UNIVERSITAT DE BARCELONA

Technological developments in Virtual Screening for the discovery of small molecules with novel mechanisms of action

Marina Miñarro Lleonar

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

Technological developments in Virtual Screening for the discovery of small-molecules with novel mechanisms of action

Marina Miñarro Lleonar

2022

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PROGRAMA DE DOCTORAT EN RECERCA
DESENVOLUPAMENT I CONTROL DE MEDICAMENTS

TECHNOLOGICAL DEVELOPMENTS IN VIRTUAL SCREENING FOR THE DISCOVERY OF SMALL MOLECULES WITH NOVEL MECHANISMS OF ACTION

Memòria presentada per Marina Miñarro Lleonar per optar al títol de doctor per la universitat de Barcelona

Xavier Barril Alonso
Director de tesis i tutor

Marina Miñarro Lleonar
Doctorand

Marina Miñarro Lleonar

2022

Als meus pares,

AGRAÏMENTS

Primer de tot vull agrair al Xavi, per donar-me la oportunitat d'entrar en aquest grup a fer el treball de final de màster i per animar-me a fer el doctorat. Gràcies per ajudar-me en tot el que ha fet falta i ensenyar-me tantes i tantes coses. Has sigut un mentor excepcional.

Al Carles, per tots els consells i ajuda que m'has donat sobretot amb la part experimental dels projectes.

M'agradaria agrair també al Jordi, encara que haguem només pogut coincidir durant dos anyets ha sigut un plaer haver treballat amb tu. Tan debò haguessis arribat abans!

Al Sergi, per estar allà els primers mesos que vaig entrar en el laboratori, per supervisar-me i ajudar-me a tirar endavant.

Serena, gracias por estar siempre allí cuando necesitaba consejo o echar unas risas

Álvaro, he tingut la sort d'estar present des del teu primer dia al laboratori i t'he pogut veure créixer i convertir-te en una peça clau. Se que t'espera un futur brillant!

Dani, moltes gràcies per tots els cafès, birres i cotilleos, per confiar en mi quan necessites parlar amb algú i sobretot per fer els dies infinitament més divertits.

I a tothom amb qui he tingut el plaer de compartir aquests anys: Míriam, Moira, Maciej, Salvo, Patricia, Varbina, Beste, Carlos, Andrea, Roger, Morena, Juan, Dylan, Alex, Bego i Ainoa. Thanks for creating this amazing workspace, you made this experience even more special!

També vull agrair a aquella gent que encara que no hagi estat present durant el dia a dia, m'heu ajudat tant o més. Jesús y Sara, por todas las cenas, tardes de juegos de mesa y esquiadas, me habéis ayudado a desconectar cuando me hacía falta y a volver con más fuerza. To Hanna, because since day one you have been the best cheerleader, and thanks for all the cheese dates! A la Marta per se una font inacabable d'historietes i companya de tonteries. A tots els Bioinfos i als Uni's crew,

per tots aquest anys d'amistat i sempre trobar moments per retrobar-nos.

A la meva família i sobretot als meus pares, perquè sense ells aquesta tesis si que no hagués sigut possible. Gràcies per tots els ànims, consells, i les tonelades de sopa i pollastre arrebossat. Sempre m'heu sapigut aconsellar i m'heu ajudar en tot el que heu pogut. Us estimo molt.

I per acabar, a tu Adrià, perquè no hagués pogut compartir aquesta aventura amb ningú més. Hi ha hagut moments bons i moments durs, però hem sobreviscut a dues tesis doctorals! Estic impacient per saber quina serà la nostre següent aventura junts. T'estimo.

ABSTRACT

Advances in structural and molecular biology have favoured the rational development of novel drugs through structure-based drug design (SBDD). Particularly, computational tools have proven to be rapid and efficient tools for hit discovery and optimization. The main motivation of this thesis is to improve and develop new methods in the area of computer-based drug discovery and to study challenging targets. Specifically, this thesis is focused on docking and Virtual Screening (VS) methodologies to be able to exploit non-standard sites, like protein-protein interfaces or allosteric sites, and discover bioactive molecules with novel mechanisms of action.

First, I developed an automatic pipeline for binding mode prediction that applies knowledge-based restraints and validated the approach by participating in the CELPP Challenge, a blind pose prediction challenge.

The aim of the first VS in this thesis is to find small molecules able to not only disrupt the RANK-RANKL interaction but also inhibit the constitutive activation of the receptor. With a combination of computational, biophysical, and cell-based assays we were able to identify the first small molecule binders for RANK that could be developed into treatment for Triple Negative Breast Cancer.

When working with novel targets, or with non-standard mechanisms of action, the relationship between binding and the biological response is unpredictable, because the latter depends on a multitude of unkown factors such as the function of the particular allosteric site, relationships with other proteins, cellular localization, etcetera. For this reason, in the next project we tested the applicability of the combination of ultrahigh-throughput VS with low-throughput high content assay. This allowed us to characterize a novel allosteric pocket in PTEN and also describe the first allosteric modulators for this protein.

Finally, as the accessible Chemical Space grows at a rapid pace, we developed an algorithm to efficiently explore ultra-large Chemical Collections using a Bottom-up approach. We prospectively validated the approach in BRD4 and identified novel BRD4 inhibitors with an affinity comparable to advanced drug candidates for this target.

Table of Contents

AGRAÏMENTS ..... **vii**

ABSTRACT ..... **ix**

List of Figures ..... ***xv***

List of Tables ..... ***xvii***

Abbreviations ..... ***xix***

Statement of Contributors ..... ***xxi***

1 Introduction ***3***

1.1 The Journey to Drug Discovery **5**

1.2 Computer-aided Drug Design in Early Drug Discovery **7**

1.3 Principles of Molecular Recognition in Protein-Ligand Binding **9**

1.3.1 Specific Protein-Ligand Interactions 10

1.3.2 Protein Flexibility 12

1.4 Principles of Drug-Receptor modulation **13**

1.5 Towards a comprehensive drug screening strategy: from in-vitro to in-silico **17**

1.5.1 The revolution of High-Throughput screening 17

1.5.2 Finding Hits in Virtual libraries: Virtual Screening 18

1.5.3 Fragment-Based drug discovery (FBDD) 19

2 Objectives ***23***

2.1 Main Objective **25**

2.2 Specific objectives **25**

3 Methods ***27***

3.1 Background on Molecular Dynamics simulations **29**

3.1.1 Molecular Dynamics Simulations with Mixed Solvents 30

3.1.2 Dynamic Undocking 31

3.2 Background on Molecular Docking **32**

3.3 Development of an automatic pipeline for participation in the celpp challenge **35**

3.3.1 Candidate preparation 35

3.3.2 Ligand preparation. 35

3.3.3 Selection of similar proteins, druggable pockets, and ligand retrieval 35
3.3.4 Ligand similarity and maximum common substructure calculation..36
3.3.5 Generation of Pharmacophoric Restraints....37
3.3.6 Molecular Docking....37
3.3.7 Pose Selection....38

3.4 Targeting rank receptor as a novel therapeutic strategy for triple negative breast cancer....39
3.4.1 Computational Methods....39
3.4.1.1 Homology modelling....39
3.4.1.2 Molecular Dynamics....39
3.4.1.3 Druggability Prediction....40
3.4.1.4 Analysis of Druggable Cavities During MD....40
3.4.1.5 Virtual Screening....41
3.4.1.6 Dynamic Undocking....42
3.4.2 Experimental methods....43
3.4.2.1 Surface Plasmon Resonance....43

3.5 Targeting pten with a combination of target-based and phenotypic screening approaches....45
3.5.1 Computational Methods....45
3.5.1.1 Target Selection....45
3.5.1.2 Virtual Screening....47
3.5.1.3 Molecular Dynamics with small molecules....49
3.5.2 Experimental methods....50
3.5.2.1 Cell culture conditions....50
3.5.2.2 Cell line characterization by Western Blot....50
3.5.2.3 Cell survival assay....51
3.5.2.4 PTEN phosphatase assay....51
3.5.2.5 pAkt ELISA....52
3.5.2.6 Surface Plasmon Resonance....53

3.6 Bottom-up exploration of the chemical space....54
3.6.1.1 Searching for Fragments in Enamine Real database and ZINC20 54
3.6.1.2 Protein Structure Selection and Preparation....54
3.6.1.3 Docking the Fragment Library....55
3.6.1.4 Filtering of Docking Results based on properties of described active Fragments....55
3.6.1.5 Fragment Clustering using Chemical Checker signatures....56
3.6.1.6 MMGBSA of Cluster Representatives....56
3.6.1.7 Dynamic Undocking....57

4 Results....59

4.1 Results for developing an automatic pipeline for the CELPP Challenge 61
4.1.1 Background on the CELPP Challenge 61
4.1.2 Overview of the Pipeline 63
4.1.3 Workflow Input Data, Data Structure and Output 64
4.1.4 Pipeline Development 65
4.1.4.1 Blast Results 65
4.1.4.2 Ligand Similarity 66
4.1.4.3 Docking Method Selection 66
4.1.4.4 Pipeline Effectiveness and Processing Time 68
4.1.5 Pipeline Validation 69
4.1.6 Challenges to Address 71
4.1.6.1 Automated Protocols 71
4.1.6.2 Scoring Challenges 72
4.1.6.3 Sampling Challenges 73

4.2 Results for targeting RANK receptor as a novel therapeutic strategy for triple-negative breast cancer 79
4.2.1 Background on RANK receptor. Implications on triple-negative breast cancer. 79
4.2.2 Druggability analysis of RANK 81
4.2.3 In Silico Identification of novel small molecules binding to RANK 83
4.2.4 SPR assay confirmed binding for some computational hits 85
4.2.5 Ongoing Cell-based experiments 91

4.3 Results for Targeting PTEN with a combination of target-based and phenotypic screening approaches 94
4.3.1 Background: Target-Based Drug Discovery vs Phenotypic Screening 94
4.3.2 Target Selection and Druggability Study of potential Tumor suppressors 96
4.3.3 Background on Phosphatase and Tensin homolog (PTEN): functions and regulation 98
4.3.4 Identification and characterization of a novel allosteric site in PTEN 100
4.3.5 Virtual Screening using Pharmacophoric Restraints 101
4.3.6 CMP1 Induces morphological changes in HCT116 PTEN +/- cell line 105
4.3.7 CMP1.3 induces polyploidization of HCT116 PTEN(+/-) 108
4.3.8 SPR confirmed binging for some of the computational hits including CMP1 and CMP1.3 109
4.3.9 CMP1 and CMP1.3 cause a morphology change independently of PTEN lipid phosphatase activity 110
4.3.10 Assessing the binding mode stability for CMP1, CMP1.2 and CMP1.3 with MD 111

4.4 Results for Bottom-Up Exploration of the Chemical Space 113
4.4.1 Background on Chemical Space Exploration 113
4.4.2 Identification of novel fragments that bind to BRD4 115
4.4.3 Fragment Growing and Experimental Validation 117

5 Discussion 121

6 Conclusions 137

6.1 General conclusions 139

6.2 Specific conclusions 139

Bibliography 141

Appendix A: Supplementary Information 167

Appendix B: Publications 193

List of Figures

Figure 1 The process of Drug Discovery 5
Figure 2 Steps in Early Drug Discovery 7
Figure 3 Frequency distribution of the most common non-covalent interactions observed in protein-ligands from the PDB..... 10
Figure 4 Geometries of π-stacking interactions. 11
Figure 5 Types of models of molecular recognition 13
Figure 6 Mechanisms of Drug-Receptor modulation with small molecules..... 14
Figure 7 Targeted Protein Degradation (TPD) mechanism of action.... 17
Figure 8 Different Strategies for Fragment to Lead 21
Figure 9 Example of an equation used to approximate atomic forces during Molecular Dynamics. 29
Figure 10 The Quasi-Bound State. 31
Figure 11 DUck set up 32
Figure 12 Example of a cavity generated with rDock..... 34
Figure 13 Surface Plasmon Resonance Assay representation..... 43
Figure 14 CellTiter-Glo assay representation. 51
Figure 15 Workflow employed for pose prediction..... 64
Figure 16 Analysis of the targets from previous CELPP weeks..... 65
Figure 17 Differences in best pose predicted for target 5p8y from CELPP week 33..... 67
Figure 18 Relation between RMSD and the MCSS 68
Figure 19 Overall view of validation set cases 71
Figure 20 Median RMSD for the submitted pose compared to the best pose generated by the pipeline. 73
Figure 21 PDB 6ok9 with the pocket detected by 3decision..... 74
Figure 22 Example of ligand binding in a PPI interface 75
Figure 23 Predictions for PDB 6dfo and hiTanimoto Receptor. 76
Figure 24 Example of side chain orientation in different PDB structures..... 77
Figure 25 Exmple of systems with other molecules in the binding site.. 78
Figure 26 Representative structure of TNFSF ligand-receptor complexes..... 80
Figure 27 RANK structural analysis..... 82
Figure 28 Pharmacophoric restraints and pocket environment..... 83
Figure 29 SPR plots for compound 1. 86
Figure 30 SPR plots for compound 2. 86
Figure 31 SPR plots for compound 3..... 87
Figure 32 SPR plots for compound 4. 87
Figure 33 SPR plots for compound 5. 88
Figure 34 SPR plots for compound 6. 88
Figure 35 SPR plots for compound 7. 89
Figure 36 SPR plots for compound 8. 89
Figure 37 SPR plots for compound 9. 90
Figure 38 SPR plots for compound 10. 90
Figure 39 HCC1954 response to Compound 1 and compound 2. 92

FIGURE 40 HEK-BLUE RESPONSE TO COMPOUND 1 AND COMPOUND 2. 92
FIGURE *41* HEK-BLUE RESPONSE FOR COMPOUND 8. 93
FIGURE 42 COMPARISON OF TARGET-BASED AND PHENOTYPIC SCREENING APPROACHES IN EARLY DRUG DISCOVERY 94
FIGURE 43 MDMIX RESULTS FOR PTEN, VHL, SMARCA4, MEN1, AND PTPRK. 98
FIGURE 44 SUMMARY OF PTEN BIOCHEMICAL FUNCTIONS, REGULATION, PHYSIOLOGICAL ROLE AND STRUCTURE 99
FIGURE 45 DRUGGABILITY STUDY OF PTEN 101
FIGURE 46 VIRTUAL SCREENING PROTOCOL USED FOR THE IDENTIFICATION OF COMPOUNDS BINDING TO THE NOVEL ALLOSTERIC SITE OF PTEN 102
FIGURE 47 MDMIX RESULTS FOR THE ALLOSTERIC SITE OF PTEN USING ISOXAZOLE AS SOLVENT. 103
FIGURE 48 SMARTS REPRESENTATION USED TO SEARCH IN ENAMINE REAL DB. 103
FIGURE 49 CELL TITTER-GLO ASSAY 106
FIGURE 50 IMMUNOHISTOCHEMISTRY FOR HCT116 PTEN (+/-) TREATED WITH CMP1, CMP1.2 AND CMP1.3. 107
FIGURE 51 IMMUNOHISTOCHEMISTRY FOR HCT116 PTEN (+/-) TREATED WITH CMP1.3. 109
FIGURE 52 PTEN ACTIVITY ASSAY 111
FIGURE 53 ANALYSIS OF THE LONG MD SIMULATIONS OF PTEN APO AND PTEN BOUND TO CMP1, CMP1.2, AND CMP1.3. 112
FIGURE 54 SIZE COMPARISON FOR DIFFERENT CHEMICAL SPACES 114
FIGURE 55 SCREENING STRATEGY FOR THE EXHAUSTIVE EXPLORATION OF THE FRAGMENT SPACE 115
FIGURE 56 CAVITY AND PHARMACOPHORIC RESTRAINTS USED DURING THE VIRTUAL SCREENING 115
FIGURE 57 A) DISTRIBUTION OF ΔG BIND FOR THE ACTIVE CHEMBL SET, THE NON-ACTIVE CHEMBL SET AND THE FRAGMENTS OBTAINED IN THE VS..... 117
FIGURE 58 PIPELINE USED FOR EXHAUSTIVE SEARCH OF THE FRAGMENT SPACE AND FOR SCAFFOLD GROWING 118
FIGURE 59 NUMBER OF COMPOUNDS FOUND AT DIFFERENT STAGES FOR THE BOTTOM-UP EXPLORATION USING DIFFERENT STARTING POINTS..... 119
FIGURE 60 RESULTS FOR THE DIVERSITY ANALYSIS OF THE EVOLVED COMPOUNDS 120

List of Tables

TABLE 1 RMSD RESULTS OBTAINED USING DIFFERENT DOCKING METHODS 67
TABLE 2 STATISTICS OF THE PIPELINE IMPLEMENTATION CELPP WEEKS 69
TABLE 3 RMSD VALUES AND PERCENTAGE OF CASES FOR EACH DOCKING PROTOCOL.....70
TABLE 4 COMPARISON BETWEEN THE SUBMITTED DOCKING METHOD VS. THE METHOD THAT YIELDS THE BEST RESULT.....76
TABLE 5 RESULTS OF W_{QB} AND SCORE.INTER FOR THE 27 PRIORITIZED COMPOUNDS. 84
TABLE 6 K_D FOR THE 10 COMPOUNDS THAT SHOWED BINDING TO RANK RECEPTOR IN SPR. 85
TABLE 7 TUSON P-VALUE AND DRUGGABLE POCKETS FOUND WITH FPOCKET FOR THE CANDIDATE TARGETS 97
TABLE 8 RESULTS OF WQB AND SCORE.INTER FOR THE 14 PRIORITIZED COMPOUNDS. 104
TABLE 9 VALUES OF W_{QB} AND SCORE.INTER FOR COMPOUNDS 1, 1.2 AND 1.3.. 108
TABLE 10 SPR RESULTS FOR PTEN COMPOUNDS 110
TABLE 11 WQB VALUE FOR THE 6 SELECTED FRAGMENTS 117

Abbreviations

ADMET	Absortion Distribution Metabolism Excretion Toxicity
BRD4	Bromodomain-containing protein 4
CADD	Computer-Aided Drug Design
CC	Chemical Checker
CELPP	Continuous Evaluation of Ligand Pose Prediction
D3R	Drug Design Resource
DSF	Diferential Scanning Fluorimetry
DUck	Dynamic Undocking
EMA	European Medicines Agency
EMT	Epithelial-Mesenchymal Transition
F2L	Fragment to Lead
FBDD	Fragment-Based Drug Discovery
FBS	Fetal Bovine Serum
FDA	Food and Drug Administration
GPU	Graphycal Processing Unit
hiResApo	High Resolution Apo
hiResHolo	high Resolution Holo
HTS	High-throughput screening
HTVS	High-throughput virtual screening
LMCSS	Largest Maximum Common Substructure
MCSS	Maximum Common Substructure
MD	Molecular Dynamics
MDmix	Mixed-solvent Molecular Dynamics
MMGBSA	Molecular Mechanics Generalized Born Surface Area
MOA	Mechanism of Action
MW	Molecular Weight
PDB	Protein Data Bank
PDD	Phenotypic Drug Discovery
POI	Protein of Interest
PPI	Protein-Protein Interaction
PTEN	Phospatase and Tensin Homolog
RANK	Receptor Activator of Nuclear factor $\varkappa$ B

RANKL	Receptor Activator of Nuclear factor κ B Ligand
RMSD	Root Mean Square Deviation
SAR	Structure-Activity Relationship
SBDD	Structure-Based Drug Discovery
SMCSS	Smalles Maximum Common Substructure
sMD	steered Molecular Dynamics
SPR	Surface Plasmon Resonance
TDD	Target-Based Drug Discovery
TNBC	Triple-negative Breast Cancer
TNFR	Tumor Necrosis Factor Receptor
TPD	Targeted Protein Degradation
TR-FRET	Time-Resolved Fluorescence Resonance Energy Transfer
VS	Virtual Screeening

Statement of Contributors

The work presented in this thesis would not have been possible without the contributions of many brilliant people, to whom I am very grateful. Due to the collaborative and interdisciplinary nature of the projects presented on this thesis, some of the data presented was acquired with the help of other people. These people are: **Roger Castaño (R.C), Álvaro Serrano(A.S)**, **Arnau Comajuncosa (A.C)**, **Andrea Bertran (A.B), Dr. Isabel Puig (I.P)**, **Dr.Eva Gonzalez Suarez (EG)** and her lab members **Lucia de Andrés Gordo(L.A), Sergio Velasco (SV) and Patricia Gonzalez Santamaria (PG)**.

In the following I state the contribution of other people to the data presented by chapter:

Targeting RANK receptor as a novel therapeutic strategy for triple-negative breast cancer: RANK Cell experiments were performed by **L.A** and **P.G** and **SV**.

Targeting PTEN with a combination of target-based and phenotypic screening approaches: R.C performed some of the Cell survival assays, the PTEN phosphatase assay, pAKT ELISA, and the Surface Plasmon Resonance assay. **Dr. I.P** performed Immunohistochemistry and took the corresponding pictures.

Bottom-Up Exploration of the Chemical Space: A.C performed the Fragment Clustering using the Chemical Checker signatures. **A.S** performed the Fragment Growing protocol. **A.B** performed all the experimental assays related to this project.

1 Introduction

1.1 THE JOURNEY TO DRUG DISCOVERY

The history of drug discovery and medicine goes back a few thousand years B.C when people would evaluate the medicinal values of some herbs and apply them as a treatment for various diseases. One would think that with all the recent technological advances and knowledge we have nowadays, the drug discovery process would be relatively easy and straightforward. Unfortunately, this is far from reality. Today, Drug Discovery is an enormous field of investigation characterized by highly complex, time-consuming, expensive multidisciplinary processes, and, more often than not, unsuccessful. In fact, it is estimated that the path from the first identification of a disease-related target to the release of a drug in the market lasts an average of 12 years [1] and costs more than 1 billion dollars [2,3].

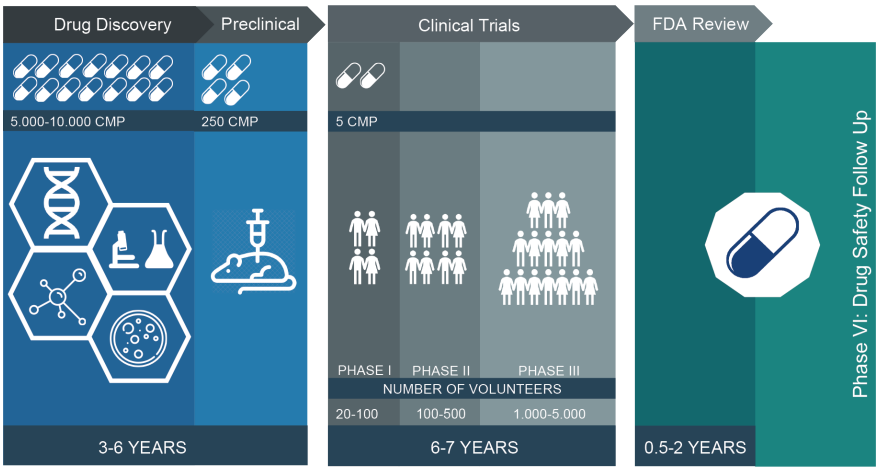


Figure 1 The process of Drug Discovery

The start of the drug discovery process reflects an unmet medical need for a particular disease, whether it would be a rare disease or one that is extremely prevalent (like COVID-19) with mild or deadly outcomes. We can summarize this process into 5 different stages [**Figure 1**]:

Early Drug Discovery: It usually starts in an academic environment, where researchers collaborate to identify and characterize potential Targets to treat a specific disease. Once the Target is properly Identified, potential leads are found and optimized for that specific target (usually done in the industry). The lead molecule must elicit the desired effect on the specific biological target implicated in the disease. The different stages of early drug discovery will be discussed in more detail in the next section.

Preclinical Research: The molecules identified in early Drug Discovery are refined, optimized, and extensively tested in the laboratory and in animal models. The aim of this step is to provide sufficient evidence of safety and efficacy before Clinical Trials in humans can begin.

Clinical Research: Selected clinical candidates are then taken to clinical trials. At this stage, a drug candidate has to pass through 3 different phases. In Phase I the drug is tested in a small group of healthy subjects, between 20 and 100, with the aim of assessing safety, identify the dose that can be given without side effects, and studying how the substance behaves in the body. In Phase II the effectiveness, tolerability, and dosage are studied in a larger group, around 100 to 500 adult patients. In Phase III (the last phase before the possible approval of a drug), the molecule is tested in thousands of patients to confirm its effectiveness and safety in many different patients. This is the stage where most failures take place, between 2011-2020 the estimated overall likelihood of approval from Phase I is around 8%[4].

Drug Review by a regulatory authority: Before a Drug or a vaccine can be distributed to the public, it needs approval from a national regulatory authority, such as the FDA in the US and the EMA in the EC.

Post-Market Drug Safety Monitoring: Once the drug is finally commercialized a series of studies are carried out to gather more comprehensive data regarding the effectiveness and safety of the new drug.

1.2 COMPUTER-AIDED DRUG DESIGN IN EARLY DRUG DISCOVERY

Between the 1950s to the 1990s computer technology progressed at an unprecedented rate [5]. That is why in the early 1980s there was an increase of interest in the impact of computational methods applied to the pharmaceutical Industry. Millions of dollars were invested in hardware and software and the need for scientists specialized in this area grew tremendously [6]. We can define Computer-Aided Drug Design (CADD) as a broad range of theoretical and computational approaches with the aim of discovering, designing, and developing therapeutic chemical agents [7]. These computational methods are mostly (but not exclusively) applied during the early phase of drug discovery with the aim of increasing the odds of finding new compounds with desirable *in vitro* and *in vivo* properties.

It is possible to define 4 steps in the Early Drug Discovery process as seen in **Figure 2**:

Figure 2 Steps in Early Drug Discovery

Target Identification and Validation Once we decide to start a drug discovery process aimed at a specific disease, one of the most important steps is the identification of the most relevant biological entities related to the development, progression or symptoms of that disease, which will be considered as targets. A good target, besides being clinically relevant, needs to be "druggable". This means that it has to be able to bind a molecule and, upon binding, elicit a biological response that is therapeutically useful and can be measured both *in vivo* and *in vitro*. At this stage, computationally we can mine available biological data, which helps in identifying and prioritizing potential disease targets [8]. Also,

Machine Learning techniques can be applied to predict drug targets by means of analyzing proteomics and chemogenomic data [9]. Once we have identified our target of interest, we need to assess its ability to bind drug-like molecules, which may be referred to as "druggability" or "bindabilility". This can be achieved by means of pocket detection programs, which look for cavities on the protein surface, or with other methods that try to identify binding hot spots (e.g. mixed-solvents MD).

Hit Identification and Validation During this step, compounds are identified (hits) through experimental or virtual screening of libraries of molecules. Here we refer as hit a compound that shows activity against the target of interest on the primary screening assay. Some of the most used computational tools for hit Identification are docking and molecular dynamics.

A common approach to identifying hits is virtual screening, where a large number of compounds are filtered with structure-based, ligand-based, or hybrid approaches [10], such that only a small subset of compounds will be tested in the experimental assay.

Another technique that has been gaining popularity is fragment screening, which is based on the use of much smaller size molecules with the aim of eventually evolving these initial fragment hits into potent drug-like molecules [11].

Hit-to-Lead When a new hit is identified, validated and selected for progression, it will enter the hit-to-lead stage. At this point, the main goal is to further improve the potency. Amongst the computational approaches for chemical optimization is the Quantitative SAR (QSAR) method, which is able to predict the activity of new analogs derived from a series of active compounds [12]. Another method used at this stage is the Free Energy Perturbation (FEP), which is able to calculate relative binding affinities for a congeneric series of ligands [13].

Lead Optimization Finally, at this stage, other drug-relevant properties are improved. The aim is to produce a series of analogs with a particular focus on improving pharmacokinetic and ADMET properties (Absorption, Distribution, Metabolism, Excretion, and

Toxicity). Although there is still a long road for improvement, new machine learning models are being developed with promising results [14–16].

1.3 PRINCIPLES OF MOLECULAR RECOGNITION IN PROTEIN-LIGAND BINDING

In any drug discovery process the end goal is that when we administer a drug, it will eventually bind to the therapeutic target with high affinity and elicit the desired biological response. One of the main drivers of the binding event is molecular recognition. Here we refer to molecular recognition as the existence of specific attractive interactions and shape complementary between two molecules [17].

During the last decades, our understanding of the types of interactions that play a role in protein-ligand binding has advanced a lot, however, the consistently accurate prediction of compound affinities still remains the Achilles' heel of CADD. We know that for a binding event to take place it needs to be associated with a negative binding free energy (ΔG), which can be seen as a measure of the stability of a protein-ligand complex or, as the binding affinity of a ligand to a given acceptor [18][**Equation 1.1**]. It depends on an enthalpic term (ΔH), defined as the changes in energy resulting from the formations of the non-covalent complex (which includes formation and rupture of interactions between the protein, the ligand and the whole molecular environment), and an entropic term (ΔS), which measures how heat energy is distributed over the thermodynamic system (related to the degrees of freedom of the system).

$$\Delta G = \Delta H - T\Delta S$$

Equation 1.1

At the moment, the use of free energy calculations is still a crude estimate of affinity useful for an enrichment of ligand candidates in virtual screenings, but not for the accurate prediction of binding affinity [19]. The main problem of the scoring methods, is that they rely on the concept of additivity, where the contributions of pairwise interactions

are treated independently of the total binding free energy. In reality, this oversimplistic representation of protein-ligand binding is not accurate as there are many factors that affect the entropy and enthalpy of the system, such as: specific intermolecular interactions, protein flexibility, the solvent effect or the role of structural waters.

1.3.1 SPECIFIC PROTEIN-LIGAND INTERACTIONS

The Protein Data Bank is one of the most important resources for structure-based drug design. With, currently 192.489 macromolecular structures, it offers a background for not only the study of structural features of biologic complexes but also for the study of nature, geometry, and frequency of atomic interactions [**Figure 3**] [20].

The most prevalent interaction is the **Hydrophobic interaction**, which is the contact between carbon, halogen, or sulfur atoms, with distances ranging from 3,7 to 4,4 Å. Between them, the most common one is the one formed by an aliphatic carbon in the receptor and an aromatic carbon in the ligand [21]. From the protein side, leucine, valine, isoleucine, and alanine side chains are the most frequently engaged in hydrophobic interactions.

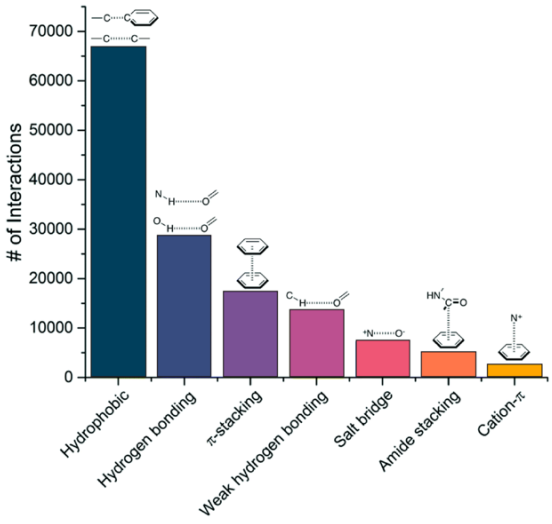


Figure 3 Frequency distribution of the most common non-covalent interactions observed in protein-ligands from the PDB. Adapted from [20]

Hydrogen bonds are the second most observed interactions, formed between two electronegative atoms (i.e nitrogen and oxygen) that share a hydrogen. Because of their sharp distance (2,7 to 3,2 Å [21]) and angular dependencies, Hydrogen bonds provide the defined geometries in biological complexes and contribute to the specificity of molecular recognition [22,23]. They can contribute to the binding energy between -1,5 to 4,7 kcal/mol [21], depending on the environment around the hydrogen bond. Usually, a hydrogen bond that is buried in a hydrophobic cavity has a higher contribution to the binding free energy than a bond that is solvent-exposed. In addition, it has also been shown that water-shielded hydrogen bonds can act as kinetic traps, slowing down the release of the ligand from the protein-ligand complex [24].

π-stacking occurs between two aromatic rings and can be considered a special case of hydrophobic interaction. It is caused by intermolecular overlapping of p orbitals in **π** conjugated systems. Depending on the arrangement of the rings we can have different types of geometries: face-to-face, edge-to-face, and parallel displaced [**Figure 4**]. The distance range between 3,4-3,8 Å [21]. The aminoacids capable of doing this interaction with the ligand are Phenylalanine, Tyrosine, Tryptophan, and Histidine.

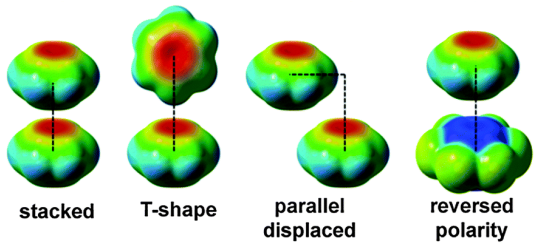


Figure 4 Geometries of π-stacking interactions. Adapted from [25]

Weak hydrogen bonds are hydrogen bonds where carbon is the hydrogen bond donor. It only occurs on conjugated carbons in strong electro-withdrawing environments. Although the magnitude of the interaction is about one-half the strength of the hydrogen bond [26], it plays an important role in processes such as protein folding [27], the

interaction of nucleic acids with proteins [28], enzyme catalysis [29], and the stabilization of protein-ligand binding complexes [30].

Salt bridges are contacts formed between positively charged and negatively charged atoms. They are similar to hydrogen bonds, but the distances are usually shorter (median of 2,79 Å) [21]. Their contribution to the binding energy is also highly dependent on the context, and, although they are stronger than neutral H-bonds, their contribution is masked due to the large energetic penalty for desolvating charged groups [31,32].The aminoacids capable of making salt bridges are: Aspartate and Glutamate (negatively charged), and Arginine, Lysine, and Histidine (positively).

Amine stacking is the interaction occurring between an amide group and an aromatic ring.

Cation-π is the interaction formed by an aromatic ring and a positively charged nitrogen atom.

There are still other interactions that are not as common, like the halogen bonds, and others that have more of a covalent character, namely the coordination complexes with metalloproteins. But there are also important indirect interactions, such as the water bridges, where a network of structurally stable water molecules is mediating the interaction between the protein and the ligand.

1.3.2 PROTEIN FLEXIBILITY

Conformational change plays a big role in many biological processes such as molecular recognition, enzymatic activity, and allosteric modulation. The first model postulated by Fischer in 1894 was the "lock and key" model [33] where both the receptor and the ligand were treated as rigid bodies. If these two rigid entities shared a complementary shape, then the interaction would be possible [**Figure 5a**]. However, we know that this model is too simple to recreate what is happening in reality [34]. Both proteins and ligands are dynamic entities that undergo certain conformational changes upon binding (i.e. differences in apo and holo protein conformations). It was not until 1958 that Koshland

proposed the induced-fit model to explain the protein conformational changes during the binding process [35]. This model suggests that, when an enzyme binds to its substrate, it optimizes the interface through physical interactions to form the final complex structure [**Figure 5b**]. This model was able to explain, for example, why are there ligands buried in protein-binding sites [36], however, it can not explain big conformational changes such as backbone motions, domain rearrangements, or disorder-to-order transitions for very flexible proteins [37]. In the 1990s the conformational-selection model was proposed by several researchers [38]. This model suggests that when the receptor is in its unbound state, it fluctuates among multiple conformational states, with their occupancy probabilities explained by their relative free energies according to the Boltzmann distribution. Only a subset of those conformations have the ability to bind to the ligand, and when bound, the distribution of probabilities is shifted towards these states [**Figure 5c**].

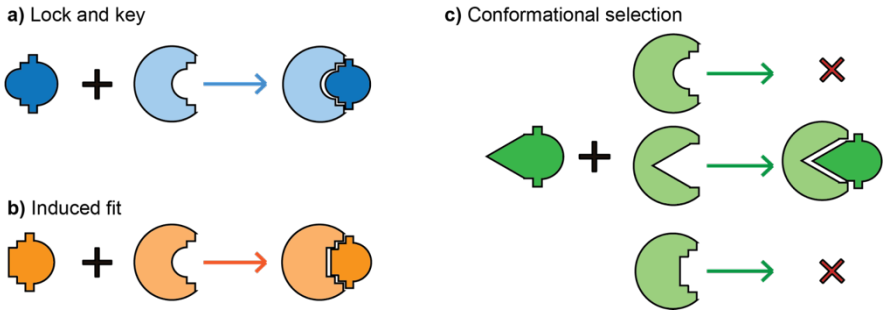


Figure 5 Types of models of molecular recognition a) Lock & key b) Induced fit, the ligand adapts the cavity conformation of the receptor c) Conformational selection: there is an ensemble of different conformations of the protein, the ligand only binds to some of the conformations shifting the equilibrium towards that conformation.

1.4 PRINCIPLES OF DRUG-RECEPTOR MODULATION

At the beginning of the early drug discovery process, there is one crucial step, assessing the druggability of the target and defining the cavity where our ligand will possibly bind. Based on where a drug binds, we

can differentiate them into orthosteric and allosteric binders. But perhaps, rather than where the small molecule is binding, the most important thing is understanding how the small molecule modulates the receptor upon binding. In this section, we will overview some of the most common mechanisms of drug-receptor modulation.

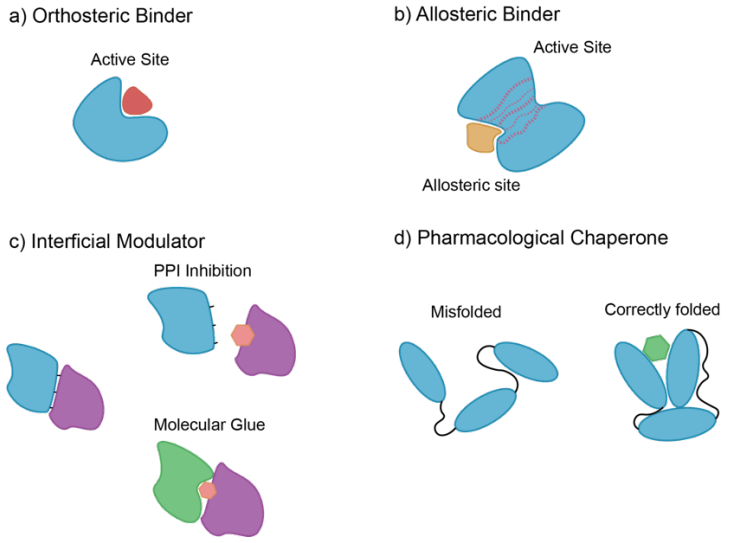


Figure 6 Mechanisms of Drug-Receptor modulation with small molecules

Orthosteric binders bind at the active site of the receptor, competing with the natural substrate [**Figure 6a**]. Traditionally, almost all the drugs that are on the market are orthosteric drugs. In the case of enzymes, orthosteric binders are almost invariably inhibitors. In the case of receptor proteins, we can classify the drugs as agonists and antagonists, depending on the effect that the drug will have over the receptor. ***Agonism*** occurs when a drug binds to a receptor and causes a biological response. If the drug is able to trigger a maximal response of the receptor, they are called *full agonists*. If they are only able to generate a fraction of the possible response of the receptor, they are called *partial agonists*. There are some occasions where an agonist binds to a receptor and causes an opposite response, these substances are called *inverse agonists*. ***Antagonism*** happens when a drug binds to the receptor and blocks or interferes with the ability of an agonist/natural substrate to

activate the receptor. The most common type of antagonism is the *reversible competitive*, where a drug competes with an agonist for its binding site and limits the amount of agonist that can bind.

One of the main problems when dealing with orthosteric binders is selectivity. Usually, binding sites across protein families are highly conserved and when we administer the drug it will bind to the target protein as well as the binding sites of homologous proteins. Besides, the function of these similar proteins can vary a lot, which will lead to unwanted side effects [39].

Allosteric binders bind into a site different from the active site of the receptor [**Figure 6b**]. When an effector binds at one site of the molecule, it causes a perturbation that leads to a functional change at another site by means of alteration of the shape or dynamics [40]. One of the most recent postulated models for allosterism is the ensemble model [41,42]. Similarly to the conformational selection model, the binding of the allosteric ligand, causes a perturbation that shifts the distribution of the conformational states.

Allosteric binders have advantages over orthosteric ligands in terms of potency, because we remove the competition with the endogenous substrate, and in terms of selectivity, because allosteric sites are less conserved across protein families. Additionally, they allow a much more precise modulation of the protein activity.

On the downside, allosteric drug discovery is challenging. Contrary to orthosteric drugs that bind to a known active site, allosteric sites are often unknown and the effects upon drug binding are difficult to predict, and even a small change in the drug-target interaction may lead to different downstream effects[43]. Complications in allosteric drug discovery can go further into the drug discovery process. Because of the high divergence rate of allosteric sites in species homologs, the translation from the initial pharmacological studies to animal models of disease can be challenging [44].

Interfacial modulators are molecules able to disrupt or stabilize PPI either by direct competition at the interface or via allosteric

destabilization binding at a protein site different from the interface [45] [**Figure 6c**]. The manipulation of PPI is an attractive mechanism of action as they are critically important in disease-specific molecular mechanisms and pathways [46]. PPI are challenging to target as they usually have small, shallow, or exposed cavities and the molecule has to compete with the protein partner, which usually have a much larger interaction area [47]. The first approved PPI inhibitors were peptides or molecules derived from natural products which led to a poor oral bioavailability and low cell-permeability. However, thanks to the greater understanding of PPI structure and energetics, PPI inhibitors have reached "drug-like" properties. Many researchers suggested that the ΔG_{bind} of PPIs was often not evenly distributed across the entire buried surface area, but rather concentrated on energetic "hot spots" that have a large contribution to ΔG_{bind} [48]. By placing the molecules at these sites PPI inhibitors with lower molecular mass could be created [49,50]. Another type of Interfacial modulators that have been gaining popularity are the ones able to stabilize PPI, called *Molecular glues (MG)*. They provide novel and additional interactions between proteins partners, over-stabilizing a (usually) pre-existing complex. This MOA has been exploited in the field of Targeted Protein Degradations (TPD), where the MG promotes the recruitment of neo-substrates by an E3 ligase facilitating the ubiquitination and posterior degradation via the proteasome.

Pharmacological chaperones. Most disease-causing mutations affect protein processing, folding, pH stability, protein aggregation, defective transport to the lysosomes, and many post-translational modifications. It has been proposed that small molecules, *pharmacological chaperones (PC)*, could be used to restore the folding, trafficking and biological activity of these non-functional proteins [**Figure 6d**]. PC stabilizes the native conformation or promotes the correct folding of the protein, resulting in an enhancement of its activity [51].

PROteolysis TArgeting Chimera molecules (PROTACs) together with MG are the two main drivers of TPD. In this case, PROTACs are hetero bifunctional molecules that bind simultaneously to a protein of interest (POI) and to an E3 ubiquitin ligase, inducing the formation of

a non-natural ternary complex that promotes the ubiquitination of the POI which is later recognized and degraded by the proteasome[52] [Figure 7]. PROTACs rely not only on the affinity of the individual warheads for the E3 or the POI, but in some cases, also on cooperativity (the formation of the ternary complex). Although TPD is a relatively new field, several PROTACs have already entered clinical trials.

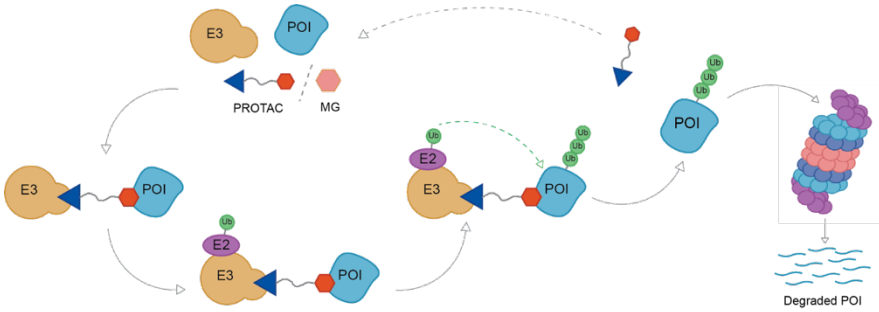


Figure 7 Targeted Protein Degradation (TPD) mechanism of action

1.5 TOWARDS A COMPREHENSIVE DRUG SCREENING STRATEGY: FROM IN-VITRO TO IN-SILICO

1.5.1 THE REVOLUTION OF HIGH-THROUGHPUT SCREENING

Current drug discovery relies on the screening of large chemical libraries against an extracellular or intracellular target to identify novel compounds with the desired MOA. However, with the genomic revolution in the 1990s, which increased enormously the number of novel targets, and the concomitant increase in the number of available chemical compounds raised the need for an automated screening process. High-throughput screening (HTS) provides a practical method to screen thousands to millions of compounds in miniaturized *in vitro*

assays very quickly against multiple targets [53]. HTS is still the gold standard in the pharmaceutical industry today, having approximately >50% success rates [54]. Very early on, people noticed that rather than the target type, the main component affecting the success rate was the content, size, and quality of the compound collections used. That is why pharma companies cherish their proprietary databases and are not available for the scientific community to exploit [55]. There have been many initiatives to use HTS outside the pharma industries [56]. Still, the cost and logistics of handling large amounts of compounds make this method unaffordable for smaller organizations and academic labs.

Thus, researchers are always finding new ways to not only expand the explored chemical space but also in an efficient and cost-effective manner. In this context virtual screening (VS) and fragment-based drug discovery (FBDD) have established themselves as the future of drug discovery.

1.5.2 FINDING HITS IN VIRTUAL LIBRARIES: VIRTUAL SCREENING

The first publication about Virtual Screening appeared in 1997, where by using molecular docking and a database of 2500 "2D molecular sketches" they successfully identified new inhibitors for trypanothione reductase (TR) [57]. Since then, the use of Virtual Screening has been shown to be an excellent alternative to HTS due to its reduced cost and the ability to exploit larger chemical collections that in result raises the probability of finding more and better hits. VS is an *in silico* technique where large databases of chemical compounds are evaluated against a molecular target using computational methods. The goal of VS is to predict the binding between ligands and a molecular target and rank them according to their binding affinity [53]. Another advantage of "going virtual" is that the compounds can be tested even before being synthesized, which avoids the costly and time-consuming task of synthesis for inactive molecules.

Amongst all the methods used for Virtual screening, molecular docking is the most used technique for its relatively low computational cost and good results [58]. Despite all the advantages of docking, it has many disadvantages which will be discussed in more detail in section 3.3. Other methods used for virtual screening are, for example, machine learning classifiers.

A recent example of docking-based virtual screening campaigns are the ones targeting the main protease of SARS-CoV-2. In early 2020, major efforts were initiated to develop new drugs to treat coronavirus infections. To that aim, many researchers made use of virtual screening strategies to successfully identify inhibitors, some of them with a broad-spectrum activity against coronaviruses [59–61]

Docking-based VS can be applied to library sizes up to 10^8 [62,63], but when dealing with larger chemical collections there are limitations related to calculation time and data management. An example of such a chemical collection is ENAMINE REAL SPACE with $31x10^9$ virtually synthesizable compounds. However, this pales in comparison to other proprietary spaces, like Merck's MASSIV space [64], with 10^{20} compounds, or Pfizer's PGVL with 10^{14} [65].

There are methods to speed up the docking calculations, for example, we can use tailored VS protocols, where molecules are discarded early in the docking calculation if they do not achieve a specific energy value. Or if we have previous information about the molecular features that a ligand should have to interact with a specific receptor (pharmacophore), select only those molecules that are able to fulfil this pattern. Still, these strategies are not enough to solve all the challenges faced when searching in these massive spaces.

1.5.3 FRAGMENT-BASED DRUG DISCOVERY (FBDD)

Fragment screening was initially developed to find hit compounds where other traditional methods (HTS) fail. FBDD is based on identifying small chemical fragments that bind weakly to a biological

target, and then growing them by evolving or combining them to obtain a drug-like compound with a higher affinity.

Fragments are small organic molecules with low molecular weight and tend to bind to the biological target with low affinity, usually in the μM to mM range. In analogy to Lipinski's *"rule of 5"* to define a drug-like molecule [66], for fragments we have the *"rule of 3"* proposed by Congreve et al, which theorize that a fragment should have: (1) a MW equal or below 300 Da, (2) at most 3 rotatable bonds, (3) a logP below 3, (4) at most three hydrogen bond donor groups and (5) at most three hydrogen bond acceptor groups [67].

A smaller chemical compound means less rotatable bonds which entails for more stable interactions. The other important property of fragments is that they are *ligand efficient*, meaning that they possess a high binding affinity per heavy atom, and thus, are ideal for optimization into clinical candidates with good *drug-like* properties.

For fragment to lead (F2L) optimization there are three main approaches: Fragment growing, Fragment linking and Fragment merging. *Fragment growing* is the most used strategy, where we have an initial fragment hit and additional chemical groups are added to it to achieve a compound with *drug-like* properties. *Fragment linking* consists on the identification of multiple fragments binding to different parts of the pocket and then linking them without affecting the binding orientation and position of the initial fragments [68]. *Fragment merging* is an approach that consists of combining information of multiple chemical hits together.

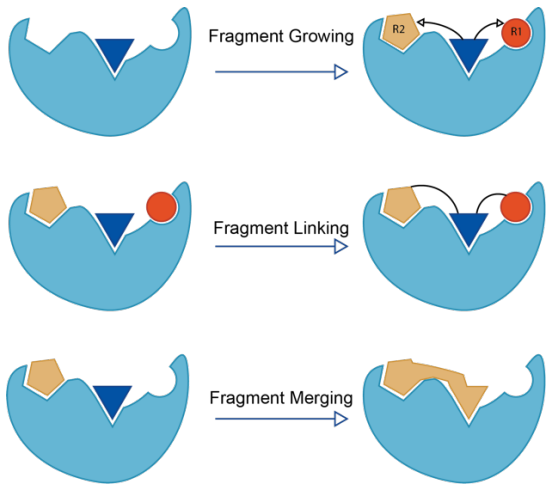


Figure 8 Different Strategies for Fragment to Lead

The interest in FBDD has grown further than finding hits that fit better challenging biding pockets, but also as an efficient way of exploring the chemical space. The chemical space grows exponentially with the number of atoms making the *drug-like* chemical space (≤35 heavy atoms) challenging to explore. However, by considering only the fragment space (≤22 heavy atoms) we can screen a bigger part of this space as it is much smaller. Then, by using F2L methods we can deeply explore the privileged areas of the *drug-like* space. This will be discussed in more detail in section 4.4.

2 Objectives

2.1 MAIN OBJECTIVE

The general objective of this work is to develop state-of-the-art methods for computer-based drug discovery and apply them to targets of pharmacological interest. In particular, we want to deepen our understanding of binding mode prediction, create new methodologies to efficiently explore the chemical space and discover bioactive molecules with novel mechanisms of action.

2.2 SPECIFIC OBJECTIVES

1. Improve docking binding mode prediction by applying knowledge-based restraints and assessing the results in blind challenges.
2. Rational discovery of small compounds to inhibit RANK protein and assess their applicability as a therapy for triple-negative breast cancer.
3. Test the feasibility of combining ultrahigh-throughput Virtual Screening with low-throughput high-content assays, validating the approach prospectively by discovering novel PTEN allosteric modulators.
4. Development of an algorithm to efficiently explore ultra-large chemical collections, validating the approach prospectively on a bromodomain protein, used as test system.

3 Methods

3.1 BACKGROUND ON MOLECULAR DYNAMICS SIMULATIONS

The initial "lock-and-key" theory postulated by Fischer in 1984 [33] has been widely abandoned in favour of binding models that not only account for conformational changes upon ligand binding [69,70], but also for the intrinsic constant protein motions[36].

Unfortunately, the calculations required to describe large systems' quantum-mechanical motions and chemical reactions are often too complex and computationally intensive for even the best supercomputers. MD simulations developed in the late 1970s [71], seek to overcome this limitation by using simple approximations based on Newtonian physics to simulate atomic motions, thus reducing the computational complexity [72]. In MD, all the forces that govern molecular systems are estimated from equations like the one shown in **Figure 9**, where bonded and non-bonded interactions are parametrized to fit quantum-mechanical calculations and experimental data.

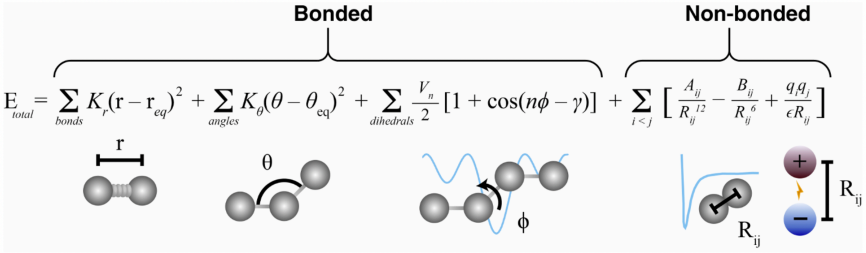


Figure 9 Example of an equation used to approximate atomic forces during Molecular Dynamics. Extracted from[72].

Many simulation programs have been developed, with the most popular being CHARMM [73], AMBER [74,75], GROMACS [75], and OpenMM [76]. Several Force Fields (the equations and parameters used to estimate de forces during MD) are commonly used in molecular dynamics simulations (i.e. AMBER [77,78], CHARMM [79], GROMOS [80] or OPLS [81]) differing in the way that they are parametrized, however, most advanced force-fields are similarly capable of reproducing experimental observables [82]

3.1.1 MOLECULAR DYNAMICS SIMULATIONS WITH MIXED SOLVENTS

When a ligand binds to a protein, it triggers a cascade of changes in the receptor and the solvent, that optimizes the packing at the interface to accommodate the ligand [83]. Being able to locate and characterize these binding sites from other areas in the protein surface is an essential step in SBDD. In this thesis, the method used to achieve this purpose is MDmix [84,85]. MDmix relies on the use of MD simulations with aqueous/organic solvent mixtures to obtain high-quality interaction maps that later can be used as a guide in ligand design.

After protein preparation, the target system is immersed in a solvent-filled truncated octahedral box constructed from replicas of a pre-equilibrated box of a selected solvent mixture. Then, the MDmix protocol performs an equilibration of the solvated system, consisting of a heating stage of 800 ps to reach 300 K in the NPT ensemble and a 1 ns stage in the NVT ensemble at 300 K. Production runs of at least 20 ns in the NPT ensemble are then carried out, storing atomic coordinates every picosecond. All non-hydrogen atoms of the protein are restrained with soft harmonic potentials (k=0.01 kcal/molÅ^2).

After the production stage, all replicas are superimposed to a reference structure. A grid with 0.5Å spacing in each direction is constructed for each one of the probes of the solvent mixture and the observed density in each grid element is compared to the expected density and converted to binding free energy (ΔG_{bind}) [**Equation 3.1**] using the inverse Boltzmann relationship.

$$\Delta G_{bind} = -k_b T \cdot ln\left(\frac{N_i}{N_0}\right)$$

Equation 3.1

Lastly, the regions of the grid with the most negative ΔG_{bind} for each probe are selected as a hotspot.

3.1.2 DYNAMIC UNDOCKING

Hydrogen bonds are not only the most important polar interaction between protein-ligand complexes but they have also proven to be key at providing structural stability to such complexes [86,87], which is the ability to form a precise and stable binding mode, thanks to the strict angular and distance dependencies between the hydrogen bond donor and acceptor. Structural stable complexes not only need to have a low ΔG_{bind}, but also present a narrow free energy minimum in the bound state [**Figure 10**]. As an example, in Figure 10 molecules 1, 2 and 3 have the same kinetics and thermodynamic constants. However, compound 1 has a steeper slope around the bound state, leading us to conclude that this compound will be more structurally stable than the other two.

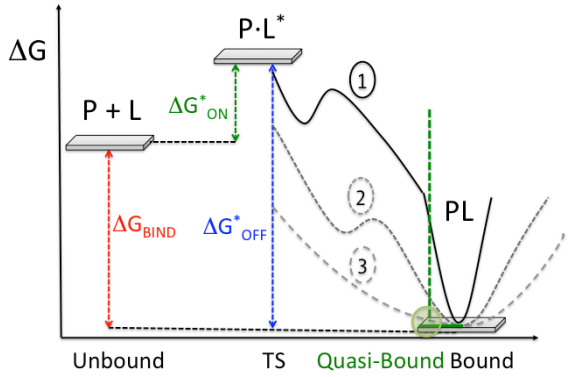


Figure 10 The Quasi-Bound State. Graphical representation of the Quasi-Bound State. Extracted from [86]

Structural stability can be quantified by displacing the ligand from its equilibrium position to a quasi-bound (QB) state where a preselected H-bond interaction has been broken.

Dynamic Undocking (DUck) makes use of Steered Molecular Dynamics (sMD) to pull the ligand from the original position to the QB state by using the distance between the selected hydrogen bond interacting atoms as the collective variable and monitoring the force required in the process to calculate the work (W_{QB}).

The DUck protocol starts by defining the hydrogen bond of interest and selecting the protein's residues that define the H-bond's local environment, referred to hereinafter as the chunk [**Figure 11a**]. After production of the chunk, DUck performs, automatic ligand parameterization in MOE, minimization, equilibration, and two sMD simulations (at two different temperatures, 300 K and 325 K), in which the distance between the interacting atoms in the ligand and protein is increased from 2.5 to 5.0Å [**Figure 11b**], and if the W_{QB} value (work necessary to break the H-bond) in the previous step reaches a predefined threshold, then the system is sampled by a short unbiased MD simulation, after which the sMD protocol is run again with the resulting new structures. The simulations are discontinued if the measured W_{QB} in any replica was below the threshold. Once a sufficient number of sMD runs are completed, the lowest W_{QB} value obtained is selected.

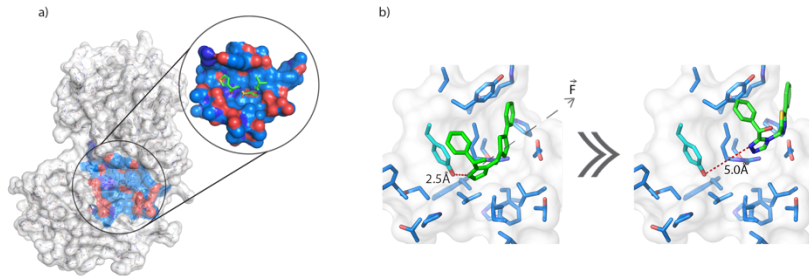


Figure 11 DUck set up a) An example of a protein Chunk b) sMD simplified.

3.2 BACKGROUND ON MOLECULAR DOCKING

Predicting how a ligand binds to a target is an essential step for SBDD, and molecular docking has become a standard tool for drug discovery [88,89]. The Docking protocol can be described as a combination of two components, a search algorithm that generates low-energy ligand conformations (poses), and a scoring function able to rank the poses generated by the search algorithm. These models can be used to interpret and guide ligand design well before the structure of the protein-ligand complex can be experimentally determined.

The perfect search algorithm would exhaustively elucidate all possible binding modes between the receptor and the ligand. However, even taking a really simple system, performing such an extensive search would require huge amounts of computational time [90] and would be infeasible for larger molecules. As a consequence, only a small amount of the conformational space is sampled by applying constraints and approximations in an attempt to locate the global minimum as efficiently as possible.

However, generating a plethora of binding modes is ineffective without a model to rank them. The scoring function (SF) should be able to distinguish the experimental binding modes from all other modes explored through the searching algorithm. We can distinguish between different classes of scoring functions: force-field based, knowledge-based, and empirical. Force-field based SF estimates the free energy with a weighted sum of several energy terms comprising inter- and intra-molecular interactions (eg. van der Waals, electrostatic interactions, hydrogen bonds). Knowledge-based SF are designed to reproduce experimental structures. They rely on the statistical analysis of intermolecular interactions within large 3D structural databases of complexes [91]. Finally, empirical scoring functions estimate the free energy of binding using a weighted sum of parameters [92]. Although proven to be a powerful asset in drug discovery campaigns, docking programs do not always find accurate ligand poses when compared to the experimental solution.

In this thesis rDock [93] is used in all the docking calculations. rDock makes use of a combination of Stochastic search techniques to generate low energy ligand poses. The standard docking protocol generates the ligand pose using 3 stages of a Genetic Algorithm [94] followed by a low temperature Monte Carlo and Simplex Minimization.

Before performing the docking calculations, a cavity needs to be defined [**Figure 12**]. In rDock this can be done with the "two sphere" method (two different-sized spheres are defined and the cavity is only accessible by the small spheres but not the large spheres) and the reference ligand

method (it creates a docking volume of a given size around a binding mode of a known ligand).

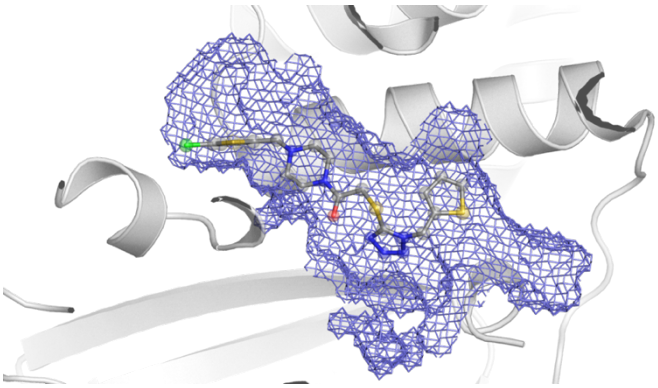


Figure 12 Example of a cavity generated with rDock

The scoring function of rDock falls in the category of empirical SF. As seen in **Equation 3.2**, rDock master socring function (S^{total}) is a weighted sum of intermolecular (S^{inter}), ligand intramolecular (S^{intra}), site intramolecular (S^{site}), and external restraints if provided ($S^{restraint}$). The weights have been extrapolated from known experimental data using a set of protein-ligand and RNA-ligand complexes.

$$S^{total} = S^{inter} + S^{intra} + S^{site} + S^{restraint}$$

Equation 3.2

$$S^{inter} = W_{vwd}^{inter} S_{vwd}^{inter} + W_{polar}^{inter} S_{polar}^{inter} + W_{repul}^{inter} S_{repul}^{inter} + W_{atom}^{inter} S_{atom}^{inter} + W_{solv} S_{solv} + W_{rot} N_{rot} + W_{const}$$

Equation 3.3

$$S^{intra} = W_{vwd}^{intra} S_{vwd}^{intra} + W_{polar}^{intra} S_{polar}^{intra} + W_{repul}^{intra} S_{repul}^{intra} + W_{dihedral}^{intra} S_{dihedral}^{intra}$$

Equation 3.4

$$S^{site} = W_{vwd}^{site} S_{vwd}^{site} + W_{polar}^{site} S_{polar}^{site} + W_{repul}^{site} S_{repul}^{site} + W_{dihedral}^{site} S_{dihedral}^{site}$$

Equation 3.5

$$S^{restraint} = W_{cavity} S_{cavity} + W_{tether} S_{tether} + W_{nmr} S_{nmr} + W_{ph4} S_{ph4}$$

Equation 3.6

Formally, solutions should be sorted based on S^{total}, but it has been show that the intramolecular term bears large error and can introduce more noise than signal to the predictions [95].

3.3 DEVELOPMENT OF AN AUTOMATIC PIPELINE FOR PARTICIPATION IN THE CELPP CHALLENGE

3.3.1 CANDIDATE PREPARATION

For each candidate structure, co-crystallized solvent and ligands are removed using Schrödinger's split structure tool [96] and only the coordinates of the receptor are kept. Subsequently, the protein preparation tool from MOE [97] is used to fix problems within the crystal structure and the Protonate 3D tool [98] is used to assign protonation states to the protein (assuming pH 7.0). All the files are saved in Tripos MOL2 format, as required by the docking program, rDock [93] . All the above steps are integrated in an SVL script for automation.

3.3.2 LIGAND PREPARATION.

We take the query ligand in SMILES string format and use the LigPrep tool from Schrödinger [99] to calculate the 3D structure with proper topology, tautomerism, bond orders, and geometry of bonds, angles, dihedrals, and rings. Also, the ionizable groups are protonated at pH 7 with a threshold of ±1 pH unit. All ligands are saved in SDF format.

3.3.3 SELECTION OF SIMILAR PROTEINS, DRUGGABLE POCKETS, AND LIGAND RETRIEVAL

One of the pillars of the whole process is being able to select good reference systems from which we can extract some restraints to guide our docking predictions. For this purpose, we have integrated into the pipeline a protocol based on the *3Decision* tool from Discngine [100].

3Decision is a web-based platform that centralizes all structural knowledge (including all the RCSB PDB dataset) to perform multiple kinds of analyses. We query 3decision using a dedicated REST API endpoint. Using as input the target sequence in FASTA format, a Blast against the database is performed to select those proteins that share a high Identity (I% > 80%). 3decision database also contains all pre-computed druggable pockets as predicted by fpocket cavity detection tool [101]. The pockets are aligned based on the sequence and superimposed to the query structure. Finally, we export all the ligands found in the aligned pockets, in a SDF file which is also converted to SMILES format using Openbabel [102]. In the case where multiple druggable pockets are detected, the corresponding docking protocol is applied to every pocket.

3.3.4 LIGAND SIMILARITY AND MAXIMUM COMMON SUBSTRUCTURE CALCULATION

After retrieving the ligands found in similar pockets, a similarity analysis is performed between the query ligand and the list of retrieved ligands using MACCS keys fingerprints and the Tanimoto coefficient scoring, which has been identified as one of the best metrics for similarity calculations [103]. The Tanimoto coefficients as well as the fingerprints were calculated using rdkit [104].

The maximum common substructure (MCSS) between the target ligand and the ligands retrieved from similar proteins was calculated using RDKit's FindMCS function [104]. As a complementary measure of similarity between the ligands and also working as a method to evaluate the robustness of the MCSS, a Tanimoto coefficient based on MCSS was calculated using **Equation 3.7** [105].

$$Tanimoto_{MCSS} = \frac{N_{AB}}{(N_A + N_B) - N_{AB}}$$

Equation 3.7

where N_A and N_B are the number of heavy atoms in molecules A and B respectively, and N_{AB} is the number of heavy atoms in the MCSS. The Tanimoto_{MCSS} can have values between 0 and 1, being 1 the value obtained when two molecules are identical.

3.3.5 GENERATION OF PHARMACOPHORIC RESTRAINTS

Ligand-based pharmacophore modelling has had a great impact in drug discovery [106].In this work this strategy is used to extract common chemical features from the aligned ligands retrieved by 3decision before elucidating the pharmacophores. The Align-it tool from Silicos-it [107] is used to generate a combination of pharmacophore points for each molecule in the set. In this work two different versions of the protocol for the generation of a consensus pharmacophore are tested. In the first version, after the generation of the pharmacophoric points for each molecule, the features that were common between molecules were selected and ranked by number of appearances and then the two highest ranked features were selected and used as mandatory pharmacophoric restraints for docking. In the second version, the ligands are first clustered based on similarity (MACCS fingerprints and Tanimoto similarity of 0,9). From each cluster the ligand corresponding to the centroid is selected, thus removing redundancy and getting a diverse set of ligands and then the pharmacophoric points are generated. From here, only the most representative points (those shared by more than 45% of the ligands) are considered as mandatory restraints. Points shared by between 20% and 44% of the ligands are considered as optional restraints. For the optional restraints, at least one of them must be fulfilled during the docking process.

3.3.6 MOLECULAR DOCKING

To define the binding site in this work we chose the reference ligand method with rDock's default parameters. From the pool of retrieved ligands, we select as reference ligand the one having the maximum sum

of MACCS Tanimoto similarity score and $Tanimoto_{MCSS}$ score. This combined score implies a similar ligand and also a similar size to the target ligand. As a result, the cavity size is adapted to the query ligand, adding another restriction level to the docking process.

rDock can perform free docking as well as different types of restraint docking. Using rDock capabilities, our pipeline can use three different docking protocols, depending on the characteristics of the system and the available information. If we find a good reference ligand ($Tanimoto_{MCSS} > 0,5$), then the pipeline will choose tethered docking, fixing the MCSS with the *sdtether* utility. Otherwise, if there is a sufficient number of diverse ligands to extract a pharmacophore (>5), a pharmacophoric restraint docking is chosen instead. Finally, unrestrained docking is used for the remaining cases. All the docking predictions use the standard rDock docking protocol (*dock.prm*).

3.3.7 POSE SELECTION

The output from the pipeline is a set of poses generated by the docking program for each candidate structure in an SDF file. Then the poses are sorted by rDock's intermolecular score (SCORE.INTER), which accounts for the protein-ligand interaction free energy. Formally, solutions should be sorted based on the total score which accounts for the intramolecular energy as well (SCORE.INTRA + SCORE.INTER), but it has been shown that the intramolecular term bears a large error and can introduce more noise than signal to the predictions [95]. Using *sdsort,* the best pose is selected and saved in an SDF file. If more than one cavity were detected, this selection protocol is then applied to each cavity. Thereafter, the cavities are ranked based on the MCSS score obtained during the Ligand similarity and MCSS calculation and then, the best poses from each cavity are ranked by rDock's SCORE.INTER. The best scoring pose from the top scoring pocket is then selected for submission. Finally, the files are transformed to the format required by CELPP submission rules: the ligand pose in MOL format and the receptor in PDB format.

3.4 TARGETING RANK RECEPTOR AS A NOVEL THERAPEUTIC STRATEGY FOR TRIPLE NEGATIVE BREAST CANCER

3.4.1 COMPUTATIONAL METHODS

3.4.1.1 HOMOLOGY MODELLING

The extracellular domain of the human RANK (UNIPROT: Q9Y6Q6; residues 30-212) was modelled using the MOE homology model tool, selecting as a template the crystalized mouse RANK protein, presenting an Identity with human RANK of 82.53% (PDB: 3ME2 (chain B) [108]).

3.4.1.2 MOLECULAR DYNAMICS

Molecular dynamics (MD) simulations were carried out to identify druggable pockets and to monitor their dynamic behavior. In all cases, simulations were carried out with the pmemd.cuda program of the Amber package and preparation with the tleap program of the AmberTools package. The protein was assigned ff14SB [109] atom types, the TIP3P [110] water model and periodic boundary conditions, with a truncated octahedron cubic box extending at least 14 Å further from the protein in each direction and dimension. The solvated system is neutralized with Na+ or Cl- ions, as needed. Equilibration consisted of a heating stage of 800 ps to reach 300 K in the NPT ensemble and a 1 ns stage in the NVT ensemble at 300 K. SHAKE [111] was applied to all bonds involving hydrogen using a 2 fs timestep. Electrostatic interactions were calculated by the particle-mesh Ewald (PME) method using constant pressure and temperature conditions. The temperature was kept constant at 300 K using a Berendsen thermostat with a 0,1 picosecond (ps) coupling constant, and the pressure at 1.0 bar using the

Berendsen barostat with a 0.5 ps time coupling constant. Van der Waals and short-range Coulomb interactions were truncated at 9Å.

3.4.1.3 DRUGGABILITY PREDICTION

Protein homology model for human RANK was protonated and checked for accuracy using the ProteinPrepare tool from MOE [97]. Then, using pyMDmix, immersed in a solvent-filled truncated octahedral box constructed from replicas of a pre-equilibrated box of solvent mixture. The solvents used were pure water, ethanol at 20% in water (ETA) and acetamide at 20% in water (MAM). Parameters for the organic solvents have been published in [112]. All non-hydrogen atoms of the protein are restrained with soft harmonic potentials (k=0,01 kcal/molÅ^2). Three independent simulations for each protein−solvent combination are carried out to obtain a total sampling of 60 ns for each system and solvent mixture.

After the production stage, all replicas are superimposed to a reference structure (backbone atoms of the protein in the homology model coordinates). Then, a grid with 0.5Å spacing in each direction is constructed for each one of the probes of the solvent mixture and the observed density in each grid element is compared to the expected density and converted to binding free energy using the inverse Boltzman relationship. Lastly, the regions of the grid with the most negative ΔG_{bind} for each probe are selected as a hot spot.

3.4.1.4 ANALYSIS OF DRUGGABLE CAVITIES DURING MD

To monitorize the volume of the putative ligand binding pockets, we generated 3 independent Molecular Dynamics simulations each 200ns-long, for a total sampling of 600ns. The trajectories were then analysed using MDpocket [113], an open-source tool based on the fpocket [101] cavity detection algorithm. The MDpocket analysis was performed for each replica with 400 snapshots equally spaced in time using the default fpocket parameters (-m 3,0, -M 6,0, -I 30, -n3).

3.4.1.5 VIRTUAL SCREENING

System preparation

The structure for the selected RANK snapshot was prepared using MOE 2016 [97] by removing water and cofactors, capping the termini and gaps, and for setting the protonation state of the protein with default settings. The cavity was defined in the prepared structure by the reference ligand method, using the hotspots identified by MDmix as atomic centers.

Docking Protocol

The virtual library of compounds consisted of ~7M compounds coming from different vendors. The library was prepared with LigPrep [114], so that at most eight stereoisomers, six tautomers and eight ring conformers would be generated and lastly, probable ionization states within the pH range of six to eight would be generated. The prepared library was docked with 3 pharmacophoric restraints, 1 H-bond acceptor at a distance of 3.1Å, from N of Cys-82, 1 Hydrophobic spot at a distance of 3.1Å from Trp-88 and another hydrophobic spot at a distance of 4.6Å from Leu-111. All the points were defined with a tolerance of 0.7 Å radius. If the feature did not adhere to the positional constraints, rDock would assign a positive (unfavourable) pharmacophore restraint score, for which the cutoff was set to 1.0. Furthermore, a high-throughput VS (HTVS) protocol was implemented (HTVS protocol in supplementary information), which consisted of three stages, for which at every stage the number of docking runs increases (up to 50 runs), and the rDock "SCORE.INTER" filter becomes stricter (-16 at 5 runs to -21 at 15 runs). If at the end of the 50 runs the molecule did not achieve a SCORE.INTER lower than -22, the molecule would be discarded.

Filtering Docking Results

For each ligand the poses are sorted by rDock's SCORE and the best one selected. All the ligands are then sorted by SCORE.INTER.norm

and clustered using Reynolds clustering in MOE, setting 0.95 Tanimoto similarity threshold using MACCS key fingerprints. Finally, for each cluster the molecule with the best SCORE.INTER.norm is selected as the cluster representative.

3.4.1.6 DYNAMIC UNDOCKING

DUck was performed on the top 2.000 compounds coming from docking, pulling from the N of the backbone of Cys-82. The first step for a DUck simulation is the definition of the chunk, that represents the local environment surrounding the residue interacting with the ligand. The sequence gaps created during the process of selecting the chunk residues were capped. For this, each section of residues was split into separate chains, and the termini of each chain were acetylated or methylated. Lastly, the chunk was checked for clashes possibly created during the capping of the chains. The chunk included the following residues: 60-70, 79-98, 111-116.

After production of the chunk, up to 50 replicas of sMD/MD were performed, during which a W_{QB} threshold of 4 kcal/mol was used, so that the simulations were discontinued if the measured W_{QB} in any replica was below the threshold.

DUck protocol uses MOE [97] to automatically prepare the scripts for the simulation and to prepare the structure (AMBER force field 99SB [115]) and ligand (Parm@Frost [116]). The simulations were performed at the Barcelona Supercomputing Center using NVIDIA Tesla M2090GPUs. The average computational time was 0.5 GPU hours per molecule.

3.4.2 EXPERIMENTAL METHODS

3.4.2.1 SURFACE PLASMON RESONANCE

Surface Plasmon Resonance (SPR) is an optical technique to measure molecular interactions. SPR occur when plane-polarized light hits a metal film (usually gold) under total reflection conditions. In an SPR experiment, one molecule (**Ligand**) is immobilized on a sensor chip, then a second molecule (**Analyte**) is passed thru a constant flow. If there is binding between the ligand and analyte the change of mass in the chip surface results in changes in the refractive index which can be recorded and displayed in a sensorgram in real time. SPR experiments can be used to measure kinetic binding constants (ka, kd) and equilibrium binding constants (affinity, Ka = 1/Kd).

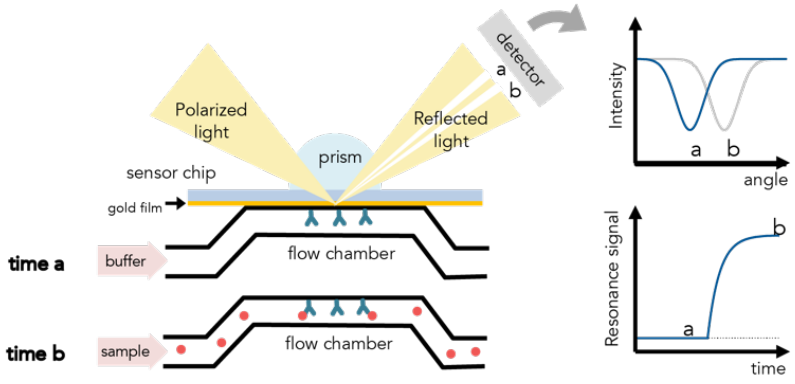


Figure 13 Surface Plasmon Resonance Assay representation. Extracted from [117]

Surface Plasmon Resonance Assay was performed using Biacore T200 SPR biosensor (Cytiva) instrument at 25ºC. A CM5 sensor chip (Cytiva) was inserted, preconditioned and normalized following the protocol proposed by the supplier. For the experiments four channels in the chip were used: two with no protein immobilized but treated to block the dextran (reference) and the other two with immobilized RANK protein (683-RK, R&D Systems). Immobilization was carried out using standard amine coupling procedure, which starts with the activation of

the carboxymethyl dextran matrix of the sensor chip with 0.1 M N-hydroxysuccinimide and 0,4 M 1-ethyl-3-(3-(dimethylamino)propyl)carbodiimide hydrochloride at a flow rate of 15 µL/min for 7 min. The immobilization was then performed at a flow rate of 5 µL/min, using a protein mixture diluted 1:100 with 10mM sodium acetate (pH 5.5). To determine the amount of protein immobilized, the following formula in **Equation 3** was used to have an expected Rmax of 50 RU (calculations for a hypothetical ligand of 500 Da):

$$RL = \frac{MW_L}{MW_a} \times \frac{R_{max}}{s}$$

Equation 3.8

RL: Response level (RU) of immobilized ligand; MWL:molecular Weight of ligand (protein); MWa: Molecular weight of analyte (compound); Rmax: Maximum binding capacity; s: Stoichiometry (number of binding sites per analyte).

Once the protein was immobilized, 1M ethanolamine hydrochloride was injected for 7 minutes at a flow rate of 15 µL/min to block activated groups of the dextran matrix. PBS (10mM phosphate, pH 7,4, 150mM NaCl) was used as an immobilization running buffer. Interaction assays were performed in a running buffer consisting of 1.05xPBS, 0.05% (v/v) tween 20, and 5% (v/v) DMSO. The 26 hits from VS were initially tested at a single dose concentration of 100uM. For that 20mM stock solutions were prepared by dissolving the ligand in 100% DMSO. For the compounds that tested positive, a bank of dilutions was prepared in 100% DMSO from which the sample concentrations were obtained by diluting them with 1.05xPBS, 0.05% (v/v) Tween-20. The concentrations considered were 500µM, 125µM, 62,5µM, 31,25µM, 15,625µM, 7,8µM. For the solvent correction, 8 dilutions that ranged from 3% to 8% DMSO in 1.05xPBS, 0,05% (v/v) Tween-20.

The Biacore T200 evaluation software 2.0 was used for data analysis. Signals were corrected for nonspecific binding to the surface by subtracting signals from a reference surface (i.e., the same

immobilization procedure without protein) from those with protein bound. Artifacts derived from DMSO interferences were corrected using a series of solvent standards (solvent correction). Background signals were corrected subtracting blank injections (blank subtraction to the injected ligand signals). To estimate binding affinity, SPR data was fitted to a single interaction model, where steady state values were extracted from the sensorgrams recorded and plotted against the different concentrations assayed.

3.5 TARGETING PTEN WITH A COMBINATION OF TARGET-BASED AND PHENOTYPIC SCREENING APPROACHES

3.5.1 COMPUTATIONAL METHODS

3.5.1.1 TARGET SELECTION

Potential targets were selected according to their TUSON-p-value. TUSON is a computational method that analyzes the likelihood that an individual gene functions as a tumor suppressor (TSG) or an oncogene (OG) based on their characteristic pattern of different types of mutation [118]. From a set of 18.682 genes, only the ones with a TSG TUSON-p-value lower than 0.005 were selected. We then selected only the ones that had a known PDB structure, discarded all the kinases and visually inspected all the structures.

Druggable Cavities Inspection

A search for putative druggable cavities was performed using fpocket [101], a pocket detection algorithm based on Voronoi tessellation. As the main measure to select the pockets we referred to the *Druggability score*, which indicates the probability of that pocket of binding a drug-like molecule, being 1 very likely to bind and 0 being likely to not bind any drug-like molecule. By using a threshold of 0.5 4 structures were selected.

Druggability Assays

Protein X-ray structures for PTPRK (2C7S [119]), PTEN (1D5R [120]), SMARCA4 (2GRC [121]), and MEN1(4GPQ [122]) were protonated and checked for accuracy using the ProteinPrepare tool from MOE [97,123], then immersed in a solvent-filled truncated octahedral box constructed from replicas of a pre-equilibrated box of solvent mixture. The solvents used were water, ethanol at 20% (ETA), acetamide 20% (MAM), isoxazole 20% (ISX). Equilibration consisted of a heating stage of 800 ps to reach 300 K in the NPT ensemble and a 1 ns stage in the NVT ensemble at 300 K. Production runs of 20 ns in the NPT ensemble are then carried out, storing atomic coordinates every picosecond. SHAKE [111] was applied to all bonds involving hydrogen using a 2 fs timestep. Electrostatic interactions were calculated by the particle-mesh Ewald (PME) method using constant pressure and temperature conditions. The temperature was kept constant at 300 K using a Berendsen thermostat with a 0,1 picosecond (ps) coupling constant, and the pressure at 1.0 bar using the Berendsen barostat with a 0,5 ps time coupling constant. Van der Waals and short-range Coulomb interactions were truncated at 9Å.

All non-hydrogen atoms of the protein are restrained with soft harmonic potentials (k=0,01 kcal/molÅ^2). Three independent simulations are carried out for each protein−solvent combination to obtain a total sampling of 60 ns for each system and solvent mixture.

After the production stage, all replicas are superimposed to a reference structure (backbone atoms of the protein in the crystallographic coordinates). Then, a grid with 0,5Å spacing in each direction is constructed for each one of the probes of the solvent mixture, and the observed density in each grid element is compared to the expected density and converted to binding free energy using the inverse Boltzmann relationship. Lastly, the regions of the grid with the most negative ΔG_{bind} for each probe are selected as a hot spot.

3.5.1.2 VIRTUAL SCREENING

System preparation

The structures were prepared using MOE 2016 [97] by removing water and cofactors, capping the termini and gaps, and for protonation with default settings. For the virtual screening, the PTEN PDB structure, 1D5R (chain A) [120] was used for docking, the cavity was defined in the prepared structure by the reference ligand method, using the hotspots identified by MDmix as reference atoms.

Docking protocol

The virtual library of compounds consisted of ~7M compounds coming from different vendors. The library was prepared with LigPrep [99] so that at most eight stereoisomers, six tautomers, and eight ring conformers would be generated and lastly, probable ionization states within the pH range of six to eight would be generated. The prepared library was docked with 3 pharmacophoric restraints, 1 H-bond acceptor at a distance of 3Å, from the hydroxyl of Tyr-164, 1 Hydrophobic spot at a distance of 4Å from the ring center of Tyr-164, and a hydrophobic spot at a distance of 4,5Å from the Cα of Arg-163. All the points were defined with a tolerance (flat-bottom restraint) of 0,7 Å radius. If the feature did not adhere to the positional constraints, rDock would assign a positive (unfavorable) pharmacophore restraint score, for which the cutoff was set to 1,0. Furthermore, a high-throughput VS (HTVS) protocol was implemented, which consisted of three stages, for which at every stage the number of docking runs increases (up to 50 runs), and the rDock "SCORE.INTER" filter becomes stricter.

Filtering Docking Results

For each ligand, the poses are sorted by rDock's SCORE and the best one selected. All the ligands are then sorted by SCORE.INTER and clustered using Reynolds clustering in MOE [97], setting 0,95 Tanimoto similarity threshold using MACCS key fingerprints [124]. Finally, for

each cluster the molecule with the best SCORE.INTER is selected as the cluster representative.

Dynamic Undocking

DUck was performed on the top 2.000 compounds coming from docking, pulling from Tyr-164, OH. The first step for a DUck simulation is the definition of the chunk, which represents the local environment surrounding the residue interacting with the ligand. The sequence gaps created during the process of selecting the chunk residues were capped. For this, each section of residues was split into separate chains, and the termini of each chain were acetylated or methylated. Lastly, the chunk was checked for clashes possibly created during the capping of the chains. The chunk included the following residues: 164-177, 188, 272-281, 318-320, 324, 345.

After production of the chunk, we performed up to fifty replicas of sMD/MD, during which a W_{QB} threshold of 3 kcal/mol was used, so that the simulations were discontinued if the measured W_{QB} in any replica was below the threshold. If the runs were completed, the lowest obtained W_{QB} value was used.

DUck protocol uses MOE [97] to automatically prepare the scripts for the simulation and to prepare the structure (AMBER force field 99SB [115]) and ligand (Parm@Frost [116]). The simulations were performed at the Barcelona Supercomputing Center using NVIDIA Tesla M2090GPUs. The average computational time was 0,5 GPU hours per molecule.

Enamine Real Sampling

Enamine Real Database with 273M Compounds (as for June 2019) was filtered with rdkit [104] to find compounds that matched the following criteria: it contains a substituted tetrazole, the molecular weight is between 250-500Da, it has between 3-8 rotatable bonds and no reactive groups. As a result, from this search 5.5M compounds were clustered using Reynolds Clustering with a Tanimoto coefficient similarity of 90% using MACCS keys fingerprints. The centroid of each cluster was

chosen as the cluster representative resulting in 189.852 compounds that were prepared, docked, and undocked following the aforementioned protocol.

3.5.1.3 MOLECULAR DYNAMICS WITH SMALL MOLECULES

PTEN structure was obtained from the PDB id 1D5R [120]. Standard protein preparation protocols were followed using MOE [97]. Duplicated proteins, crystallization buffer, compounds and salts were removed. The ff14SB [109] and gaff2 [78] forcefields were used to assign atom types for the protein and the compounds (CMP1, CMP1.2 and CMP1.3) respectively. Partial charges for the compounds were derived using the RESP [125,126] protocol at the HF/6-31G(d) level of theory, as calculated with Gaussian09. Each system was solvated on a truncated octahedral TIP3P water box of 14 Å of radius and 13 Cl- anion was added to neutralize the system. Minimization and equilibration was performed with SANDER. The systems were heated in the NVT ensemble from 100 K to 298 K in three stages of 250 ps (100K-150K, 150K-250K, 250K-298K) while retaining the harmonic restraints to the compound and the protein. Subsequently, the density of the system was equilibrated to 1 bar in the NPT ensemble during 9 stages of 250ps, where the harmonic potential lowered from 5,0 kcal mol^{-1} $Å^{-2}$ to 0 kcal mol^{-1} $Å^{-2}$.

During the equilibration and subsequent production and steered molecular dynamics trajectories, temperature control was achieved using a Langevin thermostat (with a collision frequency of 3 ps^{-1}) and a Berendesen barostat was used to control the pressure when simulating in the NPT ensemble. SHAKE [111] was applied to all atoms involving hydrogen to allow for a timestep of 2 fs and all simulations were performed with the CUDA accelerated version of PMEMD.

3.5.2 EXPERIMENTAL METHODS

3.5.2.1 CELL CULTURE CONDITIONS

Human HCT116 cells were purchased from American Type Culture Collection (ATCC) (#CRL-247) and cultured in RPMI 1614 medium supplemented with 10% fetal bovine serum (FBS) in 5% CO2 at 37°C.

3.5.2.2 CELL LINE CHARACTERIZATION BY WESTERN BLOT

Protein extracts from cell samples were obtained by homogenizing the cells in a lysis buffer containing 50 mM Tris-HCl pH 7,4, 150mM NaCl, 1% NP40, 0,25%SDS, 0,5mM DTT, 1mM NaF and protease inhibitors (Complete, Mini, EDTA-free protease inhibitor cocktail, Roche). Then the cells were rocked at 4ºC for 30 min and then centrifuged at 15.000 rpm for 15 min at 4ºC. Protein was then quantified using the Pierce BCA Protein Assay Kit from Thermo Fisher following the manufacturer's instructions.

Cell lysates were separated by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred onto a nitrocellulose membrane. After probing with primary antibodies, the membranes were incubated with horseradish peroxidase-conjugated secondary antibody and visualized by ECL (Pierce, Rockford, IL). Different exposure times of the films were used to ensure that bands were not saturated. Quantification of the films was performed by densitometry using ImageJ software (Bethesda, MD, USA).

Antibodies specific for Akt, pAkt-Ser473, PTEN, S6 Ribosomal Protein (5G10), and Phospho-S6 Ribosomal Protein (Ser240/244) were obtained from Cell Signaling Technologies. Vinculin and α-Tubulin antibodies were obtained from Sigma-Aldrich.

3.5.2.3 CELL SURVIVAL ASSAY

Cell survival was assessed with the CellTiter-Glo kit that measures ATP concentration in living cells by reacting with luciferin and quantified with luminescence reading [**Figure 14**]. Cells were plated in 96-well plates at a density of 10.000 cells per well and left overnight in the incubator to let them attach to the surface. The treatment is applied at the desired concentration in 0.5% of DMSO in the media and it is refreshed after 48h. After 72h cell survival is assayed following the manufacturer's guidelines. An equal volume of Titer-Glo reagent is added to the well and mixed with the media. The plates are incubated on a rocking platform for 2 minutes to allow lysis to occur and then 10 minutes in the dark without shaking to stabilize the luminescence signal. The plate is read with Spark 10M plate reader, with the luminescence settings at room temperature with 1s of integration time. The counts per second obtained are normalized to the DMSO control.

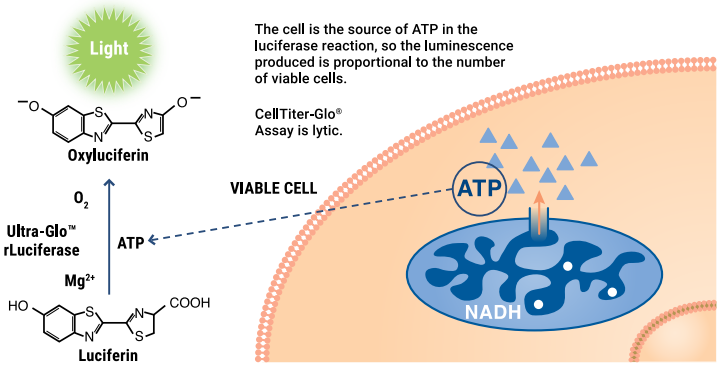


Figure 14 CellTiter-Glo assay representation. Extracted from [127]

3.5.2.4 PTEN PHOSPHATASE ASSAY

The effect of the compounds on PTEN phosphatase activity was assessed using a Malachite green-based phosphatase assay kit (Echelon Biosciences, Salt Lake City, UT, USA) following the manufacturer's instructions. PTEN enzyme and PIP3 substrate were purchased from Echelon Biosciences (Salt Lake City, UT). The method is based on the

quantification of the phosphate liberated from PI(3,4,5)P3, which forms a colored complex with molybdate-malachite green and quantified by reading the absorbance at 620 nm. Briefly, 50 ng of PTEN enzyme was diluted in TBS buffer with 10 mM DTT, then the compound dissolved in DMSO, and the substrate were added to initiate the reaction. After 45 minutes of incubation at 37C, 100 uL of malachite green solution was added and incubated for 20 minutes to allow the complex to form and the color to develop. The activity was measured as the percentage conversion of PIP3 determined by using the formula according to the manufacturer's instruction, relative to the 3.000 picomoles of the substrate at the beginning of the reaction. Control wells containing PIP3 only were set at 0% PTEN activity. The IC50 value was calculated from the dose-response curves generated using GraphPad Prism 5 software with a three-parameter fitting.

3.5.2.5 PAKT ELISA

Phospho-Akt1 levels were measured from cells treated with the different compounds at a dose that warranted more than 50% survival in 0,1% FBS RPMI 116 media. Then the cells were collected and washed with PBS for a final lysis step with RIPA buffer (10 mM Tris (pH 7,4), 100 mM NaCl, 1 mM EDTA, 1% TritonTM X-100, 0,1% SDS, 0,5% deoxycholate, and Protease and phosphatase inhibitor cocktail from Thermo scientific). The lysate was incubated on ice with occasional vortexing and debris was pelleted centrifuging at 13.000 rpm for 10 minutes at 4°C. Finally, total protein from lysates was quantified with the Pierce BCA protein assay kit from Thermo Fisher following the manufacturer's instructions. For phosphorylation measurements, AKT1 [pS473] Ultrasensitive ELISA Kit from Invitrogen was used according to the manufacturer's protocols. Briefly, 20 ug of total protein was diluted with standard diluent buffer to a final volume of 50 uL and incubated with 50 uL of primary antibody solution for 3 hours. After thoroughly washing, 100 uL of secondary antibody solution was added and incubated for 30 minutes with a second washing step afterward. Finally, chromogen solution was incubated for 30 min and the reaction

was quenched with the stop solution. Absorbance was read at 450 nm and the results were fitted to a four-parameter curve for quantification.

3.5.2.6 SURFACE PLASMON RESONANCE

Surface Plasmon Resonance Assay was performed using Biacore T200 SPR biosensor (Cytiva) instrument at 25ºC. A CM5 sensor chip (Cytiva) was inserted, preconditioned and normalized following the protocol proposed by the supplier. For the experiments four channels in the chip were used: two with no protein immobilized but treated to block the dextran (reference) and the other two with immobilized PTEN-GST (Tebu bio, 117E-3000-10UG) protein. Immobilization was carried out using standard amine coupling procedure, which starts with the activation of the carboxymethyl dextran matrix of the sensor chip with 0.1 M N-hydroxysuccinimide and 0.4 M 1-ethyl-3-(3-(dimethylamino)propyl)carbodiimide hydrochloride at a flow rate of 15 μL/min for 7 min. The immobilization was then performed at a flow rate of 5 μL/min, using a protein mixture diluted 1:100 with 10mM sodium acetate (pH 5,5). To determine the amount of protein immobilized, the following formula shown in Equation 3 was used to have an expected Rmax of 100 RU. On the High-density channel 18.737,7 RU were immobilized and on the low-density channel 6.108,2 RU.

Once the protein was immobilized, 1M ethanolamine hydrochloride was injected for 7 minutes at a flow rate of 15 μL/min to block activated groups of the dextran matrix. PBS (10mM phosphate, pH 7,4, 150mM NaCl) was used as an immobilization running buffer. Interaction assays were performed in a running buffer consisting of 1,0xPBS, 0,05% (v/v) tween 20, and 5% (v/v) DMSO using a flow rate of 60 μL/min with a contact time of 60s and dissociation time of 120s. Compounds were tested from 200 μM to 3,125 μM with 1:2 serial dilutions.

The Biacore T200 evaluation software 2,0 was used for data analysis. Signals were corrected for nonspecific binding to the surface by subtracting signals from a reference surface (i.e., the same

immobilization procedure without protein) from those with protein bound. Artifacts derived from DMSO interferences were corrected using a series of solvent standards (solvent correction). Background signals were corrected by subtracting blank injections (blank subtraction to the injected ligand signals). To estimate binding affinity, SPR data was fitted to a single interaction model, where steady state values were extracted from the sensorgrams recorded and plotted against the different concentrations assayed.

3.6 BOTTOM-UP EXPLORATION OF THE CHEMICAL SPACE

3.6.1.1 SEARCHING FOR FRAGMENTS IN ENAMINE REAL DATABASE AND ZINC20

To generate the fragment library, we used two different compound collections, Enamine Real database and ZINC20 (up to 350MW). Both collections were filtered using rdkit [104], selecting compounds with 14 or fewer heavy atoms and compounds containing at least one ring. After removing duplicates between both databases, we were left with a collection of 4.123.967 unique fragments in SMILES format.

The fragment library was prepared with Corina and ChemAxon. The library was protonated at pH 7 and tautomers were generated using ChemAxon[128]. Then Corina (version 4.4.0) [129] was used to generate stereoisomers (up to 4), ring conformations (up to 5, with a maximum strain energy of 8 kcal/mol), and to add implicit hydrogens. The final library was saved in 3D SDF. After ligand preparation, we obtained a total of 11.952.000 entries (molecular states for docking).

3.6.1.2 PROTEIN STRUCTURE SELECTION AND PREPARATION

The PDB structure of BRD4 (PDB code 4LR6 [130])was prepared with MOE [97] and set the protonation states at pH 7,0. From an internal study of conserved waters carried out overlapping the PDB structure of

BRD4, we decided to maintain 7 water molecules in the cavity (HOH 302, 305, 311, 322, 327, 331, 332). The structure was then saved in the standard Tripos MOL2 format.

3.6.1.3 DOCKING THE FRAGMENT LIBRARY

For docking of the fragment library, we used the rDock software. The cavity was defined with rbcavity using the "reference ligand method" with the cocrystallized ligand as the reference with a 6 Å radius. The prepared library was docked with 2 pharmacophoric restraints: 1 H-bond acceptor at a distance of 2Å from Nδ of Asn-140, and 1 Hydrophobic spot at a distance of 2,5Å of the crystallographic water network [131]. The acceptor point had a tolerance (flat-bottom restraint radius) of 0,5 Å and the Hydrophobic point of 1 Å. If the feature did not adhere to the positional constraints, rDock would assign a positive (unfavorable) pharmacophore restraint score, for which the cutoff was set to 1.0. Furthermore, a high-throughput VS (HTVS) protocol was implemented, which consisted of three stages, for which at every stage the number of docking runs increases (up to 15 runs), and the rDock "SCORE.INTER" filter becomes stricter.

3.6.1.4 FILTERING OF DOCKING RESULTS BASED ON PROPERTIES OF DESCRIBED ACTIVE FRAGMENTS

A set of already known fragments from CHEMBL were selected as a test set to determine the conditions for further filtering the docking results. The set consisted of 35 fragments with 14 or fewer heavy atoms and containing at least one ring. The fragments had to also comply with the interaction with Nδ of Asn-140. The test set was docked following the same protocol as the fragment library and a SCORE.INTER of -12 kcal/mol was deemed to be the most adequate to use as a threshold for filtering the docking results. Additionally, the docking solutions that did not place a hydrophobic group near the water network were discarded.

3.6.1.5 FRAGMENT CLUSTERING USING CHEMICAL CHECKER SIGNATURES

362.345 fragments were characterized using the A1-A5 Chemical Checker signaturizers [132,133] which include information related to 2D and 3D topological fingerprints, scaffolds, structural keys and broad physicochemical properties. These fragments were then clustered using the Sklearn Kmeans algorithm [134] with a fixed number of 2.000 clusters and the remaining function parameters were left by default.

For each cluster, the fragment compound closest to the cluster centroid was selected as the representative one, finally accounting for 2.000 fragments out of 362.345. The clustering process was evaluated in three different ways:

1. Comparison between distance distributions of same-cluster and different-cluster fragments and subsequent computation of the ROC Curve.

2. Comparison between the average standard deviation for each feature of same-cluster and different-cluster fragments.
3. Comparison between several molecular properties (molecular weight, logP, QED, number of donors, number of acceptors, polar surface area, number of rotatable bonds and number of aromatic groups) of the representative fragments and all the fragments.

3.6.1.6 MMGBSA OF CLUSTER REPRESENTATIVES

To further filter the docking results, we used Schrödinger's Prime MM-GBSA tool to calculate the Binding free energies of the 2.000 cluster representatives. To select a proper ΔG_{bind} threshold value, we used the same test set of already known fragments. From the test set a value of -30 kcal/mol was set as threshold and applied to the fragment library (973).

3.6.1.7 DYNAMIC UNDOCKING

DUck was performed on the 973 fragments, pulling from the Nδ of Asn140. For BRD4, the chunk was prepared manually by selecting residues within 6 Å from Asn-140, Trp81-Ala89, Lys91-Leu94, Tyr97, Ile101, Pro104- Met105, Thr131, Asn135-Tyr137, Tyr139-Asn140, Asp144-Ile146, and Met149, and water molecules 302, 305, 311, 322, 327, 331, and 332 (numbering according to PDB structure 4LR6 [131]

5 replicas of sMD/MD were performed, during which a W_{QB} threshold of 7 kcal/mol was used, so that the simulations were discontinued if the measured WQB in any replica was below the threshold.

DUck protocol uses MOE [97] to automatically prepare the scripts for the simulation and to prepare the structure (AMBER force field 99SB [115]) and ligand (Parm@Frost [116]). The simulations were performed at the Barcelona Supercomputing Center using NVIDIA Tesla M2090GPUs.

4 Results

4.1 RESULTS FOR DEVELOPING AN AUTOMATIC PIPELINE FOR THE CELPP CHALLENGE

4.1.1 BACKGROUND ON THE CELPP CHALLENGE

Computational approaches have proven to be a valuable addition to wet-lab techniques in the field of drug discovery [135]. Amongst them, we can find Structure-Based Drug Design (SBDD) methods, where the three-dimensional structure of biomolecules is used to identify small molecules that can interact with them. Predicting how a ligand binds to a target is an essential step for SBDD, and molecular docking has become a standard tool for drug discovery [89,136]. The outcome of docking is a set of proposed positions and conformations of the ligand in the binding site (poses), each with an associated score. These models can be used to interpret and guide ligand design well before the structure of the protein–ligand complex can be experimentally determined.

Nonetheless, docking programs do not always find accurate ligand poses when compared to the experimental solution. There are still challenges that need to be addressed such as receptor flexibility, proper accounting of solvation effects or better scoring functions [89]. Owing to the potential and relevance of docking for SBDD, there has been a substantial and sustained effort to improve the technique, and many docking tools have been developed, such as GLIDE [137], rDock [93], GOLD [138] and AutoDock [139]. Because different docking programs use different sampling strategies and scoring functions, it is important to be able to evaluate and compare the performance between them. To that aim, test sets are available to evaluate the performance of docking and scoring methods in binding mode, binding affinity or virtual screening tasks. Regarding the former application, multiple assessments have been performed with different evaluation benchmarks [140–145]. One of the most recent and complete studies was conducted by Wang et al. (2016)[[144], who evaluated ten different docking programs, including five commercial programs and five academic programs using a collection of 2.002 protein–ligand complexes from the PDB.

Concurrently, a strong emphasis has been put on generating highly refined test sets, which only include high-quality structures of relevant protein targets containing drug-like ligands. Some of the most-used validation datasets are CCDC/Astex [146] and Iridium [147]. Such datasets and comparative studies provide a comprehensive understanding of the advantages and limitations of each docking program and help users make more appropriate choices among available methods. However, they suffer from an important limitation: in an attempt to keep the comparison across docking programs fair, the authors of the comparative studies use standard parameters, whereas, in real-life applications, advanced users introduce substantial bias to improve performance. In consequence, such comparative studies reveal the intrinsic capabilities of the programs, which is quite different from how they are actually used in typical drug-discovery settings. In addition, as relatively small sets of well-curated protein–ligand complexes become widely adopted as test-sets, there is a risk of biasing docking programs towards those specific datasets.

The challenges organised by the Drug Design Resource (D3R) represent a welcome departure from this tendency. D3R aims to provide benchmark datasets and blinded challenges to assist in the evaluation and improvement of computational algorithms, giving participants the freedom to use the methods as they see fit, but encouraging the use of reproducible protocols. Besides the annual Grand Challenge, D3R also organises the CELPP Challenge (Continuous Evaluation of Ligand Pose Prediction) [148]. Participants in CELPP are encouraged to develop an automated workflow to generate binding mode predictions for different targets that are delivered weekly.

In this section, we describe the development of the first version of a pipeline for participation in the CELPP Challenge, as well as validation results. The main focus of our workflow is to adopt a knowledge-based approach whenever possible, trying to extract data from similar systems that are already deposited in the PDB. Depending on the amount of information available, the docking algorithm may benefit from

knowledge about the location of the binding site, specific pharmacophores or even the binding mode of specific substructures. We will describe the different options, analyse their respective performances and identify aspects that need further improvement.

4.1.2 OVERVIEW OF THE PIPELINE

One of the key aspects of this work is the automation of the process; therefore, all the steps are gathered in a combination of python, SVL and shell scripts and divided into individually functional modules corresponding to the different phases of the protocol [**Figure 15**]. There are four phases summarized here (see Method section for further details):

Phase 1: Protein analysis. Download the sequence of the query protein, and identify structures of homologous proteins in the PDB and ligands that bind to them (this is performed through a query in 3decision [100]).

Phase 2: Ligand analysis. Compute a similarity score and maximum common substructure between the query ligand and all ligands retrieved in Phase 1.

Phase 3: Pharmacophore generation. Derive, whenever possible, a pharmacophore for the ligands retrieved in Phase 1.

Phase 4: Docking. Three docking strategies are used: tethered docking (when large maximum common substructure (MCS) is shared with a reference ligand), docking with pharmacophoric restraints (if a pharmacophore could be defined in Phase 3) and docking without any restraints (in all cases).

Additionally, the process includes communication with the CELPP server to download the queries and upload the predictions.

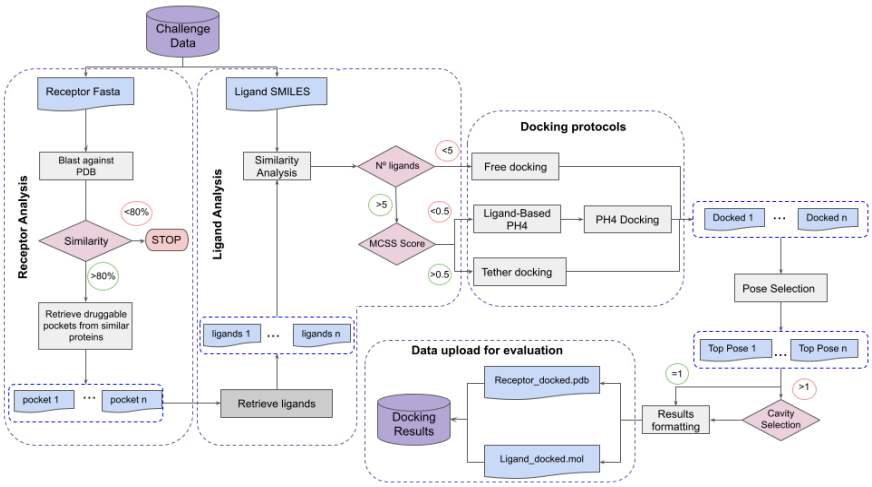


Figure 15 Workflow employed for pose prediction**.**

4.1.3 WORKFLOW INPUT DATA, DATA STRUCTURE AND OUTPUT

Each weekly CELPP data package is downloaded as a gzipped tar file that contains one directory per target. The target is a protein defined by its primary sequence. Within each directory, there is a set of structures that have the same or highly similar sequences to the target. They are provided as potential receptor structures for docking and contain the highest resolution unbound candidate protein (hiResApo), the highest resolution ligand-bound (hiResHolo), the candidate protein that contains the ligand with the largest MCSS to the target ligand (LMCSS), the candidate protein that contains the ligand with the smallest MCSS (SMCSS) and the candidate protein that contains the ligand with the highest structural similarity (based on Tanimoto score and Daylight fingerprints, as implemented by RDkit [104]) to the target ligand (hiTanimoto). Then, we find the SMILES [149], MOL file and INCHI key [150] corresponding to the target ligand. Finally, the suggested binding pocket center is also given. However, our pipeline includes a cavity detection phase, so the suggested binding pocket center will not

be used. The expected output from participants is a docked pose of the target ligand with each suggested candidate structure.

4.1.4 PIPELINE DEVELOPMENT

4.1.4.1 BLAST RESULTS

Before starting the implementation of the pipeline, we analyzed the targets from previous CELPP weeks (test set) to check how often they had high similarity homologues already deposited in the RCSB PDB. For this purpose, we ran a blast search against the RCSB PDB with two different identity thresholds: 80% and 95%. From this step, we could conclude that 100% of the targets had some close homolog structure available (>80% identity) within the RCSB PDB prior to its release. When looking for proteins with an identity higher than 95%, we obtained varying results across weeks with an average of 77,1% of positive cases [**Figure 16a**]. This mirrors the trends in the PDB, which is highly redundant in protein composition [151]. In light of the results, we set the identity threshold for blast searches in our automatic pipeline to 80%.

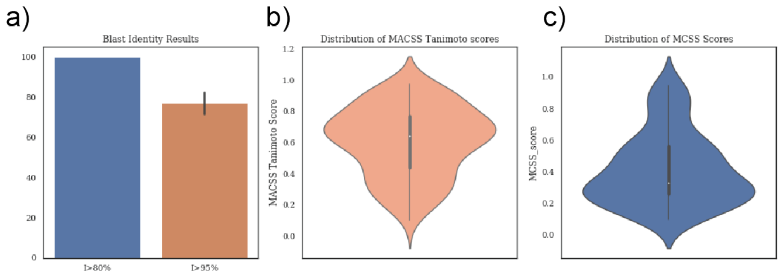


Figure 16 Analysis of the targets from previous CELPP weeks A) Histogram representing the percentage of targets for which we obtained blast results with and identity higher than 80% (blue) and 95% (marron) (B) Distribution of Tanimoto MACSS score and (C) Tanimoto MCSS scores obtained for the ligands in the test

4.1.4.2 LIGAND SIMILARITY

We analyzed the similarity between the ligands provided by CELPP and the ligands obtained by 3decision from similar proteins. After running the 3decision protocol, we were able to obtain sets of ligands for 75% of the proteins in the test set. Using MACSS keys fingerprints, we obtained a mean Tanimoto score of 0,6 with 0,008 and 0,96 being the minimum and the maximum scores obtained, respectively [**Figure 16b**]. We also took into account the size of the compared ligands and their maximum common substructure with a complementary similarity measure, the Tanimoto MCSS [105]. Its value distribution is rather different from the Tanimoto MACSS, [**Figure 16c**] with average, minimum and maximum values of 0,42, 0,1 and 0,947, respectively.

4.1.4.3 DOCKING METHOD SELECTION

Using the same target, we compared the performance of the three different docking methods (tethered, pharmacophoric restraints and free) and checked if there was any kind of correlation between the docking RMSD and the Tanimoto similarity to the reference ligands. RMSD values were calculated using the sdrmsd utility from rDock. The mean RSMD values for tethered docking, docking with pharmacophoric restraints and free docking were 2,81 Å, 2,15 Å and 2,19 Å, respectively. Thus, while the use of knowledge-based restraints improved the predictions in individual cases [**Figure 17**], the overall performance was not better [**Table 1**]. In the case of tethered docking, our analysis showed that it should only be applied when the Tanimoto MCSS is larger than 0,65, after which point almost all predictions were correct [**Figure 18a**]. Unfortunately, this applied to a small proportion of the cases (15%). Surprisingly, free docking also produced improved predictions for this set of ligands, which might be due to the similarity with the ligand of reference used to define the cavity or to the protein pre-organisation (quasi self-docking). The plot also showed that using tethered docking when the MCSS is too small leads to worse predictions

than free docking, explaining the apparently worst performance of tethered docking compared to free docking when considering the entire test set. Regarding pharmacophore-guided docking, contrary to our initial expectations, we found that there was not a significant difference in total mean RMSD between restrained and free docking (2,15 Å and 2,9 Å, respectively). This could, in part, be related to the cavity definition process, which already limits the docking space and may leave a small margin for improvement. However, it also suggested that the choice of pharmacophoric restraints was sub-optimal and had to be re-optimised. Thus, we introduced an improved pharmacophore elucidation protocol (see Methods).

Table 1 RMSD results obtained using different docking methods

	Free docking	Tethered docking	Ph4 docking
mean	2.19	2.81	2.15
median	1.96	1.71	1.63
min	0.43	0.33	0.39
max	7.41	15.07	7.41

Note:
RMSD values in Å

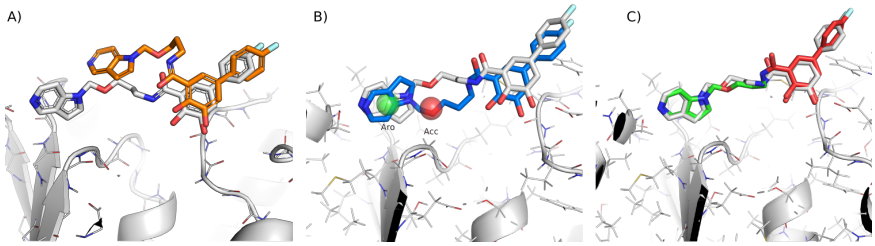


Figure 17 Differences in best pose predicted for target 5p8y from CELPP week 33. Image (A) corresponds to free docking with an RMSD of 4.09 Å. Image (B) is the best prediction obtained with pharmacophoric restraints (1.74 Å). Image (C) corresponds to the best pose.

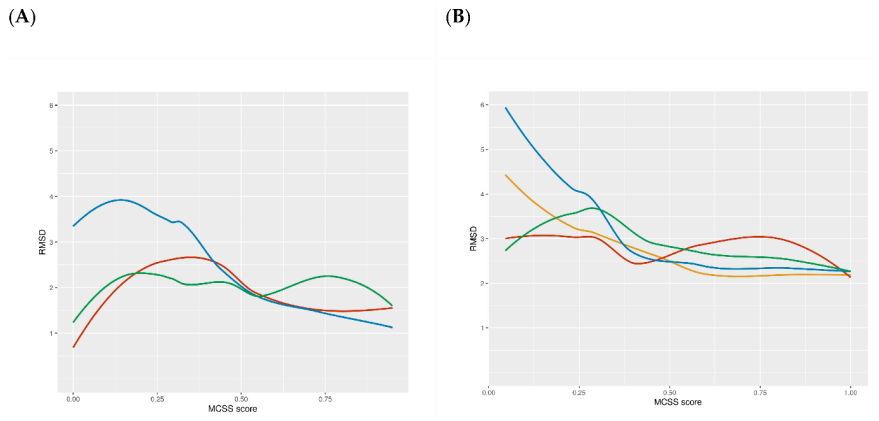


Figure 18 Relation between RMSD and the MCSS score using (A) the test set and (B) the validation set. Free docking results shown in red, docking with pharmacophoric restraints in green (version 1) and yellow (version 2; only applied to the validation set) and MCS-tehtered docking in blue.

4.1.4.4 PIPELINE EFFECTIVENESS AND PROCESSING TIME

The above-described pipeline performance was tested with a collection of pre-released CELPP weeks as well as with the weekly released CELPP set. The execution time of the whole protocol took an average 6.5 min per target. The total execution time varied each week depending on the number of released targets (26 to 68 in the period considered here) and the connection speed to 3decision (from 22s to 3min per target). The 3decision protocol could not obtain reference structures for 20% of the targets due to some internal errors of a beta version of the program or because there were no ligands found in druggable pockets from similar proteins. This last event was relatively rare, as it accounted for 25% of times that we were not able to obtain results from 3decision, or 5% of the total. Finally, the similarity analysis to the docked ligand poses took 4,8 min per target on average [**Table 2**].

Table 2 Statistics of the pipeline implementation CELPP weeks.

Weeks Stats

Week	No. of targets	3decision time	docking time	total time
Week1	31	34.00	103.00	137.00
Week2	44	103.00	174.00	277.00
Week3	27	10.00	113.00	123.00
Week4	43	118.00	176.00	294.00
Week5	29	35.00	153.00	188.00
Week6	40	182.00	265.00	447.00
Week7	68	234.00	123.00	357.00
Week8	26	102.00	111.00	213.00
Week9	28	126.00	247.00	373.00
Week10	48	158.00	382.00	540.00
Week11	50	193.00	270.00	463.00
Week12	26	137.00	716.00	853.00
mean	38	119.33	236.08	355.42

Note:
Time measured in minutes

4.1.5 PIPELINE VALIDATION

To validate the pipeline, we ran it prospectively for a total of 12 weeks. **Table 3** shows that the pharmacophoric restrained protocol was the most-used method (51% of the cases). On the other hand, free docking and tethered docking were applied in much lower percentages of cases, 35% and 13,01%, respectively. The mean RMSD value for free docking was 6,2 Å, 5,1 Å for pharmacophore-guided docking and 2,8 Å for tethered docking. However, there is a bigger difference when looking at the proportion of correctly predicted cases by each method. For free docking, only 7,9% of the cases had an RMSD value lower than 2 Å, for pharmacophore guided docking this value increased to 21,4%, and in tethered docking we reached 31,5% of correct poses.

Table 3 RMSD values and percentage of cases for each docking protocol.

	Free docking	Ph4 docking	Tethered docking
mean	6.2	5.1	2.8
Std	6.2	3.4	1.6
min	1	0.5	0.7
Q1	3.9	2.2	1.6
Q2	6.3	4.7	2.3
Q3	8.2	7.7	3.6
max	13.6	13.9	12.7
<=2Å	7.9%	21.4%	31.5%
Application rate	35%	51%	13%

Note:
RMSD values in Å

The values obtained with the validation set were much worse than the ones obtained using the test set. The main difference between the sets as that the automatic pipeline for retrieving the cavities using 3decision was not yet automatized during the development stage. In consequence, all the cavities were visually inspected and selected using the 3decision webserver. By contrast, the automatic scripts used at the validation stage to identify the docking cavity and retrieve aligned ligands from 3decision were error-prone. We also had to consider the possibility that the test set was not representative enough of the whole range of systems that can be found in the CELPP Challenge. Nonetheless, the sources of errors and the difference in performance between the test set and validation will be reviewed in the next section.

After analysing the prospective results, we wanted to review if the algorithm for docking protocol selection derived from the test set was the most adequate one. For this purpose, we applied all three protocols to all the validation set and compared the best RMSD obtained for the three methods [**Figure 18b**]. We could find some differences regarding the accuracy of the docking methods in the test set and validation sets. Tethered docking yielded better results than free docking when MCSS score ≥ 0,5 on the validation set (vs. a marginal improvement on an MCSS score ≥ 0,65 for the test set). Nonetheless, tethered docking was still the method that gave the worst results in low MCSS score values (MCSS < 0,3). As for the pharmacophore-guided docking, during the

validation phase, we improved the pharmacophoric elucidation protocol that provided consistently better results than in the test set (see Methods). It also provided improved results compared to free docking in the 0,5 to 1 MCSS score range, with a performance on par with tethered docking. In the 0,25 to 0,5 MCSS score range, pharmacophore-guided docking and free docking performed at a similar level. At lower MCSS score values, free docking outperformed pharmacophore-guided docking.

4.1.6 CHALLENGES TO ADDRESS

In this section we will describe the most important factors affecting the predictive performance of our pipeline. **Figure 19** depicts the main issues and challenges to overcome in the CELPP challenge, which will be treated in more detail in the following sections.

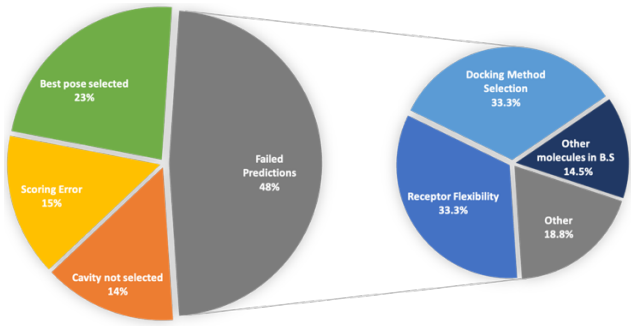


Figure 19 Overall view of validation set cases.

4.1.6.1 AUTOMATED PROTOCOLS

When testing a docking program or workflow, a crucial component that will have a big impact in the predictions is the choice of dataset [145]. Usually, the datasets to test docking programs, such as DUD-E [152] or Astex [146], are highly curated datasets, whilst the CELPP receptors are selected automatically and are not manually prepared by experts. Additionally, we have to take into account that CELPP is designed as a cross-docking challenge, which means that we have the added problem

of protein flexibility, as the used receptor may not be in the most-fitting position for the ligand. Finally, participants are given, each week, an average of 40 systems to predict and a limited amount of time (3 days), which implies that all the processes need to be automatized, leaving virtually no time for the visual inspection or study of the targets.

In consequence, the pose prediction performance is lower than for other challenges. The median prediction RMSD for the best categories (LMCSS and hiTanimoto receptors) is around 5 Å, being only 20% of the pose predictions accurate within 2Å [148], whereas reported performance for curated datasets regularly reaches the 80% [145]. Clearly, the latter reflects a best-case scenario, which means that a significant effort to improve automated target structure selection and preparation will be necessary in order to attain better results in CELPP.

4.1.6.2 SCORING CHALLENGES

Over the past years extensive efforts have been dedicated to improving the existing scoring functions, but nowadays the accuracy of most scoring functions is still a limiting factor in many drug design projects, and results require careful evaluation and post-docking analysis.

To assess the accuracy of the docking score, we selected a subset of 446 submitted cases and checked if the submitted pose is the one with the lowest RMSD compared to the crystal structure. In 208 out of 446 total cases (46,6%) the docking protocol was able to produce a correct pose (RMSD lower than 2 Å), but in 75 of them, the pose with the lowest RMSD was not ranked as the best solution by rDock's intermolecular score (SCORE.INTER). This translates to a 64% success rate when the correct pose can be generated. Note that this is close to the 76% success rate obtained on the CCDC-Astex Diverse Set, a standard test set for binding mode prediction where correct predictions can be generated for 99% of cases [93].

Figure 20 shows the median RMSD obtained with the different receptors for the submitted pose and for the best pose generated by the

pipeline. The median RMSD for the submitted pose was around 4,18 Å, whereas if we considered the best prediction, the mean decreased to 2,9 Å and the median to 2,4 Å. From these results, it is evident that the pipeline would benefit greatly from a complementary method to re-score the docking poses. An approach that presented better results in other blind challenges [24][153] was the combination of the docking scores with Dynamic Undocking (DUck) [86,87] simulations of the top-scoring poses. By combining both methods, we expected to be able to obtain a more accurate pose ranking for challenge submission.

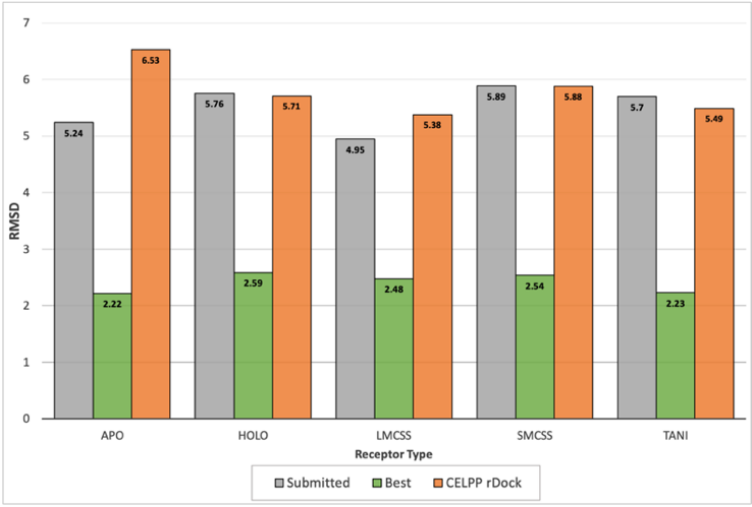


Figure 20 Median RMSD for the submitted pose compared to the best pose generated by the pipeline. CELPP rDock workflow values obtained from the D3R website [154].

4.1.6.3 SAMPLING CHALLENGES

Cavity Selection

The CELPP Challenge is designed as a pose prediction challenge and to assess the influence of receptor choice in docking performance. For that reason, the coordinates for the centre of the cavity are provided by the organisers. Nonetheless, we wanted to go one step further by creating a pipeline of general applicability and add a cavity selection step to our protocol, thus avoiding the need to pre-define the binding site.

The cavity detection is performed automatically by 3decision, and all the possible cavities are retrieved and considered for docking. The method that 3decision uses for cavity detection is fpocket, a pocket detection algorithm based on Voronoi tessellation [101]. When more than one cavity is detected, our pipeline selects the cavity based on the similarity of the ligands retrieved by 3decision with the target ligand. On average, 3,2 cavities were detected per target, but in 67 cases (14%), the correct cavity was not detected, and so the docking was carried out in the wrong cavity. **Figure 21** shows an example where 3decision only detected the cavity represented by the grey surface, missing the actual cavity represented by the green surface. In 9% of cases, the failure corresponded to shallow cavities on the protein surface that are not detected by the fpocket algorithm.

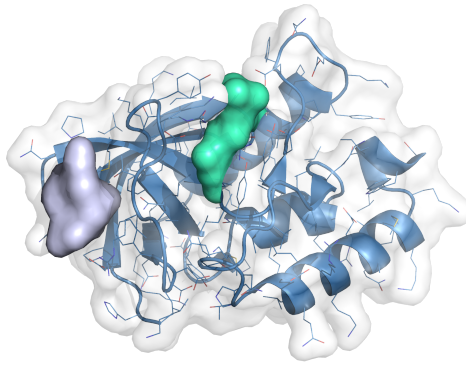


Figure 21 PDB 6ok9 [155]with the pocket detected by 3decision represented by the purple surface and the correct pocket represented by the green surface.

Another reason for not detecting the cavity correctly (14% of cases) is that the ligands bind at the interface of a dimer, but only one protein is reported in the challenge. Note that, unlike other docking challenges or scenarios, the receptors provided by CELPP are not manually curated. They rely on a fully Automated Pipeline to perform that task, which can sometimes lead to the selection of inappropriate structures (e.g., giving a monomer instead of a dimer) for obtaining an accurate ligand pose [148]. **Figure 22a** shows one such example. The remaining failures in this category were attributed to an error with the API when downloading the analysis results.

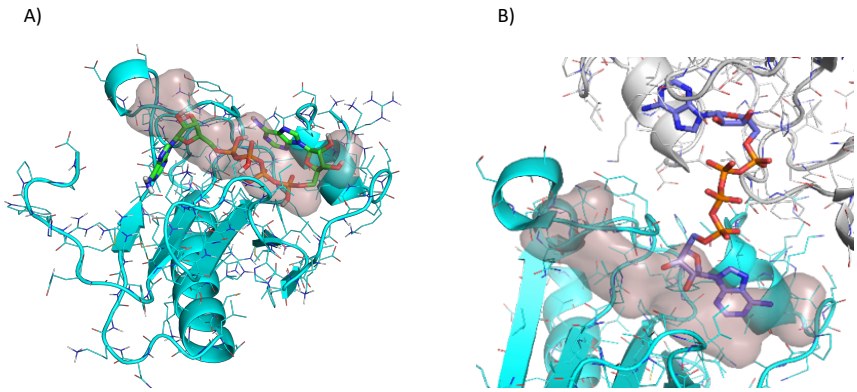


Figure 22 Example of ligand binding in a PPI interface (A) hiTanimoto receptor for target 6j65 [156]. The solution selected by the pipeline is represented as sticks. (B) Protein dimer in PDB code 6j65. The crystalized ligand is represented as sticks. For both figures, the reference cavity provided by 3decision is shown as a transparent surface.

Docking Method Selection

In our protocol we implemented three different docking strategies that were applied depending on the different set thresholds. From the 305 cases of the validation set where we did not obtain the correct pose, in 78 cases the correct binding pose had been correctly predicted by a different docking strategy.

As shown in **Table 4**, from those 78 cases, only in 9 cases the correct solution was found by free docking instead of a form of guided docking. By contrast, 26 cases could have been correctly predicted if a form of guided docking had been used instead of free docking. This analysis also reveals that the two forms of guided docking employed here are not equivalent: 27 incorrect pharmacophore-guided docking solutions were correctly predicted by tethered docking. Vice versa, 16 incorrect tethered docking solutions were correctly predicted by pharmacophore-guided docking. One such example is shown in **Figure 23**. These results suggest that all the binding poses generated by the different docking protocols should be considered, then rescored with a post-docking method to identify the best one [157].

Table 4 Comparison between the submitted docking method vs. the method that yields the best result.

		Best Prediction		
		FREE	PH4	TETHERED
Submitted	FREE		6	20
	PH4	8		27
	TETHERED	1	16	

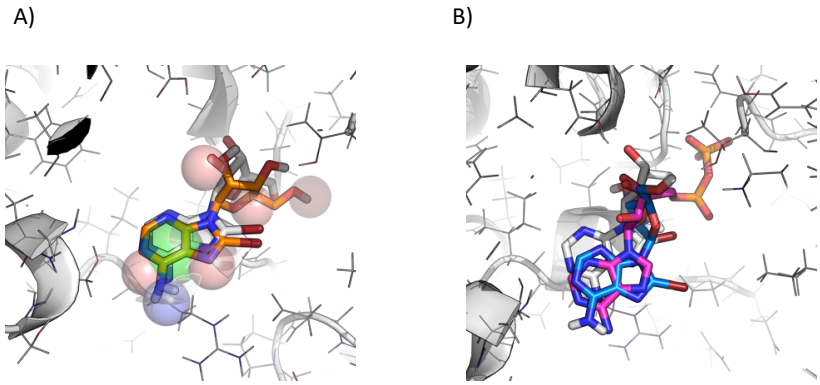


Figure 23 Predictions for PDB 6dfo and hiTanimoto Receptor using: (A) Pharmacophoric restraints. Predicted pose in orange. Pharmacophore represented as spheres. (B) Tether docking. Predicted pose in blue. Reference ligand in pink. In both cases, the crystallographic solution is shown in white for reference. The RMSD values with the predicted poses are 1,2 Å and 3,3 Å, respectively.

Receptor Flexibility

As pointed out by many previous studies [158], receptor flexibility is an important factor that can alter docking predictions. Both small changes on side-chain orientation and bigger structural changes can lead to incorrect predictions [159]. We could attest to this phenomenon when docking against the different proposed receptors. For each target, the docking protocol was run using all the receptors provided by the organisers. **Figure 20** displays the validation results categorised by the receptor. The best-performing receptor was LMCSS, which corresponds to the one hosting the ligand most similar to the query. SMCSS obtained the worst results, with a median RMSD of 5,9 Å.

As an example, **Figure 24** shows two cases where the differences in side-chain orientation of residues from the binding site are interfering with the correct binding position. In the case of 6pl1 (**Figure 24a** and **Figure 24b**), there is a difference in the conformation of a loop in the binding site of all the receptors used that cause Phe-669 (in blue) to block part of the binding site obtaining a totally different cavity. It is established that, by using a variety of receptor conformations, we increased the probability of generating a correct ligand pose, but selecting the optimal docking cavity remains a major challenge for docking methods [160,161]. This result also highlights the need to select multiple binding mode predictions, which should be re-scored with a more rigorous computational methodology.

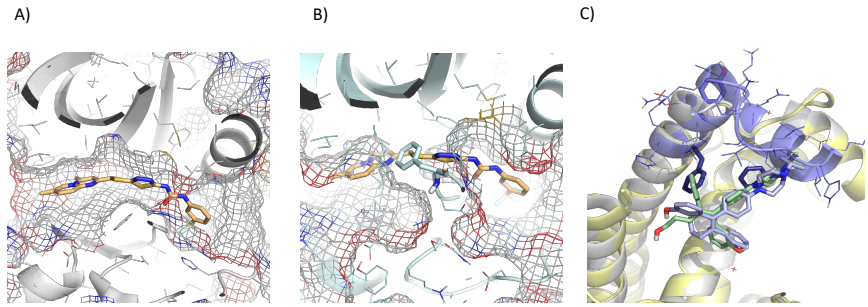


Figure 24 Example of side chain orientation in different PDB structures. (A) Differences in binding site structure organisation between 6pl1 [162] crystal and the selected hiTanimoto receptor by CELPP; the correct ligand pose is represented in beige, (B) Differences in site conformations for target 6a6k [163]between receptor hiResHolo in purple, the crystal structure in white and hiTanimoto receptor in yellow. The ligand crystal pose is represented in green and in light purple is the pose obtained using the hiResHolo receptor.

Other Molecules in the Binding Site

This pipeline was intended for general applicability, and for this reason, during the cavity preparation process all the ligands and co-solvents were removed, and only the coordinates of the receptor were kept. However, in some systems, especially enzymes, cofactors can have an important role in determining the ligand binding mode. Two such examples are provided in **Figure 25**. Lastly, the fact that there can be

other molecules in the binding site can interfere when generating the pharmacophoric restraints. As they are in the same cavity, our protocol included them in the list of retrieved ligands from similar proteins, and those are considered in the pharmacophoric restraint generation pipeline.

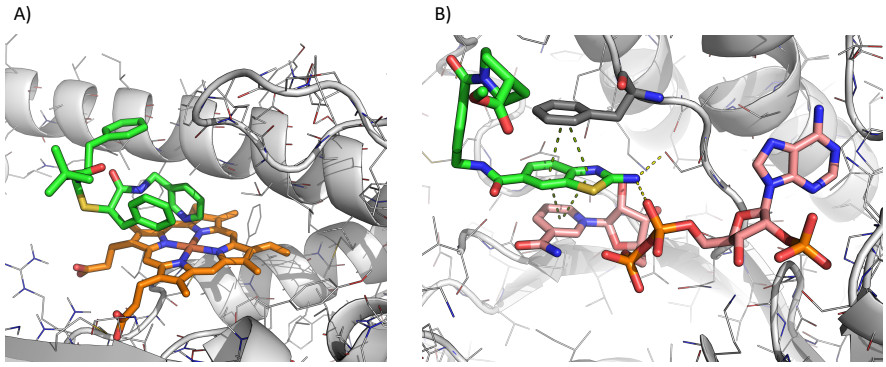


Figure 25 Example of a system with other molecules on the binding site. a) Interaction of ligand G0D (green) with heme group (orange) in PDB 6DA2 [164].Ligand belongs to a series of analogues with pyridine as a heme-ligating head that works as an inhibitor of CYP3A4 by decreasing the heme reduction rates [164]. b) Interaction of ligand EV8 (green) and NADP (pink) in PDB 6GD0 [165]. In yellow dashed lines are H-bond interactions and in green dashed lines π-interactions.

4.2 RESULTS FOR TARGETING RANK RECEPTOR AS A NOVEL THERAPEUTIC STRATEGY FOR TRIPLE-NEGATIVE BREAST CANCER

4.2.1 BACKGROUND ON RANK RECEPTOR. IMPLICATIONS ON TRIPLE-NEGATIVE BREAST CANCER.

Breast cancer is the most commonly diagnosed cancer among women worldwide. In 2020 more than two million cases were diagnosed worldwide, and although being curable in ~70–80% of patients with early-stage, non-metastatic disease, it accounted for more than half a million deaths that year. One of the main issues with breast cancer is that, on the molecular level, breast cancer is a highly heterogeneous disease, resulting in different molecular subtypes which largely influence treatment decisions. One of these subtypes is Triple-negative breast cancer (TNBC) which accounts for 15-20% of all breast cancers and is commonly diagnosed in women younger than age 40. The term TNBC refers to the fact that this specific subtype does not express human epidermal growth factor receptor type 2 (HER2), oestrogen receptor (ER), or progesterone receptor (PR) making endocrine therapies targeted to these receptors not applicable in these patients. TNBC remains a clinical challenge due to high rates of relapse, a propensity to form visceral metastases, and the lack of targeted therapies. Thus, there is a clinical unmet need to identify novel targeted therapies for the treatment of TNBC.

RANK signalling pathway, driven by the RANK receptor and its ligand RANKL, has emerged as a novel target in breast cancer. RANK is a type I transmembrane protein belonging to the Tumor Necrosis Factor Receptor Superfamily (TNFRSF). RANK extracellular domain is comprised of four tandem cysteine-rich-repeat domains (CRDs), which are characteristic of the TNFRSF proteins [108,166]. These CRDs are connected by loop regions which cause the receptor to fold into an

elongated shape and confer to the protein a high degree of flexibility [**Figure 26**]. RANKL (TNF superfamily TNFSF11), the only ligand binding to the extracellular portion of RANK, is a type II transmembrane protein belonging to the TNF family. Both the membrane-spanning and soluble forms of RANKL are assembled into functional homotrimers like other members of the TNFSF. The binding of RANKL to RANK causes trimerization of the receptor, activating the signalling pathway [167]

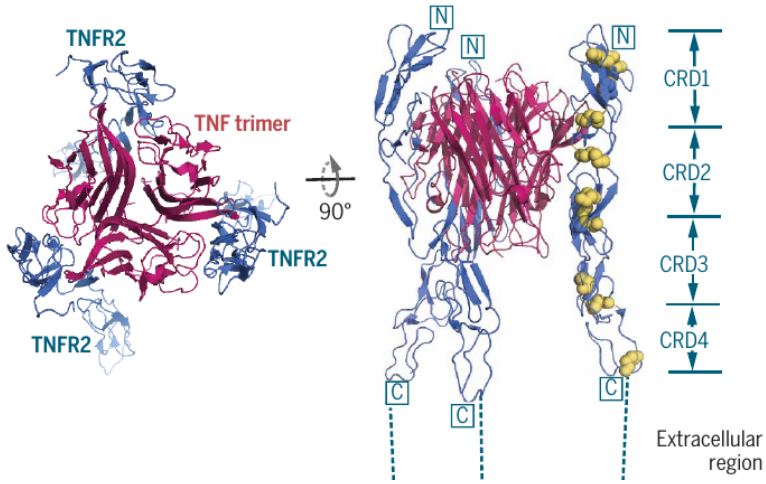


Figure 26 Representative structure of TNFSF ligand-receptor complexes. Top (left) and lateral (right) view of the structure of TNF-TNFR. A trimeric TNF ligand (magenta) is bound on the outside by three TNFR monomers (blue). In one of the TNFR monomers on the right, the four cysteine-rich domains (CRD1-CRD4) are labelled and the disulfide bonds are illustrated as yellow spheres (PDB:ID 3ALQ [168]). Extracted from [166]

In mouse models, RANK overexpression promotes mammary tumorigenesis and lung metastasis, whereas inhibition of RANK signalling in established tumours reduces breast cancer recurrence and metastasis [169,170]. Moreover, RANK overexpression in breast cancer cells leads to constitutive activation of the pathway in a RANKL-independent manner [171,172]. This RANKL-independent mechanism might explain the failure of Denosumab, an antibody that binds to RANKL and prevents its binding to RANK, to treat breast cancer patients[173].

The aim of this project is to develop small molecules that bind to the extracellular domain of RANK that, on the one hand, could inhibit the ligand-dependent signalling, acting as RANKL antagonists, and on the other hand, could abolish the constitutive activity of RANK, therefore acting as an inverse agonist.

4.2.2 DRUGGABILITY ANALYSIS OF RANK

For the identification of novel compounds with the ability of binding human RANK and disrupting the interaction with RANKL [**Figure 27a**], we first generated a homology model for the RANK human receptor. We used as a template the crystal structure of the extracellular domain of mouse RANK (PDB:3ME2 [108]) as it has a sufficiently high level of sequence identity with the human protein (~85%) to obtain a reliable model.

Starting from the homology model, MDmix was used to identify if there was any putative ligand binding site with suitable interaction hotspots that could be used as guide for virtual screening. Analysing the results for the ethanol probes, we identified a cluster of hydrophobic and polar hotspots near a CRDs domain [**Figure 27b**] (residues: 57-71 and 78-96). This region agreed partially with an already described binding site for a peptide able to bind mouse RANK and disrupt the interaction with RANKL [174]. However, the cavity identified was too small to fit any drug-like molecule. Besides, the most energetic hotspot identified was a hydrophobic hotspot derived from the interaction with TRP88 and the polar hotspots presented really week energy profile indicating that the site is not druggable because it offers scarce binding potential.

It has been shown that TNFR molecules present a great domain flexibility between the CRDs and in some cases even within each CRD [166]. Taking into consideration this high flexibility and that near the identified cavity region there is a loop comprising residues 109-117, we carried out a long MD to test if it was possible for this region to move, leaving exposed a bigger cavity.

During the 200ns MD this loop presented a high flexibility during the whole simulation. Using MDpocket [113] we detected that, due to this loop movement, a path connecting the already identified cavity and a region below the loop was created leaving exposed a bigger cavity [**Figure 27c**]. We then selected a frame that presented an open conformation of the loop and repeated the MDmix analysis with the same solvents. This time, apart from having a bigger cavity, we were able to access a region of the cavity with a really strong polar interaction point (near the N of Cys-82) that could be used with high confidence as a pharmacophoric restraint.

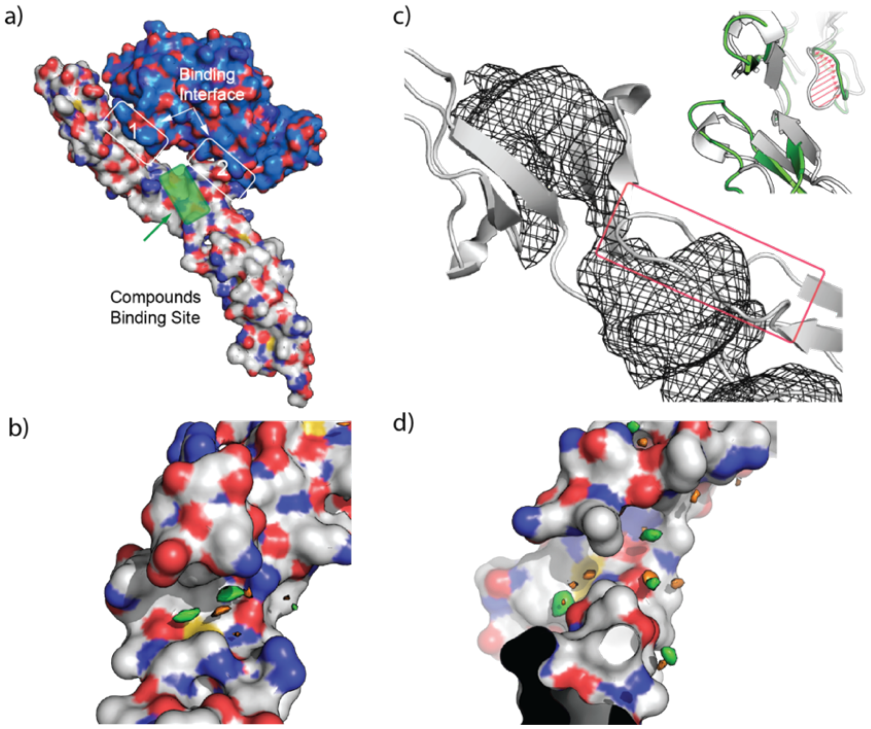


Figure 27 RANK structural analysis. a) Interaction between RANK and RANKL in pdb 3me2. b) MDmix results for the homology model of RANK. Hydrophobic binding hot spots are depicted in green, polar binding hot spots are in orange. c) The black mesh represents the transient cavity detected during MD, caused by the movement of the loop marked in red. d) MDmix results for the MD snapshot with the open conformation of the loop. Hydrophobic binding hot spots are depicted in green, polar binding hot spots are in orange.

4.2.3 IN SILICO IDENTIFICATION OF NOVEL SMALL MOLECULES BINDING TO RANK

An in-house-assembled virtual library comprising ~7M commercially available compounds was used for the virtual screening campaign against the receptor conformation that was identified during the MD. The three hotspots with the lowest binding energy identified by MDmix were used as pharmacophoric restraints to guide the docking process: an acceptor near the backbone N of Cys-82, a hydrophobic point staking with the indole ring of TRP88, and another hydrophobic spot near the side-chain of Leu-111 [**Figure 28**].

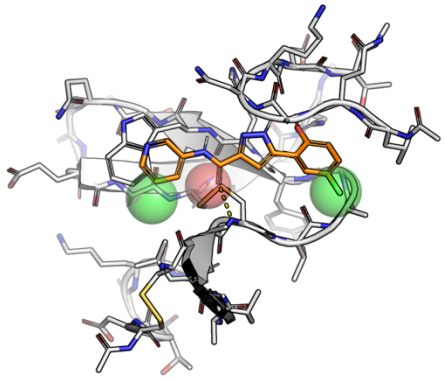


Figure 28 Pharmacophoric restraints and pocket environment. In red is depicted the acceptor pharmacophoric point and in green are depicted the hydrophobic points. A compound selected by the VS protocol is shown for illustrative purposes (orange).

After using the docking protocol detailed in the methods section (see Methods section 3.4.1.5), we obtained a total of 1.060.099 unique molecules that fulfilled the aforementioned pharmacophoric restraints. These molecules were then filtered by the rDock's SCORE.INTER, which gave us a total of 125.616 molecules with a SCORE.INTER <-25 KJ/mol. To discard very similar molecules, we performed a clustering step based on Tanimoto similarity and MACCS fingerprints [124], keeping only the best-scoring molecule as the cluster

representative. We obtained 87.653 clusters, and thus the same number of molecules corresponding to the cluster representatives. Only the top 2.000 cluster representatives (ranked by docking score) were subjected to DUck to calculate the work needed to break the key hydrogen bond with Cys-82. A W_{QB} value of 4 kcal/mol was chosen as it is a value most ligands are able to fulfill [153]. Only 186 molecules of the 2.000 (9,3%) passed this threshold. Finally, we carried out a visual inspection of the compounds. 27 compounds were selected [**Table 5**] aiming for structural diversity, however, 1 of them was no longer available for purchase and we ended up purchasing 26 compounds.

Table 5 Results of W_{QB} and SCORE.INTER for the 27 prioritized compounds.

ID	W_{QB}(kcal/mol)	SCORE.INTER (KJ/mol)
1	12,8	-30,5
2	12,0	-30,4
3	10,7	-31,1
4	8,8	-30,5
5	8,3	-28,2
6	7,8	-29,6
7	7,5	-28,3
8	7,3	-29,4
9	7,3	-30,3
10	7,2	-28,3
11	7,0	-29,9
12	7,0	-30,0
13	6,8	-30,7
14	6,7	-27,6
15	6,7	-27,9
16	6,6	-30,8
17	6,4	-31,9
18	6,4	-29,9
19	6,3	-31,2
20	6,3	-30,6
21	5,9	-30,2
22	5,8	-25,6
23	5,8	-28,3
24	5,4	-28,5
25	5,4	-31,3
26	5,0	-26,1
27	4,6	-29,2

4.2.4 SPR ASSAY CONFIRMED BINDING FOR SOME COMPUTATIONAL HITS

After the in silico virtual screening, we used SPR to evaluate experimentally the binding of our compounds to immobilized RANK. We immobilized RANK in a CM5 sensor chip with immobilization levels between 5.000 and 8.000 RU following the protocol in section 3.4.2.1. We first performed single-dose experiments at 100μM for the 26 purchased compounds from which only 10 presented a positive response and were then selected for additional SPR experiments. The K_D of the identified compounds ranged between 90μM and 7.000μM [**Table 6**]. The purpose of these experiments was to discern between true and false binders and to rank them for further activity assays performed at CNIO.

Table 6 K_D for the 10 compounds that showed binding to RANK receptor in SPR. Ch2 refers to the results obtained in the low-density Chanel. Ch4 refers to the results obtained in the high-density channel. The Chi² refers to the difference between the experimental data and the model fitted curve (a measure of the average squared residuals).

ID	K_D1 Ch2 (uM)	K_D1 Ch4 (uM)	K_D(Avg;uM)	Rmax Ch2 (RU)	Chi^2 (Ch2)	Rmax Ch4 (RU)	Chi^2 (Ch4)
1	85,6	95,9	90.8	4,54	0,94	8,36	0,28
2	105,0	149,0	127	27,09	3,36	38,25	2,07
3	45,9	335,0	191	7,53	1,33	30,86	16,90
4	219,0	168,0	194	21,82	1,97	26,02	3,35
5	147,0	109,0	128	14,80	0,36	6,36	0,78
6	318,0	320,0	319	36,11	0,55	39,20	0,67
7	241,0	306,0	274	24,32	13,10	51,85	17,40
8	327,0	368,0	347,25	48,89	4,88	57,22	3,12
9	901,0	558,0	729,00	32,43	0,01	22,49	0,28
10	1561,0	12580,0	7070,50	33,21	0,31	290,20	0,24

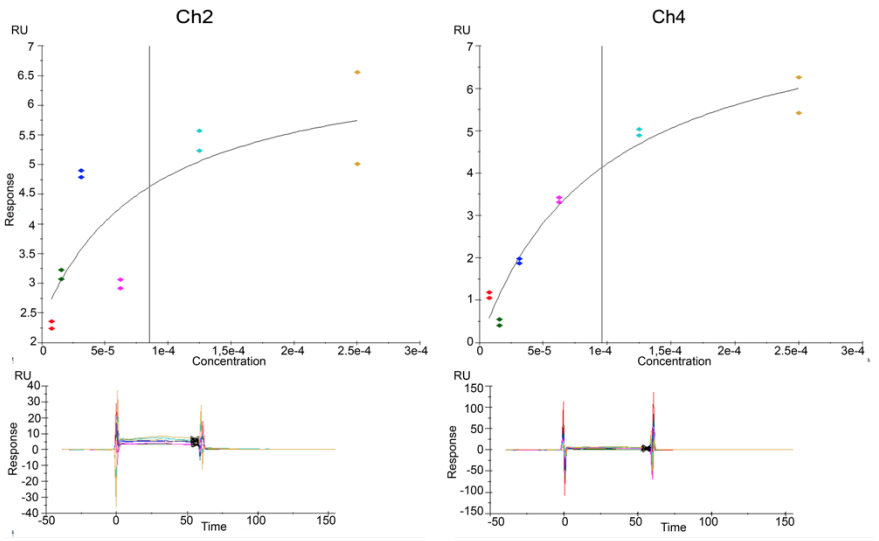


Figure 29 SPR plots for compound 1. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

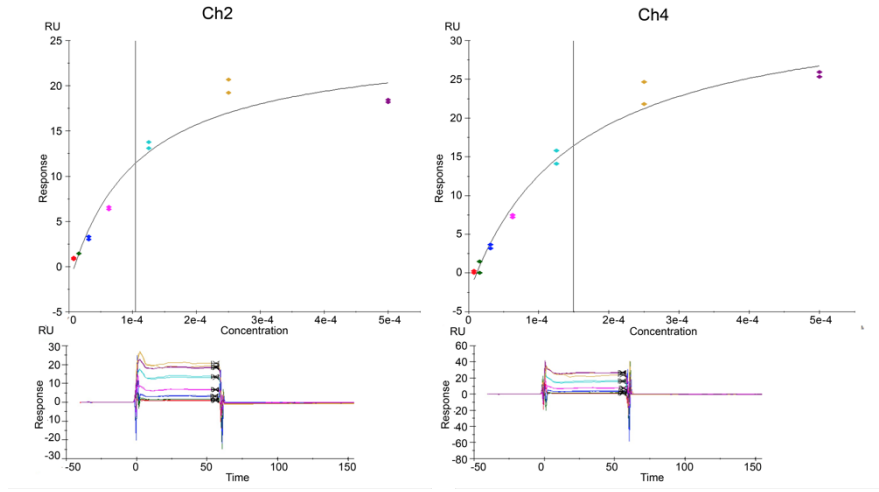


Figure 30 SPR plots for compound 2. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

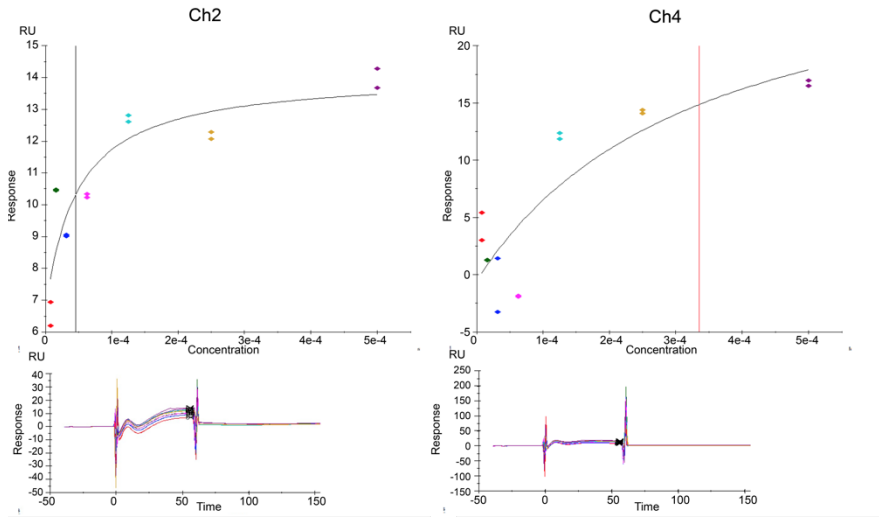


Figure 31 SPR plots for compound 3. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

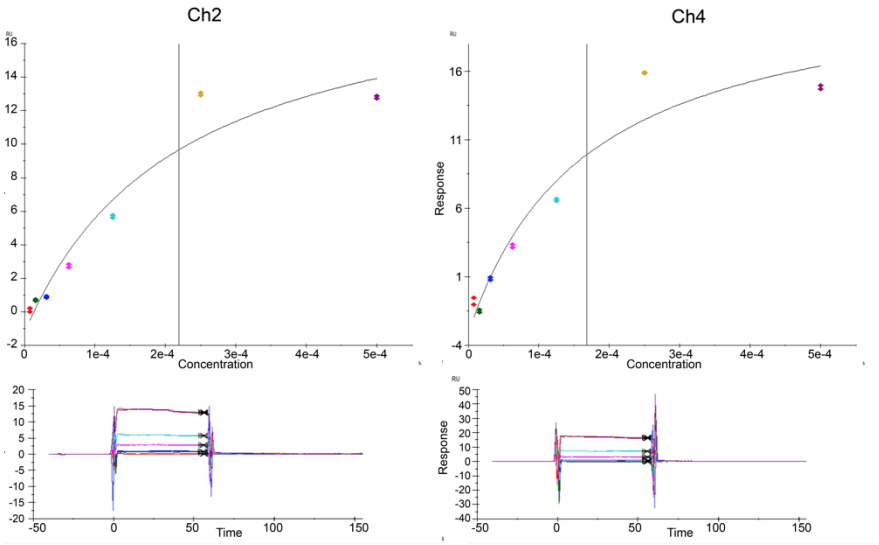


Figure 32 SPR plots for compound 4. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

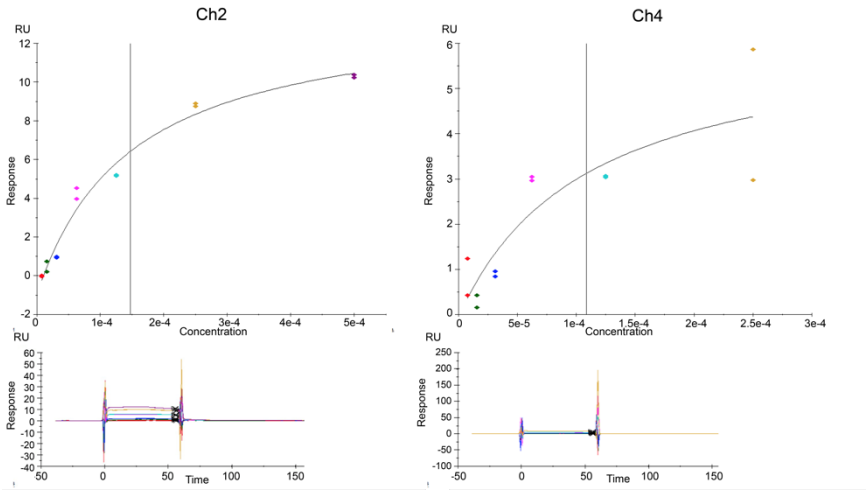


Figure 33 SPR plots for compound 5. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

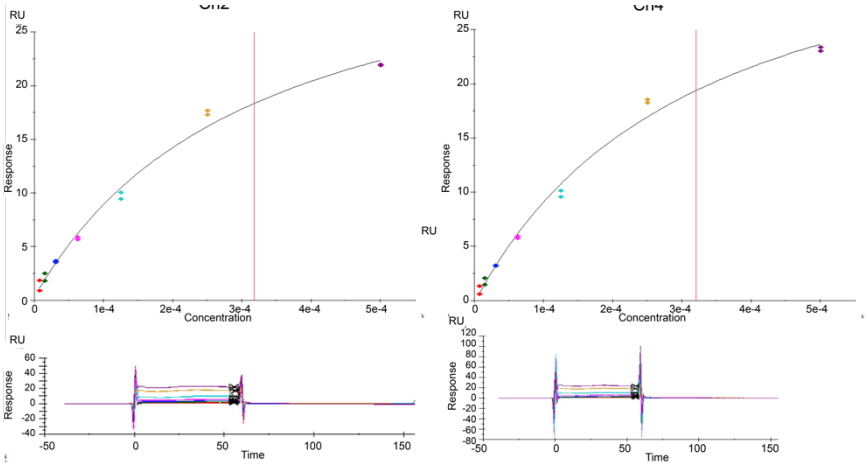


Figure 34 SPR plots for compound 6. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

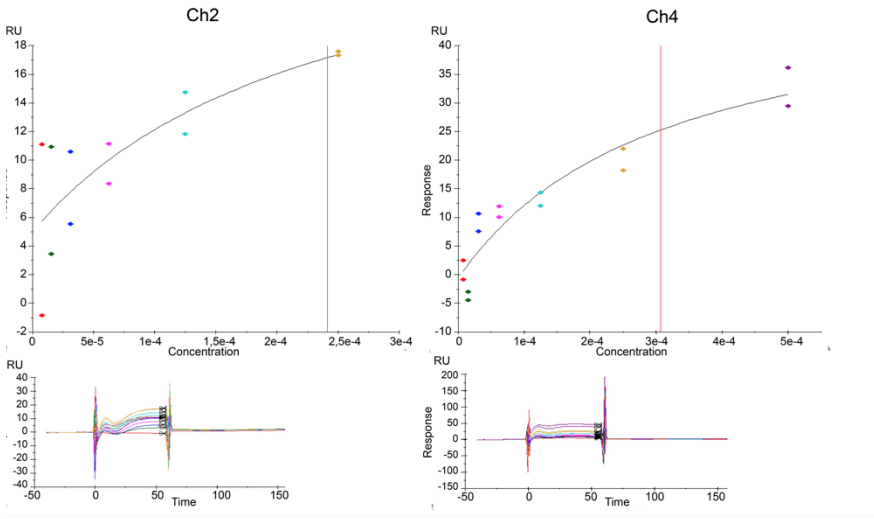


Figure 35 SPR plots for compound 7. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

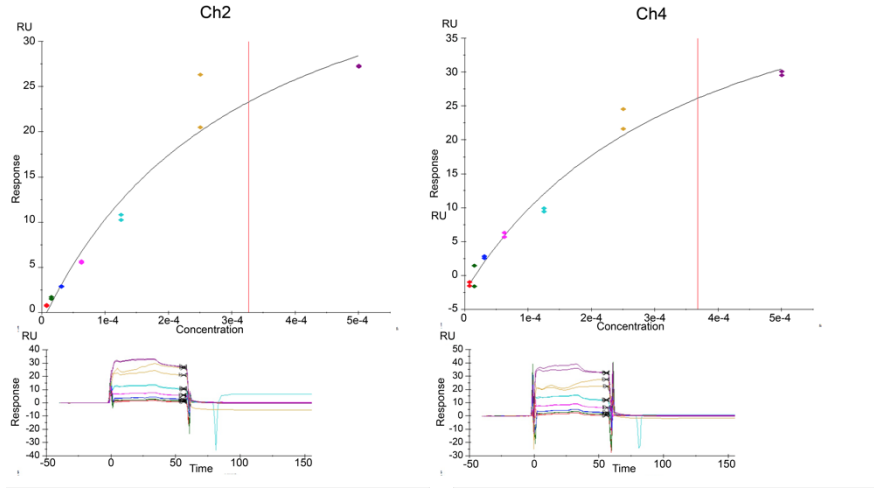


Figure 36 SPR plots for compound 8. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

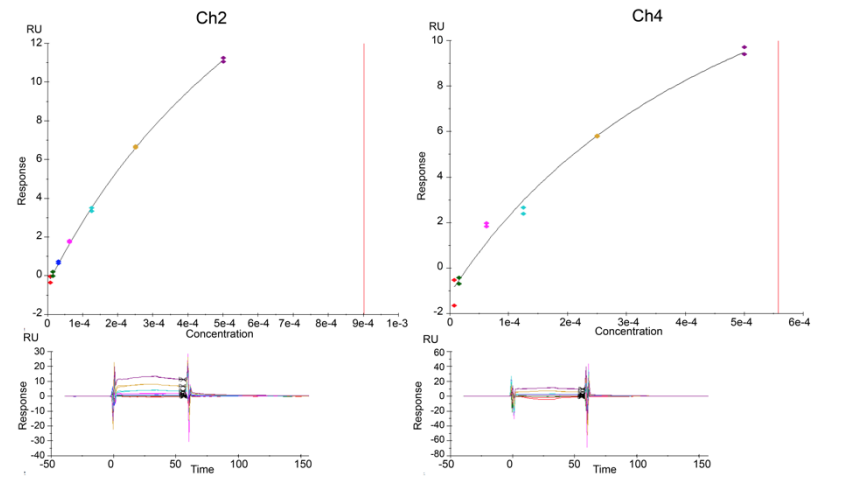


Figure 37 SPR plots for compound 9. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

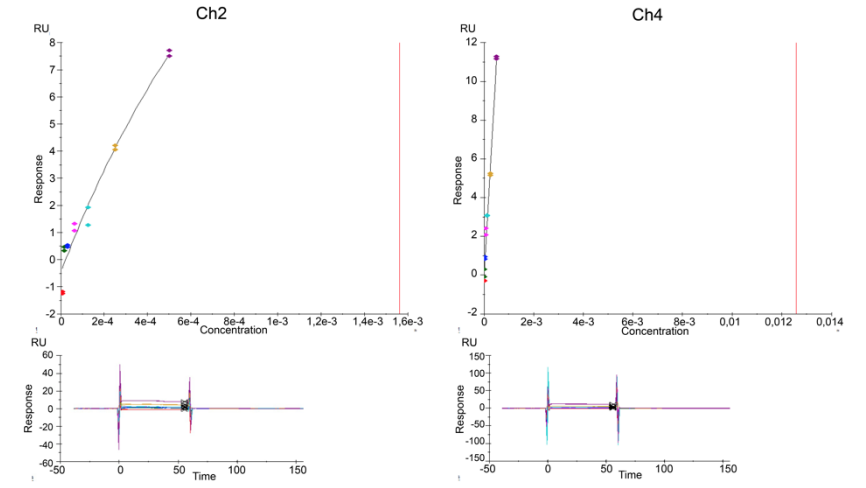


Figure 38 SPR plots for compound 10. In the top row are displayed the steady state response against the concentration for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right). In the Bottom row are displayed the sensograms for the low-density Channel (Ch2; left) and the high-density Channel (Ch4; right).

4.2.5 ONGOING CELL-BASED EXPERIMENTS

This project was done in collaboration with the group of Eva Gonzalez Suarez on the CNIO. Thanks to the success in the virtual screening, this project was awarded an ERC Proof of Concept (**ERC-PoC-2022-1; project name TargetRank**) to assess the therapeutic potential of the molecules found in the Virtual Screening and to further develop the compounds. To that aim, the group at CNIO is testing the ability of the compounds to inhibit the constitutive and the RANKL-dependent activation of RANK by means of cell-based assays.

The first assay performed was using the breast cancer cell line HCC1954 with basal levels of RANK and the same cell line where RANK was removed from the membrane (shRANK) [**Figure 39**]. The activity of RANK was measured using the level of expression of the mRNA of BiRC3, a protein involved in the RANK signaling pathway. In the left plot in **Figure 39** was tested if compounds 2 and 3 were able to inhibit the constitutive activation of RANK. For compound 2 we do not see any difference in the levels of BiRC3 compared to the control. Compound 3 seems to work as an agonist, activating the pathway and increasing the levels of BiRC3. On the other hand, the right plot in **Figure 39** shows the results for the inhibition of the RANKL-dependent activation. Compound 2, in all the concentrations tested, reduced the levels of BiRC3 mRNA, indicating that compound 2 is able to inhibit the RANKL-dependent activation. For compound 3 we did not see a decrease on BiRC3 levels.

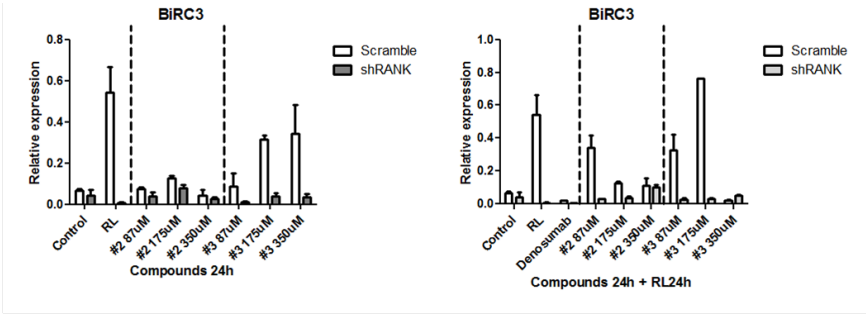


Figure 39 HCC1954 . In the plots is measured the activation of RANK by means of the increase in BiRC3 mRNA levels. The shRANK (grey bars) refers to a modified HCC1954 cell line were RANK is not expressed on the membrane. Scramble (white) refers to a HCC1945 with basal levels of RANK on the membrane. On the left is tested the ability of the compounds 2 and 3 to block the constitutive activation of RANK. On the right is tested the ability of the compounds to inhibit RANK-RANKL interaction.

To better study the effect of the molecules, the HEK-Blue cell line is being used to monitor the activation of the NF-κB and AP-1 pathways, both implicated in the RNKL/RANK signaling pathway. Only the ability to block the constitutive activation was tested for the moment. As a result, compound 1 and compound 2 induced the activation of RANK [**Figure 40**] whereas compound 8 was able to inhibit RANK in a dose-dependent manner [**Figure 41**].

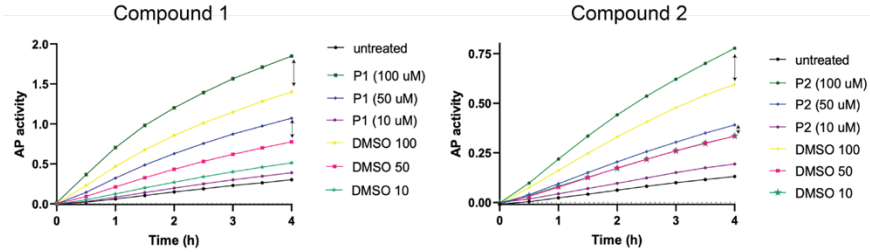


Figure 40 HEK-Blue response to Compound 1 and Compound 2. In the plots is represented the activity of Alkaline phosphatase after treating the HEK-Blue cell line with compound 1 (left plot) and compound 2 (right). In both plots, the arrows indicate the difference between the treated and the DMSO control for each concentration.

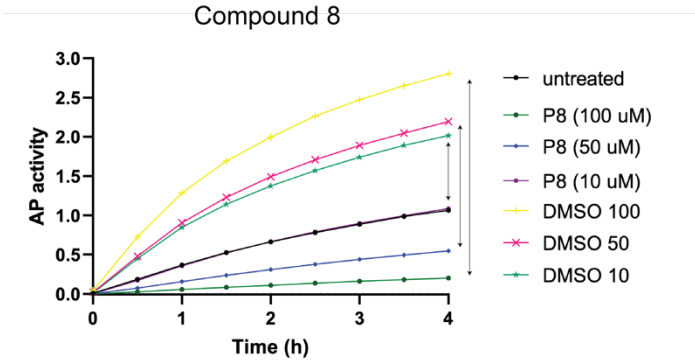


Figure 41 HEK-Blue response for Compound 8. The plot represents the activity of Alkaline phosphatase after treating the HEK-Blue cell line with compound 8. The arrows indicate the difference between the treated and the DMSO control for each concentration.

4.3 RESULTS FOR TARGETING PTEN WITH A COMBINATION OF TARGET-BASED AND PHENOTYPIC SCREENING APPROACHES

4.3.1 BACKGROUND: TARGET-BASED DRUG DISCOVERY VS PHENOTYPIC SCREENING

Historically, new medicines were discovered through observation of their therapeutic effect either directly in humans as part of traditional medicine or in models of disease. However, with the advent of the molecular biology revolution of the 1980s and the sequencing of the human genome in 2021, the focus shifted towards the study of specific molecular targets.

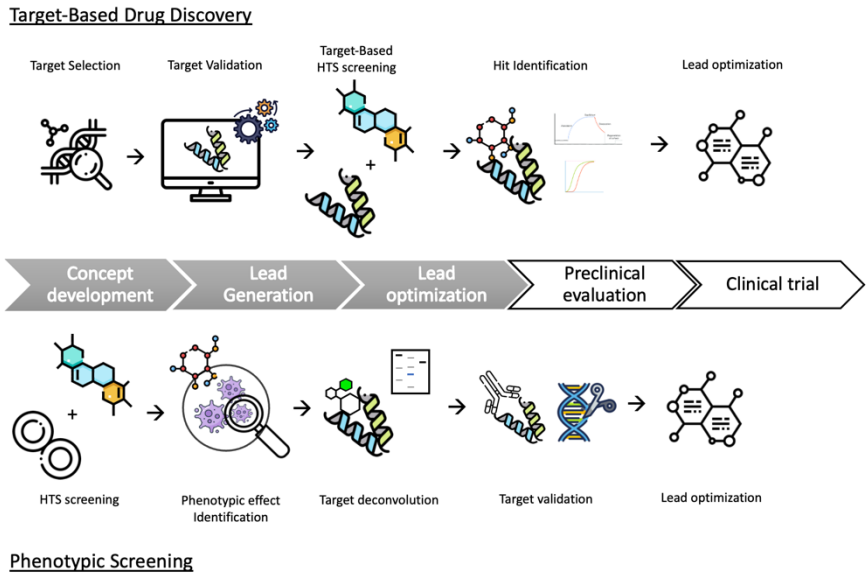


Figure 42 Comparison of target-based and phenotypic screening approaches in early drug discovery.

Target-based drug discovery (TDD) [**Figure 42**] approaches offer a rational pathway for drug discovery [175]. It is based on the premise

that, once a biological component has been identified as disease-modifying, it is possible to work with the component in its purified forms instead of in the complex natural environment (cell, tissue, or organism). Biophysical and biochemical assays make it possible to detect molecules that bind to and alter biological activity, respectively. This offers several desirable possibilities. For instance, the assays can be set up in a high-throughput mode, facilitating the hit identification phase. Additionally, it is possible to carry out the preliminary stages of optimization using the in vitro assay. In this way, when the molecules are tested in cell- or organism-based assays, they are already sufficiently potent to exert a measurable effect. An example of a successful TDD approach is the identification of the 5HT2a receptor as a key molecular target involved in psychosis [176]. After the discovery of the target, the drug pimavanserin was identified as an inverse agonist and approved by the US Food and Drug Administration (FDA) to treat Parkinson's disease psychosis in 2016 [177]. On the downside, setting up the assays is a costly and lengthy process involving the production and purification of the biological component, followed by assay development and validation. This demands an up-front commitment to the project that often exceeds the capabilities of a mid-sized academic lab.

Furthermore, there is a growing interest in non-standard mechanisms of action, such as allosterism or conformational trapping. In such cases, the relationship between binding and the biological response is unpredictable, because the biological response (if any) will depend on the biological function of the particular allosteric site, which is generally unknown. Investing resources in the development of binding and/or biochemical assays may be, not only expensive but also the wrong strategy here [178]. Instead, one should focus on the biological outcome from the onset of the project, pursuing only molecules that 1) modify the behaviour of the biological system, and 2) do so through direct interaction with the intended target.

Furthermore, in recent years there has been a revival in interest in phenotypic drug discovery (PDD) [**Figure 42**] approaches [179–184] following the observation that the majority of the first-in-class drugs

approved by the FDA between 1999 and 2008 were discovered following a PDD strategy without a previous drug target hypothesis. In the case of PDD a physiologically relevant biological system or cellular signalling pathway is directly targeted to identify biologically active compounds that yield the desired phenotypic effect. The majority of successful drug discovery programs combine target knowledge and functional cellular assays to identify drug candidates with the most advantageous molecular mechanism of action. One of the main challenges in PDD is the relation between the disease model and the biology of the disease in humans (chain of translatability). Another challenge for PDD is the development of disease-relevant cell systems that are valid for high throughput hit identification [185–187]. However, technological advances in cell and molecular biology are enabling the development of models that are likely to strengthen the chain of translatability even in model systems that have reduced physiological complexity, by closely modelling the disease-relevant cell or cells and tissue, and/or focusing on the molecular and mechanistic phenotype.

In this project, we want to assess the gained benefit of combining these two seemingly antagonistic strategies (TDD and PDD). The target-based approach, allow us to screen huge virtual collections of chemical compounds, selecting molecules that can Bind an allosteric site of functionally central proteins (hubs). Experimental evaluation of the compounds in a phenotypic assay will allow us to focus on sites and molecules that elicit a biological response, discarding candidates that either do not bind or do not cause a functional effect.

4.3.2 TARGET SELECTION AND DRUGGABILITY STUDY OF POTENTIAL TUMOR SUPPRESSORS

First, we identified unexplored targets with therapeutic potential where the proposed approach could provide an advantage over the standard target-based approach [188]. The targets should be disease-associated, have structural information, allosteric cavities and, if possible, a strong

indication that their modification will elicit an identifiable change in gene expression profile.

To select our POI we evaluated the TUSON-p-values from a list of 18.641 gens evaluated by *T.Davoli et al.* [118]. Analysis done in the same article showed that there is no clear cutoff for predicting cancer drivers. Instead, there is a continuum decreasing the probability of a given gene being a cancer driver. By setting the p-value at 0,005, we obtained a list of 366 possible targets. From those 366 genes, we kept only genes coding for a protein with 3D structure and discarded kinases, nuclear receptors, and other well-known targets. At the end of the filtering process, we had 14 target candidates. To investigate their druggability we first did a search for druggable pockets with fpocket, and discarded the structures where no pocket with a druggability score larger than 0,5 was found, which resulted in 9 proteins with at least 1 druggable pocket [**Table 7**].

Table 7 TUSON p-value and Druggable Pockets found with fpocket for the candidate targets.

Gene Name	TUSSON p-value	Druggable Pockets
PTEN	3,57e-101	2
PBRM1	5,37e-67	1
RB1	2,62e-66	3
VHL	4,27e-62	2
NF1	3,81e-59	0
KDM6A	3,33e-46	1
FBXW7	6,03e-29	0
SMARCA4	3,83e-09	1
DDX3X	2,63e-06	0
THRA	5,86e-08	1
MEN1	6,87e-08	1
DNMT3A	2,97e-08	0
TET2	4,67e-07	0
PTPRK	6,60e-04	1

After visual inspection of the putative pockets, PTEN, VHL, SMARCA4, MEN1, and PTPRK were selected for further druggability studies. In silico solvent mapping was performed with MDmix to explore the potential to interact with small molecules [**Figure 43**]. The systems were tested with a set of molecules containing polar and non-

polar groups, which recapitulate the most common moieties of drug-like ligands. Of those 5 targets, MEN1 was discarded as it did not have any clear combination of hotspots that were adequate for a virtual screening campaign. Of the other 4, PTPRK and PTEN were the ones that showed, an allosteric site with mainly hydrophobic hotspots and some key polar hotspots that could be used as pharmacophoric restraints. As PTEN had a much more significant association with cancer, we selected this protein as our target.

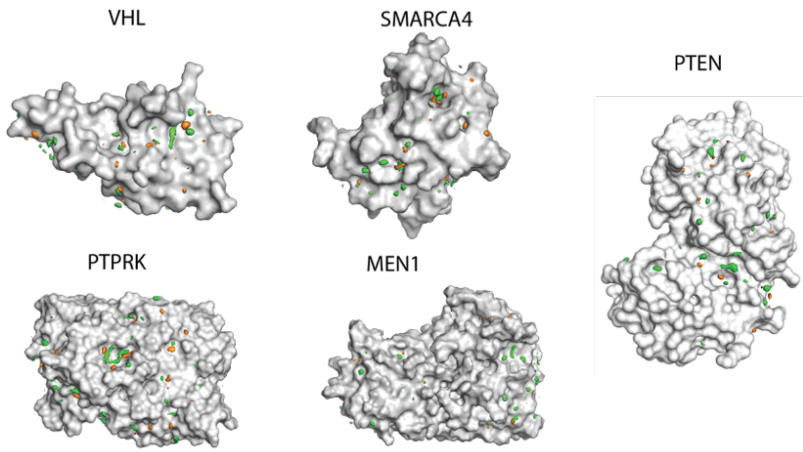


Figure 43 MDmix results for PTEN, VHL, SMARCA4, MEN1, and PTPRK. In green are depicted the hydrophobic hotspots and in orange the polar hotspots.

4.3.3 BACKGROUND ON PHOSPHATASE AND TENSIN HOMOLOG (PTEN): FUNCTIONS AND REGULATION

PTEN is one of the most frequently mutated tumor suppressors in human cancer. Even a small decrease in PTEN levels or activity increases the risk of tumor progression and development [189–193]. Structurally, PTEN is composed of five functional domains: a short N-terminal phosphatidylinositol (PtdIns)(4,5)P-binding domain (PBD), a catalytic phosphatase domain, a C2 lipid/membrane-binding domain, a C-terminal tail containing Pro, Glu, Ser, and Thr sequences and a class I PDZ-binding motif. PTEN exerts its tumor-suppressive functions in

a lipid phosphatase-dependent, protein phosphatase-dependent, or scaffold-dependent manner.

PTEN can be found both in the cytoplasm and nucleus where it is involved in many processes [**Figure 44**]. In the cytoplasm, it has a big influence on cell motility and polarity processes by contributing to establishing a gradient of PIP2-PIP3, it also plays a role in cell division and in the regulation of the tumor microenvironment. We can also find PTEN in the mitochondria, where it regulates homeostasis. In the nucleus, PTEN is involved in the control of genomic stability by maintaining centromere stability, and positively regulating DNA double-strand break repair. It also is involved in the induction of apoptosis, cell cycle arrest and senescence.

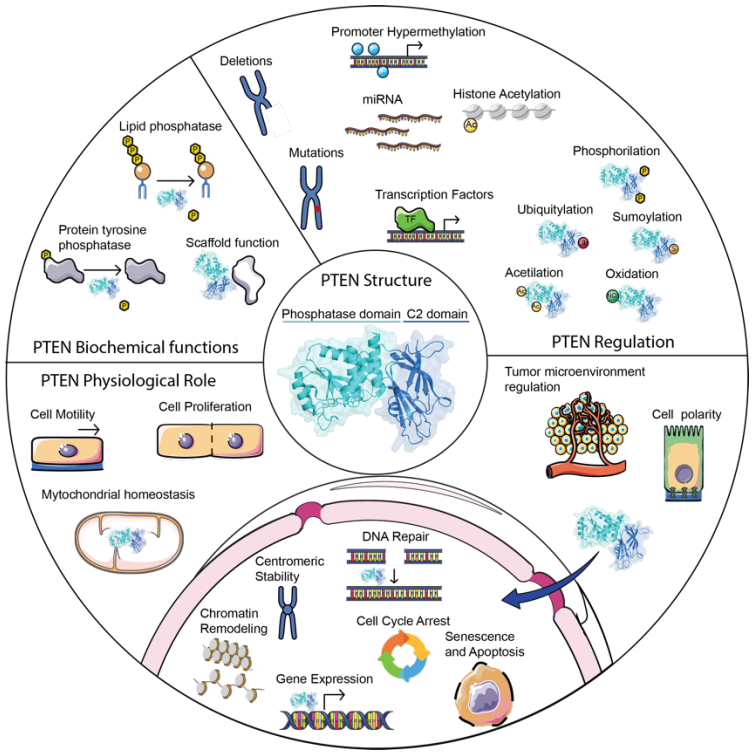


Figure 44 Summary of PTEN Biochemical functions, Regulation, Physiological Role and Structure.

Not only PTEN is involved in a plethora of processes, but it also counts with really complex regulation, comprising Post Transcriptional modifications (i.e, Ubiquytilation, Phosphorylation, Sumoylation), Epigenetic control (i.e Histone Acetylation), and Genetic alterations that affect its function. The centrality and pleiotropy of this target means that binding to an allosteric site may have unpredictable consequences, making it an ideal target for our in vitro to phenotypic approach.

4.3.4 IDENTIFICATION AND CHARACTERIZATION OF A NOVEL ALLOSTERIC SITE IN PTEN

The MDmix analysis identified multiple interaction hotspots over the surface, however, only near the hinge we saw a cluster of mainly hydrophobic and polar hotspots overlapping with one of the pockets identified earlier with fpocket. If we take a look at the residues forming the pocket, we can see that it is heavily populated by aminoacids with hydrophobic side chains (i.e Leu, Tyr, Phe, Ile, Val) which has been reported to be a common trait across allosteric sites [194].

The most energetic hotspots identified were one hydrophobic in close proximity of Tyr-164 and Phe-266, another hydrophobic one near Arg-160, and finally, one polar hotspot making an interaction with the hydroxyl of Tyr-164. These 3 hotspots were then selected as pharmacophoric constraints to be used during the virtual screening.

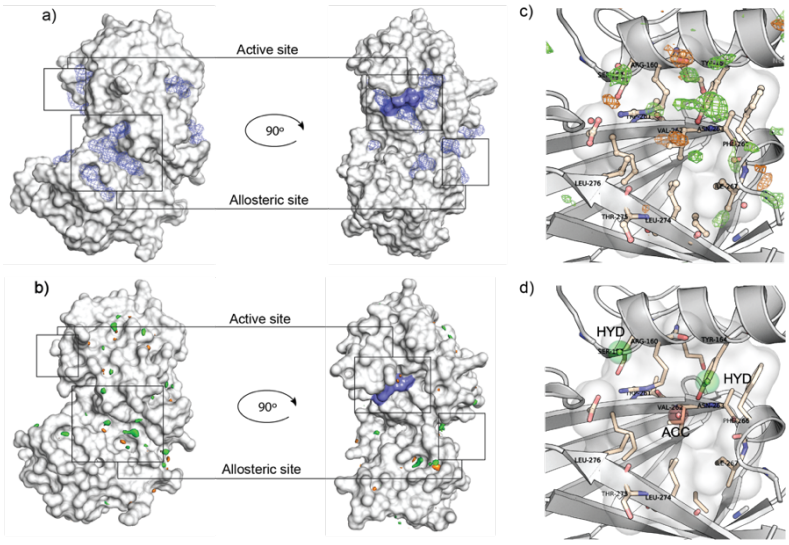


Figure 45 Druggability Study of PTEN a) fpocket results b) hotspots identified with MDmix in green are depicted hydrophobic hotspots and in orange polar hotspots c)Distribution of hotspots of the allosteric binding site d) Selected pharmacophoric points to be used in the virtual Screening.

4.3.5 VIRTUAL SCREENING USING PHARMACOPHORIC RESTRAINTS

The protocol used for the virtual screening is summarized in **Figure 46**. A virtual library composed of ~7M commercially available compounds assembled in-house from several vendors was used for the virtual screening campaign. After using the docking protocol detailed in the methods section, we obtained a total of 48.831 molecules that fulfilled the pharmacophoric restraints and had a SCORE.INTER <-25 KJ/mol. To decrease the number of candidate molecules while maintaining the chemical diversity we performed a clustering step where molecules with a Tanimoto similarity, based on MACCS keys fingerprints, higher or equal than 0,95 are grouped in the same cluster and only the molecule with the best docking score is selected as a cluster representative. With this clustering step we obtained 30.733 clusters, reducing the number of candidate molecules by 37%.

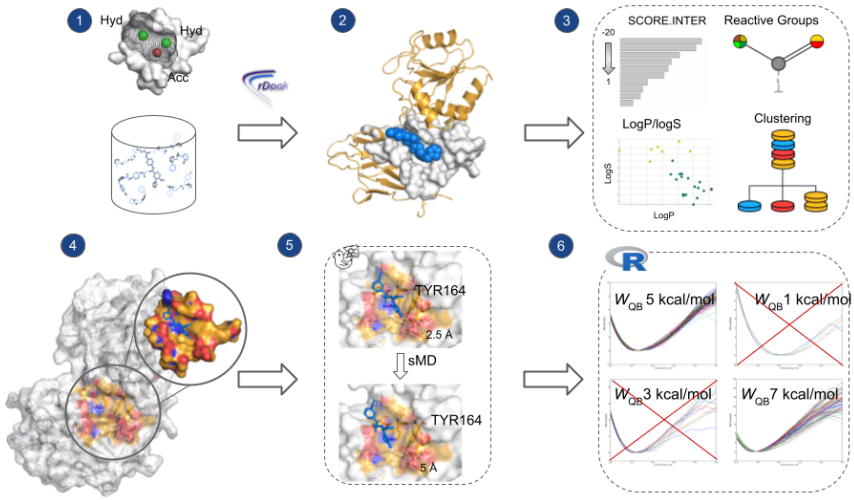


Figure 46 Virtual Screening protocol used for the identification of compounds binding to the novel allosteric site of PTEN.

The 2.000 top-scoring molecules, were subjected to the DUck protocol, pulling from the hydrogen bond formed between the ligands and the hydroxyl of Tyr164. Interestingly, only 67 ligands surpassed the W_{QB} threshold of 3 Kcal/mol, which is a really low threshold for drug-like molecules, and only 24 had a W_{QB} value larger than 4 Kcal/mol. However, we noticed that the majority of molecules showing good interaction scores in the DUck step contained a substituted tetrazole making the interaction with Tyr-164, which can indicate that the ligands with this structure are preferred in this cavity.

To assess this result, we used MDmix and selected from the already pre equilibrated solvent mixtures the organic molecule with the most similar structure to tetrazole isoxazole. We then examined the energy grid obtained to see if we could report high-affinity interactions between the solvent and the Tyr-164. In **Figure 47** we can see that the most energetic hotspots are the ones interacting with Tyr-164 (marked with the red square), and the one that is driving the interaction is the polar spot (purple).

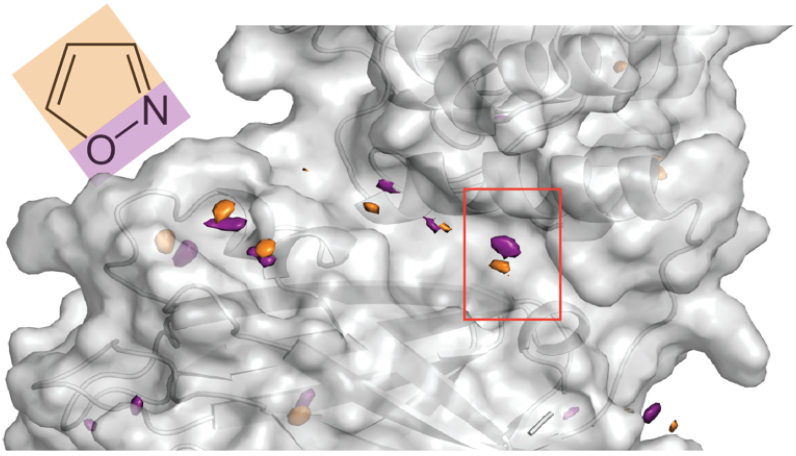


Figure 47 MDmix results for the allosteric site of PTEN using isoxazole as solvent. The MDmix hotspots are represented in pymol with an isosurface with a contour level of 1.5.

Having this hypothesis confirmed, we now wanted to search in a larger compound database for compounds containing this group to see if we could find other ligands with better affinity than the ones that we already identified. The database that we chose to explore was ENAMINE REAL db (ERdb), with about 1.2B compounds (as by July 2020). Using the pattern illustrated in **Figure 48** we were able to retrieve about 7.5M compounds that contained a substituted tetrazole.

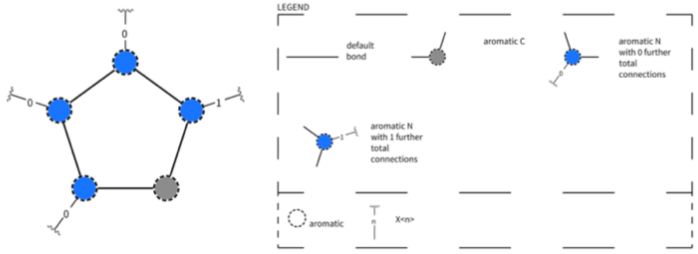


Figure 48 Smarts representation used to search in Enamine Real DB. Picture created by SMARTSviewer [195] with the SMARTS: c1[n;X2][n;X2][n;X2][n;X3]1.

Then we filtered all the compounds that presented reactive groups and that did not fit into the Lipinsky rules, obtaining a collection of 5.5M compounds. To ensure that we had a diverse set we applied again a clustering step using a Tanimoto similarity of 0,9 using MACCS keys

fingerprints. After the clustering we obtained a diverse collection of 189.852 compounds that, after ligand preparation to generate protomers, tautomers, esteroisomers and ring conformations, resulted in 520.290 molecular states to be docked.

We applied the same docking protocol and filters described above to this focused compound collection, resulting in 202 compounds that surpass the W_{QB} threshold of 3 kcal/mol and 22 compounds had a W_{QB} > 5kcal/mol. After visual inspection a diverse set of 15 ligands coming from the initial collection (6) and from the filtered version of ERdb (8) were selected, from which 14 were available for purchase [**Table 8**].

Table 8 Results of WQB and SCORE.INTER for the 14 prioritized compounds. The Collection HTSDB makes reference to the in-house collection and ER makes reference to the filtered version of ENAMINE REAL Db.

ID	MW(Da)	W_{QB}(kcal/mol)	SCORE.INTER (KJ/mol)	Collection
1	455,0	9,5	-31,9	HTSDB
2	361,4	7,2	-28,1	HTSDB
3	332,8	5,0	-25,1	HTSDB
4	322,4	4,8	-26,8	HTSDB
5	337,4	4,3	-27,4	HTSDB
6	437,6	4,0	-33,5	HTSDB
7	466,3	8,7	-29,2	ER
8	492,5	7,8	-30,1	ER
9	496,6	7,6	-28,8	ER
10	453,5	6,0	-29,4	ER
11	406,5	5,6	-28,6	ER
12	374,4	5,5	-32,9	ER
13	446,3	5,3	-28,2	ER
14	419,4	5,1	-29,3	ER

4.3.6 CMP1 INDUCES MORPHOLOGICAL CHANGES IN HCT116 PTEN +/- CELL LINE

To measure the cytotoxicity of the compounds and as a primary screen to test the effect of the compounds over the cell lines, we performed a CellTiter-Glo assay. All 14 molecules were tested at 3 different concentrations 10μM, 25μM and 50μM and the treatment was refreshed at 24h and cell viability was measured after 72h of treatment. We first performed the assay with HCT116 PTEN (+/-) and HCT116 PTEN (-/-) cell lines and compared if we had a differential effect on proliferation [**Figure 49a**]. As a positive control we included two previously described PTEN inhibitors SF1670 and VO/OH [196–198]. From this first analysis we saw a clear cytotoxic effect for compound and 8 even in the lowest concentration tested for both cell lines and was discarded from further analysis (not shown in **Figure 49**). In general, under these experimental conditions for the majority of the compounds, including the positive control, we saw no clear difference in proliferation between the cell lines. Only in compound 1, 7, and 9 we see a difference in proliferation at the highest concentration, meaning that these compounds had a more noticeable effect on the HCT116 PTEN (+/-) cell line than on the HCT116 PTEN (-/-). In the light of this results, we decided to further study the proliferation differences between the PTEN (+/-) and PTEN(-/-). We compared the proliferation rate on DMSO between the cell lines, and, as shown in **Figure 49d**, the complete loss of functional PTEN has no noticeable effect on overall proliferation. Interestingly, the HCT116 cell line has mutations in KRAS and PI3K genes which cause an overactivation of pathways related to cell proliferation [199]. We hypothesize that the effect of these mutations in synergy with the growth factors present in FBS could be overcoming the effect of the compounds over this cell line. To test this hypothesis, we decided to reduce the serum concentration from 10% FBS to 0.1% FBS and repeat the CellTiterGlo assay [**Figure 49b**].

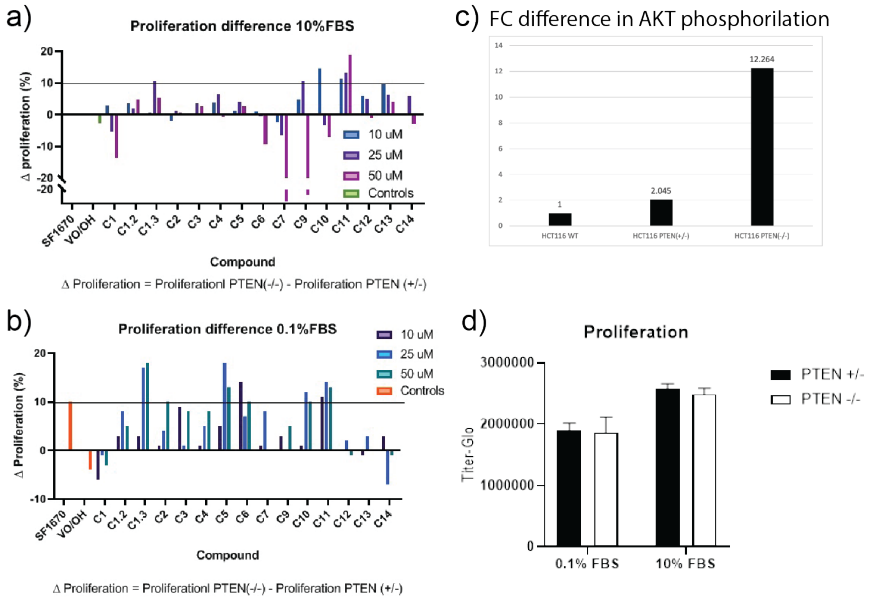


Figure 49 Cell Titter-Glo assay a) Proliferation differences between HCT116(+/-) and HTC116(-/-) using 10% FB when treated with the compounds b) Proliferation differences between HCT116 (+/-) and HCT116(-/-) using 0.1% FBS when treated with the compounds c) Difference in AKT phosphorilation between HCT116 cell lines d) Basal proliferation of HCT116 (+/-) and HCT116 (-/-).

For compound 1 (from here on refered as CMP1) in all the concentrations tested, we saw a clear change in HCT116 PTEN (+/-) cell morphology, which could be linked to an Epithelial-mesenchymal transition (EMT). EMT is a biological process that allows a polarized epithelial cell to undergo a series of biological changes that enable it to assume a mesenchymal phenotype [200]. In many studies the activation of the EMT program has been proposed as the turning point for the acquisition of a malignant phenotype in many epithelial cancer cells [200–204]. More importantly, it has been reported by previous studies that PTEN loss induces EMT and cancer stem cell activity in human colon cancer cell lines such as HCT116 [205–208]. To confirm the EMT we tracked the levels of E-cadherin and Vimentin as the loss of E-cadherin and an increase in Vimentin have been linked to EMT [201,202]. We can clearly see that, when treated with CMP 1, there is a drastic decrease in E-cadherin [**Figure 50a**] levels and an increase in

Vimentin [**Figure 50b**] in HCT116 PTEN(+/-) cells compared to the control. It is also worth noting the big difference in cell morphology when we compare the treated cells to the control.

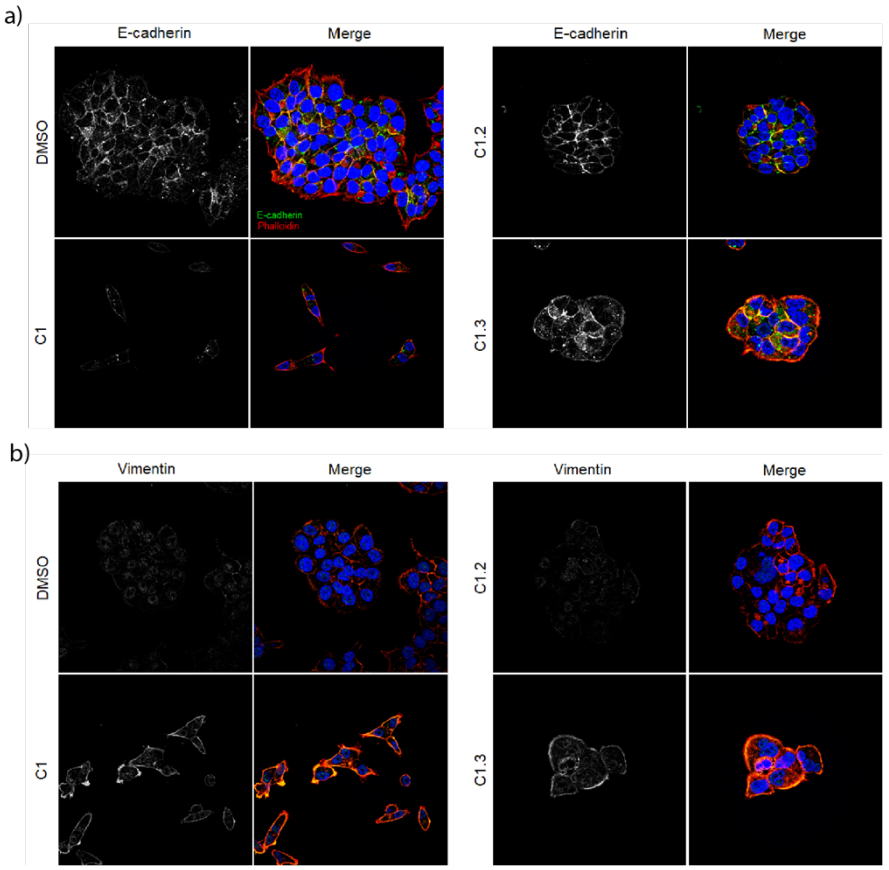


Figure 50 Immunohistochemistry for HCT116 PTEN (+/-) treated with CMP1, CMP1.2 and CMP1.3. Nuclei are dyed in blue, actin filaments in red and E-cadherin in green.

4.3.7 CMP1.3 INDUCES POLYPLOIDIZATION OF HCT116 PTEN(+/-)

In view of the effect that CMP1 has on the cell line, we wanted to see if some analogues for this compound could have a similar effect. To that end, we revisited the results from the virtual screening and looked for compounds that shared a similar scaffold to compound 1 and that passed all the post docking filters. From this step, we selected two additional compounds, CMP1.2 and CMP1.3 that although showing lower W_{QB} values and lower SCORE.INTER than CMP1 [**Table 9**] they shared a common scaffold and the predicted binding mode by docking was similar.

Table 9 Values of W_{QB} and SCORE.INTER for compounds 1, 1.2 and 1.3. HTSDB referes to the in-house collection of compounds.

ID	MW(Da)	W_{QB}(kcal/mol)	SCORE.INTER (KJ/mol)	Collection
1	455,01	9,5	-31,9	HTSDB
1_2	353,44	7,6	-28,0	HTSDB
1_3	345,44	4,2	-25,1	HTSDB

Both compounds were then tested first with the CellTiter-Glo assay and then the same markers for EMT were evaluated. From the proliferation results, we did not see any clear difference between cell lines in any of the concentrations tested [**Figure 49**]. As for the markers related to EMT (E-cadherin and Vimentin), for CMP1.2 the levels of E-cadherin and Vimentin were similar to the control [**Figure 50**]. For CMP1.3 we see an increase in Vimentin compared to the control but not a decrease in E-cadherin. However, we can see that, after the treatment, the cells showed a different phenotype where we can observe an enlargement of the cell and the presence of more than one nuclei [**Figure 51**]. It has been shown in previous studies that PTEN deficiency is related to polyploidization, as it is able to regulate and control chromosomal stability during cell division [209].

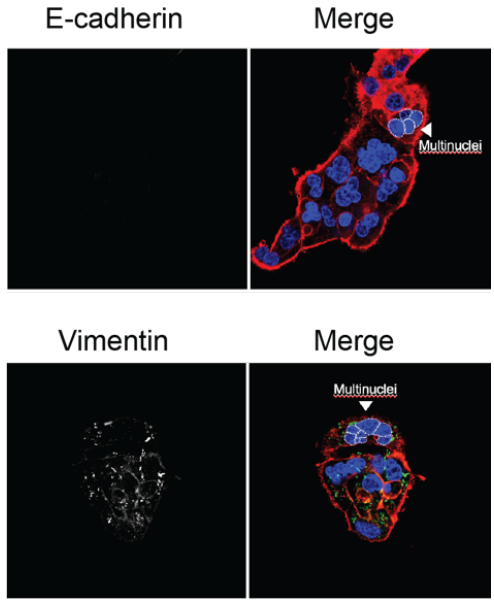


Figure 51 Immunohistochemistry for HCT116 PTEN (+/-) treated with CMP1.3. Nuclei are dyed in blue, actin filaments in red and E-cadherin in green.

4.3.8 SPR CONFIRMED BINGING FOR SOME OF THE COMPUTATIONAL HITS INCLUDING CMP1 AND CMP1.3

Due to PTEN's multiple roles and implications in many biological pathways we wanted to assess via SPR biosensor experiments the binding of all the tested compounds even if they did not show any phenotypic effect on the cell line. We immobilized PTEN in a CM5 sensor chip with immobilization levels between 5.000 and 8.000 RU following the protocol in section 3.5.2.6. In **Table 10** are summarized the SPR results for all the compounds. Compounds CMP1, CMP1.2, CMP1.3, CMP8 are the ones that showed a better affinity for PTEN, having a Kd of 65,37μM, 25,02μM, 18.82μM and 62,84μM respectively. Interestingly, with compounds CMP.6 and CMP.7, we observed that the Rmax obtained double the expected Rmax in both cases. This could indicate that these two compounds have a 2:1 stoichiometry, corresponding to a two-binding site model. On the other hand,

compounds CMP2, CMP3, CMP5, CMP9, and CMP12 showed little to no response. Compound CMP4, CMP11, CMP13, and CMP14 showed nice kinetics profiles but we were not able to reach saturation. CMP10 needed also higher concentrations to reach saturation, however, this compound was only tested at a maximum concentration of 20µM because it precipitates. This is in accordance with the cell assays as at 10µM already formed crystals at 1% of DMSO.

Table 10 SPR results for PTEN compounds

ID	K_D(uM)	SD (uM)	Rmax (RU)	Chi^2
1	65	±15	46,70	3,41
1.2	25	±33	5,50	1,92
1.3	19	±4,2	16,90	3,57
2	-	-	-	-
3	279	±21	81,83	3,56
4	81	±41	79,40	132,00
5	235	±25	83,10	7,95
6	163	±16	167,90	15,70
7	162	±39	104,00	21,20
8	55	±8,5	41,70	0,83
9	382	±15	122,20	1,19
10	19	±0,97	111,70	3,30
11	160	±78,95	20,70	0,77
12	357	±45	214,00	15,80
13	95	±17	35,90	3,72
14	24	±16,55	12,00	2,97
SF1670	9	±3,7	34,20	4,09

4.3.9 CMP1 AND CMP1.3 CAUSE A MORPHOLOGY CHANGE INDEPENDENTLY OF PTEN LIPID PHOSPHATASE ACTIVITY

One of the main and most well-known functions of PTEN is the lipid phosphatase activity, by which it regulates the levels of PI(3,4,5)P3 in vivo. Not only that, but the lipid phosphatase activity has also been proven to be critical for the tumor suppressor function of PTEN [210].

To assess if our compounds had an effect on the lipid phosphatase activity of PTEN we performed a Malachite Green assay using PIP3 diC8 as substrate and quantified the free phosphate resulting from PTEN enzymatic activity. Strikingly, only CMP14 showed a significant decrease in substrate conversion, 20% decrease compared to the DMSO [**Figure 52**]. The positive controls SF1670 and VO-OH showed a decrease of more than 40% compared to the DMSO.

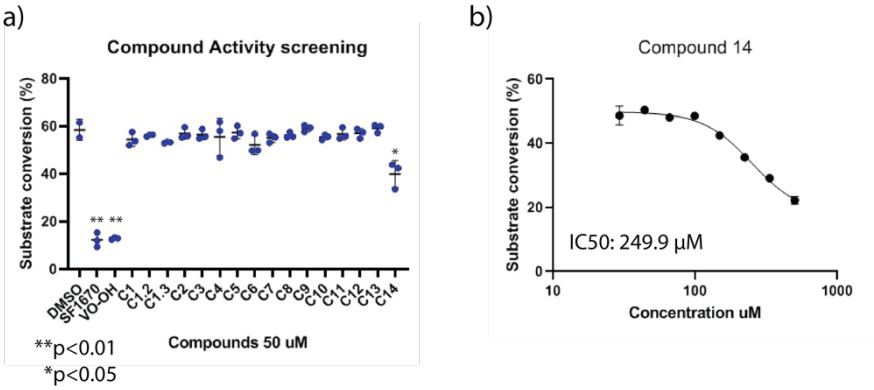


Figure 52 PTEN activity assay a) PTEN activity assay results b) Dose-Response of compound 14.

4.3.10 ASSESSING THE BINDING MODE STABILITY FOR CMP1, CMP1.2 AND CMP1.3 WITH MD

To evaluate the stability of the proposed binding mode and assess the dynamics of PTEN, independent MD simulations were run for the Apo structure of PTEN and PTEN bound with CMP1, CMP1.2, and CMP1.3. During the MD simulations of PTEN bound with the compounds, the 3 compounds stayed in the binding site for the whole duration of the simulation. CMP1.2 appeared to be the one with the most stable binding mode, with RMSD fluctuations between 0.5 Å and 1.5 Å. CMP1 and CMP1.3 present a less stable binding mode, with RMSD fluctuations between 1. Å and 3.0 Å. However, in all the cases the interaction with Tyr-164 is maintained for the whole duration of the MD simulation [**Figure 53** top], thus validating the binding mode provided by the virtual screening protocol.

Figure 53 Analysis of the long MD simulations of PTEN apo and PTEN bound to CMP1, CMP1.2, and CMP1.3. Above is represented the RMSD of the ligands during the MD simulation. Below is represented the Root Mean Square Fluctuation for each residue in PTEN.

The root mean square fluctuation (RMSF) of PTEN was calculated to check if there were any differences in the dynamics of the protein when it was bound to the compounds [**Figure 53** bottom]. Compared to PTEN Apo, we could not see any difference in the RMSF when PTEN is bound to the compounds.

4.4 RESULTS FOR BOTTOM-UP EXPLORATION OF THE CHEMICAL SPACE

4.4.1 BACKGROUND ON CHEMICAL SPACE EXPLORATION

Drug discovery starts with identifying a "hit" compound that, after a long and expensive optimization, evolves into a drug candidate. In order to produce new drugs more cheaply and quickly, researchers need to make the drug discovery cycle more efficient. For this reason, big pharmaceutical companies have invested heavily in HTS collections, that nowadays can contain up to 5-million compounds [**Figure 54**], which is a small fraction of the enormous chemical space estimated to be 10^{20} to 10^{60} million available drug-like compounds. On-demand chemical collections emerged a few years ago, pioneered by the chemical supplier Enamine, which created the first catalogue in 2015, with 600 million compounds (https://enamine.net/compound collections/real-compounds). These collections were initially dismissed because chemical synthesis outcomes are uncertain and people assumed that the synthetic success rate would be low. Nowadays, the Enamine collection boasts 32 billion compounds and the company has repeatedly demonstrated the ability to deliver >90% of the requested compounds in 4 weeks at very competitive prices (10 mg for 120€ to 170€).

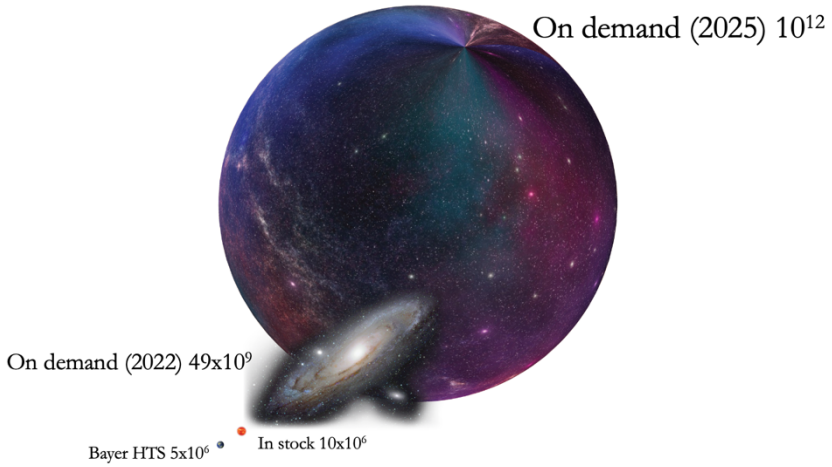


Figure 54 Size Comparison for different chemical spaces.

However, with increasingly bigger libraries, the computational time and cost of the exploration itself become the next bottleneck. But this bottleneck was recently removed with the introduction of the combinatorial search of the chemical space. This approach involves the fragmentation of the database into a collection of building blocks and reactions, which allows for extremely fast searches without the need of enumerating the individual compounds. One of the first methods developed to explore this fragmented space is FTrees-FS [211], a pharmacophoric-style similarity search method that uses a reduced graph representation of the molecules.

Nonetheless, docking-based VS of this huge number of compounds is unattainable, even with parallel cloud computing capabilities. To address this challenge, we have conceived a novel strategy that explores the chemical universe from the bottom up. First, we performed a systematic search of the fragment space (up to 14 heavy atoms), identifying the most promising scaffolds making the key interactions with the receptor. We then search for drug-like molecules that contain this initial scaffold. This allows us to focus only on the most promising areas of the vast chemical space, maximizing the success probability of the selected compounds and reducing immensely the computational cost.

4.4.2 IDENTIFICATION OF NOVEL FRAGMENTS THAT BIND TO BRD4

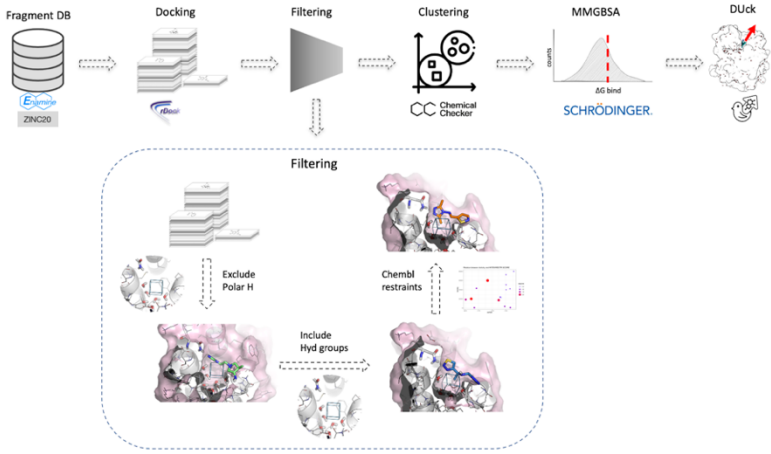


Figure 55 Screening strategy for the exhaustive exploration of the fragment space.

To exhaustively explore the fragment space, we followed the protocol depicted in **Figure 55**. First, we performed a virtual screening using the library of ~11M fragments described in the Methods section [see methods section 3.6.1.1] From previous studies done in the group, we know that BRD4 has a highly preserved water network inside the binding site that favors the placement of a hydrophobic group and the main hydrogen bond is made with Asn-140 [131] [**Figure 56**].

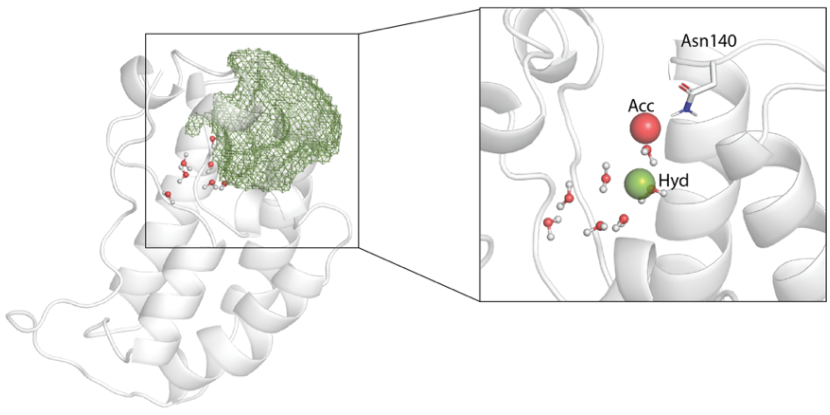


Figure 56 Cavity and pharmacophoric restraints used during the Virtual Screening.

With that in mind, we decided to apply a filtering strategy favoring the placement of a hydrophobic group near the water network. First, all the ligand poses with polar groups near the net of waters were removed. Then, we selected only the ligand poses that placed a hydrophobic group at the bottom of the cavity. Finally, using as a reference the SCORE.INTER values obtained from fragments deposited in CHEMBL with reported activity for BRD4, we selected only those ligands with a SCORE.INTER<-12. At the end of the filtering stage, we were left with 362.345 unique fragments. Even though, we were able to filter out around 89% of the fragments from the initial fragment database a clustering process was needed to reduce the number of putative fragments and to ensure chemical diversity. To that end, we applied a K-means clustering with 2.000 clusters based on the chemical signatures from the Chemical Checker [see methods] and selected only the cluster representative. Then, for the 2.000 cluster representatives, we used MMGBSA calculations to assess the binding free energy using a threshold value of $\Delta G_{bind} \leq$ -30 kcal/mol (value obtained using the set of active fragments from CHEMBL), which discarded 51,3% of the fragments [**Figure 57a**]. As the last step, we used DUck on the remaining 973 fragments to assess structural robustness, as a combined rDock and DUck approaches have been shown to reduce false positives and retrieve higher experimental hit rates[86,153]. By using a value of W_{QB} >7 kcal/mol we filtered out 99,3% of the fragments having only 7 possible fragment candidates. The top 6 candidates were selected for fragment growing [**Table 11**].

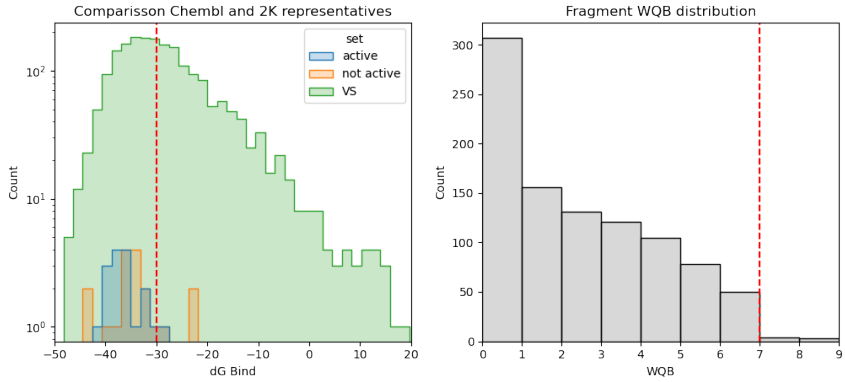


Figure 57 a) Distribution of ΔG_{bind} for the active Chembl set, the non-active Chembl set and the fragments obtained in the VS. In red is represented the threshold value b) Distribution of WQB values for the VS fragments. In red is represented the threshold value.

Table 11 WQB value for the 6 selected fragments.

ID	W_{QB}(kcal/mol)	SCORE.INTER (KJ/mol)
ZINC001234567238	8,7	-13,6
Z2844153759	8,5	-13,5
Z2613868824	8,2	-12,5
Z2844149527	7,7	-12,2
Z4227620390	7,7	-17,2
ZINC000012396782	7,2	-12,8

4.4.3 FRAGMENT GROWING AND EXPERIMENTAL VALIDATION

To test our approximation of fragment growing three different starting points were used: I) Scaffolds extracted from already known drug candidates. II) Experimentally validated fragment hits III) Computational fragment hits.

For each initial scaffold, we perform a scaffold search on Enamine REALSpace using SpaceMacs [212]. At the end of this process, we obtain a scaffold-focused library of druglike compounds. Then a hierarchical HTVS comprised of 4 steps is performed [**Figure 58**]. First, the library of ~10M druglike compounds is docked with rDock restraining the scaffold corresponding to the initial fragment. The

resulting ligands are then clustered by k-means using the Chemical Checker fingerprints. The cluster representatives are then ranked through the ΔG_{bind} and W_{QB} obtained by MMGBSA and DUck respectively. The threshold W_{QB} value was adjusted according to the W_{QB} value obtained for the input molecule.

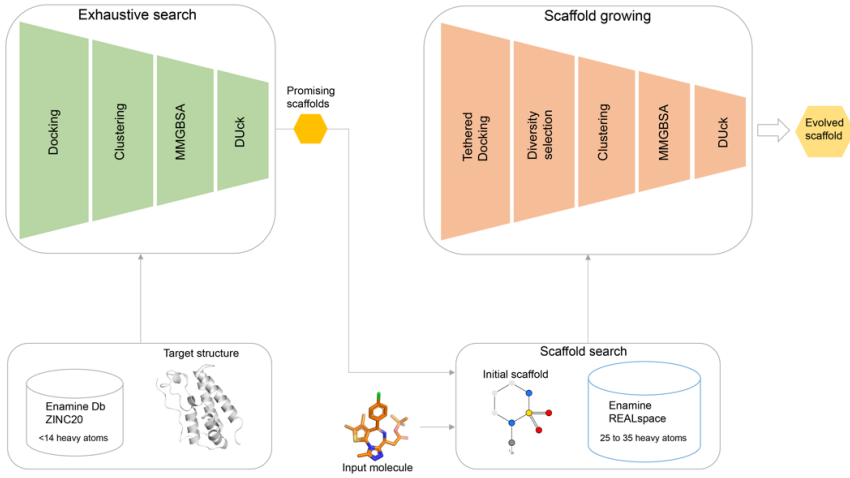


Figure 58 Pipeline used for exhaustive search of the fragment space and for Scaffold growing.

The first experiment for binding assessment was differential scanning fluorimetry (DSF), a thermal shift assay that measures changes in protein stability upon binding. The compounds were tested at 1uM and selected as positive the ones that caused a shift in ΔTm higher than 1ºC. We also applied an orthogonal technique for binding assessment based on time-resolved fluorescence resonance energy transfer (TR-FRET). In DSF a total of 48 compounds showed a change in ΔTm higher than 1ºC, which corresponds to a hit rate of 47%. In TR-FRET we obtained 11 compounds that had an IC_{50} between 1-100nM. Interestingly, 7 of these compounds come from the computational fragment hits, which shows a much higher hit rate (28%) compared to the experimental fragment hits (8%) and to the scaffolds coming from drug candidates (7%) [**Figure 59**].

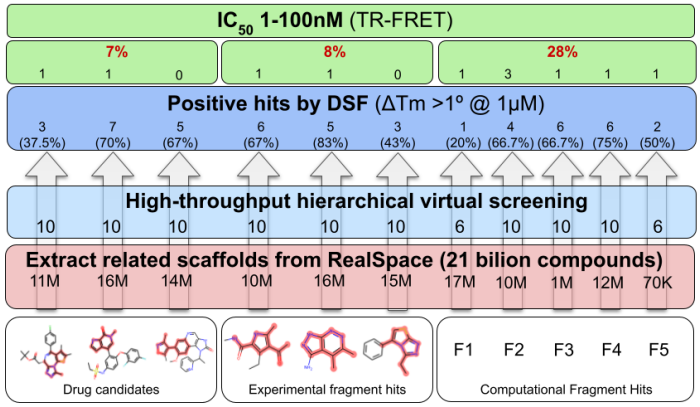


Figure 59 Number of compounds found at different stages for the bottom-up exploration using different starting points:1)Scaffolds from Drug candidates 2)Experimental fragment hits 3) Computational Fragment Hits.

Finally, we were also interested in assessing the chemical diversity of these 11 compounds by comparing them with I) known BRD4 binders from CHEMBL and II) a random set of 50.000 molecules from the Chemical Checker Universe. In **Figure 60a** we can see that the 11 compounds (red dots) are distributed across the projected chemical space (CC random molecules in grey), and not only near the areas enriched with the known BRD4 binders (blue). We also tested the difference between the CC molecules, BRD4 molecules, and the 11 compounds to the set of BRD4 known binders and check the similarity between them [**Figure 60**] using different metrics. In **Figure 60c** is depicted the Tanimoto similarity using Morgan fingerprints, which accounted for the 2D chemical similarity. We can see that the distribution of 11 molecules (orange) is much more similar to the distribution obtained with the CC molecules, than the one obtained with the BRD4 molecules. Finally, to further characterize the compounds we used the CC Chemistry Space (A*) accounting for various chemical properties (i.e 3D similarity, physicochemical parameters). Using the CC signatures, we can see that, while the chemical space for the Brd4 binders are clustered in a similar region, our molecules explore a more diverse chemical space (not biased to the brd4 chemical space) similar to the expected background distribution [**Figure 60d**].

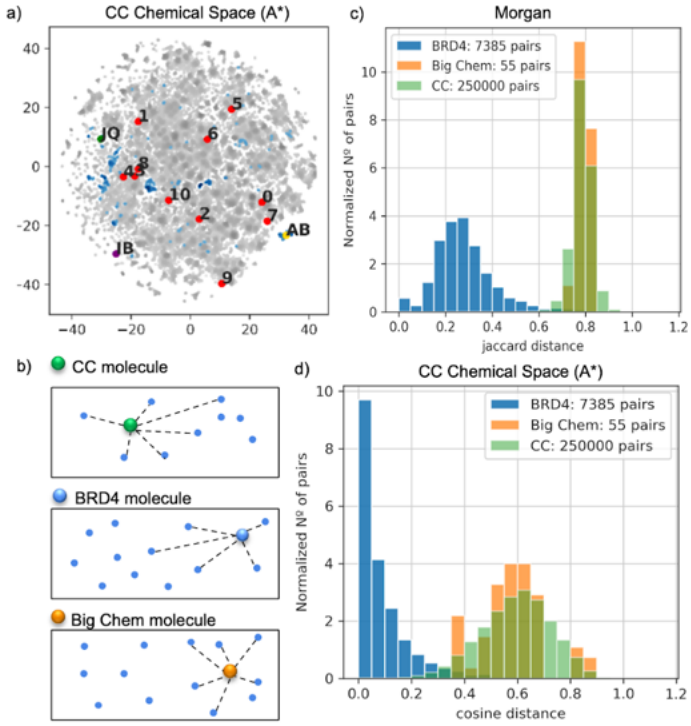


Figure 60 Results for the diversity analysis of the evolved compounds a)t-SNE representation of the CC Chemical space (grey), known BRD4 inhibitors (blue) and compounds found during the bottom-up exploration (red) b)Representation of the cosine distance calculation c) Tanimoto similarity using Morgan fingerprints d) Cosine distance using the CC (A*) fingerprints.

5 Discussion

The process of Drug Discovery is a highly complex, time-consuming, expensive and multidisciplinary process that, more often than not, results in failure. Computational methods are applied with the aim of increasing the odds of finding new compounds with desirable in vitro and in vivo properties thus, helping reduce attrition rates, costs and time needed to release a drug into the market.

The main focus of this thesis is the process of Virtual Screening, and how we can adapt it to obtain the most accurate results, even for challenging targets. To that aim, first and foremost we have to understand the limitations as well as the strengths of the VS method that we wish to apply. In the context of this thesis, all the VS performed are docking-based using rDock software.

In Chapter 4.1 we introduced the importance of quantifying the performance of docking software in real scenarios, which is essential to understand their limitations, manage expectations and guide future developments. With the CELPP Challenge, the D3R consortium aims to provide a fast-growing validation set that better captures all the complexity in a real drug discovery setting. In this Chapter, we presented an initial version of our pipeline for participation in the CELPP Challenge, which applies different knowledge-based docking approaches depending on the already available information on PDB.

To provide a baseline performance, the CELPP team has developed four workflows based on different docking programs, one being rDock. The rDock workflow represents a default implementation of the method without any optimization and using the cavity defined by the challenge. Our protocol has the added challenge of detecting the cavity automatically, but if we consider only the cases where the cavity is correctly predicted, we can observe a significant gain of performance of our protocol relative to the baseline, with improvements in the median RMSD value ranging from 1.0Å to 2.6Å, depending on the docking cavity. This confirms that gathering information from already deposited complexes in PDB and transforming it into the appropriate restraints benefits the docking process greatly.

Our final goal is to evolve this platform into a docking server where more rigorous, but also more computationally demanding methods, could be applied (e.g. Molecular Dynamics). Nonetheless, some additional points need to be revised. The first one is cavity detection and characterization. 3decision has proven to be a valuable tool for our pipeline being able to identify possible binding sites for the majority of targets. However, there are some cases where the 3decision protocol is not able to retrieve the correct pocket because they are shallow cavities or the receptor structure is ill-defined. In this first version of the pipeline, targets where there is no pocket information are neglected and no docking protocol is applied. For these situations, we could use a local implementation of fpocket [101] to check whether there are in fact no possible druggable cavities. Another option would be using Molecular Dynamics with aqueous/organic solvent mixtures (MDmix)[213,214] to identify possible binding sites. Nonetheless, we would like to add the option of taking the cavity coordinates as a reference. With this, we would separate the cavity-finding problem from the docking problem, reduce execution time and increase the predictive power when the binding site is already known.

A second point to revisit is the choice of receptor structure. As discussed, protein flexibility is an important aspect to consider in a drug discovery setup. Proteins can adapt their structure to the bound ligand, so using an apo structure or one in a complex with a very different compound degrades the performance of the docking program. One way to mitigate this effect would be to use different conformations of the receptor and select the one with the better score as the optimal structure [215].

A third aspect is the management of "third-party" molecules in the binding site, namely cofactors and water molecules. In this initial version of the pipeline, all systems are processed and prepared in the same way, stripping the binding site of all non-protein molecules. However, we detected several cases where docking failed owing to missing cofactor molecules that should be considered as part of the receptor. This can be solved with a curated list of cofactors that should

not be removed. Water molecules are frequently found at the protein-ligand interface, mediating hydrogen bonds between the partners. By keeping these structural waters on the binding site, the ligand pose predictions can be more accurate.

We will also continue to monitor the performance of restrained and unrestrained docking in prospective CELPP predictions. As previously shown, by using the MCSS score we are able to determine which is the docking method that performs best for each case. Initially, we applied a rather restrictive cutoff of 0,65, which included only 13% of the total cases. After considering all the participation cases, we were able to determine better ranges of application for each type of docking protocol, which presently is set to 0,5 and includes 31% of cases.

As far as the creation of the pharmacophores, in cases where, due to lack of pre-existing information, and ligand-based pharmacophores cannot be extracted we could make use of hot spots derived from the structure. Such hot spots can be identified by their ability to bind small organic co-solvents [3,5]. The afore-mentioned MDmix method not only identifies binding sites, but can also elucidate binding hot spots [213] that can be used as pharmacophoric restraints for docking. The addition of this methodology to our workflow would also allow us identify non-displaceable water molecules and re-assess the druggability of the pockets selected by 3decision.

All in all, from the participation in the CELPP challenge we highlighted some of the major challenges related to the prediction of protein-ligand complexes. The overall performance of docking in the CELPP challenge, with overall success rates of 20%, provides a sobering perspective of the state of the art in automated binding mode prediction. This is not strictly related to the rDock program, as the baseline performance with other software, as provided by the CELPP team, ranges between 20% and 30%. This is in stark contrast with the expectation (in this thesis and in the scientific literature) that docking can be an effective virtual screening tool. If docking cannot even predict how true ligands bind, how will it be able to distinguish such true ligands from non-binders? The solution to this apparent contradiction lies in

the fact that, for virtual screening applications, there is a large human intervention to select the most adequate conformation, define key interaction points to bias the scoring function, and eliminate unreasonable solutions. Equally important, docking is not the final selection criteria, but a very useful first step that enables various post-docking methods. All of this becomes evident when performing a docking-based VS, like the ones performed in Chapter 4.2, with the discovery of the first small molecules binding to RANK, and in Chapter 4.3, where we identified, characterized, and designed compounds binding to an allosteric site in PTEN.

In 2010 Gonzalez Suarez described the key role of RANK in the development of Breast Cancer [169,170]. Since then, her group and others have been studying the effectiveness of RANKL inhibitors like denosumab (used in the treatment of osteoporosis and bone metastasis) as a treatment for breast cancer. Unfortunately, this therapy has shown limited therapeutical effects in clinical trials with breast cancer patients.

Interestingly, Gonzalez Suarez has observed in mice that the inhibition of RANK signaling pathway is much more effective than acting upon RANKL [173]. RANK, like many other TNFR, presents two different activation pathways, one dependent on RANKL, and the other is based on the oligomerization of the receptor independently of RANKL (constitutive activation).

Due to the recently proven interest in RANK as a target for TNBC, in Chapter 4.2 I described the implementation of a VS campaign to find small molecules binding to RANK that not only can inhibit the RANK-RANKL interaction, but also avoid the constitutive activation of the receptor.

As we could attest during the participation in the CELPP Challenge, (discussion of chapter 4.1), the selection of the structure used for docking has to be chosen with care as it can have an outstanding impact on the results. In the case of RANK, there is the added challenge of not having a crystal structure for the human receptor, which forced us to create a homology model based on the mouse RANK receptor. Homology modeling is considered the most accurate computational

tool to determine the 3D structure [216]. However, the sequence similarity level between the template and the target sequence is an important factor to generate 3D structures with high accuracy [217]. The sequence similarity between the template and the target sequence for the RANK homology model was >80%, a value high enough to have a high accuracy model.

Because there are no other described small molecules that bind to the receptor, we applied MDmix to get a thorough druggability analysis and identify any putative cavity, and also to find hotspots that could be used as pharmacophoric restraints. From that analysis, we could not find any druggable cavity nor any combination of hotspots that could be used as pharmacophores. The PDB structure used to perform the homology model corresponded to the RANK-RANKL dimer, which presented a closed conformation of RANK, which is needed to properly interact with RANKL. As TNFR molecules present great domain flexibility, we hypothesized that a region comprising residues 109-117, which was close to the RANK-RANKL binding interface, was flexible enough to move leaving exposed a putative druggable cavity.

Although MDmix is an MD-based method, it is not adequate to assess any big conformational change of the receptor as the length of the MD is relatively short (20ns x 3replicas) and light restraints on the Heavy Atoms are applied during the simulation to avoid protein conformational sampling but ensuring convergence of the solvent exchange process [218]. Consequently, we performed a series of long unrestrained MD simulations, which confirmed this region's high degree of mobility when the receptor is not bound to RANKL. From the MD simulation, we selected a representative snapshot and repeated the MDmix analysis. This time, the structure presented a promising cavity with two hydrophobic and one polar hotspots that could be used as pharmacophoric restraints.

We selected a collection of ~7M compounds coming from different vendors and performed the VS campaign using the pharmacophoric restraints found with MDmix, an acceptor near N of Cys-82, a hydrophobic point near Trp-88, and another hydrophobic spot near

Leu-111. From this process, only 27 compounds passed the docking and undocking filters (see section 3.4.1.5). We must point out that, although the conformation extracted from the MD presented better druggability, the selected binding site is still really shallow and solvent-exposed which makes it difficult to find really strong binders. On the other hand, the fact that we are targeting an extracellular domain implies that eventual ligands would not need to permeate the cell membrane and could have physicochemical properties beyond the rule of five. Furthermore, the targeted protein conformation implies that eventual ligands could inhibit through non-competitive mechanism (e.g. conformational trapping), which is less demanding in terms of binding affinity.

We chose SPR to validate their interaction and quantify the binding affinity of the compounds. We confirmed binding for 10 compounds with KD values ranging from 90 μM to 7 mM. Notwithstanding that the affinity values are on the weaker side, these are the first reported small molecule binders for the RANK receptor. From here on, some efforts have already started to optimize the potency of the compounds, by trying to grow the compounds so they can reach other interesting hotspots found during the druggability analysis with MDmix. It is important to note that the domain responsible for the oligomerization and subsequent activation of the RANKL-independent mechanism (PLAD domain), is located in the intracellular region of the receptor, making really difficult to predict if the ligands will be able to reduce the constitutive activity.

Currently, assays in cancer cell lines are being carried out in the CNIO to test the ability of the compounds to block both activation pathways, with promising preliminary results for some of the compounds.

In the RANK project, we have adopted the traditional pathway of following up the virtual screening hits with a biophysical or biochemical assay, then proceeding to cell-based assays. While this makes sense for drug discovery projects that pursue the traditional mechanism of action of competitive inhibition, it may be counterproductive when dealing with a non-standard mechanism of action. Instead, it may be more

effective to look for functional effects in relevant biological systems. To illustrate this point, in chapter 4.3 we have set the objective of discovering allosteric modulators of a protein with key regulatory function, but without making any pre-assumption about the consequence of such allosteric binding. We have focused our attention on tumor suppressor proteins because they play a central regulatory role and there is a lack of chemical compounds that specifically modulate their function.

Our selection procedure led us to PTEN. PTEN regulation is really complex, and even a small decrease in its levels or activity increases the risk of tumor progression and development [193]. Additionally, PTEN is also involved in restraining several cellular regeneration processes [192]. In the nervous system, where tissue (re)growth is limited, PTEN inhibition has been shown to promote axon regeneration after crush injuries in both optical and spinal neurons [219–221]. Additionally, PTEN deletion has benefits in cardiomyocyte survival by preventing ischemia and limiting reperfusion [222]. PTEN dose reduction may also have an application in the context of Alzheimer's disease. The inhibition of PTEN at the synapses affected by β-amyloid aggregation leads to reductions in cognitive deficiencies [223]. In this regard, reducing the functional dose of PTEN is risky, as its prominent role as tumor suppressor has to be considered. That is why PTEN deletion has raised some concerns over its therapeutic suitability [219,224–226]. To address these issues, potent and selective inhibitors are required that allow selective, short-term PTEN inhibition. Unfortunately, the compounds available today are not selective as they have been shown to have an effect over other phosphatases, leading to unwanted side effects [198]. In this case, allosteric regulation of the protein would solve some of the issues regarding selectivity as allosteric sites are less conserved than orthosteric sites. Other targeting paradigms are also being considered like targeted-RNA degradation (RIBOTAC) of miR-21 (a regulator of the expression of PTEN and other proteins) [227,228]. In summary, modulation of PTEN, either activating or inhibiting some of its functions, offers potential therapeutic opportunities, and there is a lack

of specific modulators of this protein that could be used as chemical probes.

From the cavity search and druggability analysis performed with MDmix, a putative allosteric site was identified at the hinge region between the C2 and Phosphatase domains. The hotspot with the best energy value for this region corresponded to a polar hotspot interacting with the hydroxyl of Tyr-164. In a later study, Nira Smith et al employing MD simulations and network proximity assays, studied the conformational dynamics of PTEN germline mutations associated with cancer and autism. This study highlights the inter-domain region, the same region that we identified with MDmix, as a crucial region that participates in both the stability and the overall dynamics of the protein, having some key residues working as key functional centers that might govern long-range allosteric regulation [229].

Initially, we performed the VS using the same compound collection as we used in chapter 4.2. Strikingly, only 24 compounds showed WQB values larger than 4 kcal/mol, which is a relatively low value for drug-like molecules. However, we noticed that the majority of molecules showing good WQB values in DUck contained a substituted tetrazole making the interaction with Tyr-177, which can indicate that the ligands with this structure are preferred in this cavity.

There is some evidence from the literature that, in ligand-based virtual screening at least, increasing the size of the search space does lead to an increase in the number of hits [230]. We decided to search in a bigger chemical space like ENAMINE Real Db for compounds having the substituted tetrazole. Following this strategy and applying the same VS protocol to the new set of compounds, we were able to identify ligands with better WQB values. At the end of the VS, 14 compounds were selected to be tested with the phenotypic approach.

Initially, we considered a cell growth assay as screening method, but later found out that cell growth was equivalent in WT, heterozigous and PTEN KO cells. We noted that the cell lines used (HCT116) have mutations in KRAS and PI3K genes, which cause overactivation of the pathways related to cell proliferation. We hypothesized that the effect

of these mutations in synergy with the growth factors present in the FBS could be overcoming the lower dose (hterozygous cell line) and even the absence (KO cell line) of PTEN. We then repeated the experiment with 0,1% FBS, but even in the absence of growth factors, all cell lines were growing at the same rate. However, in the course of this assay, we noted a striking morphological change in the heterozygous cells upon treatment with some of our compounds. For CMP1 the morphological changes could be attributed to an EMT process, previously linked to PTEN inhibition [205–208]. Indeed, when treated with CMP1 the levels of E-cadherin, responsible for the establishment and maintenance of epithelial cell morphology, decreased and the levels of Vimentin, a class-III intermediate filament found in non-epithelial cells, increased. For CMP1.3 we detected another type of morphological change also closely related to PTEN inhibition. In this case, we observed an enlargement of the cell and the presence of more than one nuclei [209,231].

An important reason to avoid biophysical assays as primary screening is that they require good quality protein, which is often difficult to obtain. Exceptionally, the protein PTEN can be readily purchased, thus we were able to test the binding of all the 14 selected compounds through SPR, to obtain an orthogonal readout that will be compared to the phenotypic effect on the cell-based assays. Out of the 14 compounds, we detected binding for 8 of them, with KD ranging from 18 μM to 160 μM. Remarkably, CMP1.3 had the lowest KD value (18 μM) and CMP1 ranked in the top 4. Non-binders – as expected – did not show any biological effect. However, the SPR-derived Kd was not a good predictor of phenotypic outcome. This suggests that, if it is practical to perform a binding assay a primary screening, it could be used as a first-pass filter, but it would be advisable to test all compounds in the phenotypic assay.

As a second comparator, we have assessed the compounds in a biochemical assay. As PTEN's best-known function is its phosphatase activity, it would be tempting to use such biochemical assay as a primary screening. However, we find that only one compounds (CMP14) has

any effect on this assay, but this activity does not translate into any observable phenotypic consequence. Even more interestingly, CMP1 and CMP1.3, which showed binding in SPR, did not affect the phosphatase activity of PTEN. The fact that with SPR we had a clear signal but the activity of PTEN is not disrupted, confirms that we are not targeting the active site and suggests that we are regulating PTEN through an allosteric mechanism independent of the phosphatase activity. But, of course, we will need to assess the enzymatic activity in the cell context, considering the multitude of PTEN substrates, not just PIP3. This will be addressed in future work.

Looking retrospectively, we have seen that the combination of target-based and phenotypic screening offers an advantage when working with novel MOA for challenging targets with intricate regulations like PTEN. The TDD approach has allowed us to select a small number of drug-like ligands (14), which facilitates the thorough assessment of the effect of the compounds on the cell lines. Additionally, we have attested that when dealing with a novel allosteric site, the effects of the compounds are almost unpredictable. As an example, two analogues like CMP1 and CMP1.3, which share the same binding motive to the receptor, exert different phenotypic effects.

We have also seen that, apart from the selection of the receptor and the druggability study needed before performing a VS, the other paramount player is the molecular database used. For RANK and PTEN we followed a similar approach, we exhaustively searched in moderately large chemical databases for ligands with drug-like properties. However, when targeting PTEN, we have seen that exploring bigger screening collections, if they maintain a high diversity, increases the probability of finding better hits.

These collections were initially treated with skepticism because chemical synthesis outcomes are uncertain and people assumed that the synthetic success rate would be low and delivery would take a long time. Nowadays, ENAMINE has a set of well-established and optimized reactions that enables a fast synthesis with a success rate of >80% and guaranteed purity of >90%. Today, ENAMINE contains 31 billion

compounds and is expected to reach the trillion scale (1012) in a couple of years, growing towards the theoretically accessible chemical space [232].

Despite the potential of these chemical collections, with increasingly bigger libraries, the computational time and cost of an exhaustive exploration become unfeasible. Standard methods require vast computational resources that scale linearly with the growing number of compounds. For these reasons, it is still necessary to develop HTVS protocols capable of navigating these massive collections.

In Chapter 4.4 we describe the development of a novel strategy that explores the chemical universe from the bottom up. By first performing a systematic exploration of the low molecular weight chemical space, it allows us to rapidly focus on the most privileged areas of the chemical space to be further explored following an FBDD approach. We validated our protocol by prospectively finding novel BRD4 inhibitors.

As FBDD success depends largely on the quality of the initial hit, we decided to explore three different starting points I) from known drug scaffold II) experimentally validated fragments III) computational fragment hit obtained from an exhaustive screening.

FS collections are usually designed to provide uniform coverage of the chemical space [233], giving great importance to a diversity-based design of the collection. However, these collections have a typical size of 10^3 which only represents a sample of the 17 million fragments that can be synthesized on demand[or the 10^9 theoretically possible fragments [234]. To be able to exhaustively explore the fragment space, we extracted the fragment-sized molecules containing less than 14 heavy atoms and at least one ring from ENAMINE REAL Db and ZINC20. Thus obtaining an FS collection of 4.1 million unique fragments.

Virtual FS was carried out by docking, biased by pharmacophoric interaction points derived from the natural substrate preferences. Besides, as we have seen with the participation in the CELPP Challenge, often gathering previous information about your system can yield better

results [235,236]. We selected a set of active fragments in CHEMBL to have an idea which are the values of SCORE.INTER, ΔG_{bind} and WQB that we could expect.

After preparing the compounds, HTVS with rDock was performed, using the energy values obtained from the CHEMBL active compounds. Thanks to the work done by a previous colleague, we know that a hydrophobic hot spot is preferred near the structural network of waters at the bottom of the BRD4 cavity [131]. However, rDock counts the presence of waters as a possible source for H-bonds, thus positioning polar groups or positively charged substituents close to the waters. For that reason, all these molecules were filtered out.

Nevertheless, docking of fragment molecules is often regarded as particularly challenging for two main reasons: first, fragments may be more promiscuous in their binding modes than larger "drug-like" molecules [237,238]; and second, docking scoring functions are inaccurate even for large molecules, and are likely to be still less accurate for fragments [239–241]. Thus, Docking is not a reliable method to prioritize fragments. For that reason, and in contrast to other published HTVS campaigns [62,242,243], we apply a hierarchy of increasingly sophisticated computational methods, which maximizes the success probability of the selected compounds. The resulting docking hits were clustered using the CC fingerprints and using a k-means algorithm with 2.000 clusters. Dynamic Undocking and MMGBSA are quick and powerful tools to enhance the results from docking. Specially DUck, as we have demonstrated its potential in discriminating False Positives [86]. A final set of 6 compounds passed all the filters and were selected as candidates for fragment evolution.

As for the Fragment Growing Stage, we used SpaceMACS to perform substructure searches in ENAMINA REALspace so to obtain scaffold-focused libraries of around 10M to 20M compounds for each initial fragment. Scaffold searches took an average of 10 hours (using 32 CPU multithread node). Benchmarks done by changing the library size, scaffold and parameters showed that computational cost (as well as memory requirements) scaled with the output library size more than the

database size. Thus, when the combinatorial databases inevitably grow, the performance of this crucial step will remain stable.

This, however, arises a concern on the chemical diversity of the combinatorial databases have. As all compounds are generated from a limited amount of building blocks, molecules are bound to be similar. Several studies have shown that the overlap between different combinatorial databases is minimal [232], consequently, it would be advisable to do the substructure search in more than one database. Finally, not all compounds from the database were optimal for drug discovery. PAINS and druglike property filtering was still needed before docking.

To validate the compounds we selected two orthogonal biophysical assays, DSF and TR-FRET. In DSF a total of 48 compounds showed a change in ΔTm higher than 1ºC, which corresponds to a hit rate of 47%. In TR-FRET we obtained 11 compounds that had an IC50 between 1-100nM. Interestingly, there aren't any trends pointing at drug derivate scaffolds having a higher hit ratio than any computational scaffold. Even, in TR-FRET from the 11 compounds that showed an IC50 between 1-100nM, placing them in a similar range than other BET inhibitors that are in clinical phases (JQ1 with 10,7nM, IBET 151 with 20-100nM and ABBV-075 with 1-2,2nM) [244]. We also assessed the chemical diversity of the 11 compounds by comparing them with known BRD4 binders from CHEMBL and also a random set of 50.000 molecules. All the compounds were distributed across the chemical space, and not only in areas enriched with known BRD4 binders. This indicates that we were able to explore a more diverse chemical space, not biased to the "BRD4 Chemical Space".

Finally, we have initiated a collaboration with the group of Maria Garcia-Alai in the EMBL to perform crystallization with the active compounds.

With this project we showed that the massive chemical libraries offer opportunities for finding highly active compounds. Additionally, we have demonstrated that a bottom-up approach enables us to explore the chemical space with minimal computational resources. This approach

could open the path for computational campaigns to aim for lead-like affinities directly in the hit discovery phase. Allowing to reduce significantly the efforts and shift the attention towards optimization of other important properties (e.g toxicity, availability).

6 Conclusions

6.1 GENERAL CONCLUSIONS

This thesis described the application of novel methodologies in Virtual Screening with the aim of discovering bioactive molecules better, more effectively, even against challenging targets and novel mechanisms of action.

6.2 SPECIFIC CONCLUSIONS

1. We developed an automatic pipeline for participation on the CELPP Challenge that makes use of knowledge-based restraints to improve the docking predictions. The pipeline is able to generate predictions for most of the proposed targets as well as obtain poses with low RMSD values when compared to the crystal structure. Besides, our pipeline highlights some major challenges in the automatic prediction of protein-ligand complexes that need to be carefully considered in SBDD.
2. We described the first small molecule inhibitors targeting RANK protein. Preliminary results indicate that the compounds are able to block the RANKL-dependent and the constitutive activation of the RANK signalling pathway.
3. We described a protocol that combines ultrahigh-throughput Virtual Screening with low-throughput high-content assays. We tested the approach with PTEN, which led as to describing the first allosteric modulators for this protein.
4. We developed an algorithm to efficiently explore ultra-large chemical collections. We show that the bottom-up exploration of the chemical space is an efficient approach towards finding potent hits in combinatorial libraries. We identified BRD4 inhibitors with potencies comparable to advanced drug candidates such as JQ1.

BIBLIOGRAPHY

1. Gail A. Van Norman. Drugs, Devices, and the FDA: Part 1 An Overview of Approval Processes for Drugs. JACC: Basic to Translational Sciences. 2016;1(3):170–9.
2. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May 1;47:20–33.
3. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. Vol. 323, JAMA - Journal of the American Medical Association. American Medical Association; 2020. p. 844–53.
4. Clinical development success rates and contributing factors (2011-2020) [Internet]. [cited 2022 Jul 1]. Available from: https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020
5. Collen MF. The Origins of Informatics. Journal of the American Medical Informatics Association. 1994;1(2):91–107.
6. Drie JH. Computer-aided drug design: The next 20 years. J Comput Aided Mol Des. 2007 Oct;21(10–11):591–601.
7. Prieto-Martínez FD, López-López E, Eurídice Juárez-Mercado K, Medina-Franco JL. Computational Drug Design Methods—Current and Future Perspectives. In: In Silico Drug Design. Elsevier; 2019. p. 19–44.
8. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. Vol. 14, Drug Discovery Today. 2009. p. 147–54.
9. Agamah FE, Mazandu GK, Hassan R, Bope CD, Thomford NE, Ghansah A, Chimusa ER. Computational/in silico methods in drug target and lead prediction. Brief Bioinform. 2020 Sep 1;21(5):1663–75.
10. Dr. Siju E. N.* DrGRRAPPFT, DPP, DrNH and KR. CADD: Pharmacological Approaches in Drug Discovery. World Journal of Pharmacy and Pharmaceutical Science. 2017;6:892–908.
11. Hoffer L, Renaud JP, Horvath D. Fragment-Based Drug Design: Computational and Experimental State of the Art.

Vol. 14, Combinatorial Chemistry & High Throughput Screening. 2011.
12. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'Min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR modeling: Where have you been? Where are you going to? Vol. 57, Journal of Medicinal Chemistry. American Chemical Society; 2014. p. 4977–5010.
13. Wu D, Zheng X, Liu R, Li Z, Jiang Z, Zhou Q, Huang Y, Wu XN, Zhang C, Huang YY, Luo H bin. Free energy perturbation (FEP)-guided scaffold hopping. Acta Pharm Sin B. 2022 Mar 1;12(3):1351–62.
14. Jiménez-Luna J, Pérez-Benito L, Martínez-Rosell G, Sciabola S, Torella R, Tresadern G, de Fabritiis G. DeltaDelta neural networks for lead optimization of small molecule potency. Chem Sci. 2019;10(47):10911–8.
15. Erikawa D, Yasuo N, Sekijima M. MERMAID: an open source automated hit-to-lead method based on deep reinforcement learning. J Cheminform. 2021 Dec 1;13(1).
16. Green H, Durrant JD. DeepFrag: An Open-Source Browser App for Deep-Learning Lead Optimization. J Chem Inf Model. 2021 Jun 28;61(6):2523–9.
17. Hann MM, Leach AR, Harper G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. J Chem Inf Comput Sci. 2001;41(3):856–64.
18. Du X, Li Y, Xia YL, Ai SM, Liang J, Sang P, Ji XL, Liu SQ. Insights into protein–ligand interactions: Mechanisms, models, and methods. Vol. 17, International Journal of Molecular Sciences. MDPI AG; 2016.
19. Kuhn B, Fuchs JE, Reutlinger M, Stahl M, Taylor NR. Rationalizing tight ligand binding through cooperative interaction networks. J Chem Inf Model. 2011;51(12):3180–98.
20. Ferreira De Freitas R, Schapira M. A systematic analysis of atomic protein-ligand interactions in the PDB. Medchemcomm. 2017;8(10):1970–81.
21. Bissantz C, Kuhn B, Stahl M. A medicinal chemist's guide to molecular interactions. J Med Chem. 2010;53(14):5061–84.

22. Nittinger E, Inhester T, Bietz S, Meyder A, Schomburg KT, Lange G, Klein R, Rarey M. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces. J Med Chem. 2017 May 25;60(10):4245–57.
23. Steiner T. The Hydrogen Bond in the Solid State. Angewante Chemie. 2002;41:48–76.
24. Schmidtke P, Javier Luque F, Murray JB, Barril X. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. J Am Chem Soc. 2011 Nov 23;133(46):18903–10.
25. Matthews RP, Welton T, Hunt PA. Competitive pi interactions and hydrogen bonding within imidazolium ionic liquids. Physical Chemistry Chemical Physics. 2014 Feb 21;16(7):3238–53.
26. Vargas R, Garza J, Dixon DA, Hay BP. How strong is the C(α)-H···O=C hydrogen bond? J Am Chem Soc. 2000 May 17;122(19):4750–5.
27. Aravinda S, Shamala N, Bandyopadhyay A, Balaram P. Probing the Role of the C-H···O Hydrogen Bond Stabilized Polypeptide Chain Reversal at the C-terminus of Designed Peptide Helices. Structural Characterization of Three Decapeptides. J Am Chem Soc. 2003 Dec 10;125(49):15065–75.
28. Mandel-Gutfreund Y, Margalit H, Jernigan RL, Zhurkin VB. A Role for CH...O Interactions in Protein - DNA Recognition.
29. Horowitz S, Trievel RC. Carbon-oxygen hydrogen bonding in biological structure and function. Vol. 287, Journal of Biological Chemistry. 2012. p. 41576–82.
30. Pierce AC, Sandretto KL, Bemis GW. Kinase inhibitors and the case for CH···O hydrogen bonds in protein-ligand binding. Proteins: Structure, Function and Genetics. 2002 Dec 1;49(4):567–76.
31. Hendsch ZS, Tidor B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. Vol. 3, Protein Science. Cambridge University Press; 1994.
32. Carey DW, Joel F Schildbach, Robert T. Sauer. Are buried salt bridges important for protein stability and

conformational specificity? Nat Struct Biol. 1995;2(2):122–8.
33. Fischer E. Influence of Configuration on the Action of Enzymes. J Am Chem Soc. 1894;3:2985–93.
34. Jennifer Lippincott-Schwartz, Erik Snapp, Anne Kenworthy. Studying protein dynamics in living cells. Nature Reviews molecular Cell Biology . 2001;2:444–56.
35. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. PNAS [Internet]. 1958;44(2):98–104. Available from: https://www.pnas.org
36. Teague SJ. Implications of protein flexibility for drug discovery. Vol. 2, Nature Reviews Drug Discovery. European Association for Cardio-Thoracic Surgery; 2003. p. 527–41.
37. Bahar I, Chennubhotla C, Tobi D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. Vol. 17, Current Opinion in Structural Biology. 2007. p. 633–40.
38. Wolynes &, Karplus & Shakhno-Vitch ; Karplus ; Folding funnels, binding funnels, and protein function. Lazaridis & Karplus. Gruebele & Wolynes; 1989.
39. Nussinov R, Tsai CJ. The Different Ways through Which Specificity Works in Orthosteric and Allosteric Drugs. Curr Pharm Des. 2012 Feb 20;18(9):1311–6.
40. Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. Vol. 153, Cell. Elsevier B.V.; 2013. p. 293–305.
41. Hilser VJ, Wrabl JO, Motlagh HN. Structural and energetic basis of allostery. Vol. 41, Annual Review of Biophysics. 2012. p. 585–609.
42. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. Vol. 508, Nature. Nature Publishing Group; 2014. p. 331–9.
43. Sadowsky JD, Burlingame MA, Wolan DW, Mcclendon CL, Jacobson MP, Wells JA. Turning a protein kinase on or off from a single allosteric site via disulfide trapping. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1102376108
44. Smith NJ, Milligan G. Allostery at G protein-coupled receptor homo- and heteromers: Uncharted pharmacological

landscapes. Vol. 62, Pharmacological Reviews. 2010. p. 701–25.
45. Fischer G, Rossmann M, Hyvönen M. Alternative modulation of protein-protein interactions by small molecules. Vol. 35, Current Opinion in Biotechnology. Elsevier Ltd; 2015. p. 78–85.
46. Mabonga L, Kappo AP. Protein-protein interaction modulators: advances, successes and remaining challenges. Vol. 11, Biophysical Reviews. Springer Verlag; 2019. p. 559–81.
47. Raj M, Bullock BN, Arora PS. Plucking the high hanging fruit: A systematic approach for targeting protein-protein interactions. Bioorg Med Chem. 2013 Jul 15;21(14):4051–7.
48. Ran X, Gestwicki JE. Inhibitors of protein–protein interactions (PPIs): an analysis of scaffold choices and buried surface area. Vol. 44, Current Opinion in Chemical Biology. Elsevier Ltd; 2018. p. 75–86.
49. Oltersdorf T, Elmore SW, Shoemaker AR, Armstrong RC, Augeri DJ, Belli BA, Bruncko M, Deckwerth TL, Dinges J, Hajduk PJ, Joseph MK, Kitada S, Korsmeyer SJ, Kunzer AR, Letai A, Li C, Mitten MJ, Nettesheim DG, Ng SC, Nimmer PM, O'Connor JM, Oleksijew A, Petros AM, Reed JC, Shen W, Tahir SK, Thompson CB, Tomaselli KJ, Wang B, Wendt MD, Zhang H, Fesik SW, Rosenberg SH. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature. 2005 Jun 2;435(7042):677–81.
50. Cukuroglu E, Engin HB, Gursoy A, Keskin O. Hot spots in protein-protein interfaces: Towards drug discovery. Prog Biophys Mol Biol. 2014;116(2–3):165–73.
51. Díaz JCL, del Castillo JC, Rodriguez-López EA, Alméciga-Díaz CJ. Advances in the development of pharmacological chaperones for the mucopolysaccharidoses. Vol. 21, International Journal of Molecular Sciences. MDPI AG; 2020.
52. Sun X, Gao H, Yang Y, He M, Wu Y, Song Y, Tong Y, Rao Y. Protacs: Great opportunities for academia and industry. Vol. 4, Signal Transduction and Targeted Therapy. Springer Nature; 2019.

53. Maia EHB, Assis LC, de Oliveira TA, da Silva AM, Taranto AG. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. Vol. 8, Frontiers in Chemistry. Frontiers Media S.A.; 2020.
54. MacArron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DVS, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS. Impact of high-throughput screening in biomedical research. Vol. 10, Nature Reviews Drug Discovery. 2011. p. 188–95.
55. Volochnyuk DM, Ryabukhin S v., Moroz YS, Savych O, Chuprina A, Horvath D, Zabolotna Y, Varnek A, Judd DB. Evolution of commercially available compounds for HTS. Vol. 24, Drug Discovery Today. Elsevier Ltd; 2019. p. 390–402.
56. Janzen WP. Screening technologies for small molecule discovery: The state of the art. Vol. 21, Chemistry and Biology. Elsevier Ltd; 2014. p. 1162–70.
57. Horvath D. A Virtual Screening Approach Applied to the Search for Trypanothione Reductase Inhibitors [Internet]. 1997. Available from: https://pubs.acs.org/sharingguidelines
58. Meng XY, Zhang HX, Mezei M, Cui M. Molecular Docking: A powerful approach for structure-based drug discovery. Current Opinion Computer Aided Drug Design. 2011;7(2):146–57.
59. Luttens A, Gullberg H, Abdurakhmanov E, Vo DD, Akaberi D, Talibov VO, Nekhotiaeva N, Vangeel L, de Jonghe S, Jochmans D, Krambrich J, Tas A, Lundgren B, Gravenfors Y, Craig AJ, Atilaw Y, Sandström A, Moodie LWK, Lundkvist Å, van Hemert MJ, Neyts J, Lennerstrand J, Kihlberg J, Sandberg K, Danielson UH, Carlsson J. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. J Am Chem Soc. 2022 Feb 23;144(7):2905–20.
60. Guo S, Xie H, Lei Y, Liu B, Zhang L, Xu Y, Zuo Z. Discovery of novel inhibitors against main protease (Mpro) of SARS-CoV-2 via virtual screening and biochemical evaluation. Bioorg Chem. 2021 May 1;110.

61. Kanhed AM, Patel D v., Teli DM, Patel NR, Chhabria MT, Yadav MR. Identification of potential Mpro inhibitors for the treatment of COVID-19 by using systematic virtual screening approach. Mol Divers. 2021 Feb 1;25(1):383–401.
62. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Algaa E, Tolmachova K, Tolmachev AA, Shoichet BK, Roth BL, Irwin JJ. Ultra-large library docking for discovering new chemotypes. Nature. 2019 Feb 14;566(7743):224–9.
63. Barril X. Computer-aided drug design: time to play with novel chemical matter. Vol. 12, Expert Opinion on Drug Discovery. Taylor and Francis Ltd; 2017. p. 977–80.
64. Knehans T and KFM and KH and SH and HA and RF and EJ and LC and KM. Merck AcceSSible InVentory (MASSIV): In silico synthesis guided by chemical transforms obtained through bootstrapping reaction databases. Abstracts of Papers of the American Chemical Society. 2017;254.
65. Hu Q, Peng Z, Sutton SC, Na J, Kostrowicki J, Yang B, Thacher T, Kong X, Mattaparti S, Zhou JZ, Gonzalez J, Ramirez-Weinhouse M, Kuki A. Pfizer global virtual library (PGVL): A chemistry design tool powered by experimentally validated parallel synthesis information. ACS Comb Sci. 2012 Nov 12;14(11):579–89.
66. Lipinski CA, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Vol. 23, Advanced Drug Delivery Reviews. 1997.
67. Congreve M, Carr R, Murray C, Jhoti H. A 'Rule of Three' for fragment-based lead discovery? Drug Discov Today. 2003 Oct 1;8(19):876–7.
68. Lamoree B, Hubbard RE. Current perspectives in fragment-based lead discovery (FBLD). Vol. 61, Essays in Biochemistry. Portland Press Ltd; 2017. p. 453–64.
69. Ahmad E, Rabbani G, Zaidi N, Khan MA, Qadeer A, Ishtikhar M, Singh S, Khan RH. Revisiting ligand-induced conformational changes in proteins: Essence, advancements, implications and future challenges. Vol. 31, Journal of Biomolecular Structure and Dynamics. 2013. p. 630–48.

70. Gutteridge A, Thornton J. Conformational changes observed in enzyme crystal structures upon substrate binding. J Mol Biol. 2005 Feb 11;346(1):21–8.
71. McCammon JAGBR. Dynamics of folded proteins. Nature. 1977;267:585–90.
72. Durrant Jacob D, McCammon J Andrew. Molecular dynamics simulations and drug discovery. BMC Biology . 2011;
73. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. J Comput Chem. 2009 Jul 30;30(10):1545–614.
74. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. Vol. 26, Journal of Computational Chemistry. 2005. p. 1668–88.
75. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation PROGRAM SUMMARY Title of program: GROMACS version 1.0 [Internet]. Vol. 91, Computer Physics Communications. 1995. Available from: http://rugmd0,chela,rug.nl/'gmx/gmx,egi
76. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol. 2017 Jul 1;13(7).
77. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem SOC. 1995;117:5179–97.

78. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and Testing of a General Amber Force Field. J Comput Chem. 2004;25:1157–74.
79. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010 Mar;31(4):671–90.
80. Christen M, Hünenberger PH, Bakowies D, Baron R, Bürgi R, Geerke DP, Heinz TN, Kastenholz MA, Kräutler V, Oostenbrink C, Peter C, Trzesniak D, van Gunsteren WF. The GROMOS software for biomolecular simulation: GROMOS05. Vol. 26, Journal of Computational Chemistry. 2005. p. 1719–51.
81. Beveridge DL, Jorgensen WL. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society, . 1988;110(6):1657–66.
82. Beauchamp KA, Lin YS, Das R, Pande VS. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. J Chem Theory Comput. 2012 Apr 10;8(4):1409–14.
83. Mattos C, Bellamacina CR, Peisach E, Pereira A, Vitkup D, Petsko GA, Ringe D. Multiple solvent crystal structures: Probing binding sites, plasticity and hydration. J Mol Biol. 2006 Apr 14;357(5):1471–82.
84. Alvarez-Garcia D, Schmidtke P, Cubero E, Barril X. Extracting Atomic Contributions to Binding Free Energy using Molecular Dynamics Simulations with Mixed Solvents (MDmix). Curr Drug Discov Technol. 2021 Dec 24;19(2).
85. Alvarez-garcia D, Barril X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. 2014;
86. Ruiz-Carmona S, Schmidtke P, Luque FJ, Baker L, Matassova N, Davis B, Roughley S, Murray J, Hubbard R, Barril X. Dynamic undocking and the quasi-bound state as

tools for drug discovery. Nat Chem [Internet]. 2016;9(3):201–6. Available from: http://www.nature.com/doifinder/10.1038/nchem.2660
87. Majewski M, Barril X. Structural Stability Predicts the Binding Mode of Protein-Ligand Complexes. J Chem Inf Model. 2020;60(3):1644–51.
88. Yuriev E, Holien J, Ramsland PA. Improvements, trends, and new ideas in molecular docking: 2012-2013 in review. Journal of Molecular Recognition. 2015;28(10):581–604.
89. Śledź Pawełand Caflisch A. Protein structure-based drug design: from docking to molecular dynamics. Curr Opin Struct Biol. 2018;48:93–102.
90. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. Vol. 16, Journal of Computer-Aided Molecular Design. 2002.
91. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. J Med Chem. 1999 Mar 11;42(5):791–804.
92. Eldridge MD, Murray CW, Auton TR, Paolini G v, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. Vol. 11, Journal of Computer-Aided Molecular Design. 1997.
93. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, Barril X, Hubbard RE, Morley SD. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol. 2014;10(4):e1003571.
94. Oshiro CM, Kuntz ID, Dixon JS. Flexible ligand docking using a genetic algorithm. Vol. 9, Journal of Computer-Aided Molecular Design. 1995.
95. Tirado-Rives J, Jorgensen WL. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. J Med Chem. 2006;49(20):5880–4.
96. Schrödinger LLC. Small-Molecule Drug Discovery Suit 2018-1. New York, NY. 2018;
97. Chemical Computin Group. MOE. 1010 Sherbooke St. West, Suite# 910; 2016.

98. Labute P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins: Structure, Function and Bioinformatics. 2009;75(1):187–205.
99. Schrödinger LLC. LigPrep, version 2.5. New York, NY. 2011;
100. Eric Le Roux. 3Decision. Discngine;
101. le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10(1):168.
102. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An Open chemical toolbox. J Cheminform. 2011;3(10):1–14.
103. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform [Internet]. 2015;7(1):1–13. Available from: http://dx.doi.org/10.1186/s13321-015-0069-3
104. Landrum G. Rdkit: Open-source cheminformatics software. URL http://www rdkit org/, https://github com/rdkit/rdkit. 2016;
105. Bostro J, Hogner A, Schmitt S. Do Structurally Similar Ligands Bind in a Similar Fashion? 2006;6716–25.
106. Yang SY. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. Drug Discov Today [Internet]. 2010;15(11–12):444–50. Available from: http://dx.doi.org/10.1016/j.drudis.2010.03.013
107. Taminau J, Thijs G, de Winter H. Pharao: Pharmacophore alignment and optimization. J Mol Graph Model. 2008;27(2):161–9.
108. Liu C, Walter TS, Huang P, Zhang S, Zhu X, Wu Y, Wedderburn LR, Tang P, Owens RJ, Stuart DI, Ren J, Gao B. Structural and Functional Insights of RANKL–RANK Interaction and Signaling. The Journal of Immunology. 2010 Jun 15;184(12):6910–9.
109. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015 Jul 7;11(8):3696–713.

110. Mark P, Nilsson L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. Journal of Physical Chemistry A. 2001 Nov 1;105(43):9954–60.
111. Kräutler V, van Gunsteren WF, Hünenberger PH. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. Vol. 22, J Comput Chem. 2001.
112. Arcon JP, Defelipe LA, Modenutti CP, López ED, Alvarez-Garcia D, Barril X, Turjanski AG, Martí MA. Molecular Dynamics in Mixed Solvents Reveals Protein-Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. J Chem Inf Model. 2017 Apr 24;57(4):846–63.
113. Schmidtke P, Bidon-chanal A, Luque FJ, Barril X. MDpocket : open-source cavity detection and characterization on molecular dynamics trajectories. 2011;27(23):3276–85.
114. Manual U. LigPrep 2.3.
115. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins: Structure, Function and Bioinformatics. 2010 Jun;78(8):1950–8.
116. Bayly C, McKay Daniel. An Informal AMBER Small Molecule Force Field : parm@Frosst [Internet]. 2010 [cited 2022 Sep 15]. Available from: http://www.ccl.net/cca/data/parm_at_Frosst/
117. Harvard Medical School. Surface Plasmon Ressonance [Internet]. [cited 2022 Sep 15]. Available from: https://cmi.hms.harvard.edu/surface-plasmon-resonance
118. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ. Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. Cell [Internet]. 2013;155(4):948–62. Available from: http://dx.doi.org/10.1016/j.cell.2013.10.011
119. Eswaran J, Debreczeni JÉ, Longman E, Barr AJ, Knapp S. The crystal structure of human receptor protein tyrosine phosphatase κ phosphatase domain 1. Protein Science. 2006 Jun;15(6):1500–5.

120. Jie-Oh Lee, Haijuan Yang, Maria-Magdalena Georgescu, Antonio Di Cristofano, Tomohiko Maehama, Yigong Shi, Jack E. Dixon, Pier Pandolfi, Nikola P. Pavletich. Crystal Structure of the PTEN Tumor Suppressor:Implications for Its Phosphoinositide PhosphataseActivity and Membrane Association. Cell. 1999;99:323–34.
121. Singh M, Popowicz GM, Krajewski M, Holak TA. Structural ramification for acetyl-lysine recognition by the bromodomain of human BRG1 protein, a central ATPase of the SWI/SNF remodeling complex. ChemBioChem. 2007 Jul 23;8(11):1308–16.
122. Shi A, Murai MJ, He S, Lund G, Hartley T, Purohit T, Reddy G, Chruszcz M, Grembecka J, Cierpicki T. Structural insights into inhibition of the bivalent menin-MLL interaction by small molecules in leukemia. In: Blood. American Society of Hematology; 2012. p. 4461–9.
123. Labute P. Protonate 3D: assignment of macromolecular protonation state and geometry. 2008;
124. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002 Nov;42(6):1273–80.
125. Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J Comput Chem. 2000 Jan 30;21(2):132–46.
126. Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem. 2002 Dec;23(16):1623–41.
127. Johanna Lee and Mariel Mohns. How to Choose a Cell Viability or Cytotoxicity Assay [Internet]. Promega Corporation. 2019 [cited 2022 Sep 15]. Available from: https://www.promega.es/resources/guides/cell-biology/cell-viability/#introduction-to-cell-viability-assays-196d8754-2cfd-4b5b-90ce-abe03b005b26
128. chemAxon. cxcalc, version 20.21.0 [Internet]. Available from: http://www.chemaxon.com
129. CORINA.3D Structure Generator, Version 4.4.0 [Internet]. Available from: https://mn-am.com/products/corina/

130. Gehling VS, Hewitt MC, Vaswani RG, Leblanc Y, Coîté A, Nasveschuk CG, Taylor AM, Harmange JC, Audia JE, Pardo E, Joshi S, Sandy P, Mertz JA, Sims RJ, Bergeron L, Bryant BM, Bellon S, Poy F, Jayaram H, Sankaranarayanan R, Yellapantula S, Bangalore Srinivasamurthy N, Birudukota S, Albrecht BK. Discovery, design, and optimization of isoxazole azepine BET inhibitors. ACS Med Chem Lett. 2013 Sep 12;4(9):835–40.
131. Piticchio SG, Martínez-Cartró M, Scaffidi S, Rachman M, Rodriguez-Arevalo S, Sanchez-Arfelis A, Escolano C, Picaud S, Krojer T, Filippakopoulos P, von Delft F, Galdeano C, Barril X. Discovery of Novel BRD4 Ligand Scaffolds by Automated Navigation of the Fragment Chemical Space. J Med Chem. 2021 Dec 23;64(24):17887–900.
132. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, Juan-Blanco T, Aloy P. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol [Internet]. 2020;38(9):1087–96. Available from: http://dx.doi.org/10.1038/s41587-020-0502-7
133. Bertoni M, Duran-Frigola M, Badia-i-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llergo A, Brun-Heath I, Villegas N, de Herreros AG, Aloy P. Bioactivity descriptors for uncharacterized chemical compounds. Nat Commun. 2021 Dec 1;12(1).
134. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot andÉdouardand M, Duchesnay andÉdouard, Duchesnay EDOUARDDUCHESNAY Fré. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot [Internet]. Vol. 12, Journal of Machine Learning Research. 2011. Available from: http://scikit-learn.sourceforge.net.

135. Yu W, MacKerell AD. Computer-Aided Drug Design Methods. Antibiotics. 2017;85–106.
136. Yuriev E, Holien J, Ramsland PA. Improvements, trends, and new ideas in molecular docking: 2012-2013 in review. Journal of Molecular Recognition. 2015;28(10):581–604.
137. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. 2004;1739–49.
138. Jones G, Willett P, Glen RC, Leach AR, Taylor R, Uk KBR. Development and Validation of a Genetic Algorithm for Flexible Docking. 1997;
139. Olson OT and AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem. 2010;31(2):455–61.
140. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins: Structure, Function and Genetics. 2004;56(2):235–49.
141. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J. On evaluating molecular-docking methods for pose prediction and enrichment factors. J Chem Inf Model. 2006;46(1):401–15.
142. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem. 2006;49(20):5912–31.
143. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. J Chem Inf Model. 2009;49(6):1455–74.
144. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, Tian S, Hou T. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem

Chem Phys [Internet]. 2016;18(18):12964–75. Available from: http://xlink.rsc.org/?DOI=C6CP01555G
145. Corbeil CR, Williams CI, Labute P. Variability in docking success rates due to dataset preparation. J Comput Aided Mol Des. 2012;26(6):775–86.
146. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW. Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem. 2007;50(4):726–41.
147. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein-ligand structures in drug discovery. Drug Discov Today [Internet]. 2012;17(23–24):1270–81. Available from: http://dx.doi.org/10.1016/j.drudis.2012.06.011
148. Wagner JR, Churas CP, Liu S, Swift R v., Chiu M, Shao C, Feher VA, Burley SK, Gilson MK, Amaro RE. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. Structure [Internet]. 2019;27(8):1326-1335.e4. Available from: https://doi.org/10.1016/j.str.2019.05.012
149. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
150. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier. J Cheminform. 2015;7(1):23.
151. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A. 2014;111(10):3733–8.
152. A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos RS and JPO. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. Nucleic Acids Res. 2014;42:1083–90.
153. Ruiz-Carmona S, Barril X. Docking-undocking combination applied to the D3R Grand Challenge 2015. J Comput Aided Mol Des. 2016;30(9):805–15.
154. https://drugdesigndata.org/about/celpp2-charts. 2021.

155. Jeliazkov JR, Robinson AC, García-Moreno BE, Bergera JM, Gray JJ. Toward the computational design of protein crystals with improved resolution. Acta Crystallogr D Struct Biol. 2019 Nov 1;75:1015–27.
156. Yu J, Liu Z, Liang Y, Luo F, Zhang J, Tian C, Motzik A, Zheng M, Kang J, Zhong G, Liu C, Fang P, Guo M, Razin E, Wang J. Second messenger Ap4A polymerizes target protein HINT1 to transduce signals in FcεRI-activated mast cells. Nat Commun. 2019 Dec 1;10(1).
157. Varela-Rial A, Majewski M, de Fabritiis G. Structure based virtual screening: Fast and slow. Wiley Interdiscip Rev Comput Mol Sci. 2021;12(2):1–17.
158. Feixas F, Lindert S, Sinko W, McCammon JA. Exploring the role of receptor flexibility in structure-based drug discovery. Biophys Chem. 2014;186:31–45.
159. Kumar A, Zhang KYJ. A cross docking pipeline for improving pose prediction and virtual screening performance. J Comput Aided Mol Des. 2018;32(1):163–73.
160. Abagyan R, Rueda M, Bottegoni G. Recipes for the selection of experimental protein conformations for virtual screening. J Chem Inf Model. 2010 Jan 25;50(1):186–93.
161. Barril X, Morley SD. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J Med Chem. 2005 Jun 30;48(13):4432–43.
162. Subramanian G, Zhu Y, Bowen SJ, Roush N, White JA, Huczek D, Zachary T, Javens C, Williams T, Janssen A, Gonzales A. Lead identification and characterization of hTrkA type 2 inhibitors. Bioorg Med Chem Lett. 2019 Nov 15;29(22).
163. Kim J, Song J, Ji HD, Yoo EK, Lee JE, Lee SB, Oh JM, Lee S, Hwang JS, Yoon H, Kim DS, Lee SJ, Jeong M, Lee S, Kim KH, Choi HS, Lee SW, Park KG, Lee IK, Kim SH, Hwang H, Jeon YH, Chin J, Cho SJ. Discovery of Potent, Selective, and Orally Bioavailable Estrogen-Related Receptor-γ Inverse Agonists to Restore the Sodium Iodide Symporter Function in Anaplastic Thyroid Cancer. J Med Chem. 2019 Feb 28;62(4):1837–58.
164. Samuels ER, Sevrioukova I. Structure-Activity Relationships of Rationally Designed Ritonavir Analogues: Impact of Side-

Group Stereochemistry, Headgroup Spacing, and Backbone Composition on the Interaction with CYP3A4. Biochemistry. 2019 Apr 16;58(15):2077–87.
165. Linciano P, Pozzi C, Iacono L dello, di Pisa F, Landi G, Bonucci A, Gul S, Kuzikov M, Ellinger B, Witt G, Santarem N, Baptista C, Franco C, Moraes CB, Müller W, Wittig U, Luciani R, Sesenna A, Quotadamo A, Ferrari S, Pöhner I, Cordeiro-Da-Silva A, Mangani S, Costantino L, Costi MP. Enhancement of Benzothiazoles as Pteridine Reductase-1 Inhibitors for the Treatment of Trypanosomatidic Infections. J Med Chem. 2019 Apr 25;62(8):3989–4012.
166. Vanamee ÉS, Faustman DL. Structural principles of tumor necrosis factor superfamily signaling [Internet]. Vol. 11, Sci. Signal. 2018. Available from: https://www.science.org
167. Renema N, Navet B, oise Heymann MF, Lezot F, Heymann D. RANK – RANKL signalling in cancer. Biosci Rep. 2016;36:1–17.
168. Mukai Y, Nakamura T, Yoshikawa M, Yoshioka Y, Tsunoda SI, Nakagawa S, Yamagata Y, Tsutsumi Y. Solution of the Structure of the TNF-TNFR2 Complex. Science Signaling [Internet]. 2010 Nov 16;3(148). Available from: www.SCIENCESIGNALING.org
169. Gonzalez-Suarez E, Jacob AP, Jones J, Miller R, Roudier-Meyer MP, Erwert R, Pinkas J, Branstetter D, Dougall WC. RANK ligand mediates progestin-induced mammary epithelial proliferation and carcinogenesis. Nature. 2010 Nov 4;468(7320):103–7.
170. Yoldi G, Pellegrini P, Trinidad EM, Cordero A, Gomez-Miragaya J, Serra-Musach J, Dougall WC, Muñoz P, Pujana MA, Planelles L, González-Suárez E. RANK signaling blockade reduces breast cancer recurrence by inducing tumor cell differentiation. Cancer Res. 2016 Oct 1;76(19):5857–69.
171. Palafox M, Ferrer I, Pellegrini P, Vila S, Hernandez-Ortega S, Urruticoechea A, Climent F, Soler MT, Muñoz P, Viñals F, Tometsko M, Branstetter D, Dougall WC, González-Suárez E. RANK induces epithelial-mesenchymal transition and stemness in human mammary epithelial cells and promotes tumorigenesis and metastasis. Cancer Res. 2012 Jun 1;72(11):2879–88.

172. Pasquale P, Alex C, Marta Ines G, William DC, Purificaciõn M, Miguel Angel P, Eva GS. Constitutive activation of RANK disrupts mammary cell fate leading to tumorigenesis. Stem Cells. 2013 Sep;31(9):1954–65.
173. Coleman R, Finkelstein DM, Barrios C, Martin M, Iwata H, Hegg R, Glaspy J, Periañez AM, Tonkin K, Deleu I, Sohn J, Crown J, Delaloge S, Dai T, Zhou Y, Jandial D, Chan A. Adjuvant denosumab in early breast cancer (D-CARE): an international, multicentre, randomised, controlled, phase 3 trial. Lancet Oncol. 2020 Jan 1;21(1):60–72.
174. Téletchéa S, Stresing V, Hervouet S, Baud'Huin M, Heymann MF, Bertho G, Charrier C, Ando K, Heymann D. Novel RANK antagonists for the treatment of bone-resorptive disease: Theoretical predictions and experimental validation. Journal of Bone and Mineral Research. 2014;29(6):1466–77.
175. Croston GE. The utility of target-based discovery. Expert Opin Drug Discov [Internet]. 2017;12(5):427–9. Available from: http://dx.doi.org/10.1080/17460441.2017.1308351
176. Weiner DM, Burstein ES, Nash N, Croston GE, Currier EA, Vanover KE, Harvey SC, Donohue E, Hansen HC, Andersson CM, Spalding TA, Gibson DFC, Krebs-Thomson K, Powell SB, Geyer MA, Hacksell U, Brann MR. 5-Hydroxytryptamine 2A Receptor Inverse Agonists as Antipsychotics [Internet]. Vol. 299, THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS. 9321. Available from: http://jpet.aspetjournals.org
177. Vanover KE, Weiner DM, Makhay M, Veinbergs I, Gardell LR, Lameh J, del Tredici AL, Piu F, Schiffer HH, Ott TR, Burstein ES, Uldam AK, Thygesen MB, Schlienger N, Andersson CM, Son TY, Harvey SC, Powell SB, Geyer MA, Tolf BR, Brann MR, Davis RE. Pharmacological and behavioral profile of N-(4-fluorophenylmethyl)-N-(1-methylpiperidin-4-yl)-N′-(4-(2-methylpropyloxy)phenylmethyl) carbamide (2R,3R)-dihydroxybutanedioate (2:1) (ACP-103), a novel 5-hydroxytryptamine2A receptor inverse agonist. Journal of

Pharmacology and Experimental Therapeutics. 2006 May;317(2):910–8.
178. Sams-Dodd F. Target-based drug discovery: Is something wrong? Drug Discov Today. 2005;10(2):139–47.
179. Swinney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov. 2011;10(7):507–19.
180. Mullard A. The phenotypic screening pendulum swings. Nat Rev Drug Discov [Internet]. 2015;14(12):807–9. Available from: http://dx.doi.org/10.1038/nrd4783
181. Kotz J. Phenotypic screening, take two. Nature [Internet]. 2012;5(15):3. Available from: http://www.nature.com/scibx/journal/v5/n15/pdf/scibx.2012.380.pdf
182. Haasen D, Schopfer U, Antczak C, Guy C, Fuchs F, Selzer P. How Phenotypic Screening Influenced Drug Discovery: Lessons from Five Years of Practice. Assay Drug Dev Technol. 2017;15(6):239–46.
183. Swinney DC. Phenotypic vs. Target-based drug discovery for first-in-class medicines. Vol. 93, Clinical Pharmacology and Therapeutics. 2013. p. 299–301.
184. Eder J, Sedrani R, Wiesmann C. The discovery of first-in-class drugs: Origins and evolution. Nat Rev Drug Discov. 2014;13(8):577–87.
185. Esch EW, Bahinski A, Huh D. Organs-on-chips at the frontiers of drug discovery. Nat Rev Drug Discov. 2015;14(4):248–60.
186. Heath JR, Ribas A, Mischel PS. Single-cell analysis tools for drug discovery and development. Nat Rev Drug Discov [Internet]. 2016;15(3):204–16. Available from: http://dx.doi.org/10.1038/nrd.2015.16
187. Ståhlberg A, Kubista M, Åman P. Profiling and Its Potential Diagnostic Applications. 2011;735–40.
188. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, Jadhav A, Jensen LJ, Johnson GL, Karlson A, Leach AR, Ma'ayan A, Malovannaya A, Mani S, Mathias SL, McManus MT, Meehan TF, von Mering C, Muthas D, Nguyen DT, Overington JP, Papadatos G, Qin J, Reich C, Roth BL, Schürer SC, Simeonov A, Sklar LA, Southall N, Tomita S,

Tudose I, Ursu O, Vidović D, Waller A, Westergaard D, Yang JJ, Zahoránszky-Köhalmi G. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov [Internet]. 2018;17(5):317–32. Available from: http://dx.doi.org/10.1038/nrd.2018.14
189. Papa A, Wan L, Bonora M, Salmena L, Song MS, Hobbs RM, Lunardi A, Webster K, Ng C, Newton RH, Knoblauch N, Guarnerio J, Ito K, Turka LA, Beck AH, Pinton P, Bronson RT, Wei W, Pandolfi PP. Cancer-Associated PTEN Mutants Act in a Dominant-Negative Manner to Suppress PTEN Protein Function. Cell [Internet]. 157(3):595–610. Available from: http://dx.doi.org/10.1016/j.cell.2014.03.027
190. Lee YR, Pandolfi PP. PTEN Mouse Models of Cancer Initiation and Progression. Cold Spring Harb Perspect Med. 2019;a037283.
191. Leslie NR, Batty IH, Maccario H, Davidson L, Downes CP. Understanding PTEN regulation: PIP2, polarity and protein stability. Oncogene. 2008;27(41):5464–76.
192. McLoughlin NM, Mueller C, Grossmann TN. The Therapeutic Potential of PTEN Modulation: Targeting Strategies from Gene to Protein. Cell Chem Biol [Internet]. 2018;25(1):19–29. Available from: https://doi.org/10.1016/j.chembiol.2017.10.009
193. Lee Y ru, Chen M, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. Nat Rev Mol Cell Biol [Internet]. 2018;19(September). Available from: http://dx.doi.org/10.1038/s41580-018-0015-0
194. Yang JS, Seo SW, Jang S, Jung GY, Kim S. Rational engineering of enzyme allosteric regulation through sequence evolution analysis. PLoS Comput Biol. 2012 Jul;8(7).
195. SmartsView.
196. Alimonti A, Nardella C, Chen Z, Clohessy JG, Carracedo A, Trotman LC, Cheng K, Varmeh S, Kozma SC, Thomas G, Rosivatz E, Woscholski R, Cognetti F, Scher HI, Pandolfi PP. A novel type of cellular senescence that can be enhanced in mouse models and human tumor xenografts to suppress prostate tumorigenesis. Journal of Clinical Investigation. 2010 Mar 1;120(3):681–93.

197. H.-Y. FU, L. SHEN, X.-S. GA, D.-X. CUI, Z.-Y. CUI. SF1670 inhibits apoptosis and inflammation via the PTEN/Akt pathway and thus protects intervertebral disc degeneration. Eur Rev Med Pharmacol Sci. 2020;24:86494–8702.
198. Spinelli L, Lindsay YE, Leslie NR. PTEN inhibitors: An evaluation of current compounds. Adv Biol Regul [Internet]. 2015;57:102–11. Available from: http://dx.doi.org/10.1016/j.jbior.2014.09.012
199. Kim MJ, Lee SJ, Ryu JH, Kim SH, Kwon IC, Roberts TM. Combination of KRAS gene silencing and PI3K inhibition for ovarian cancer treatment. Journal of Controlled Release. 2020 Feb 1;318:98–108.
200. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. Journal of Clinical Investigation. 2009 Jun 1;119(6):1420–8.
201. Roche J. The epithelial-to-mesenchymal transition in cancer. Vol. 10, Cancers. MDPI AG; 2018.
202. Their JP. Epithelial-mesenchymal transitions in tumor progression. Vol. 2, Nature Reviews Cancer. European Association for Cardio-Thoracic Surgery; 2002. p. 442–54.
203. Yang J, Weinberg RA. Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis. Vol. 14, Developmental Cell. 2008. p. 818–29.
204. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Vol. 144, Cell. 2011. p. 646–74.
205. Bowen KA, Doan HQ, Zhou BP, Wang Q, Zhou Y, Rychahou PG, Evers BM. PTEN Loss Induces Epithelial-Mesenchymal Transition in Human Colon Cancer Cells.
206. Qi Y, Liu J, Chao J, Scheuerman MP, Rahimi SA, Lee LY, Li S. PTEN suppresses epithelial–mesenchymal transition and cancer stem cell activity by downregulating Abi1. Sci Rep. 2020 Dec 1;10(1).
207. Li Y, Hu Q, Li C, Liang K, Xiang Y, Hsiao H, Nguyen TK, Park PK, Egranov SD, Ambati CR, Putluri N, Hawke DH, Han L, Hung MC, Danesh FR, Yang L, Lin C. PTEN-induced partial epithelial-mesenchymal transition drives diabetic kidney disease. Journal of Clinical Investigation. 2019 Mar 1;129(3):1129–51.

208. Qi Y, Liu J, Chao J, Greer PA, Li S. PTEN dephosphorylates Abi1 to promote epithelial morphogenesis. Journal of Cell Biology. 2020 Sep 7;219(9).
209. Zhang Z, Hou SQ, He J, Gu T, Yin Y, Shen WH. PTEN regulates PLK1 and controls chromosomal stability during cell division. Cell Cycle. 2016 Sep 16;15(18):2476–85.
210. Myers MP, Pass I, Batty IH, van der Kaay J, Stolarov JP, Hemmings BA, Wigler MH, Downes CP, Tonks NK. The lipid phosphatase activity of PTEN is critical for its tumor supressor function [Internet]. Vol. 95, Biochemistry. 1998. Available from: www.pnas.org.
211. Rarey M, Stahl M. Similarity searching in large combinatorial chemistry spaces. Vol. 15, Journal of Computer-Aided Molecular Design. 2001.
212. Schmidt R, Klein R, Rarey M. Maximum Common Substructure Searching in Combinatorial Make-on-Demand Compound Spaces. J Chem Inf Model. 2021;
213. Seco J, Luque FJ, Barril X. Binding site detection and druggability index from first principles. J Med Chem. 2009;52(8):2363–71.
214. Alvarez-Garcia D, Barril X. Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. J Med Chem. 2014;57(20):8530–9.
215. Novoa EM, de Pouplana LR, Barril X, Orozco M. Ensemble docking from homology models. J Chem Theory Comput. 2010;6(8):2547–57.
216. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. Chem Biol Drug Des. 2019 Jan 1;93(1):12–20.
217. Rost B, Sander C. Bridging the Protein Sequence-Structure Gap by Structure Predictions. Annu Rev Biophys Biomol Struct [Internet]. 1996;25:113–49. Available from: www.annualreviews.org/aronline
218. Alvarez-Garcia D, Barril X. Relationship between protein flexibility and binding: Lessons for structure-based drug design. J Chem Theory Comput. 2014;10(6):2608–14.
219. Liu K, Lu Y, Lee JK, Samara R, Willenberg R, Sears-Kraxberger I, Tedeschi A, Park KK, Jin D, Cai B, Xu B,

Connolly L, Steward O, Zheng B, He Z. PTEN deletion enhances the regenerative ability of adult corticospinal neurons. Nat Neurosci. 2010 Sep;13(9):1075–81.
220. Takehara-Nishiuchi K, McNaughton BL. Promoting Axon Regenerationin the Adult CNS by Modulationof the PTEN/mTOR Pathway. Science (1979). 2008 Nov 7;322(5903):960–3.
221. Sun F, Park KK, Belin S, Wang D, Lu T, Chen G, Zhang K, Yeung C, Feng G, Yankner BA, He Z. Sustained axon regeneration induced by co-deletion of PTEN and SOCS3. Nature. 2011 Dec 15;480(7377):372–5.
222. Ruan H, Li J, Ren S, Gao J, Li G, Kim R, Wu H, Wang Y. Inducible and cardiac specific PTEN inactivation protects ischemia/reperfusion injury. J Mol Cell Cardiol. 2009 Feb;46(2):193–200.
223. Knafo S, Sánchez-Puelles C, Palomer E, Delgado I, Draffin JE, Mingo J, Wahle T, Kaleka K, Mou L, Pereda-Perez I, Klosi E, Faber EB, Chapman HM, Lozano-Montes L, Ortega-Molina A, Ordóñez-Gutiérrez L, Wandosell F, Viña J, Dotti CG, Hall RA, Pulido R, Gerges NZ, Chan AM, Spaller MR, Serrano M, Venero C, Esteban JA. PTEN recruitment controls synaptic and cognitive function in Alzheimer's models. Nat Neurosci. 2016 Feb 23;19(3):443–53.
224. Pun RYK, Rolle IJ, LaSarge CL, Hosford BE, Rosen JM, Uhl JD, Schmeltzer SN, Faulkner C, Bronson SL, Murphy BL, Richards DA, Holland KD, Danzer SC. Excessive Activation of mTOR in Postnatally Generated Granule Cells Is Sufficient to Cause Epilepsy. Neuron. 2012 Sep 20;75(6):1022–34.
225. Williams MR, De-Spenza T, Li M, Gulledge AT, Luikart BW. Hyperactivity of newborn pten knock-out neurons results from increased excitatory synaptic drive. Journal of Neuroscience. 2015 Jan 21;35(3):943–59.
226. Gutilla EA, Steward O. Selective neuronal PTEN deletion: Can we take the brakes off of growth without losing control? Vol. 11, Neural Regeneration Research. Editorial Board of Neural Regeneration Research; 2016. p. 1201–3.

227. Zhang P, Liu X, Abegg D, Tanaka T, Tong Y, Benhamou RI, Baisden J, Crynen G, Meyer SM, Cameron MD, Chatterjee AK, Adibekian A, Childs-Disney JL, Disney MD. Reprogramming of Protein-Targeted Small-Molecule Medicines to RNA by Ribonuclease Recruitment. J Am Chem Soc. 2021 Aug 25;143(33):13044–55.
228. Costales MG, Aikawa H, Li Y, Childs-Disney JL, Abegg D, Hoch DG, Pradeep Velagapudi S, Nakai Y, Khan T, Wang KW, Yildirim I, Adibekian A, Wang ET, Disney MD. Small-molecule targeted recruitment of a nuclease to cleave an oncogenic RNA in a mouse model of metastatic cancer. PNAS [Internet]. 2020 Feb 4;117(5):2407–11. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1914286117
229. Smith IN, Thacker S, Seyfi M, Cheng F, Eng C. Conformational Dynamics and Allosteric Regulation Landscapes of Germline PTEN Mutations Associated with Autism Compared to Those Associated with Cancer. Am J Hum Genet. 2019 May 2;104(5):861–78.
230. Grebner C, Malmerberg E, Shewmaker A, Batista J, Nicholls A, Sadowski J. Virtual screening in the cloud: How big is big enough? J Chem Inf Model. 2020 Sep 28;60(9):4274–82.
231. Brandmaier A, Hou SQ, Shen WH. Cell Cycle Control by PTEN. Vol. 429, Journal of Molecular Biology. Academic Press; 2017. p. 2265–77.
232. Bellmann L, Penner P, Gastreich M, Rarey M. Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. J Chem Inf Model. 2022 Feb 14;62(3):553–66.
233. Keseru GM, Erlanson DA, Ferenczy GG, Hann MM, Murray CW, Pickett SD. Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia. Vol. 59, Journal of Medicinal Chemistry. American Chemical Society; 2016. p. 8189–206.
234. Visini R, Awale M, Reymond JL. Fragment Database FDB-17. J Chem Inf Model. 2017 Apr 24;57(4):700–9.
235. Jacquemard C, Drwal MN, Desaphy J, Kellenberger E. Binding mode information improves fragment docking. J Cheminform. 2019;11(1).

236. Miñarro-Lleonar M, Ruiz-Carmona S, Alvarez-Garcia D, Schmidtke P, Barril X. Development of an Automatic Pipeline for Participation in the CELPP Challenge. Int J Mol Sci. 2022;23(9).
237. Chen Y, Shoichet BK. Molecular docking and ligand specificity in fragment-based inhibitor discovery. Nat Chem Biol. 2009;5(5):358–64.
238. Teotico DG, Babaoglu K, Rocklin GJ, Ferreira RS, Giannetti AM, Shoichet BK. Docking for fragment inhibitors of AmpC-lactamase [Internet]. Available from: www.pnas.orgcgidoi10.1073pnas.0813029106
239. Marcou G, Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model. 2007;47(1):195–207.
240. Hubbard RE, Chen I, Davis B. Informatics and modeling challenges in fragment-based drug discovery. Current opinion in drug discovery & development [Internet]. 2007 May;10(3):289—297. Available from: http://europepmc.org/abstract/MED/17554855
241. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Vol. 11, Drug Discovery Today. 2006. p. 580–94.
242. Gorgulla C, Boeszoermenyi A, Wang ZF, Fischer PD, Coote PW, Padmanabha Das KM, Malets YS, Radchenko DS, Moroz YS, Scott DA, Fackeldey K, Hoffmann M, Iavniuk I, Wagner G, Arthanari H. An open-source drug discovery platform enables ultra-large virtual screens. Nature. 2020 Apr 30;580(7805):663–8.
243. Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton AT, Ban F, Stern A, Cherkasov A. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nature Protocols 2022 [Internet]. 2022;1–26. Available from: https://www.nature.com/articles/s41596-021-00659-2
244. Lu T, Lu W, Luo C. A patent review of BRD4 inhibitors (2013-2019). Vol. 30, Expert Opinion on Therapeutic Patents. Taylor and Francis Ltd; 2020. p. 57–81.

APPENDIX A: SUPPLEMENTARY INFORMATION

LIST OF SUPPLEMENTARY FIGURES

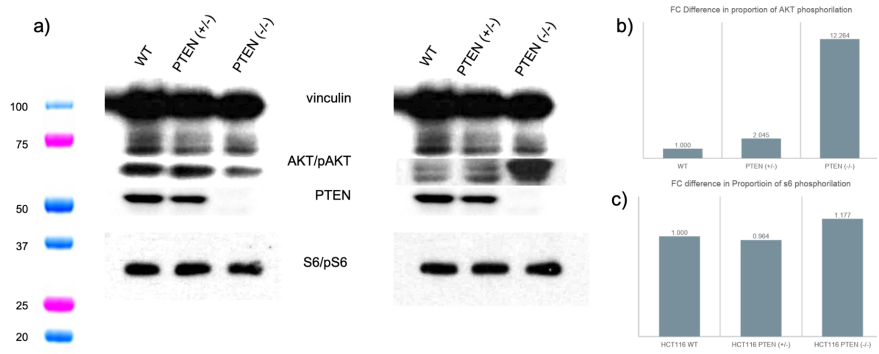
A. 1 HTVS filter file used for the Virtual Screening of RANK. 171
A. 2 HTVS filter file used for the Virtual Screening of PTEN. 171
A. 3 Characterization of the HCT116 PTEN(+/-), HCT116 PTEN(-/-) and HCT116 WT Cell lines. 172
A. 4 Western blot conditions used during the Characterization of the HCT116 PTEN(+/-), HCT116 PTEN(-/-) and HCT116 WT Cell lines. 172
A. 5 Structure of described PTEN inhibitors. 173
A. 6 SPR Plot for SF1670. 173
A. 7 SPR plot for VO-OH. 174
A. 8 SPR plot for Compound 1. 174
A. 9 SPR plot for Compound 1.2. 175
A. 10 SPR plot for Compound 1.3. 175
A. 11 SPR plot for Compound 2. 176
A. 12 SPR plot for Compound 3. 176
A. 13 SPR plot for Compound 4. 177
A. 14 SPR plot for Compound 5. 177
A. 15 SPR plot for Compound 6. 178
A. 16 SPR plot for Compound 6, Two binding site Model. 178
A. 17 SPR plot for Compound 7. 179
A. 18 SPR plot for Compound 7, Two binding site Model. 179
A. 19S PR plot for Compound 8. 180
A. 20 SPR plot for Compound 9. 180
A. 21 SPR plot for Compound 10. 181
A. 22 SPR plot for Compound 11. 181
A. 23 SPR Kinetic profile for compound 11. 182
A. 24 SPR plot for Compound 12. 182
A. 25 SPR plot for Compound 13. 183
A. 26 SPR plot for Compound 14. 183
A. 27 SPR Kinetic profile for compound 14. 184
A. 28 HTVS filter file used for the fragment Virtual Screening of BRD4. 185
A. 29 Summary of DSF and TR-FRET for JQ1. 186
A. 30 Summary of DSF and TR-FRET for BC-11D. 186
A. 31 Summary of DSF and TR-FRET for BC-15D. 187
A. 32 Summary of DSF and TR-FRET for BC-16A. 187
A. 33 Summary of DSF and TR-FRET for BC-14E. 188
A. 34 Summary of DSF and TR-FRET for BC-12D. 188
A. 35 Summary of DSF and TR-FRET for BC-14A. 189
A. 36 Summary of DSF and TR-FRET for BC-14B. 189
A. 37 Summary of DSF and TR-FRET for BC-15A. 190
A. 38 Summary of DSF and TR-FRET for BC-17C. 190
A. 39 Summary of DSF and TR-FRET for BC-02E. 191
A. 40 Summary of DSF and TR-FRET for BC-05D. 191
A. 41 Summary of DSF and TR-FRET for BC-15C. 192
A. 42 Summary of DSF and TR-FRET for BC-07C. 192

```
5
if - -16 SCORE.INTER 1.0 if - SCORE.NRUNS  5 0.0 -1.0,
if - 1 SCORE.RESTR.PHARMA 1.0 if - SCORE.NRUNS  5 0.0 -1.0,
if - -21 SCORE.INTER 1.0 if - SCORE.NRUNS  15 0.0 -1.0,
if - 1 SCORE.RESTR.PHARMA 1.0 if - SCORE.NRUNS  3 0.0 -1.0,
if - SCORE.NRUNS 49 0.0 -1.0,
2
- SCORE.INTER -10,
- SCORE.RESTR.PHARMA 1,
```

A. 1 HTSV filter file used for the Virtual Screening of RANK. The first 5 lines correspond to the running filters and the last 2 correspond to the writing filters

```
5
if - -18 SCORE.INTER 1.0 if - SCORE.NRUNS  5 0.0 -1.0,
if - 1 SCORE.RESTR.PHARMA 1.0 if - SCORE.NRUNS  5 0.0 -1.0,
if - -23 SCORE.INTER 1.0 if - SCORE.NRUNS  15 0.0 -1.0,
if - 1 SCORE.RESTR.PHARMA 1.0 if - SCORE.NRUNS  3 0.0 -1.0,
if - SCORE.NRUNS 49 0.0 -1.0,
2
- SCORE.INTER -10,
-SCORE.RESTR.PHARMA 1
```

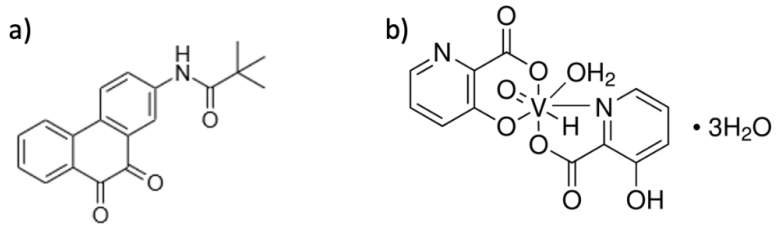
A. 2 HTSV filter file used for the Virtual Screening of PTEN**.** The first 5 lines correspond to the running filters and the last 2 correspond to the writing filters.



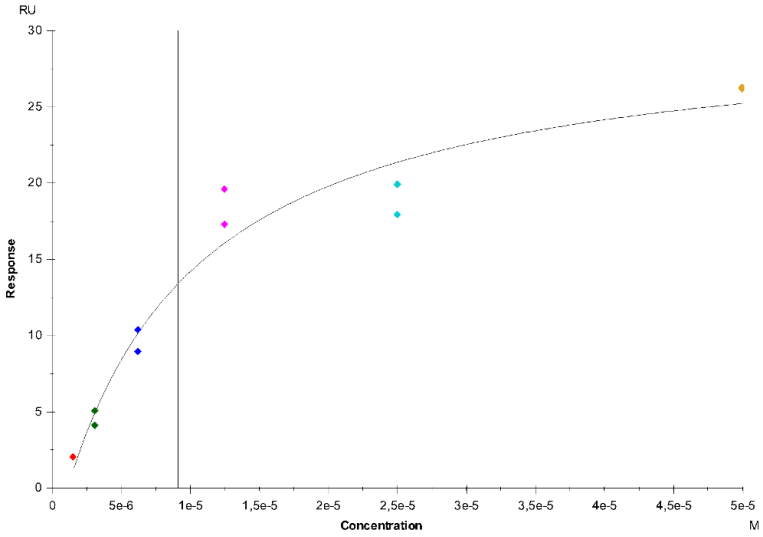
A. 3 Characterization of the HCT116 PTEN(+/-), HCT116 PTEN(-/-) and HCT116 WT Cell lines. a) Western blot. On the left panel the bands correspond to the unphosphorilated state of AKT and S6. On the right panel the bands correspond to the phosphorilated state of AKT and S6. On both panels are also displayed PTEN and vinculine levels b) Fold Change difference in the phosphorilation of AKT between the three cell lines c) Fold Change difference in the phosphorilation of S6 between the three cell lines

	Prot	Ab1	Ab2	ECL
Vinculin	5µg	1/10000 (O.N 4ºC)	1/10000 (1h RT)	1’
Tubuline	5µg	1/5000 (30’ RT)	1/5000 (30’ RT)	1’
S6	5µg	1/30000 (30’ RT)	1/40000 (30’ RT)	1’
pS6	5µg	1/20000 (30’ RT)	1/20000 (30’ RT)	1’
AKT	5µg	1/10000 (O.N 4ºC)	1/10000 (1h RT)	1’
pAKT	26,6µg	1/2000 (O.N 4ºC)	1/5000 (1h RT)	5’
PTEN	20µg	1/2000 (O.N 4ºC)	1/5000 (1h RT)	5’

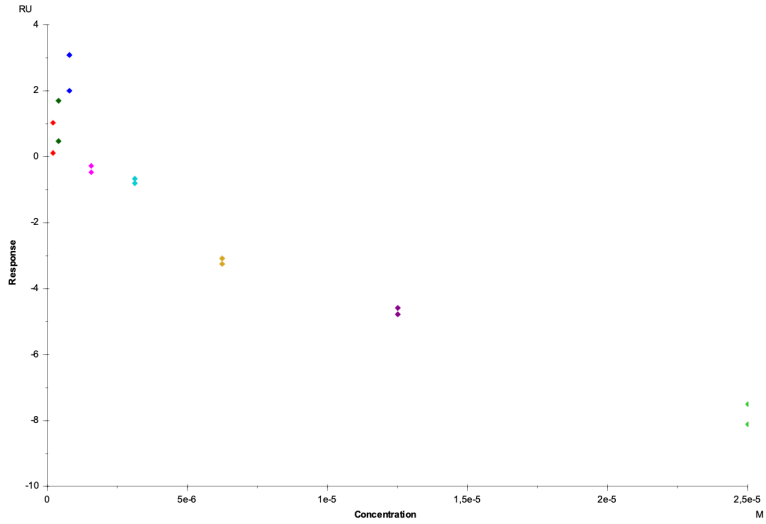
A. 4 Western blot conditions used during the Characterization of the HCT116 PTEN(+/-), HCT116 PTEN(-/-) and HCT116 WT Cell lines. Prot column refers to the amount of protein charged for the lecture of each protein. Ab1 column corresponds to the dilution of the primary antibody used. Ab2 column corresponds to the dilution of the secondary antibody used. ECL corresponds to the minutes of incubation with the Thermo Scientific SuperSignal ECL substrate.



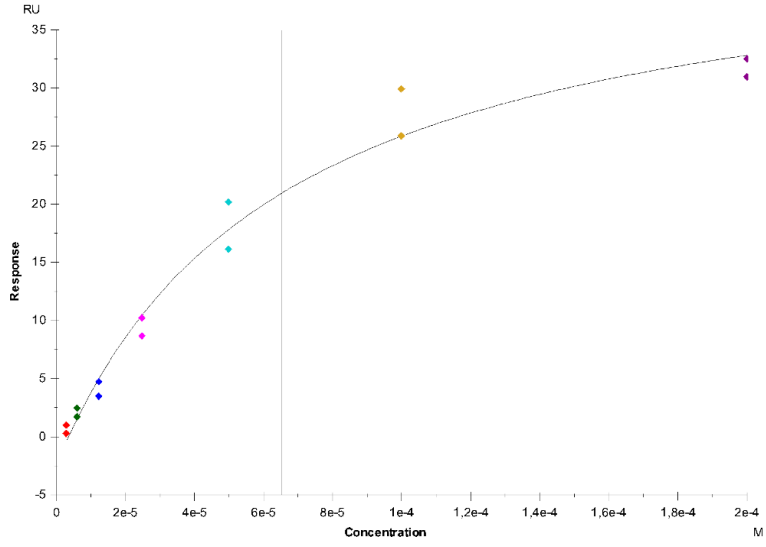
A. 5 Structure of described PTEN inhibitors. a) Structure of SF1670 b) Structure of VO-OH.



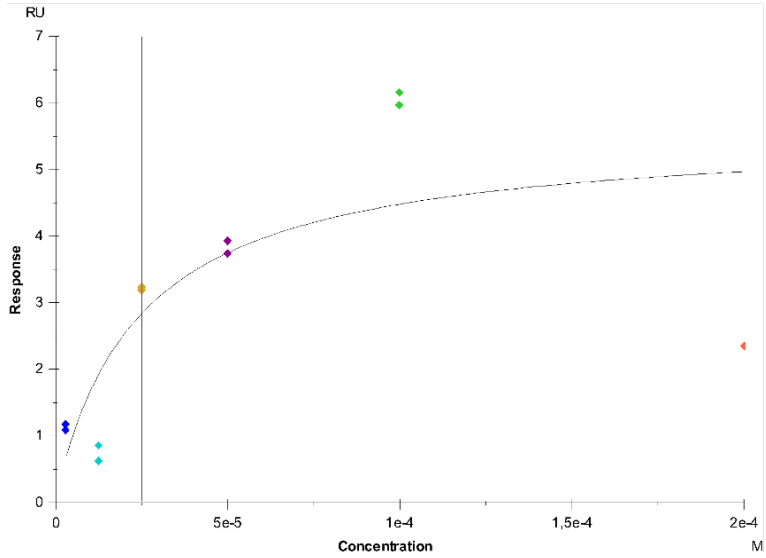
A. 6 SPR Plot for SF1670. Steady-state response against concentration to determine the binding affinity of control compound SF1670 against PTEN.



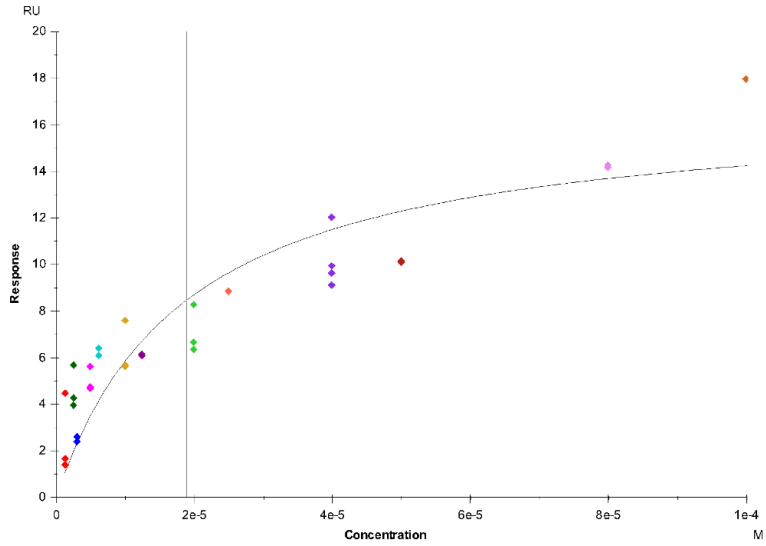
A. 7 SPR plot for VO-OH. Steady-state response against concentration to determine the binding affinity of control compound VO-OH against PTEN. There is no interaction between VO-OH and PTEN due to the degradation of the compound batch.



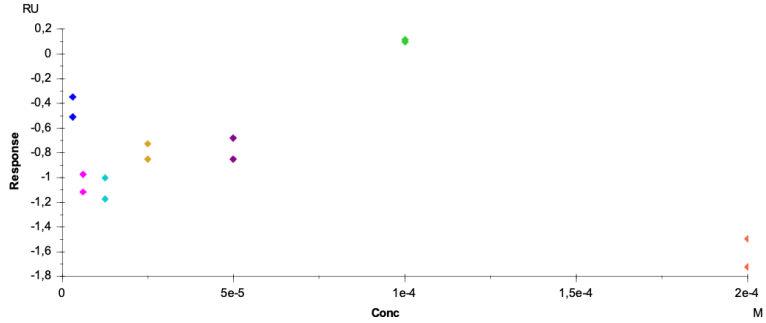
A. 8 SPR plot for Compound 1. Steady-state response against concentration to determine the binding affinity of control compound Compound 1 against PTEN.



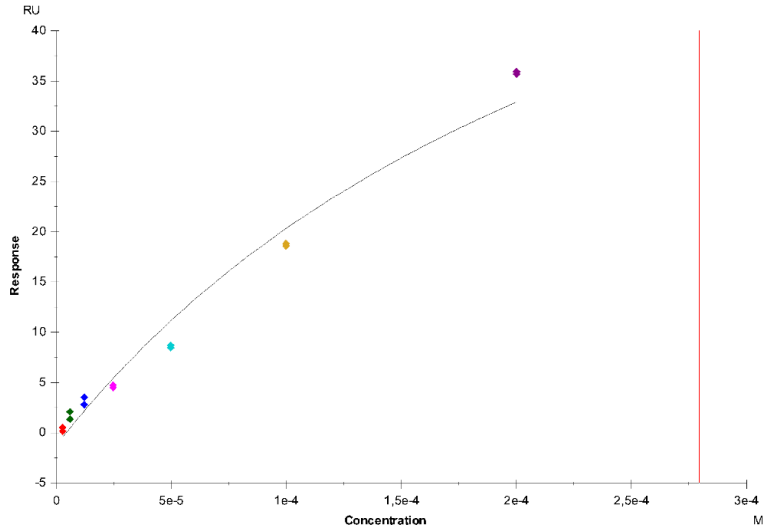
A. 9 SPR plot for Compound 1.2. Steady-state response against concentration to determine the binding affinity of control compound Compound 1.2 against PTEN.



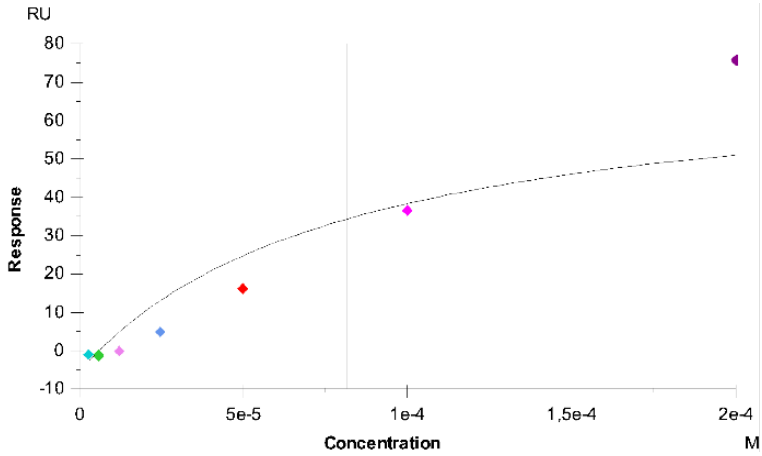
A. 10 SPR plot for Compound 1.3. Steady-state response against concentration to determine the binding affinity of control compound Compound 1.3 against PTEN.



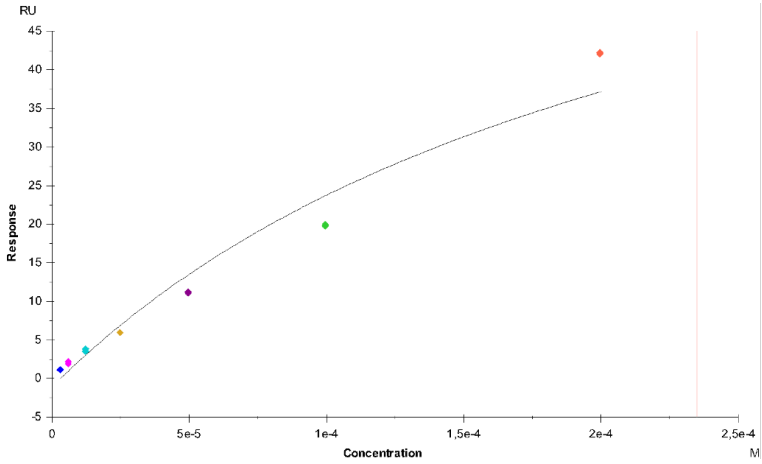
A. 11 SPR plot for Compound 2. Steady-state response against concentration to determine the binding affinity of control compound Compound 2 against PTEN. The response does not saturate and the linear signal is not consistent. RU is in the negative range. It was not possible to determine a KD for this compound.



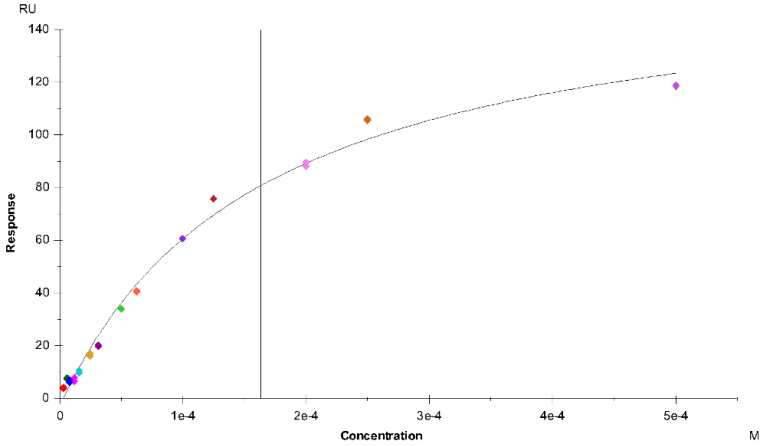
A. 12 SPR plot for Compound 3. Steady-state response against concentration to determine the binding affinity of control compound Compound 3 against PTEN.



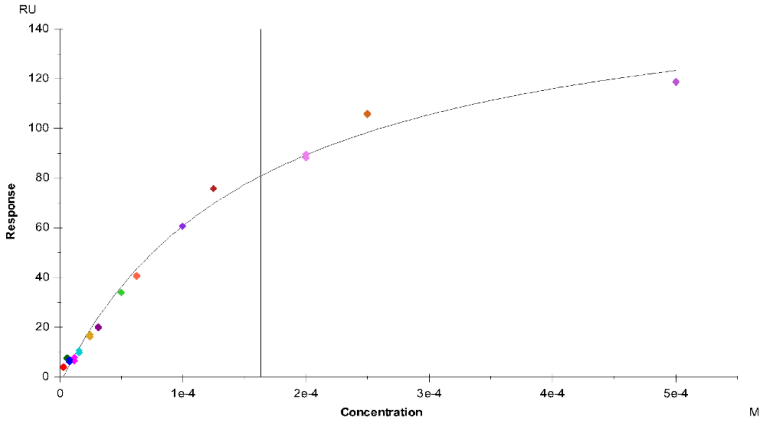
A. 13 SPR plot for Compound 4. Steady-state response against concentration to determine the binding affinity of control compound Compound 4 against PTEN.



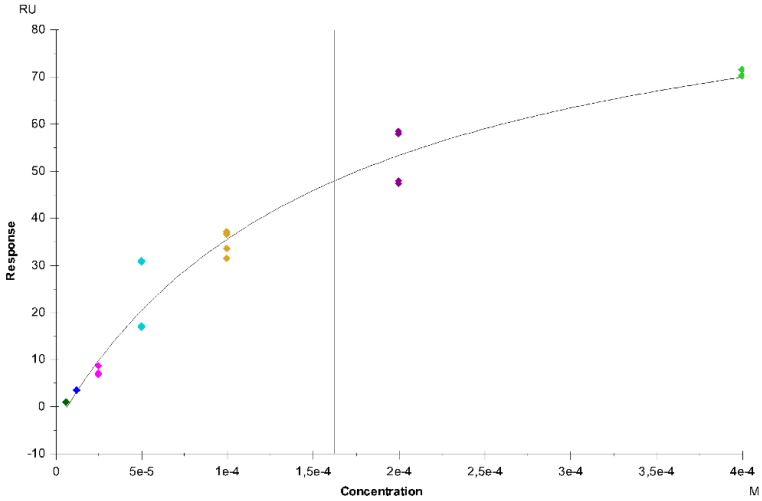
A. 14 SPR plot for Compound 5. Steady-state response against concentration to determine the binding affinity of control compound Compound 5 against PTEN.



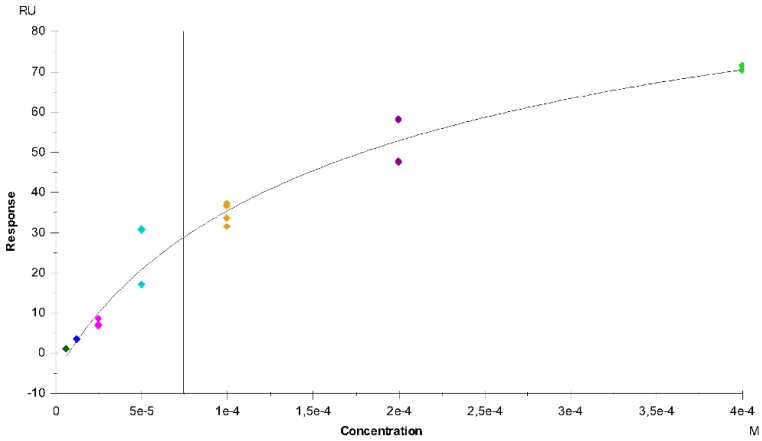
A. 15 SPR plot for Compound 6. Steady-state response against concentration to determine the binding affinity of control compound Compound 6 against PTEN. Rmax is double than expected, possible if the stoichiometry is 2:1.



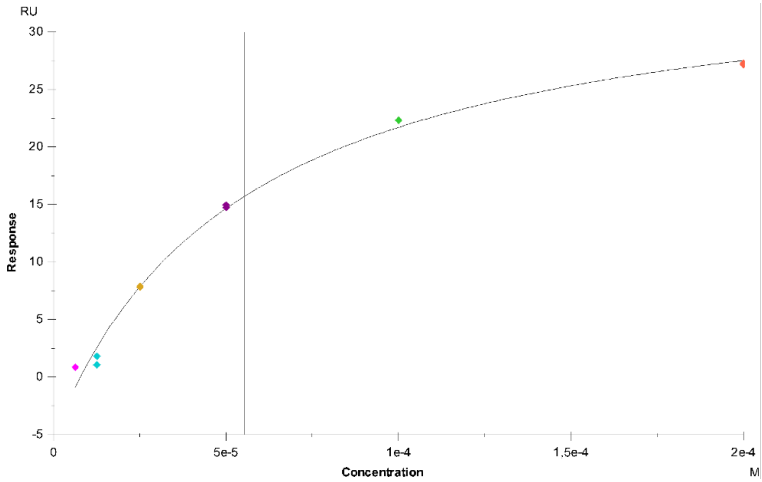
A. 16 SPR plot for Compound 6, Two binding site Model. Steady-state response against concentration to determine the binding affinity of control compound Compound 6 against PTEN fitted with a 2:1 binding model.



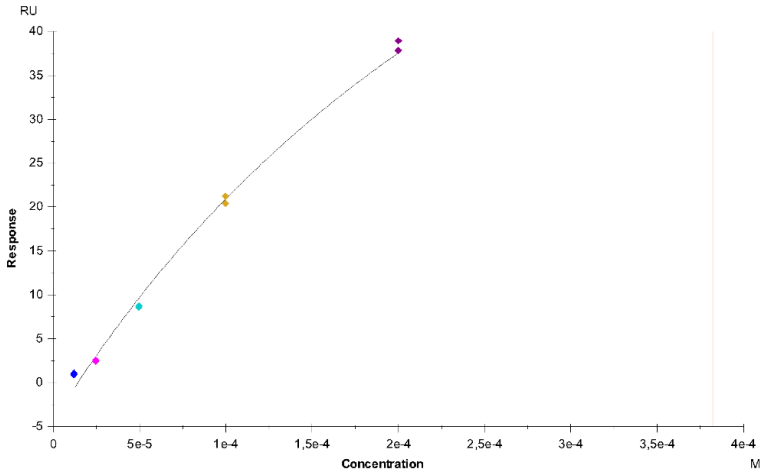
A. 17 SPR plot for Compound 7. Steady-state response against concentration to determine the binding affinity of control compound Compound 7 against PTEN. Rmax is double than expected, possible if the stoichiometry is 2:1.



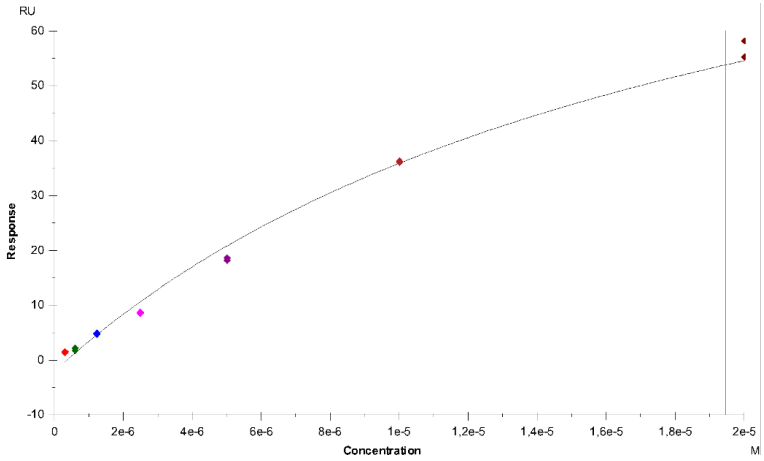
A. 18 SPR plot for Compound 7, Two binding site Model. Steady-state response against concentration to determine the binding affinity of control compound Compound 7 against PTEN fitted with a 2:1 binding model.



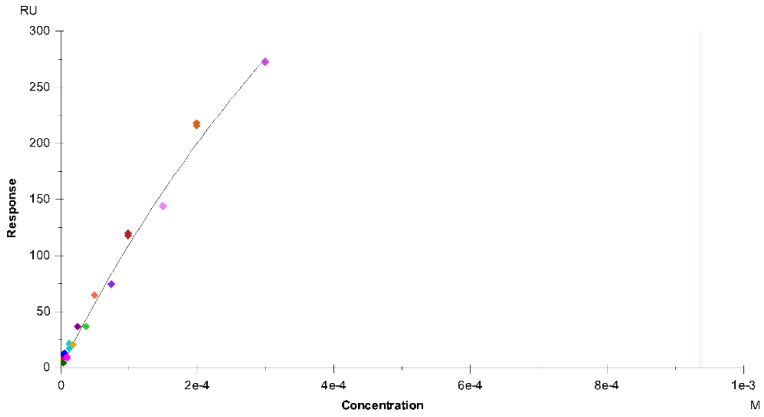
A. *19*S PR plot for Compound 8. Steady-state response against concentration to determine the binding affinity of control compound Compound 8 against PTEN.



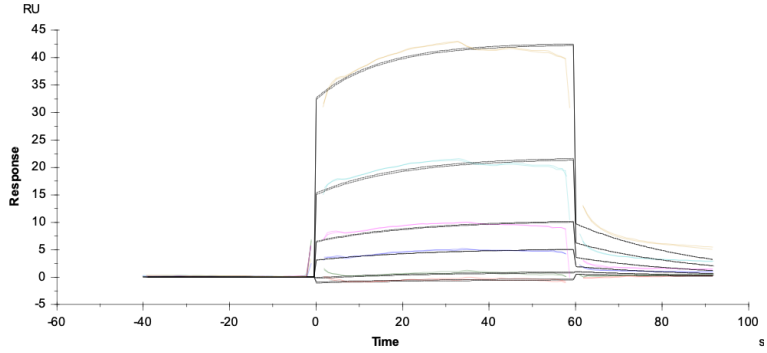
A. 20 SPR plot for Compound 9. Steady-state response against concentration to determine the binding affinity of control compound Compound 9 against PTEN.



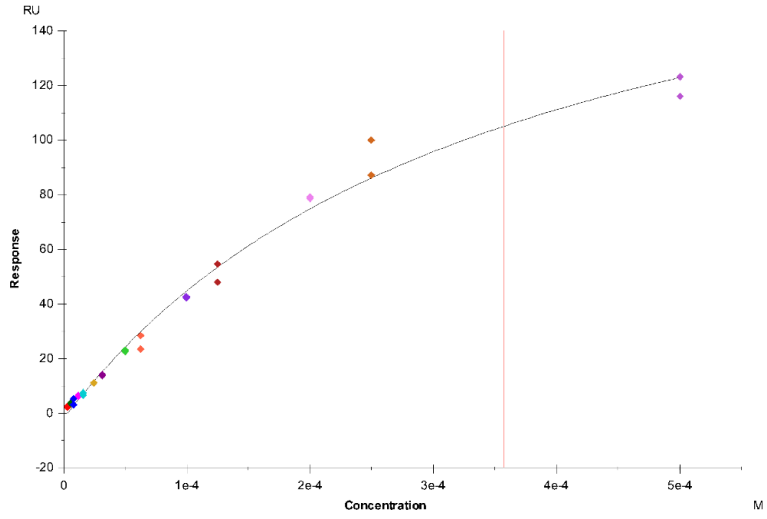
A. 21 SPR plot for Compound 10. Steady-state response against concentration to determine the binding affinity of control compound Compound 10 against PTEN. Tested at a maximum concentration of 20μM because it precipitates.



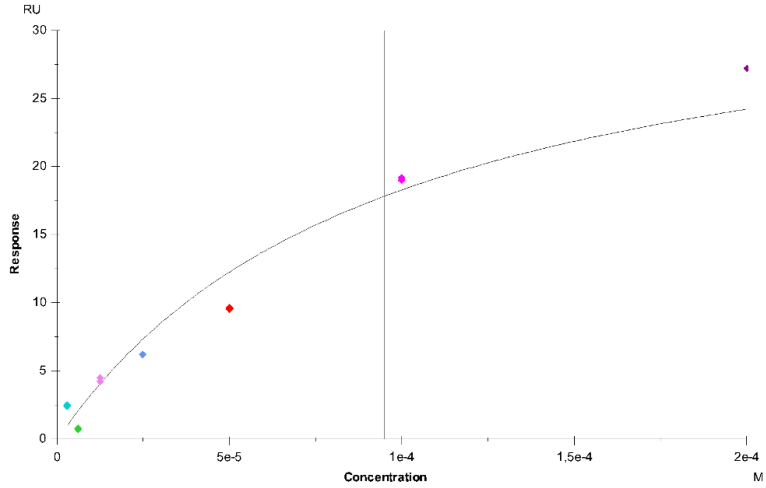
A. 22 SPR plot for Compound 11. Steady-state response against concentration to determine the binding affinity of control compound Compound 11 against PTEN. Saturation not reached giving higher RU than expected.



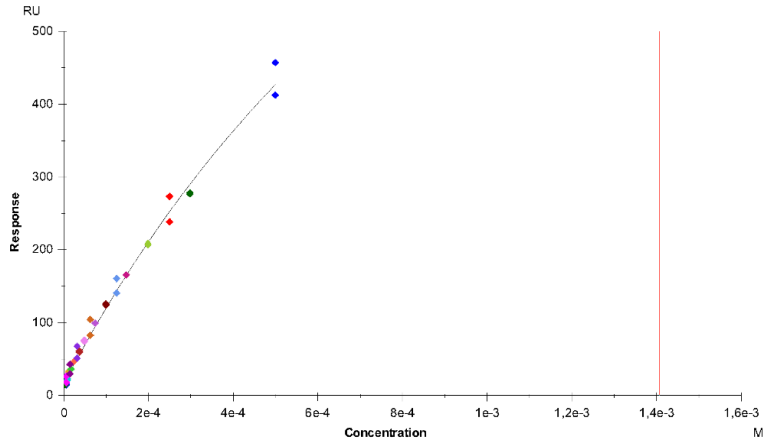
A. 23 SPR Kinetic profile for compound 11. Kinetic profile of compound 11 against PTEN.



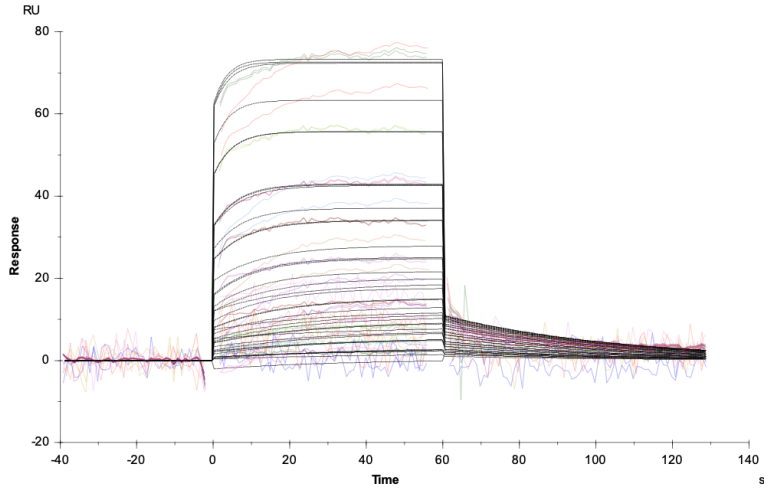
A. 24 SPR plot for Compound 12. Steady-state response against concentration to determine the binding affinity of control compound Compound 12 against PTEN.



A. 25 SPR plot for Compound 13. Steady-state response against concentration to determine the binding affinity of control compound Compound 13 against PTEN.



A. 26 SPR plot for Compound 14. Steady-state response against concentration to determine the binding affinity of control compound Compound 14 against PTEN. Saturation not reached giving higher RU than expected.



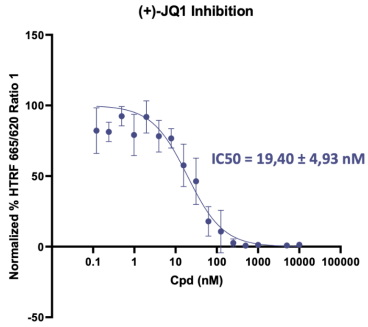
A. 27 SPR Kinetic profile for compound 14. Kinetic profile of compound 14 against PTEN. The spikes on the signal are due to problems in data collection with the instrument.

```
5
if - -3 SCORE.INTER 1.0 if - SCORE.NRUNS  3 0.0 -1.0,
if - 2 SCORE.RESTR.PHARMA 1.0 if - SCORE.NRUNS  3 0.0 -1.0,
if - -6 SCORE.INTER 1.0 if - SCORE.NRUNS  8 0.0 -1.0,
if - 1 SCORE.RESTR.PHARMA 1.0 if - SCORE.NRUNS 8 0.0 -1.0,
if - SCORE.NRUNS 15 0.0 -1.0,
2
- SCORE.INTER -8,
- SCORE.RESTR.PHARMA 1,
```

A. 28 HTVS filter file used for the fragment Virtual Screening of BRD4. The first 5 lines correspond to the running filters and the last 2 correspond to the writing filters.

(+)-JQ1

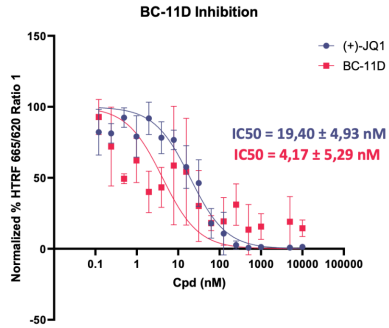
Data	Value	SD	R2
ΔTm @ 1 µM	2,49	0,1628	
TR-FRET IC50 (nM) t=1.5h	19,40	4,93	0,8981
Solubility problems	No		
logS predicted			



A. 29 Summary of DSF and TR-FRET for JQ1. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-11D (Comp1_25-30_2394209)

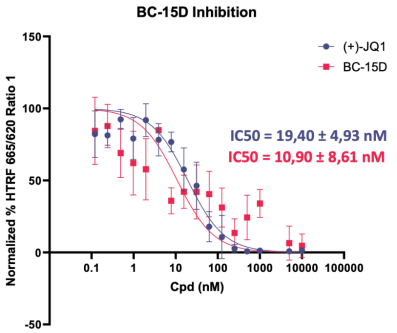
Data	Value	SD	R2
ΔTm @ 1 µM	0,90	0,2000	
TR-FRET IC50 (nM) t=1.5h	4,174	5,29	0,06617
Solubility problems	No		
logS predicted	-3,31		



A. 30 Summary of DSF and TR-FRET for BC-11D. The initial scaffold for this compound is the Computational fragment 1. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-15D (Comp2_306939)

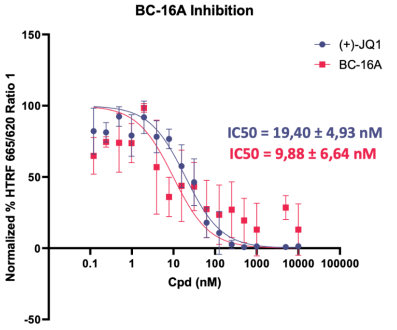
Data	Value	SD	R2
ΔTm @ 1 µM	1,96	0,1783	
TR-FRET IC50 (nM) t=1.5h	10,90	8,61	0,3388
Solubility problems	Yes		
logS predicted	-3,87		



A. 31 Summary of DSF and TR-FRET for BC-15D. The initial scaffold for this compound is the Computational fragment 2.In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-16A (Comp2_30626)

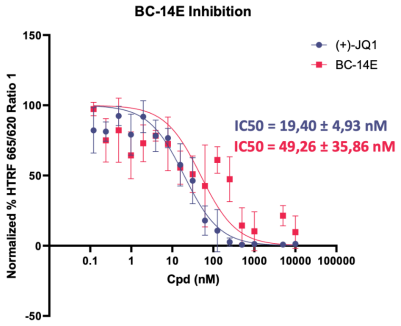
Data	Value	SD	R2
ΔTm @ 1 µM	2,19	0,1673	
TR-FRET IC50 (nM) t=1.5h	9,88	6,64	0,3113
Solubility problems	No		
logS predicted	-3,79		



A. 32 Summary of DSF and TR-FRET for BC-16A. The initial scaffold for this compound is the Computational fragment 2. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-14E (Comp2_9173395)

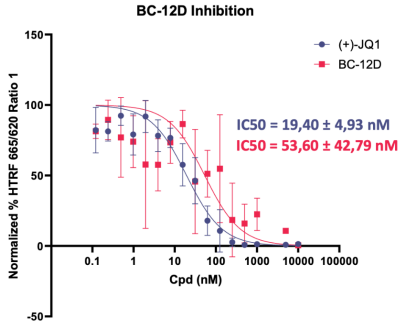
Data	Value	SD	R2
ΔTm @ 1 µM	0,91	0,0759	
TR-FRET IC50 (nM) t=1.5h	49,26	35,86	0,4104
Solubility problems	No		
logS predicted	-0,927		



A. 33 Summary of DSF and TR-FRET for BC-14E. The initial scaffold for this compound is the Computational fragment 2. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-12D (Comp5_81428)

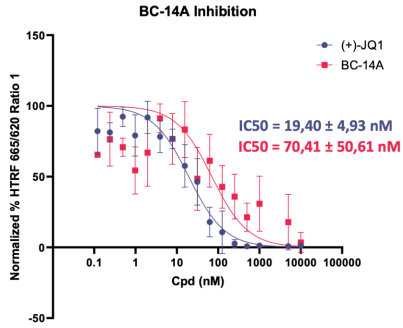
Data	Value	SD	R2
ΔTm @ 1 µM	0,79	0,2096	
TR-FRET IC50 (nM) t=1.5h	53,60	42,79	0,4411
Solubility problems	Yes		
logS predicted	-5,15		



A. 34 Summary of DSF and TR-FRET for BC-12D. The initial scaffold for this compound is the Computational fragment 5. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-14A (Comp6_112847)

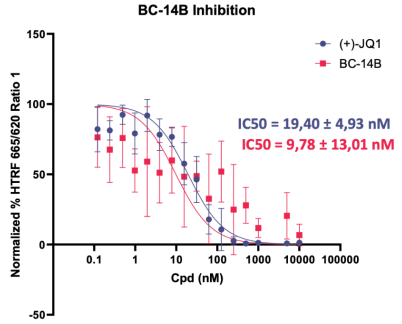
Data	Value	SD	R2
ΔTm @ 1 µM	1,29	0,3659	
TR-FRET IC50 (nM) t=1.5h	70,41	50,61	0,2244
Solubility problems	No		
logS predicted	-5,5		



A. 35 Summary of DSF and TR-FRET for BC-14A. The initial scaffold for this compound is the Computational fragment 6. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-14B (Comp6_15593)

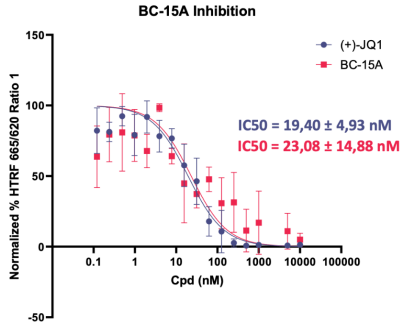
Data	Value	SD	R2
ΔTm @ 1 µM	1,68	0,2446	
TR-FRET IC50 (nM) t=1.5h	9,783	13,01	-0,1267
Solubility problems	No		
logS predicted	-4,32		



A. 36 Summary of DSF and TR-FRET for BC-14B. The initial scaffold for this compound is the Computational fragment 6. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-15A (Comp6_882887)

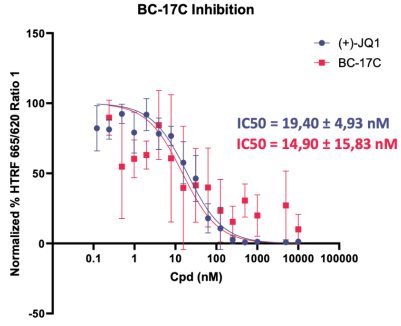
Data	Value	SD	R2
ΔTm @ 1 µM	0,92	0,2128	
IC50 TR-FRET (nM) t=1.5h	23,08	14,88	0,4376
Solubility problems	Yes		
logS predicted	-4,75		



A. 37 Summary of DSF and TR-FRET for BC-15A. The initial scaffold for this compound is the Computational fragment 6. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET

BC-17C (JQ1_10314792)

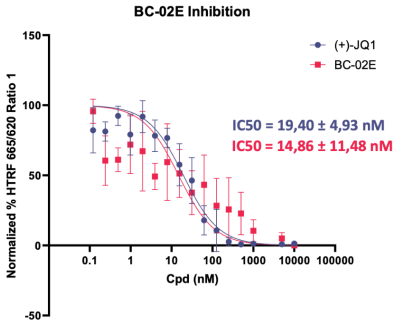
Data	Value	SD	R2
ΔTm @ 1 µM	1,23	0,3800	
IC50 TR-FRET (nM) t=1.5h	14,90	15,83	0,1362
Solubility problems	No		
logS predicted	-4,63		



A. 38 Summary of DSF and TR-FRET for BC-17C. The initial scaffold for this compound is JQ1. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-02E (4LZS_1272046)

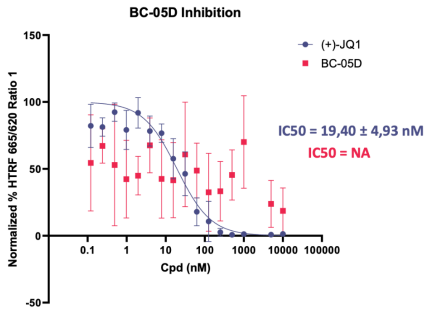
Data	Value	SD	R2
ΔTm @ 1 µM	0,90	0,0452	
IC50 TR-FRET (nM) t=1.5h	14,86	11,48	0,3484
Solubility problems	No		
logS predicted	-4,69		



A. 39 Summary of DSF and TR-FRET for BC-02E. The initial scaffold for this compound is the crystalized fragment in PDB 4LZ6. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-05D (6ZED_708831)

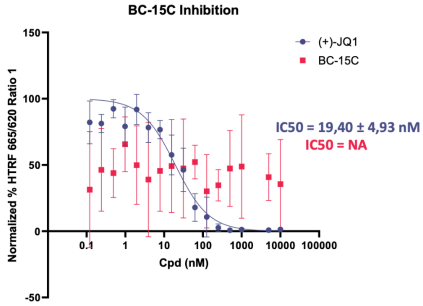
Data	Value	SD	R2
ΔTm @ 1 µM	2,09	0,0326	
IC50 TR-FRET (nM) t=1.5h	NA	NA	-1,352
Solubility problems	Yes		
logS predicted	-3,99		



A. 40 Summary of DSF and TR-FRET for BC-05D. The initial scaffold for this compound is the crystalized fragment in PDB 6ZED.. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-15C (Comp4_3128559)

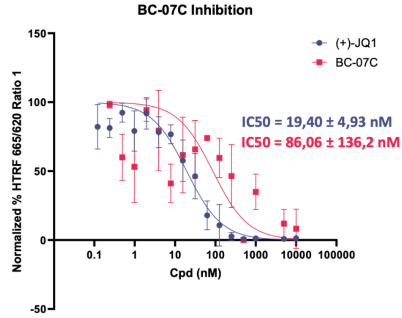
Data	Value	SD	R2
ΔTm @ 1 µM	0,75	0,5757	
IC50 TR-FRET (nM) t=1.5h	NA	NA	-1,757
Solubility problems	No		
logS predicted	-3,76		



A. 41 Summary of DSF and TR-FRET for BC-15C. The initial scaffold for this compound is the Computational fragment 4. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

BC-07C (ABBV_89798)

Data	Value	SD	R2
ΔTm @ 1 µM	1,07	0,0110	
IC50 TR-FRET (nM) t=1.5h	86,06	136,2	-0,2002
Solubility problems	No		
logS predicted	-4,13		



A. 42 Summary of DSF and TR-FRET for BC-07C. The initial scaffold for this compound is ABBV1. In the left table are displayed the ΔTm at 1µM obtained with DSF, the IC50 value obtained with TR-FRET at 1.5h and if any solubility problems were observed for the compound. On the right is depicted the Dose-Response curve and the IC50 value obtained with TR-FRET.

APPENDIX B: PUBLICATIONS

International Journal of
Molecular Sciences

MDPI

Article

Development of an Automatic Pipeline for Participation in the CELPP Challenge

Marina Miñarro-Lleonar [1], Sergio Ruiz-Carmona [2], Daniel Alvarez-Garcia [3], Peter Schmidtke [4] and Xavier Barril [1,3,5,*]

[1] Pharmacy Faculty, University of Barcelona, Av. de Joan XXIII 27-31, 08028 Barcelona, Spain; mminarro@ub.edu
[2] Baker Heart and Diabetes Institute, Melbourne 3004, Australia; sruizcarmona@gmail.com
[3] GAIN Therapeutics, Parc Cientific de Barcelona, Baldiri i Reixac 10, 08029 Barcelona, Spain; dalvarez@gaintherapeutics.com
[4] Discngine S.A.S., 79 Avenue Ledru Rollin, 75012 Paris, France; peter.schmidtke@discngine.com
[5] Catalan Institute for Research and Advanced Studies (ICREA), Passeig de Lluis Companys 23, 08010 Barcelona, Spain
* Correspondence: xbarril@ub.edu

Abstract: The prediction of how a ligand binds to its target is an essential step for Structure-Based Drug Design (SBDD) methods. Molecular docking is a standard tool to predict the binding mode of a ligand to its macromolecular receptor and to quantify their mutual complementarity, with multiple applications in drug design. However, docking programs do not always find correct solutions, either because they are not sampled or due to inaccuracies in the scoring functions. Quantifying the docking performance in real scenarios is essential to understanding their limitations, managing expectations and guiding future developments. Here, we present a fully automated pipeline for pose prediction validated by participating in the Continuous Evaluation of Ligand Pose Prediction (CELPP) Challenge. Acknowledging the intrinsic limitations of the docking method, we devised a strategy to automatically mine and exploit pre-existing data, defining—whenever possible—empirical restraints to guide the docking process. We prove that the pipeline is able to generate predictions for most of the proposed targets as well as obtain poses with low RMSD values when compared to the crystal structure. All things considered, our pipeline highlights some major challenges in the automatic prediction of protein–ligand complexes, which will be addressed in future versions of the pipeline.

Keywords: docking; D3R; automated pipeline; pocket detection; binding mode prediction

Citation: Miñarro-Lleonar, M.; Ruiz-Carmona, S.; Alvarez-Garcia, D.; Schmidtke, P.; Barril, X. Development of an Automatic Pipeline for Participation in the CELPP Challenge. *Int. J. Mol. Sci.* **2022**, *23*, 4756. https://doi.org/10.3390/ijms23094756

Academic Editors: Gary A. Piazza and Jia-Zhong Li

Received: 1 April 2022
Accepted: 21 April 2022
Published: 26 April 2022

1. Introduction

Computational approaches have proven to be a valuable addition to wet-lab techniques in the field of drug discovery [1]. Amongst them, we can find Structure-Based Drug Design (SBDD) methods, where the three-dimensional structure of biomolecules is used to identify small molecules that can interact with them. Predicting how a ligand binds to a target is an essential step for SBDD, and molecular docking has become a standard tool for drug discovery [2,3]. The outcome of docking is a set of proposed positions and conformations of the ligand in the binding site (poses), each with an associated score. These models can be used to interpret and guide ligand design well before the structure of the protein–ligand complex can be experimentally determined.

Nonetheless, docking programs do not always find accurate ligand poses when compared to the experimental solution. There are still challenges that need to be addressed such as receptor flexibility, proper accounting of solvation effects or better scoring functions [3]. Owing to the potential and relevance of docking for SBDD, there has been a substantial and sustained effort to improve the technique, and many docking tools have been developed, such as GLIDE [4], rDock [5], GOLD [6] and AutoDock [7]. Because different docking

programs use different sampling strategies and scoring functions, it is important to be able to evaluate and compare the performance between them. To that aim, test sets are available to evaluate the performance of docking and scoring methods in binding mode, binding affinity or virtual screening tasks. Regarding the former application, multiple assessments have been performed with different evaluation benchmarks [8–13]. One of the most recent and complete studies was conducted by Wang et al. (2016), who evaluated ten different docking programs, including five commercial programs and five academic programs using a collection of 2002 protein–ligand complexes from the PDB. Concurrently, a strong emphasis has been put on generating highly refined test sets, which only include high-quality structures of relevant protein targets containing drug-like ligands. Some of the most-used validation datasets are CCDC/Astex [14] and Iridium [15]. Such datasets and comparative studies provide a comprehensive understanding of the advantages and limitations of each docking program and help users make more appropriate choices among available methods. However, they suffer from an important limitation: in an attempt to keep the comparison across docking programs fair, the authors of the comparative studies use standard parameters, whereas in real-life applications, advanced users introduce substantial bias to improve performance. In consequence, such comparative studies reveal the intrinsic capabilities of the programs, which is quite different from how they are actually used in typical drug-discovery settings. In addition, as relatively small sets of well-curated protein–ligand complexes become widely adopted as test-sets, there is a risk of biasing docking programs towards those specific datasets.

The challenges organised by the Drug Design Resource (D3R) represent a welcome departure from this tendency. D3R aims to provide benchmark datasets and blinded challenges to assist in the evaluation and improvement of computational algorithms, giving participants the freedom to use the methods as they see fit, but encouraging the use of reproducible protocols. Besides the annual Grand Challenge, D3R also organises the CELPP Challenge (Continuous Evaluation of Ligand Pose Prediction) [16]. Participants in CELPP are encouraged to develop an automated workflow to generate binding mode predictions for different targets that are delivered weekly.

In this article, we describe the development of the first version of a pipeline for participation in the CELPP Challenge, as well as validation results. The main focus of our workflow is to adopt a knowledge-based approach whenever possible, trying to extract data from similar systems that are already deposited in the PDB. Depending on the amount of information available, the docking algorithm may benefit from knowledge about the location of the binding site, specific pharmacophores or even the binding mode of specific substructures. We will describe the different options, analyse their respective performances and identify aspects that need further improvement.

2. Results and Discussion

The goal of this work was to create an automated workflow for protein–ligand pose prediction. It must be able to extract information from related complexes deposited in the PDB and to use it in different docking protocols. Throughout this work, a test set consisting of structures released in previous weekly CELPP challenges was used to design the protocol and for benchmarking.

2.1. Overview of the Pipeline

One of the key aspects of this work is the automation of the process; therefore, all the steps are gathered in a combination of python, SVL and shell scripts and divided into individually functional modules corresponding to the different phases of the protocol (Figure 1). There are four phases summarized here (see Method section for further details):

Phase 1: Protein analysis. Download the sequence of the query protein, identify structures of homologous proteins in the PDB and ligands that bind to them (this is performed through a query in 3decision [17]).

Phase 2: Ligand analysis. Compute a similarity score and maximum common substructure between the query ligand and all ligands retrieved in Phase 1.

Phase 3: Pharmacophore generation. Derive, whenever possible, a pharmacophore for the ligands retrieved in Phase 1.

Phase 4: Docking. Three docking strategies are used: tethered docking (when large maximum common substructure (MCS) is shared with a reference ligand), docking with pharmacophoric restraints (if a pharmacophore could be defined in Phase 3) and docking without any restraints (in all cases).

Additionally, the process includes communication with the CELPP server to download the queries and upload the predictions.

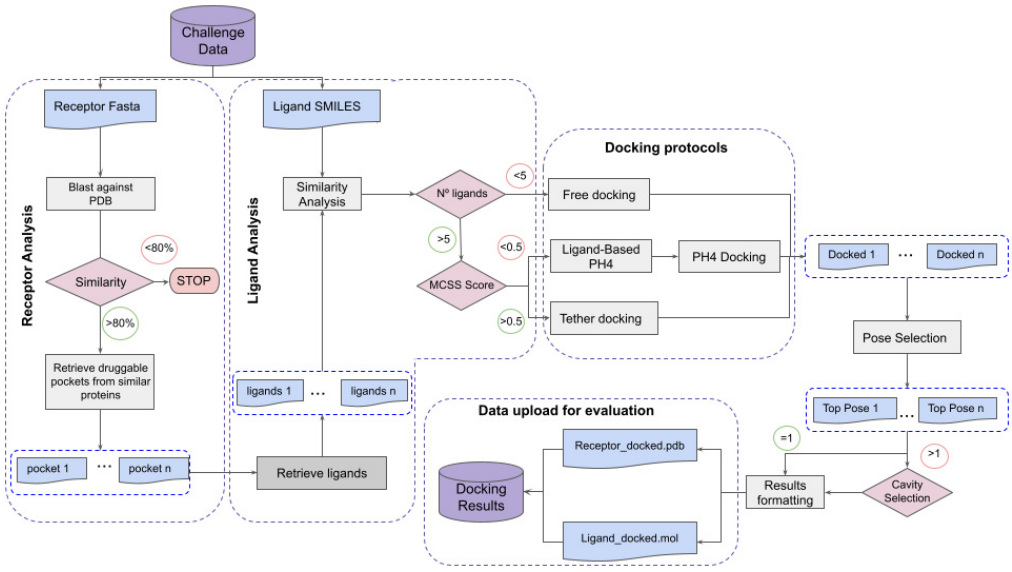


Figure 1. Workflow employed for pose prediction.

2.2. Workflow Input Data, Data Structure and Output

Each weekly CELPP data package is downloaded as a gzipped tar file that contains one directory per target. The target is a protein defined by its primary sequence. Within each directory, there is a set of structures that have the same or highly similar sequences to the target. They are provided as potential receptor structures for docking and contain the highest resolution unbound candidate protein (hiResApo), the highest resolution ligand-bound (hiResHolo), the candidate protein that contains the ligand with the largest MCSS to the target ligand (LMCSS), the candidate protein that contains the ligand with the smallest MCSS (SMCSS) and the candidate protein that contains the ligand with the highest structural similarity (based on Tanimoto score and Daylight fingerprints, as implemented by RDkit [18]) to the target ligand (hiTanimoto). Then, we find the SMILES [19], MOL file and INCHI key [20] corresponding to the target ligand. Finally, the suggested binding pocket centre is also given. However, our pipeline includes a cavity detection phase, so the suggested binding pocket centre will not be used. The expected output from participants is a docked pose of the target ligand with each suggested candidate structure.

2.3. Pipeline Development

2.3.1. Blast Results

Before starting the implementation of the pipeline, we analysed the targets from previous CELPP weeks (test set) to check how often they had high similarity homologues already deposited in the RCSB PDB. For this purpose, we ran a blast search against the

RCSB PDB with two different identity thresholds: 80% and 95%. From this step, we could conclude that 100% of the targets had some close homolog structure available (>80% identity) within the RCSB PDB prior to its release. When looking for proteins with an identity higher than 95%, we obtained varying results across weeks with an average of 77, 1% of positive cases (Figure 2A). This mirrors the trends in the PDB, which is highly redundant in protein composition [21]. In light of the results, we set the identity threshold for blast searches in our automatic pipeline to 80%.

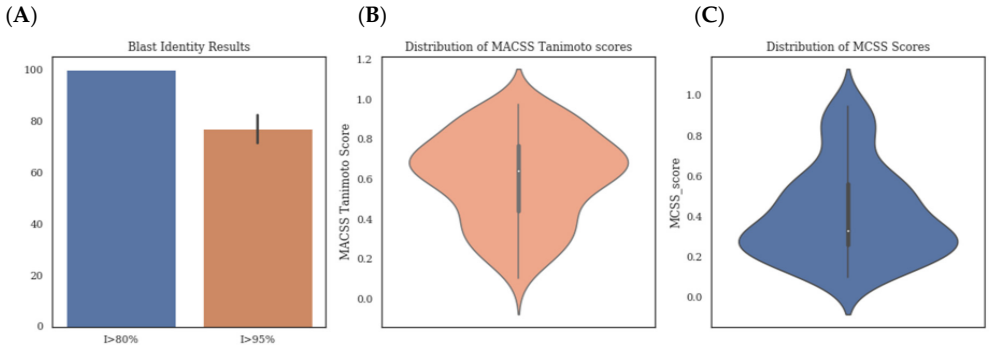


Figure 2. (**A**) Histogram representing the percentage of targets for which we obtained blast results with and identity higher than 80% (blue) and 95% (marron) (**B**) Distribution of Tanimoto MACSS score and (**C**) Tanimoto MCSS scores obtained for the ligands in the test set.

2.3.2. Ligand Similarity

We analysed the similarity between the ligands provided by CELPP and the ligands obtained by 3decision from similar proteins. After running the 3decision protocol, we were able to obtain sets of ligands for 75% of the proteins in the test set. Using MACSS keys fingerprints, we obtained a mean Tanimoto score of 0.6 with 0.008 and 0.96 being the minimum and the maximum scores obtained, respectively (Figure 2B). We also took into account the size of the compared ligands and their maximum common substructure with a complementary similarity measure, the Tanimoto MCSS [22]. Its value distribution is rather different from the Tanimoto MACSS, (Figure 2C) with average, minimum and maximum values of 0.42, 0.1 and 0.947, respectively.

2.3.3. Docking Method Selection

Using the same target, we compared the performance of the three different docking methods (tethered, pharmacophoric restraints and free) and checked if there was any kind of correlation between the docking RMSD and the Tanimoto similarity to the reference ligands. RMSD values were calculated using the sdrmsd utility from rDock. The mean RSMD values for tethered docking, docking with pharmacophoric restraints and free docking were 2.81 Å, 2.15 Å and 2.19 Å, respectively. Thus, while the use of knowledge-based restraints improved the predictions in individual cases (Figure 3), the overall performance was not better (Table 1). In the case of tethered docking, our analysis showed that it should only be applied when the Tanimoto MCSS is larger than 0.65, after which point almost all predictions were correct (Figure 4A). Unfortunately, this applied to a small proportion of the cases (15%). Surprisingly, free docking also produced improved predictions for this set of ligands, which might be due to the similarity with the ligand of reference used to define the cavity or to the protein pre-organisation (quasi self-docking). The plot also showed that using tethered docking when the MCSS is too small leads to worse predictions than free docking, explaining the apparently worst performance of tethered docking compared to free docking when considering the entire test set. Regarding pharmacophore-guided docking, contrary to our initial expectations, we found that there was not a significant difference in total mean RMSD between restrained and free docking (2.15 Å and 2,19 Å,

respectively). This could, in part, be related to the cavity definition process, which already limits the docking space and may leave a small margin for improvement. However, it also suggested that the choice of pharmacophoric restraints was sub-optimal and had to be re-optimised. Thus, we introduced an improved pharmacophore elucidation protocol (see Methods and results below).

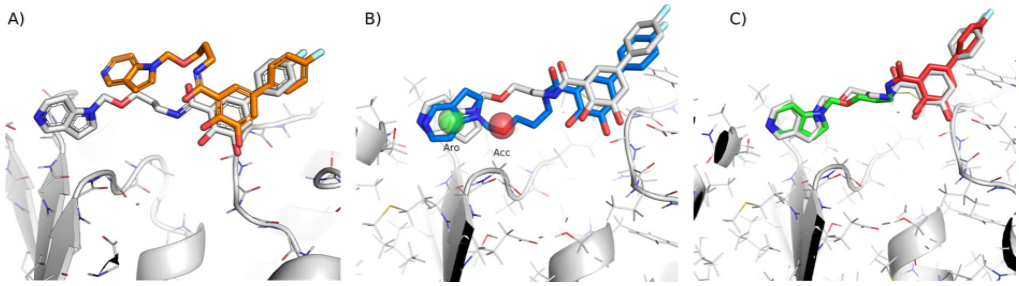


Figure 3. Differences in best pose predicted for target 5p8y from CELPP week 33. Image (**A**) corresponds to free docking with an RMSD of 4.09 Å. Image (**B**) is the best prediction obtained with pharmacophoric restraints (1.74 Å). Image (**C**) corresponds to the best pose using tethered docking, obtaining an RMSD of 0.95 Å. The red substructure indicates the tethered atoms.

Table 1. RMSD results obtained using different docking methods.

	Free Docking	Tethered Docking	Ph4 Docking
Mean	2.19	2.81	2.15
Median	1.96	1.71	1.63
Min	0.43	0.33	0.39
max	7.41	15.07	7.41

RMSD values in Å.

2.3.4. Pipeline Effectiveness and Processing Time

The above-described pipeline performance was tested with a collection of pre-released CELPP weeks as well as with the weekly released CELPP set. The execution time of the whole protocol took an average 6.5 min per target. The total execution time varied each week depending on the number of released targets (26 to 68 in the period considered here) and the connection speed to 3decision (from 22 s to 3 min per target). The 3decision protocol could not obtain reference structures for 20% of the targets due to some internal errors of a beta version of the program or because there were no ligands found in druggable pockets from similar proteins. This last event was relatively rare, as it accounted for 25% of times that we were not able to obtain results from 3decision, or 5% of the total. Finally, the similarity analysis to the docked ligand poses took 4.8 min per target on average (Table 2).

2.4. Pipeline Validation

To validate the pipeline, we ran it prospectively for a total of 12 weeks. Table 3 shows that the pharmacophoric restrained protocol was the most-used method (51% of the cases). On the other hand, free docking and tethered docking were applied in much lower percentages of cases, 35% and 13.01%, respectively. The mean RMSD value for free docking was 6.2 Å, 5.1 Å for pharmacophore-guided docking and 2.8 Å for tethered docking. However, there is a bigger difference when looking at the proportion of correctly predicted cases by each method. For free docking, only 7.9% of the cases had an RMSD value lower than 2 Å, for pharmacophore guided docking this value increased to 21.4%, and in tethered docking we reached 31.5% of correct poses.

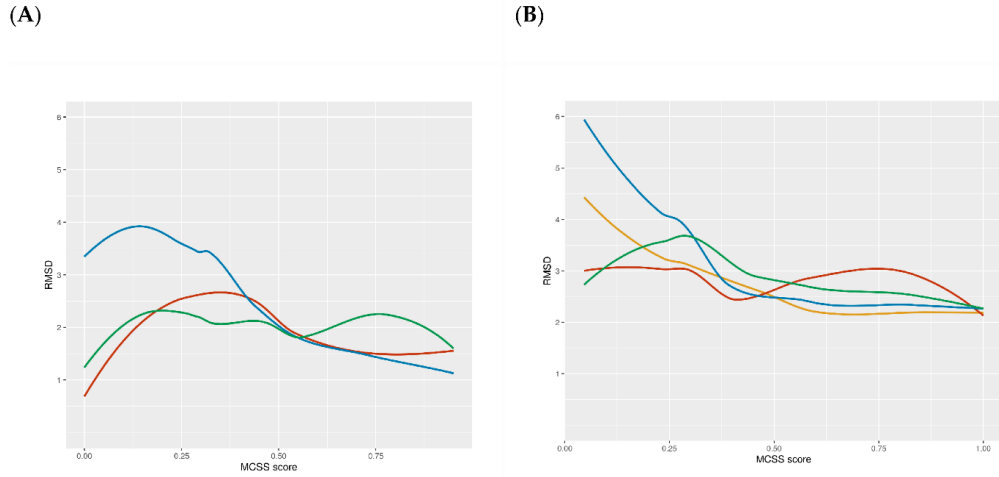


Figure 4. Relation between RMSD and the MCSS score using (**A**) the test set and (**B**) the validation set. Free docking results shown in red, docking with pharmacophoric restraints in green (version 1) and yellow (version 2; only applied to the validation set) and MCS-tethered docking in blue.

Table 2. Statistics of the pipeline implementation CELPP weeks.

	No. of Targets	3decision Time	Docking Time	Total Time
Week1	31	34	103	137
Week2	44	103	174	277
Week3	27	10	113	123
Week4	43	118	176	294
Week5	29	35	153	188
Week6	40	182	265	447
Week7	68	234	123	357
Week8	26	102	111	213
Week9	28	126	247	373
Week10	48	158	382	540
Week11	50	193	270	463
Week12	26	137	716	853
Mean	38	119.33	236.08	355.42

Time measured in minutes.

The values obtained with the validation set were much worse than the ones obtained using the test set. The main difference between the sets as that the automatic pipeline for retrieving the cavities using 3decision was not yet automatized during the development stage. In consequence, all the cavities were visually inspected and selected using the 3decision webserver. By contrast, the automatic scripts used at the validation stage to identify the docking cavity and retrieve aligned ligands from 3decision were error-prone. We also had to consider the possibility that the test set was not representative enough of the whole range of systems that can be found in the CELPP Challenge. Nonetheless, the sources of errors and the difference in performance between the test set and validation will be reviewed in the next section.

Table 3. RMSD values and percentage of cases for each docking protocol.

	Free Docking	Ph4 Docking	Tethered Docking
Mean	6.2	5.1	2.8
Std	6.2	3.4	1.6
Min	1	0.5	0.7
Q1	3.9	2.2	1.6
Q2	6.3	4.7	2.3
Q3	8.2	7.7	3.6
max	13.6	13.9	12.7
≤2Å	7.9%	51%	13%
Application rate	35%	51%	13%

RMSD values in Å.

After analysing the prospective results, we wanted to review if the algorithm for docking protocol selection derived from the test set was the most adequate one. For this purpose, we applied all three protocols to all the validation set and compared the best RMSD obtained for the three methods (Figure 4). We could find some differences regarding the accuracy of the docking methods in the test set and validation sets. Tethered docking yielded better results than free docking when MCSS score $\geq$ 0.5 on the validation set (vs. a marginal improvement on an MCSS score $\geq$ 0.65 for the test set). Nonetheless, tethered docking was still the method that gave the worst results in low MCSS score values (MCSS < 0.3). As for the pharmacophore-guided docking, during the validation phase, we improved the pharmacophoric elucidation protocol that provided consistently better results than in the test set (see Methods). It also provided improved results compared to free docking in the 0.5 to 1 MCSS score range, with a performance on par with tethered docking. In the 0.25 to 0.5 MCSS score range, pharmacophore-guided docking and free docking performed at a similar level. At lower MCSS score values, free docking outperformed pharmacophore-guided docking.

2.5. Challenges to Address

In this section we will describe the most important factors affecting the predictive performance of our pipeline. Figure 5 depicts the main issues and challenges to overcome in the CELPP challenge, which will be treated in more detail in the following sections.

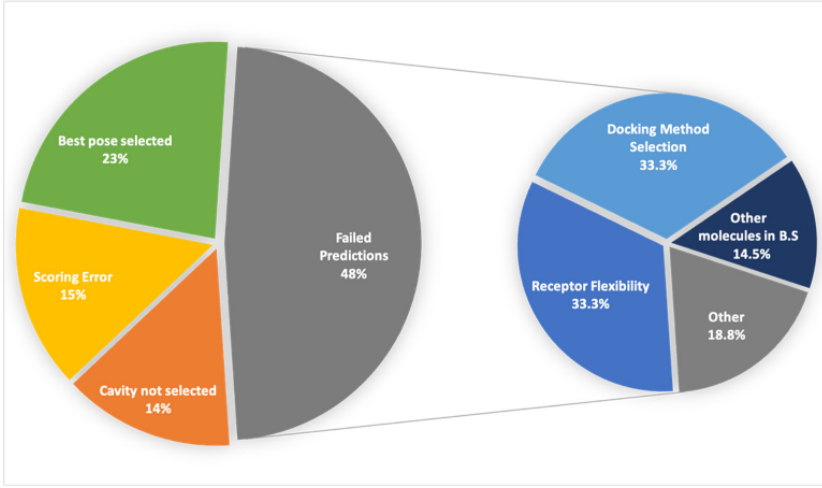


Figure 5. Overall view of validation set cases.

2.5.1. Automated Protocols

When testing a docking program or workflow, a crucial component that will have a big impact in the predictions is the choice of dataset [13]. Usually, the datasets to test docking programs, such as DUD-E [23] or Astex [14], are highly curated datasets, whilst the CELPP receptors are selected automatically and are not manually prepared by experts. Additionally, we have to take into account that CELPP is designed as a cross-docking challenge, which means that we have the added problem of protein flexibility, as the used receptor may not be in the most-fitting position for the ligand. Finally, participants are given, each week, an average of 40 systems to predict and a limited amount of time (3 days), which implies that all the processes need to be automatized, leaving virtually no time for the visual inspection or study of the targets.

In consequence, the pose prediction performance is lower than for other challenges. The median prediction RMSD for the best categories (LMCSS and hiTanimoto receptors) is around 5 Å, being only 20% of the pose predictions accurate within 2 Å [17], whereas reported performance for curated datasets regularly reaches the 80% [13]. Clearly, the latter reflects a best-case scenario, which means that a significant effort to improve automated target structure selection and preparation will be necessary in order to attain better results in CELPP.

2.5.2. Scoring Challenges

Over the past years extensive efforts have been dedicated to improving the existing scoring functions, but nowadays the accuracy of most scoring functions is still a limiting factor in many drug design projects, and results require careful evaluation and post-docking analysis.

To assess the accuracy of the docking score, we selected a subset of 446 submitted cases and checked if the submitted pose is the one with the lowest RMSD compared to the crystal structure. In 208 out of 446 total cases (46.6%) the docking protocol was able to produce a correct pose (RMSD lower than 2 Å), but in 75 of them, the pose with the lowest RMSD was not ranked as the best solution by rDock's intermolecular score (SCORE.INTER). This translates to a 64% success rate when the correct pose can be generated. Note that this is close to the 76% success rate obtained on the CCDC-Astex Diverse Set, a standard test set for binding mode prediction where correct predictions can be generated for 99% of cases [5].

Figure 6 shows the median RMSD obtained with the different receptors for the submitted pose and for the best pose generated by the pipeline. The median RMSD for the submitted pose was around 4.18 Å, whereas if we considered the best prediction, the mean decreased to 2.9 Å and the median to 2.4 Å. From these results, it is evident that the pipeline would benefit greatly from a complementary method to re-score the docking poses. An approach that presented better results in other blind challenges [24] was the combination of the docking scores with Dynamic Undocking (DUck) [25,26] simulations of the top-scoring poses. By combining both methods, we expected to be able to obtain a more accurate pose ranking for challenge submission.

2.5.3. Sampling Challenges

Cavity Selection

The CELPP Challenge is designed as a pose prediction challenge and to assess the influence of receptor choice in docking performance. For that reason, the coordinates for the centre of the cavity are provided by the organisers. Nonetheless, we wanted to go one step further by creating a pipeline of general applicability and add a cavity selection step to our protocol, thus avoiding the need to pre-define the binding site. The cavity detection is performed automatically by 3decision, and all the possible cavities are retrieved and considered for docking. The method that 3decision uses for cavity detection is fpocket, a pocket detection algorithm based on Voronoi tessellation [27]. When more than one cavity is detected, our pipeline selects the cavity based on the similarity of the ligands retrieved

by 3decision with the target ligand. On average, 3.2 cavities were detected per target, but in 67 cases (14%), the correct cavity was not detected, and so the docking was carried out in the wrong cavity. Figure 7 shows an example where 3decision only detected the cavity represented by the grey surface, missing the actual cavity represented by the green surface. In 9% of cases, the failure corresponded to shallow cavities on the protein surface that are not detected by the fpocket algorithm.

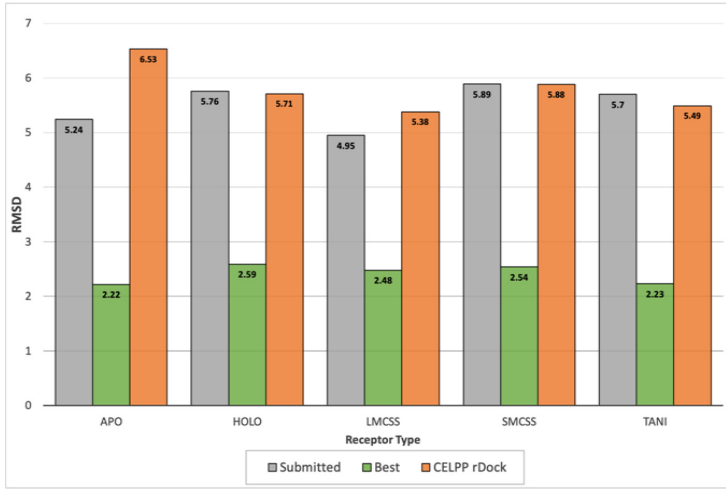


Figure 6. Median RMSD for the submitted pose compared to the best pose generated by the pipeline. The CELPP rDock workflow values are obtained from the D3R website (https://drugdesigndata.org/about/celpp2-charts accessed on 1 May 2021).

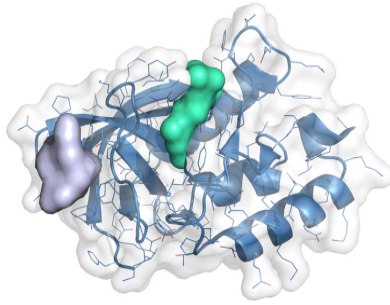


Figure 7. PDB 6ok9 with the pocket detected by 3decision represented by the purple surface and the correct pocket represented by the green surface.

Another reason for not detecting the cavity correctly (14% of cases) is that the ligands bind at the interface of a dimer, but only one protein is reported in the challenge. Note that, unlike other docking challenges or scenarios, the receptors provided by CELPP are not manually curated. They rely on a fully Automated Pipeline to perform that task, which can sometimes lead to the selection of inappropriate structures (e.g., giving a monomer instead of a dimer) for obtaining an accurate ligand pose [17]. Figure 8A shows one such example. The remaining failures in this category were attributed to an error with the API when downloading the analysis results.

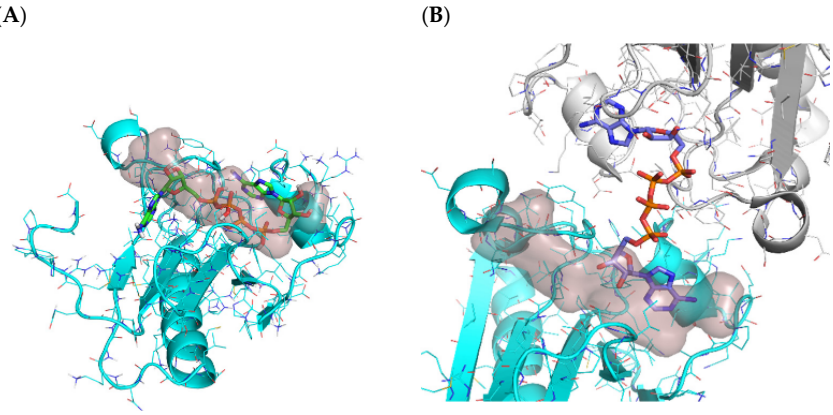


Figure 8. (**A**) hiTanimoto receptor for target 6j65. The solution selected by the pipeline is represented as sticks. (**B**) Protein dimer in PDB code 6j65. The crystalized ligand is represented as sticks. For both figures, the reference cavity provided by 3decision is shown as transparent surface.

Docking Method Selection

In our protocol we implemented three different docking strategies that were applied depending on the different set thresholds. From the 305 cases of the validation set where we did not obtain the correct pose, in 78 cases the correct binding pose had been correctly predicted by a different docking strategy.

As shown in Table **??**, from those 78 cases, only in 9 cases the correct solution was found by free docking instead of a form of guided docking. By contrast, 26 cases could have been correctly predicted if a form of guided docking had been used instead of free docking. This analysis also reveals that the two forms of guided docking employed here are not equivalent: 27 incorrect pharmacophore-guided docking solutions were correctly predicted by tethered docking. Vice versa, 16 incorrect tethered docking solutions were correctly predicted by pharmacophore-guided docking. One such example is shown in Figure 9. These results suggest that all the binding poses generated by the different docking protocols should be considered, then rescored with a post-docking method to identify the best one [28].

Table 4. Comparison between the submitted docking method vs. the method that yields the best result.

		Best Prediction		
		Free	**Ph4**	**Tethered**
Submitted	Free		6	20
	Ph4	8		27
	Tethered	1	16	

Receptor Flexibility

As pointed out by many previous studies [29], receptor flexibility is an important factor that can alter docking predictions. Both small changes on side-chain orientation and bigger structural changes can lead to incorrect predictions [30]. We could attest to this phenomenon when docking against the different proposed receptors. For each target, the docking protocol was run using all the receptors provided by the organisers. Figure 6 displays the validation results categorised by the receptor. The best-performing receptor was LMCSS, which corresponds to the one hosting the ligand most similar to the query. SMCSS obtained the worst results, with a median RMSD of 5.9 Å.

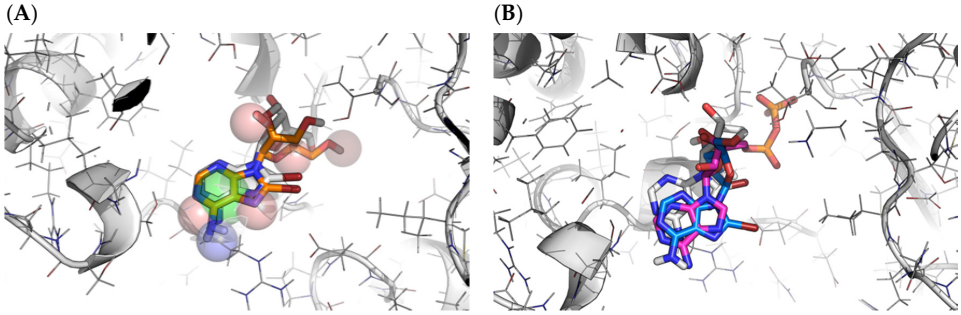


Figure 9. Predictions for PDB 6dfo and hiTanimoto Receptor using: (**A**) Pharmacophoric restraints. Predicted pose in orange. Pharmacophore represented as spheres. (**B**) Tether docking. Predicted pose in blue. Reference ligand in pink. In both cases, the crystallographic solution is shown in white for reference. The RMSD values with the predicted poses are 1.2 Å and 3.3 Å, respectively.

As an example, Figure 10 shows two cases where the differences in side-chain orientation of residues from the binding site are interfering with the correct binding position. In the case of 6pl1 (Figure 10A), there is a difference in the conformation of a loop in the binding site of all the receptors used that cause Phe 669 (in blue) to block part of the binding site obtaining a totally different cavity. It is established that, by using a variety of receptor conformations, we increased the probability of generating a correct ligand pose, but selecting the optimal docking cavity remains a major challenge for docking methods [31,32]. This result also highlights the need to select multiple binding mode predictions, which should be re-scored with a more rigorous computational methodology.

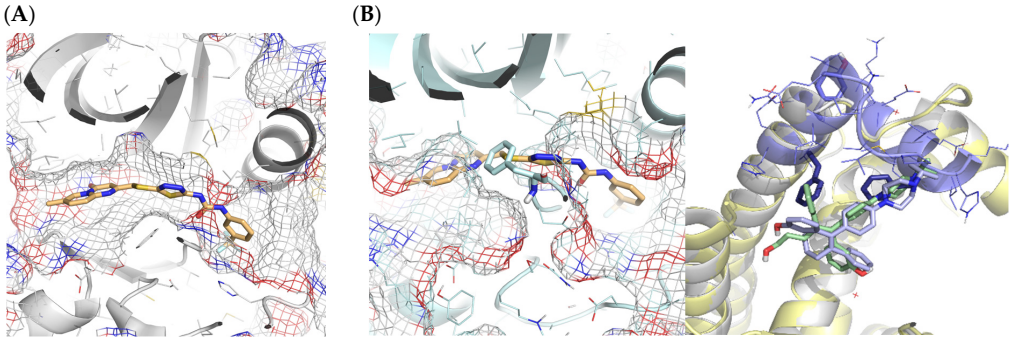


Figure 10. (**A**) Differences in binding site structure organisation between 6pl1 crystal and the selected hiTanimoto receptor by CELPP; the correct ligand pose is represented in beige, (**B**) Differences in site conformations for target 6a6k between receptor hiResHolo in purple, the crystal structure in white and hiTanimoto receptor in yellow. The ligand crystal pose is represented in green and in light purple is the pose obtained using the hiResHolo receptor.

Other Molecules in the Binding Site

This pipeline was intended for general applicability, and for this reason, during the cavity preparation process all the ligands and co-solvents were removed, and only the coordinates of the receptor were kept. However, in some systems, especially enzymes, cofactors can have an important role in determining the ligand binding mode. Two such examples are provided in Figure 11. Lastly, the fact that there can be other molecules in the binding site can interfere when generating the pharmacophoric restraints. As they are in the same cavity, our protocol included them in the list of retrieved ligands from similar proteins, and those are considered in the pharmacophoric restraint generation pipeline.

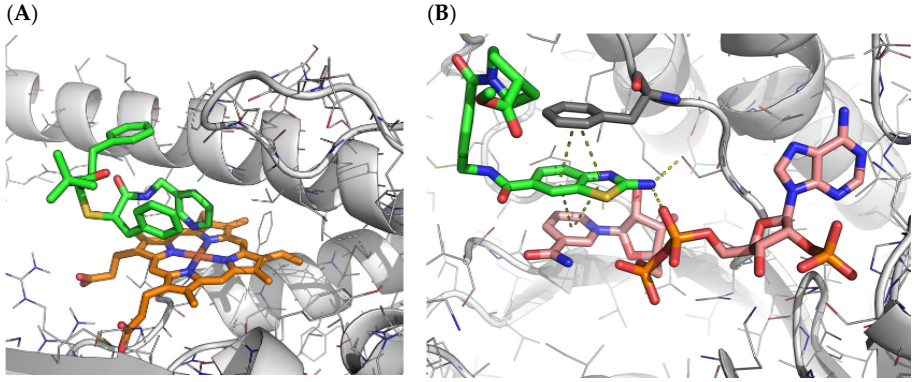


Figure 11. (**A**) Interaction of ligand G0D (green) with heme group (orange) in PDB 6DA2 [33].Ligand belongs to a series of analogues with pyridine as a heme-ligating head that works as an inhibitor of CYP3A4 by decreasing the heme reduction rates [33]. (**B**) Interaction of ligand EV8 (green) and NADP (pink) in PDB 6gd0. In yellow dashed lines are H-bond interactions and in green dashed lines π interactions.

3. Materials and Methods

3.1. Candidate Preparation

For each candidate structure, co-crystallized solvent and ligands were removed using Schrödinger's split structure tool [34], and only the coordinates of the receptor were kept. Subsequently, the protein preparation tool from MOE [35] was used to fix problems within the crystal structure, and the Protonate 3D tool [36] was used to assign protonation states to the protein (assuming pH 7.0). All the files were saved in Tripos MOL2 format, as required by the docking program, rDock [5]. All the above steps were integrated in an SVL script for automation purposes.

3.2. Ligand Preparation

We took the query ligand in SMILES string format and used the LigPrep tool from Schrödinger [37] to calculate the 3D structure with the proper topology; tautomerism; bond orders and geometry of bonds, angles, dihedrals and rings. Additionally, the ionizable groups were protonated at pH 7 with a threshold of $\pm$1 pH unit. All ligands were saved in SDF format.

3.3. Selection of Similar Proteins, Druggable Pockets and Ligand Retrieval

One of the pillars of the whole process was being able to select good reference systems from which we could extract some restraints to guide our docking predictions. For this purpose, we integrated into the pipeline a protocol based on the 3decision tool from Discngine; 3decision [17] is a web-based platform that centralizes all structural knowledge (including all the RCSB PDB dataset) to perform multiple kinds of analyses. We queried 3decision using a dedicated REST API endpoint. Using as input the target sequence in FASTA format, a blast against the database was performed to select those proteins that share a high identity (I% > 80%). The 3decision database also contains all pre-computed druggable pockets as predicted by the fpocket cavity detection tool [27]. The pockets are aligned based on the sequence and superimposed to the query structure. Finally, we exported all the ligands found in the aligned pockets in an SDF file, which was also converted to SMILES format using Openbabel [38]. In the case where multiple druggable pockets were detected, the corresponding docking protocol was applied to every pocket.

3.4. Ligand Similarity and Maximum Common Substructure Calculation

After retrieving the ligands found in similar pockets, a similarity analysis was performed between the query ligand and the list of retrieved ligands using MACCS keys fingerprints and the Tanimoto coefficient scoring, which has been identified as one of best metrics for similarity calculations [39]. The Tanimoto coefficients as well as the fingerprints were calculated using rdkit [19].

The maximum common substructure (MCSS) between the target ligand and the ligands retrieved from similar proteins was calculated using RDKit's FindMCS function [19]. As a complementary measure of similarity between the ligands, and also working as a method to evaluate the robustness of the MCSS, a Tanimoto coefficient based on MCSS was calculated using Equation (1) [22].

$$Tanimoto_{MCSS} = \frac{N_{AB}}{(N_A + N_B) - N_{AB}} \quad (1)$$

where N_A and N_B are the number of heavy atoms in molecules A and B, respectively, and N_{AB} is the number of heavy atoms in the MCSS. The $Tanimoto_{MCSS}$ can have values between 0 and 1, 1 being the value obtained when two molecules are identical.

3.5. Generation of Pharmacophoric Restraints

Ligand-based pharmacophore modelling has had a great impact in drug discovery [40]. In this work, this strategy was used to extract common chemical features from the aligned ligands retrieved by 3decision before elucidating the pharmacophores. The Align-it tool from Silicos-it [41] was used to generate a combination of pharmacophore points for each molecule in the set. In this work two different versions of the protocol for the generation of a consensus pharmacophore were tested. In the first version, after the generation of the pharmacophoric points for each molecule, the features that were common between molecules were selected and ranked by number of appearances, and then the two highest ranked features were selected and used as mandatory pharmacophoric restraints for docking. In the second Version, the ligands were first clustered based on similarity (MACCS fingerprints and Tanimoto similarity of 0.9). From each cluster, the ligand corresponding to the centroid was selected, thus removing redundancy and obtaining a diverse set of ligands, and then the pharmacophoric points were generated. From here, only the most-representative points (those shared by more than 45% of the ligands) were considered as mandatory restraints. Points shared by between 20% and 44% of the ligands were considered optional restraints. For the optional restraints, at least one of them needed to be fulfilled during the docking process.

3.6. Molecular Docking

To perform all the docking processes, we used rDock [5], a fast, versatile and open-source docking program. To run rDock, we needed the prepared receptor structure and a definition of the binding site. To define the binding site in this work, we chose the reference ligand method with rDock's default parameters. From the pool of retrieved ligands, we selected as a reference ligand the one having the maximum sum of the MACCS Tanimoto similarity score and $Tanimoto_{MCSS}$ score. This combined score implies a similar ligand and also a similar size to the target ligand. As a result, the cavity size was adapted to the query ligand, adding another restriction level to the docking process.

After ligand preparation, rDock is able to explore exocyclic bond rotations on the fly using a genetic algorithm together with rotations and translations. Conveniently, rDock can perform free docking as well as different types of restraint docking. Using rDock capabilities, our pipeline could use three different docking protocols, depending on the characteristics of the system and the available information. If we found a good reference ligand ($Tanimoto_{MCSS} > 0.5$), then the pipeline would choose tethered docking, fixing the MCSS with the *sdtether* utility. Otherwise, if there was a sufficient number of diverse ligands to extract a pharmacophore (>5), a pharmacophoric restraint docking was chosen

instead. Finally, unrestrained docking was used for the remaining cases. All the docking predictions used the standard rDock docking protocol (*dock.prm*).

3.7. Pose Selection

The output from the pipeline was a set of poses generated by the docking program for each candidate structure in an SDF file. Then, the poses were sorted by rDock's intermolecular score (SCORE.INTER), which accounts for the protein–ligand interaction's free energy. Formally, solutions should be sorted based on the total score, which accounts for the intramolecular energy as well (SCORE.INTRA + SCORE.INTER), but it has been shown that the intramolecular term bears a large error and can introduce more noise than signal to the predictions [42]. Using *sdsort*, the best pose was selected and saved in an SDF file. If more than one cavity was detected, this selection protocol was then applied to each cavity. Thereafter, the cavities were ranked based on the MCSS score obtained during the Ligand similarity and MCSS calculation, and then the best poses from each cavity were ranked by rDock's SCORE.INTER. The best scoring pose from the top scoring pocket was then selected for submission. Finally, the files were transformed to the format required by CELPP submission rules: the ligand pose in MOL format and the receptor in PDB format.

4. Conclusions

Quantifying the performance of docking software in real scenarios is essential to understanding their limitations, managing expectations and guiding future developments. With the CELPP Challenge, D3R aimed to provide a fast-growing validation set that better captures all the complexity in a real drug-discovery setting. Here we presented an initial version of our pipeline for participation on the CELPP Challenge, which applies different knowledge-based docking approaches depending on the already available information on PDB.

To provide a baseline performance, the CELPP team developed four workflows based on different docking programs, one being rDock. The rDock workflow represents a default implementation of the method without any optimisation and using the cavity defined by the challenge. Our protocol had the added challenge of detecting the cavity automatically, but when we considered only the cases where the cavity was correctly predicted, we observed a significant performance of our protocol relative to the baseline, with improvements in the median RMSD value ranging from 1.0 Å to 2.6 Å, depending on the docking cavity (Figure 6). This confirms that gathering information from already-deposited complexes in PDB and transforming them into the appropriate restraints benefits the docking process greatly.

Our final goal was to evolve this platform into a docking server where more rigorous, but also more computationally demanding methods, could be applied (e.g., molecular dynamics). Nonetheless, there are some additional points that need to be revised. The first one is cavity detection and characterization. For our pipeline being able to identify possible binding sites for the majority of targets, 3decision has proven to be a valuable tool. However, there are some cases where the 3decision protocol is not able to retrieve the correct pocket because they are shallow cavities or the receptor structure is ill-defined. In this first version of the pipeline, targets where there is no pocket information are neglected, and no docking protocol is applied. For these situations, we could use a local implementation of fpocket [38] to check whether there are, in fact, no possible druggable cavities. Another option would be using molecular dynamics with co-solvent/water mixtures (MDmix) [43,44] to identify possible binding sites. Nonetheless, we would like to add the option of taking the cavity coordinates as a reference. With this, we would separate the cavity-finding problem from the docking problem, reduce execution time and increase the predictive power when the binding site is already known.

A second point to revisit is the choice of receptor structure. As discussed, protein flexibility is an important aspect to consider in a drug-discovery setup. Proteins can adapt their structures to the bound ligand, so using an apo structure or one in a complex with a very different compound degrades the performance of the docking program. One way to

mitigate this effect would be to use different conformations of the receptor and select the one with the better score as the optimal structure [45].

A third aspect is the management of 'third-party' molecules in the binding site, namely cofactors and water molecules. In this initial version of the pipeline, all systems are processed and prepared in the same way, stripping the binding site of all non-protein molecules. However, we detected several cases where docking failed owing to missing cofactor molecules that should be considered part of the receptor. This can be solved with a curated list of cofactors that should not be removed. Water molecules are frequently found at the protein–ligand interface, mediating hydrogen bonds between the partners. By keeping these structural waters on the binding site, the ligand pose predictions can be more accurate.

We will also continue to monitor the performance of restrained and unrestrained docking in prospective CELPP predictions. As previously shown, by using the MCSS score, we are able to determine which is the docking method that performs best for each case. Initially, we applied a rather restrictive cutoff of 0.65, which included only 13% of the total cases. After considering all the participation cases, we were able to determine better ranges of applications for each type of docking protocol, which presently is set to 0.5 and includes 31% of cases.

As far as the creation of the pharmacophores, in cases where, due to a lack of pre-existing information when ligand-based pharmacophore cannot be extracted, we could make use of hot spots derived from the structure. Such hot spots can be identified by their ability to bind small organic co-solvents [43,46]. By performing molecular dynamics with co-solvent/water mixtures (MDmix), we can identify binding sites and hot spots [47] that could be used as pharmacophoric restraints for docking. The addition of this methodology to our workflow would also allow us to assess the druggability of the pockets selected by 3decision.

Author Contributions: Conceptualization, X.B., S.R.-C. and M.M.-L.; methodology, M.M.-L.; software, D.A.-G., P.S. and M.M.-L.; validation, M.M.-L.; formal analysis, M.M.-L.; investigation, M.M.-L.; resources, D.A.-G. and P.S.; data curation, M.M.-L.; writing—original draft preparation, M.M; writing—review and editing, X.B. and D.A.-G.; visualization, M.M.-L.; supervision, X.B. and S.R.-C.; project administration, X.B.; funding acquisition, X.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (RTI2018-096429-BI00) and the Catalan government (2014SGR1189).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the systems mentioned in this paper were provided by the CELPP organizers. Data regarding the CELPP challenge can be found in https://drugdesigndata.org/about/celpp2-charts (accessed on 31 March 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yu, W.; MacKerell, A.D. Computer-aided drug design methods. In *Antibiotics*; Humana Press: New York, NY, USA, 2017; pp. 85–106.
2. Yuriev, E.; Holien, J.; Ramsland, P.A. Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. *J. Mol. Recognit.* **2015**, *28*, 581–604. [CrossRef] [PubMed]
3. Śledź, P.; Caflisch, A. Protein structure-based drug design: From docking to molecular dynamics. *Curr. Opin. Struct. Biol.* **2018**, *48*, 93–102. [CrossRef] [PubMed]
4. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749. [CrossRef] [PubMed]

5. Carmona, S.R.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A.B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R.E.; Morley, S.D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10*, e1003571.
6. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748. [CrossRef]
7. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef]
8. Perola, E.; Walters, W.P.; Charifson, P.S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins Struct. Funct. Bioinform.* **2004**, *56*, 235–249. [CrossRef]
9. Chen, H.; Lyne, P.D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2005**, *46*, 401–415. [CrossRef]
10. Warren, G.L.; Andrews, C.W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2005**, *49*, 5912–5931. [CrossRef]
11. Cross, J.B.; Thompson, D.C.; Rai, B.K.; Baber, J.C.; Fan, K.Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474. [CrossRef]
12. Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: The prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.* **2016**, *18*, 12964–12975. [CrossRef]
13. Corbeil, C.R.; Williams, C.I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput. Mol. Des.* **2012**, *26*, 775–786. [CrossRef]
14. Hartshorn, M.J.; Verdonk, M.L.; Chessari, G.; Brewerton, S.C.; Mooij, W.T.; Mortenson, P.N.; Murray, C.W. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741. [CrossRef]
15. Warren, G.L.; Do, T.; Kelley, B.P.; Nicholls, A.; Warren, S.D. Essential considerations for using protein–ligand structures in drug discovery. *Drug Discov. Today* **2012**, *17*, 1270–1281. [CrossRef] [PubMed]
16. Wagner, J.R.; Churas, C.P.; Liu, S.; Swift, R.V.; Chiu, M.; Shao, C.; Feher, V.A.; Burley, S.K.; Gilson, M.K.; Amaro, R.E. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *Structure* **2019**, *27*, 1326–1335.e4. [CrossRef] [PubMed]
17. Le Roux, E.; Schmidtke, P. 3Decision (Version 2021.3.1) [Computer Software]. Discngine. Available online: https://3decision.disngine.cloud (accessed on 31 March 2022).
18. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
19. Landrum, G.; Kelley, B.; Tosco, P.; Vianello, R.; Turk, S.; Swain, M.; Pahl, A.; Fuller, P.; Wójcikowski, M.; Sforna, G.; et al. rdkit/rdkit: 2016_09_4 (Q3 2016) Release. 2017. Available online: https://zenodo.org/record/268688#.Ymc9o9pByUk (accessed on 5 February 2017).
20. Heller, S.R.; McNaught, A.; Pletnev, I.V.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Chemin.* **2015**, *7*, 1–34. [CrossRef]
21. Khafizov, K.; Madrid-Aliste, C.; Almo, S.C.; Fiser, A. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3733–3738. [CrossRef] [PubMed]
22. Boström, J.; Hogner, A.; Schmitt, S. Do Structurally Similar Ligands Bind in a Similar Fashion? *J. Med. Chem.* **2006**, *49*, 6716–6725. [CrossRef] [PubMed]
23. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594. [CrossRef]
24. Ruiz-Carmona, S.; Barril, X. Docking-undocking combination applied to the D3R Grand Challenge 2015. *J. Comput. Mol. Des.* **2016**, *30*, 805–815. [CrossRef]
25. Carmona, S.R.; Schmidtke, P.; Luque, F.J.; Baker, L.; Matassova, N.; Davis, B.; Roughley, S.; Murray, J.; Hubbard, R.; Barril, X. Dynamic undocking and the quasi-bound state as tools for drug discovery. *Nat. Chem.* **2016**, *9*, 201–206. [CrossRef]
26. Majewski, M.; Barril, X. Structural Stability Predicts the Binding Mode of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 1644–1651. [CrossRef]
27. Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinform.* **2009**, *10*, 1–11. [CrossRef] [PubMed]
28. Varela-Rial, A.; Majewski, M.; de Fabritiis, G. Structure based virtual screening: Fast and slow. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *12*, 1–17. [CrossRef]
29. Feixas, F.; Lindert, S.; Sinko, W.; McCammon, J.A. Exploring the role of receptor flexibility in structure-based drug discovery. *Biophys. Chem.* **2013**, *186*, 31–45. [CrossRef]
30. Kumar, A.; Zhang, K.Y.J. A cross docking pipeline for improving pose prediction and virtual screening performance. *J. Comput. Mol. Des.* **2017**, *32*, 163–173. [CrossRef]
31. Barril, X.; Morley, S.D. Unveiling the Full Potential of Flexible Receptor Docking Using Multiple Crystallographic Structures. *J. Med. Chem.* **2005**, *48*, 4432–4443. [CrossRef] [PubMed]

32. Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2009**, *50*, 186–193. [CrossRef]
33. Samuels, E.R.; Sevrioukova, I. Structure–Activity Relationships of Rationally Designed Ritonavir Analogues: Impact of Side-Group Stereochemistry, Headgroup Spacing, and Backbone Composition on the Interaction with CYP3A4. *Biochemistry* **2019**, *58*, 2077–2087. [CrossRef] [PubMed]
34. Schrödinger, L. Small-Molecule Drug Discovery Suite 2018-1. New York, NY, USA, 2018. Available online: https://www.macinchem.org/blog/files/1ed80631e38d91494a9921f6344cac55-1411.php (accessed on 31 March 2022).
35. *Molecular Operating Enviroment*; MOE 2006.08; Chemical Computing Group: Montreal, QC, Canada, 2006.
36. Labute, P. *Protonate 3d: Assignment of Macromolecular Protonation State and Geometry*; Chemical Computing Group Inc.: Montreal, QC, Canada, 2008.
37. *LigPrep*; Version 3.0; Schrödinger: Mannheim, Germany, 2014.
38. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef] [PubMed]
39. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20. [CrossRef] [PubMed]
40. Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discov. Today* **2010**, *15*, 444–450. [CrossRef] [PubMed]
41. Taminau, J.; Thijs, G.; De Winter, H. Pharao: Pharmacophore alignment and optimization. *J. Mol. Graph. Model.* **2008**, *27*, 161–169. [CrossRef]
42. Tirado-Rives, J.; Jorgensen, W.L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein−Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880–5884. [CrossRef] [PubMed]
43. Alvarez-Garcia, D.; Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **2014**, *57*, 8530–8539. [CrossRef]
44. Seco, J.; Luque, F.J.; Barril, X. Binding Site Detection and Druggability Index from First Principles. *J. Med. Chem.* **2009**, *52*, 2363–2371. [CrossRef] [PubMed]
45. Novoa, E.M.; Pouplana, L.R.D.; Barril, X.; Orozco, M. Esemble docking from homology models. *J. Chem. Theory Comput.* **2010**, *6*, 2547–2557. [CrossRef]
46. Alvarez-Garcia, D.; Barril, X. Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design. *J. Chem. Theory Comput.* **2014**, *10*, 2608–2614. [CrossRef]
47. Bajusz, D.; Rácz, A.; Héberger, K. Comparison of data fusion methods as consensus scores for ensemble docking. *Molecules* **2019**, *24*, 2690. [CrossRef]

Lenalidomide stabilizes protein-protein complexes by turning labile intermolecular H-bonds into robust interactions.

Marina Miñarro-Lleonar,[1,2,3] Andrea Bertran-Mostazo[1, 3], Jorge Duro,[1,2] Xavier Barril,[1,2,3,4] Jordi Juárez-Jiménez,[1,2]**

[1]Unitat de Fisicoquímica, Departament de Farmàcia i Tecnologia Farmacéutica, i Fisicoquímica. Facultat de Farmacia I Ciències de l'Alimentació. Universitat de Barcelona (UB). Av. Joan XXIII, 27-31, 08028 Barcelona, Spain.

[2]Institut de Química Teòrica i Computacional (IQTC), Universitat de Barcelona (UB), Barcelona, Spain.

[3]Institut de Biomedicina, Universitat de Barcelona (UB), Barcelona, Spain

[4] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Spain.

Abstract

Targeted protein degradation (TPD) is emerging as a very promising strategy to modulate protein activities in several diseases, spearheaded by anti–myeloma drugs lenalidomide and pomalidomide. It has been recently demonstrated that the mechanism of action of these drugs involves the increased degradation of several proteins, including the transcription factors Ikaros and Aiolos as well as the enzyme Caseine Kinase 1α (CK1α). It has been shown that lenalidomide and pomalidomide are able to stabilize the complex between the E3 ligase Cereblon ($CRL4^{CRBN}$) and the aforementioned proteins, while, remarkably, the stability of the protein-protein interaction is very low. Even though the structures for these complexes have

been determined, there are no evident interactions that can account for the high formation efficiency of the ternary complex. In this work, we have leveraged Molecular Dynamics to shed light into the molecular determinants underlying the stabilization effect exerted by lenalidomide in the complex between $CRL4^{CRBN}$ and CK1α. Furthermore, we evaluated the effect that different mutations of CK1α in the stability of the ternary complex $CRL4^{CRBN}$–lenalidomide–CK1α and provide a thermodynamic and kinetic rational for the stabilization effect. These results pave the way to further understand cooperativity effects in drug–induced protein–protein complexes and could help in the future design of improved targeted molecular degraders.

Introduction

The concept of molecular glues (MGs) was introduced by Zheng and co–workers[1] to describe the stabilizing effect of the plant hormone auxin on several complexes with the SCF^{TIR1} ubiquitin ligase complex. Subsequently, it has been revealed that this mechanism is quite common in nature[2–5] and that even a number of widely used drugs such as the immunosuppressant drug Cyclosporin A[6] or the anti–cancer agents Paclitaxel[7] or Indisulam[8] share a similar mechanism of action. These findings have spured interest in leveraging selective stabilization of protein–protein interaction in drug discovery. However, it has been repeatedly noted in the literature[9–11] that MGs discovery is too reliant on serendipity and that the future development of successful MGs as therapeutic agents ought to shift into more rational approaches and further understanding of the molecular mechanisms underpinning the ligand–induced stabilization of protein–protein interactions. Contrary to traditional drug discovery, which focuses on the formation of binary complexes, rational development of MGs will require detailed understanding of the formation of ternary complexes, which often imply non–additive mechanisms.[12] The physicochemical factors underlying these mechanisms are usually difficult to anticipate from structural analysis or common Computer Aided Drug Design protocols such

as docking or Virtual Screening.[3] Nonetheless, they are critical to the selective stabilization of protein–protein complexes, and must be understood to fully exploit the therapeutic opportunities offered by MGs.

A landmark example of the potential of MGs to impact human health is provided by thalidomide derivatives lenalidomide and pomalidomide, so called IMiDs, widely used in the treatment of multiple myeloma. Only recently it was described that these molecules induce the ubiquitination and degradation of the transcription factors Ikaros (IKZF1) and Aiolos (IKZF3)[13] and the enzyme Casein Kinase 1α (CK1α)[14] by stabilizing the complex of these proteins with the E3 ligase CRBN, which is the substrate receptor of the CUL4–RBX1–DDB1 ubiquitin ligase complex (CRL4). IMiDs are accommodated in a tryptophan cage in the substrate binding domain of CRBN[15] and structural evidence has shown that IKZF1,[16] CK1α[17] and other proteins[18] bind to the CRBN–IMiD interface, establishing a set of protein–protein interactions through a β–hairpin loop structure that contains a Gly residue on the apex.[5,16,17] In a recent work, Cao et. al.[19] estimated that pomalidomide stabilizes the IKZF1–CRBN complex by around fourfold, while lenalidomide stabilizes the CK1α–CRBN complex by around 30–fold. The authors also proposed that instead of creating new sets of interactions, MGs in general, and IMiDs in particular, must be able to stabilize pre–existing protein–protein interactions. Analysis of the structural data available seems to support this hypothesis, as the direct intermolecular interaction between lenalidomide–CK1α[17] and pomalidomide–IKZF1[16] are rather unremarkable and thus, cannot account for the increase in stability of the ternary complex. These observations highlight the importance of the non–additive mechanism at play in these interactions.[12] In this work, we use biomolecular simulations to evidence that the stabilization effect exerted by lenalidomide in the complex between CRBN and CK1α relies on the its ability to increase the structural stability of three key H–bonds at the CRBN–CK1α interface. Using data for four different mutants of CK1α we demonstrate that the robustness of

these three H–bonds directly correlates with the stability of the ternary CRBN–lenalidomide–CK1α complex, even when mutations do not directly disturb the ability of either protein to establish these interactions. The underlying mechanism is proposed to depend on the capacity of lenalidomide to provide hydrophobic shielding to pre–existing protein–protein hydrogen bonds, thus increasing the structural, kinetic and thermodynamic stability of the complex. We anticipate that this may be a general mechanism that can be exploited for the future rational development of MG.

Results

Presence of lenalidomide results in stronger H–bond interactions at the CRBN–CK1α interface.

Examination of the protein–protein interface of the CRBN–lenalidomide–CK1α complex reveals that there are three protein–protein hydrogen bonds between the 36–42 β–hairpin loop of CK1α and the C–terminal domain of CRBN (Supplementary Figure S1). Namely, the side–chains of CRBN residues Asn351, His357 and Trp400 engage the backbone carbonyl oxygens of the CK1α residues Ile37, Thr38 and Asn39 respectively. These interactions hereafter referred to as $CRBN^{Asn351}$–$CK1\alpha^{Ile37}$, $CRBN^{His357}$–$CK1\alpha^{Thr38}$ and $CRBN^{Trp400}$–$CK1\alpha^{Asn39}$ respectively, have been demonstrated to be key for the recruitment of CK1α by CRBN.[17] However, there is no evident factor precluding the formation of these interactions in the absence of lenalidomide, which is in line with the hypothesis of stabilization of pre–existing protein–protein complexes put forward by Cao and co–workers. Previous works have shown that most stable receptor–ligand complexes display, at least, one robust and hard–to break intermolecular H–bond,[20–22] and the importance of these interactions has also been highlighted as a main player in protein structural stability.[23,24] Therefore, we investigated the energetic cost

of independently breaking each of the H–bonds identified at the CRBN–CK1α interface (Figure 1) combining Steered Molecular Dynamics and the Jarzynski's equality.[25,26]

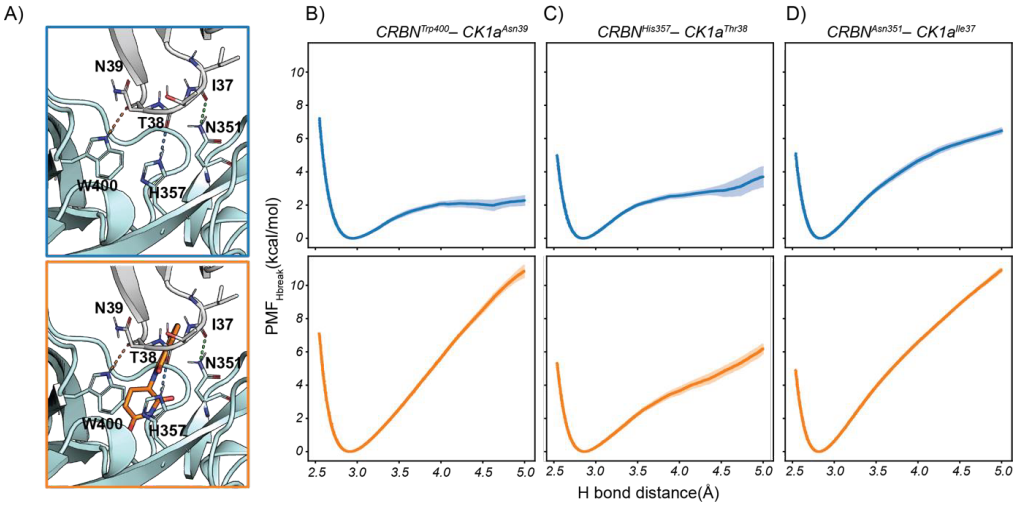


Figure 1: H–bond dissociation energy profiles in the presence and absence of lenalidomide at the CRBN–CK1α interface. **A**. Detailed view of the CK1α–CRBN dimeric interface (top) and its ternary complex with lenalidomide (bottom). **B.** Energy profile of the $CRBN^{Trp400}$– $CK1\alpha^{Asn39}$ H–bond in the absence (top) and presence (bottom) of lenalidomide. **C.** Energy profile of the $CRBN^{His357}$ – $CK1\alpha^{Thr38}$ H–bond in the absence (top) and presence (bottom) of lenalidomide. **D.** Energy profile of the $CRBN^{Asn351}$ – $CK1\alpha^{Ile37}$ H–bond in the absence (top) and presence (bottom) of lenalidomide.

Although convergence of sampling is usually a concern when applying the Jarzynski relationship, we consider that the reduced number of degrees of freedom that the system may access during sampling of the rupture of a given H–bond (where donor and acceptor are pulled apart from 2.5 Å to 5 Å, *vide infra*) will allow to calculate Potentials of Mean Force along the separation distance between donor and acceptor (hereafter referred to as PMF of H–bond breakage or PMF_{HB_break}) of sufficient accuracy as to distinguish strong from weak H–bond interactions, similarly to what us and others have previously reported in the literature for other systems.[21,27,28] By comparing the values of PMF_{HB_break}, we established that, in the absence of lenalidomide, the stronger H–bond is $CRBN^{Asn351}$–$CK1\alpha^{Ile37}$ (PMF_{HB_break} = 6.4 +/- 0.1

kcal/mol) followed by the $CRBN^{His357}$–$CK1\alpha^{Thr38}$ interaction (PMF_{HB_break} = 3.7 +/- 0.6 kcal/mol) and the $CRBN^{Trp400}$–$CK1\alpha^{Asn39}$ interaction (PMF_{HB_break} = 2.3 +/- 0.3 kcal/mol). The presence of lenalidomide at the interface causes a large increase in the energy necessary to break the three H–bonds, with estimated PMF_{HB_break} of 10.9 +/- 0.1, 6.3 +/- 0.3 and 11.0 +/- 0.4 kcal/mol for the $CRBN^{Asn351}$–$CK1\alpha^{Ile37}$, $CRBN^{His357}$–$CK1\alpha^{Thr38}$ and $CRBN^{Trp400}$–$CK1\alpha^{Asn39}$ interactions respectively (Table 1). We examined the 3D–structure of the ternary complex to obtain clues about the stabilization of the investigated H–bonds. The only direct H–bond between lenalidomide and CRBN (the carbonyl group of the oxoisoindol moiety with the side–chain of Asn351) is insufficiently connected to the protein–protein H–bonds to suggest that it can cause a concerted change in the interaction network. Instead, the increased stability may be explained by the change of local environment around the H–bonds. Indeed, it has been previously shown that incoming water molecules catalyze the rupture of solvent exposed H–bonds, by decreasing the energetic barrier required to bring apart donor and acceptor.[20,29] Based on these observations, we hypothesized that the main role of lenalidomide will be to create a hydrophobic environment around the protein–protein interface that effectively shields the H–bonds from incoming water molecules.

Reinforced H–bonds display increased hydrophobic shielding at the CRBN–CK1α interface.

To probe our hypothesis that lenalidomide stabilizes the CRBN–CK1α complex mainly by hydrophobic shielding effects, we studied the changes on the local environment of the three key H–bonds at the interface upon binding of the MG. The radial distribution function (RDF) provides the average number of water molecules found around a certain atom with respect to what would be expected on the bulk solvent during the course of a Molecular Dynamics simulation.

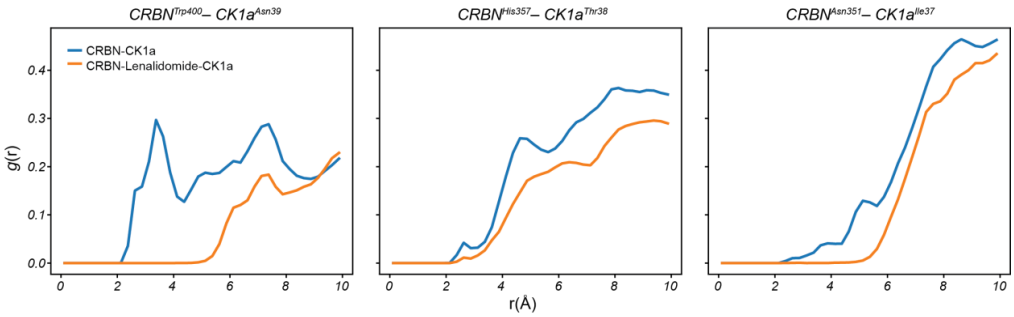


Figure 2: Radial Distribution Function (RDF) of water molecules around the backbone carbonyl oxygen of CK1α involved on the key CRBN–CK1α H–bonds. The blue line represents values for the two–body complex CRBN–CK1α and the orange represents values for the three–body complex CRBN–lenalidomide–CK1α

Therefore, it can be used as a proxy to estimate the solvent exposure of certain atoms or residues. We determined the RDF of the backbone carbonyl groups of CK1α in molecular dynamics of both the CRBN–CK1α and CRBN–lenalidomide–CK1α complexes (Figure 2) in which the H–bond distances for the three key interactions was kept between 2.5 and 3.5 Å using flat bottom restraints (see the methods section for further details). As expected for atoms at the interface of the protein–protein complex, their water exposure is relatively low. However, there was a noticeable reduction on the RDF around the backbone carbonyl of Asn39 (from 0.3 in the binary complex to 0 in the ternary complex). Furthermore, the RDF around the $CRBN^{Asn351}$ – $CK1\alpha^{Ile37}$ and the $CRBN^{Trp400}$– $CK1\alpha^{Asn39}$ H–bonds for radii below 5 Å drops to 0 in the presence of lenalidomide. On the other hand, the reductions in the RDF for the $CRBN^{His357}$ – $CK1\alpha^{Thr38}$ H–bond – the interaction that is least reinforced by the presence of lenalidomide – is relatively minor.

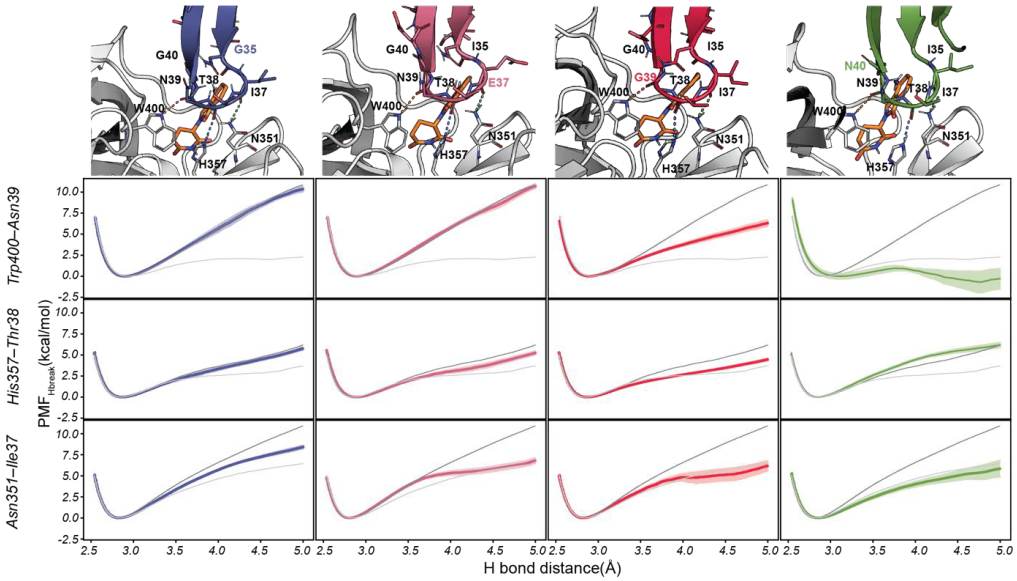


Figure 3: H–bond dissociation energy profiles in the presence and absence of lenalidomide at the CRBN–CK1α interface. Detailed view of the CRBN–lenalidomide–CK1α interface for ^{I35G}CK1α, ^{I37E}CK1α, ^{N39G}CK1α and ^{G40N}CK1α and associated PMF_{HB_break} profiles for CRBN^{Trp400}–CK1α^{Asn39} (Top), CRBN^{His357}–CK1α^{Thr38} (middle) and CRBN^{Asn351} –CK1α^{Ile37} (bottom). PMF_{HB_break} profiles for CRBN–lenalidomide–^{wt}CK1α (dark grey) and the CRBN–^{wt}CK1α (light grey) are included for reference

H–bond robustness correlates with the measured stability of CRBN–lenalidomide–^{MUT}CK1α complexes.

Petzold et. al. reported that the ternary complexes between CRBN–lenalidomide and CK1α mutants ^{I35G}CK1α, ^{I37E}CK1α, ^{N39G}CK1α and ^{G40N}CK1α displayed decreasing stability.[17] We therefore investigated whether the robustness of the H–bonds at the interface on these ternary complexes was diminished with respect to the CRBN–lenalidomide–CK1α complex (Figure 3 and Table 1).

Analysis of the energetic profiles of H–bond breakage showed that ^{N39G}CK1α and ^{G40N}CK1α displayed the greatest alterations, both with ΔPMF_{HB_break} in excess of 4 kcal/mol for the

CRBN^{Asn351}–CK1α^{Ile37} and CRBN^{Trp400}–CK1α^{Asn39} interactions. In the case of the CRBN^{His357}–CK1α^{Thr38} interaction, there was a lesser reduction in ^{N39G}CK1α (ΔPMF_{HB_break} = 1.7 kcal/mol) , while in the case of ^{G40N}CK1α the latter interaction was not affected. In fact, besides ^{N39G}CK1α, the effect of mutations on the CRBN^{His357}–CK1α^{Thr38} interaction was within the estimated uncertainty margins (ΔPMF_{HB_break} of 0.0 ± 0.6, -0.4 ± 0.5 and 1.0 ± 0.5 kcal/mol for the ^{G40N}CK1α, ^{I35G}CK1α and ^{I37E}CK1α mutants respectively). For the ^{I35G}CK1α and ^{I37E}CK1α mutants, only the CRBN^{Asn351} –CK1α^{Ile37} was significantly weakened with respect the *wt* complex, displaying ΔPMF_{HB_break} of 2.5 ± 0.3 kcal/mol and 4.1 ± 0.4 kcal/mol respectively. Therefore, all mutants displayed as or even more robust H–bonds, on average, than the complex between CRBN and CK1α without lenalidomide, but the profile of dissociation energy with respect to the *wt* ternary complex was weakened for at least one of the H–bonds in all the cases.

Table 1. Summary of the absolute and relative PMF_{HB_break} values (in kcal mol^{-1}) for the CRBN– CK1α systems considered in this work. Relative values (in parentheses) are showed with respect to the CRBN–LEN–CK1α complex. Error estimates were obtained by bootstrapping ten times the W profiles used to estimate the PMF.

H–bond (CRBN–CK1α)	**^{wt}CK1α**	**^{wt}CK1α No LEN**	**^{I35G}CK1α**	**^{I37E}CK1α**	**^{N39G}CK1α**	**^{G40N}CK1α**
Trp400–Asn39	10.9 ± 0.3	2.3 ± 0.3 (-8.6 ± 0.6)	10.3 ± 0.3 (-0.6 ± 0.6)	10.7 ± 0.3 (-0.2 ± 0.6)	6.3 ± 0.4 (-4.6 ± 0.7)	-0.4 ± 1.2 (-11.3 ± 1.5)
His357–Thr38	6.2 ± 0.3	3.7 ± 0.6 (-2.5 ± 0.9)	5.8 ± 0.2 (-0.4 ± 0.5)	5.2 ± 0.2 (-1.0 ± 0.5)	4.5 ± 0.1 (-1.7± 0.4)	6.2 ± 0.3 (0.0 ± 0.6)
Asn351–Ile37	10.9 ± 0.1	6.5 ± 0.2 (-4.4 ± 0.3)	8.4 ± 0.2 (-2.5 ± 0.3)	6.8 ± 0.3 (-4.1 ± 0.4)	6.2 ± 0.6 (-4.7 ± 0.7)	5.8 ± 1.0 (-5.1 ± 1.1)
$\sum PMF_{HB\_break}$	28.0 ± 0.7	12.4 ± 1.1 (-15.6 ± 1.8)	24.5 ± 0.7 (-3.5 ± 1.4)	22.7 ± 0.8 (-5.3 ± 1.5)	16.9 ± 1.1 (-11.1 ± 1.8)	12.4 ± 2.4 (-15.6 ± 3.1)

While there is no experimental value that can be linked directly with the calculated ΔPMF_{HB_break} for the breaking of singular H–bonds, we hypothesized that the observed variation in the energy required for breaking the three interactions at the CRBN–CK1α

interface may inform about the stability of the resulting ternary complex with lenalidomide. To probe this possibility, we first established that there was no co–dependence between the breakages of the three hydrogen bonds (Figure S2), and therefore, the energy required to break all three bonds could be approximated as the addition of the individual PMF_{HB_break} values. We found that the sum of PMF_{HB_break} for the three key H–bonds in each of the complexes between CRBN and CK1α was correlated with its estimated binding affinity ($R^2 = 0.94$, Figure 4).

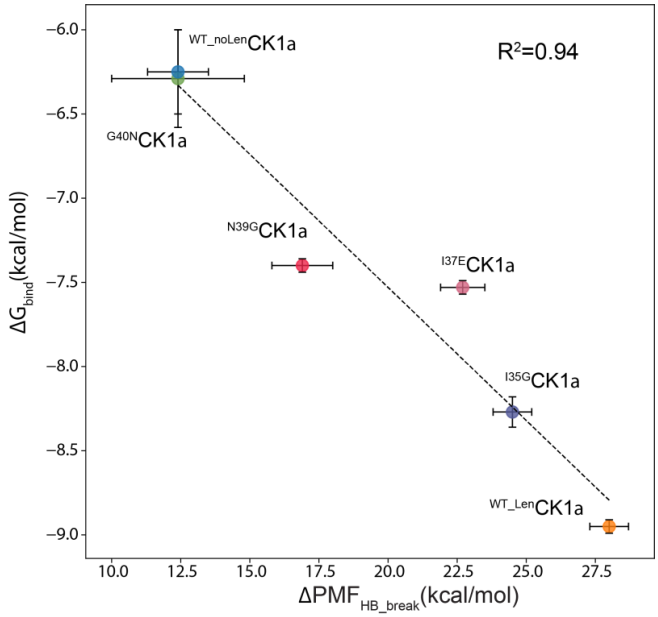


Figure 4: Correlation plot between the sum of the PMF_{HB_break} for the three H–bonds for different variants of CK1α with respect to ΔG_{bin} calculated from K_D estimations. The PMF_{HB_break} was taken at the end point of the PMF profile and error bars were obtained by bootstrapping of the W profiles. PMF_{HB_break} values are reported in table 1. The ΔG_{bin} was obtained by transforming the K_Ds fitted using data from reference[17] and error bars were obtained by error propagation. The reproduced [CRBN]–520/490 nm TR–FRET Ratio plot is provided in supplementary figure S7 and experimental values are provided in supplementary tables S1 and S2.

Weakening of the three H–bond interactions stems from better accessibility of water molecules to the protein–protein interface.

Having stablished the correlation between the strength of the hydrogen bonds at the CRBN–lenalidomide–CK1α interface and the stability of the ternary complex, we next investigated the molecular determinants that could account for the reduced strength of the hydrogen bonds displayed by the four single point CK1α mutants. First, we determined the RDF of water molecules around the H–bonds and compared them with the RDF profiles obtained for the binary and ternary complexes of CK1α (Supplementary Figure S3). All the mutants displayed RDF profiles closer to the ternary complex than to the binary complex. Nevertheless, the profiles obtained for the ternary complex involving the ^{N39G}CK1α mutant was very different that the one obtained for the wild type CK1α, with increased RDF values with respect to the latter in the areas of the first and second solvation shell for both the CRBN^{His357}–CK1α^{Thr38} and the CRBN^{Trp400}–CK1α^{Asn39} H–bonds, while the remaining H–bond (the furthest from the mutation point) only displayed differences beyond 6 Å. A similar pattern was observed on the profiles obtained for the ^{I35G}CK1α and ^{I37E}CK1α mutants, where the closest H–bond was the most affected by the change, although in these cases the differences were only observed on the second solvation shell region. In contrast with the stark decrease in PMF_{HB_break}, the profiles for the remaining mutant ^{G40N}CK1α were indistinguishable from the profiles of the ternary complex with the wild type CK1α. Intrigued by this apparent discrepancy, we visualized the trajectories and identified that, regardless of the system involved, low breaking profiles corresponded with those in which at least one water molecule entered the protein–protein interface from the bulk and established an H–bond with the carbonyl atom previously involved in the protein–protein interaction, while high work profiles corresponded with H–bond breakages in which water molecules did not access the protein–protein interface or did not stablish an H–bond. (Supplementary Movie S1). We hypothesized that the higher rate of access of water molecules to the protein–protein interface in the case of the ^{G40N}CK1α maybe related to a worse hydrophobic packing of lenalidomide's core against the bulkier and more flexible

Asn sidechain than against the Gly residue in position 40. We therefore measured the average distance between lenalidomide's centre of mass and the alpha carbon of residue 40 of CK1α in all the mutants and in the wild type (Figure S4). The average distance was estimated to be ca. 4.8 Å for all the systems but ^{G40N}CK1α, in which the average distance was closer to 5.8 Å.

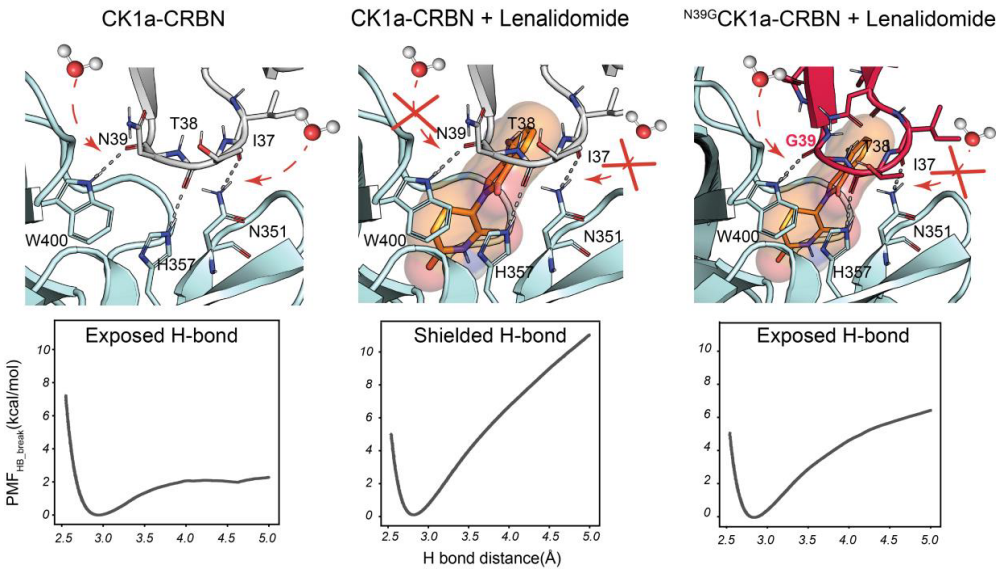


Figure 5: Proposed mechanism underlying the stabilization of the CRBN–CK1α complex by lenalidomide and the effect of mutations in the CK1α sequence. Lenalidomide hinders water accessibility to the CRBN–CK1α interface, increasing the strength of H–bonds. Mutations that alter water accessibility to the interface diminish the stability of the ternary complex.

Considering the results, we propose that lenalidomide (and by extension other IMiDs) enable the degradation of CK1α and other CRBN neo–substrates by strengthening the pre–existing H–bonds at the interface, which results in a complex stable enough as to be tagged by ubiquitination (Figure 5). The reinforcement of the H–bonds seems to be related to the ability of IMiDs to hinder access of water molecules to the protein–protein interface, and hence, their effectiveness is very susceptible to single point mutations that increase the flow of water into the interface, either by means of local or long–range effects.

Discussion

The work demonstrates that the presence of lenalidomide at the CRBN–CK1α interface results in a significant increase in the free energy required to break three key H–bond interactions at the protein–protein interface (Figure 1), as well as highlighting the sensitivity of this effect to point mutations of one of the partners, even when these mutations do not directly hinder the formation of the H–bonds (Figure. 3 and Table 1). Interestingly, we detect an important correlation (squared Pearson R value of 0.94) between the cumulative strength of the three hydrogen bonds and the energy of binding derived from the observed K_D. In principle, there is no reason why the binding energies derived from K_D measurements (an equilibrium property) and the breaking energies of H–bonds (which as computed are an out of equilibrium property) should be correlated. However, we propose that the correlation is not spurious and instead reflects two key mechanistical aspects of the interaction between CRBN and CK1α. First, is that the CK1α point mutations studied are not likely to affect the k_{on} of the complexes, which makes the observed decreases of affinity almost exclusively dependent on changes of the k_{off}. Second, and more crucially, the outstanding correlation between the free energy of H–bonds rupture and the observed affinity indicates that the dissociation of these complexes follows a rather simple two state mechanism, where breaking the H–bonds at the interface is the rate limiting step. Under these circumstances, the PMF_{HB_break} is the major contributor to changes in the $k_{off.}$ And can inform about the equilibrium constant. This observation, together with the dramatic effect of lenalidomide, underscores the potential that rationally designed MGs could hold for the modulation of protein–protein interactions in biomedical and biotechnological settings.

Regarding the underlying mechanism, we have shown that, when bound to the CRBN–CK1α interface, lenalidomide severely hinders water accessibility to the key protein–protein hydrogen bonds, as demonstrated by the stark decrease on the RDF value. (Figure 2) This

hydrophobic shielding effect seems a main driver in the stabilization effect triggered by lenalidomide, and thus could be considered to play a major role in the non–additive effects observed for this compound. It has been previously reported that relatively minor alterations of the H–bond environment can significantly alter H–bond lifetimes.[20,29] This effect is entirely consistent with the stabilization of pre–existing interactions put forward by Cao and co–workers and it is expected that similar mechanism underlies the degradation of other CRBN neo–substrates such as Ikaros and Aiolos and that is shared by other IMiDs such as pomalidomide (Figure 5). Beyond CRBN related systems, by analysing the crystallographic structures available in the PDB, we hypothesize that a similar effect underlies the recently described Cannabidiol–dependent stabilization of a dual–nanobody sensor[19] (PDBid 7TE8) and the long–standing puzzle of the Fucsicoccin–dependent stabilization of interactions involving 14-3-3 proteins (PDBid: 3P1S)[30] (Figure S5). Interestingly, evaluating water accessibility to the protein interface is not enough to anticipate H–bond strength. While the changes triggered by the ^{I35G}CK1α, ^{I37E}CK1α and ^{N39G}CK1α mutations can be rationalised on the basis of local changes to the environment of the H–bond, the behaviour observed for ^{G40N}CK1α is rather unexpected, as an increase in the size and hydrophobicity of the sidechain results in better access of water molecules to the protein–protein interface during the H–bond rupture process, that is not anticipated by RDF profiles of the complexes in equilibrium. Therefore, our results stress that, though often neglected, changes in the protein–protein interactions caused by the presence of MGs are as important as the direc interactions between the MGs and the proteins. We postulate that instead of solely focusing in maximizing affinity, computer–aided drug design strategies for MGs should also aim at maximizing protein–protein interactions by hydrophobic shielding of polar interactions. Analogous strategies should also be investigated for other types of interactions. In this work we demonstrate that an easy–to–implement SMD–based protocol is enough to predict stabilization of H–bonds which, in this

particular system, offer an excellent predictor of the thermodynamic stability of the ternary complex. It remains to be investigated if these results will transfer to other MGs systems, but the incorporation of this strategy in drug design workflows may assist much–needed rational approaches to the design of future MGs

Methods

Molecular simulations setup. Lenalidomide was built using the Molecular Operating Environment software package.[31] Models for the CRBN and CK1α were built starting from the crystallographic structure PDB id. 5FQD,[32] downloaded from the Protein Data Bank.[33–35] Standard protein preparation protocols were followed, including the removal of duplicated proteins, crystallization buffer compounds and salts. Additionally, the DNA Damage–Binding Protein 1 was removed in all systems and the appropriate capping groups were added to the terminal residues of CRBN. Mutants of CK1α were obtained with the mutagenesis wizard tool of PyMOl.[36,37] The ff14SB[38]and gaff2[39] forcefields were used to assign atom types for the protein and the lenalidomide respectively. Partial charges for lenalidomide were derived using the RESP[40,41] protocol at the HF/6-31G(d) level of theory, as calculated with Gaussian09. The Zn^{2+} cation bound to CRBN was modelled using the out of center dummy model[42] Each system, was solvated on a truncated octahedral box of TIP3P[43,44] water molecules and the appropriate number of counterions were added to achieve charge neutrality, accounting for simulations systems of approximately 100000 atoms. Each system was then minimized in three stages: first, the position of water molecules was minimized combining 3500 steps of steepest descent and 6500 steps of conjugate gradient, while the position of the proteins and ligand atoms was restrained using a harmonic potential with force constant of 5.0 kcal mol^{-1} $Å^{-2}$. Next, side chains and water molecules were minimized using 4500 steps of steepest descent, followed by 7500 steps of conjugate gradient while the atoms of lenalidomide and the Zn^{2+} cation were restrained

with a harmonic potential using the same force constant. The systems were then heated in the NVT ensemble from 100 K to 298 K in three stages of 250 ps (100K–150K, 150K–250K, 250K–298K), while retaining the harmonic restraints to lenalidomide and the Zn^{2+} cation and subsequently their density was equilibrated to 1 bar for 1 ns in the NPT ensemble. During the equilibration and subsequent production and steered molecular dynamics trajectories, temperature control was achieved using a Langevin thermostat (with a collision frequency of 3 ps^{-1}) and a Berendesen barostat was used to control the pressure when simulating in the NPT ensemble. SHAKE[45] was applied to all atoms involving hydrogen to allow for a timestep of 2 fs and all simulations were performed with the CUDA accelerated version of PMEMD.[46]

Steered Molecular Dynamics protocol. The stability of each H–bond in each system was assessed using 100 independent SMD trajectories conducted in three stages. First, new velocities were assigned to the equilibrated structure using a different random seed number at 298 K. Subsequently an MD trajectory was performed for 10 ns, using flat–bottom restraints to keep the three protein–protein H–bonds at the interface between 2.5 and 3.5 Å, using a force constant of 60 kcal/mol $Å^2$. Second, the final configuration of each trajectory was then used as a starting structure for a short (1 ns) SMD simulation in which the donor and acceptor involved in one of the H–bonds were brought to a distance of 2.5 Å. Third, a 5 ns–long SMD trajectory was started, in which the distance between donor an acceptor was increased at a rate of 0.5 Å/ns, using a spring constant of 500 kcal/ mol $Å^2$ to ensure the applicability of the stiff spring approximation.[47] The PMF_{HB_break} was then computed leveraging the Jarzynki's equality[48,49] (1).

$$e^{-\Delta G/k_B T} = \langle e^{-W_i/k_B T} \rangle \quad (1)$$

were the right–hand term corresponding to the ensemble average of exponential work values obtained in non–equilibrium conditions. From the above equation, for every increase of 0.0005 Å in the H–bond distance, the PMF_{HB_break} was obtained using expresion (2)

$$PMF_{HB\_break} = -k_B T \ ln \sum_{i=1}^{N} \frac{e^{W_i^{HB\_break}/k_B T}}{N} \quad (2)$$

Where $W_i^{HB\_break}$ refers to the work value of the ith independent SMD trajectory and N is the number of independent SMD trajectories (N=100 in this work). Error estimations for the PMF_{HB_break} profiles were obtained by bootstrapping ten times at each distance point the set of W^{HB_break} values. Convergence of the PMF_{HB_break} at 5 Å of H–bond (Figure S6) distance was evaluated combining subsampling and bootstrapping.

Calculation of water radial distribution function. The radial distribution function of water molecules around the backbone carbonyl oxygen of the CK1α residues involved in the interaction with CRBN was calculated was calculated using cpptraj[50,51], for a range between 0 and 10 Å from the atom of interest and with a bin spacing value of 0.1 Å.

Experimental data sourcing and analysis. Time–resolved fluorescence resonance energy transfer (TR/FRET) data points were extracted from Petzold et. al.[32] using WebPlotDigitizer v4.5[52] and analysis was performed with the Graphpad Prism 8 software.[53] Data points were adjusted to a non–linear regression curve achieving binding saturation. The maximum ratio value obtained for the CRBN–CK1α–lenalidomide ternary complex was use as constrained maximum signal (Ymax) in all the conditions to determine the K_D.

References

1 X. Tan, L. I. A. Calderon-Villalobos, M. Sharon, C. Zheng, C. V Robinson, M. Estelle and N. Zheng, *Nature*, 2007, **446**, 640–645.

2 B. Z. Stanton, E. J. Chory and G. R. Crabtree, *Science (1979)*, , DOI:10.1126/science.aao5902.

3 S. A. Andrei, E. Sijbesma, M. Hann, J. Davis, G. O'Mahony, M. W. D. Perry, A. Karawajczyk, J. Eickhoff, L. Brunsveld, R. G. Doveston, L. G. Milroy and C. Ottmann, *Expert Opin Drug Discov*, 2017, 12, 925–940.

4 L. G. Milroy, T. N. Grossmann, S. Hennig, L. Brunsveld and C. Ottmann, *Chem Rev*, 2014, 114, 4695–4748.

5 Y. Che, A. M. Gilbert, V. Shanmugasundaram and M. C. Noe, *Bioorg Med Chem Lett*, 2018, **28**, 2585–2592.

6 Q. Huai, H.-Y. Kim, Y. Liu, Y. Zhao, A. Mondragon, J. O. Liu and H. Ke, *Proceedings of the National Academy of Sciences*, 2002, **99**, 12037–12042.

7 P. B. Schiff and S. B. Horwitz, *Proceedings of the National Academy of Sciences*, 1980, **77**, 1561–1565.

8 D. E. Bussiere, L. Xie, H. Srinivas, W. Shu, A. Burke, C. Be, J. Zhao, A. Godbole, D. King, R. G. Karki, V. Hornak, F. Xu, J. Cobb, N. Carte, A. O. Frank, A. Frommlet, P. Graff, M. Knapp, A. Fazal, B. Okram, S. Jiang, P.-Y. Michellys, R. Beckwith, H. Voshol, C. Wiesmann, J. M. Solomon and J. Paulk, *Nat Chem Biol*, 2020, **16**, 15–23.

9 M. Słabicki, Z. Kozicka, G. Petzold, Y.-D. Li, M. Manojkumar, R. D. Bunker, K. A. Donovan, Q. L. Sievers, J. Koeppel, D. Suchyta, A. S. Sperling, E. C. Fink, J. A. Gasser, L. R. Wang, S. M. Corsello, R. S. Sellar, M. Jan, D. Gillingham, C. Scholl, S. Fröhling, T. R. Golub, E. S. Fischer, N. H. Thomä and B. L. Ebert, *Nature*, 2020, **585**, 293–297.

10 N. S. Scholes, C. Mayor-Ruiz and G. E. Winter, *Cell Chem Biol*, 2021, **28**, 1048–1060.

11 C. Mayor-Ruiz, S. Bauer, M. Brand, Z. Kozicka, M. Siklos, H. Imrichova, I. H. Kaltheuner, E. Hahn, K. Seiler, A. Koren, G. Petzold, M. Fellner, C. Bock, A. C.

Müller, J. Zuber, M. Geyer, N. H. Thomä, S. Kubicek and G. E. Winter, *Nat Chem Biol*, 2020, **16**, 1199–1207.

12 L. Brunsveld, Y. Higuchi, L.-G. Milroy, S. A. Andrei, C. Ottmann and P. J. de Vink, *Chem Sci*, , DOI:10.1039/c8sc05242e.

13 A. K. Gandhi, J. Kang, C. G. Havens, T. Conklin, Y. Ning, L. Wu, T. Ito, H. Ando, M. F. Waldman, A. Thakurta, A. Klippel, H. Handa, T. O. Daniel, P. H. Schafer and R. Chopra, *Br J Haematol*, 2014, **164**, 811–821.

14 J. Krönke, E. C. Fink, P. W. Hollenbach, K. J. MacBeth, S. N. Hurst, N. D. Udeshi, P. P. Chamberlain, D. R. Mani, H. W. Man, A. K. Gandhi, T. Svinkina, R. K. Schneider, M. McConkey, M. Järås, E. Griffiths, M. Wetzler, L. Bullinger, B. E. Cathers, S. A. Carr, R. Chopra and B. L. Ebert, *Nature*, 2015, **523**, 183–188.

15 E. S. Fischer, K. Böhm, J. R. Lydeard, H. Yang, M. B. Stadler, S. Cavadini, J. Nagel, F. Serluca, V. Acker, G. M. Lingaraju, R. B. Tichkule, M. Schebesta, W. C. Forrester, M. Schirle, U. Hassiepen, J. Ottl, M. Hild, R. E. J. Beckwith, J. W. Harper, J. L. Jenkins and N. H. Thomä, *Nature*, 2014, **512**, 49–53.

16 Q. L. Sievers, G. Petzold, R. D. Bunker, A. Renneville, M. Słabicki, B. J. Liddicoat, W. Abdulrahman, T. Mikkelsen, B. L. Ebert and N. H. Thomä, *Science (1979)*, , DOI:10.1126/science.aat0572.

17 G. Petzold, E. S. Fischer and N. H. Thomä, *Nature*, 2016, **532**, 127–130.

18 M. E. Matyskiela, T. Clayton, X. Zheng, C. Mayne, E. Tran, A. Carpenter, B. Pagarigan, J. McDonald, M. Rolfe, L. G. Hamann, G. Lu and P. P. Chamberlain, *Nat Struct Mol Biol*, 2020, **27**, 319–322.

19 S. Cao, S. Kang, H. Mao, J. Yao, L. Gu and N. Zheng, *Nat Commun*, 2022, **13**, 1–14.

20 P. Schmidtke, F. Javier Luque, J. B. Murray and X. Barril, *J Am Chem Soc*, 2011, **133**, 18903–18910.

21 S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N. Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard and X. Barril, *Nat Chem*, 2017, **9**, 201–206.

22 M. Majewski, S. Ruiz-Carmona and X. Barril, *Commun Chem*, 2019, **2**, 110.

23 G. G. Ferenczy and M. Kellermayer, *Comput Struct Biotechnol J*, 2022, **20**, 1946–1956.

24 R. Vogel, M. Mahalingam, S. Lüdeke, T. Huber, F. Siebert and T. P. Sakmar, *J Mol Biol*, 2008, **380**, 648–655.

25 S. Park and K. Schulten, *Journal of Chemical Physics*, 2004, **120**, 5946–5961.

26 H. Xiong, A. Crespo, M. Marti, D. Estrin and A. E. Roitberg, *Theor Chem Acc*, 2006, **116**, 338–346.

27 F. Colizzi, R. Perozzo, L. Scapozza, M. Recanatini and A. Cavalli, *J Am Chem Soc*, 2010, **132**, 7361–7371.

28 R. C. Bernardi, M. C. R. Melo and K. Schulten, *Biochim Biophys Acta*, 2015, **1850**, 872–877.

29 L. M. Nilsson, W. E. Thomas, E. V Sokurenko and V. Vogel, *Structure*, 2008, **16**, 1047–1058.

30 C. Anders, Y. Higuchi, K. Koschinsky, M. Bartel, B. Schumacher, P. Thiel, H. Nitta, R. Preisig-Müller, G. Schlichthörl, V. Renigunta, J. Ohkanda, J. Daut, N. Kato and C. Ottmann, *Chem Biol*, , DOI:10.1016/j.chembiol.2013.03.015.

31 Molecular Operating Environment (MOE), *Scientific Computing & Instrumentation*, 2019, 32.

32 G. Petzold, E. S. Fischer and N. H. Thomä, *Nature*, 2016, **532**, 127–130.

33 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res*, 2000, 28, 235–242.

34 S. K. Burley, H. M. Berman, C. Christie, J. M. Duarte, Z. Feng, J. Westbrook, J. Young and C. Zardecki, *Protein Science*, 2018, **27**, 316–330.

35 D. S. Goodsell, C. Zardecki, L. Di Costanzo, J. M. Duarte, B. P. Hudson, I. Persikova, J. Segura, C. Shao, M. Voigt, J. D. Westbrook, J. Y. Young and S. K. Burley, *Protein Science*, 2020, 29, 52–65.

36 W. L. Delano, *CCP4 Newsletter on protein crystallography*.

37 W. L. DeLano, *Schrödinger LLC*, 2020.

38 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J Chem Theory Comput*, 2015, **11**, 3696–3713.

39 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J Comput Chem*, 2004, **25**, 1157–1174.

40 C. C. I. Bayly, P. Cieplak, W. D. Cornell and P. a Kollman, *J Phys Chem*, 1993, **97**, 10269–10280.

41 T. Fox and P. a. Kollman, *J Phys Chem B*, 1998, **102**, 8070–8079.

42 F. Duarte, P. Bauer, A. Barrozo, B. A. Amrein, M. Purg, J. Åqvist and S. C. L. Kamerlin, *Journal of Physical Chemistry B*, 2014, **118**, 4351–4362.

43 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J Chem Phys*, 1983, **79**, 926–935.

44 P. Mark and L. Nilsson, *Journal of Physical Chemistry A*, 2001, **105**, 9954–9960.

45 V. Kräutler, W. F. Van Gunsteren and P. H. Hünenberger, *J Comput Chem*, 2001, **22**, 501–508.

46 R. Salomon-Ferrer, A. W. Goetz, D. Poole, S. Le Grand and R. C. Walker, *J Chem Theory Comput*, 2013, **9**, 3878–3888.

47 S. Park, F. Khalili-Araghi, E. Tajkhorshid and K. Schulten, *Journal of Chemical Physics*, , DOI:10.1063/1.1590311.

48 C. Jarzynski, *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, , DOI:10.1103/PhysRevE.56.5018.

49 C. Jarzynski, *Phys Rev Lett*, , DOI:10.1103/PhysRevLett.78.2690.

50 D. R. Roe and T. E. Cheatham, *J Comput Chem*, , DOI:10.1002/jcc.25382.

51 D. R. Roe and T. E. Cheatham, *J Chem Theory Comput*, , DOI:10.1021/ct400341p.

52 A. Rohatgi, *Pacifica, California, USA*.

53 H. Motulsky, *GraphPad Software Inc*.

AUTHOR INFORMATION

Corresponding Author

* Jordi Juárez-Jimenez – Unitat de Fisicoquímica, Departament de Farmàcia i Tecnologia Farmacéutica, i Fisicoquímica. Facultat de Farmacia I Ciències de l'Alimentació. Universitat de Barcelona (UB). Av. Joan XXIII, 27-31, 08028 Barcelona, Spain.
Email: jordi.juarez@ub.edu

* Xavier Barril – Unitat de Fisicoquímica, Departament de Farmàcia i Tecnologia Farmacéutica, i Fisicoquímica. Facultat de Farmacia I Ciències de l'Alimentació. Universitat de Barcelona (UB). Av. Joan XXIII, 27-31, 08028 Barcelona, Spain.
Email: xbarril@ub.edu

Author Contributions

Marina Miñarro-Lleonar: Investigation, Formal analysis, Software, Visualization, Writing-Review and Editing. **Andrea Bertran-Mostazo:** Investigation, Formal analysis, Writing-Review and Editing. **Jorge Duro:** Investigation. **Xavier Barril:** Conceptualization, Resources, Funding acquisition, Supervision, Writing Review and Editing. **Jordi Juárez-Jiménez:**

Conceptualization, Investigation, Software, Formal analysis, Funding acquisition, Supervision, Project administration, Writing – Original Draft.

ACKNOWLEDGMENT

This work received funding from the research project PDI2020-115683GA-100 ("Proyectos de I+D+i - Modalidad Generación de Conocimiento") financed by MCIN/AEI/10.13039/501100011033 and from the research project RTI2018-096429-N-I00 (Proyectos I+D+i – Modalidad Retos Investigación" financed by MCIN/AEI /10.13039/501100011033/ FEDER "Una manera de hacer Europa". X.B. and J.J-J. are members of the Computational Biology Drug Design Consolidated Research Group supported by the Generalitat de Catalunya (2017SGR1746) A.B-M. is supported by the predoctoral fellowship PRE2019-087468 financed by MCIN/AEI /10.13039/501100011033/ and FSE "El FSE invierte en tu futuro". We thankfully acknowledge access to the Marenostrum 4 HPC facilities granted through the Red Española de Supercomputación (BCV-2019-2-0021 and BCV-2019-3-0012).