, J. Vielmetter, and R. Balhorn, “Identification of ligands that target the HCV-E2 binding site on CD81,” *J. Comput. Aided. Mol. Des.*, vol. 27, no. 4, pp. 337–346, 2013, doi: 10.1007/s10822-013-9649-3.

[84] BioSolveIT GmbH, “SeeSAR version 10.2,” 2020. www.biosolveit.de/SeeSAR.

[85] Macs in Chemistry, “Drug Design.” https://www.macinchem.org/blog/files/category-drug-design.php.

[86] T. Lengauer and M. Rarey, “Computational methods for biomolecular docking,” *Curr. Opin. Struct. Biol.*, vol. 6, no. 3, pp. 402–406, 1996, doi: 10.1016/S0959-440X(96)80061-3.

[87] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, “Docking and scoring in virtual screening for drug discovery: Methods and applications,” *Nat. Rev. Drug Discov.*, vol. 3, no. 11, pp. 935–949, 2004, doi: 10.1038/nrd1549.

[88] T. Mostashari-Rad *et al.*, “Study of CXCR4 chemokine receptor inhibitors using QSPR and molecular docking methodologies,” *J. Theor. Comput. Chem.*, vol. 18, no. 4, p. 1950018, 2019, doi: 10.1142/S0219633619500184.

[89] G. Rastelli, “Emerging topics in Structure-based virtual screening,” *Pharm. Res.*, vol. 30, no. 5, pp. 1458–1463, 2013, doi: 10.1007/s11095-013-1012-9.

[90] W. L. Jorgensen, “Rusting of the lock and key model for protein-ligand binding,” *Science*, vol. 254, no. 5034, pp. 954–955, 1991, doi: 10.1126/science.1719636.

[91] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, “Protein-ligand docking: Current status and future challenges,” *Proteins Struct. Funct. Bioinforma.*, vol. 65, no. 1, pp. 15–26, 2006, doi: 10.1002/prot.21082.

[92] D. S. Goodsell and A. J. Olson, “Automated docking of substrates to proteins by simulated annealing,” *Proteins Struct. Funct. Bioinforma.*, vol. 8, no. 3, pp. 195–202, 1990, doi: 10.1002/prot.340080302.

[93] G. R. Smith and M. J. E. Sternberg, “Prediction of protein-protein interactions by docking methods,” *Curr. Opin. Struct. Biol.*, vol. 12, no. 1, pp. 28–35, 2002, doi: 10.1016/S0959-440X(02)00285-3.

[94] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, “Principles of docking: An overview of search algorithms and a guide to scoring functions,” *Proteins Struct. Funct. Genet.*, vol. 47, no. 4, pp. 409–443, 2002, doi: 10.1002/prot.10115.

[95] J. Wang, Y. Deng, and B. Roux, “Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials,” *Biophys. J.*, vol. 91, no. 8, pp. 2798–2814, 2006, doi: 10.1529/biophysj.106.084301.

[96] R. G. Coleman, M. Carchia, T. Sterling, J. J. Irwin, and B. K. Shoichet, “Ligand Pose and Orientational Sampling in Molecular Docking,” *PLoS One*, vol. 8, no. 10, p. e75992, 2013, doi: 10.1371/journal.pone.0075992.

[97] A. R. Leach, B. K. Shoichet, and C. E. Peishoff, “Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps,” *J. Med. Chem.*, vol. 49, no. 20, pp. 5851–5855, 2006, doi: 10.1021/jm060999m.

[98] C. F. Wong, “Flexible receptor docking for drug discovery,” *Expert Opin. Drug Discov.*, vol. 10, no. 11, pp. 1189–1200, 2015, doi: 10.1517/17460441.2015.1078308.

[99] T. J. A. Ewing and I. D. Kuntz, “Critical evaluation of search algorithms for automated molecular docking and database screening,” *J. Comput. Chem.*, vol. 18, no. 9, pp. 1175–1189, 1997, doi: 10.1002/(SICI)1096-987X(19970715)18:9<1175::AID-JCC6>3.0.CO;2-O.

[100] Ajay and M. A. Murcko, “Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes,” *J. Med. Chem.*, vol. 38, no. 26, pp. 4953–4967, 1995, doi: 10.1021/jm00026a001.

[101] O. Trott and A. J. Olson, “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading,” *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2009, doi: 10.1002/jcc.21334.

[102] R. Wang, L. Lai, and S. Wang, “Further development and validation of empirical scoring functions for Structure-based binding affinity prediction,” *J. Comput. Aided. Mol. Des.*, vol. 16, no. 1, pp. 11–26, 2002, doi: 10.1023/A:1016357811882.

[103] H. J. C. Berendsen and W. F. van Gunsteren, “Practical Algorithms for Dynamic Simulations,” *Proc. Int. Sch. Phys. “Enrico Fermi,”* pp. 43–65, 1985.

[104] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, “LINCS: A Linear Constraint Solver for molecular simulations,” *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997, doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[105] S. Miyamoto and P. A. Kollman, “Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models,” *J. Comput. Chem.*, vol. 13, no. 8, pp. 952–962, 1992, doi: 10.1002/jcc.540130805.

[106] J. Hermans, H. J. C. Berendsen, W. F. Van Gunsteren, and J. P. M. Postma, “A consistent empirical potential for water–protein interactions,” *Biopolymers*, vol. 23, no. 8, pp. 1513–1518, 1984, doi: 10.1002/bip.360230807.

[107] D. van der Spoel *et al.*, “Gromacs User Manual version 4.0,” 2005, [Online]. Available: www.gromacs.org.

[108] A. K. Malde *et al.*, “An Automated force field Topology Builder (ATB) and repository: Version 1.0,” *J. Chem. Theory Comput.*, vol. 7, no. 12, pp. 4026–4037, 2011, doi: 10.1021/ct200196m.

[109] S. Canzar *et al.*, “Charge group partitioning in biomolecular simulation,” *J. Comput. Biol.*, vol. 20, no. 3, pp. 188–198, 2013, doi: 10.1089/cmb.2012.0239.

[110] K. B. Koziara, M. Stroet, A. K. Malde, and A. E. Mark, “Testing and validation of the Automated Topology Builder (ATB) version 2.0: Prediction of hydration free enthalpies,” *J. Comput. Aided. Mol. Des.*, vol. 28, no. 3, pp. 221–233, 2014, doi: 10.1007/s10822-014-9713-7.

[111] J. B. J. Chapman, R. E. Cohen, A. V. Kimmel, and D. M. Duffy, “Improving the Functional Control of Aged Ferroelectrics Using Insights from Atomistic Modeling,” *Phys. Rev. Lett.*, vol. 119, no. 17, p. 177602, 2017, doi: 10.1103/PhysRevLett.119.177602.

[112] J. N. Israelachvili, *Intermolecular and Surface Forces: Third Edition*. Academic Press, 2011.

[113] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. Van Gunsteren, “A generalized reaction field method for molecular dynamics simulations,” *J. Chem. Phys.*, vol. 102, no. 13, pp. 5451–5459, 1995, doi: 10.1063/1.469273.

[114] T. Darden, D. York, and L. Pedersen, “Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems,” *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993, doi: 10.1063/1.464397.

[115] G. A. Cisneros, M. Karttunen, P. Ren, and C. Sagui, “Classical electrostatics for biomolecular simulations,” *Chem. Rev.*, vol. 114, no. 1, pp. 779–814, 2014, doi: 10.1021/cr300461d.

[116] T. Darden, L. Perera, L. Li, and P. Lee, “New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations,” *Structure*, vol. 7, no. 3, pp. R55–R60, 1999, doi: 10.1016/S0969-2126(99)80033-1.

[117] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, “A smooth particle mesh Ewald method,” *J. Chem. Phys.*, vol. 103, no. 19, pp. 8577–8593, 1995, doi: 10.1063/1.470117.

[118] G. Bussi, D. Donadio, and M. Parrinello, “Canonical sampling through velocity rescaling,” *J. Chem. Phys.*, vol. 126, no. 1, p. 014101, 2007, doi: 10.1063/1.2408420.

[119] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, 1984, doi: 10.1063/1.448118.

[120] GROMACS development Team, “Introduction — GROMACS 2020 documentation,” 2020. http://manual.gromacs.org/2020/reference-manual/introduction.html.

[121] H. B. Curry, “The method of steepest descent for non-linear minimization problems,” *Q. Appl. Math.*, vol. 2, no. 3, pp. 258–261, 1944, doi: 10.1090/qam/10667.

[122] C. Cook, W. Jin, and S. X. D. Tan, “GPU-based ising computing for solving balanced min-cut graph partitioning problem,” *ArXiv*, 2019.

[123] N. Homeyer and H. Gohlke, “Free energy calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area method,” *Mol. Inform.*, vol. 31, no. 2, pp. 114–122, 2012, doi: 10.1002/minf.201100135.

[124] P. A. Kollman *et al.*, “Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models,” *Acc. Chem. Res.*, vol. 33, no. 12, pp. 889–897, 2000, doi: 10.1021/ar000033j.

[125] S. Miyamoto and P. A. Kollman, “What determines the strength of noncovalent association of ligands to proteins in aqueous solution?,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 18, pp. 8402–8406, 1993, doi: 10.1073/pnas.90.18.8402.

[126] N. Froloff, A. Windemuth, and B. Honig, “On the calculation of binding free energies using continuum methods: Application to MHC class I protein-peptide interactions,” *Protein Sci.*, vol. 6, no. 6, pp. 1293–1301, 1997, doi: 10.1002/pro.5560060617.

[127] R. Kumari, R. Kumar, and A. Lynn, “G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations,” *J. Chem. Inf. Model.*, vol. 54, no. 7, pp. 1951–1962, 2014, doi: 10.1021/ci500020m.

[128] A. Ganesan, M. L. Coote, and K. Barakat, “Molecular dynamics-driven drug discovery: leaping forward with confidence,” *Drug Discov. Today*, vol. 22, no. 2, pp. 249–269, 2017, doi: 10.1016/j.drudis.2016.11.001.

[129] S. Genheden and U. Ryde, “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities,” *Expert Opin. Drug Discov.*, vol. 10, no. 5, pp. 449–461, 2015, doi: 10.1517/17460441.2015.1032936.

[130] M. K. Gilson and B. Honig, “Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis,” *Proteins Struct. Funct. Bioinforma.*, vol. 4, no. 1, pp. 7–18, 1988, doi: 10.1002/prot.340040104.

[131] R. C. Rizzo, T. Aynechi, D. A. Case, and I. D. Kuntz, “Estimation of absolute free energies of hydration using continuum methods: Accuracy of partial charge models and optimization of nonpolar contributions,” *J. Chem. Theory Comput.*, vol. 2, no. 1, pp. 128–139, 2005, doi: 10.1021/ct050097l.

[132] D. Sitkoff, K. A. Sharp, and B. Honig, “Accurate calculation of hydration free energies using macroscopic solvent models,” *J. Phys. Chem.*, vol. 98, no. 7, pp. 1978–1988, 1994, doi: 10.1021/j100058a043.

[133] W. Clark Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, “Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics,” *J. Am. Chem. Soc.*, vol. 112, no. 16, pp. 6127–6129, 1990, doi: 10.1021/ja00172a038.

[134] H. Gohlke, C. Kiel, and D. A. Case, “Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes.,” *J. Mol. Biol.*, vol. 330, no. 4, pp. 891–913, 2003, doi: 10.1016/S0022-2836(03)00610-7.

[135] B. Honig and A. Nicholls, “Classical electrostatics in biology and chemistry,” *Science*, vol. 268, no. 5214, pp. 1144–1149, 1995, doi: 10.1126/science.7761829.

[136] K. A. Sharp and B. Honig, “Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation,” *J. Phys. Chem.*, vol. 94, no. 19, pp. 7684–7692, 1990, doi: 10.1021/j100382a068.

[137] Sebastian Raschka, “Molecular docking, estimating free energies of binding, and AutoDock’s semi-empirical force field,” *Sebastian Raschka’s Website*, 2014, doi: 10.13140/2.1.5097.1528.

[138] E. A. Coutsias, C. Seok, and K. A. Dill, “Using quaternions to calculate RMSD,” *J. Comput. Chem.*, vol. 25, no. 15, pp. 1849–1857, 2004, doi: 10.1002/jcc.20110.

[139] N. Rosário-Ferreira *et al.*, “In silico end-to-end protein-ligand interaction characterization pipeline: the case of SARS-CoV-2,” Submited for Publication, 2020.

[140] ChemAxon, “Molecule file conversion with Molconverter,” [Online]. Available: https://docs.chemaxon.com/display/docs/Molecule_file_conversion_with_Molconverter.html.

[141] ChemAxon, “Marvin.” https://chemaxon.com/products/marvin.

[142] Morris, “What is the format of a PDBQT file?,” 2007. http://autodock.scripps.edu/faqs-help/faq/what-is-the-format-of-a-pdbqt-file/.

[143] G. Ren, V. S. Reddy, A. Cheng, P. Melnyk, and A. K. Mitra, “Visualization of a water-selective pore by electron crystallography in vitreous ice,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 4, pp. 1398–1403, 2001, doi: 10.1073/pnas.98.4.1398.

[144] H. Park, J. Lee, and S. Lee, “Critical assessment of the automated AutoDock as a new docking tool for virtual screening,” *Proteins Struct. Funct. Genet.*, vol. 65, no. 3, pp. 549–554, 2006, doi: 10.1002/prot.21183.

[145] A. S. F. Oliveira, A. M. Baptista, and C. M. Soares, “Inter-domain communication mechanisms in an ABC importer: A molecular dynamics study of the MalFGK 2E complex,” *PLoS Comput. Biol.*, vol. 7, no. 8, p. e1002128, 2011, doi: 10.1371/journal.pcbi.1002128.

[146] B. Hess, “P-LINCS: A parallel linear constraint solver for molecular simulation,” *J. Chem. Theory Comput.*, vol. 4, no. 1, pp. 116–122, 2007, doi: 10.1021/ct700200b.

[147] BioSolveIT, “FTrees,” 2019. https://www.biosolveit.de/FTrees/.

[148] P. Gatti-Lafranconi, “Pairwise distances,” 2019. https://www.pymolwiki.org/index.php/Pairwise_distances.

[149] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings,” *Adv. Drug Deliv. Rev.*, vol. 46, no. 1–3, pp. 3–26, 2001, doi: 10.1016/S0169-409X(00)00129-0.

[150] A. Leo, C. Hansch, and D. Elkins, “Partition coefficients and their uses,” *Chem. Rev.*, vol. 71, no. 6, pp. 525–616, 1971, doi: 10.1021/cr60274a001.

[151] T. Sander, J. Freyss, M. Von Korff, and C. Rufener, “DataWarrior: An open-source program for chemistry aware data visualization and analysis,” *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 460–473, 2015, doi: 10.1021/ci500588j.

[152] E. Migliati *et al.*, “Inhibition of aquaporin-1 and aquaporin-4 water permeability by a derivative of the loop diuretic bumetanide acting at an internal pore-occluding binding site,” *Mol. Pharmacol.*, vol. 76, no. 1, pp. 105–112, 2009, doi: 10.1124/mol.108.053744.

[153] P. Kolb and J. Irwin, “Docking Screens: Right for the Right Reasons?,” *Curr. Top. Med. Chem.*, vol. 9, no. 9, pp. 755–770, 2009, doi: 10.2174/156802609789207091.

[154] A. Kamegawa, Y. Hiroaki, K. Tani, and Y. Fujiyoshi, “Two-dimensional crystal structure of aquaporin-4 bound to the inhibitor acetazolamide,” *Reprod. Syst. Sex. Disord.*, vol. 65, no. 2, pp. 177–184, 2016, doi: 10.1093/jmicro/dfv368.

[155] B. Aikman, A. De Almeida, S. M. Meier-Menches, and A. Casini, “Aquaporins in cancer development: Opportunities for bioinorganic chemistry to contribute novel chemical probes and therapeutic agents,” *Metallomics*, vol. 10, no. 5, pp. 696–712, 2018, doi: 10.1039/c8mt00072g.

[156] D. F. Savage and R. M. Stroud, “Structural Basis of Aquaporin Inhibition by Mercury,” *J. Mol. Biol.*, vol. 368, no. 3, pp. 607–617, 2007, doi: 10.1016/j.jmb.2007.02.070.

[157] C. Esteva-Font, B. J. Jin, S. Lee, P. W. Phuan, M. O. Anderson, and A. S. Verkman, “Experimental evaluation of proposed small-molecule inhibitors of water channel aquaporin-1,” *Mol. Pharmacol.*, vol. 89, no. 6, pp. 686–693, 2016, doi: 10.1124/mol.116.103929.

[158] Y. Sonntag *et al.*, “Identification and characterization of potent and selective aquaporin-3 and aquaporin-7 inhibitors,” *J. Biol. Chem.*, vol. 294, no. 18, pp. 7377–7387, 2019, doi: 10.1074/jbc.RA118.006083.

[159] S. Ferreira and B. L. Victor, “Improving the druggability of Aquaporin-1 for future drug discovery campaigns,” *MOL2NET 2019, Int. Conf. Multidiscip. Sci. 5th Ed. Sess. EJIBCE-03 Struct. Comput. Biol. Lisboa, Port.*, 2020, doi: 10.3390/mol2net-05-06774.

[160] A. Gaulton *et al.*, “ChEMBL: A large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1100–D1107, 2012, doi: 10.1093/nar/gkr777.

[161] Y. Hirano, N. Okimoto, I. Kadohira, M. Suematsu, K. Yasuoka, and M. Yasui, “Molecular mechanisms of how mercury inhibits water permeation through aquaporin-1: Understanding by molecular dynamics simulation,” *Biophys. J.*, vol. 98, no. 8, pp. 1512–1519, 2010, doi: 10.1016/j.bpj.2009.12.4310.

[162] D. C. Thompson, C. Humblet, and D. Joseph-McCarthy, “Investigation of MM-PBSA rescoring of docking poses,” *J. Chem. Inf. Model.*, vol. 48, no. 5, pp. 1081–1091, 2008, doi: 10.1021/ci700470c.

[163] T. Hou, J. Wang, Y. Li, and W. Wang, “Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. the accuracy of ranking poses generated from docking,” *J. Comput. Chem.*, vol. 32, no. 5, pp. 866–877, 2011, doi: 10.1002/jcc.21666.

[164] G. Poli, C. Granchi, F. Rizzolio, and T. Tuccinardi, “Application of MM-PBSA methods in virtual screening,” *Molecules*, vol. 25, no. 8, p. 1971, 2020, doi: 10.3390/molecules25081971.

[165] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, “Molecular properties that influence the oral bioavailability of drug candidates,” *J. Med. Chem.*, vol. 45, no. 12, pp. 2615–2623, 2002, doi: 10.1021/jm020017n.

Appendix

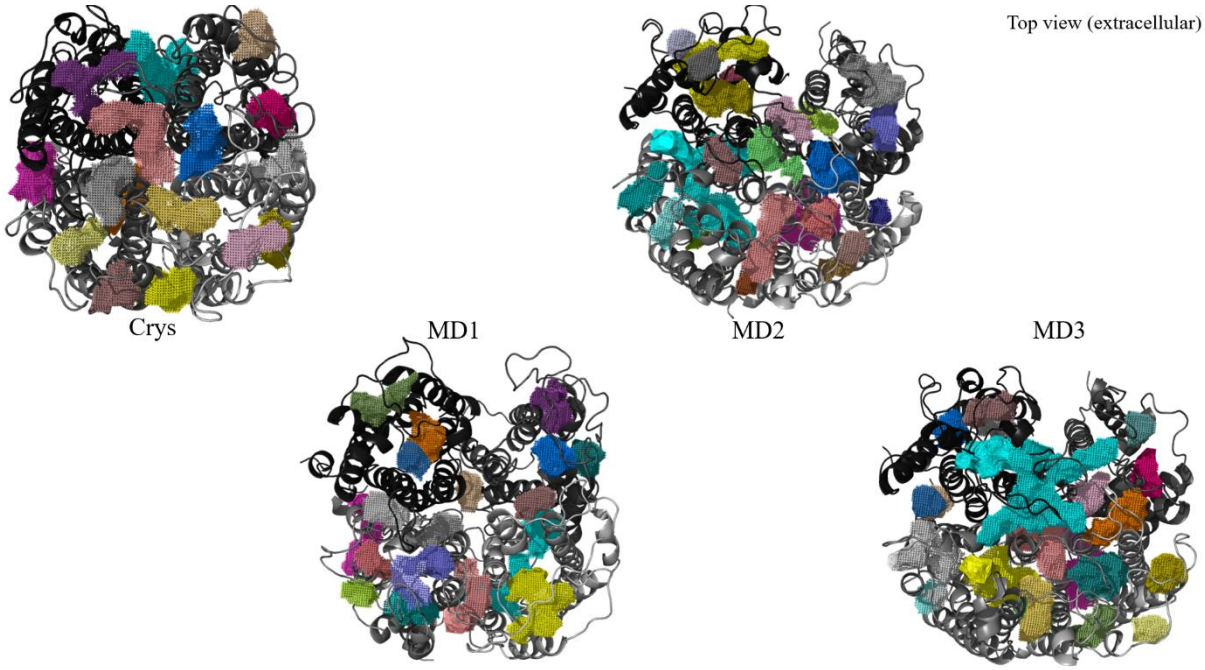

Figure A.1. Top view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Automated Pocket Detection using DogSiteScorer (colored mesh).

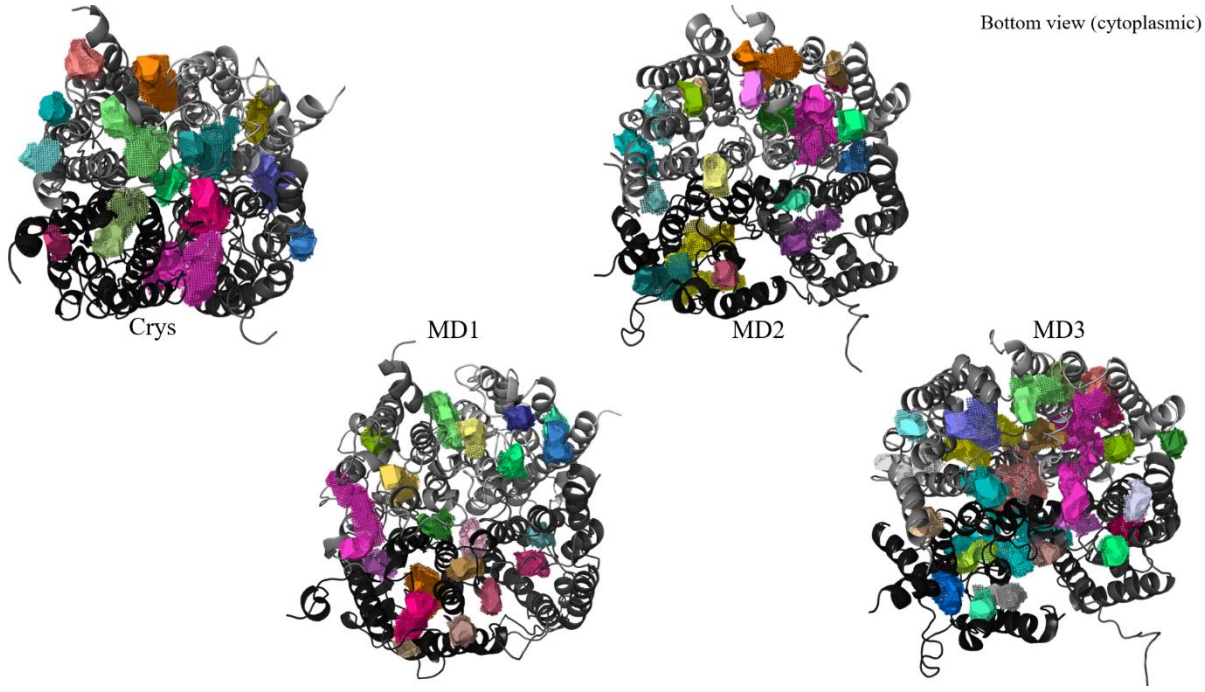

Figure A.2. Bottom view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Automated Pocket Detection using DogSiteScorer (colored mesh).

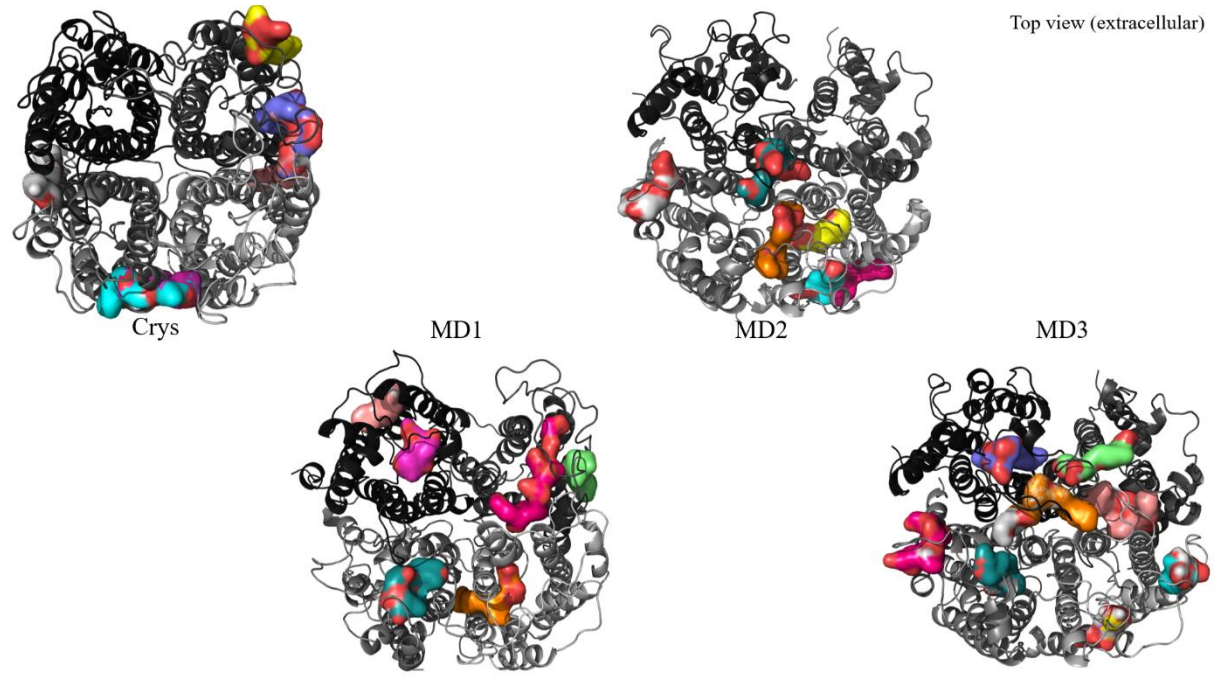

Figure A.3. Top view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Automated Pocket Detection using AutoLigand (colored surface).

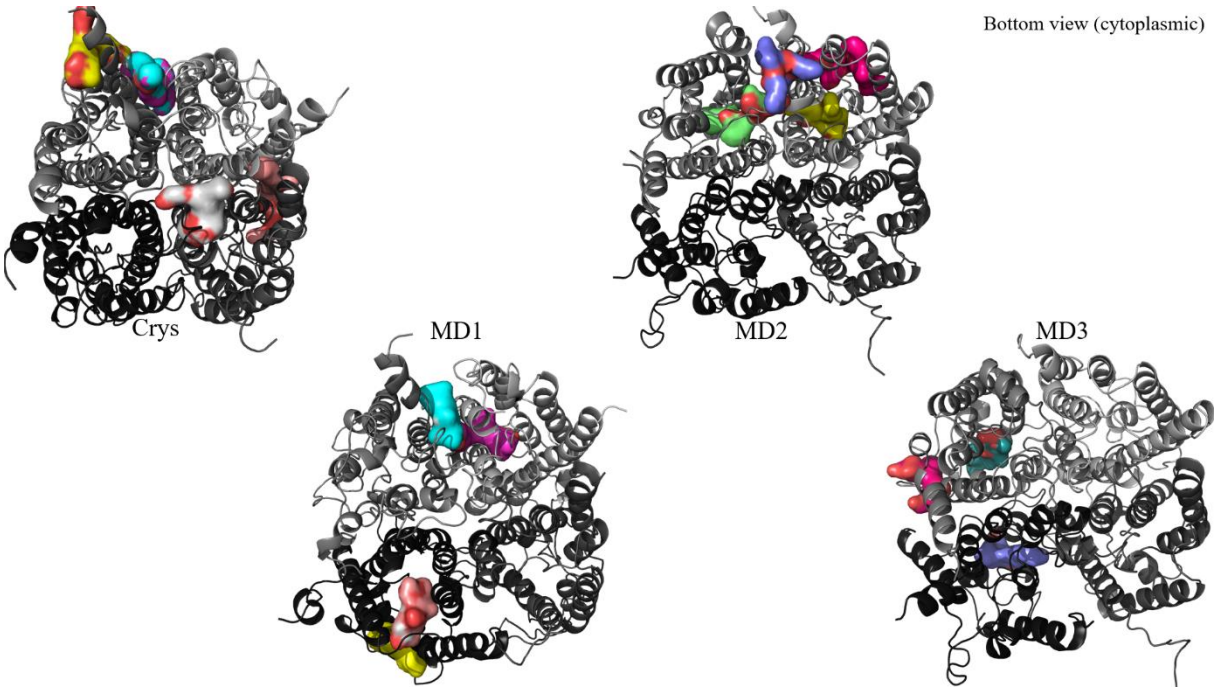

Figure A.4. Bottom view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Automated Pocket Detection using AutoLigand (colored surface).

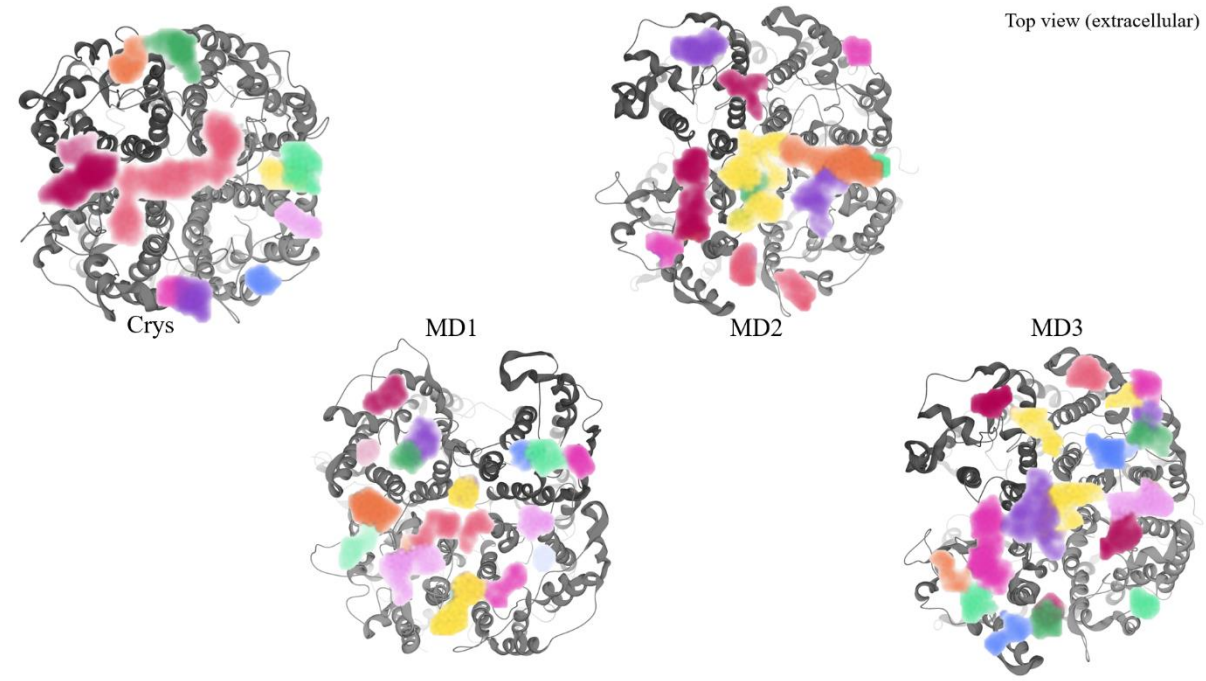

Figure A.5. Top view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Automated Pocket Detection using SeeSar (colored surface).

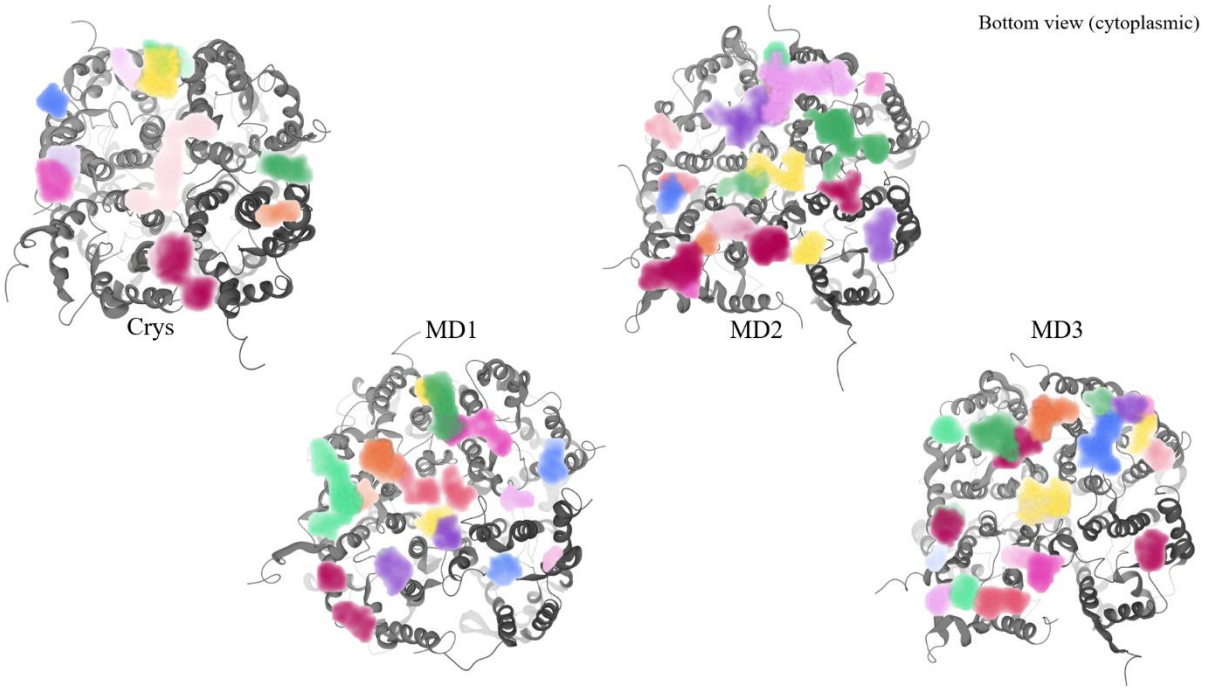

Figure A.6. Bottom view of the Crystallographic structure of AQP1 (Crys) and the three MD snapshots (MD1, MD2, and MD3) after undergoing Automated Pocket Detection using SeeSar (colored surface).

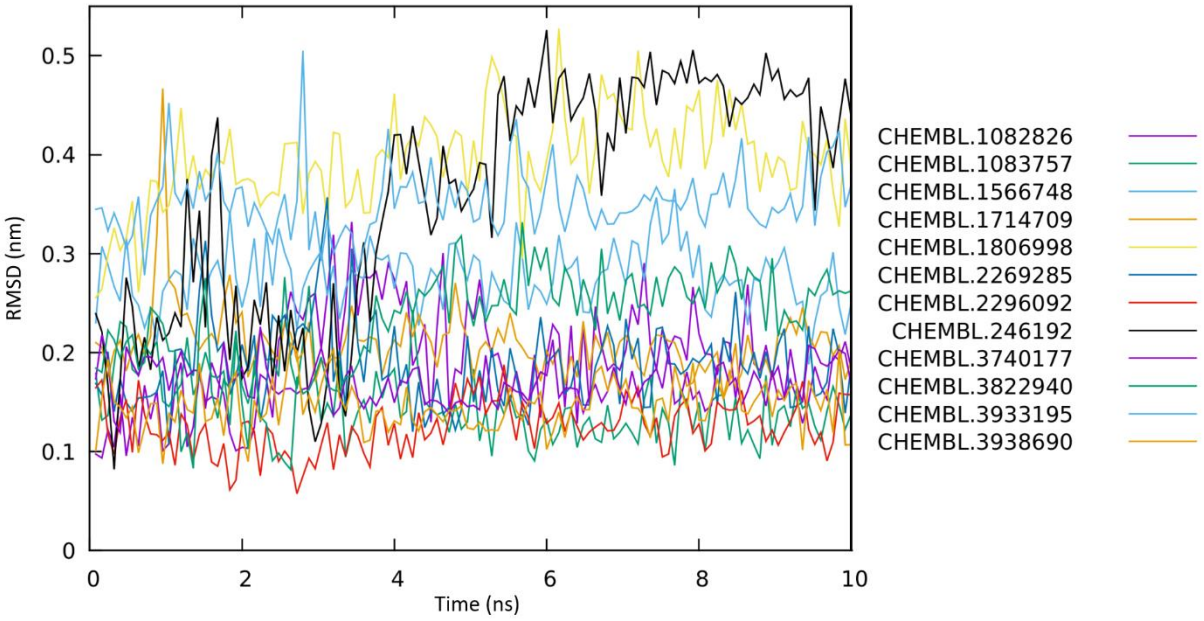

Figure A.7. Representation of the RMSD (nm) variation through simulation time (10 ns) of twelve compounds bound at the B_NPA binding site.

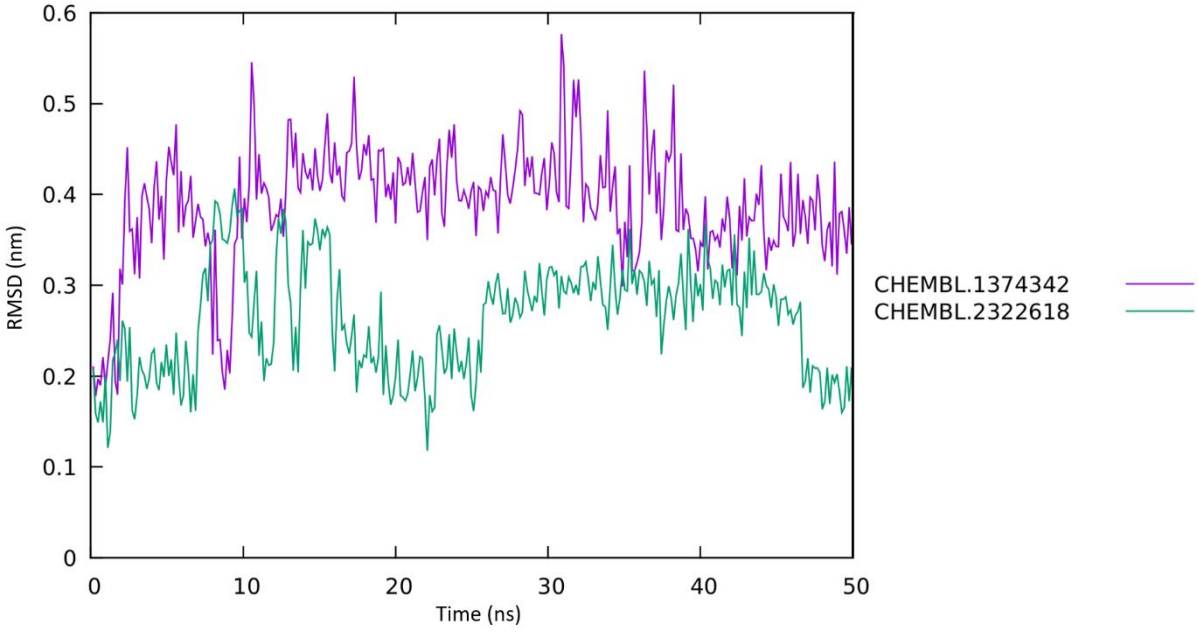

Figure A.8. Representation of the RMSD (nm) variation through simulation time (50 ns) of two compounds bound at the B_NPA binding site.

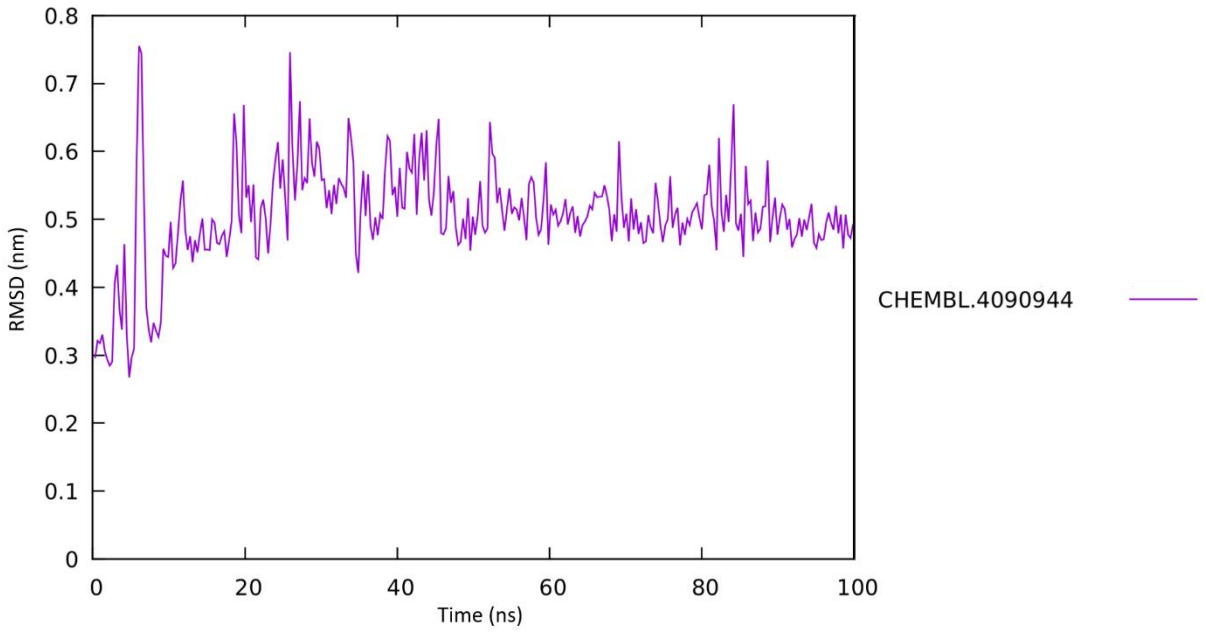

Figure A.9. Representation of the RMSD (nm) variation through simulation time (100 ns) of one compound bound at the B_NPA binding site.

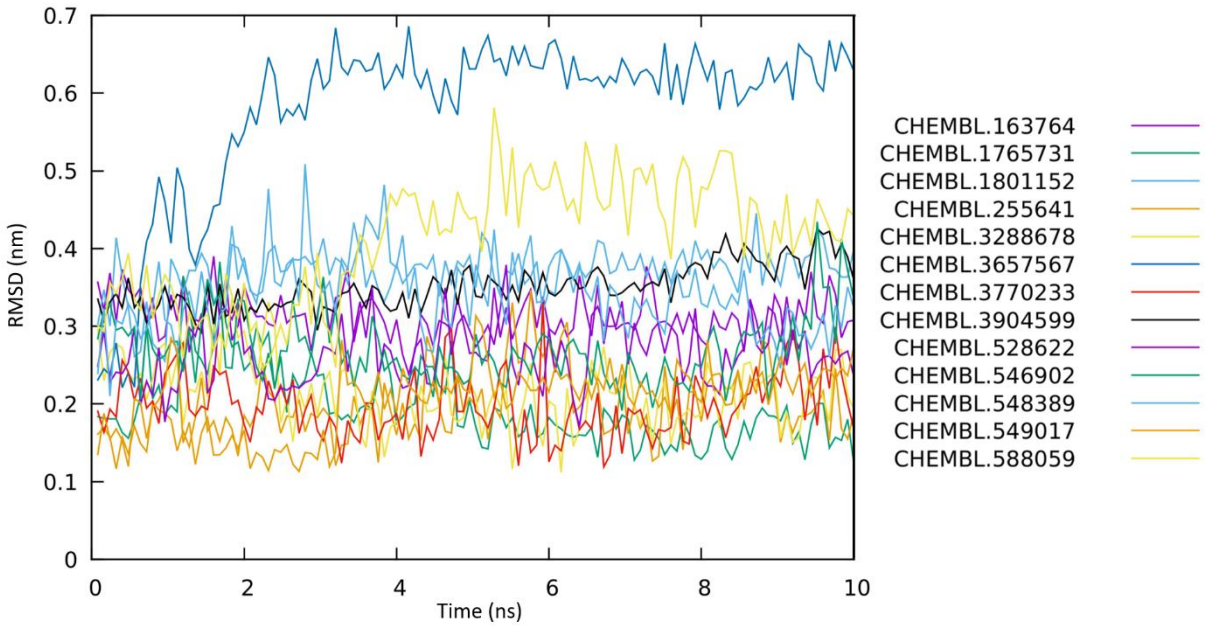

Figure A.10. Representation of the RMSD (nm) variation through simulation time (10 ns) of thirteen compounds bound at the C_ar/R binding site.

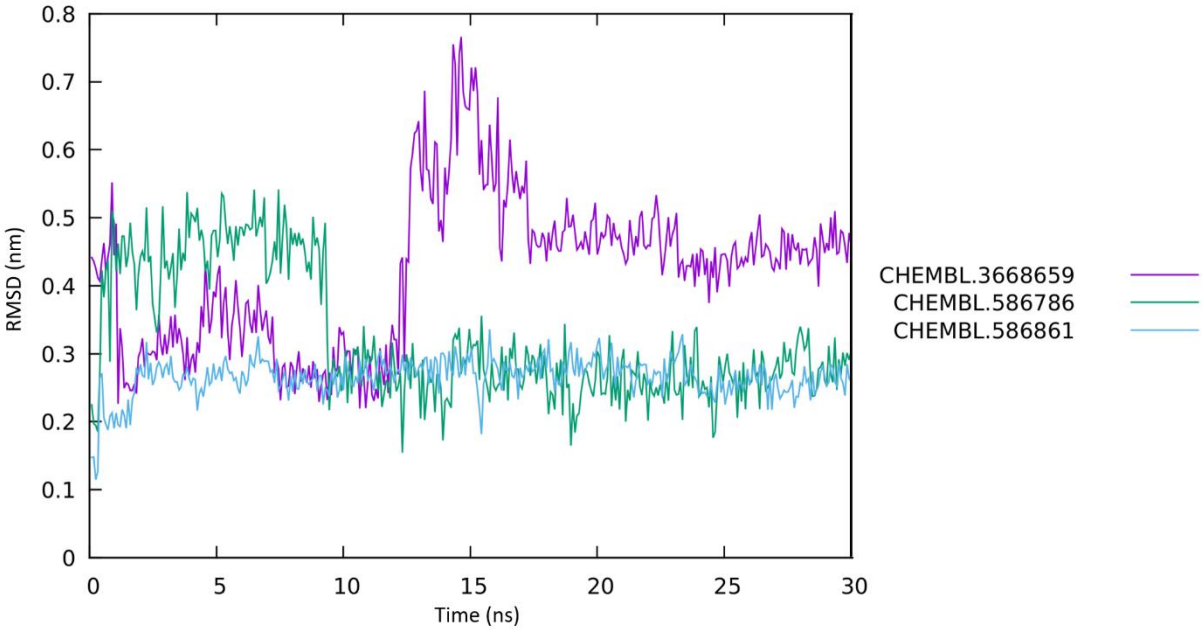

Figure A.11. Representation of the RMSD (nm) variation through simulation time (30 ns) of three compounds bound at the C_ar/R binding site.

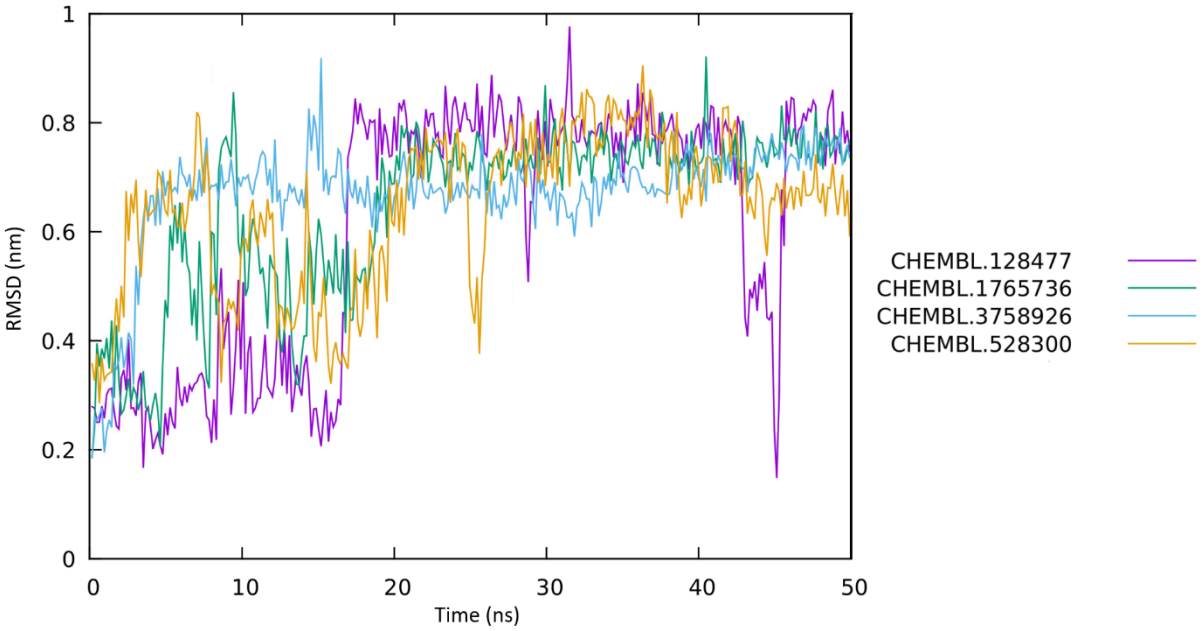

Figure A.12. Representation of the RMSD (nm) variation through simulation time (50 ns) of four compounds bound at the C_ar/R binding site.

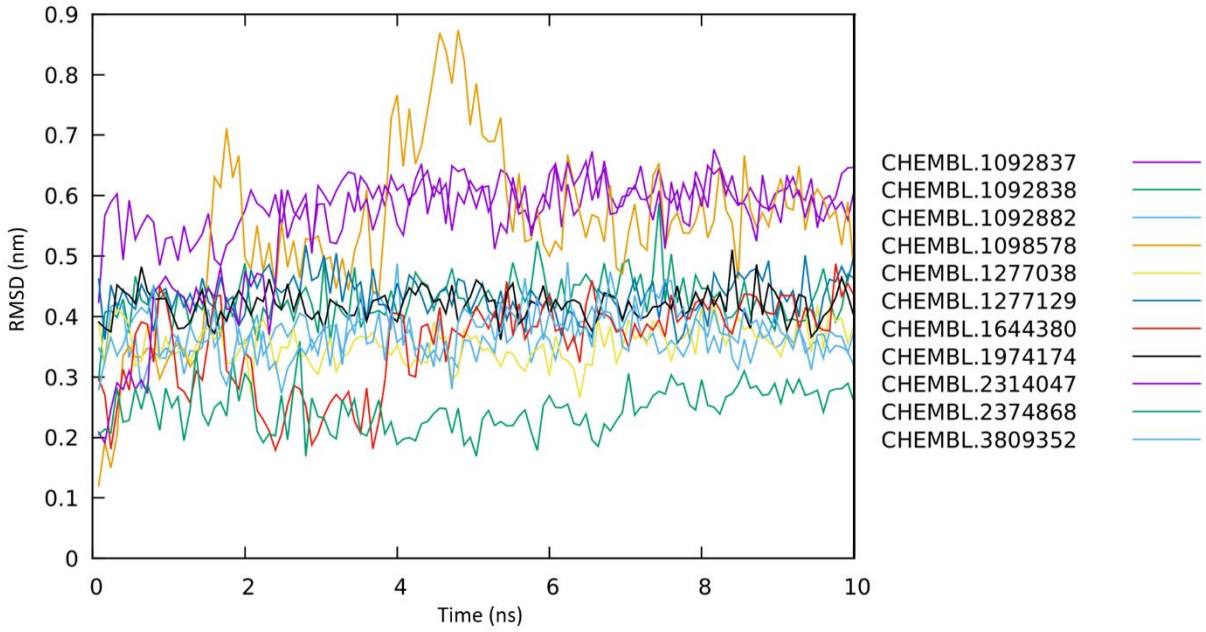

Figure A.13. Representation of the RMSD (nm) variation through simulation time (10 ns) of eleven compounds bound at the C_NPA binding site.

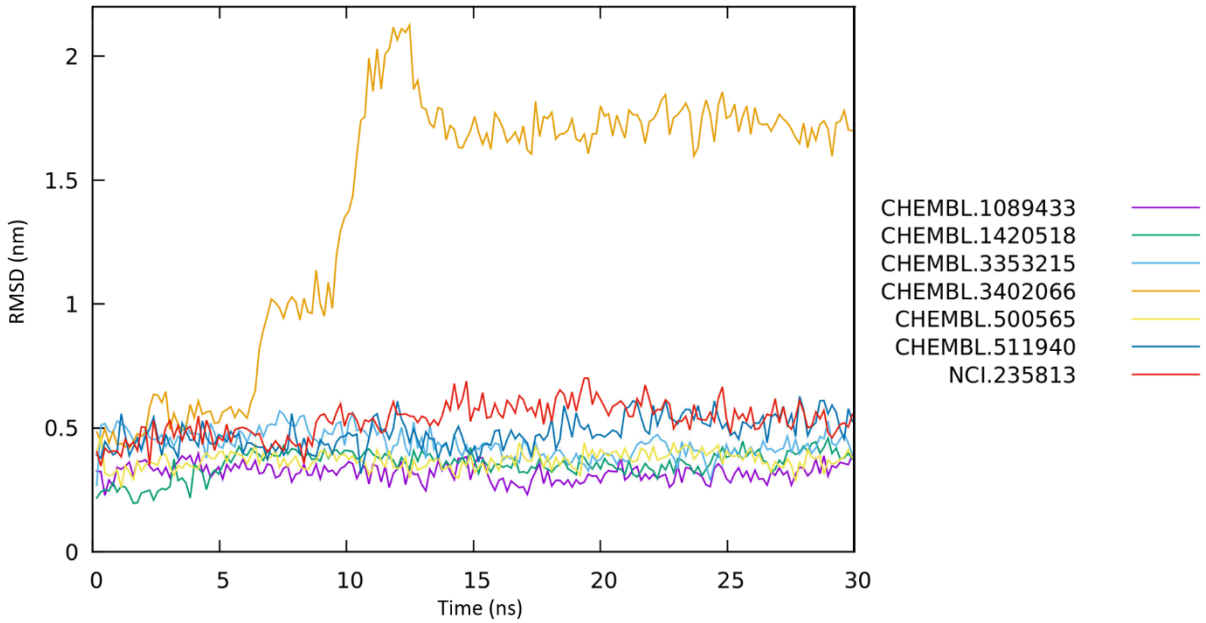

Figure A.14. Representation of the RMSD (nm) variation through simulation time (30 ns) of seven compounds bound at the C_NPA binding site.

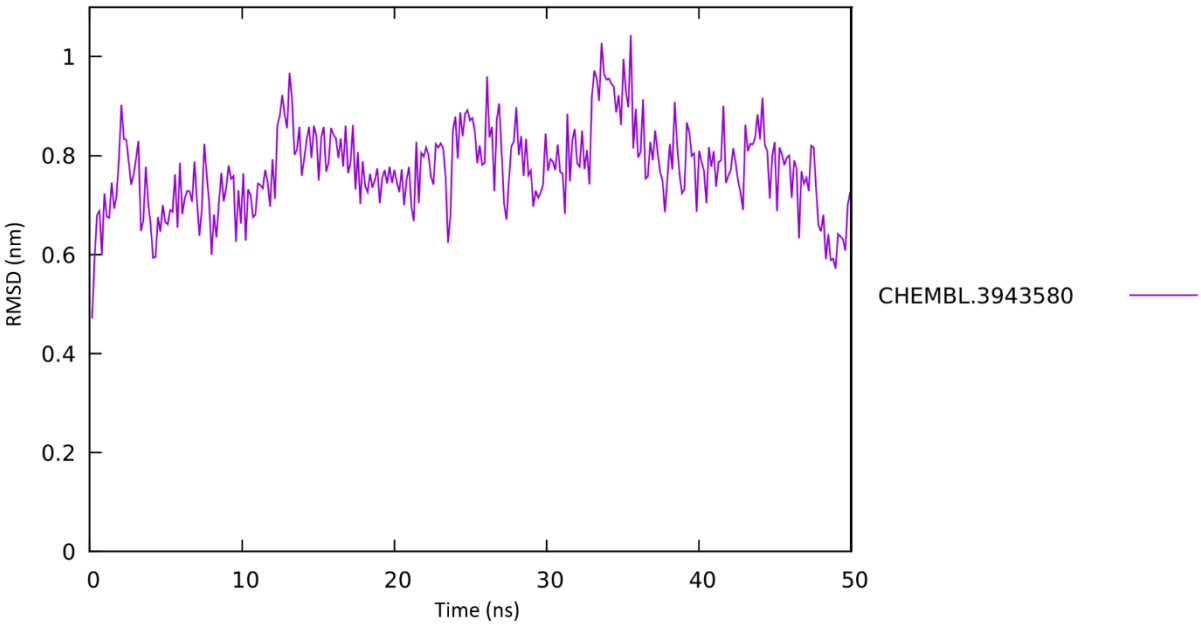

Figure A.15. Representation of the RMSD (nm) variation through simulation time (50 ns) of one compound bound at the C_NPA binding site.

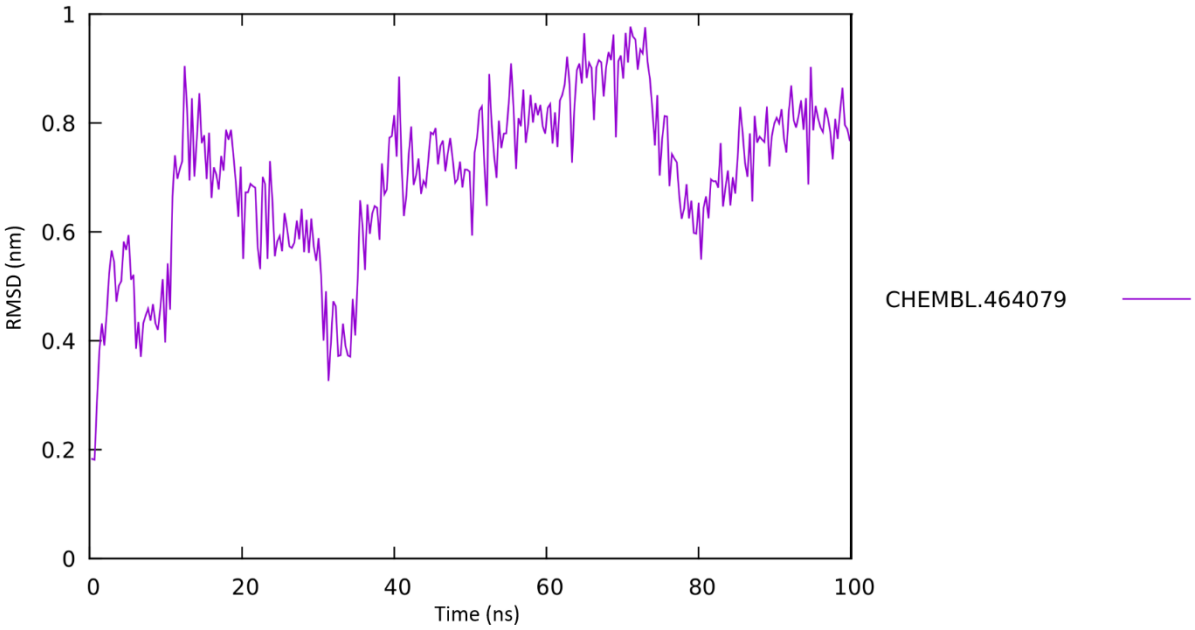

Figure A.16. Representation of the RMSD (nm) variation through simulation time (100 ns) of one compound bound at the C_NPA binding site.

Table A.1. Molweight, cLogP, cLogS, Hydrogen Acceptor, Hydrogen Donors, Rotatable Bonds, and Aromatic Rings values, for the top 5 compounds, based on the MM/PBSA ranking, for the B_NPA binding site.

B_NPA							
Ligand	Mol. Weight (daltons)	cLogP	cLogS	H-Acceptors	H-Donors	Rotatable Bonds	Aromatic Rings
CHEMBL4090944	481.59	4.23	-5.30	8.00	0.00	4.00	2.00
CHEMBL3740177	490.51	4.13	-3.57	8.00	1.00	4.00	3.00
CHEMBL1083757	440.55	3.31	-5.72	6.00	1.00	2.00	4.00
CHEMBL1383242	465.44	4.14	-9.10	8.00	0.00	3.00	5.00
CHEMBL4060285	481.59	4.23	-5.30	8.00	0.00	4.00	2.00

Table A.2. Molweight, cLogP, cLogS, Hydrogen Acceptor, Hydrogen Donors, Rotatable Bonds, and Aromatic Rings values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_ar/R binding site.

C_ar/R							
Ligand	Mol. Weight (daltons)	cLogP	cLogS	H-Acceptors	H-Donors	Rotatable Bonds	Aromatic Rings
CHEMBL255641	435.49	3.72	-6.71	9.00	3.00	7.00	5.00
CHEMBL3904599	488.55	3.27	-8.88	8.00	2.00	5.00	5.00
CHEMBL3668659	492.54	4.02	-7.58	9.00	3.00	5.00	4.00
CHEMBL128477	486.53	4.70	-7.90	8.00	2.00	7.00	6.00
CHEMBL1801152	481.55	4.41	-7.50	7.00	2.00	6.00	3.00

Table A.3. Molweight, cLogP, cLogS, Hydrogen Acceptor, Hydrogen Donors, Rotatable Bonds, and Aromatic Rings values, for the top 5 compounds, based on the MM/PBSA ranking, for the C_NPA binding site.

C_NPA							
Ligand	Mol. Weight (daltons)	cLogP	cLogS	H-Acceptors	H-Donors	Rotatable Bonds	Aromatic Rings
CHEMBL500565	484.59	1.91	-4.31	7.00	2.00	1.00	0.00
CHEMBL2374868	484.67	4.53	-6.24	5.00	0.00	0.00	0.00
CHEMBL464079	494.55	2.01	-5.88	8.00	2.00	3.00	3.00
CHEMBL2314047	453.54	4.22	-6.09	6.00	0.00	4.00	4.00
CHEMBL1092882	488.51	2.01	-6.71	9.00	2.00	2.00	4.00

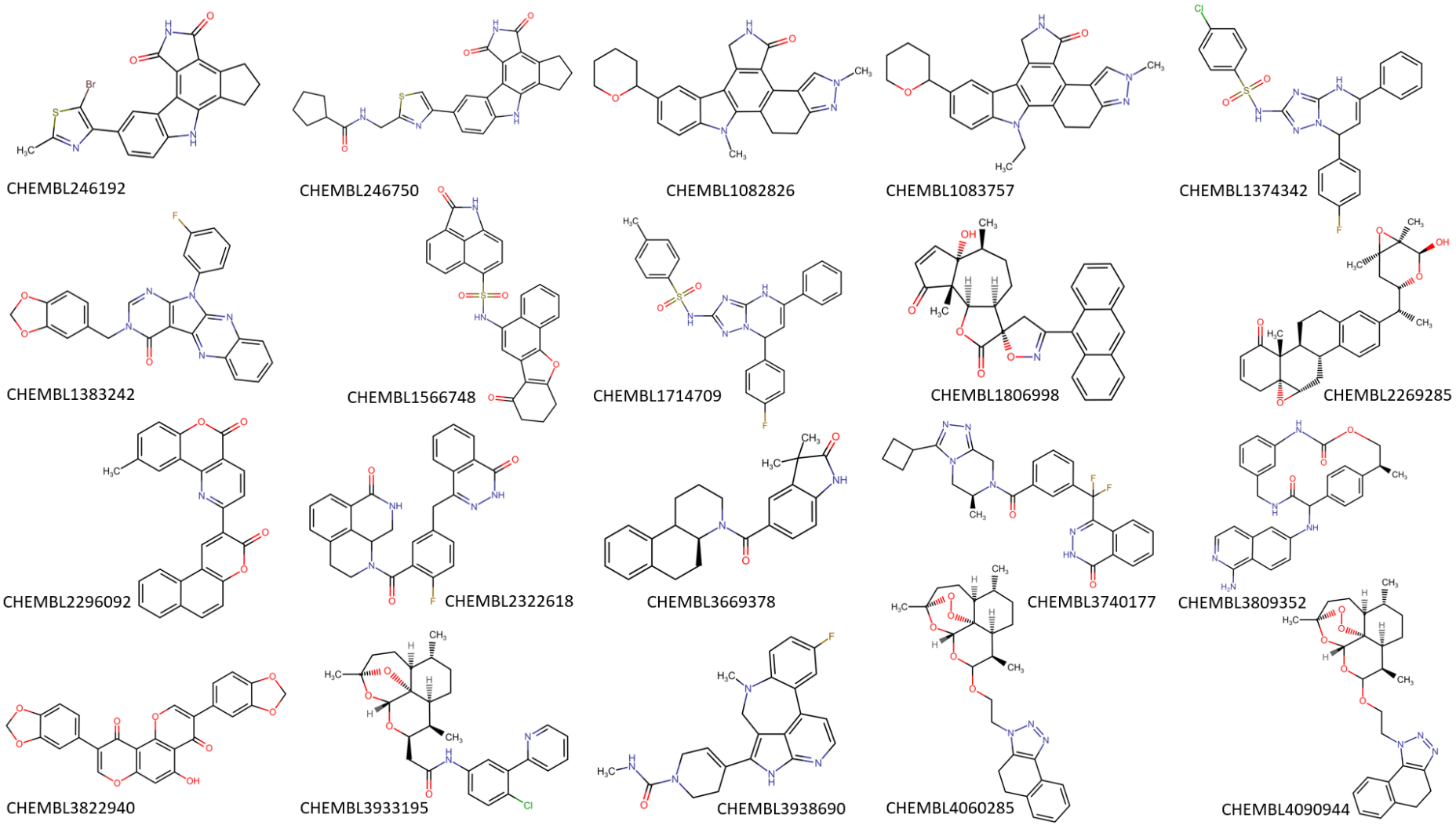

Figure A.17. Top 20 compounds for the B_NPA binding site.

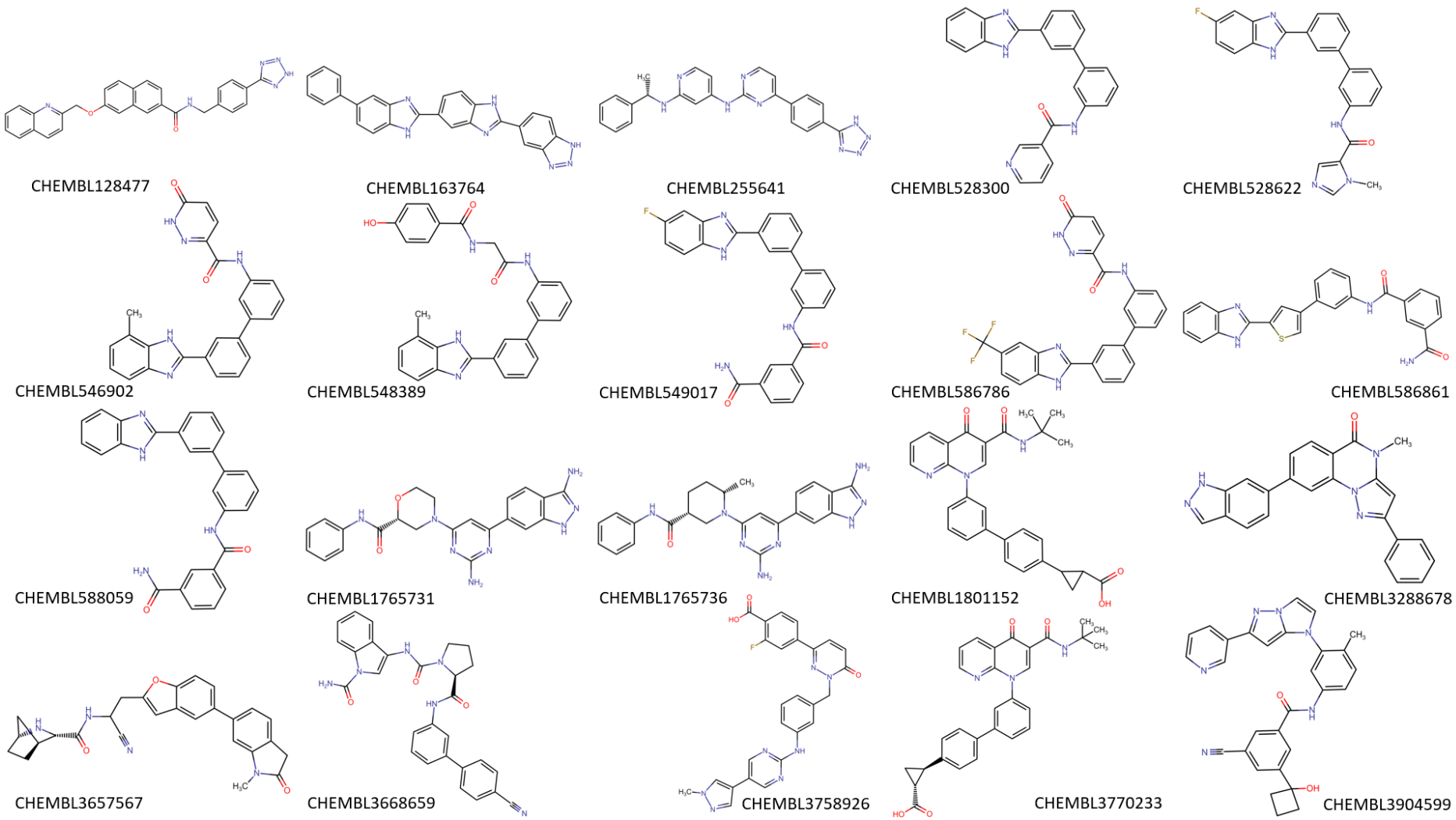

Figure A.18. Top 20 compounds for the C_ar/R binding site.

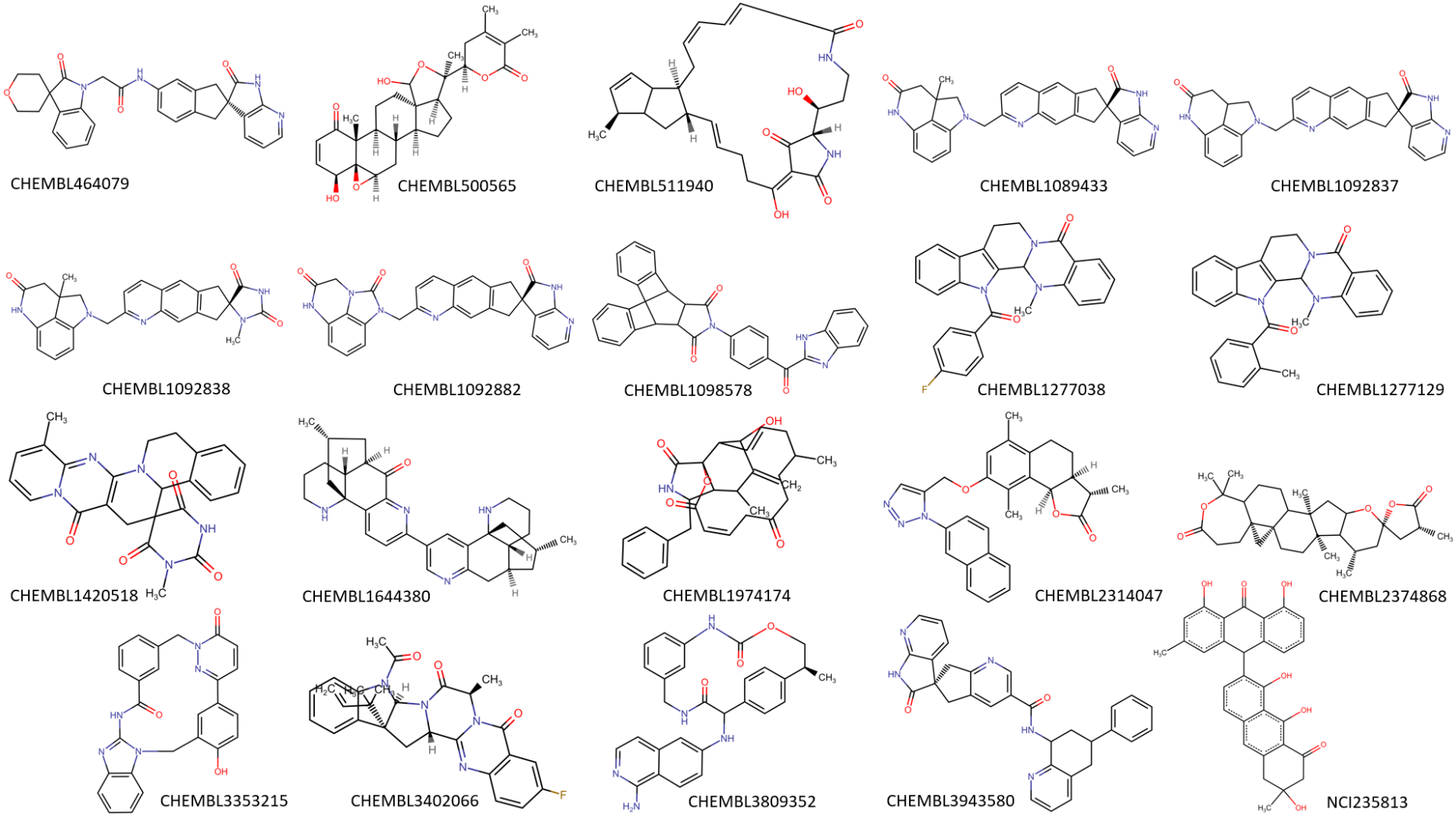

Figure A.19. Top 20 compounds for the C_NPA binding site.