Polytechnic of Turin

Master's Degree in Biomedical Engineering

A.a. 2024/2025

Graduate session December 2025

Virtual Screening in Search of Inhibitors of UBR1 E3 Ligase

Supervisors:	**Candidate:**
Prof. Jack A. TUSZYNSKI	**Chiara COLUCCI**
Prof. Richard FAHLMAN	
Dr. Paola VOTTERO	

Abstract

Multiple myeloma (MM) is a haematological malignancy that accounts for approximately 10% of all neoplasms of this type. The disease is characterized by a clonal proliferation of plasma cells, leading to an abnormal increase in monoclonal immunoglobulins in the serum. Initial therapy for active MM involves a combination of drugs, including proteasome inhibitors, immunomodulatory agents and corticosteroids.

The management of relapsed/refractory MM is more complex as MM does not follow a single pattern of oncogenic pathway alteration. Although advances have significantly improved prognosis, patients have a five-year survival rate of around 60%. For this reason, MM remains a malignancy in continuous therapeutic evolution, and the current challenge is to develop new strategies that reduce minimal residual disease.

In this sense, PROTAC (PROteolysis TArgeting Chimera) is part of the paradigm shift adopted in recent decades, which is based on targeted protein degradation and exploits the mechanism of the ubiquitin-proteasome system. It allows proteins to be degraded through the action of an enzymatic system followed by destruction within the 26S proteasome. This technology consists of a synthetic molecule composed of a warhead and an anchor, connected by a linker.

This project aimed to select potential compounds that could successfully bind to the UBR1 E3 Ligase protein, in order to constitute the anchor component of a PROTAC. To achieve this goal, a combination of computational methods was employed.

Since the crystallographic structure of UBR1 has not yet been fully resolved, the UBR-box domain was modeled by homology, using the 3NY3 structure as template. In order to obtain more meaningful results, the model of the entire UBR1 protein predicted by AlphaFold2 was also taken into consideration.

Molecular dynamics simulations were performed on both models to extract the most representative structures of the trajectory for use in subsequent stages.

Potential binding sites that could accommodate the ligands were then identified. After selecting the most promising pocket on the homology model of the UBR-box domain of UBR1 from among those found, also supported by visualization of the electrostatic potential map, a structure-based pharmacophore model was constructed.

The virtual screening phase was carried out using the NCI Diversity Set and the pharmacophore hypothesis was used as a filter.

The hits found were preliminarily analyzed in terms of pharmacokinetic properties and safety profile, which allowed the identification of 10 potential ligands.

Molecular docking represented another key step, enabling the evaluation of the binding affinity between the UBR1 E3 Ligase target and the potential ligands. To support a more robust selection of final candidates, the most promising hits were used as a starting point for a 95% similarity search on ZINC-22, which led to the construction of a docking set of more than 130 compounds. The properties of all of them were then analyzed in detail.

Through the evaluation of the features and the results obtained in the docking phase, 28 compounds were selected, divided by priority, for the subsequent experimental testing phase.

Acknowledgements

My sincere thanks go to Prof. Tuskynski, who gave me the opportunity to carry out this project and to delve into a field rich in new knowledge that I hope to continue exploring in the future.

I would also like to thank Prof. Fahlman for laying the foundations for this study and for making it possible to perform the experimental tests to validate the results at the University of Alberta laboratory.

My heartfelt thanks go to Paola for her constant support.

Finally, to my family and friends, I am deeply grateful for your love, encouragement, and unwavering support.

Table of Content

Abstract I

Acknowledgements II

Acronyms V

1. Introduction 1

1.1 Multiple Myeloma and Current Treatment Strategies 1

1.1.1 Current Therapeutic Strategies in Multiple Myeloma 3

1.1.2 Emerging Approaches: PROTACs in Multiple Myeloma Treatment 5

1.2 The Ubiquitination Proteasome System (UPS) 8

1.2.1 Mechanism of Action 8

1.2.2 26S Proteasome 10

1.2.3 Deubiquitination 11

1.2.4 Alteration in UPS Circuit 12

1.3 PROTACs: Proteolysis Targeting Chimeras 14

1.3.1 Different Generations of PROTAC 15

1.3.2 Interest of Pharmaceutical Company 17

1.3.3 Advantagies and Disadvantagies 18

1.4 E3 Ligase and PROTAC Anchors 20

1.4.1 E3 Ligase Families 20

1.4.2 Case Studies 21

1.4.3 UBR1 E3 Ligase 22

1.5 Computational Approaches in Drug Discovery 25

1.5.1 Protein Data Bank and AlphaFold 25

1.5.2 Virtual Libraries 27

1.5.3 Modular Synthon-Based Approaches 30

1.5.4 Structure-Based Virual Screening 31

1.5.5 Molecular Dynamics as a Tool for Model Refinement 32

2. Materials & Methods 35

2.1 Protein Structure Generation 35

2.1.1 Homology Model Construction 36

2.1.2 AlphaFold Model 41

2.1.3 Comparison Between Models 42

2.2 Molecular Dynamic Simulations 43

2.2.1 Energy Minimization 43

2.2.2 Equilibration Phases 45

2.2.3 Production Simulation and Calculation of Parameters for Data Analysis 46

2.2.4 Extraction of the Most Representative Structure 47

2.3 Binding Site Prediction and Evaluation 49
2.3.1 Search for Binding Pockets 49
2.3.2 Electrostatic Potential Map 51
2.3.3 Structure-Based Pharmacophore Model 53
2.4 Virtual Screening 56
2.4.1 Ligand Preparation 56
2.4.2 Ligand Screening 58
2.4.3 Preliminary Evaluation of Compound Properties 60
2.5 Molecular Docking and Compounds Selection 65
2.5.1 Similarity Search and Docking Set Construction 65
2.5.2 Docking with GNINA 68
2.5.3 In-Depth ADMET Evaluation 72
2.5.4 Selection of Candidates 76
3. Results 78
3.1 Analysis of the models 78
3.1.1 Homology Model 78
3.1.2 AlphaFold Model 84
3.1.3 Comparison Between the Obtained Models 88
3.2 Molecular Dynamics Simulations Outputs 90
3.3 Pocket identification 94
3.3.1 FTSite output 94
3.3.2 MEPs Visualization 97
3.3.3 Pharmacophore Model Based on the Selected Pocket 100
3.4 First phase of identifying potential ligands 103
3.4.1 Virtual Screening Hits 103
3.4.2 Features of Top Hit Compounds 105
3.5 Identification of Final Candidates 110
3.5.1 Docking Results 110
3.5.2 ADMET Predictor v12.0 Results 113
3.5.3 Ligands selected for experimental tests 116
4. Discussion and Future Directions 121
Appendix A 125
Appendix B 126
Appendix C 127
Appendix D 133
Bibliography 135

Acronyms

3DEM cryo-electron microscopy

ADMET Absorption, Distribution, Metabolism, Excretion, and Toxicity

AF AlphaFold

AF2 AlphaFold2

AF3 AlphaFold3

AI Artificial Intelligence

APC/C Anaphase-Promoting Complex/Cyclosome

APBS Adaptive Poisson–Boltzmann Solver

AR Androgen Receptor

ASCT Autologous stem cell transplantation

ATP adenosine triphosphate

CADD computer-assisted drug discovery

CELMoDs cereblon E3 ligase modulating drugs

CNNs Convolutional Neural Networks

CRBN Cereblon

CRL4 Cullin-RING 4 ubiquitin-ligase

CRLs Cullin-RING ligases

CRS Cytokine Release Syndrome

D-box destruction box

DDD Drug Discovery and Development

DL Deep Learning

DTP Developmental Therapeutics Programme

DUBs Deubiquitinating enzymes

E1 Ubiquitin-activating enzymes

E2 Ubiquitin-conjugating enzyme

E3 Ubiquitin-Ligase enzyme

ECFP Extended-Connectivity Fingerprints

EMBL-EBI European Molecular Biology Laboratory's European Bioinformatics Institute

ER Estrogen Receptor

FDA US Food and Drug Administration

FLC Free Light Chains

GMQE Global Model Quality Estimate

GPCRs G protein-coupled receptors

GPUs Graphics Processing Units

GUI Graphical User Interface

HAC Heavy Atom Count

HBAs Hydrogen Bond Acceptors

HBDs Hydrogen Bond Donors

HECT Homologous to the E6-AP C-Terminus

HPC High-Performance Computing

hERG Human Ether-à-go-go-Related Gene

HTS High-Throughput Screening

IMiDs immunomodulating agents

MD Molecular Dynamics

MEP Electrostatic Potential Map

MGUS monoclonal gammopathy of unknown significance

ML Machine Learning

MM Multiple myeloma

mmCIF macromolecular Crystallographic Information File

MX X-ray crystallography

NCI US National Cancer Institute

NIH National Institutes of Health

NK Natural Killer

NMR Nuclear Magnetic Resonance

PAE Predicted Aligned Error

PDB Protein Data Bank

PFS Progression-Free Survival

PhosTAC phosphorylation targeting chimeras

PI Proteasome Inhibitors

PME Particle Mesh Ewald

POI Protein Of Interest

PPI Protein Protein Interaction

PROTACs PROteolysis TArgeting Chimeras

QMEANDisCo Qualitative Model Energy Analysis - Distance Constraints

QSAR Quantitative Structure-Activity Relationship

RIBOTAC Ribonuclease-Targeting Chimera

RING Really Interesting New Gene

RMSD Root Mean Square Deviation

RRMM refractory relapsed multiple myeloma

SBD Substrate Binding Domain

SBVS Structure-Based Virtual Screening

SEER National Cancer Institute's Surveillance, Epidemiology, and End Results Program

SMM Smoldering Multiple Myeloma

sPCL secondary Plasma Cell Leukemia

SAR Structure-Activity Relationship

TPD Target Protein Degradation

UPS Ubiquitin-Proteasome System

VdW Van der Waals

VHL Von Hippel-Lindau

VS Virtual screening

1. Introduction

1.1 Multiple Myeloma and Current Treatment Strategies

Multiple myeloma (MM) accounts for 10% of haematological malignancies and is a disease characterised by a clonal proliferative condition of plasma cells, defined by an abnormal increase in monoclonal immunoglobulin levels. Based on data from 2018 to 2021 (excluding 2020 due to COVID), roughly 0.8% of men and women will be diagnosed with myeloma at some point in their lives.

Each year, for every 100,000 men and women, there were 7.3 new cases and 2.9 deaths, while around 37.6% of people diagnosed with this disease do not survive more than 5 years [1]. Myeloma is most frequently diagnosed among people aged 65–74 and its incidence varies according to ethnicity: African Americans are twice as likely as whites to develop the disease, while Asians have a lower rate.

There was no particular flaw found in the 2011 myeloma genome sequencing and a 2018 study reveals that there is no single way in which oncogenic pathways are altered in MM, but different genetic mechanisms can act [2], [3]. Whole exome sequencing also showed comparable results, and all patients upon diagnosis had many distinct subclones, including mutations in the driver genes KRAS, NRAS, and BRAF.

The preponderance of clones varies according on the development of the disease and the response to treatment. This discovery has altered the understanding of myeloma from a disease that progresses linearly and becomes more resistant to treatment to one in which a "Darwinian" or "branching" process takes place, whereby chemotherapy may suppress some clones but ultimately leads to the dominance of new, chemotherapy-resistant clones [4].

Premalignant neoplastic plasma cell growth, known as monoclonal gammopathy of unknown significance (MGUS) or smoldering multiple myeloma (SMM) based on disease severity, is hypothesized to precede all MM cases.

Serum monoclonal protein < 30 g/L, absence or mild presence of monoclonal light chains in urine, lack of lytic bone lesions, anemia, hypercalcemia, or renal failure due to gammopathy are the characteristics that identify MGUS. If a bone marrow examination is performed, the plasma cell count must be <10%. MGUS is more frequent than MM, involving around 3% of people over 50 years old [5].

However, most MGUS cases remain stable and do not progress to MM.

At the time of diagnosis, it is often difficult to predict which patients will develop myeloma or related plasma cell disorders. The presence of organ damage such as hypercalcemia, kidney failure, anemia and lytic bone lesions collectively known as CRAB, represent the usual criteria for diagnosing active multiple myeloma.

Recently, thanks to data from studies on patients with an asymptomatic but potentially evolving form of SMM, The International Myeloma Working Group has updated these criteria to allow rapid detection of the individuals who are at high risk of progression.

These new parameters include the identification of a concentration of free light chains (FLC) greater than 100 mg/L, a percentage of plasma cells in the bone marrow greater than 60% and MRI evidence of multiple lytic lesions.

Patients with high-risk SMM have an estimated 80% probability of progression to MM within the next two years. High-risk SMM is characterized by ≥10% marrow plasma cells and at least one of the following: increasing or high paraprotein levels (>30 g/L), FLC concentration between 8 and 100 mg/L, immunoparesis of uninvolved immunoglobulins, presence of circulating plasma cells, abnormal plasma cell phenotype, high-risk cytogenetic abnormalities, or PET/MRI abnormalities.

Several clinical studies are currently investigating whether early intervention in these patients can delay or prevent the progression of symptomatic MM [6], [7].

The following table highlights the main characteristics and differences between the various conditions:

Feature	MGUS	SMM	Active MM
CRAB	Absent	Absent	Present (classical criterion)
Monoclonal protein	< 3 g/dL	≥ 3 g/dL	Any of them
Bone marrow plasma cells	< 10%	10 – 60%	≥ 60% (biomarker criterion)
FLC ratio	It can be altered but not pronounced	8 – 100	≥ 100
Lytic lesion	Absent	Absent	Multiple
Risk of progression	~1%/year	high risk: ~80% at 2 years	Full-blown disease, treatment required

Table 1.1: *Main differences between MGUS, SMM and active MM.*

Over the past few decades, MM sufferers' prognosis has considerably improved. According to data from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER), patients diagnosed between 2012-2018 had a five-year survival rate of 58%, compared to 32% for patients diagnosed in 1995-1997.

Modern treatment of MM allows early cytoreduction and clinical improvement in almost all patients. However, while a small percentage of patients with MM may recover or achieve indefinite remission of the disease with first-line treatment, most patients relapse and if they develop treatment-resistant disease, it is eventually fatal [8].

For this reason, MM is an oncological disease in continuous therapeutic evolution. The ongoing challenge is to develop new therapeutic strategies that can offer more personalized, longer-lasting care and reduce the residual minimal disease.

1.1.1 Current Therapeutic Strategies in Multiple Myeloma

Currently, the standard of care for initial treatment of MM is the administration of three or four drugs (triplets or quadruplets) that include proteasome inhibitors (PI), immunomodulating agents (IMiDs), corticosteroids and, more recently, monoclonal antibodies.

The PIs block the proteolytic complex involved in the degradation of ubiquitinated proteins; in fact, the malignant plasma cells are particularly sensitive to this mechanism due to the high protein production.

One of the main PIs is bortezomib, approved by the US Food and Drug Administration (FDA) in 2003 [9]. This drug has been shown to be effective in the relapsing/refractory phase, so other PIs have also been studied at all stages of MM therapy. It was initially administered intravenously, but to reduce neurological toxicity it was then administered subcutaneously.

A second-generation proteasome inhibitor is carfilzomib, approved by the FDA in 2012 [10]. This is more potent than the previous one because it binds irreversibly, but both are associated with cardiovascular toxicity.

Ixazomib (since 2015) has a better tolerability profile, so it is also used in fragile patients. This is the only oral drug, but it binds reversibly [11].

Another type of targeted molecular therapy are the Cereblon-binding agents, called IMiDs (immunomodulatory imide drugs) or CELMoDs (cereblon E3 ligase modulating drugs).

IMiDs determine the degradation of key transcription factors (IKZF1, IKZF3). For example, the degradation of IKZF3, in turn, significantly reduces the transcription of IRF4, a critical transcription factor for the survival of MM cells. These act by binding to cereblon, a Cullin-RING 4 ubiquitin-ligase (CRL4) complex (an E3 ubiquitin ligase complex).

One of the IMiDs is thalidomide, approved by the FDA in 2006, but only since 2010 has it been clarified that it could bind to cereblon. However, this is associated with peripheral neuropathy.

Lenalidomide (since 2006) is a second-generation agent; it is used in both the first line of treatment and maintenance, but has a renal-dependent pharmacokinetics.

In patients refractory to lenalidomide and bortezomib, the third-generation drug pomalidomide can be used. The latter and lenalidomide were structurally derived from thalidomide, and both have a myelosuppression toxicity profile [12].

Advanced clinical studies are still ongoing for CELMoDs like mezigdomide and iberdomide. It is too early to evaluate if such compounds perform better than current IMiDs and so influence their widespread adoption but studies are promising [13].

Synthetic corticosteroids like dexamethasone have immunosuppressive and anti-inflammatory effects. It has a direct effect on malignant plasma cells and improves the therapeutic efficiency of certain treatment combinations used to treat myeloma [14].

Autologous stem cell transplantation (ASCT) is another first-line therapy that allows patients eligible for transplantation to achieve a deeper and longer-lasting remission.

The VRd induction regimen of bortezomib, lenalidomide, and dexamethasone is used to prepare patients for transplantation [15]. Post-transplant maintenance is performed with lenalidomide as it has confirmed a significant increase in Progression-Free Survival (PFS). For patients who are not eligible for transplantation, due to age or comorbidity, the use of daratumumab-containing regimens such as DRd (daratumumab, lenalidomide, dexamethasone) is indicated, which has shown significant improvements in PFS and quality of life [16].

Management of refractory relapsed multiple myeloma (RRMM) is complex and depends on several factors: duration of response to previous therapy, type of drugs used, presence of cumulative toxicity and biological characteristics of the disease.

In relapsed patients the sequential use of combinations with known drugs has been shown to be effective. An example is the combination of daratumumab with pomalidomide and dexamethasone (DPd) is used in lenalidomide-refractory patients. Carfilzomib is also successfully used in combination with lenalidomide or dexamethasone.

The combination IRd includes Ixazomib together with lenalidomide and dexamethasone. It offers a favourable tolerability profile and simplified treatment management.

Immunotherapy is breaking new ground in the treatment of MM, with an increasing number of therapeutic options.

Some of them are the anti-CD38 monoclonal antibodies daratumumab and isatuximab, which have shown high efficacy in both first-line therapy and relapsed patients. They bind to different parts of CD38 but both act with multiple mechanisms such as cell-mediated antibody-dependent cytotoxicity, phagocytosis, complement-mediated cytotoxicity, and direct cell death.

Isatuximab has also been shown to work in patients refractory to daratumumab. Elotuzumab, an antibody directed against SLAMF7 that acts by promoting the immune response by activating natural killer (NK) cells, has shown efficacy in relapsed patients. A highly expressed antigen on malignant plasma cells is the B-cell maturation antigen (BCMA).

The CAR-T cells target precisely this and represent an important evolution of immunotherapy. In treated patients, the FDA-approved products are idecabtagene vicleucel (ide-cel) and ciltacabtagene autoleucel (cilta-cel), but their use is conditioned by logistical difficulties, prolonged production times and risk of serious negative consequences including neurotoxicity and cytokine release syndrome (CRS).

To address these limitations, "off-the-shelf" therapies such as bispecific antibodies are being developed, which are much more selective than previous ones without the need for customized production. Among these, we find teclistamab, which was the first approved by the FDA and talquetamab. These drugs show effective responses even in patients already undergoing multiple treatment lines. New therapies such as trispecific antibodies, CAR-NK cells and therapeutic vaccines are also emerging but are still in the experimental stage [17].

Other emerging therapies for the treatment of MM are BCL-2 inhibitors, such as venetoclax. These have shown particular efficacy in patients with translocation t(11;14), paving the way for personalized treatments based on genetic profile. This drug induces apoptosis of cancer cells and is used as maintenance therapy after autologous stem cell transplantation.

Selinexor is a selective inhibitor of XPO1 (also known as CRM1), a protein responsible for the nuclear export of cancer suppressor proteins such as p53. By blocking XPO1, it promotes the nuclear accumulation of these proteins and induces apoptosis in cancer cells. For patients with refractory and multiline MM, it is authorized when used in conjunction with dexamethasone [8].

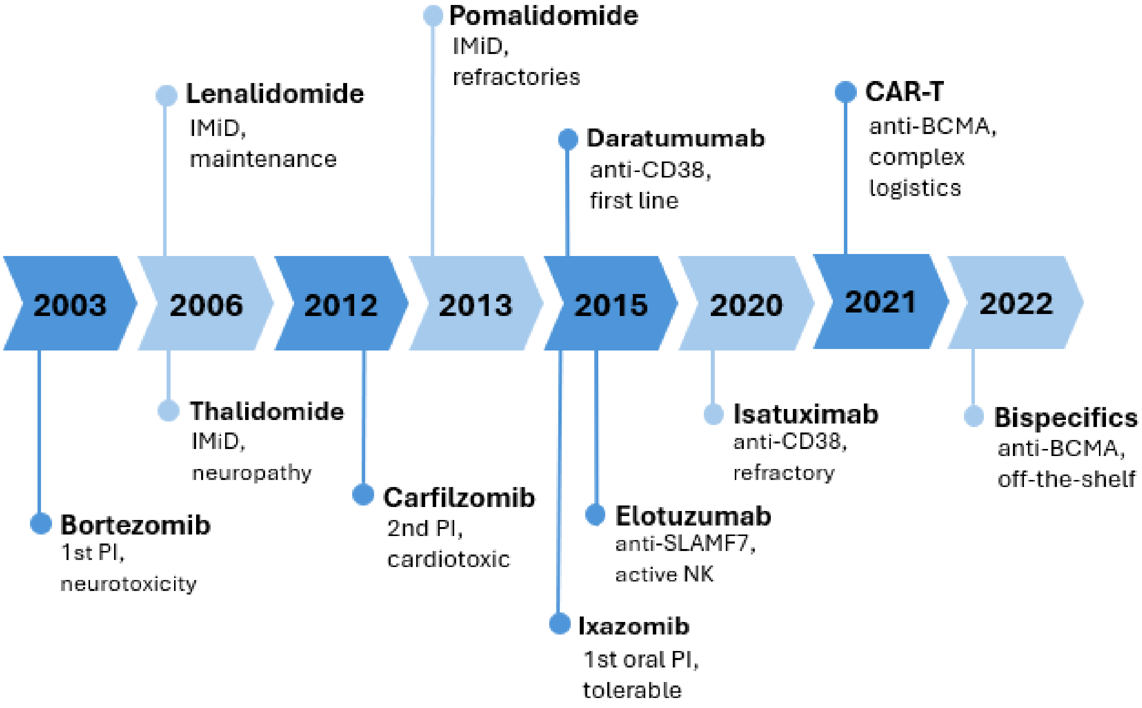


Figure 1.1: *Timeline of therapeutic advances in MM: from the first IMiDs and the introduction of PI such as bortezomib (2003), to second- and third-generation agents, monoclonal antibodies (daratumumab, isatuximab, elotuzumab), and most recently CAR-T cells and bispecific antibodies targeting BCMA. The figure highlights the chronological sequence of drug approvals.*

1.1.2 Emerging Approaches: PROTACs in Multiple Myeloma Treatment

In recent years, an innovative therapeutic approach has aroused great interest in the treatment of multiple myeloma: PROTACs (PROteolysis TArgeting Chimeras). They have strong activity against tumours, such as prostate cancer, leukaemia and lymphoma.

In the case of MM, PROTACs ARV 825 and ARV 763, which target BET proteins which contain bromodomains, for degradation. They have been shown to be associated with a decrease in MYC and Akt/mTOR. The downregulation of the latter, together with altered expression of genes such as IGLL5, IRF4, PRDM1/BLIMP-1 and XBP1, play key role in the control of cell proliferation and for this reason they are involved in the disease progression.

ARV 825, inducing degradation of BET proteins via CRBN E3 ligase, has been shown to be active against primary myeloma cells both in vitro and in vivo and can overcome drug resistance. This PROTAC inhibited cell proliferation of human MM cell lines, arrested the cell cycle and induced apoptosis of MM cells in vivo. CRBN expression is particularly relevant in MM pathology, as well as in other haematological malignancies and the higher the expression of CRBN increases, the greater the sensitivity to ARV 825 of MM patients also increases.

Resistance to lenalidomide, pomalidomide and to ARV 825-resistant MM cells, can be overcome thanks to a different PROTAC, the MZ1, based on the E3 ligase VHL [18], [19].

Other pomalidomide-based PROTACs that recruit the E3 ligase CRBN as a therapeutic target have also been found to be functional in inhibiting the proliferation of hematopoietic cancer cells. Examples are PROTACs with significant CDK6 degradation potential built by linking the CDK6 inhibitor palbociclib and the recruiter of this E3 ligase. To aid the deduction of the best design strategy and its future optimization, a library of CDK6 degraders was developed containing details such as linker length, spatial orientation and binding affinity.

One functional degrader that has been developed is PROTAC CP-10, derived from palbociclib, which has been shown to be able to degrade mutated and overexpressed CDK6. The results indicate that this approach may lead to potential future clinical applications, both in MM and in treating other cancers [20].

PROTACs targeting the degradation of the kinases CDK8 and CDK19, which are crucial for the regulation of transcription and stability of cyclin C (CCNC), have also been developed.

Although CDK8/19 inhibitors have shown some efficacy in solid tumour and leukaemia models, their activity is limited to the catalytic functions of the kinases, leaving non-enzymatic functions unaffected. To overcome this limitation, PROTACs combining selective CDK8/19 inhibitors with ligands for the CRBN E3 ligase have been developed, resulting in an effective degradation of both CDK8 and CDK19.

Transcriptomic studies confirmed that these degraders replicate most of the inhibitor-induced effects, but with the addition of CCNC degradation - a critical event, as MM shows a higher dependence on CCNC than other tumour types. In MM cell lines with high CCNC dependence, CDK8/19 PROTACs proved to have superior efficacy compared to clinically approved drugs. These data underline the therapeutic potential of PROTACs oriented towards CDK8/19 degradation in MM subtypes with specific molecular vulnerabilities.

PROTACs based on derivatives of the immunomodulator lenalidomide offer new opportunities for selective degradation.

Lenalidomide acts as a molecular glue that recruits neosubstrates to degradation via the CRL4 CRBN complex, and is the basis for many PROTACs directed against oncogenic targets. However, the degradation profile of neosubstrates by lenalidomide and its derivatives may include proteins involved in embryonic development and tumour progression. Recent studies have shown that targeted structural modifications, such as fluorination at position 6 of the lenalidomide ring, can optimise selectivity towards therapeutically relevant neosubstrates such as IKZF1, IKZF3 and CK1α, all implicated in the pathogenesis of MM and myelodysplastic syndromes with chromosome 5q deletion.

PROTACs built on these modified derivatives have been shown to induce targeted and highly selective degradation, with a significant anti-proliferative effect on MM cell lines, superior lenalidomide. This rational engineering approach of PROTACs, focused on neosubstrate selectivity, represents a promising strategy to improve therapeutic efficacy and reduce side effects related to non-specific degradation [21].

Another interesting development in the field of MM therapy and its aggressive leukemic evolution, secondary plasma cell leukemia (sPCL), is the use of PROTACs to degrade the antiapoptotic BCL-XL, frequently expressed in these malignant cells.

sPCL is a rare and difficult to treat medical condition, with very limited therapeutic options, especially in the refractory phase. While BH3 mimetics such as A1155463 are able to inhibit BCL-XL and induce cell death in some sPCL lines, their use is strongly limited by haematological toxicity, particularly thrombocytopenia, due to the off-target effect on platelets.

To overcome this problem, PROTAC DT2216 has been developed, a selective degrader of BCL-XL that exploits the E3 ligase VHL, thus avoiding the toxic effect on platelets. Studies on myeloma and sPCL cell lines have shown that DT2216 is capable of degrading BCL-XL to nanomolar concentrations, inducing apoptosis and activation of BAX- and BAK-mediated cell death cascades. The sensitivity to DT2216 strongly correlates with that to A1155463, confirming that in those patients with dependence on BCL-XL, targeted degradation via PROTAC may represent an effective and safer therapeutic option. The rational for focused clinical studies in individuals with BCL-XL-dependent sPCL is supported by the evident therapeutic benefit of DT2216's lack of thrombocytopenia [22].

Considering these discoveries and the potential of these new technologies, the choice of E3 ligases plays a fundamental role, as they represent the anchor through which PROTACs exert their function. Through their modulation, it is possible to explore new therapeutic strategies based on targeted protein degradation. In this context, it has been decided to investigate the UBR1 ligase, as it emerges as a potential innovative therapeutic target.

1.2 The Ubiquitination Proteasome System (UPS)

In recent decades, more precisely since since the late 90s, there has been a paradigm shift in pharmacology from the simple concept of function inhibition to an approach that aims to selectively remove pathological proteins from cells. This is the idea behind the therapeutic solution of Target Protein Degradation (TPD), which exploits the cells' natural disposal mechanisms.

There are two main mechanisms of protein degradation, which are independent but interconnected: Lysosome-Mediated Degradation and the Ubiquitin-Proteasome System (UPS). The first one is responsible for the degradation of membrane proteins, extracellular or even entire organelle compartments, long-lived, insoluble protein aggregates and intracellular parasites. It occurs via endocytosis, phagocytosis or autophagy and can exploit these mechanisms to digest proteins with acid hydrolases. There are many technologies associated with this mechanism, including for instance LYTACs (Lysosome Targeting Chimeras), AUTACs (Autophagy Targeting Chimeras) and ATTECs [23].

The mechanism that is explored in this section is UPS as it is exploited by the technology proposed in this project.

1.2.1 Mechanism of Action

This is the process by which most intracellular proteins are degraded in a controlled manner, in eukaryotic cells. This mechanism degrades approximately 85% of the body's proteins. All facets of the cellular metabolic networks associated with either healthy or diseased processes involve this system, either directly or indirectly.

It acts through an enzyme system consisting of the activation enzyme E1, the conjugating enzyme E2 and the one responsible for specificity, E3 ligase. This process is followed by destruction in 26S proteasome.

Ubiquitin is found in all eukaryotic cells and is a small essential protein of 76 residues. It has a compact structure, with a 5-stranded antiparallel b-sheet crossed by a single helix.

In the ubiquitination process, the most important roles are played by the C-terminal carboxylic group of G76 and the primary amino group in each of the seven lysine residues (K48, K63, K6, K11, K27, K29 and K33). The Lys48 conjugated residue, however, is primarily responsible for the proteasome's destruction, as it signals proteins for proteasomal degradation.

Specifically, in the ubiquitination process, in which the target substrate is marked with ubiquitin, the three enzymes involved act as follows:

- **E1** is the first to initiate this process by activating ubiquitin. Through an ATP-dependent mechanism, it aims to form a thioester-reactive intermediate E1-ubiquitin. A thioester bond is created between the activated glycine residue and a cysteine residue on its own catalytic site as a result of E1 adenylating the C-terminal glycine of ubiquitin;
- **E2** (ubiquitin-conjugating enzyme), a ubiquitin-carrier protein, receives the active ubiquitin by transthiolation;

- Subsequently, the enzyme **E3 ligase** comes into action. Depending on the E3 ligase domain, ubiquitin transfer can occur either directly onto the substrate (RING-type) or via a transient E3 ubiquitin intermediate (HECT-type). The ubiquitin molecule is transferred onto the target protein by E3, which is also in charge of recruiting it.

Usually, a lysine residue of the target protein's substrate, most especially a ε-NH2 group, leading to the formation of an isopeptide bond between ubiquitin and the substrate protein. It may be that the substrate of the protein to be tagged does not contain any lysine residues. In this case, the role of attaching the thioester bond is performed by the N-terminal α-amino group of the target protein [24].

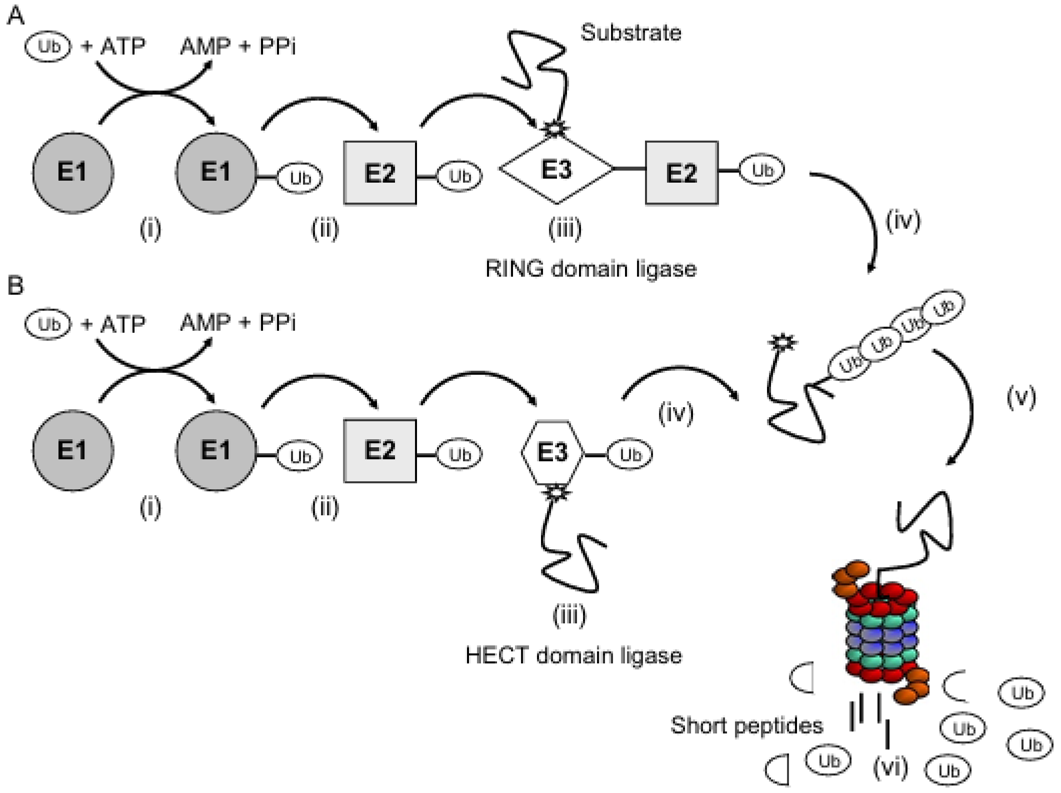


Figure 1.2: *The UPS. (Ai and Bi) show the ATP-dependent activation of ubiquitin that is catalyzed by E1 enzyme. (Aii and Bii) illustrate the transfer of the activated ubiquitin to E2. (Aiii) shows the case of a RING domain ligase,in which the ubiquitin-charged E2 binds to the E3 ligase that carries the substrate protein and (Aiv) the transfer of the activated ubiquitin moiety directly to the substrate. Biii displays the case of an HECT domain ligase, the ubiquitin is transferred from the ubiquitin-charged E2 to a conserved Cys residue in the E3 ligase and then to the substrate. (Biv) describes the steps that are repeated in order to produce the polyubiquitin chain. (A, Bv and A, Bvi) illustrate that the ubiquitinated substrate is then degraded to short peptides by the 26S proteasome with release of free and reusable ubiquitin. Some of the ubiquitin molecules are degraded in this process along with the substrate [25].*

To explain the assembly of polyubiquitin chains, several models have been proposed including the model “hybrid model”, “seesaw”, “indexation” and “hit and run”, even if in vivo there are variations.

According to this standard model, ubiquitination is a repeated cycle, thus producing not only one molecule but chains of polyubiquitin on the target protein. It has been seen that the length of this chain can vary considerably. Other types of enzymes, E4, called chain elongation factors are also proposed but, actually, in some essays they have been shown to act as E3. For this reason and because the tertiary structure of their U-box is very similar to that of the RING domain (really interesting new gene), it is thought that only E3 are able to use the ubiquitin thioester-bound HECT domain of their E3 cognates.

However, for the 26S proteasome to recognize and process proteins efficiently in the phase of degradation, at least four ubiquitin residues are required.

Finally, it should be mentioned that the ubiquitin signal is not always connected to a polyubiquitin chain. In order to provide structural variety and regulate different biological processes, ubiquitin may also be conjugated to a protein as a single moiety in a process known as monoubiquitination or as numerous ubiquitin moieties in a process known as multiple monoubiquitination. These are different processes, which do not affect the UPS mechanism.

After the construction of the ubiquitin molecule chain on the target protein, the 26S proteasome comes into play and degrades the substrate with ubiquitin into small peptides, 8 to 10 amino acids long. Ubiquitin molecules are partly degraded in this process, but others are released, ready to be reused.

1.2.2 26S Proteasome

The 26S proteasome is found in both cytoplasm and cell nucleus, and it is a key component of the UPS for its role in the degradation of ubiquitin-marked proteins. It is a macromolecular multi-enzimatic structure of 2.4 MDa which consists of two main subunits. The catalytic core is 20S and to it one or two regulatory particles 19S, also called PA700, can be bound to form the complex 26S.

The 20S complex of 700-kDa consists of four heptameric rings, of which the two outer ones are formed by seven alpha subunits (α1-7) and two inner ones composed of seven beta subunits (β1-7). The latter two rings contain the catalytic subunits responsible for breaking down the target proteins into small peptides. Specifically, they are β1 that has caspase-like activity (PGPH), β2 with tryptic activity and β5 with chymotryptic activity. In inflammatory or immune conditions, the catalytic subunits can be replaced by induced by interferon-γ, generating the so-called immunoproteasome (β1i, β2i, β5i).

The key regulatory activity is carried out by the 19S (also known as RP/PA700) complex of about 1 MDa. This consists of two subcomplexes: the base, which contains six ATPasic subunits (Rpt1-Rpt6) and some non-catalytic structural subunits (Rpn1, Rpn2), and the lid, formed by several subunits including Rpn3, Rpn5-Rpn9, Rpn11. The Rpn10 and Rpn13 receptors recognize ubiquitinated proteins so that the 19S proteasome can unfold them in an ATP-dependent way and allow them to enter the catalytic channel of the 20S.

During this process, the ubiquitin chains are removed before the protein is degraded, preventing the destruction of the ubiquitin itself. However, some molecules can still be degraded.

The 26S proteasome is extremely dynamic; besides the typical shape with two 19S units, there are hybrid variants that regulate the specificity and efficiency of the degradation process, such as 19S-20S-11S or 19S-20S-19S. In addition, 20S can also act independently under some physiological or tissue conditions, degrading oxidized or damaged proteins without ubiquitination.

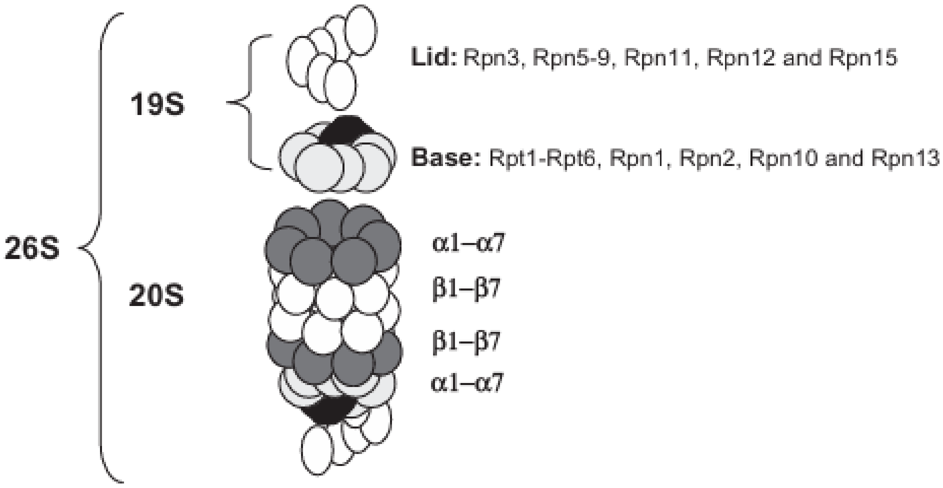

26S
19S
20S
Lid: Rpn3, Rpn5-9, Rpn11, Rpn12 and Rpn15
Base: Rpt1-Rpt6, Rpn1, Rpn2, Rpn10 and Rpn13
α1–α7
β1–β7
β1–β7
α1–α7

Figure 1.3: *Schematic representation of the 20S proteasome and the 19S regulatory particle [25].*

1.2.3 Deubiquitination

Ubiquitination is a reversible process, therefore deubiquitination can occur. The latter mechanism ensures the recovery of ubiquitin by enzymes that remove it from target proteins or polyubiquitin chains, thus modulating their stability, function and targeted proteins' fate.

These enzymes are called deubiquitinating enzymes (DUBs) and are classified into different families. The largest family consists of the Ubiquitin Specific Proteases (USPs), but there are also the JAB1/MPN/Mov34 Metalloenzymes (JAMMs) that are often associated with the proteasome, the Ubiquitin C-terminal Hydrolases (UCHs), the Ovarian Tumor Proteases (OTUs) and the Josephin domain DUBs.

Deubiquitination is critical and crucial because it controls the activation of transcription factors, the response to oxidative stress and the activity of oncosuppressors. An example of the latter function is USP29 which, under stress conditions, can stabilise p53 and promote apoptosis. For this reason, DUBs are emerging as new pharmacological targets in the treatment of various conditions, including cancer and neurodegenerative diseases.

A DUB with a key role in the context of the proteasome is Rpn11. It is a member of the JAMM family and is particularly important. It removes ubiquitin chains in a processive manner that is coordinated with substrate entry into the catalytic channel. This prevents the proteasome from being blocked by substrates that are still ubiquitinated.

The USP14 and UCH37 DUBs, on the other hand, act earlier by removing ubiquitin chains before the degradation of the targeted protein. In some cases, they can "rescue" the target protein from degradation [25], [26].

1.2.4 Alteration in UPS Circuit

Any alteration in the UPS circuit can favour the development of diseases, as this process is fundamental for cell metabolism and is involved in several functions. UPS is crucial for cells to exit mitosis; in fact it allows the degradation of e.g. cyclin B and the CDK inhibitor p27kip1 at precise moments in the cell cycle.

The E3 ligases involved in this process are the SCF complexes that regulate entry into S-phase and recognise substrates after phosphorylation, and the Anaphase-Promoting Complex/Cyclosome (APC/C). It is precisely the latter that is important for the separation of sister chromatids, the exit from mitosis and the degradation of cell cycle regulators containing the destruction box (D-box). The D-box is a sequence of nine amino acid residues that signals the decay of mitotic cyclic proteins at the end of mitosis. This is recognised by APC/C, so the proteins are labelled with ubiquitin and degraded by the proteasome [27].

UPS triggers an inflammatory response by degrading the previously phosphorylated inhibitor-κB. In the cytosol, this inhibitor is in fact bound to NF-κB, a key transcription factor involved in the inflammatory response. Once free of the inhibitor, it is able to enter the cell nucleus and induce the expression of several genes.

This system is also essential for the generation of antigenic peptides to be presented on Major Histocompatibility Complex class I (MHC I), a molecule that is present on the surface of all nucleated cells and that exhibits peptides that derive from the degradation of endogenous proteins in cytotoxic T lymphocytes ($CD8^+$).

Another function of UPS is the elimination of misfolded proteins through interaction with members of the heat shock family and cofactors. This involves CHIP, an E3 ligase that interacts with Hsc70 facilitating autophagy. UPS controls the removal of misfolded proteins in the ER that are translocated into the cytoplasm, deglycosylated and finally degraded.

Diseases such as cystic fibrosis (e.g. ΔF508 mutation in CFTR) are caused by premature degradation of otherwise functioning proteins. In addition, there exist genetic diseases such as Angelman syndrome and Liddle syndrome that lead to mutations in E3 ligases. In the first case, the mutation occurs in E6-AP, which is also involved in p53 and leads, among other symptoms, to severe intellectual disability. In the second case, the mutation is in NEDD4 and causes excessive sodium retention, hypertension and electrolyte imbalances [28].

UPS plays an important role in preventing the genesis of cancer, in fact it regulates the cell concentration of p53, a transcription factor and tumour suppressor encoded by the gene TP53. It is involved in DNA repair, cell cycle arrest and in the induction of cell death, to prevent the propagation of mutated DNA. In MM, p53 is more frequently inactivated not through mutation, but through chromosomal deletions, epigenetic alterations or negative regulation by proteins such as MDM2 and MDM4. In fact, MDM2 is an E3 ligase that uses UPS to target p53. MDM4 enhances the effect of MDM2 even though it lacks ligase activity. The proteasome inhibitor Bortezomib prevents MDM2-mediated p53 degradation, stabilising it and promoting apoptosis of malignant cells. While effective in many patients, Bortezomib's utility is often limited by the development of resistance [29].

In order to maintain fast protein turnover and get rid of misfolded proteins caused by oncogene-driven stress, several oncogenic pathways take advantage of the UPS dysregulation. Other E3 ligases such as Skp2, COP1 and Pirh2, are overexpressed in various tumours and promote the degradation of tumour suppressors or cell cycle regulators such as p27 and p21. Conversely, a loss of function of E3 ligases such as Fbw7 can occur, leading to the accumulation of oncogenic substrates [25].

It follows from this that if the UPS mechanism is not functioning properly, it can cause several issues including dysregulation of the cell cycle, an incorrect inflammatory and immune response, and promote the survival of cells with damaged DNA.

These insights paved the way for new and targeted therapeutic strategies. Among these, a revolutionary technology is PROTACs that exploit the UPS mechanism to induce the selective degradation of disease-relevant proteins. In this way, it is possible to switch from passive inhibition to active removal of oncogenic drivers or tumour suppressors.

Therefore, pharmacological reprogramming of UPS via PROTACs represents a continuation of current therapeutic strategies and an opportunity to achieve higher specificity, lower toxicity, broader efficacy in genetically different myeloma subtypes, and overcoming resistance to currently used drugs.

The following sections will explore in more depth the mechanism, design principles and therapeutic potential of PROTACs, with a focus on their application in targeting E3 ligases such as UBR1, the candidate explored in this project.

1.3 PROTACs: Proteolysis Targeting Chimeras

By taking advantage of the UPS, heterobifunctional substances known as Proteolysis Targeting Chimeras (PROTACs) can specifically cause protein degradation. Sakamoto and Crews first introduced the idea of this technology in 2001, when they created the first PROTAC molecule, PROTAC-1. They demonstrated that methionine aminopeptidase-2 (MetAP-2), an enzyme known to inhbit angiogenesis, can be bound to SCFβ-TRCP, then tagged with ubiquitin and at the end successfully degraded. They proposed that PROTACs may be helpful research tools for altering the phenotype of cells via the targeted deletion of certain proteins or therapeutic approaches that concentrate on removing disease-promoting proteins [30].

Currently, strategies such as targeted drug therapies and gene therapy are already in place to treat complex disorders such as neurodegenerative diseases, autoimmune diseases and cancer. However, the drugs are mainly small molecules regulating proteins or enzymes and are based on traditional drug design principles as most diseases are accompanied by alterations in specific proteins. Moreover, patients can develop resistance to these drugs, especially in the advanced stage of some cancers.

In this sense, PROTACs can overcome the limitations of traditional therapeutic approaches and act on protein degradation in a targeted manner. This is the reason they have great potential in biological research and drug development.

A PROTAC molecule is structurally made up of three distinct components:

- The **warhead**: it recruits the protein of interest (POI) and can be designed to be selective only for the desired one
- The **anchor**: it is the portion of the synthetic molecule consisting of an E3 ligand that binds to the substrate binding domains (SBD) of the E3 ligase. This part of the PROTAC is the one that exploits the UPS mechanism to mark the POI with ubiquitin and allows its degradation to take place.
- The **linker**: it acts as a bridge and binds the warhead to the anchor, forming a ternary complex (POI:PROTAC:E3 Ligase). This portion is also essential because several studies have shown that its length and composition strongly influence the biological activity and physicochemical properties of the entire molecule [31].

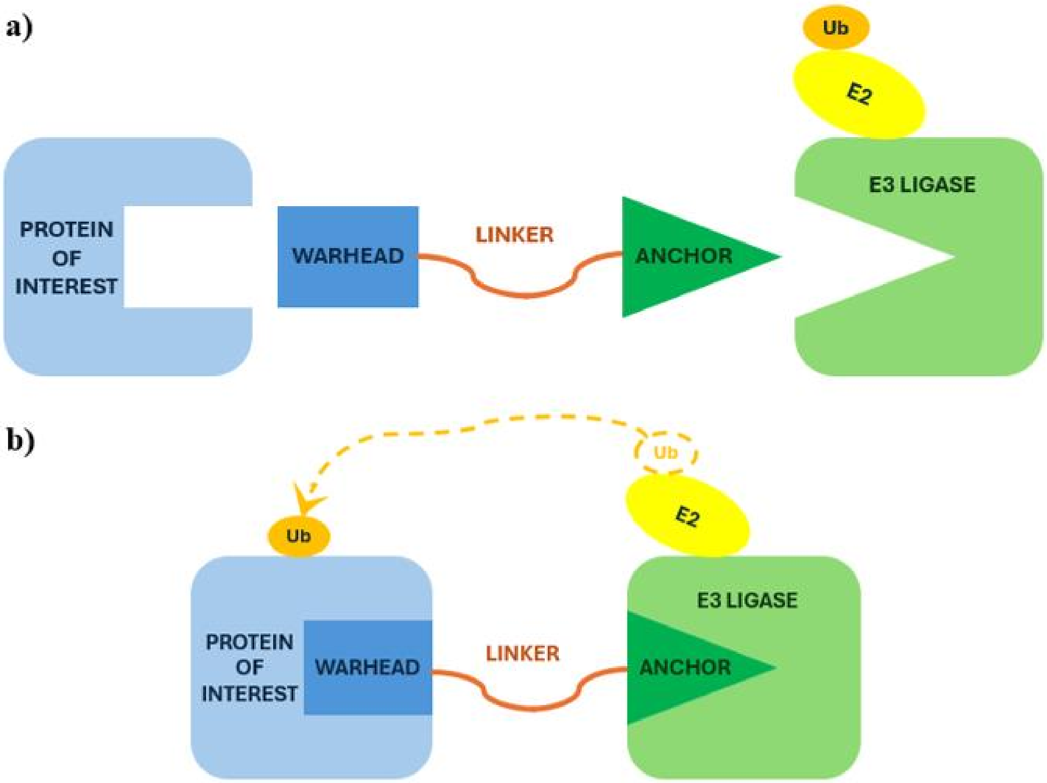


Figure 1.4: *Schematic image of the fundamental parts that constitute a PROTAC: the warhead, the linker and the anchor. In illustration a), E2 is bound to the E3 ligase enzyme, to which ubiquitin is in turn bound. Picture b) shows the bond between the warhead and the protein of interest and that between the anchor and the E3 ligase. In this process, ubiquitin moves from E2 to the protein of interest.*

1.3.1 Different Generations of PROTAC

PROTAC-1 developed by Sakamoto and Crews belongs to the first generation of this technology. In order to achieve the desired effects, however, the peptide molecules had to be microinjected directly into the cell because they could not permeate into it in any other way, and they also required phosphorylation to function and in general they were very unstable. Subsequent to this first report, other PROTACs were studied and developed that proved to work in an intact cell, especially in the specific degradation of androgen receptors (ARs) and estrogen receptors (ERs).

New E3 ligases were then introduced as Von Hippel-Lindau (VHL) such as in PROTAC for HIF-1α. In this case, microinjection was not necessary, did not require phosphorylation-dependent degrons for substrate recognition and offered greater flexibility. Although these initial technologies induced target degradation, they were functional only in the low-micromolar range and so they were not suitable for therapeutic development.

To overcome the limitations of the first PROTACs, the second generations were designed. The evolution over generations reflects the refinement of the concept, the resolution of technical constraints and the expansion of therapeutic applications.

Since 2008, entirely synthetic instead of peptide-based PROTACs began to be developed, and small molecule ligands began to be used for the recruitment of both POI and E3 ligase. The first non-peptide PROTAC used Nutlin, an MDM2-p53 PPI inhibitor, to recruit the E3 ligase MDM2 and degrade AR. It proved the viability of this method in terms of cell permeability, even though micromolar doses were still necessary.

Notable among the E3 ligases are CRBN (cereblon), which is bound to thalidomide and IMiDs, VHL, which remains widely used due to the availability of synthetic ligands (e.g. VH032) and IAP (inhibitors of apoptosis) that have a RING domain that presents the E3 ubiquitin ligase activity. Benefits from this introduction include enhanced cell permeability and greater metabolic stability.

PhosphoPROTACs were created in 2013, and their ability to stop tumor development in murine models provided the first proof that this technology functioned animal models, not only in cultured cells. The main feature is that the effect depends on the activation of receptor tyrosine kinase (RTK). It is designed to be activated only in response to a cellular pathway signal, making it controllable in time and space. A signal activates a kinase that phosphorylates the target protein at one or more key residues, and the phosphorylation creates a recognisable motif (degron) for a specific E3 ligase that recognises the sequence. In this way, the POI can be ubiquitinated and subsequently degraded.

Further progress has been made with HaloPROTACs, composed of a VHL ligand and a chloroalkane linker to covalently bind to the POI HaloTag7 protein. They have been used to degrade GFP-HaloTag fused proteins, with nanomolar efficiency (DC50 ~19 nM). In 2015, RIPK2 PROTAC was the first PROTAC VHL small molecule discovered with very potent activity (DC50 = 1.4 nM) and catalytic and E3-dependent action was confirmed. BET/BRD targeting PROTACs are instead against BRD4 epigenetic proteins with VHL or CRBN and are very effective in transcriptional regulation. CLIPTACs, developed by Astex, are also formed intracellularly and degrade targets such as BRD4 and ERK1/2 [32].

Between 2017 and 2018, the third generation emerged, with the introduction of PROTACs that can be activated, conditioned, or equipped with more sophisticated control mechanisms. One problem with PROTACs is that they can also act on non-target tissues and have toxic side effects, becoming dangerous. For this reason, some studies have tried to make this technology time and space controllable. One solution identified was using photo-caged PROTAC, the first of which were opto-PROTAC and pc-PROTAC. This technology works by binding a group of molecular species called photocages to the PROTAC, which are photodegraded when illuminated. This allows controlling the moment and point where the PROTAC is released.

For instance, in 2019, the team led by Wenyi Wei and Ian Jin converted the ALK and dBET1 inhibitors into photo-controlled PROTACs known as Opto-DALK and Opto-DBET1. The latter proved to be less toxic and more suitable for use in precision medicine although there was still a problem. In fact, this photodegradation process of the photocage was irreversible.

However, some studies have succeeded in overcoming this limitation by designing the groups to allow PROTACs to isomerise photochemically under the irradiation of light at different wavelengths. In this way the PROTAC can be activated and deactivated in a controlled manner.

The ability of RNA-PROTACs to degrade RNA binding proteins was introduced in 2019. They allow the degradation of RBPs, proteins found in cells that can bind to specific RNAs by exploiting the oligonucleotide sequence as a POI ligand. Alterations in RBPs are in fact associated with the development of several diseases. Taking the stem protein Lin28 as an example, it possesses a "zinc finger" domain in its C-terminal region, which allows it to bind microRNAs containing the specific sequence 5'-AGGAGAU-3'. This sequence was used to design a ligand capable of recognizing Lin28. To recruit the VHL E3 ligase, a peptide derived from the transcription factor HIF-1α, known for its affinity towards VHL, was used instead. By combining and optimizing these two elements, an RNA-PROTAC was developed that can selectively bind to Lin28 (in competition with its natural microRNAs) and lead it to degradation via UPS.

Another related technology introduced in 2021 is phosphorylation targeting chimeras (PhosTAC) but instead of recruiting the E3 ligase and degrading the POI, these regulate their activity by phosphorylating them. In recent years, beyond PROTACs, novel related technologies have emerged. For example, the Disney group developed the first RIBOTAC, a small molecule capable of recruiting RNase L (an endogenous cellular nuclease) and causing it to interact with a specific RNA. Once this complex is formed, the RNA is cut by RNase L, leading to degradation through natural cellular RNA turnover mechanisms.

More recently, the same team took an RTK inhibitor molecule used in oncology, Dovitinib, and reconverted it into a RIBOTAC designed to target pathogenic RNA. Compared to the already known molecule, it showed a 2500-fold greater selectivity towards the target RNA and reduced toxicity, since it acts exclusively on the pathological RNA [33].

1.3.2 Interest of Pharmaceutical Company

Currently, this new therapeutic approach has been used to degrade target proteins related to numerous diseases, including neurodegenerative diseases, viral infections, immune system disorders, and tumors. Thanks to their use, it is possible to eliminate proteins and therefore regulate signaling pathways that would not have been reachable with traditional approaches. They have drawn a lot of interest from the pharmaceutical and biotechnology industries as well as from academics.

Astrazeneca's PROTACs, which target B-cell lymphoma 6, are one example. The pharmaceutical company GKS has instead developed them for the target of both PCAF/GCN5, transcription coactivators, and IRAK4, important in the immune response. Pfizer has focused on targeting a factor important in lymphocyte signaling, BTK, while Boehringer Ingelheim on a factor involved in cell adhesion, FAK. In addition to therapeutic applications, AbbVie has shown that PROTACs, like other drugs, can also induce resistance mechanisms. In the meantime, Promega has introduced systems to monitor in real time the kinetics of degradation and the mechanism of action of PROTACs inside living cells. The pharmaceutical company Arvinas has instead developed particularly advanced PROTACs.

The first is ARV-110, an androgen receptor degrader active even against mutant variants that has shown good safety profiles in phase I of clinical trials. It is in phase II clinical trials for castration-resistant prostate tumor (mCRPC) [34].

In patients with ER-positive/HER2-negative breast cancer, ARV-471 is being tested, designed to degrade the estrogen receptor. In 2024, it received FDA Fast Track Designation for the Treatment of these patients and a phase III study is currently underway evaluating its efficacy in combination with palbociclib, another breast cancer drug.

In addition to these that are in advanced clinical trials, other PROTACs developed also show significant effects in reducing tumor volume and its regression in cellular and animal models. Other targets are epigenetic proteins such as those of the BET family (BRD2, BRD3, BRD4), involved in the transcriptional regulation of numerous oncogenic genes. Kinases, targets such as BCR-ABL (in leukemia), BTK (in CLL) and FAK (in metastases) were effectively degraded with PROTAC. A transcription factor considered “undruggable”, STAT3, was also successfully degraded [35].

1.3.3 Advantagies and Disadvantagies

The advantages of using this new therapeutic approach with unique biological and chemical characteristics are many, especially in the treatment of tumors. First of all, the degradation process is "event-driven" and PROTACs act as catalysts in this mechanism.

Doses can be reduced and even the frequency of administration can be lower than conventional drugs, in fact a single molecule can degrade multiple POIs. Ligands with high selectivity towards the target protein or the E3 ligase even with low affinity can still be efficient and successfully eliminate all the functions of the POI. The ability to break down proteins that would otherwise be regarded as "undruggable" is one of the biggest benefits. Many transcription factors such as c-Myc or RNA-binding proteins such as IGF2BPs do not have pockets accessible to traditional drugs. However, these proteins are fundamental both in the development and progression of the tumor. This technology instead uses oligonucleotides or low affinity ligands as baits to avoid having to necessarily identify well-defined pockets. Furthermore, by rapidly degrading the POI via UPS, they reduce the risk of compensatory feedback of the target protein, ensuring that therapeutic efficacy is not reduced and side effects are not increased. With small molecule inhibitors, however, this does not happen.

The main problem is the permeability inside the cells that likely involves passive diffusion and sometimes active transport, but their penetration mechanism has not yet been fully clarified. However, this represents a crucial factor for the effectiveness of this technology as it must necessarily access the intracellular UPS system. The known PROTACs have large dimensions with weights between 1000 and 2000 Da and a large polar surface [36]. Liquid chromatography coupled to mass spectrometry (LC-MS) is usually used to measure the permeability of molecules inside the cell, but due to the low intracellular concentrations, this technology is not suitable for PROTACs.

Different tests have been developed by different research groups to perform this verification. Zeng and colleagues introduced a test that provides the possibility of measuring cell permeability quantitatively and classifying them in real time. This method is based on the interaction of PROTAC with CRBN. Foley and his research group also designed a test that uses molecules with a chloroalkane tag that bind to HaloTag-GFP fusion proteins.

In general, however, these tests can introduce errors and an ideal method does not yet exist, but further studies in this direction are needed [35]. To overcome the permeability drawback, the CLIPTAC strategy has been proposed. It involves dividing the PROTAC into two smaller precursors that only rejoin once inside the cell, or, more simply, reducing the molecular weight. To reduce polarity, flexible linkers could be inserted that allow the creation of intramolecular hydrogen bonds. Nanoparticles or liposomes could also be used to facilitate the entry of the PROTAC inside the cell. As for the anchor that recruits the E3 ligase, permeable peptides could be inserted such as polyariginine sequences.

Another issue to be aware in PROTAC design is the Hook effect which causes this technology to have a non-linear dose-dependent behavior. It occurs when a decrease in degradation activity is observed at high doses of the substance. It happens that instead of forming the ternary complex, the PROTAC can bind individually the POI or the E3 ligase and this leads to the construction of inefficient binary complexes [PROTAC–POI] or [PROTAC–E3].

However, this does not mean that the PROTAC is ineffective but this is due to saturation and therefore greater attention is required in determining the optimal dose in vitro and in vivo [36].

Despite that, the successes in the degradation of several different POIs involved in multiple diseases demonstrates the potential of PROTACs. In addition to their ability to degrade notoriously difficult targets, these molecules can also overcome drug resistance, target non-enzymatic functions of proteins, and offer reversible chemical control.

PROTACs therefore represent a promising therapeutic strategy not only for tumors sensitive to current treatments, but also for those refractory and without effective clinical options.

However, high affinity ligands for POI and E3 ligase are necessary for their creation. It is possible to acquire undesirable off-target effects if the specificity is inadequate. For this reason, their design and optimization are essential.

1.4 E3 Ligase and PROTAC Anchors

E3 ligases affect most biological processes in eukaryotes, precisely because, together with E1 and E2, they are involved in the UPS mechanism of targeted protein degradation. It is responsible for the specificity of the substrate and, therefore, plays a key role in this process. It regulates the last step of the enzymatic cascade and is responsible for bringing the POI closer together and transferring ubiquitin to it. This feature can be exploited by new therapeutic approaches. By targeting specific E3 ligases, the selectivity of treatments could be increased, resulting in more effective treatments with fewer side effects.

1.4.1 E3 Ligase Families

Only 1-2 E1 and 10-20 E2 enzymes are known, while more than 600 E3 ligases are recognized. The latter can be divided into four main groups according to their composition. Each group has a low sequence homology with the others.

HECT family:

Many E3 ligases belong to the HECT family, i.e homologous to the E6AP carboxyl terminus. E2 transfers the ubiquitin molecule to this conserved HECT domain. A cysteine residue accommodates the molecule that will be used for POI marking and is necessary for thioester bond formation. In turn, this family can be divided into other subgroups, depending on the N-terminal domain, which is responsible for substrate recognition. There are 9 E3 ligases of the Nedd4 subgroup and they possess the C2 domain, which serves to facilitate ubiquitination of the target protein, and the WW domain. Another family stands out because it has one or more RLD domains that are important for interaction with chromatin and are involved in the regulation of Ran GTPases. This subgroup has 6 members. The E6AP ligase contains an AZUL (zinc finger) domain, which is crucial for its catalytic activity, and is the founding member of the HECT-type E3 ligases. The HUWE1 ligase contains domains detected in tumours, such as WWE and UBA.

RING family:

Unlike this family, there is another in which ubiquitin is transferred directly from E2 enzyme to substrate, without the formation of an intermediate bond: the E3 ligase only acts as a bridge. These ligases contain a RING (Really Interesting New Gene) domain and take their name from this. They have a domain containing zinc finger motifs for structural coordination and binding to E2. There are monomeric or multi-subunit RINGs. COP1, MDM2 and TRAF6 belong to the first category and possess domains for self-ubiquitination and substrate binding. On the other hand, some complexes that are crucial in controlling the cell cycle are larger in size because they consist of several subunits and use a scaffold protein, such as the Cullin-RING ligases (CRLs) that use cullines to assemble the different parts and allow easier interaction between the substrate and E2. Some E3 ligases fall into this category, such as SCF (Skp1/Cullin/F-box), which is among the largest, and the APC/C complex consisting of 19 subunits. The activity of the E3 ligases belonging to this family can be regulated through phosphorylation, interaction with small molecules or by attaching a small NEDD8 protein (neddylation).

RBR Family:

The E3 ligases RBR, RING-between-RING, possess three different catalytic domains: one that recruits E2 with the ubiquitin molecule (RING1), one intermediate (IBR) and the last one that contains the catalytic cysteine (RING2). This family has some properties similar to RINGs and others to HECTs. Similar to the latter, the ubiquitin transfer mechanism occurs in two steps whereby ubiquitin is transferred from the E2 to the RING2 domain and then to the substrate. Unlike HECTs, the E3s of the RBR family tend to form linear ubiquitin chains. An example of this is the Linear Ubiquitin Chain Assembly Complex (LUBAC), which consists of several subunits such as HOIP, HOIL-1L, Parkin and SHARPIN, and is able to modulate NF-κB signalling by this mechanism.

U-box ligases:

The U-box ligases are another, albeit small subfamily of E3 ligases. They contain RING-like domains of around 70 amino acids. This domain serves the transfer of ubiquitin from E2 to the substrate lysine and is therefore crucial for the performance of enzymatic activity. Also this family is crucial for post-translational protein quality monitoring [37].

1.4.2 Case Studies

Crews and his research group analyzed the differences between two PROTACs: one that acts by exploiting VHL and one with CRBN. The linker used was the same for both. It was observed that the endogenous substrates degradable by CRBN were more random whilst those of VHL were well defined. Crews and his group also investigated other E3 ligases and their ligands such as KEAP1, RNF4, RNF114, DCAF15, and DCAF16 that have shown increased ability to degrade non-natural substrates. The CRBN ligand library has been expanded with new thalidomide analogues such as CC-122, CC-220, and CC-885 [35]. The fact that CRBN ligases can promote the recruitment of neo-substrates can at the same time lead to unwanted off-target effects. For this reason, methods such as immunoblotting were used [32].

In general, the recruited E3 ligase can also influence the binding site between the PROTAC and the POI, thus conditioning the activity of the entire molecule, increasing it, decreasing it, or leaving it unchanged. For this reason, it is crucial to carefully choose the E3 ligase to use for this therapeutic approach [35]. On the other hand, one of the problems encountered for example with the ligands used for cIAP1 that made their use limited was the specificity and the self-degradation of the ligase itself. There was a study which aimed to analyze six different E3 ligases, chosen so that they represented the three main classes. Among these, β-TRCP, already used in the past, and Parkin, which represents a promising example of recruitable E3, emerged [32].

Since these enzymes are tissue-tumor specific, the identification of E3 ligases and their ligands should consider the pathological cellular context in which they will act, to increase efficiency and reduce side effects. Despite their importance, very few have been studied and employed in PROTACs, less than 10 out of all those discovered [36]. This represents a limitation for the evolution of this technology.

Currently, clinical-stage PROTACs aim to exclusively recruit E3 ligases such as VHL and CRBN because initial research efforts focused more on this direction. On average, TPD investigations have utilized fewer than 2% of the hundreds of E3 ligases found in the human genome. Developing new E3 ligases may be a solution to the problems of off-target toxicity, reducing side effects and acquired drug resistance. For example, some patients with MM may manifest genetic alterations of CRBN and may render PROTACs based on this E3 ligase ineffective. This could also allow access to a larger set of degradable proteins, as target selectivity varies greatly between different E3 ligases [38]. Thus, it is highlighted how urgent it is to find novel E3 ligases and choose new ligands. This project aims to identify new possible ligands of UBR1 E3 ligase that can be the anchor of this new promising technology.

1.4.3 UBR1 E3 Ligase

Protein quality is crucial for healthy cells. All misfolded proteins must necessarily be eliminated because they are potentially toxic. Until 2008, the ligases involved in the degradation mechanism of these proteins in the endoplasmic reticulum were known, but their particular role in the cytoplasm had not yet been identified. That year, a study revealed the potential of UBR1 as an E3 ligase capable of acting in this sense, thus responsible for quality control of misfolded cytoplasmic proteins and their degradation (QC).

UBR1 in yeast has a molecular mass of about 225 kDa, while in humans, the UBR1 gene (Gene ID: 197131) located on chromosome 15 (15q15.1), encodes a protein of approximately 200 kDa consisting of 1758 amino acids. The research was conducted by means of genetic screening on Saccharomyces cerevisiae and it was discovered that UBR1 represented a key component for the degradation of the chimeric misfolded protein ΔssCL·myc [39], [40].

The researchers revealed that this E3 ligase, a member of the RING family, is a component of the N-rule pathway, "a proteolytic pathway whose physiological targets include proteins with destabilizing N-terminal residues" [41]. This rule links the identity of a protein's N-terminal residue to its average in vivo life; in fact, it serves as a "life or death" signal for the protein. If the protein starts with a destabilising amino acid, it is recognised as being eliminated and degraded by UPS. It is a highly conserved pathway and in eukaryotes is part of the ubiquitin system, comprising two branches that act together to degrade the majority of cellular proteins:

- N-terminal acetylated residues are recognised by the Ac/N-end rule;
- Non-acetylated residues, such as Arg, Lys, Leu, Phe, Trp, Tyr, Ile, are identified by the Arg/N-end rule.

The UBR1 E3 ligase acts as a N-recognin and belongs to the second pathway that involves arginylation, i.e. the addition of an arginine to the N-terminal residue. Non-acetylated N-terminal residues can be first-class (basic, such as Arg, Lys, His) and hydrophobic and bulky then second-class (such as Leu, Phe, Trp, Tyr, Ile). There is a further subdivision that speaks of three categories of N-terminal residues in the Arg/N-end rule. Arg, Lys, Leu and Trp are recognised directly by E3 ligases of the UBR type and are therefore called primary destabilising residues. Asp and Glu, on the other hand, require modifications such as arginylation to become like the previous ones. Tertiary residues such as Cys, Gln and Asn must undergo modifications such as oxidation or deamidation before they can be arginylated.

In addition to the degradation of potentially toxic proteins, the Arg/N-end rule pathway contributes to the regulation of protein homeostasis, the response to oxidative stress, the control of the cell cycle and meiosis, neurogenesis and cardiovascular development.

The actual recognition of the degron by UBR1 occurs via very specific binding sites:

- **Type-1** is for basic residues: It is located in the UBR box, a domain of about 70-80 amino acids, which characterises all members of the UBR family. It can bind Type-1 destabilising residues: Arg, Lys, His. Its crystal structure has been found in both yeast and mammals, and it has been noted that it can bind the positively charged groups of the side chains of basic residues by means of a negative electrostatic pocket. The binding is highly specific. Biochemical data confirm its specificity and affinity towards N-degron substrates as UBR1 can directly interact with N-terminal peptides containing destabilising residues with low micromolar affinity (Kd ≈ 1 μM).

- **Type-2 is** for hydrophobic residues: It is located close to Type-1 and, although its crystal structure has not yet been determined, targeted mutagenesis and binding assays have shown it to be an independent domain. It recognises hydrophobic and bulky residues at the N-terminal position, such as Leu, Phe, Trp, Tyr and Ile.

- The third binding site is allosteric and is normally inactive under basal conditions. It serves to extend the substrate spectrum and to respond dynamically to the presence of multiple degradative signals. This occurs through the recognition of non-N-terminal internal degrons in proteins such as Cup9, a transcriptional repressor in yeast.

In mammals, there are also other homologues that may lack both type 1 and type 2 sites, examples include Ubr2, Ubr4 and Ubr5. The latter does not have the type 2 site. In addition, some members containing the UBR domain (Ubr3, Ubr6, Ubr7) do not participate in N-degron recognition, so they cannot be defined as N-recognins. Ubr2 is highly conserved as Ubr1 and performs similar functions as it is also an N-recognin.

Ubiquitination is enabled by the formation of the complex called the UBR1-UFD4 double E3 system, which is thought to be highly conserved in all eukaryotes. UFD4 is an HECT E3 ligase that, although it is not a N-recognin, facilitates the addition of K48-linked polyubiquitin chains to POI, increasing the efficiency of the degradation process. This means that Ubr1 is also able to interact with substrates of the UFD pathway, suggesting a functional overlap between the two degradative pathways and flexibility of action.

In 2023, to facilitate the development of these therapies, Liu and his group developed a web portal to help rapidly identify E3 ligases with promising activity against desired targets. Specifically, they characterised 7 aspects: chemical ligandability, expression patterns, protein-protein interactions (PPI), structure availability, functional essentiality, cellular localisation and PPI interface [38]. UBR1 was classified with a ‘confidence score’ of 5/6, indicating a high level of experimental evidence and biological annotation in the context of the UPS. According to this portal, it is expressed in several human cell types and tissues, including liver, kidney, pancreas and brain. Its expression has been confirmed at both bulk and single-cell RNA-seq levels, and its subcellular localisation is predominantly cytoplasmic, consistent with its role in cytosolic protein quality control (QC) [42].

The crucial role of UBR1 in the protein degradation system is supported by the fact that its alterations can generate cell toxicity, inflammatory response and metabolic dysregulation. As a result, mutations at the type 1 or type 2 sites impair its ability to recognise specific N-degrons, leading to abnormal accumulation of proteins that should be degraded.

The ability to read N-terminal and internal degradation signals, thanks to its three distinct sites, makes UBR1 an example of an E3 ligase with multivalent recognition capability that is crucial in proteostasis regulation. This justifies its interest and use in PROTACs. This project aims to identify potential ligands for its use as an anchor of this new technology.

1.5 Computational Approaches in Drug Discovery

On average, it takes 15 years and US$2 billion to discover and develop a small molecule drug. In the 1970s, the concept of computer-assisted drug discovery (CADD) was introduced and is still an integral part of the process. Although there have been periods of both enthusiasm and disillusionment, it has been shown that this new approach leads to a reduction in preclinical efforts, which alone account for 43% of the pharmaceutical industry's expenditure, with a reduction in costs and development time. Most resources are invested in clinical trials, but their failure rate is still 90%, which is very high.

Advances in biotechnology and life sciences have been remarkable in recent decades, but this achievement is partly explained by issues arising from the early stages of drug discovery and development (DDD). Problems may occur from inadequate target validation or sub-optimal ligand properties. If their properties could be more strongly confirmed, the failure rate of clinical trials would decrease markedly. Validation of characteristics such as pharmacokinetics (PK), ADMET properties (excretion and toxicity, absorption, distribution and metabolism profiles), also using a computational approach, facilitates the development of safer, more effective and affordable drugs.

In recent years, pharmaceutical and biotechnology companies have been building physics-based approaches, deep learning (DL) and artificial intelligence (AI), investing millions of dollars and hiring the first computational chemists. These efforts are producing an increasing number of clinical candidates in a very short time frame: 1-2 months to go from target to lead compound and less than a year to go from target to clinical trial.

Other factors also contribute to this result, such as new technologies enabling automation in crystallography, microcrystallography and cryo-electron microscopy have made it possible to obtain the 3D structures of most clinical targets. One example is the structural progress and provision of 3D models of G protein-coupled receptors (GPCRs) and other membrane proteins that mediate the action of over 50% of drugs. These models are used for ligand screening and lead optimisation [43].

1.5.1 Protein Data Bank and AlphaFold

The three-dimensional structures of biological macromolecules obtained through these techniques are included in the Protein Data Bank (PDB), the first and main public database founded in 1971. In 1999, the archive contained 10,000 structures, whereas today it has grown to 230,000 entries, including proteins, nucleic acids, macromolecular machines and their complexes with small ligand molecules among which approximately 191,000 structures determined by X-ray crystallography (MX). In the 1980s, with the advent of Nuclear Magnetic Resonance (NMR) spectroscopy, it became possible to measure protein dynamics and study intrinsically disordered proteins. Currently, there are more than 14,000 structures determined using this method.

The first structure in the PDB discovered using cryo-electron microscopy (3DEM) dates back to the 1990s, and by the end of February 2025, 24,379 structures had been archived. This technique does not require crystals and is suitable for larger macromolecular systems and heterogeneous samples. Thanks to significant improvements in the technology, there has been a significant improvement in resolution. Starting in 2014, with the discovery of the structure of a Nup-84 sub-complex of the Saccharomyces cerevisiae nuclear pore complex, approaches combining information found using multiple methodologies, both biophysical and computational (integrative/hybrid methods, IHM), began to be used. These structures were deposited in the PDB-Dev prototype database, but at the end of 2024 they were officially integrated into the Protein Data Bank under the new section “PDB-IHM structures”.

Now each model has both a PDB-Dev identifier and a standard PDB identifier, and the platform has been renamed PDB-IHM to reflect its focus on complex structural models obtained using multidisciplinary approaches. For publication in scientific journals and for funding, data deposition, which was initially voluntary, has become mandatory.

In 2003, in order to maintain a single, global and consistent version of the PDB, the international consortium Worldwide Protein Data Bank (wwPDB) was established. It is made up of US, European and Asian organisations and promotes the adoption of common formats, shared scientific guidelines and data interoperability. The structures deposited are validated and made available according to the FAIR, FACT and TRUST principles, which promote quality, transparency and sustainability in scientific data management, ensuring its reuse, accuracy and accountability in the long term.

Initially, the 3D coordinates of atoms in the first structures were stored using the Diamond format, which was simple and suitable for the computers of the time. The “classic” PDB format was then introduced, which contained information with a fixed length of 80 characters but did not allow complex or very large structures to be represented.

In the 1990s, the mmCIF (macromolecular Crystallographic Information File) format was created, which can represent complex relationships between atoms, residues and chains without limits and is ideal for automatic analysis by software. In 2014, the mmCIF format was adapted to contain structures identified using different techniques and is currently called PDBx/mmCIF.

The system used to store the structural data obtained is the unique, global OneDep platform, which performs automatic validation and identifies each structure with a unique PDB code. PDB bio-curators manually examine each structure to correct any errors, standardise names and formats, and add metadata. The average one-time cost to deposit a new PDB structure in 2023 was approximately $420 (less than 1% of the estimated determination cost), while the annual storage cost is approximately $10 per structure.

PDB serves as a model for international collaboration in data management. It has now become crucial for drug discovery and development because it provides insight into the actual structure of the target. It has accelerated the development of structural biology, as new structures can be reconstructed more quickly from those already known.

In addition, 3D structures allow us to understand how proteins and viruses work, as has been the case with HIV-1 and SARS-CoV-2, and to design drugs that bind specifically to these molecules. More than 70% of small molecule drugs approved between 2010 and 2018, and all new cancer drugs between 2019 and 2023, were developed with the help of PDB structures. The fact that the PDB is free has greatly stimulated the biotech industry: companies and researchers can use it freely to develop new products, patents and drugs. A 2017 economic study estimated that the use of PDB data via the RCSB.org website alone generated an aggregate economic value of approximately $9.2 billion.

The existence of the PDB archive has also been of fundamental importance for the evolution of structural bioinformatics, homology modelling, computational docking and de novo protein structure prediction. The accuracy of structural prediction has been revolutionised thanks to AlphaFold2 and RoseTTAFold, expanding the number of structures available for structure-based ligand discovery. These tools learn from thousands of real examples in PDB the implicit rules that guide protein folding. Today, RCSB.org not only shows experimental structures but also includes these computational models [44]. Although there have been debates about the usefulness or not of these models, recent results on docking campaigns have shown equal effectiveness between those that used the predicted structures and others that used experimentally found structures [45].

An early version of AlphaFold was introduced and tested in the 13th Critical Assessment of Protein Structure Prediction (CASP13). This system, which preceded AlphaFold2, already used machine learning, but it was only the latter that enabled a significant breakthrough in protein structure prediction, being the first computational technique that can consistently predict protein structures with atomic precision even when no comparable structures are available.

This model, developed by DeepMind, integrates new neural network architectures and training procedures based on evolutionary, physical and geometric constraints of protein structures that allow the direct prediction of the 3D coordinates of all heavy atoms in a protein. This neural model consists of two stages: Evoformer and Structure Module [46]. To understand in detail how it works, see Subchapter 2.1.2.

AlphaFold3 (AF3) is an evolution of AlphaFold2 and its prediction capabilities to include complexes of almost all molecular types present in the PDB. Proteins, small molecules, ions, nucleic acids (DNA, RNA), and modified residues are examples of this. This is because it far exceeds the accuracy of previous specialised tools for protein-ligand, protein-nucleic acid and antibody-antigen interactions. AF3 replaces AF2's Evoformer with a simpler “pairformer” module, as it allows less direct dependence on multiple sequence alignments than the previous method. In addition, it directly predicts raw atomic coordinates with a diffusion module, instead of the AF2 structure module that operated on specific frames for amino acids and side chain torsion angles [47].

1.5.2 Virtual Libraries

Another factor is the significant change over the years in the quantity and accessibility of chemical molecules. Until a few years ago, pharmaceutical companies could only test a few million molecules from physical libraries, i.e. from compounds that they actually had in their laboratories or could buy, thus with limited access and high costs.

Today, virtual libraries exist, and thanks to computational chemistry and 3D models, it is possible to digitally simulate the behaviour of billions of molecules, thus much faster, by exploring enormous chemical spaces equipped with software for virtual screening.

For this approach to be effective, it is essential to expand the size of the screening libraries as much as possible. This has long been the impediment for identifying new successful and target-selective ligands. If the selection is not effective, it is necessary to spend years to perform the optimization. A typical high-throughput screening (HTS) campaign uses about 50,000-500,000 compounds but a secondary validation is needed, in fact, this method can give rise to false positive or false negative results. DNA-encoded libraries (DELs) were also constructed by conjugating ligands with unique sequences via a linker. However, this limits the number of chemical reactions that can be used and can produce incorrect results due to blocking of imported portions for binding or non-specific interaction of DNA tags. Further validation is therefore required with this method as well. Hence, it was thought to employ machine learning (ML) models that were trained on the outcomes of DELs for every target from on-demand chemical spaces.

The expansion of libraries to millions of compounds is only possible for big pharmaceutical companies, but they are still too few to solve the problem. Only recently has the expansion to billions of compounds (gigaspace scale) occurred, which can change the method of drug discovery. As a result, extremely strong ligands with affinities frequently in the mid-nanomolar or sub-nanomolar range have been discovered.

These libraries are on-demand: a compound is synthesized only when it is needed, such as after being identified as a potential candidate in a virtual screening. This eliminates the need to physically synthesize and store millions or billions of compounds that may never be used. Furthermore, when a molecule is found to be active towards the target, optimisation of properties such as affinity, selectivity and stability is carried out. Through chemical modifications, by varying functional groups, side chains or rings, a so-called SAR (Structure-Activity Relationship) analysis is performed, which consists of testing these new compounds. With traditional methods, this is done by means of resource-intensive custom syntheses in the laboratory. In on-demand libraries these compounds are already present, so they can be ordered directly or used for in silico simulations: this method is called SAR-by-catalogue. However, a proper gigascale library must contain compounds with chemical diversity, i.e. with different scaffolds, shapes and functionalities, and must not resemble those already present in databases of known molecules. This increases the likelihood of discovering unique and patentable molecules.

Example of such libraries is Enamine REAL, the first of its kind developed in 2017 and grown from 170 million to 5.5 billion compounds in 2022 and has currently reached 76 billion compounds [48].

Other commercially available libraries are CHEMriya with 55 billion compounds and GalaXi with 2.1 billion [49]. These libraries were able to expand further with the addition of new predefined reactions (or transformations) and a specific set of building blocks, being handled as non-enumerated chemical spaces. Pfizer's PGVL is the first virtual chemical space. Compounds have less than 10% overlap and are diverse.

Chemioinformatics tools have been developed to navigate them without being enumerated, e.g. fragment-based chemical similarity searches or more complex 3D techniques like the Rapid Isostere Discovery Engine (RIDE), which is based on atomic property fields. The generation of hypothetically synthesisable compounds using generative chemistry based on deep learning (DL) is also part of an alternative approach and is used in the de novo design of ligands. GDB-17, which has 166.4 billion molecules with up to 17 atoms (C, N, O, S, and halogens), and GDB-18, which contains about 10^{13} compounds, are two examples. However, the fact that success rates must be validated remains in order not to lose the advantage of low time and cost: this is the case for libraries such as Enamine REAL Space, but not for others [43].

Enamine (REAL), WuXi (GalaXi), and also Mcule (Ultimate) are an important part of the ZINC-22: a freely available database of small chemical compounds, primarily designed for ligand discovery through virtual screening and molecular docking. This also includes the stock portion of the previous ZINC20. It has evolved from a database of millions to billions of molecules. At the end of 2023, it contained over 37 billion commercially available 2D compounds, of which over 4.5 billion were constructed in biologically relevant “ready-to-dock” 3D formats but it adds approximately 300 million new molecules per month. It includes physical properties relevant to docking, such as conformations, partial atomic charges, cLogP values, and solvation energies. It is accessible online via an easy-to-use graphical user interface (GUI), Cartblanche, which allows you to quickly identify molecules similar to a given structure. You can also search by ZINC ID and Supplier Code.

ZINC-22 has been reorganised as a set of many smaller databases, which can be prepared asynchronously, concurrently and scalably. Molecules are grouped into ‘tranches’ based on heavy atom count (HAC), lipophilicity (calculated Log P), charge, and size. Despite rapid growth, the chemical diversity of the database continues to increase. This database includes compounds covering a wide range of molecular sizes: up to 29 heavy atoms (HAC), with a potential extension up to HAC34. More than 95% of compounds up to HAC24 and more than 80% up to HAC25 are represented by pre-generated 3D structures. This makes it particularly suitable for structure-based virtual screening [50].

The challenges faced by computational approaches are still many. Chemical spaces have to maintain a high diversification but must also be “drug likeness”. If this happens, they are more likely to contain millions of initially active “hits” and thousands of optimised “leads” molecules for any given target. This makes further medicinal chemistry work toward potential drugs easier.

Accuracy is a key parameter because it protects against false positives. To have 10,000 false hits, in a library of 10 billion compounds, would be enough to have a false positive rate of one in a million and cause artefacts in candidate selection. To avoid this problem, strategies are adopted such as using two different scoring functions to evaluate the same molecule and choosing those that score well in both, and since artefacts tend to cluster in similar compounds, it is more functional to select highly different hits. Since docking or scoring scores are not perfect, one can choose from several score ranges (top, middle, borderline), thus increasing the probability of identifying true actives. Before moving on to experimental testing, a manual review is always recommended to check binding modes, looking for unusual interactions, docking artefacts, unrealistic poses or steric collisions. If there are regions with low score discrimination to which no real hits belong, the function can be improved to more accurately separate active and inactive.

The large number of libraries also brings a great advantage in handling false negatives: the discarding or classification as unpromising of some molecules that are true actives is well tolerated. If we assume there are 1 million potential true hits in a 10 billion database and screening misses 50% of them (i.e. 500,000 molecules), that still leaves another 500,000.

The speed for handling gigascale libraries must be appropriate, as if one wanted to screen 10 billion compounds on a single CPU core and if docking a compound took 10 seconds per CPU core, it would take over 3,000 years at a cost of about a million dollars on cloud computing at the cheapest CPU rates. Indeed, with the increasing size of libraries, however, computational time and cost are becoming the main bottleneck of the screening process [43].

Because of the limited resources, it was decided to use the NCI Diversity Set, a collection of chemical compounds provided by the US National Cancer Institute (NCI), as the library for virtual screening. It is selected to maximise chemical diversity as it contains a representative subset of the NCI's larger collection of molecules (Developmental Therapeutics Program, DTP) enabling rapid virtual screening [51]. To expand the series of hits and start SAR optimisation, a similarity search was then performed in ZINC-22. The rationale and characteristics of the NCI Diversity Set, the use of ZINC-22 for similarity search and the strategy adopted in this project will be further explored in Chapter 2.

1.5.3 Modular Synthon-Based Approaches

An approach that aims to address the challenges of exploring immense chemical spaces and the need for easily synthesizable compounds is the modular approach based on synthons. A synthon is a structural unit or molecular fragment considered as a building block in the design of more complex molecules. It is based on the idea of designing drugs with desired properties using these building blocks as pieces of a puzzle. The main obstacle that slowed down the development of this method was the customised synthesis of the compounds, but thanks to V-SYNTHES (Virtual Synthon Hierarchical Enumeration Screening) technology, this limitation has been overcome.

The method involves the preparation of a minimum library of representative fragments from the synthons available in REAL Space. These are all tested for a given scaffold attachment point while the other points are temporarily blocked with small inert groups to reduce computational complexity and prevent reactive groups in the building blocks from otherwise creating strong but false interactions. The fragments are then subjected to a docking-based screening to select only those with the best scores. The process is repeated for the remaining binding positions. Next, the top-ranked compounds are subjected to more accurate docking and then filtered according to the variety of desired properties.

The advantages of this approach are computational efficiency, promising high hit rates (23% hit rate for submicromolar ligands) even with robust chemical novelty as the initial selection does not depend on known ligands. Once candidates have been identified, analogues can be explored, making optimisation straightforward because it allows a broad exploration of structure-activity relationships (SAR-by-catalogue), reducing costly custom syntheses [43] [45].

1.5.4 Structure-Based Virual Screening

Another type of advanced computational method is one of the most established and it can be used to find hits and leads. It is based on receptor structure and is called Structure-Based Virtual Screening (SBVS). Thanks to the advent of gigascale libraries, it is used in the early phase of drug discovery to predict the quality of the bond, and its purpose is to explore and categorise a large chemical space very quickly. SBVS exploits the three-dimensional structural information of a protein target to identify and optimise molecules, taken from virtual libraries, that can successfully bind to it. The system also allows the prediction of a binding score, giving the possibility to prioritise and order compounds according to the results, particularly relevant as the size of libraries increases.

Several techniques can be adopted to increase the speed and accuracy of the final results, including molecular mechanics in internal coordinates, using force fields to calculate energy and modelling interactions between atoms. This approach allows for the rapid exploration of flexible conformations of the ligand during docking. Based instead on geometric matching are the empirical 3D shape-matching approaches that compare the three-dimensional shape of the ligand with that of the binding site or with known ligands. Alternatively, a multi-step “docking funnel” method can be adopted, in which the number of ligands involved is progressively reduced by applying different, more complex and accurate methods.

The introduction and use of this computational technology was also driven by the structural revolution, in fact, the automation of crystallography, microcrystallography and cryo-EM technologies were essential in this process. Without this crucial step, it would not have been possible to enrich the libraries with high-resolution 3D structures of clinically relevant chemical targets [43]. Moreover, 3D structures often reveal the target in a state or molecular complex relevant to its biological function. In order to achieve a more accurate screening, with fewer false negatives and greater reliability in predicting binding, the most promising starting point is the use of a “holo” structure, i.e. bound to a ligand. The latter, in this configuration, is already located within the binding pocket, which, being modelled around it, is more realistic, compact and suitable to accommodate molecules, reflecting the natural interactions between ligand and protein. This is not the case if one starts with an “apo” structure that, not being bound to any ligand, might have the binding pocket partially collapsed or closed [52].

To achieve the goal of this project, this approach was chosen because it is able to separate a small fraction of potential ligands from a large number of compounds that lead to failure [52]. Moreover, in experimental tests, other structure-based prospective screening campaigns have shown success rates of 10-40%, producing hits with binding affinities in the 0.1-10 μM range. Currently, campaigns using the three-dimensional structure of protein targets have been shown to work on almost all classes of targets including enzymes, signalling proteins, nuclear receptors even on GPCRs, which have historically been difficult to study computationally due to structural complexity and conformational dynamics. To evaluate the performance of virtual screening algorithms and compare them objectively, public competitions such as the D3R Grand Challenge are organised. The results of these challenges, used as benchmarks, showed a steady improvement in binding mode and binding affinity prediction mainly due to improved algorithms, force fields, scoring functions and artificial intelligence [43].

1.5.5 Molecular Dynamics as a Tool for Model Refinement

For some chemical systems, predicting relevant properties is difficult or too costly to manage experimentally. In these cases, molecular dynamics (MD) simulations can be used to complement experimental data. This is a versatile and powerful computational methodology that continues to provide crucial insights in a wide range of scientific contexts. This approach enables characterisation of membrane structure, organisation and permeability, lipid-protein and lipid-drug interactions, protein-ligand interactions, and protein structure and dynamics. MD simulations have long been employed in the discovery of new drugs, such as enzyme inhibitors and therapeutic targets for cancer [53].

It is a deterministic approach that forecasts future states by using the system's current states. Its actual dynamics can therefore be calculated from pre-existing structural information. This is possible thanks to Newton's equations of motion, which can be used to calculate atomic forces and predict changes in the position of atoms over time. For this reason, MD simulations are used to investigate the conformational transition of proteins caused by mutations or ligand binding/unbinding. It provides detailed atomic insights that are difficult to obtain with traditional biochemical or pathological experiments and it offers non-static snapshots that show the movement of each amino acid at the atomic level. It is capable of replicating both in vitro and in vivo environments, for instance at different pH levels, in the presence of water and ions, at different salt or ion concentrations, in the presence of a lipid bilayer and other biological constituents.

It is used to examine protein function, determine their stability and analyse enzymatic reactions. Specifically, MD simulations have been applied to the analysis of membrane proteins, including their structure and organisation, permeability, and lipid-protein interactions. They have been used to evaluate conformational changes between active and inactive states in studies on G protein-coupled receptors (GPCRs), also caused by lipids. MD has also been used extensively to understand the relationship between SARS-CoV-2 and the human host, particularly the structural and conformational basis of new mutations in the virus's spike (S) protein and their binding to the ACE2 receptor. It has also been used to screen for inhibitors of the S protein.

These methods constitute an effective guide for scientific research and facilitate the advancement of medical treatments [54]. As with the advent of gigascale libraries, MD simulations are possible thanks to the growing availability of experimentally determined protein structures, but also to the wide availability of graphics processing units (GPUs), which allow simulations to be performed locally. Over the years, from 1980 to the present day, their use has increased exponentially: by 2021, there were almost 18,000 published studies using MD simulations. It is currently common practice to simulate proteins with hundreds of amino acid residues surrounded by water and salt for time scales of 10-100 nanoseconds. GROMACS, AMBER, CHARMM, DL_POLY, NAMD and LAMMPS are examples of widely available user-friendly platforms for this purpose [55].

MD simulations are therefore a fundamental tool for understanding the relationship between protein structure and function and guiding drug development. In this project, they were used to analyse conformational fluctuations in both the homology model of the UBR-box of UBR1 E3 Ligase and the complete protein predicted by AlphaFold2.

2. Materials & Methods

The project aims to identify small molecules that binds to UBR1 E3 Ligase, a key component of the ubiquitination pathway. These molecules will serve as the anchor of a synthetic PROTAC, designed to promote the targeted degradation of proteins associated with MM.

To achieve this, a combination of computational methods was employed, ranging from protein structure modelling and refinement to virtual screening and molecular docking. The following sections describe in detail the materials and methodologies adopted in each step of the workflow.

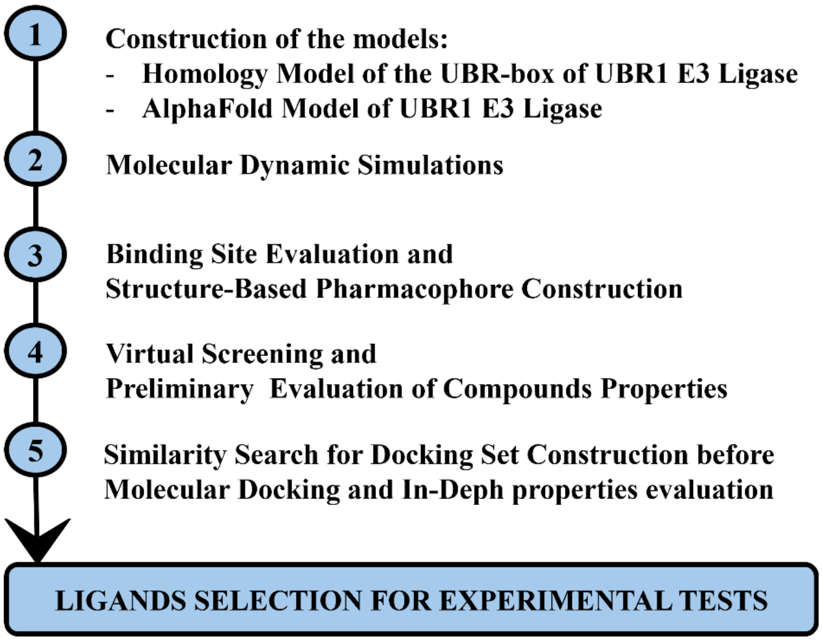

1
Construction of the models:
- Homology Model of the UBR-box of UBR1 E3 Ligase
- AlphaFold Model of UBR1 E3 Ligase
2
Molecular Dynamic Simulations
3
Binding Site Evaluation and
Structure-Based Pharmacophore Construction
4
Virtual Screening and
Preliminary Evaluation of Compounds Properties
5
Similarity Search for Docking Set Construction before
Molecular Docking and In-Deph properties evaluation
LIGANDS SELECTION FOR EXPERIMENTAL TESTS

Figure 2.1: *Summary of the project's methodological process with a brief description of the various stages involved, from the initial model construction phase to the ligands selection for experimental tests.*

2.1 Protein Structure Generation

The advent of new experimental techniques has led to a structural revolution that has enabled the discovery of an increasing number of protein architectures (see subchapter 1.5).

Despite this, not all structures have been experimentally resolved. This is the case with the human E3 ligase UBR1, whose crystallographic structure has not yet been completely determined and deposited in the RCSB PDB. Only some of its domains can be found in this database. Examples relating to Homo sapiens are:

- 3NY1, which represents the UBR-box domain of the UBR1 protein with a refinement resolution of 2.085 Å;
- 5TDC, which shows the UBR-box domain of UBR1 in complex with a methylated peptide with a refinement resolution of 1.607 Å.

They were released in 2010 and 2017, respectively.

- 9V0K, which also shows the UBR box but with a refinement resolution of 1.54 Å was released more recently, in June 2025.

All structures mentioned were determined using the X-ray diffraction method.

Other configurations have been determined in Saccharomyces cerevisiae S288C with electron microscopy method, such as 7MEX (structure of the yeast UBR1 in complex with UBC2 and N-degron) and 7MEY (structure of the yeast UBR1 in complex with UBC2 and monoubiquitinated N-degron), released in 2021. However, they have a refinement resolution greater than 3 Å [56].

Since there is no complete and accurate experimental crystallographic structure of UBR1, two different approaches were used to obtain its model:

- the use of **homology modelling** for the UBR-box domain;
- structure prediction with **AlphaFold2** for the entire protein.

Both methods are discussed in detail below.

2.1.1 Homology Model Construction

It was decided to use a related protein, UBR2. It has a similar sequence and structure, making it a good model for modelling UBR1. This approach is justified by the need for a more accurate three-dimensional representation for virtual screening and docking studies.

Specifically, the 3NY3 structure was chosen from the RCSB PDB database, which represents the structure of the UBR2 UBR-box in complex with 1 type N-degron. This structure was selected because the specific UBR-box domain is responsible for the specificity of the ligase as it is involved in recognizing protein degradation signals. Furthermore, despite being deposited in 2010, it has a resolution of 1.60 Å. The method used for its determination is X-ray diffraction. The study by Matta-Camacho and colleagues reports the crystal structure of 3NY3. The new protein structure adopts a previously undescribed fold and is stabilised by three zinc ions to form a binding pocket for type 1 N-degrons. N-terminal arginine is preferred, according to NMR studies [57].

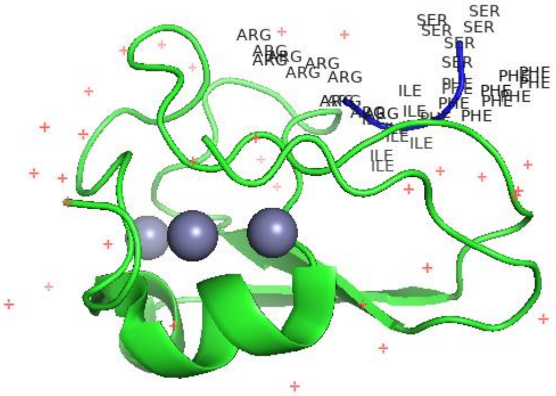


Figure 2.2: *Structure of 3NY3: UBR2 UBR-box in complex with 1 type N-degron. The UBR-box is shown in green, the three zinc ions in grey, the water molecules in red and the N-degron in blue. N-degron is composed of the residues Arginine (ARG), Isoleucine (ILE), Phenylalanine (PHE), and Serine (SER). They are also called RIFS, and they are included in the B-chain. The image was created using PyMOL software.*

Protein sequence of 3NY3 in *FASTA* format (canonical) from PDB is:

>3NY3_1|Chain A|E3 ubiquitin-protein ligase UBR2|Homo sapiens (9606)
GPLGSLCGRVFKVGEPTYSCRDCAVDPTCVLCMECFLGSIHRDHRYRMTTSGGG
GFCDCGDTEAWKEGPYCQKHE
>3NY3_2|Chain B|N-degron|null RIFS

Preparation of the template structure:

Before proceeding with the actual homology modelling, it was necessary to remove the bound peptide, the N-degron. For this purpose, PyMOL software was used. The peptide was simply selected and then deleted using the 'remove' command.

This resulted in the following peptide sequence:
>3ny3_clean_A
LCGRVFKVGEPTYSCRDCAVDPTCVLCMECFLGSIHRDHRYRMTTSGGGGFCDC
GDTEAWKEGPYCQKHE

Subsequently, the number of residues loaded from the 'clean' structure by the peptide was determined. Each residue contributes with an entire charge:

- ASP (Aspartic Acid) and GLU (Glutamic Acid) have -1 each;
- LYS (Lysine) and ARG (Arginine) have +1 each;
- Each zinc ion has +2.

By using PyMOL, it was possible to obtain the number of positive and negative residues, obtaining:

- 86 negative residues;
- 82 positive residues;
- 3 zinc ions (as shown in figure 2.2)

The total residues charge is -4, but adding the charge of the 3 zinc ions (+6) gives a total system charge of +2. From this, we can deduce that the system is not fully balanced. The total charge should be close to zero to ensure electrostatic equilibrium, in fact, when preparing a structure, excess charge can cause numerical instability in calculations. Adding counterions is an important step to solve this issue. For this reason, GROMACS software suite was used.

At first, chain ID of 3 zinc ions was modifies from «A» to «Z» because the command 'pdb2gmx' cannot read different components in the same ID chain. The .pdb file of the structure was converted to the GROMACS format using the AMBER99SB-ILDN force field as it is capable of recognising zinc ions [58]. The solvent model chosen (water) was TIP3P. Then, a cubic box was created for the simulation and the system was subsequently solvated, specifying spc216.gro as the solvent configuration file. The ions.mdp file was used to prepare the .tpr input file for adding ions to the system. Verlet was used as the scheme for calculating short-range interactions. Coulomb interactions were calculated using the Particle Mesh Ewald (PME) method. Next, it was decided to add ions to the solvent.

Specifically, four chlorine ions (each with a charge of -1) were added to the solvent. The output file in GROMACS format was then converted to pdb, which is easier to display.

As a result of this process, the structure of 3NY3, previously cleansed of the peptide, was neutralised. For details of the commands used and the ions.mdp file, see Appendix A.

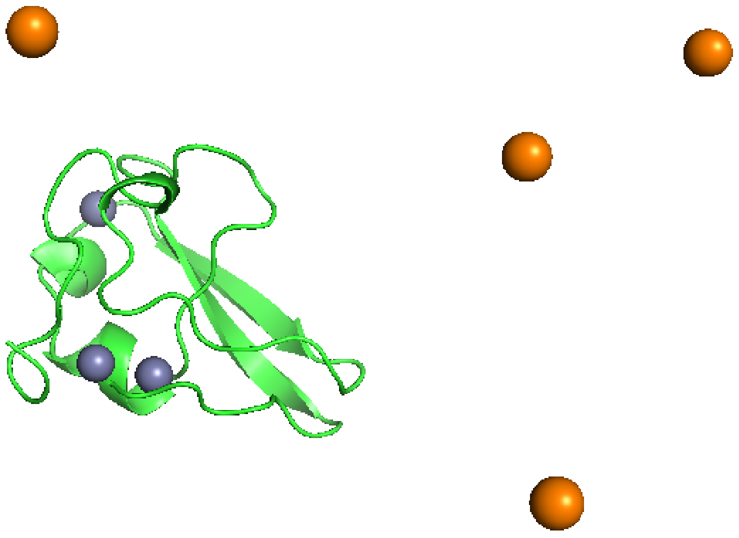


Figure 2.3: *Image of the structure of 3NY3, without the bound peptide and after neutralisation. The UBR-box is shown in green, the three zinc ions in grey and the four chlorine ions in orange. File 3ny3_now.pdb. Image constructed using PyMOL software.*

Sequence Alignment:

The sequence alignment step is crucial to ensure that 3NY3 and UBR1 have sufficient similarity. A good starting point is to achieve 40% sequence identity, but to create a human protein, at least 70% is expected [59].

Key data for this process were:

- ✓ the protein sequence of HUMAN E3 ubiquitin-protein ligase UBR1 in FASTA format (canonical) taken from Uniprot [60];
- ✓ the protein sequence of neutralised 3NY3, without the bound peptide, in FASTA format (canonical) found with PyMol.

For the sequences of these entities, see Appendix B.

EMBOSS Water online was used in this step. It determines the local alignment of two sequences using the Smith-Waterman method, which has been adapted for performance improvements [61]. This algorithm is based on the concept of dynamic programming, in which a problem is divided into simpler subproblems and their solutions are stored in memory. It was developed in 1981 and is a local alignment algorithm that aims to discover the best fragment between two sequences. This aspect is crucial in this project because 3NY3 represents only one domain of the E3 ligase under consideration.

The process consists of two macrophases. The first is filling the alignment matrix: the target sequence is placed horizontally (St), while the query sequence (Sq) is placed vertically. Each cell H(i,j) receives a score calculated based on whether there is a match of identical amino acids/nucleotides (positive score), mismatch (negative score), and penalties for insertions or deletions (gaps). As the matrix is populated, the value of each new cell depends only on its neighbors: the one diagonally across, the one above, and the one to the left. In the original approach the matrix is calculated using linear gap costs, while Gotoh added affined gap costs.

$$H_{i,j} = \begin{cases} H_{i-1,j-1} + M(S_t[i], S_q[j]) \\ E_{i,j} \\ F_{i,j} \\ 0 \end{cases}$$

$$E_{i,j} = max\begin{cases} E_{i,j-1} - \delta \\ H_{i,j-1} - \Delta - \delta \end{cases}$$

$$F_{i,j} = max\begin{cases} F_{i,j-1} - \delta \\ H_{i,j-1} - \Delta - \delta \end{cases}$$

The alignment score of St[i] and Sq[j] is shown by the term $H_{i-1,j-1} + M(S_t[i], S_q[j])$ where $M(S_t[i], S_q[j])$ is the scoring matrix. The effect of the preceding column and row on the present score is shown by Ei,j and Fi,j, respectively. *Δ* represents the gap open, while δ the extension penalty.

Define when the index i or j is less than 1, Hi,j, Ei,j, and Fi,j should all equal 0. We suppose that the optimal alignment score is stored in cell (i', j') when the alignment matrix computation is finished. The cell with the highest number tells where the best bit of similarity between the two sequences is located. Beginning at (i', j'), the second macrophase, the backtracking phase, would continue until it reached a cell where the value was zero. This path corresponds to the best local alignment. The process is time-consuming, especially in the initial phase. Therefore, parallel computing methods have been introduced to accelerate the operation [62].

Beside the use of this software, NCBI BLAST was initially tried, but 3NY3's sequence was too short to allow for meaningful alignment.

The results obtained with EMBOSS water were satisfactory; specifically, the sequence identity was 77.1%. For this reason, we proceeded with the construction of the homology model using the prepared 3NY3 structure as a template.

Model generation with Swiss-MODEL:

Homology modelling technique allows the construction of a protein ("target") 3D model starting from homologous proteins ("templates"), whose structure has been characterised experimentally. This tool is used to bridge the gap between known protein sequences and experimentally determined structures. Despite advances, the latter are inferior to high-throughput methods for screening protein-protein interactions.

Obtaining three-dimensional structures allows us to understand their function at the molecular level, and quaternary structures provide insight into the functioning of biological systems, how protein complexes and networks operate, and how they can be modulated. Furthermore, models can be generated by users without specific computational experience, while still being reliable and allowing for easy visualization and interpretation of results. This is enabled by fully automated workflows and servers.

In the case of this project, this method was used to predict the structure of the UBR1 protein, employing the prepared structure of 3NY3 as template. Swiss-MODEL was he pivotal software for this step and, specifically, the "User Template" mode was employed. The ProMod3 modelling engine, introduced in June 2016, serves as the foundation. It replaced the previous ProMod-II package, significantly increasing the accuracy of the generated models. It was designed to offer rapid and flexible prototyping for future modelling developments in SWISS-MODEL. It uses the OpenStructure as computational structural biology framework to connect the various modelling steps: from alignment to loop building, up to the final quality evaluation. It allows for the integrated management of sequences, structures, experimental data and simulations.

First, the program aligns the sequences of the two entities and, if this is successful, the structurally equivalent regions will be modelled with high reliability. Conserved regions are copied directly, while variable ones are reconstructed either using fragment databases or ab initio modelling methods, which predict properties based solely on atomic coordinates and the fundamental laws of quantum mechanics.

Amino acid side chains that differ from the template are repositioned using rotamer libraries. The model is then optimised and estimates of the quality of the 3D model are provided as output [63].

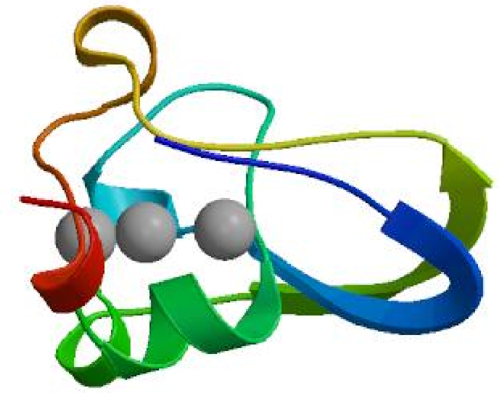


Figure 2.4: *Homology model of UBR1 E3 Ligase created using Swiss-MODEL software, based on the 3NY3 structure template.*

For a more detailed discussion of the sequence alignment results and the analysis of the generated protein structure, see subchapter 3.1.1.

2.1.2 AlphaFold Model

To evaluate the E3 Ligase UBR1 in its entirety, since its structure has not yet been determined experimentally, the three-dimensional model predicted by AlphaFold2 was considered. It was obtained in PDB format from the AlphaFold Protein Structure Database portal [64]. It has been developed by DeepMind in collaboration with the European Bioinformatics Institute (EMBL-EBI), a world-leading provider of biological data resources and tools. AlphaFold is an artificial intelligence algorithm that can predict a protein's three-dimensional structure from amino acid sequences alone. It stands out for its ability to predict structures with atomic accuracy, even in the absence of known homologous structures.

The success of this method has been demonstrated and validated in the 14th Critical Assessment of Protein Structure Prediction (CASP14), where it significantly outperformed other methods. From July 2021 to January 2023, it predicted 200 million amino acid sequences, with 2 million users in 190 countries. It means that nearly all "known" proteins with sequences in the UniProt database are included in this.

AlphaFold is based on neural networks and, like all AI methods, requires extensive training data to make accurate predictions. For this purpose, DeepMind trained it on publicly available data, such as those managed and supported by EMBL-EBI. However, this approach also leverages other sources of information, such as experimentally determined structures from the PDB, protein sequences and annotations from UniProt, and metagenomics data from MGnify. However, the researchers who worked on the project explain that public data were of fundamental importance and that the same is true for its future development.

To predict protein structure, it requires the amino acid sequence and aligned sequences of homologs as input. This neural model consists of two stages: Evoformer and Structure Module. The first is the heart of the neural network, which builds MSA representation and Pair representation from the protein sequence or multiple sequence alignments (MSAs). This provides information on the evolution and possible spatial relationships between pairs of residues. The Structure Module introduces the 3D coordinates. After all this, an iterative refinement (Recycling) takes place in which the output is reinserted as input. AlphaFold2 also provides an indication of the confidence in the prediction, useful for understanding where the structure is most reliable: pLDDT (predicted Local Distance Difference Test) and the TM-score. The structures predicted by AlphaFold2 had a median backbone accuracy of 0.96 Å r.m.s.d.95, a substantial improvement over the 2.8 Å r.m.s.d.95 of the second-best method [46].

The considered model refers to the Homo sapiens (Human) protein with UniProt code Q8IWV7. It is composed of 1749 amino acids: for its sequence see Appendix B. Since the value of the Average pLDDT parameter is sufficiently high, the model was considered suitable for further studies. Thanks to this approach it was possible to obtain the Predicted Aligned Error (PAE) graph and the AlphaMissense Pathogenicity Heatmap. For structure analysis, see subchapter 3.1.2.

The exact delimitation of the N-domain and the RING-H2 domain in the human UBR1 E3 ligase is not currently clearly defined in the literature. Studies on yeast Ubr1 suggest that the N-domain occupies residues 313–560, and it is plausible that the corresponding region in human UBR1 is located within the first 700 residues. The RING finger, known for its catalytic role, is located in the C-terminal portion. However, the position of the UBR-box (highlighted in orange in Figure 2.5), which occupies residues 97–168, is known.

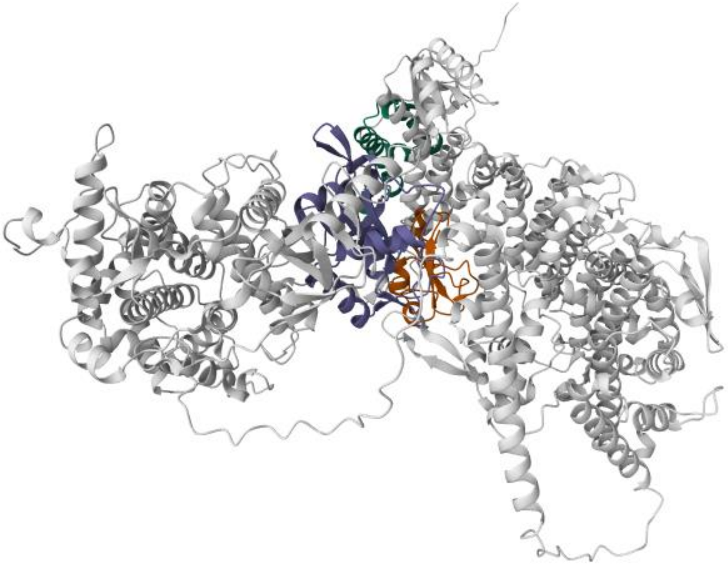


Figure 2.5: *Model of the human E3 ubiquitin-protein ligase UBR1 predicted by AlphaFold2 [64]. The UBR-box domain is highlighted in orange and is classified as Domain 2 in the AlphaFold database. The two domains in green and purple are classified as Domain 1 and Domain 3, respectively, but no matches were found for these in the structural domains cataloged in CATH.*

2.1.3 Comparison Between Models

At the end of the construction phase, two models were obtained:

- ✓ The homology model of the UBR-box domain of UBR1 E3 Ligase;
- ✓ The model of the entire UBR1 E3 Ligase protein predicted by AlphaFold2.

To assess the similarity between the models and perform a structural comparison, PyMOL, an open-source molecular modeling and graphics software, was used.

The two structures, in PDB format, were loaded as input into the software. They were then aligned using the 'align' command. This method was chosen because it provides the Root Mean Square Deviation (RMSD) value, which measures the average distance between the atoms of the molecules. It was decided to keep the default 'cycle' parameter, i.e. 5, to provide a more representative measure of conserved regions, discarding those related to loops, more flexible regions, or non-conserved regions. For an analysis of the results of this step, see subchapter 3.1.3.

2.2 Molecular Dynamic Simulations

To verify the conformational stability of the obtained models, explore the flexibility of the residues and finally obtain representative structures of the proteins to be used for virtual screening and docking studies, molecular dynamics simulations were conducted. It was chosen to run the simulations on both models in order to obtain a comparison with regards to the values of RMSD, secondary structure and in terms of energy. Additionally, the UBR1 E3 Ligase model predicted by AlphaFold2 was also analyzed to check whether the position of the binding sites found on the homology model in the next step was consistent. For further information on why MD simulations are a useful tool for research purposes, see subchapter 1.5.5.

To perform this step of the project, GROMACS (GROningen MAchine for Chemical Simulations) was used, specifically version 2024.4. This is a free and open-source software suite for high-performance molecular dynamics and output analysis. It was possible to use it on high-performance computing (HPC) resources, specifically on the Narval and Cedar clusters made available by the Digital Research Alliance of Canada. Cedar, as of September 2025, has been decommissioned and all files have been moved to the Fir cluster. Execution times were significantly reduced compared to running on local machines because the computational problem is divided into smaller parts that are performed simultaneously by each node.

For both models, the simulations were performed under near-physiological conditions. A constant temperature of 310 K was chosen to mimic the human body temperature. The ionic concentration was configured to 0.15 M and corresponds to the physiological ionic strength. The pH, however, was not directly set as it depends on the protonation state of the ionizable residues and is defined during system preparation. In this case, the residues were kept in their standard forms at physiological pH (approximately 7.0). The steps executed are analyzed in detail below.

2.2.1 Energy Minimization

The first phase was system preparation. The model's file was input in PDB format and converted to GROMACS format. A topology file was also generated, containing information on forcefield parameters, features (e.g., mass and charge), bonds between atoms, and angles. Hydrogens present in the PDB were ignored, and heavy hydrogens were used to increase simulation stability. The AMBER99SB-ILDN forcefield was selected because it is optimized for flexible protein regions and capable of recognizing zinc ions, and the TIP3P water model, an explicit three-site model compatible with the former, was chosen. The protein was then inserted and centered in a cubic box with a minimum distance of 1.0 nm from the walls to leave sufficient space for water and ions. This was filled with water molecules (spc216 model), and the topology file was updated. The topology file is then verified, expanded from a molecular definition to an atomic one, and the TPR file needed for ion insertion is created. This step also includes the em.mdp file, which provides the configuration parameters but, in this step, it does not yet use the ones related to minimization. The 'maxwarn' option was used to ignore non-critical warnings. The total charge of the protein was neutralized, and a salt concentration of 0.15 M was then reached by adding ions to the solvent group. Specifically, 33 water molecules were removed and replaced with 14 sodium ions and 19 chlorine ions.

Subsequently, the potential energy of the system was minimized to prepare the system for dynamical study. Locating the minimum points on the potential energy surface is important because they correspond to the conditions in which the biological system operates. To reach the minimum closest to the point on the surface, an energy minimization algorithm was used. Finding an energy minimum translates into smaller forces acting on the atoms. These would in fact lead to large accelerations, the breaking of the structure, and finding oneself in a very different region of the state space. The ideal would be to find a global minimum, but this requires considerable computational power. For this reason, a first-order derivative method was used: Steepest Descent, defined in the em.mdp file. It uses the Arbitrary Step Approach, which, after deciding the direction in which to move along the gradient, proceeds in steps. In terms of steps, it is longer than a line search, but a single one is much faster. The position x at step k+1 is given by:

$$x_{k+1} = x_k + \lambda_k s_k$$

where λk is the step length and sk is the direction. When a step produces an increase in energy, it is assumed to have skipped a minimum and therefore must return to the previous step. This algorithm works very well when the minimum is far away: when it is very close, it loses efficiency because it may not realize it has skipped one because it has already entered the next.

However, as Steepest Descent allows for rapid movement from the starting point, which is usually far from the minimum, and because it requires less computational power than higher-order algorithms, it was chosen for this stage of the project. The number of steps was set to 50,000 to achieve the right balance between resolution and processing speed. It should be noted that minimization algorithms do not explore any of the minimum and are insufficient to calculate the macroscopic properties of the system. Molecular dynamics methods, however, allow us to explore configurations along the entire potential energy curve and better describe the partition function. The minimizer had a step length of 0.01 nm and stopped when the maximum force reached a value of 1,000 kJ/mol/nm. To handle short-range interactions (cutoff within 1 nm), the Verlet cutoff scheme was used. The neighbour list, updated every 10 steps, was constructed using a Grid method (the space is divided into cells). Periodic boundary conditions (PBCs) were configured to simulate an "infinite system" by replicating the box. For long-range electrostatic interactions, the Particle-Mesh Ewald (PME) method was chosen, so the potential is calculated as the sum of the interactions across all periodic boxes and is very precise. Potential-shift-Verlet causes the potential to continuously go to zero right at the cutoff. Van der Waals interactions were calculated only within the specified cutoff (1 nm), beyond which the contribution was ignored. The Potential-shift-Verlet also ensured that the VdW potential smoothly canceled at the cutoff, avoiding numerical artifacts.

After setting these parameters, the potential energy minimization was performed and the output file with the minimized structure was em.gro.

This procedure was performed both for the homology model of the UBR-box of UBR1 E3 ligase and for the structure of the entire protein predicted by AlphaFold2. Furthermore, the same parameters were set to ensure comparable results. Table 2.1 shows the simulation results for both models.

Model	Homology Model	AlphaFold Model
Number of steps in which the Steepest Descent converges to Fmax < 1000	671	3467
Potential energy	$-2.34 \cdot 10^5$	$-9.37 \cdot 10^6$
Maximum force	$9.26 \cdot 10^2$ on atom 1015	$9.50 \cdot 10^2$ on atom 16451
Norm of Force	28.76	7.62

Table 2.1: *Output of energy minimization with the Steepest Descent method for homology model of the UBR-box of UBR1 E3 ligase and for the structure of the entire protein predicted by AlphaFold2.*

2.2.2 Equilibration Phases

It was proceeded with the NVT equilibration phase and then with the NPT phase for both models with the same configuration. During the first of these, the file obtained from the minimization and the nvt.mdp file were given as input. This phase served to stabilize the system temperature at 310 K by using the V-rescale thermostat (Bussi thermostat) with a fast coupling constant of 1 ps and keeping the volume fixed. The simulation duration, performed using a Leapfrog integrator for molecular dynamics, was 100 ps (50,000 steps with 0.002 as the dt). This algorithm, like all those used for simulations, uses numerical integration of the equations of motion since there is no analytical solution unless strong approximations are used.

The criteria it must meet are:

- Conservation of energy and momentum;
- Computational efficiency;
- Stability, allowing the use of sufficiently long integration timesteps;
- The assumption that positions, velocities, and accelerations can be approximated by a Taylor series expansion.

In particular, Leapfrog algorithm first calculates the velocities at time t + δt/2, which are then used to calculate the positions at time t + δt. The advantage is that the velocities are calculated directly, but not simultaneously with the positions for the same instant of time. The coordinates, velocities, energies, and log were written every 1 ps. Also in this case, for the non-bonding terms (cutoff 1 nm), the Verlet cutoff scheme was used with the Grid method to update the neighbor list every 10 steps. For the long-range electrostatic terms, PME was used. The velocities were generated from the Maxwell distribution.

The second equilibration step involved stabilizing the protein at constant pressure by adding the barostat and adjusting the system to the desired density. Water, in fact, can be too compressed or too rarefied. The volume of the box is then varied until the average pressure reaches the desired value (1 bar). Position restraints were also set on the protein here to avoid distortions during solvent adaptation.

The simulation lasted 100 ps, like the previous phase. The changes made compared to the previous configuration were to continue the simulation from the nvt.mdp file and activate the pressure coupling parameters. The Parrinello-Rahman barostat chosen was very accurate. In this case, the pressure value is not expressed directly but is set indirectly in the isothermal compressibility parameter, which measures how much the volume of a material changes when the pressure varies. The formula is as follows:

$$\kappa_T = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T$$

where V is the volume, P is the pressure and T is the temperature. For water at 300 K and 1 bar, the experimental isothermal compressibility is approximately 4.5·10-5 bar-1. With this value, the behavior of the solvent reflects real water well. No new velocities were generated because they resumed from those obtained from the previous phase.

2.2.3 Production Simulation and Calculation of Parameters for Data Analysis

At the end of these steps, the structures were ready for the actual production simulation, in which their temporal evolution in water and salts was observed. GROMACS was given the mdsimulation.mdp file, which contains the simulation configuration parameters, and the npt.gro file, automatically generated at the end of the constant-pressure equilibration phase. The simulation was carried over from the previous NPT.

The same integration parameters, constraints, thermostat, and barostat used in the NVT and NPT equilibration phases were maintained for both models: a leap-frog integration algorithm with a 2 fs integration step, a V-rescale thermostat at 310 K applied separately to the protein and solvent, an isotropic Parrinello-Rahman barostat at 1 bar, and electrostatic and van der Waals cutoffs of 1.0 nm with long-range PME processing. The changes made concern the positions no longer being constrained (the DPOSRES parameter was removed from the md_simulation.mdp file) and the much larger number of steps, which allows for a simulation on a time scale useful for the design and for analyzing the molecular dynamics of the system. A 100 ns simulation (50,000,000 steps) was performed for the homology model and a 10 ns simulation (5,000,000 steps) for the protein predicted by AlphaFold2. A different simulation time was adopted due to the size of the latter and the high computational cost otherwise required.

The Root Mean Square Deviation (RMSD) of the backbone was then calculated compared to the reference structure to assess the protein's overall stability throughout the simulation. To analyze the conformation of the φ and ψ torsion angles during the simulation, a Ramachandran diagram was generated.

The evolution of secondary structures was assessed using the DSSP tool integrated into GROMACS, which assigns the presence of α-helices, β-sheets, or loops in each frame of the trajectory. To reduce the amount of data and computation time, the DSSP was not calculated on each frame of the trajectory, but on snapshots taken at regular intervals of 100 ps. This resulted in 1000 frames for the 100 ns trajectory and 100 frames for the 10 ns trajectory. The trends of the various energy components (potential, kinetic and total) were then extracted from the energy file. For an analysis of the results obtained, see subchapter 3.2.

For more details on the GROMACS commands and files used in this phase of the project, see Appendix C. As an example, only the commands related to the simulations performed on the structure of the homology model of the UBR-box of UBR1 E3 Ligase are reported, but the same ones (with different file names) were also used for the model of the whole protein predicted by AlphaFold2.

2.2.4 Extraction of the Most Representative Structure

At this point, the most representative structure was extracted for the UBR-box homology model of UBR1 E3 Ligase. Indeed, during an MD simulation, the protein explores many conformations in space (microstates of state space), due to its intrinsic flexibility. However, for the subsequent phases of the project, such as searching for binding pockets, pharmacophore creation, virtual screening, and docking, it was decided to search for the structure that best represents the reality, under physiological conditions, of the average state of the protein.

A new group called 'Protein_ZN' was created, in addition to the existing ones, containing only the protein and zinc ions using the GROMACS 'make_ndx' command. The solvent was not included in the structure extrapolation because the software for identifying the binding sites is based on the geometry of the pockets and water molecules, by filling them, can disturb their identification.

Subsequently, the second part of the trajectory (50,000 ps), where the protein is expected to be most stabilized, was extracted. To do this, the 'trjconv' command was used, given as input the input files of the topology and initial coordinates, the complete simulation trajectory, and the created index file. The newly created group was then selected to obtain a new trajectory (md_50_100ns_Pr_ZN.xtc) containing only the protein and zinc ions, only from the time interval between 50 and 100 ns.

The 'cluster' command was used to create the clusters, inserting the filtered trajectory as input. The clustering algorithm was GROMOS, which groups frames based on an RMSD threshold, set to 0.55 Å. If two conformations have an RMSD lower than this threshold, they belong to the same cluster. The Protein_ZN group was used to consider only the protein and zinc ions. The cutoff was chosen by setting different values and observing the number of clusters created. Below are some examples.

Cutoff	Number of clusters	Comment
0.350	1	too low
0.060	1	too low
0.055	**4**	**chosen**
0.050	178	too high
0.040	23676	too high

Table 2.2: *Result of the number of clusters for different cut-off values in the clustering process of the protein configurations in the second half of the trajectory, based on the RMSD value.*

With the chosen cutoff, four clusters were obtained, meaning the protein assumed four main conformations between 50 and 100 ns. Appendix C details the commands used in this phase as well. Please note that the GROMACS commands used in this phase are based on the knowledge acquired during the 'Multiscale' course. The most populated cluster represents the dominant conformational state, from which the centroid, the most representative structure, was extracted. This was precisely the conformation used for the subsequent phases of the project.

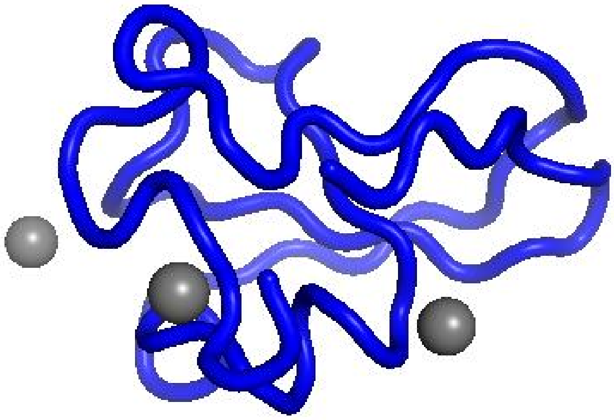


Figure 2.6: *Most representative structure of the UBR1 E3 ligase UBR-box model, extracted from the most populous cluster in the second half of the MD simulation trajectory. Image created with PyMOL. In blue it can be seen the structure of the protein and in grey the three zinc ions.*

2.3 Binding Site Prediction and Evaluation

In this phase of the project, potential binding pockets suitable for hosting the ligand were identified and analyzed on both models. The goal was to select the most suitable one on the homology model, comparing it with those of the model predicted by AlphaFold2, to conduct subsequent virtual screening and docking studies. To achieve this goal, an Electrostatic Potential Map (MEP) was first constructed for both structures. Subsequently, these were entered into a software that predicts their binding sites. The outputs were compared with each other and with the previous MEP. The execution method is explained in detail below.

2.3.1 Search for Binding Pockets

At this stage of the project, it was necessary to find possible binding sites between the protein and the future selected ligand. In fact, as explained in detail in subsection 2.3.3, the characteristics of the pocket influence the choice of ligand. For this purpose, FTSite was used, a computational algorithm that is part of the FTMap web server family. The study by Jones, Jindal and colleagues demonstrated that it is an effective tool that enhance the binding sites prediction even for high-precision protein models produced by deep learning techniques like AlphaFold2. It can be also used with experimentally generated structures. For these reasons, this method, like FTMap, is especially pertinent in the context of drug development as it is utilized for computational structural modeling to comprehend biological processes, structure, and function.

The algorithm requires the PDB file of the structure of a protein, DNA or RNA as input. In this case, the process was carried out both on the homology model of the UBR-box of UBR1 E3 Ligase and on the model of the entire protein predicted by AlphaFold2. The most representative structures obtained from the molecular dynamics analysis were loaded in the software, suitably prepared by removing water molecules and retaining only the protein and any relevant cofactors.

First, FTMap comes into action, sampling millions of positions of 16 small organic molecules (probes) that differ in size, shape and polarity. To find the most favourable positions for each type of probe on the protein surface, their poses are evaluated using a detailed energy expression similar to molecular mechanics. This is followed by a clustering phase based on average energy, in which the probes with the most favourable positions are grouped together. Here, it is essential to define consensus sites, i.e. regions that link different clusters of probes and define binding hotspots. The more probes there are, the more important the site is. This determines the site's capability to be targeted by small molecules with high affinity and therefore its druggability. FTSite is responsible for combining the nearby hotspots identified by FTMap and thus generating the binding pockets in their entirety. The regions indicated may be orthosteric or allosteric sites. Both algorithms can take small conformational changes into account [65].

Other software such as fpocket, P2Rank, SiteFinder, or FINDSITE are often used to find binding pockets on the surface of proteins. However, FTSite offers good accuracy without taking evolutionary factors into account, and it does not rely on surrogate indicators of ligand binding propensity such pocket volume, cavity depth, or the capacity to bind nonpolar spheres [66]. Moreover, the choice of software fell on FTSite because it uses the docking approach of small molecular probes with different chemical-physical properties, as explained above. This means that this method does not require input information on the ligand but is based on the analysis of the three-dimensional structure of the target. In fact, at this stage of the project, no possible ligand had yet been identified, as this is precisely what we wanted to achieve with this process.

The outputs were visualized using PyMOL. For both models, three possible binding pockets were found, representing the most likely positions for ligand binding. The electrostatic potential maps were compared with the regions where binding pockets were identified for both models. In addition, the positions of the binding sites obtained on the homology model of the UBR-box were compared with those predicted on the entire UBR1 E3 Ligase. The analysis, discussed in detail in subsection 3.3.2, showed that one of the three predicted pockets on the structure of the homology model was the one that housed the bound peptide in the 3NY3 structure (the template used to create it). For this reason, this site was chosen for the prosecution of the project. The selected pocket is shown in Figure 2.7.

In support of this choice, it should be noted that FTSite, in the model predicted by AlphaFold2, highlighted a binding pocket in the UBR-box domain in the same region. These comparisons were made possible using PyMOL's 'align' command. The initial 3NY3 structure was aligned, first with that of the homology model and then with that predicted by AlphaFold, both of which were output by FTSite.

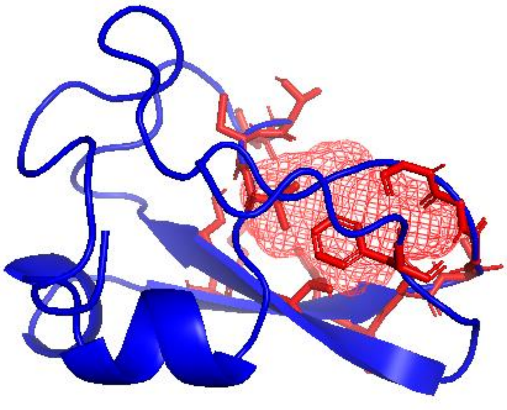


Figure 2.7: *3D visualization of the selected binding pocket, predicted by FTSite, of the homology model of UBR1 E3 Ligase's UBR-box domain (site_select_1). The protein structure is shown in blue (ribbon). The residues that delimit the pocket are highlighted in red (stick) and are those most likely to be involved in interactions with potential ligands. The red mesh represents the volume of the cavity calculated by the software. Image created with PyMOL.*

The residues belonging to the specific binding site chosen were found with PyMOL, using the command:

```
iterate site_select_1, print (resn, resi, chain)
```

where the selected pocket is defined by the name site_select_1. 126 atoms were iterated and the residues identified are as follows:

PHE 103, THR 109, SER 111, PRO 119, THR 120, CYS 121, VAL 122, LYS 139,
HIS 141, THR 142, SER 143, THR 144, GLY 145, GLY 146, GLY 147

For the sake of completeness, the residues belonging to the other two pockets are reported in subsection 3.3.1, identified using the same approach as above (obviously changing the name of the selected object). For site_select_2 93 atoms were iterated, while for site_selec_3 70. See this section also to observe where these regions are located on the protein structure.

2.3.2 Electrostatic Potential Map

The MEPs were constructed using PyMOL's APBS (Adaptive Poisson–Boltzmann Solver) Electrostatics plugin. Instead of explicitly representing each water molecule, it treats the solvent as a continuous medium with dielectric properties (continuous solvation). To do this, it must be given as input a complete continuous structure with all the atoms present and the force field parameters, i.e., partial charges and atomic radii.

The limitation, however, is that not all proteins from the PDB satisfy these conditions and do not contain this information, which is essential for constructing electrostatic fields. To overcome this problem and to act as a bridge between the PDB structure file and APBS, PDB2PQR was developed: a software that automates structure preparation. It allows you to generate an APBS-compatible PQR (Position, Quote, and Radius) file, similar to the PDB but containing all the missing information. It allows to complete side chains or residue fragments by adding atoms (not hydrogens), determine protonation states by estimating the pH and assigning hydrogens so that hydrogen bonds are physically correct and assign atomic charges and radii from various force fields [67], [68].

In this specific case, the function that allows PROPKA to be used to assign a protonation state at pH 7 was enabled. It is an empirical method integrated into PDB2PQR that calculates the pKa values of amino acids very quickly and accurately. It takes into account the effects of desolvation, hydrogen bonds and electrostatic interactions. The Amber forcefield was used to assign charges and atomic radii, and an internal output naming scheme was set. Furthermore, several options were selected:

- ensuring that new atoms are not reconstructed too close to existing ones;
- optimizing the hydrogen bonding network;
- create an APBS input file (to use it in the next step);
- add/keep chain IDs in the PQR file;
- remove the waters from the output file.

The same parameters and options were set for both structures. The PQR format files of the models were thus generated from the PDB2PQR software.

At the end of this process, the resulting files were uploaded to PyMOL. Opening the APBS Electrostatics plugin revealed a message clarifying that no further model preparation was necessary as the selection already contained charges and radii. This means that the work done with the PDB2PQR software had achieved the desired results. Explicit solute and solvent models require sampling and equilibration, while implicit ones do not. Nevertheless, the latter manage to ensure robust and qualitative efficiency and accuracy.

The chosen software uses an implicit solvation model based on the Poisson–Boltzmann (PB) equation. It solves the partial differential equation and yields a global solution for the electrostatic potential, inside and around a biomolecule:

$$-\nabla * \ \epsilon\nabla\phi - \sum_{i}^{M} c_i q_i e^{-\beta(q_i\phi+V_i)} = \rho$$

The parameter describing the potential is ϕ. The number of mobile ionic species in the solvent, i, ranges from 1 to M, qi are their charges and ci are their concentrations. The biomolecular structure is described by three entities. Vi is the steric ion-solute interaction potential. ϵ is the function of the dielectric constant that changes depending on what we consider: inside the protein it takes on a lower value (defined by em), while in the solvent it is higher (defined by es). At the surface of the protein, the dielectric function changes drastically, creating the boundary conditions of the PB equation. This boundary depends on the position of the protein atoms and the atomic radius. ρ depends on the partial charges assigned by the force field and represents the distribution of atomic charges. On the molecule, they are seen as discrete points placed in the coordinates of the centre of the atoms, mathematically defined as the sum of Dirac delta. Beta is equal to 1/kT (the inverse of thermal energy) where k is Boltzmann's constant and T is the temperature.

By solving this equation, it is possible to describe how the electrostatic potential around the protein immersed in water varies at the set ionic concentration. It should be noted that PB theory is an approximation, but despite this, it is used in various fields that require the determination of global electrostatic properties. Applications include visualisation, structural analysis and diffusion simulations [69].

In the specific case of the project, the linearised PE equation (LPBE), valid when $\beta q_i \phi$ is much less than 1, was solved using the grid-based finite difference PMG solver. The parameters set were as follows: The grid and resolution were set automatically according to the system dimensions. The parameters set were as follows:

- internal dielectric constant of 2, as it accurately represents the hydrophobic environment and low polarizability of the protein core;
- external dielectric constant of 78, as that of water at room temperature, reflecting water's ability to shield electrostatic interactions;
- temperature of 310 K to make the model consistent with biological conditions;
- concentration 0.15 M (ionic species Na^+ and Cl^-), as physiological salinity with typical radii.

The electrostatic potential map is read using a three-component colour coding applied to the molecular surface: red, white and blue. It was produced for both models so that they could be compared with each other and in order to analyse more comprehensively the regions where potential binding pockets were then identified. PyMOL's "set transparency" command was also used. This made it possible to view the protein structure simultaneously with the map, allowing for a detailed understanding of which parts of the protein were negative, neutral, and positive. For results interpretation, see subchapter 3.3.2.

2.3.3 Structure-Based Pharmacophore Model

After constructing the target model and finding a possible binding pocket on its surface, it was time to search for potential small molecules capable of binding correctly to the chosen site. The selection of possible ligands is in fact the primary objective of the project, as they will constitute the anchor of the PROTAC.

The method used to pursue this objective is virtual screening. However, a preliminary phase was carried out in which a pharmacophore model was constructed based on the binding pocket identified by FTSite. In fact, virtual screening relies solely on computing power, and each molecule is simply "thrown" at the target to see which one binds best. The construction of the pharmacophore in the preliminary phase offers numerous advantages, such as improved accuracy: the selected compounds are more likely to fit well into the binding pocket. It identifies the essential physical and chemical characteristics required for the binding of a molecule to a given receptor and so, in this case, for ligand selection.

In the definition of a pharmacophore, the International Union of Pure and Applied Chemistry (IUPAC) refers to steric and electronic characteristics, i.e. those relating to the size and shape of the molecule (its geometry), the distribution of charges and the ability to form bonds (bond sectors location). These are required to activate (or prevent) a biological target's biological response and to guarantee appropriate supramolecular interactions with that target. The characteristics are translated into geometric entities that constitute the pharmacophore query:

- hydrogen bond acceptors (HBAs) are points in space where the candidate molecule must have a group capable of accepting a hydrogen bond;
- hydrogen bond donors (HBDs) are those where there must be an atom capable of donating a hydrogen;
- ionisable groups, positive (PI) or negative (NI), are positions where it is important for a charge to be present for electrostatic interactions;
- areas where non-polar or aromatic groups (AR) are needed, useful for stacking or hydrophobic (H) interactions;
- exclusion volumes (XVOL), i.e. regions of space that must remain free without any bound atoms as they represent steric hindrances of the protein.

During the virtual screening phase, the software uses this set of spatial rules to create a filter: only compounds that correctly position these features in space (respecting distances and geometries) pass the selection and are considered potential ligands.

This approach reduces the number of compounds on which these subsequent analyses must be performed, helping to reduce the computational cost of the process. In fact, the compound libraries used for drug discovery are often very large. Consequently, although

CADD methods such as virtual screening are already employed to shorten development times, the use of pharmacophore models as queries enables the process to be accelerated further.

Another advantage is the possibility of finding different chemotypes as virtual screening hits, i.e. structurally different molecules, because the drug-like model design process is not based on actual atoms or specific chemical groups. It can be used to find chemically divergent molecules that trigger similar biological events on the same target. Thus, this approach makes it possible to drastically reduce the number of molecules to be subjected to subsequent stages of analysis, such as molecular docking, while maintaining structural diversity.

There are two main approaches that differ based on the input entered into the software. Ligand-Based Pharmacophore Modelling is used when the structure of the target is unknown but certain active compounds that bind to the same protein target are available. In this situation, the common characteristics of interest are extracted and the pharmacophore is created. This method can in turn lead to two different approaches: quantitative structure-activity relationship (QSAR) or quantitative structure-property relationship (QSPR) modeling.

In the case of this project, however, the Structure-Based Pharmacophore Modelling method was used because a 3D model of the target protein was constructed, which can be used to identify ligands. If a possible ligand had been available, the pharmacophore model would have been more accurate and of higher quality. Information could have been obtained regarding its bioactive 3D conformation, highlighting the functional groups directly involved in interactions with the target. Furthermore, information on spatial restrictions due to the shape of the binding site (exclusion volumes) could have been added, which in this case was not possible [59] [70].

BioLuminate 5.8, a software developed by Schrödinger, was used to design the structure-based pharmacophore model. As it was not possible to purchase their software package, a one-month temporary licence was requested and granted. Initially, LigandScout 4.5 was tried, but even though a temporary licence was obtained, the features needed to pursue the project's goals were blocked. Thus, the structure of the homology model of the UBR-box of UBR1 E3 Ligase was entered as input in BioLuminate 5.8, more precisely the centroid of the largest cluster extracted from the trajectory of the molecular dynamic simulation. In addition, the residues of the chosen bonding pocket were defined manually. To see the list of defined residues, belonging to the site_select_1 pocket, read subchapter 2.3.2.

All features identified by the software have been selected:

- ✓ acceptor of hydrogen bond sites (A1 and A2);
- ✓ donor of hydrogen bond sites (D3, D4 and D5).

The result can be obtained in various formats such as .mae, .mol2, .sdf or .smi (and also in their zipped form) so that the pharmacophore model can be used in different software for subsequent analysis. This model will from now on be identified by the name Pharmacophore_1.

Using the same approach, another pharmacophore model was created based on the structure related to the pocket defined as site_select_2. Due to its proximity to the region where the peptide is bound in the 3NY3 template structure, even though it is not the same, it was considered useful for possible future projects and therefore taken into account at this stage. During the design phase, the corresponding residues have been then inserted and the following properties identified by the software have been selected:

- ✓ donor of hydrogen bond sites (D3, D4 and D5)
- ✓ hydrophobic region (H3 and H4)
- ✓ ring feature/aromatic interaction (R5).

This model will from now on be identified by the name Pharmacophore_2.

The Pharmacophore_1 model was selected for the advancement of the project and therefore to be used as a hypothesis in virtual ligand screening. For the analysis of pharmacophore models, see subsection 3.3.3.

To validate the pharmacophore model, software exists to calculate sensitivity, specificity, enrichment factor and GH score using sets of experimentally known active molecules and inactive or decoy molecules. ROC curves can also be constructed and AUC (Area Under Curve) calculated, which allows us to understand how well the model can distinguish between active and inactive molecules. Fisher's randomization test determines whether there is a substantial or random relationship between pharmacophore properties and biological activity. To perform these analyses, powerful tools such as LigandScout, Discovery Studio or MOE must be used. However, these are commercial products and it was not possible to obtain a licence allowing these specific functions to be utilised. For this reason, the chosen model was validated by applying it directly in the subsequent virtual screening phase and analysing the results obtained.

2.4 Virtual Screening

The virtual screening phase was the core of the project, as it allowed the identification of potential ligands most likely to bind to the target, the UBR-box domain of the UBR1 E3 Ligase protein. The first step was to select the library for ligand search. They were then prepared using dedicated software and the screening was performed by entering the designed pharmacophore model as a filter. Finally, an assessment was made of the properties of the compounds that obtained the highest scores. The different sub-phases are explained in detail below.

2.4.1 Ligand Preparation

The library considered was the National Cancer Institute (NCI) Diversity Set. The NCI is one of the agencies of the National Institutes of Health (NIH) which is involved in research and development for the prevention, diagnosis and treatment of cancer. The Developmental Therapeutics Programme (DTP) is an internal programme within the NCI that is part of the Division of Cancer Treatment and Diagnosis (DCTD). It provides databases and resources for academic and industrial researchers and offers large-scale biological testing (in vitro and in vivo screening). It is committed to collecting, maintaining and distributing libraries of chemical compounds such as the NCI compound repository. The latter is a repository of pure natural products and synthetic substances synthesised or donated by academic laboratories, pharmaceutical companies and research groups. It serves as a "chemical warehouse" for screening new molecules with potential anti-tumour or therapeutic activity and researchers can use it for non-clinical research. Over 200,000 compounds have been submitted to the DTP for biological study or, in certain situations, synthesized under the DTP's auspices, making it an incredibly varied library. With the exception of shipping fees, samples sourced from the NCI/DTP Open Chemical Repository are freely accessible, but only in the United States (at the time of writing) [71]. In addition, the DPT provides sets with high chemical diversity, such as the NCI Diversity Sets.

Although official collections such as Diversity Set VII contain 1,581 compounds, some versions of NCI Diversity Set II are reported in the literature as containing 1,974 compounds [72]. The latter version was used in this project [73]. The Diversity Set selection is composed by extracting representative subsets from the Chemical Repository collection.

The construction process begins with an initial filtering stage in which molecules with undesirable properties, such as those that are chemically unstable, too large or too reactive, are eliminated. Next, the chemical and structural diversity is analysed, identifying compounds that represent the greatest number of different chemical scaffolds. This ensures that molecules that are too similar to each other are not included. The set created with this approach, while containing far fewer molecules than the previous one, maintains a high degree of variability among the structures.

Given its small size, virtual screening has a low computational cost. Computational limitations and time constraints, as well as its balance between the number of molecules and coverage of the chemical space, made it perfect for the purposes of this study.

Another reason for the choice of this particular set of compounds was that it was thought that, once the final ligands had been identified, they would be freely available for testing in the laboratory. Unfortunately, as already mentioned, the NCI does not currently accept requests for compounds or plate sets from countries other than the United States and US Territories. However, they have been sent by other suppliers.

The preparation phase for the Diversity Set ligands, prior to the actual virtual screening, was performed using Maestro 14.3 software, developed by Schrödinger. As with BioLuminate 5.8 software, it was possible to obtain a one-month provisional licence with the necessary functions. Specifically, the Ligprep tool was employed. It generates accurate and realistic 3D structures of small molecules from 1D files such as SMILES strings or 2D planar structures in formats such as SDF and MOL. Corrections can be made to reflect real-world "chemical space" to prepare molecules for subsequent steps such as docking, molecular dynamics simulations, or virtual screening, as in this case. It generates tautomers due to the movement of protons in a solution, different protonation/ionization states, calculated as a function of the set pH and considers different stereoisomers. In fact, the same molecule can exist in multiple chemical forms in solution. Ignoring them could result in missing potential interactions with the target. It also takes into account possible ring conformations. LigPrep also performs a filter, discarding molecules that do not meet certain chemical or pharmacological criteria, such as Lipinski's rules. This was also done during the construction of the Diversity Set from the Repository collection. Furthermore, to ensure realistic geometries compatible with docking and molecular dynamics software, energy minimization is performed on the generated 3D molecular structures.

Despite its multiple functionalities, it has a high computational efficiency: it is able to process approximately one ligand/second on a single CPU, with the possibility of parallelization on HPC clusters. At the end of the process, the software outputs a file in Maestro or SDF format (depending on the choice) with the ligands ready to be used for the subsequent phases [74].

Initially, it was necessary to remove the structure identified by code NSC 84460 from the set because it contained an arsenic atom that was difficult for the software to process. A total of 1,973 compounds were therefore prepared. The new set in SDF format was loaded into the software and the relevant parameters were configured:

- ✓ Maximum ligand size of 500 atoms to avoid molecules that are too large or unrealistic;
- ✓ OPLS4 force field for the energy minimization of ligands, Schrodinger's latest version, accurate for pharmaceutical chemistry;
- ✓ Ionization generating possible states at pH 7.40 ± 2.00;
- ✓ Epik method for ionization to have high accuracy;
- ✓ Include original state in order to also preserve the input structure;
- ✓ Enabled ‘Desalt’ to remove any inorganic ions or small unwanted fragments;
- ✓ Enabled ‘Generate tautomers’ to generate all relevant tautomeric forms;

- ✓ Selected ‘Retain specified chiralities (vary other chiral centers)’ to maintain the chiral centres already defined in the input, but allowing the software to vary the unspecified chiral centres;
- ✓ Set a maximum of 32 stereoisomers per ligand;
- ✓ Chosen to obtain the output in Maestro format so that the file can be easily used for virtual screening using the same software.

This process generated several files, including one containing the ligands ready for the next screening phase in MAEGZ format, a compressed molecular model file created by Schrödinger specifically for its Maestro molecular modelling software.

2.4.2 Ligand Screening

Virtual screening methods are fundamental for identifying new bioactive substances because they increase the speed of the process through computational simulations. As already mentioned, these approaches select molecules that are most likely to bind correctly to the target from large virtual libraries of compounds. They are able to discard structures that could be toxic or have unfavourable properties such as potency, affinity and selectivity (pharmacodynamics) and absorption, metabolism and bioavailability (pharmacokinetics).

It is common practice, when performing virtual screening on a very large library of compounds such as PDB, PubChem, ChEMBL, Zinc or Drugbank, to apply a filter, even before starting the process, which analyses and discards molecules that do not have the desired ADMET profile and do not comply with Lipinski's rule of five. This approach reduces the computational weight of the simulation. Due to the origin of the set of ligands considered in this study, it was not necessary to apply this filter before starting the virtual screening. The computational resources were considered adequate to perform this type of simulation. The evaluation of the above properties was conducted downstream of the process and was used as a criterion for selecting ligands, both for the expansion phase of the docking set and for the final selection for experimental testing (see subsections 2.5.1 and 2.5.3, respectively).

This is in line with some protocols reported in the literature and it should be underlined that the choice of the type of approach to be used is in the hands of the researcher. The pharmacophore model plays a key role in the virtual screening process. It provides the hypothesis that the software uses to filter the collection of ligands and identify “hits”, i.e. molecules that match the characteristics required to be potentially active against the target. Since the pharmacophore represents chemical functionalities and spatial relationships rather than actual chemical groups, the results may include “hits” with different structures.

The Phase tool of the Maestro 14.3 software was used to implement this phase of the project. It is dedicated to pharmacophore modelling, allowing for its generation, validation and even screening. It is able to construct feature maps by evaluating which ones are common in a set of ligands (ligand-based) or deriving them from a given site (structure-based) and expressing them as geometric entities. It can use known receptor-ligand complexes to derive a hypothesis that respects both geometry and observed interactions or create hybrid models.

In this project, it was necessary to use only the Phase Ligand Screening functionality, as the pharmacophore model had been created previously with BioLuminate 5.8 software. This tool enabled the ligands in the NCI Diversity Set to be compared with the structure-based pharmacophore model and sorted according to their fitness score. It is reported that it is also possible to perform screening on complete databases of compounds available for purchase from Enamine, MilliporeSigma, MolPort, and Mcule.

The parameters were configured as follows:

- ✓ Number of conformers per ligand set to 50 to achieve a good compromise between not losing relevant conformations in order to sample the conformational space sufficiently well and computational cost;
- ✓ Phase Screen Score as scoring function which evaluates how well a ligand fits all features, penalties for mismatches or clashes, vector orientation;
- ✓ Default match rejection criteria;
- ✓ Intersite distance matching tolerance of 2.0 Å, i.e. the tolerance with which each pharmacophore feature can deviate from the ideal position, set to allow flexibility in alignment.
- ✓ Limit CPU time for matching to 0.1 seconds/molecule;
- ✓ Sort hit by decreasing Phase Screen Score
- ✓ Structure output in Maestro format.

The Phase tool supports two main search modes. In standard mode, the pharmacophore model is compared with a set of pre-calculated conformers for each molecule in the database or file. In ‘On-the-fly’ search, conformers are generated in memory as needed, using the fast search method. In this case, the first modality was used. To find the match between a ligand and the pharmacophore hypothesis, the software imposes the single tolerance defined by the user on each inter-site distance (distance between pairs of pharmacophore sites) in the hypothesis. It is also possible to decide whether all sites of the pharmacophore must be respected or whether it is sufficient to respect only a subset of m sites out of a total of k. In this case, however, the distances between those sites must still be respected and there must be m(m-1)/2 geometric relationships to be respected. This allows potentially valid molecules not to be excluded. The least-squares fitting mathematical method is applied whenever a ligand meets the criteria for aligning its points with the pharmacophore model. In this regard, a positional tolerance can also be used which, if not respected, can cause the match to be discarded. If the match is kept, it is evaluated using a fitness score and the ligand conformers are ordered from best to worst. The overall hit list has a maximum size, and when it is full, the matches with the lowest scores are eliminated [75].

From this process, 6082 structures were obtained, grouped in a file in MAEGZ format. They were sorted from the structure with the best score to the worst. The number of configurations is high due to the number of conformers set during the configuration phase. For the analysis of the results and to know the score obtained for the structures, refer to subchapter 3.4.1. For the prosecution of the project, it was decided to take into consideration the first 100 configurations found with the highest scores.

2.4.3 Preliminary Evaluation of Compound Properties

Since the generation of 50 conformers for each ligand was set during the virtual screening, the top 100 configurations that obtained the highest Phase Screen Scores in the virtual screening included multiple conformers belonging to the same compound. To reduce redundancy and evaluate only the most promising unique compounds, for each ligand represented by multiple conformers, only the one with the highest Phase Screen Score was considered.

In this way, the selection was reduced to 23 unique compounds, on which a preliminary evaluation of pharmacokinetic properties, and drug-likeness was conducted using Chemaxon Playground Calculator [76]. It is an easy-to-use web application that allows you to directly upload the compound you want to evaluate. Alternatively, it is possible to draw them with MarvinJS canvas or there are some examples of molecules to start with without drawing or uploading. It is possible to select and set the order of the available calculations and forecasts.

The characteristics that were displayed are the following:

- Protonation, i.e. pKa, microspecies, predominant species at a given pH;
- Element Analysis, i.e., molecular weight and exact molecular weight, nominal mass, composition, mass spectrum, and formula;
- Naming, both IUPAC and traditional names;
- Point values of physicochemical properties, i.e. intrinsic solubility (logS), logP, logD at pH 7.4, most basic pH and most acidic pH;
- The pH dependent lipophilicity, i.e. the graph representing the logD as a function of the pH variation with the possibility of setting a certain pH value and obtaining the specific logD value;
- The pH dependent solubility, i.e. the graph showing the logS as a function of the pH variation with the possibility of setting a certain pH value and obtaining the specific logS value;

Furthermore, thanks to the ADMET Plugin Group that leverages the power of machine learning methods on selected datasets it was possible to obtain two different models regarding the prediction of Human Ether-à-go-go-Related Gene (hERG) inhibition to eliminate the risk of cardiotoxicity:

- Continuous Activity Model from which the pActivity value, which depends on IC_{50}, is obtained;
- Classification Model in which the compound is classified as toxic or safe.

To estimate the prediction reliability of both types of models, in addition to the point value or classification result, the plugin also outputs a graph or colored bar representing the Applicability Domain. An indicator (green/yellow/red) or a numerical score reflecting the molecule's position with respect to the training domain is displayed. This helps understand how similar the molecule being analyzed is to the molecules on which the model was trained. Another useful feature allows the visualization of the most similar molecules in chemical structure among those present in the hERG model's training dataset. For these, the plugin displays the 2D chemical structure, the known experimental pActivity value or classification (depending on the model type) and the similarity value.

After analyzing the obtained properties for each of the 23 compounds, 10 were selected for the next steps. Table 2.1 lists the most significant selection criteria.

Property	Criteria
Microspecies at pH 7.4	higher percentage of neural form
Molecular Weight	less than 650 Da, relaxed Lipinski's rule of 5
Intrinsic Solubility Log(S) and Log(S) at pH 7,4	values greater than -4
Log(D) at pH 7.4	as close as possible to 1-3 values
Log(P)	values similar to log(D)
hERG (Activity Model)	pIC_{50} values less than 5 μM (IC_{50} greater than 10 μM)
hERG (Class Probability)	safety percentage greater than 70%
Prediction Class	classified as safe

Table 2.1: *List of the main criteria used in selecting the compounds to be used to compose the docking set, starting from the conformers with the highest Phase Screen Score in the first 100 hits of the virtual screening. All properties were obtained using Chemaxon Playground Calculator web application.*

It should be noted that, as a result of employing the aforementioned approach, all compounds among the top 100 hits of the virtual screening, including the various conformers, have a Phase Screen Score greater than 1.358. In addition, the number of conformers obtained in the top 100 hits considered was also counted. These parameters were taken into account in the selection process, but not in a stringent manner.

Here are some details about the criteria that were applied. The pKa is the negative logarithm of the acid dissociation constant (Ka) of an acid (or ionisable site) and indicates the tendency of the site to lose a proton. In this case, microscopic pKa was analysed, which differs from macroscopic pKa, which defines the overall equilibrium between states with different numbers of protons. In fact, the former distinguishes between specific microstates, i.e. states that differ in terms of the proton on a specific site, while keeping the other ionisable sites fixed. In systems with multiple ionisable sites, there are many possible microstates with different microscopic pKa values. If a molecule has n ionisable sites, in theory it can have up to 2n micro-species, varying which site is protonated/deprotonated.

As the pH varies, the distribution between the micro-species changes, as can be seen from the titration curves. The web application used in this step estimates how many fractions of each micro-species are present at a given pH, according to the calculated pKa, and it is possible to see which of these is predominant. In the context of predicting the binding affinity between the target and a possible ligand, this aspect is relevant because the charge and protonation of a molecule strongly alter its geometry, charge distribution, electrostatic interactions, and ability to form bonds. There are number of studies emphasise this aspect [77]. For this reason, it is useful to generate different micro-species during preparation for docking. In this context, it was decided to evaluate this property at this specific stage of the project in order to identify compounds that had a predominant neutral micro-species at physiological pH. The binding pocket of the UBR-box of UBR1 appears to be neutral overall, with no charged residues in key regions.

Lipinski's rule of 5 suggests a molecular weight of less than 500 Da for drug-like molecules, but in certain contexts this criterion can be relaxed [78]. Lipinski's rules are in fact based on a statistical trend observed among oral drugs and PROTAC technology could also be injected intravenously or directly intracellularly. In order to increase chemical diversity and not exclude potentially active compounds that could be optimised later, it was therefore decided to set a molecular weight limit of 650 Da.

Intrinsic solubility, often defined as the logarithm to base 10 of molar solubility, measures the intrinsic capability of a substance to dissolve in water for structural reasons (hydrophobicity, ability to form H bonds) and is independent of pH. It is the solubility of the neutral form and therefore reflects the chemical “nature” of the compound, not its state of ionisation. Solubility at pH 7.4, on the other hand, indicates total solubility considering the contribution of ionised forms at physiological pH. Compounds with a Log(S) value greater than -4 were chosen because lower values cause dissolution problems: a threshold used in lead-likeness filters and ADMET screening to avoid candidates that are too hydrophobic or crystalline. In addition, most approved drugs exhibit Log(S) value between -1 and -5 [79]. This criterion maintains a good probability of bioavailability without excessively penalising more hydrophobic molecules.

Log(D) measures the lipophilicity of an ionisable molecule at a given pH and is calculated as the base-10 logarithm of the concentration ratio of a molecule between an organic phase (n-octanol) and an aqueous phase (buffer at a defined pH). It takes into account the ionised micro-species present at that pH. The chosen range of values (between 1 and 3) is based on statistical analyses of thousands of approved drugs and has been considered suitable for achieving the right balance between water solubility and lipophilicity, i.e. similar to lipid membranes and hydrophobic sites but not too much to precipitate or bind excessively to lipids/plasma. It is therefore compatible with the “drug-like space” of most protein ligands. Furthermore, working within this narrow range reduces the likelihood of many ADMET parameters presenting problems. Log(P) represents the octanol/water ratio but, unlike the previous parameter, refers to the neutral form of the molecule. Most literature uses both of these parameters to obtain an accurate prediction of a compound's drug-likeness, but in some studies Log(D) is considered preferable [80].

In the case of this project, since compounds characterised by a prevalence of the neutral form at physiological pH (7.4) were selected, the two parameters can be considered almost equivalent, and the contribution of the ionised forms is neglectable.

A specific example of lipophilicity-related toxicology concerns the hERG channel (KCNH2), a gene that codes for a voltage-dependent potassium channel (Kv11.1) present in myocardial cells. It is essential for the proper functioning of the heart and specifically for the repolarisation of the cardiac action potential. It is one of the main causes of drug candidate abandonment, even in advanced clinical phases, as off-target molecules can bind non-specifically to this channel. This process inhibits K^+ flow, prolonging the QT interval, which can cause serious arrhythmias or torsades de pointes, which can be fatal. For this reason, all modern ADMET screening software also applies a filter on hERG.

A drug's hERG activity (Act) indicates its ability to block hERG channels. In experimental contexts, this parameter is determined using various electrophysiological methods and measured as the Inhibitory Concentration (IC50 or Ki) values but, in the literature, it is used as the negative of its logarithm to base 10 (pActivity). IC50 is the concentration of an enzyme inhibitor required to inhibit 50% of the target under examination in vitro. Chemaxon hERG Predictor is a statistical QSAR model based on machine learning which uses Random Forest and Conformal Prediction algorithms. In the hERG Activity Model, which performs a quantitative estimate of inhibitory potency, IC_{50} values greater than 10 μM, i.e. pIC_{50} values less than 5, are generally considered safe, ensuring a low probability of hERG inhibition. This threshold value is used in several drug development studies and represents the standard safety benchmark. Regarding the prediction class in the classification model, the distinction between “safe” and “toxic” is also based on this IC50 threshold value [81], [82].

Only compounds classified as safe were selected, i.e. compounds predicted not to block the Kv11.1 channel. More specifically, compounds with a safety probability greater than 70% were selected so that the measurement would be closer to the probability that a predicted safe compound is truly safe.

For a detailed description of the characteristics of each of the 23 compounds, see subchapter 3.4.2. It should be noted that the 10 compounds selected at this stage, listed in Table 2.2 and marked with a green tick, do not simultaneously meet all the requirements defined above. The choice was made by favouring candidates that presented an optimal compromise between the pharmacokinetic properties analysed, solubility, lipophilicity and safety profile (hERG), in line with a rational approach typical of the preliminary stages of drug discovery.

Compound identifier	Selection
NSC 143101	✓
NSC 18695	✗
NSC 25485	✗
NSC 100858	✗
NSC 111702	✓
NSC 12161	✓
NSC 527017	✗
NSC 67608	✗
NSC 72234	✗
NSC 319758	✗
NSC 133118	✓
NSC 143099	✓

Compound identifier	Selection
NSC 188491	✓
NSC 86005	✓
NSC 52902	✓
NSC 45741	✗
NSC 287050	✓
NSC 44138	✗
NSC 94017	✗
NSC 610930	✗
NSC 255980	✗
NSC 154829	✗
NSC 2561	✓

Table 2.2: *List of the 23 compounds from the NCI Diversity Set among the top 100 hits of the virtual screening, identified by the acronym "NCS" and a numerical code. The candidates evaluated as presenting the best compromise between pharmacokinetic properties, solubility, lipophilicity and hERG safety profile are marked with a green tick and are those taken into consideration in the molecular docking phase. Rejected compounds are marked with a red "x".*

2.5 Molecular Docking and Compounds Selection

Following the selection of 10 compounds based on pharmacokinetic and safety properties, molecular docking was performed to evaluate the binding affinity between the candidates and the target, the UBR-box domain of UBR1 E3 ligase. The docking set was expanded through similarity search to include compounds structurally similar to the selected candidates, and Gnina, which combines traditional and convolutional neural network-based scoring functions, was then employed. A detailed evaluation of ADMET properties was performed, and all results were integrated to select the compounds with the best global profile for subsequent experimental testing. Finally, to explore emerging methodologies, a comparison was conducted using AI-based docking approaches.

2.5.1 Similarity Search and Docking Set Construction

The 10 compounds identified by the evaluation of pharmacokinetic properties and hERG toxicity profile were obtained from virtual screening on the NCI Diversity Set and therefore they all belong to this database. Although its very small size was useful in terms of computational weight, it was considered appropriate to expand the docking set by means of similarity search, i.e. a search for structural analogues. This approach made it possible to increase the structural and chemical coverage around the best hits identified, including compounds with possible functional variations useful for improving affinity or ADMET properties. The relationship with the pharmacophore model and therefore with the chosen UBR-box binding pocket was maintained, while at the same time exploring structures that could be more stable or more easily synthesised.

Several studies in the literature present a similar workflow, although they do not directly use similarity search to expand the initial docking set. One example is the work of Patidar and colleagues, who used this approach to identify new candidates (structural analogues) from a much larger virtual library, the ZINC database. They started with an active lead found in the previous step in order to find a compound with higher affinity [83]. Furthermore, there is evidence that using ultra-large libraries for docking overcomes the limitations of smaller physical libraries, revealing new chemical structures that would otherwise be inaccessible. No work was found in the literature that precisely followed the steps of this project, but due to the potential benefits explained above, it was chosen to follow it anyway.

For this purpose, Cartblanche22 was employed, a website provided by the Irwin and Shoichet Laboratories of the Department of Medicinal Chemistry at the University of California, San Francisco (UCSF). It is an advanced GUI used to access and query ZINC-22 database, and it has been specifically developed to facilitate the selection and extraction of molecules for use in virtual screening and docking studies.

Cartblanche22 has many features: the one used in this project is molecular similarity search. This tool, called Analog By Catalog (ABC), allows you to quickly identify commercially available molecules similar to a given structure and uses the Smallworld method, which is based on graph edit distance. The SmallWorld technology from NextMove Software is an important advance in terms of how computers compare molecules.

Cartblanche is built using the Python Flask framework. It is possible to target specific types of analogues, such as substructures or scaffolds, using advanced features, or find many compounds and their close analogues at the same time using bulk search. To quickly find molecules containing a given substructure or molecular pattern expressed as SMILES or SMARTS, the Arthor tool can be used. This allows users to identify and interactively view up to 20,000 molecules. It allows searching by ZINC code (Lookup by ZINC ID), which allows you to consult up to a thousand identifiers simultaneously, belonging to both ZINC-22 and the previous versions ZINC-20/ZINC-15, and searching by supplier code (Lookup by supplier code). The Lookup by multiple SMILES mode enables parallel searches on up to a thousand chemical structures in SMILES format. The user can specify the degree of correspondence using the distance (dist) and anonymous distance (adist) parameters. Additionally, it is possible to filter and select portions of the available chemical space based on various molecular parameters, including:

- Heavy Atom Count (HAC);
- lipophilicity (calculated logP value);
- net charge of the molecule;
- file format (mol2, SDF, PDBQT, SMILES, or DB2), in both 2D and 3D representation.

Subsets of the database can be selected and downloaded using the Tranche Browser. Once the selection is complete, a script can be downloaded to access or download the selected molecules in the various formats. It also acts as a shopping cart tool for managing and prioritizing the molecules to be purchased and an approximate price estimate can be calculated to provide guidance (although the actual price is only provided by the supplier) [50].

In the current project each of the 10 compounds identified so far and their conformers present in the first 100 hits of the virtual screening were entered as input to Cartblanche22, in SMILES format. The database chosen to search for similar compounds was ZINC, updated to the second quarter of 2022, containing approximately 1.6 billion unique chemical compounds. The search was carried out on the entire set of molecules present in the database, without limitations to specific sub-collections and therefore in its most extensive version. In fact, “ZINC-All-22Q2-1.6B” was selected in the interface. This database represents one of the largest collections of commercially available compounds. The chemical similarity parameters set are reported in Table 2.3.

Parameter	Value	Description
Distance	12	Computed dissimilarity using molecular fingerprint comparison. A value of zero indicates an exact match
Anonymous Distance	4	It ignores the identity of atoms, evaluating only the topology of the molecule For instance, a value of 1 indicates the opening or closing of an aromatic ring, the addition or removal of a single atom, or the increase or decrease in the length of a chain.
Terminal	4 Up 4 Down	Sliders that control the degree of tolerance to structural changes in specific regions of the molecule (e.g., termini, rings, ligands, mutations, substitutions, or hybridization).
Ring	4 Up 4 Down	
Linker	4 Up 4 Down	
Mutation	4 Up 4 Down	
Substitution	4	
Hybridization	4	

Table 2.3: *List of parameters set in the Cartblance22 GUI to regulate the degree of chemical similarity between the query molecule and the molecules present in the ZINC-22 database, with their values and descriptions.*

The values set define a medium-high similarity search, which allows for minimal topological variations while maintaining the general structure of the starting molecule. This strategy allows for the exploration of potentially more active or synthesizable structural analogues. This made it possible to identify and download, in the SMILES format, structures with a similarity greater or equal to 95% to the query molecules.

This data represents the Tanimoto Coefficient (Tc, similarity value) calculated using the ECFP4 fingerprint. The latter is a circular fingerprint, i.e. a binary representation of the chemical structure. The Extended-Connectivity Fingerprints (ECFP) family, whose generation is based on the Morgan algorithm, records the environment of each atom up to a predetermined radius. ECFPs, in particular, represent circular atomic environments and are one of the most popular methods for similarity search, virtual screening and QSAR analysis. The number following the acronym ECFP indicates the diameter of the circular atomic environment considered, so ECFP4 is the version of ECFP in which the fingerprint is calculated using a diameter of 4. This means that it captures the chemical environment of each atom up to two bonds away, encoding the structure in a sensitive but not overly specific way to achieve a good compromise between specificity and generality.

The process for calculating Tc is as follows: the query molecule and the target molecule are converted into a fingerprint (ECFP4), i.e. a binary vector in which each bit indicates the presence or absence of a certain chemical fragment. The two vectors are compared, and Tc is calculated as:

$$T_c = \frac{c}{a + b - c}$$

where a corresponds to the number of bits set to 1 in fingerprint A, b is the number of bits set to 1 in fingerprint B and c is equal to the number of bits set to 1 in both fingerprints (A and B). Its value is always between 0 and 1 [84] [85].

It should be noted that the one conducted here is not a massive processing. If more complex searches are required (e.g., more than 1000 molecules or with more sophisticated criteria), it is possible to download a portion of the database locally and analyse it with tools such as RDKit or other chemoinformatics software. Initially, an attempt was made to do this, but the size of the database proved to be too large to be managed locally.

Through this approach, a total of 124 compounds were found to be 95% similar to the 10 compounds used as queries. The docking set thus consisted of 134 candidates. For each query molecules, an average of a dozen compounds was found, with the exception of NSC 143099, for which no molecules were found that were at least 95% similar to it.

The list of candidates identified at the end of this phase can be found in Appendix D.

2.5.2 Docking with GNINA

Molecular docking is of fundamental importance in drug discovery, as it is a computational approach used to predict non-covalent interactions between molecules, more specifically between a protein receptor and a small ligand. It is able to simulate the physical-chemical principles that govern these interactions (van der Waals forces, hydrogen bonds, electrostatic and hydrophobic interactions).

Due to its purpose, this tool was deemed essential for evaluating the pose and conformation of each of the 134 ligands included in the docking set constructed in the previous phase of the project within the binding pocket of the UBR-box domain of the selected UBR1 E3 ligase. Simultaneously, the binding affinity can be estimated for each pair of molecules. The predicted positions are in the minimum energy state. This tool is used for the virtual screening of large compound libraries.

The docking process consists of two main phases called Sampling and Scoring. The first explores the possible poses of the ligand within the receptor binding pocket and generates a set of plausible orientations. Both the receptor and the ligand can be flexible, so the conformational space to be searched is very large. For this reason, it is necessary to limit it: the receptor is typically kept rigid. However, it should be noted that certain more advanced software (such as Gnina) allows the flexibility of some side chains to be explored. Sampling strategies can vary and can be more or less computationally expensive. Systematic strategies exhaustively explore all configurations and are very accurate, while stochastic strategies are based on random or evolutionary algorithms (e.g. Monte Carlo, genetic algorithms). To refine the pose, they can be based on molecular dynamics or energy gradients.

At the end of this phase, each sampled conformation is assigned a numerical score using a function to evaluate its fitness. A ranking is created based on the probability of the pose being correct, which determines which ones will be retained. This process is essential for determining bond affinity and exploiting the configuration obtained to optimise the lead compounds.

Scoring functions can vary depending on the approach used. Those based on physical force use the energetics of interactions, such as Coulomb and Van der Waals forces, to calculate scores. Knowledge-based functions use statistics derived from a set of known binding structural data, for example, such as the PDB database. Empirical functions combine manually selected energy terms, with the weights of each term determined by fitting to experimental data. X-Score and AutoDock Vina, as well as many other docking software packages, use the latter approach. Finally, there are scoring functions based on machine learning/deep learning, as explained below.

The process outputs one or more poses ranked by score, which ideally reflect the probability that the pose represents the native complex.

Gnina was the next-generation open-source software chosen to carry out this phase of the project and evaluate the poses and binding affinity between the receptor and the ligands. It was developed as an extension of Smina which in turn is a fork of AutoDock Vina, but unlike the latter it is enhanced with convolutional neural networks (CNNs) as scoring functions, thus based on deep learning, to improve the evaluation of binding poses. The sampling engine therefore remains the classic docking (Vina) with stochastic optimization. Gnina performs flexible docking of the ligand with a rigid receptor, similar to Vina, but, thanks to deep learning, takes into account complex three-dimensional features (such as interaction density, spatial distribution of atoms, and hydrophobic potential) that traditional empirical functions do not capture well.

Gnina, in version 1.0.1, was used on the Cedar cluster, made available by the Digital Research Alliance of Canada, which was decommissioned in September 2025 and all files were transferred to the Fir cluster. In this way, as in the case of molecular dynamics simulations, the execution time was significantly reduced compared to running on local machines. The following inputs were entered into the software:

- **the prepared receptor file**, i.e. the homology model of the UBR-box of UBR1 E3 Ligase, in PDBQT format (centroid_prepared.pdbqt in Figures 2.8 in blue and 2.9)

This file was constructed from the PDB format file with the most representative structure of the second half of the molecular dynamics trajectory, also used for the search for binding pockets. The solvent molecules were not present, unlike the three zinc ions, which were to be retained. At this stage, it was important to add correctly protonated hydrogens to achieve a pH of 7.4, as this changes the charge of some residues such as Asp, Glu, Lys, and His. To achieve this, ChimeraX software was employed, setting the following options:

✓ Protonation States for: histidine
✓ Residue-name-based (HIS/HID/HIE/HIP)
✓ Do not protonate electronegative atom near metals
✓ Also consider H-bonds

In ChimeraX, the exact pH (e.g., 7.4) is not set directly with a number, but the effect is automatically approximated using the previous options, assigning protons to simulate a physiological environment.

- **The files of the 134 ligands** in SDF format but in individual files.

The SMILEs of potential candidates derived from both the virtual screening and similarity search phases were entered into an Excel file, which was then converted to CSV format. From there, using RDKit, it was possible to generate the three-dimensional coordinates of all ligands, saving them individually in SDF format files, ready for use by Gnina.

- **The configuration file**, which specifies the receptor and ligand files, the coordinates of the centre of the box and its dimensions, and several other parameters.

To find and establish the coordinates to be set as the centre of the box, the receptor file in PDBQT format and the FTSite output file were superimposed in PyMOL, clearly showing the pocket chosen for ligand binding. Using the two commands:

  - print cmd.centreofmass('pocket');
  - get_extent pocket,

it was possible to obtain the coordinates of the centre and the dimensions of the pocket, respectively. The latter were then increased to allow all poses, even those of larger ligands, to fit into the box so that it correctly covers the protein binding pocket. The box is shown in Figure 2.8.

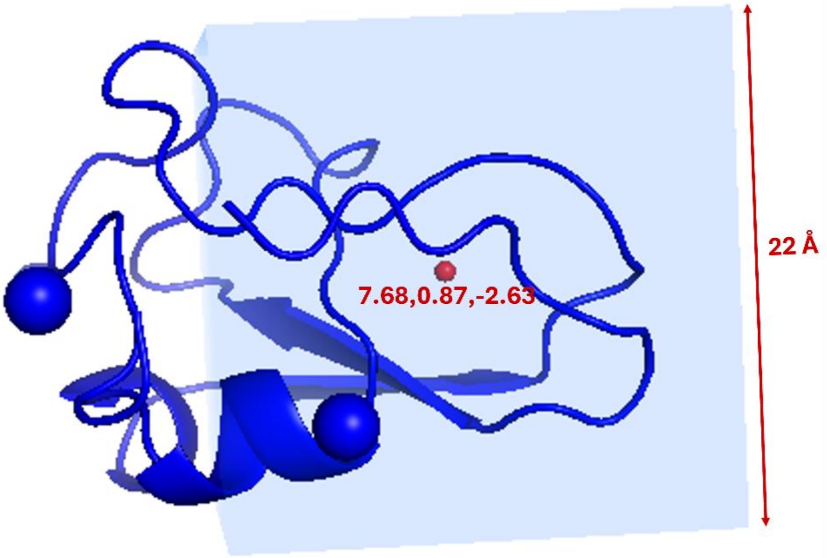


Figure 2.8: *Image of the homology model of the UBR-box domain of the UBR1 E3 Ligase protein prepared for docking (in blue) and of the box (in light blue with centre and dimensions specified in red) set in the configuration file.*

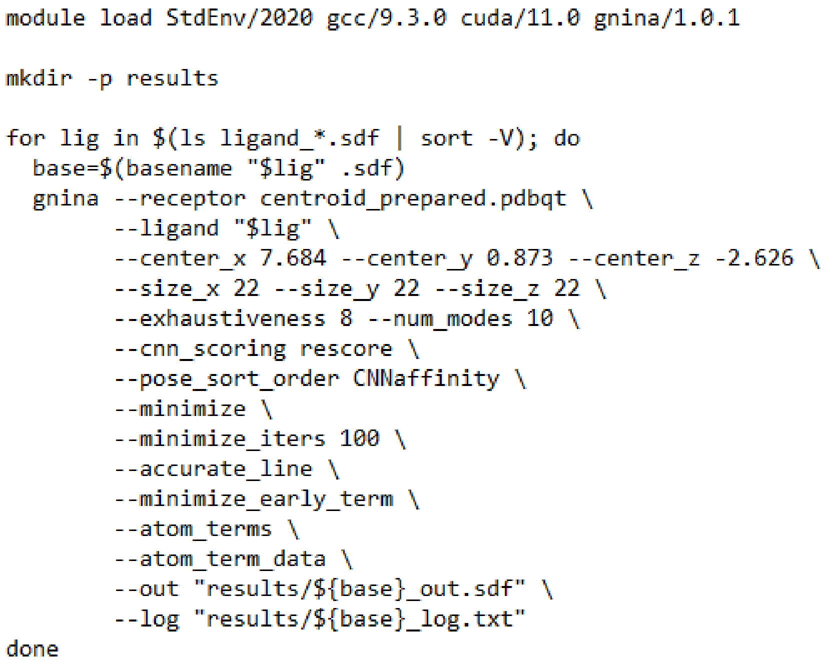
Furthermore, since Gnina does not automatically align or centre ligands with respect to the binding pocket or box, all ligands were translated with code executed on RDKit so that their centroid fell within the centre of the box. In Gnina, with the 'centre_*' command, it was possible to define each of the three dimensions of the grid box, i.e. the area explored by the

docking, with ‘size_*’ its dimensions. The box set up in this way allowed a large area to be explored, ensuring complete coverage of the binding site but without excessive computational cost.
The exhaustiveness parameter controls the depth of exploration of the conformational space. Higher values indicate a more thorough but slower search. Here it was set to 8: a good compromise for multiple screenings. The maximum number of poses saved per ligand was 10. The “rescore” function (default, recommended by Gnina) was used for the reclassification of the final poses as it is less computationally expensive. The final poses were sorted according to the affinity score predicted by the CNN. The following parameters were set to control the energy efficiency of the poses:

- ‘minimise’ activates local energy minimisation after docking;
- ‘minimise_iters’, i.e. number iterations of steepest descent, has been set to 100 to achieve a balance between accuracy and time;
- ‘accurate_line’ uses a more precise optimisation algorithm;
- ‘minimise_early_term’ to stop minimisation if the energy variation is negligible, reducing calculation time;
- ‘atom_terms’ and ‘atom_term_data’ include detailed atomic energy terms in the output, useful for subsequent analysis.

The output files with docked and minimised poses were saved in SDF format in the ‘results’ created folder. Thanks to the LOG file, it was possible to view the scores, the parameters, the execution times, and the other details of the docking process.

The code was iterated on all SDF files of the 134 ligands. The sort -V command sorts them numerically since they were saved with progressive numbering. The base command was used to create the output names, and each ligand was saved in a dedicated output file [86].



```
module load StdEnv/2020 gcc/9.3.0 cuda/11.0 gnina/1.0.1

mkdir -p results

for lig in $(ls ligand_*.sdf | sort -V); do
  base=$(basename "$lig" .sdf)
  gnina --receptor centroid_prepared.pdbqt \
        --ligand "$lig" \
        --center_x 7.684 --center_y 0.873 --center_z -2.626 \
        --size_x 22 --size_y 22 --size_z 22 \
        --exhaustiveness 8 --num_modes 10 \
        --cnn_scoring rescore \
        --pose_sort_order CNNaffinity \
        --minimize \
        --minimize_iters 100 \
        --accurate_line \
        --minimize_early_term \
        --atom_terms \
        --atom_term_data \
        --out "results/${base}_out.sdf" \
        --log "results/${base}_log.txt"
done
```

Figure 2.9: *Configuration file used as input for the Gnina software, specifying the receptor file, the ligand files, the centre and dimensions of the binding pocket box, the docking parameters, and information for saving the results.*

The CNN models employed in Gnina were trained using data including the PDBbind database, which annotates structural data with experimental binding affinity data. During training, the network "learns" to distinguish between correct poses, similar to the real crystallographic structure, and incorrect poses, i.e., those with unfavorable geometries or orientations. The generated model is thus able to automatically evaluate the new poses produced by docking, recognizing the most realistic or energetically favorable ones.

CNNs work as follows. The layers apply filters (or kernels) that run over the input to detect local patterns, each producing an activation map that indicates where that pattern was found. Initially, the filters recognize simple features (e.g., distances between atoms or small geometric patterns), but as we delve into deeper layers, the true convolution operation occurs, combining this information to identify more complex patterns (e.g., hydrophobic interactions, H-bonds, coordination geometries). Subsequently, to reduce complexity and make the network more robust to small spatial variations, it can reduce the spatial size of the maps, retaining only the most relevant information (pooling operation).

The affinity score obtained is the numerical prediction produced by the final layers of the network, which combine all the detected characteristics.

Compared to traditional approaches, CNNs applied in this context offer greater accuracy in distinguishing correct from incorrect poses, the ability to generalize to new targets and direct learning from experimental data. Performance increases for redocking from 58% for Vina to 73% for Gnina Default Ensemble when the binding site is defined, while for cross-docking they increase from 27% to 37%. When using GPU with the "rescoring" option, the increase in computation time is only two seconds [87].

The parameters evaluated at the end of this process are as follows:

- ✓ Affinity [kcal/mol];
- ✓ CNNscore;
- ✓ CNNaffinity [kcal/mol].

For their analysis and their detailed description, see the subchapter 3.5.1.

A Python code made it possible to automatically extract these output parameters, contained in the individual LOG files for each ligand, and insert them all into a CSV file. Thanks to this process, it was possible to analyse and compare the possible candidates more easily.

2.5.3 In-Depth ADMET Evaluation

As previously explained, selecting molecules with favorable profiles in silico allows experimental efforts to be focused on more promising candidates. A first, more general approach to this evaluation was made possible by the web application Chemaxon Playground Calculator, which included the 23 candidates selected at the end of the virtual screening phase.

This process, being performed individually for each compound, was time-consuming, and therefore it was not feasible to use it in the same way for the docking set expanded with similarity search. It was therefore necessary to use a tool able of rapidly processing a large number of molecules, returning a complete set of ADMET predictions in a reasonable amount of time.

For this purpose, ADMET Predictor v12.0 developed by Simulations Plus was employed, which proved suitable due to its high-throughput prediction capabilities. In fact, it allows for the rapid prediction of over 175 ADMET properties.

ADMET Predictor is a commercial AI/ML modelling platform useful for predicting the absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of molecules. When providing a prediction for a property, this software also returns confidence indicators that give a measure of how much the model “trusts” the predicted value, allowing high-confidence predictions to be distinguished from less reliable ones. This makes it possible to check whether a molecule is out of domain and, if so, to treat it with caution.

In version v12.0 (also known as AP12), significant improvements have been made compared to previous versions. Examples include expanded models for bio-relevant solubility, clearance, permeability, CYP induction, integrated HT-PBPK simulations, and high-throughput liver injury prediction (DILI) modules. At the end of the process, the predictor returns a tabular report with all the predicted properties.

This tool has been validated in several publications and is widely used in industry as a fundamental step in computational discovery. One example is the study by Maral Aminpour and colleagues, who used version 9.5 of the software to predict the ADMET profiles of 90 potential drugs in order to mitigate failures in the advanced stages due to pharmacokinetics or toxicity [88].

The Excel file used as input to the predictor was constructed so that the first column contained the IDs of the docking set compounds and the corresponding SMILES format in the second. The generated CSV file contained the compounds on each row and the corresponding property result in each column. To make the data easier to view, the CSV file was converted to Excel using the Datablist web platform, which, among its many features, allows data to be extracted using AI. This made it possible to choose which of the numerous properties to keep for subsequent evaluations and which to discard. Indeed, it was deemed impossible to compare all the properties generated by the ADMET Predictor across all 134 candidates.

Twenty-three properties were selected for analysis and can be divided into four macrocategories: Lipinski’s rule of five, toxicity, permeability and absorption, and metabolism. The list, thus divided, is visible in the Table 2.4.

Macro-category	Property	Description	Ideal values or condition
Lipinski’ s Rule of Five	MWt	Molecular weight of the molecule	< 500 Da
	HBA	Number of groups capable of accepting H-bonds	≤ 10
	HBD	Number of groups capable of donating H-bonds	≤ 5
	RuleOf5	It indicates whether the molecule violates the rule in terms of weight, HBD, HBA, logP	0 violations

Macro-category	**Property**	**Description**	**Ideal values or condition**
Lipinski’ s Rule of Five continued from previous page)	RuleOf5_Code	It indicates which Lipinski rules have been violated. ! H: too many hydrogen bond donors (HBD > 5) ! Hb: too many hydrogen bond acceptors (HBA > 10) ! Mw: molecular weight (MWt > 500 Da) ! NO: logP too high (logP > 5)	Empty space
	T_PSA	Topological Polar Surface Area that indicates the polarity and lipid solubility of a molecule	≤ 140 Å² for oral absorbability; ≤ 90–100 Å² for good permeability
Toxicity	hERG_Filter	Qualitative classification of the risk of hERG channel blockade	No
	hERG_pIC50	Prediction of hERG channel inhibition expressed as pIC_{50}	< 5 (i.e. IC_{50} > 10 µM) for low risk
	Repro_Tox	It assess the risk of reproductive toxicity	Nontoxic
	TOX_Risk	General toxicity risk index	Lowest value
Permeability and Absorption	%Fa_hum-1.0	Estimated percentage of human intestinal absorption	Ideally > 50%
	S+logD	Logarithm of the distribution coefficient (D) indicating the lipophilicity of an ionisable molecule	Between 1 and 3
	S+logP	Logarithm of the octanol/water partition coefficient (P) which indicates the lipophilicity (or hydrophobicity) of a substance	Between 1 and 3 (for neutral forms equal to logD)
	S+MDCK	Prediction of intestinal permeability by MDCK (Madine-Darby Canine Kidney) cells in vitro	Higher values
	S+Peff	Prediction of human intestinal effective permeability (in cm/s $\times 10^{-4}$)	> 1 good < 0.1 low
	S+Sw	Predicted water solubility, expressed as logarithm	> –4
	BBB_Filter	Ability to cross the blood-brain barrier (BBB)	Low
	LogBB	Logarithm of the blood/brain concentration ratio indicating how much of the molecule enters the brain	< –1 poor entrance

Macro-category	Property	Description	Ideal values or condition
Metabolism	CYP2D6_Inh	Cytochrome P450 2D6 inhibition index indicating risk of metabolic drug interactions	No (non-inhibitor)
	CYP3A4_Inh	Cytochrome P450 3A4 inhibition index	No (non-inhibitor)
	S+CL_Metab	Predicted metabolic clearance which indicates if the molecule is rapidly metabolized	No (not significantly metabolised)
	S+CL_Renal	Predicted renal clearance indicating if it is eliminated by the kidneys	No (is not eliminated to any significant extent via the kidneys)
	S+CL_Uptake	Predicted clearance via uptake (e.g., transport-mediated, hepatic excretion)	No clearance for cellular uptake not significant

Table 2.4: *List and brief description of the ADMET properties evaluated with the ADMET Predictor software (developed by SimulationPlus) for each of the 134 ligands that comprise the docking set. For each property, the value or an ideal condition is reported. They are divided into four macro-categories: Lipinski's rule of five, toxicity, permeability and absorption, and metabolism.*

The evaluation of the properties of the ligands composing the docking set mentioned above, along with the analysis of the docking output parameters, can be found in subchapter 3.5. From this, it was possible to identify the ligands for subsequent experimental tests.

Since it was not possible to identify compounds that simultaneously met all the criteria reported in Table 2.4, the selection focused on those that satisfied the greatest number of parameters and showed a better overall profile, giving priority to the properties most relevant for the design of the PROTAC.

It should be emphasized that although integrating early ADMET modeling is a recognized step in drug discovery workflows to avoid advancing candidates with unacceptable pharmacokinetic or safety profiles, in silico predictions cannot replace experimental testing. Indeed, while they are useful as a priority filter, any selected molecule must be experimentally validated.

2.5.4 Selection of Candidates

The docking output parameter values and the properties of each of the 134 ligands were entered into an Excel spreadsheet for easier comparison.

Among these, 28 compounds were selected for subsequent experimental testing at the University of Alberta laboratory.

The main criterion used was the Affinity parameter, which for the selected compounds ranges between –4 and –7 kcal/mol.

Due to time and cost constraints, they were divided into three groups:

- **First choice**, those deemed best in terms of receptor affinity, chemical properties and safety profile, and therefore most likely to be active towards the target;
- **Second choice**, to be tested after the first group;
- **Third choice**, considered valid candidates but last to be tested.

The number of selected candidates was chosen based on the testing laboratory's recommendation. Three candidates were identified for each compound family.

A family is defined as the group containing ligands at least 95% similar (searched on ZINC-22) to the compound selected at the end of the virtual screening phase (derived from the NCI Diversity Set). For this reason, there are 10 families. However, the candidates for testing are 28 because compound NSC 143099 did not contain any compounds that met the established similarity threshold, and therefore, this candidate is the only one present in the family. It was therefore included only in the first-choice group.

3. Results

This chapter presents the results obtained in each phase of the project, from the construction of the two target models to the selection of ligands for experimental testing. The workflow already used for Chapter 2 was followed.

3.1 Analysis of the models

Below, the two target models are analysed: the homology model of the UBR-box domain of UBR1 E3 Ligase and the model of the entire protein predicted by AlphaFold2. They are then compared to understand how they differ from each other.

3.1.1 Homology Model

The first step of the model construction was the sequences alignment of the prepared structure of 3NY3 in FASTA format (canonical) found with PyMol and the one of HUMAN E3 ubiquitin-protein ligase UBR1 in FASTA format (canonical) taken from Uniprot.

The results obtained with EMBOSS water software are summarised below:

Parameter	Value
Length	70
Identity	54/70 (77.1%)
Similarity	61/70 (87.1%)
Gaps	0/70 (0.0%)
Score	337.0

Table 3.1: *Results of the alignment of the 3NY3 and UBR1 sequences obtained with EMBOSS Water.*

The outputs indicate that 3NY3 is at least identical to UBR1 in 77.1% of amino acids. This means that many amino acid positions are perfectly conserved between the two structures. The high similarity shows that most residues are chemically similar, with few substitutions that could influence the local configuration. This suggests that function and conformation could be similar in the two protein segments. There are no gaps in this alignment of 70 amino acids, indicating that the local structure of this region may be well represented without discontinuities or significant changes in conformation. Finally, a score of 337.0 means a strong alignment. For these reasons, 3NY3 structure thus prepared, was considered suitable as a template for constructing the homology model of the E3 ligase UBR1.

The model generated with Swiss-MODEL has the following characteristics:

Feature	Result
Oligo-State	Monomer
Ligands	3 x ZN
Seq Identity	77.94%
GMQE	0.04
Coverage	0.04
QMEANDisCo Global	0.75 ± 0.11

Table 3.2: *Summary of the most significant parameters describing the homology model of UBR1 E3 Ligase obtained with Swiss-MODEL, based on the 3NY3 structure.*

The generated model is a single polypeptide chain. The presence of three zinc ions indicates that they were retained in the structure.

The sequence identity is high (77.94%), roughly the same as that found with EMBOSS-Water (see Table 3.1).

The Global Model Quality Estimate (GMQE) is a score that predicts model quality based on coverage and sequence identity between the target and the model. A score close to 1.0 is considered excellent. The value found of 0.04 is rather low. In this case, many portions of the complete UBR1 protein are missing from the 3NY3 structure, as it only represents the UBR-box domain. In fact, the generated model is monomeric, whereas UBR1 is a multimeric protein. This inevitably leads to a decrease in coverage (the "coverage" item in the results has the same value as GMQE). Therefore, despite this low value, it was decided to continue the project with this model. However, to overcome this problem and analyse UBR1 in its entirety, its complete model predicted by AlphaFold2 was also investigated (see subchapter 2.2).

The Qualitative Model Energy Analysis - Distance Constraints (QMEANDisCo global score) evaluates the way the interatomic distances in the model match the ensemble data derived from protein structures that have been established experimentally and are homologous to the target sequence. The global value obtained (0.75 ± 0.11) is considered reasonable. Values closer to 1 indicate high confidence. The extremities have values that fall below 0.6, but this is because the extremities are often flexible and more difficult to model.

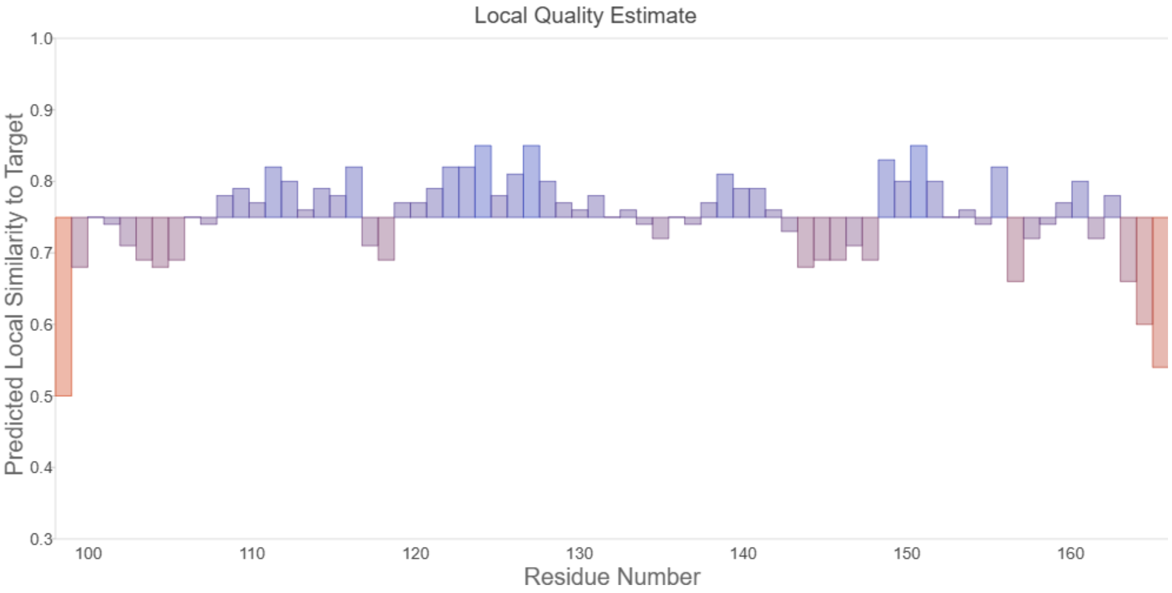


Figure 3.1: *The graph shows the global QMEANDisCo value (y-axis) for each residue of the model (x-axis).*

The Z-score analysis QMEAN is generally calculated for all models except those using AlphaFold DB models. It is predicated on the linear combination of four possible mean strength statistics. Z scores are used to compare the five scores with those predicted from similarly sized experimentally determined structures (see the graph a in Figure 3.2). Therefore, a "native-like" structure is reflected by Z scores around 0.0, while a QMEAN Z score below -4.0 often denotes a low-quality model. In this case, the Cβ which evaluates the geometric deviations related to the Cβ atoms is close to zero. The All Atom parameter, which measures the interactions between all atoms, indicates that no major anomalies exist. Solvation indicates small, but not critical, discrepancies in the hydrophobic/hydrophilic distribution. Torsion indicates that the torsional angles fall in favorable regions. The global QMEAN (–1.05) indicates that the structure is reasonably realistic. In general, the values range between -0.42 and -1.07.

The graph b in Figure 3.2 represents the comparison between the generated model and experimentally determined structures. The overall Z-score is close to 0, so the model is consistent with real proteins and no serious deviations are observed. It is observed that the generated protein is particularly long compared to the length of those it was compared to. However, for overall estimates of model quality, it is preferable to consult the overall GMQE and QMEANDisCo scores [89].

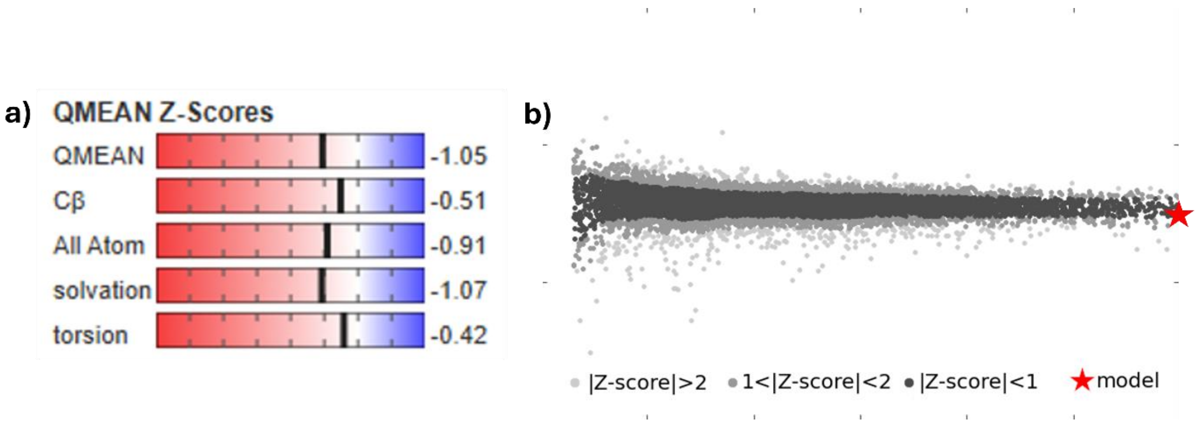


Figure 3.2: *Graphs generated with Swiss-MODEL. Graph a) shows the different QMEAN score parameters compared to those of known experimental structures from the PDB database: QMEAN, Cβ, All Atom, Solvation, and Torsion. Graph b) illustrates the Z-score of high-quality structures present in the PDB compared to the generated model, represented by a red star. The number of protein residues is on the x-axis, while the QMEAN score is on the y-axis. An experimental protein structure is shown by each point. Grey experimental structures have a |Z-score| between 1 and 2, whereas black dots have a QMEAN score within 1 standard deviation of the mean (|Z-score| between 0 and 1). Light grey are experimental constructions that deviate significantly further from the mean.*

The Ramachandran plot of the generated model was analyzed. It was conceived in 1963 by G. N. Ramachandran and colleagues, and it is still a useful tool for checking the plausibility of protein models. It provides a measure of structural quality, showing which regions of the structure might be less reliable. For each amino acid residue present within the main chain, the plot indicates all sterically allowed combinations of the peptide bond torsion angles φ (phi) and ψ (psi).

There is a division between allowed and prohibited regions. The first appear darker and do not cause steric collisions between the backbone atoms or between the backbone and the β-carbon atoms of the side chain. They therefore correspond to stable regions such as α-helices (with φ-ψ values close to -57° and -45°) and β-sheets (with φ-ψ values close to -135° and -135°). Then, there are lighter areas that are considered unreachable due to steric clashes. If there are residues that fall into these regions, it means they are less reliable and therefore the model requires further checks and refinements.

It is currently the basis of many protein structural validation software programs such as PROCHECK and MolProbity (integrated into Swiss-MODEL), which are able to define regions based on a large amount of experimental data from resolved, high-resolution structures. Beyond this purpose, this tool can also be used for initiation modelling, which predicts the three-dimensional structure of proteins solely from the amino acid sequence, or for defining secondary structure. It is also possible to refine theoretical models and experimental structures by deriving mean force potentials [90].

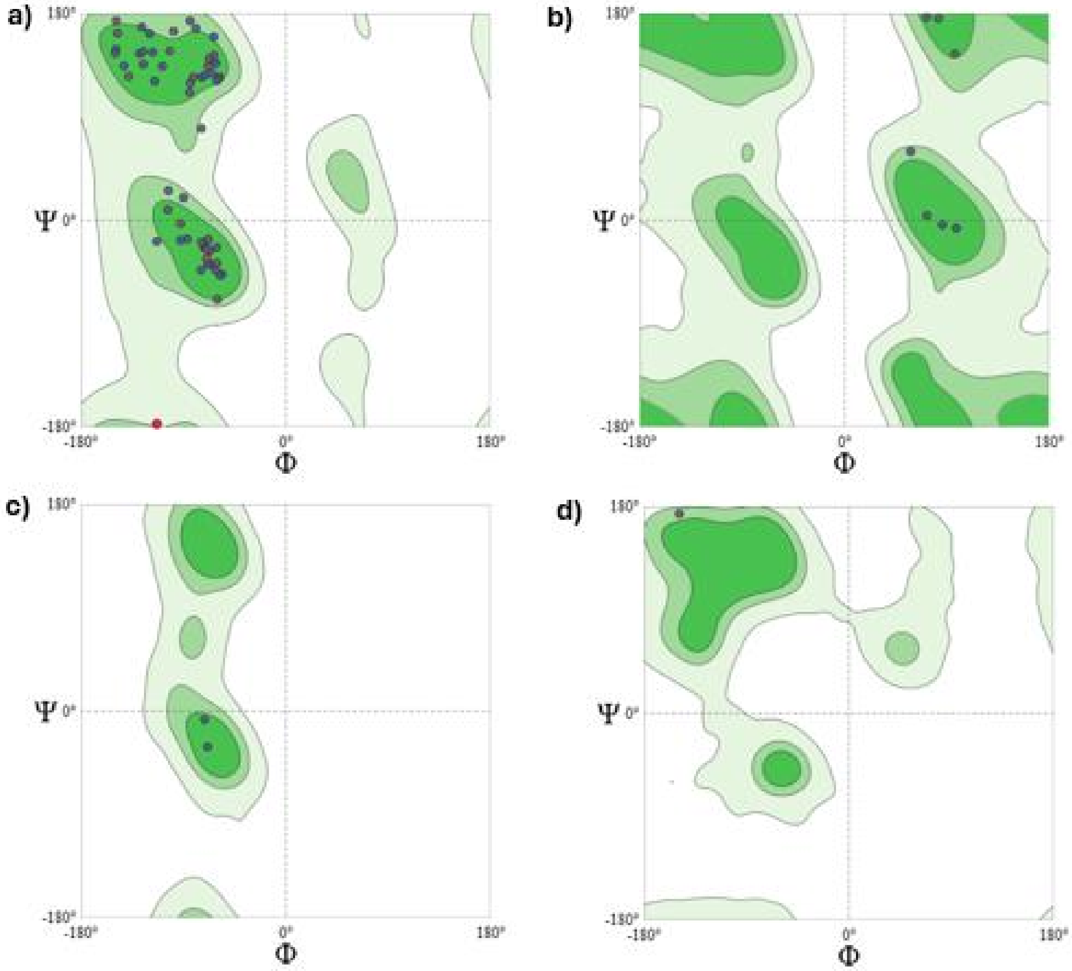


Figure 3.3: *Ramachandran plots of the UBR1 homology model, built with MolProbity, integrated into Swiss-MODEL.*
Graph a) represents the general (no Proline or Glycine).
Graph b) shows only Glycine residues.
Graph c) displays only Proline residues.
Graph d) reveals only Pre-Proline.
The torsion angle φ (phi) is on the x-axis while the torsion angle ψ (psi) is on the y-axis. The dark green regions indicate sterically allowed regions, the light green regions show allowed but less frequent regions, and the white regions represent prohibited regions. Each point on the graph represents a residue.

Analyzing the Ramachandran plot of the homology model of UBR1 E3 Ligase generated with Swiss-MODEL, it can be noted that most of the residues fall in the dark green regions, i.e., the most conformationally favoured ones. This confirms the correct folding and is consistent with what is expected for a real protein. Specifically:

- in the upper left area of the graph, clusters corresponding to β-sheets are visible;
- in the lower left area, there is a group of residues related to α-helices;
- very few points are located on the border between the permitted and the non-permitted zone. This was considered acceptable since, in experimental structures, one or two residues usually belonging to the more flexible ends are difficult to model.

Parameter	Value
MolProbity Score	1.13
Clash Score	0.99
Ramachandran Favoured	97.01%
Ramachandran Outliers	0.00%
Rotamer Outliers	1.67%
C-Beta Deviations	0
Bad Bonds	0 / 543
Bad Angles	12 / 732

Table 3.3: *Summary table of parameters obtained through structural validation with MolProbity, integrated into Swiss-MODEL.*

The MolProbity Score combines several quality parameters such as crush score, Ramachandran, and rotamers. The value found, which should be as low as possible, represents good quality and indicates that the structure found is comparable to high-resolution crystallographic structures. The low value found reflects the fact that, for the construction of the model, we started from a template structure with a high crystallographic resolution.

The Clash Score measures excessively close steric contacts between atoms, and more specifically, van der Waals sphere overlap greater than 0.4Å. This parameter must also be as low as possible, as is the case here.

The Ramachandran Favored parameter shows the percentage of residues in the most favorable φ/ψ angles (dark green regions in the Ramachandran plot, see Figure 3.3). Ideally, a percentage greater than 98% would be achieved, but the result obtained, 97.01%, is still excellent as long as steric collisions do not occur.

As a result, there are no Ramachandran Outliers, i.e. residues in forbidden regions. In general, however, a percentage of 0.05% is permitted.

The percentage of side chains in unlikely conformations is also low. Deviations of the Cβ atom greater than 0.25 Å from the ideal position are called C-β deviations. They are not present in the generated model, as is the case with an ideal model. The same happens with the number of bonds with anomalous lengths (Bad Bonds parameter). Bond angles out of tolerance (greater than 4σ deviations from ideal) are only 1.64%.

Looking at Figure 3.4, it can be noticed that the generated homology model does not differ significantly from the prepared 3NY3 structure, except for the more flexible ends. Template selection is therefore a crucial step. Having a structure with low refinement resolution inevitably leads to inaccuracies in the resulting structure, lowering the model's quality.

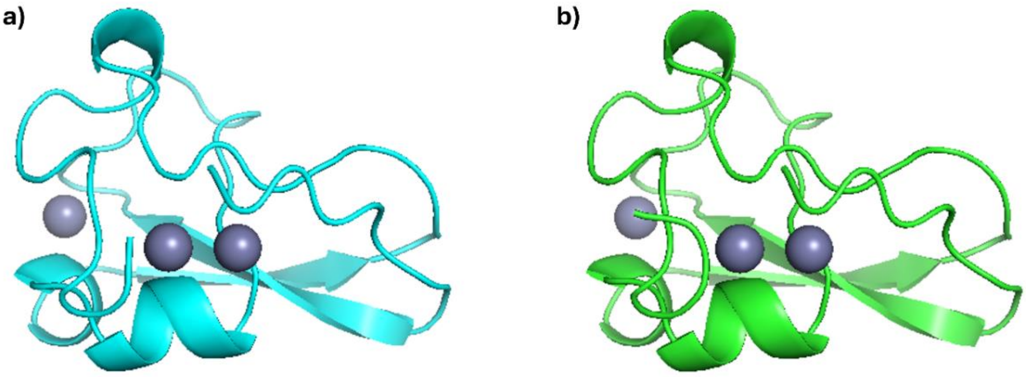


Figure 3.4: *Comparison between the structure a) of the homology model of UBR1 generated with Swiss-MODEL and b) the experimentally characterized structure of 3NY3 used as template.*

Overall, the structural analysis showed very good quality values. It was therefore decided to continue the project using this model, but for further refinement and optimization, a molecular dynamics simulation was performed. For an analysis of the results of this step, see subchapter 3.2. Since the results presented match the model of only one domain of UBR1 E3 Ligase, it was decided to also investigate the structure of the complete protein predicted by AlphaFold2 (for more information, see subchapter 3.1.2).

3.1.2 AlphaFold Model

The Average pLDDT is the average confidence level for the residues of the entire protein and allows us to evaluate the quality of the model predicted by AlphaFold2. The value found is 84.7, which is very high, indicating that the prediction is very accurate. The value of the pLDDT parameter measures the local confidence level for each residue and corresponds to the model's predicted score on the LDDT-Cα metric.

Figure 3.5 displays the predicted model of UBR1 E3 Ligase where residues have been colored according to their pLDDT value. Note that most residues, specifically 50.3%, have a pLDDT value greater than 90. These regions are expected to be modeled with high accuracy and may be appropriate for applications requiring this characteristic, such as binding site analysis. They are colored blue in Figure 2.5. Some residues (37.4%), colored light blue, have a pLDDT value between 70 and 90 and are expected to be modeled well. In these regions, unphysical bond lengths and clashes do not usually appear. Residues colored yellow and orange are few, indicating that few regions with pLDDTs are modeled with low confidence. In particular, the residuals with low accuracy are 6.7% while those with very low accuracy are 5.6%. Yellow regions should be treated with caution, while orange regions should not be interpreted. It has been demonstrated that a pLDDT value less than 50 is a reasonably strong predictor of disorder, suggesting that the region is unstructured under physiological conditions or structured only as part of a complex. Since nearly 90% of the residuals are predicted with high confidence, it can be argued that the resulting model can be considered for further studies.

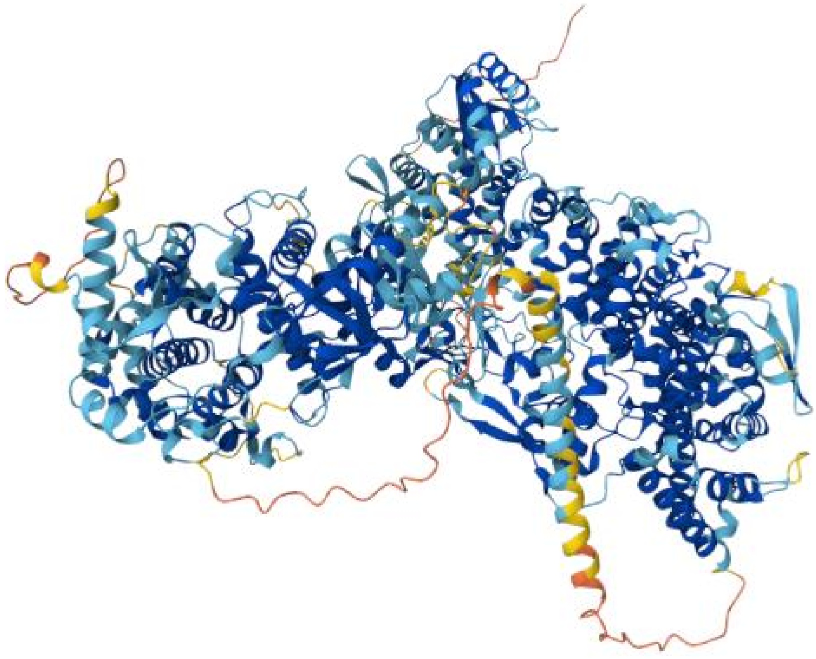


Figure 3.5: *E3 ubiquitin-protein ligase UBR1 predicted by AlphaFold2 [64]. The confidence level of the AlphaFold model prediction is represented by the pLDDT value, which ranges from 0 to 100 and is color-coded. Residues with pLDDTs above 90 are shown in blue and correspond to regions predicted with very high confidence. Values between 70 and 90 are shown in light blue and indicate high, but slightly lower, confidence. Residues colored in yellow have pLDDTs between 50 and 70 and exhibit low confidence. Values below 50 are shown in orange and reflect very low confidence.*

Within the predicted structure, it is possible to obtain information on the reliability of the relative position of two residues thanks to the Predicted Aligned Error (PAE). The graph, shown in Figure 3.6, indicates the expected distance error in Ångströms (Å), ranging from 0 Å to an arbitrary cutoff of 31 Å, as a shade of green. The color at a given point (x,y) expresses the expected distance error at the position of residue x when the predicted and actual structures are aligned on residue y. This does not indicate whether it is a contact map or inter-residue distance map. A dark green color indicates a low error (low PAE) and therefore a good prediction, while a lighter color shows a high error (high PAE) and suggests that the domains may move relative to each other. The main diagonal is always dark because a residue with respect to itself has zero error.

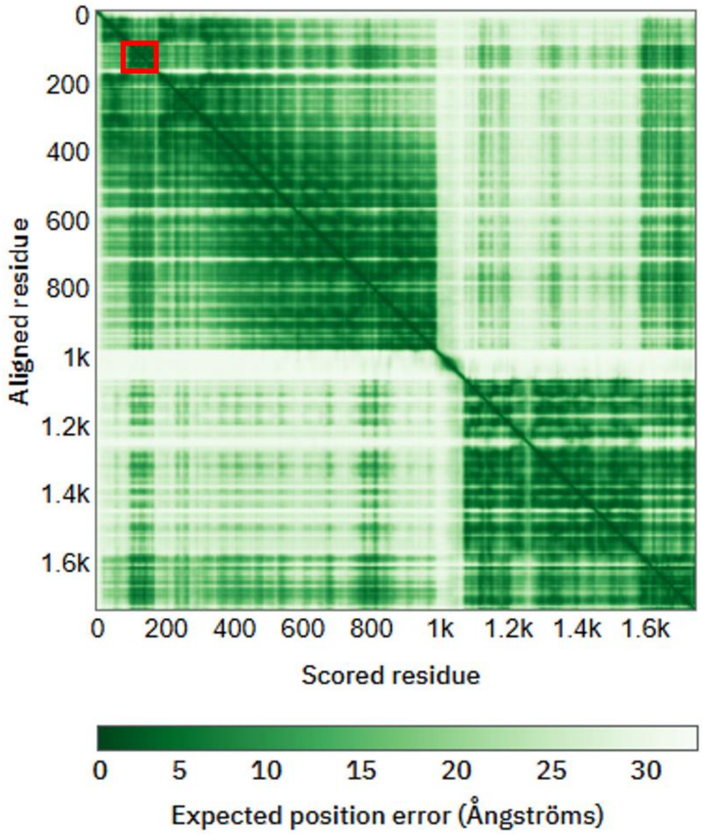


Figure 3.6: *Predicted Aligned Error (PAE) plot of the UBR1 E3 Ligase model generated by AlphaFold2 [64]. Dark green regions indicate high relative confidence, while lighter regions indicate low confidence. The x-axis (scored residue) represents the residue "chosen" to measure the error, and the y-axis (aligned residue) indicates the residue against which the error is calculated. Residues belonging to the UBR-box are highlighted by the red square.*

From Figure 3.6, it can be seen that several darker regions exist along the diagonal. The structure is therefore organized into several domains predicted with good internal confidence. An example is the UBR-box who has an average expected position error less than 5 Å. Its residues are outlined in red. This is consistent with the fact that experimental data for homologous structures in this domain are also available. However, several lighter areas also appear, indicating uncertainty about their relative arrangement by AlphaFold. This means that the individual domains were correctly predicted, but their relative positions could vary from those depicted. Therefore, there is considerable interdomain flexibility, and their bonds could be disordered. The results are nevertheless consistent with what is expected for a very long protein (more than 1,700 amino acids) and composed of multiple distinct regions.

AlphaMissense is an artificial intelligence model used to classify the effects of all 216 million possible single amino acid sequence substitutions in 19,233 canonical human proteins. It is based on AlphaFold2 and works by exploiting its ability to model protein structure and learn evolutionary constraints from related sequences. It provides an indication of which mutations are most likely to underlie human diseases, such as rare genetic disorders or developmental disorders. To do this, it classifies missense mutations as likely pathogenic, likely benign, or uncertain and produces a score that estimates the likelihood of a variant being pathogenic. This tool helps highlight the function of crucial regions of the protein and, when combined with other information, helps understand which mutations could cause a certain disease. However, it does not predict changes in protein structure or stability. For each point, the graph shows, in addition to the score, the name of the variant.

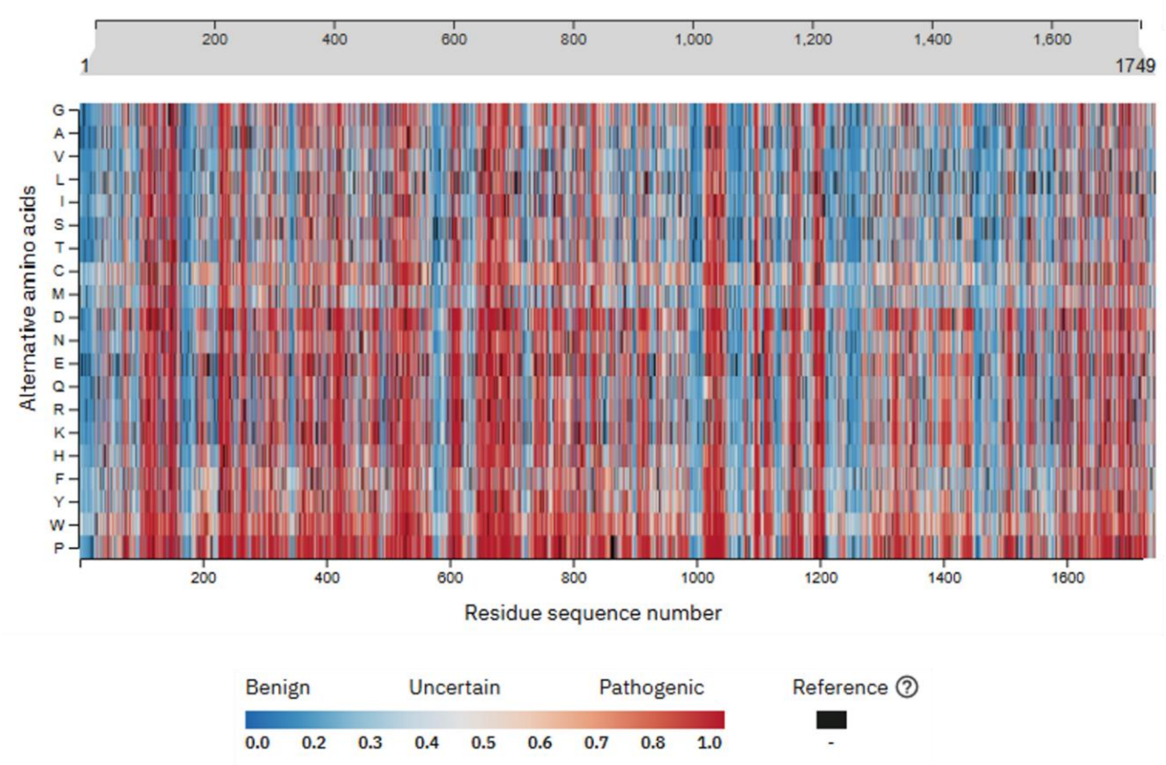


Figure 3.7: *AlphaMissense Pathogenicity Heatmap of the UBR1 E3 Ligase model generated by AlphaFold2 [64]. On the x-axis all the residues of the protein are represented (1749) while on the y-axis the 20 possible alternatives amino acids are visible. Each point (x,y) on the graph is associated with a score from 0 to 1. Points closest to zero are colored blue and are associated with benign variants. As the value approaches 1, the colors become warmer, indicating that the substitution leads to pathogenicity. The black lines (reference) indicate the original amino acid at each position in the protein.*

The graph in Figure 3.7 shows that not all regions of the UBR1 E3 ligase protein tolerate mutations equally. Many residues have been assigned warm colors. These are likely key regions for the protein's structural or catalytic function: even small variations can cause pathogenicity. An example is the UBR-box, located between residues 97 and 168, which is noted to be particularly critical regardless of the amino acid substituted; in fact, the corresponding columns are completely red. It is a highly conserved domain, essential for recognition by UBR1. However, there are also regions highlighted with cooler colors that appear less conserved and more flexible, where mutations are more likely benign.

Given all the results obtained, but especially the high pLDDT value for most of the protein residues, the model was considered valid for further analysis.

3.1.3 Comparison Between the Obtained Models

To verify the similarity between the two structures, the homology model of UBR-box of UBR1 E3 Ligase obtained with Swiss-MODEL was compared with that predicted by AlphaFold2, using the PyMOL software. The alignment result is shown in Figure 3.8.

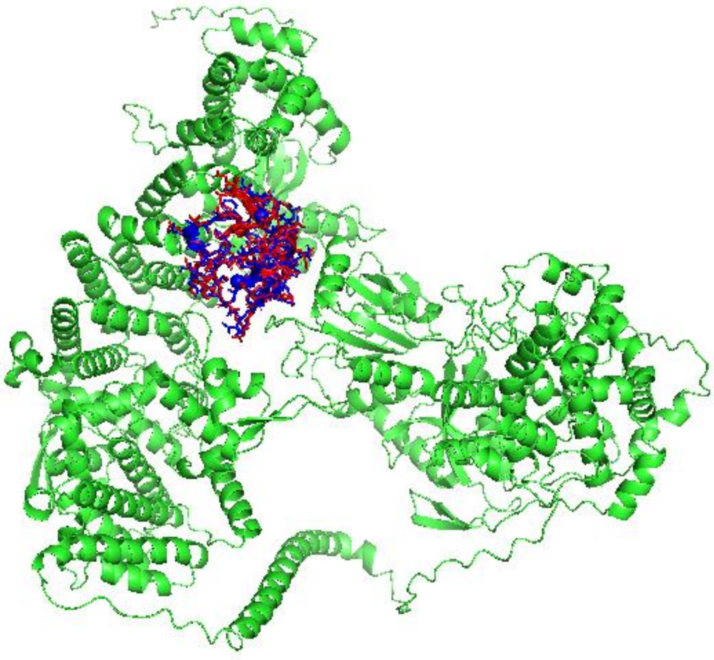


Figure 3.8: *Structural comparison between the homology model of UBR1 based on the 3NY3 template structure, shown in blue, and the model predicted by AlphaFold2. The red region represents the UBR-box of the latter, while the green domains belong to the remaining part of the protein. Image created using PyMOL software.*

In the first alignment phase, the software reads the score matrix to compare equivalent residues between the two sequences. The result was the comparison of 72 residues belonging to both structures, obtaining a score of 413,000, which represents the raw alignment score. This is very high, indicating a good match.

During the atomic superposition phase, 529 matching atoms were found between the two models. Subsequently, the RMSD was calculated through refinement interactions. By default, the software performs 5 cycles during which it progressively discards some outliers that cause this parameter to increase excessively. This eliminates some atoms with excessive deviations, improving the alignment of the remaining structures. At this stage, it was decided to leave the cycle parameter equal to 5 to provide a more representative measure of conserved regions, discarding those related to loops, more flexible regions, or non-conserved regions. The result was the progressive elimination of a total of 102 atoms, resulting in a final alignment of 427 equivalent atoms.

The calculated RMSD value was 0.491 Å, indicating that the two structures are very similar. A value less than 2 Å is considered acceptable. The formula for calculating the RMSD is as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\| r_i^{(A)} - r_i^{(B)} \right\|^2}$$

where N is the number of atoms used and $r_i^{(A)}$ and $r_i^{(B)}$ are the coordinates of the two corresponding atoms.

This approach therefore demonstrated that homology modeling produced a highly reliable structure for the UBR-box domain, which represents the structural core of the protein. The geometry of the α-helices and β-sheets was correct. It should be noted, however, that it is not as reliable as an experimental measurement since these are predictions. For molecular docking studies, these outcomes are of fundamental importance as correct modeling leads to more consistent ligand-protein interactions. The model was therefore considered sufficiently robust to be used in subsequent simulations.

For the sake of completeness, we also report the global RMSD value calculated by setting the cycle parameter equal to zero. In this case, the software does not exclude any corresponding atoms and performs the calculation with all 529. This results in an overall similarity value for the two structures, which will be higher than the previous one, since it also considers non-conserved regions. The RMSD value obtained with this method is 1.084 Å, in line with what has just been explained.

3.2 Molecular Dynamics Simulations Outputs

Molecular dynamics simulations, as explained in subchapter 2.2, were performed both on the homology model of the UBR-box domain of UBR and on the structure of the entire UBR1 E3 Ligase predicted by AlphaFold 2. The computational times for the two processes were significantly different, as the entire protein (1749 residues) is obviously much larger and more structured than the UBR-box domain alone (69 residues). For this reason, the homology model could be simulated in 100 ns, while the entire UBR1 could be simulated in just 10 ns.

Calculating the RMSD of the backbone along the entire trajectory allowed us to assess the stability of the protein's main structures during MD simulations. If this parameter's behavior over time is erratic and the values are high, it means the protein has undergone significant structural changes.

A low and stable RMSD is therefore generally desirable, and this is precisely what is evident in Figure 3.9. For both models, the results show a nearly constant RMSD value, with an average value slightly less than 0.24 Å. However, the graph in Figure 3a, relating to the MD simulation of the homology model of the UBR-box, shows greater oscillation than that in Figure 3b, relating to the entire UBR1 protein. This means that the former structure undergoes greater movements, even though around the same equilibrium point. The UBR-box domain is therefore subject to greater oscillations than the entire protein, which is larger, more structured, and more stable. It should be noted, however, that the difference between the two graphs could also be due to the different number of steps set, resulting in different simulation times. However, this suggests that the structures should not undergo any changes in shape during the simulation.

It should also be noted that the RMSD values detected are comparable in magnitude to the value obtained by aligning the two models, before the MD simulations.

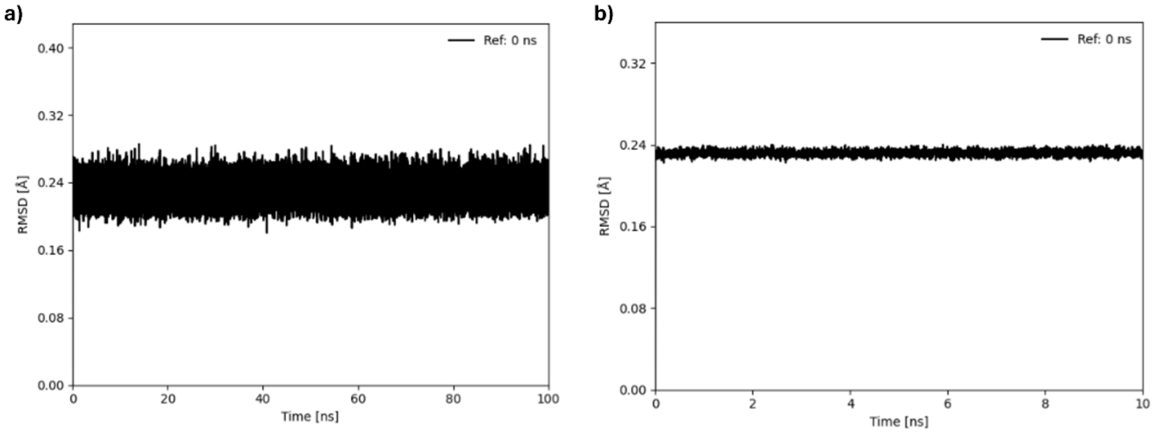


Figure 3.9: *The graphs show the RMSD value of the backbone calculated over the entire trajectory of the MD simulations. The x-axis represents the time in ns, while the y-axis represents the RMSD value in Å. Graph a) relates to the MD performed on the homology model of the UBR-box domain of UBR1, while b) concerns to the MD performed on the entire UBR1 protein predicted by AlphaFold2.*

The graphs showing the secondary structure during the simulation are reported in Figure 3.10. They display that the conformation of both models remained fairly constant. Even though the homology model represents a single domain of the entire UBR1, its residues in the model predicted by AlphaFold do not necessarily have the same conformations. They may be influenced by other residues in neighboring regions.

Some portions, of the UBR-box domain, remain in exactly the same conformation throughout the simulation. Examples include residues 25 to 30, which remain α-helices, or residues 10 to 15, which remain β-strands. Some groups of residues, however, undergo a conformational change: residues 45, 46, and 47, for example, transform from strands to loops at fairly regular intervals.

The structure of the entire UBR1 protein is more difficult to analyze due to the large number of domains and therefore residues it comprises, but it is still possible to observe that there are no major conformational changes. To compensate for this, the Ramachandran plot is also presented below. However, it is consistent with the trend in the RMSD value obtained.

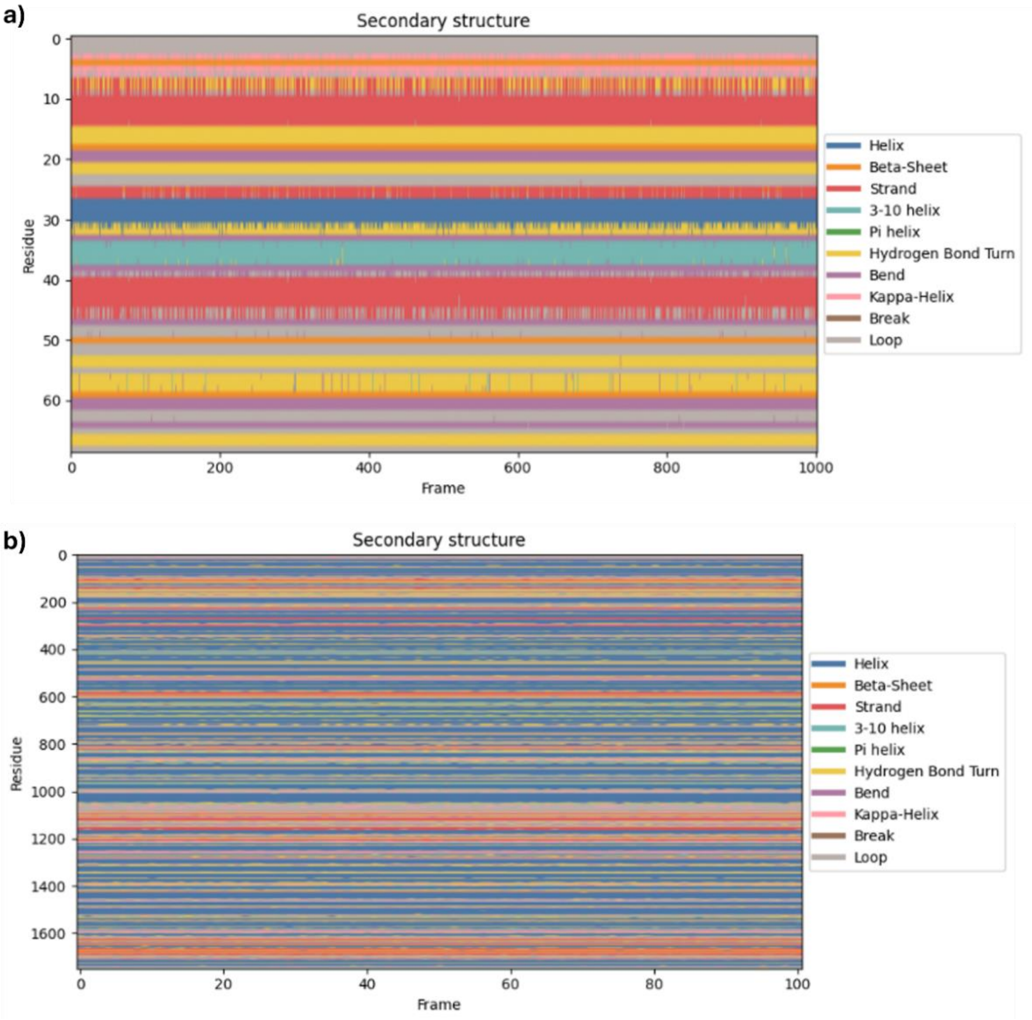


Figure 3.10: *Graphs showing the evolution of secondary structure over time. The x-axis represents the number of frames considered, while the y-axis represents the residues. The color represents the conformation assumed by the different residues over time, as can be seen in the legend on the right. Graph a) relates to the MD performed on the homology model of the UBR-box domain of UBR1, while b) concerns the MD performed on the entire UBR1 protein predicted by AlphaFold2.*

The Ramachandran plot was used to verify whether the protein residues fall within allowed regions. Graphs a) and b) in Figure 3.11 relate, respectively, to the first and last frames of the MD simulation trajectory performed on the UBR-box model. It can be noted that they are very similar to the graph shown in Figure 3.3a, as the latter also relates to the homology model, but was built with a different software.

The graphs presented here were created in this way to allow comparison of the same structure at the beginning and end of the simulation. As can be seen, they are very similar to each other, with the exception of a single residue that falls within a prohibited region, which can be considered an outlier or a residue located in a flexible loop region. However, it does not affect the overall structure integrity. Therefore, it can be concluded that the structure remains stable during the simulation and that the model does not require further verification and refinement. In fact, most residues fall within the blue regions, considered permitted. Only a few fall within light blue zones or lie on the border between these regions, which are still permitted but less frequent.

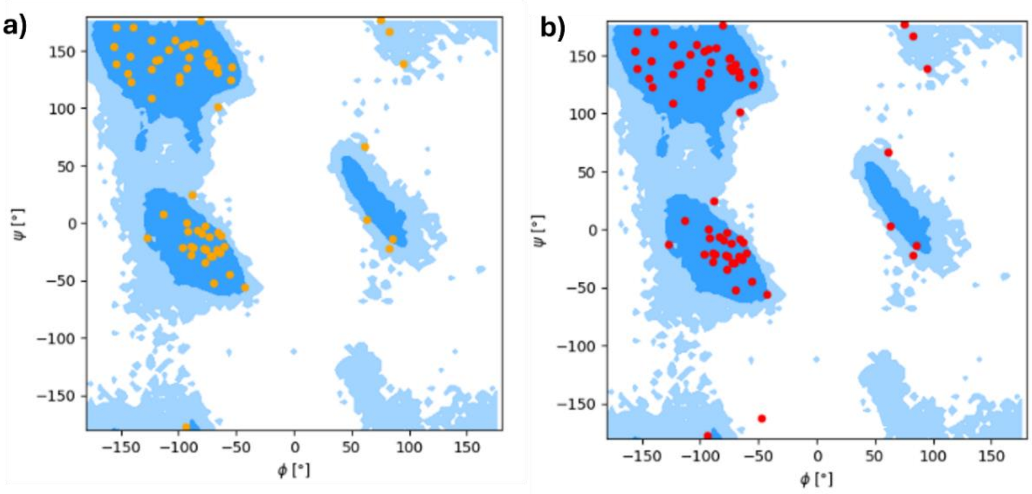


Figure 3.11: *Ramachandran plots related to the UBR-box homology model. The torsion angle φ (phi) is on the x-axis while the torsion angle ψ (psi) is on the y-axis. The blue regions indicate sterically allowed regions, the light blue areas show allowed but less frequent regions, and the white regions represent prohibited regions. Each point on the graph represents a residue. The graph a) shows the first frame of the MD simulation trajectory, while graph b) reports the last one.*

Figure 3.12 also shows the Ramachandran plot constructed over the entire MD simulation of the structure of the entire UBR1 protein to observe whether residues fell into prohibited regions during the process. It can be seen that this did not occur except for very few residues that may be part of loops. It is concluded that, even though the MD simulation was performed over 10 ns, the structure is likely correct, stable and thermodynamically favored. It could therefore be used for further, longer MD simulations and subsequent studies.

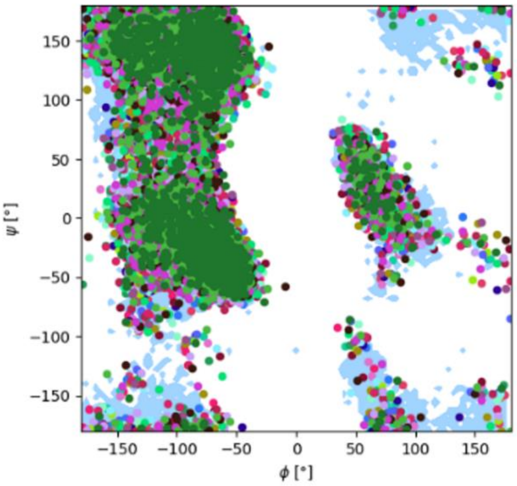


Figure 3.12: *Ramachandran plots related to the UBR1 protein predicted by AlphaFold2. It is built on the entire MD simulation trajectory. The torsion angle φ (phi) is on the x-axis while the torsion angle ψ (psi) is on the y-axis. The blue areas indicate sterically allowed regions and the white areas represent prohibited regions. Each point on the graph represents a residue.*

Finally, the energy trend over time was analyzed for both models and are illustrated in Figure 3.13. The average values found were as follows:

- For the homology model of the UBR-box domain:
  E_pot ≈ -2.06×10^5 kJ/mol; E_kin ≈ 3.87×10^4 kJ/mol; E_tot ≈ -1.68×10^5 kJ/mol.
- For the model of the entire UBR1 protein:
  E_pot ≈ -7.67×10^6 kJ/mol; E_kin ≈ 1.48×10^6 kJ/mol; E_tot ≈ -6.19×10^6 kJ/mol.

The values of the two models differ by approximately one order of magnitude, but they are not comparable because they are two systems with different dimensions and number of interactions.

Note, however, that E(t) oscillate around a constant mean value, without continuous rises or falls. The stable kinetic energy indicates that the temperature is well controlled by the thermostat. The fact that the potential energy does not drift suggests that the structure is neither collapsing nor exploding and that the integrator used (leap-frog) is adequate for the system. Thermodynamic equilibrium has therefore been reached and maintained. This is an excellent result, especially for the UBR-box model, as it provides a reliable basis for subsequent pocket detection.

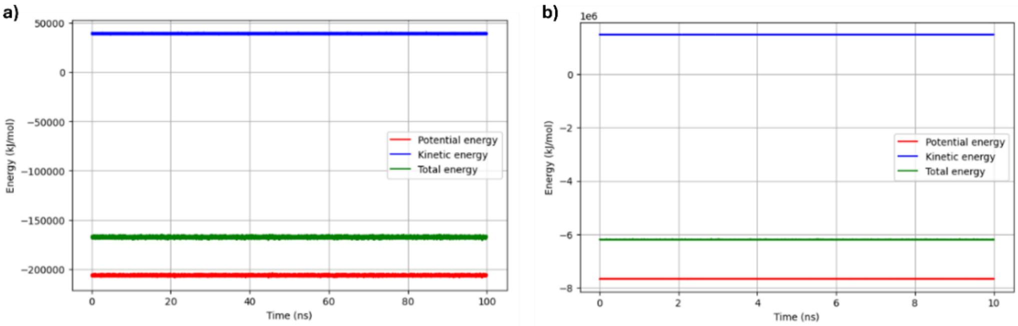


Figure 3.13: *Trends of potential (red), kinetic (blue), and total (green) energies, on the y-axis in kJ/mol, during the simulation times, on the x-axis in ns. Graph (a) refers to the simulation performed on the homology model of the UBR-box, while graph (b) refers to the entire UBR1 protein.*

3.3 Pocket identification

This section explains the rationale behind the choice of the specific pocket used to create the pharmacophore model and subsequent steps. The three potential binding sites of the UBR-box domain homology model of UBR1 E3 Ligase were compared with each other and with the pockets found in the whole protein model predicted by AlphaFold2. This was also possible thanks to the MEP displayed for both models.

3.3.1 FTSite output

The FTSite software has predicted three potential binding pockets on the homology model structure of the UBR1 E3 Ligase UBR-box, which can be seen in Figure 3.14. The objective of this phase is to identify the most promising pocket of the three.

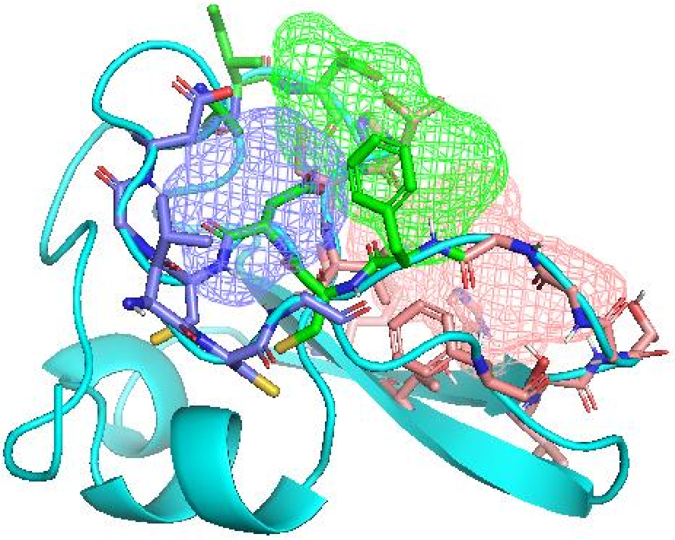


Figure 3.14: *Position of the three binding pockets on the homology model of the UBR-box domain of UBR1 E3 Ligase (in light blue): site_select_1 in pink, site_select_2 in green and site_select_3 in blue. They have been all predicted by the FTSite software and have been visualised using PyMOL.*

The residues belonging to the binding site can be found in Figure 3.15 and they are as follows.

For site_select_1 region they are:

PHE 103, THR 109, SER 111, PRO 119, THR 120, CYS 121, VAL 122, LYS 139, HIS 141, THR 142, SER 143, THR 144, GLY 145, GLY 146, GLY 147

The residues belonging to the site_select_2 region are:

PHE 103, ILE 117, ASP 118, THR 120, CYS 121, VAL 122, GLY 147, PHE 148, CYS 149, ASP 150, ALA 156

Those belonging to the site_select_3 pocket are:

LEU 98, CYS 99, GLY 100, PHE 148, CYS 149, ASP 150, CYS 151, GLY 152, ASP 153

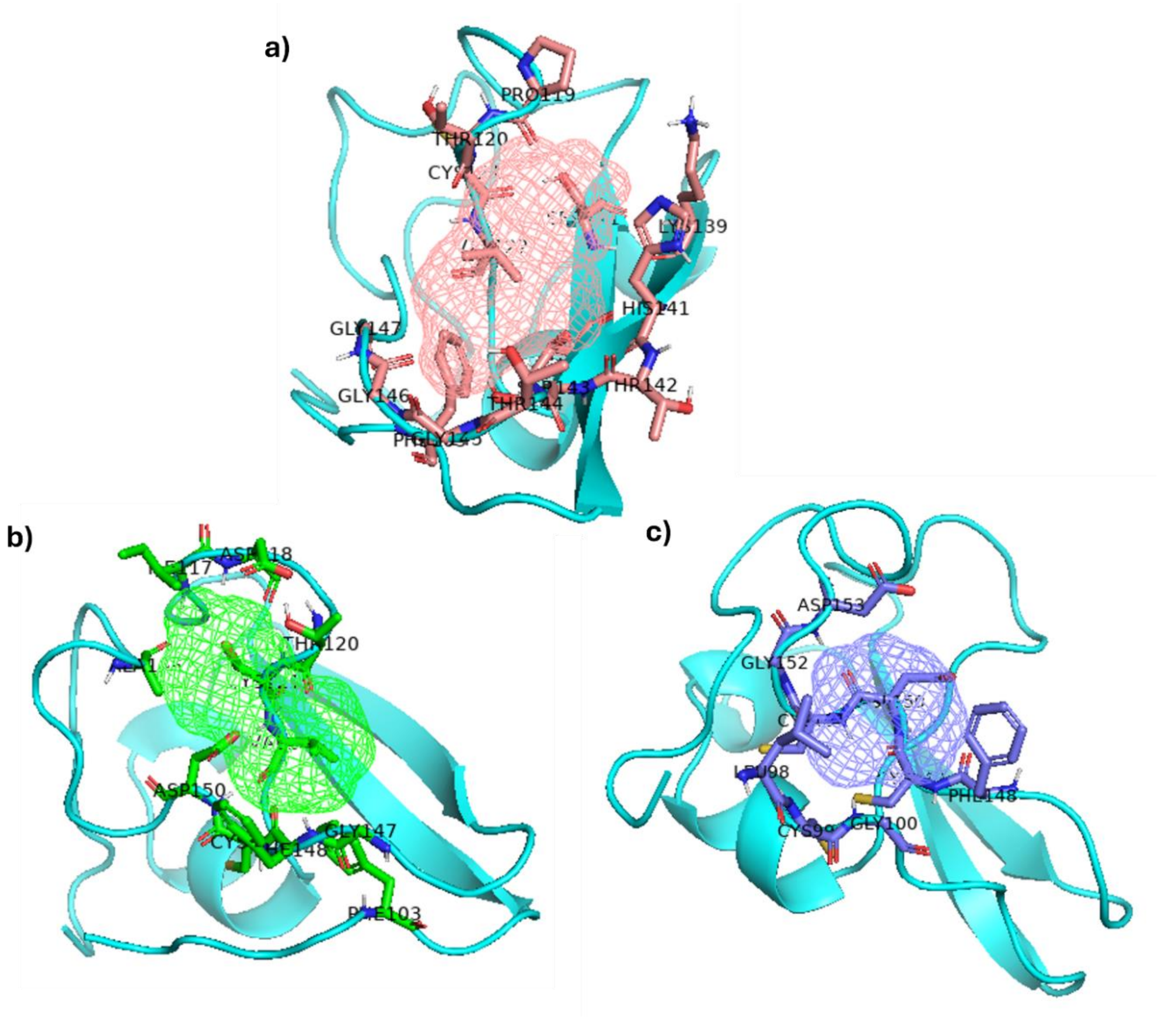


Figure 3.15: *Residues (in stick format) related to the pockets termed site_select_1 visible in pink in figure a), site_select_2 shown in green in figure b) and site_select_3 in blue in figure c). In all the images the names of the corresponding residues are labelled. Hydrogen atoms are everywhere visible in grey, oxygen atoms in red, nitrogen atoms in blue, and sulphur atoms in yellow, according to the CPK colouring convention. The coloured meshes represent the binding pockets surfaces. The UBR-box domain of UBR1 E3 Ligase is always shown in light blue.*

The software indicated these three sites, and although it did not provide a score, the numbering provided gives an indication of their priority. Therefore, the site marked with the label site_select_1 is the most likely one. Each binding site obtained is now analysed in detail.

The site_select_1 pocket is composed of both polar residues, such as Ser, Thr, His, Lys, and hydrophobic residues, i.e. Phe, Val, Pro. This suggests that it is amphipathic, ideal for accommodating molecules with both hydrophilic and hydrophobic regions.

There are consecutive glycine residues (Gly145–Gly146–Gly147) that indicate greater local flexibility compared to other portions. In fact, glycine, among the 20 amino acid residues, is the one with the lowest molecular weight and simplest side chain, consisting of a hydrogen atom. This may favour the conformational adaptation of the site to the ligand.

In addition, cysteine (Cys121) can form covalent disulphide bonds thanks to the thiol group present in the side chain. Histidine (His141), through the imidazole ring, can form a variety of specific interactions such as hydrogen bonds: it can act as both a donor and an acceptor. It can also form ionic interactions with other charged residues and bind metal ions.

This set of characteristics increases selectivity and effectiveness, creating a molecular environment that ensures that the bond occurs only with the correct substrate.

The site_select_2 pocket partially overlaps the previous one but is more nonpolar because it contains more hydrophobic residues (Phe, Ile, Val, Ala). For this reason, it is potentially more suitable for lipophilic ligands. It contains two cysteine residues (Cys121, Cys149) and two aspartic acid residues (Asp118, Asp150) that can form disulfide bonds or local polar interactions. However, these could introduce structural rigidity.

The site_select_3 pocket consists of fewer residues than the previous ones, making it smaller. It also appears to be more closed, embedded in the structure and therefore less accessible. Multiple cysteine residues are present that can form disulfide bonds or be involved in metal coordination. The two aspartic acid residues can introduce localized negative charges.

To make a more informed choice about which pocket to use for the next steps, it was decided to use FTSite to evaluate the structure of the entire UBR1 E3 Ligase protein predicted by Alphafold2. In this case too, the software returned three potential binding sites. One of these is part of the UBR-box domain, while the other two are located in distinct regions of the protein and will not be considered here. The first of them is shown with a zoom in Figure 3.16. The pocket predicted on the UBR-box domain does not exactly match any of the three sites found previously, but is located in a region adjacent to site_select_1. This information confirms that this specific protein domain can be considered reliable for use as a PROTAC target.

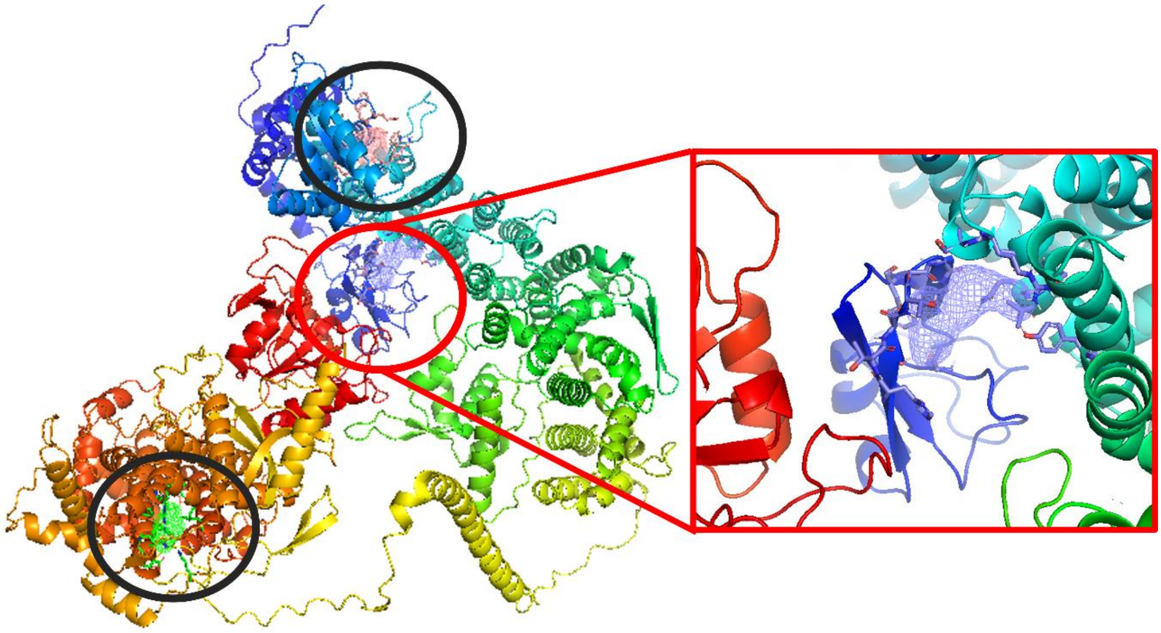


Figure 3.16: *Position of the three binding pockets on the entire UBR1 E3 Ligase model predicted by AlphaFold2. Binding pockets on various portions of the protein are circled in black, while the one found on the UBR-box domain is circled in red. A zoom of the latter is shown on the right. They have been all predicted by the FTSite software and have been visualised using PyMOL.*

Another interesting aspect was the comparison between the pockets found and the 3NY3 template structure from which the homology model was built. After aligning the structures (the template and the FTSite output), it was noted that the peptide bound to the 3NY3 protein, which had been eliminated initially, is located near the pocket named site_select_1. The alignment is visible in Figure 3.17.

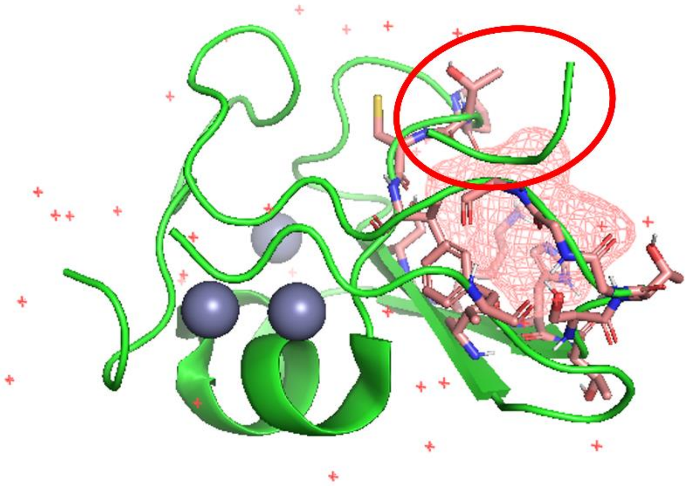


Figure 3.17: *The image represents the alignment between the 3NY3 template structure used to build the homologys model (in green) and the FTSite output. For efficiency, only the pocket named site_select_1, displayed in pink, closest to the bound peptide (circled in red), has been visualized. They have been visualised using PyMOL*

3.3.2 MEPs Visualization

First, a potential map of the UBR-box domain of UBR1 E3 Ligase was constructed, visible in Figure 3.9. The physiological condition, i.e. temperature of 310 K, pH 7 and ionic concentration 0.15 M was represented. Protein dielectric (internal) was set to 2, while solvent dielectric (external) was set to 78.

The map is color-coded.

- **Red regions** represent negative portions typically dominated by negatively charged residues, such as aspartic acid (Asp) and glutamic acid (Glu). These are favorable for binding ligands or molecules that have positively charged groups (e.g., amines).
- **Blue regions** are dominated by positively charged residues, such as arginine (Arg), lysine (Lys), and histidine (His) (depending on the pH). They can attract molecules with negatively charged groups (e.g., phosphates or carboxylates).
- **White regions** present neutral electrostatic potential or they are zones that are at the boundary between positive and negative charges. These regions are generally less involved in strong electrostatic interactions.

This is key information for analysing the map. Figure 3.18 displays the potential map across the entire UBR-box. The structure of the entire protein is visible in transparency, allowing for a better correlation between the electrostatic potential value and its regions.

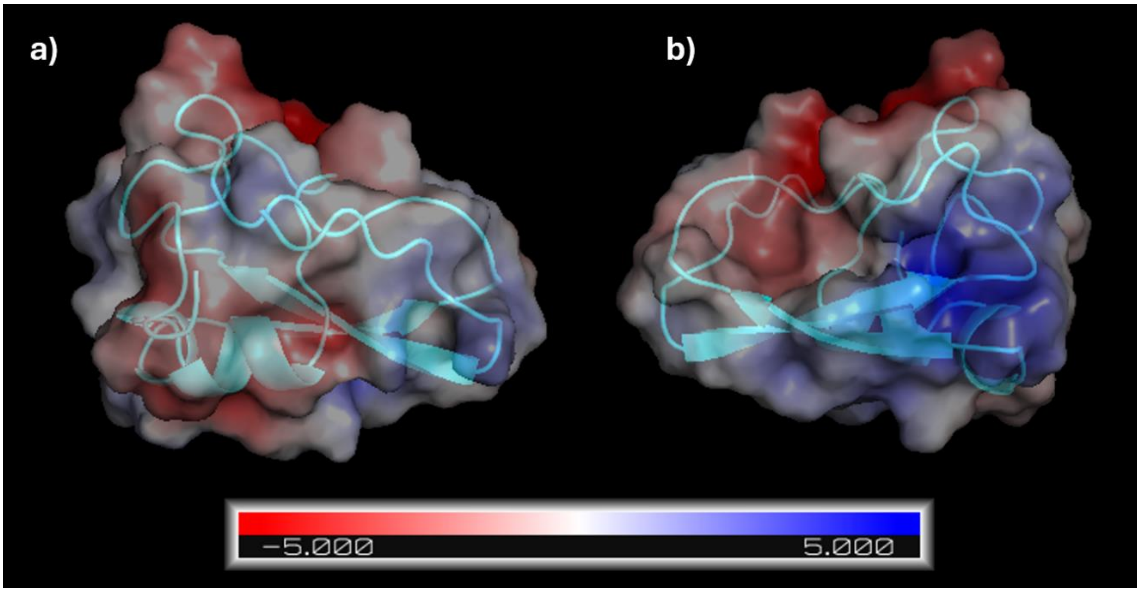


Figure 3.18: *Electrostatic potential maps of the homology model of the UBR-box domain of UBR1 E3 Ligase, constructed using the APBS Electrostatics plugin in PyMOL. Image a) shows one side of the receptor, while images b) displays it rotated by 180°. The protein structure is visualised in transparency so that the positive (blue), negative (red) and neutral (white) portions can be more easily recognised. The value is expressed in units kT/e.*

In Figure 3.19, the three different binding pockets are visible, colored according to the electrostatic potential value.

In general, it is known that:

- Acidic residues such as aspartic (Asp) and glutamic acid (Glu) contribute to negative regions.
- Basic residues such as lysine (Lys), arginine (Arg), and histidine (His), when protonated, contribute to positive regions.
- Phenylalanine (Phe), leucine (Leu), isoleucine (Ile), valine (Val), proline (Pro) and glycine (Gly) are aliphatic and hydrophobic (nonpolar) residues that make regions more neutral.
- polar, uncharged residues, such as serine (Ser), threonine (Thr), asparagine (Asn), glutamine (Gln) and cysteine (Cys), can create neutral regions, but with small variations [91].

With this information, it is possible to correlate the MEP value to the residues that constitute the three different binding sites. The pocket named site_select_1 (Figure 3.19a) is predominantly white with some red shades. This indicates that the pocket is mostly neutral or at most slightly negative. This occurs because residues like Asp or Glu, which are strong acids that do not produce a dominant red color, are absent. Lys and His residues theoretically contribute a positive charge to the region, but their effect is not significant, and therefore their effect could be overshadowed or masked by the others. Some uncharged polar residues like Thr and Ser contain oxygens that contribute to regions of slightly negative potential around the oxygens, although not to the same extent as Asp and Glu residues. Residues like

Phe, Val, and Gly tend to form amphipathic/neutral regions, and it is precisely this characteristic that prevails.

On the contrary, the pocket labeled site_select_2 (Figure 3.19b) is a deeper red, making it quite negative. It contains two well-identified, strong acidic residues, Asp118 and Asp150. There are also nonpolar residues (Phe, Ile, Val, Ala), but they do not mitigate the charge. The same is true for Cys and Thr, which are neutral but do not neutralize the strong negativity of the Asps.

The pocket site_select_3 (Figure 3.19c) is white with some red shades, making it neutral or slightly negative. Here too, there are two Asp residues (Asp150 and Asp153) that make negative contributions. However, the pocket is more prominent than the others and positioned toward the inside, so these residues may be partially shielded or in contact with other chains, and the net negative contribution may be lower. Furthermore, the pocket also contains many nonpolar residues such as Leu and Phe and more than one Cys, as well as many Glys, making the region compact.

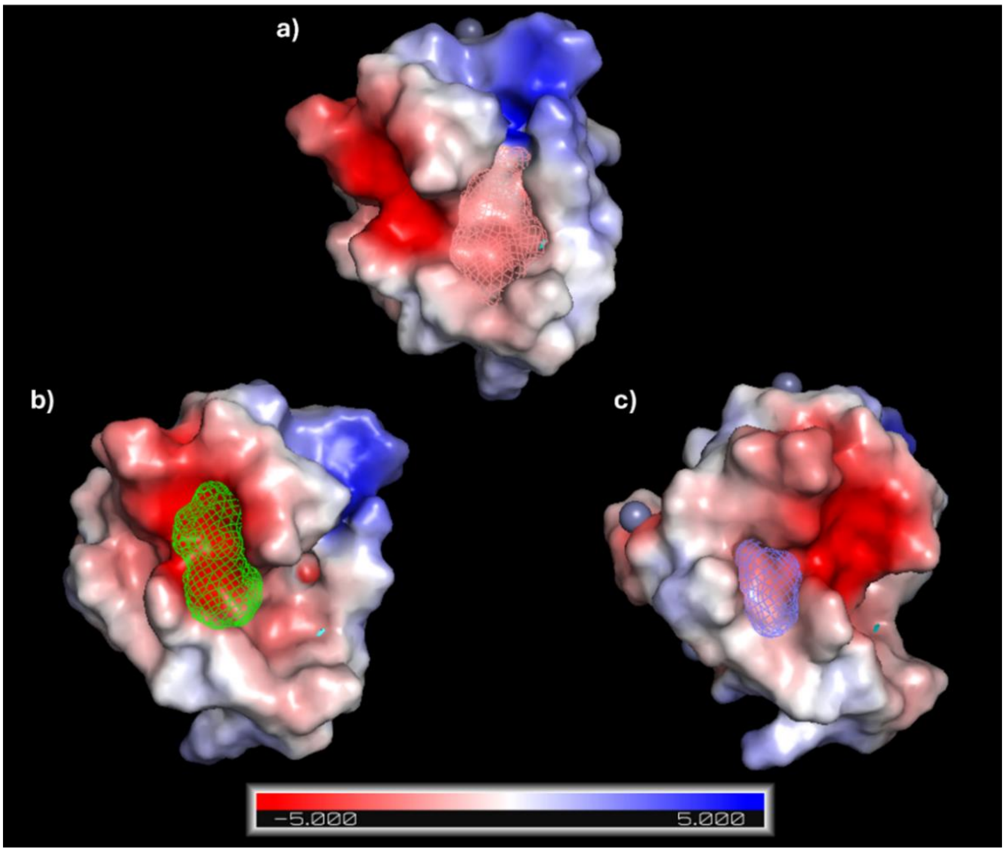


Figure 3.19: *The three images represent the electrostatic potential maps with a focus on the three binding pockets predicted by FTSite on the structure of the homology model of the UBR-box domain of UBR1 E3 Ligase, constructed using the APBS Electrostatics plugin in PyMOL The images: a) illustrate the MEP of the site_select_1 pocket, b) show the MEP of the site_select_2 pocket and c) of site_select_3. The positive regions are visualised in blue, the negative ones in red and the neutral in white. The value is expressed in units kT/e.*

Based on the information obtained, the pocket called site_select_1 was chosen. FTSite prioritized it over the other two and it has a more flexible structure. Given its amphipathic nature, it is ideal for hosting molecules with hydrophilic and hydrophobic regions and is easily accessible. It is exposed enough to allow access to small molecules and also defined to allow recognizable contact points. It therefore presents a good balance between cavity exposure and depth. Furthermore, since the peptide bound to the 3NY3 structure is very close to it, it was deemed the most suitable for accommodating the anchor of a PROTAC.

For the sake of completeness, the MEP of the structure of the entire UBR1 protein predicted by AlphaFold2 is also reported in Figure 3.20. As can be seen, alternating charged (positively or negatively) and neutral regions create a heterogeneous surface. This may be explained by the fact that the different domains of the protein perform various, even complementary, functions.

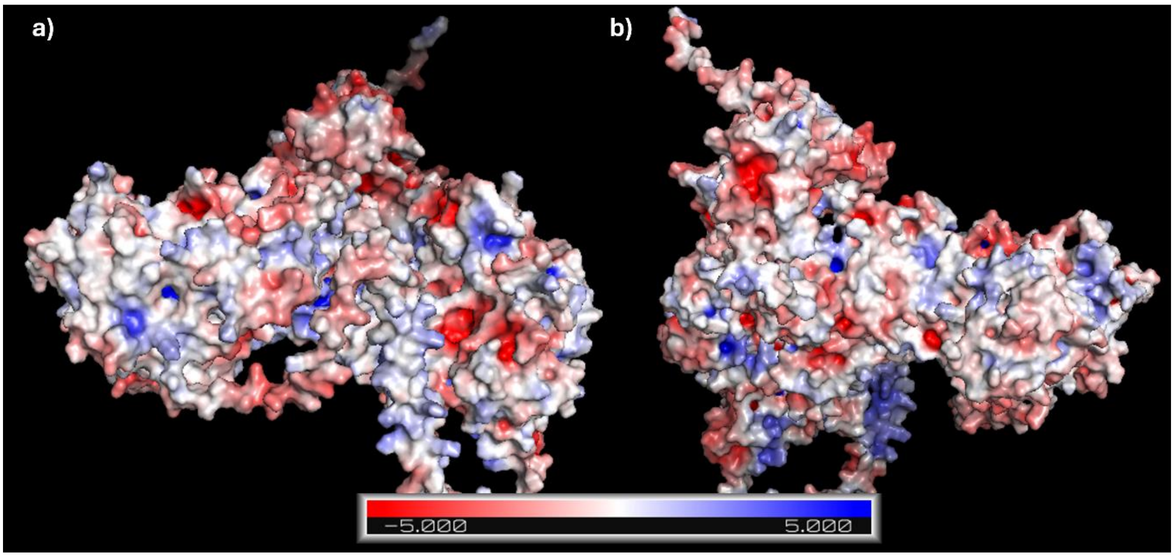


Figure 3.20: *Electrostatic potential maps of the UBR1 E3 Ligase model predicted by AlphaFold2, constructed using the APBS Electrostatics plugin in PyMOL. Image a) shows one side of the protein, while images b) displays it rotated by 180°. The positive regions are showed in blue, the negative once are displayed in red and neutrals are white. The value is expressed in units kT/e.*

3.3.3 Pharmacophore Model Based on the Selected Pocket

Based on the structure of the selected pocket, the pharmacophore model was built, visible in Figure 3.21. It reproduces the characteristics of the selected binding site (in Figure 3.15a and 3.19a) and was used as a filter in the virtual screening phase.

The features that **the ligand** must have were identified by the BioLuminate 5.8 software and they are as follows:

- A1 and A2: acceptor of hydrogen bond sites;
- D3, D4 and D5: donor of hydrogen bond sites.

The spheres surrounding the colored dots represent the spatial volume within which the ligand should position compatible functional groups. Hydrogen bond acceptors are sites on the ligand containing, for example, oxygen atoms from carboxyl groups and exhibit a partial negative charge. They can bind to portions of the receptor with amino or hydroxyl groups. In contrast, hydrogen bond donors are regions on the ligand that have a partial positive charge, containing amino or hydroxyl groups, and bind to, for example, carboxylic acid groups on the receptor's pocket.

The arrows indicate the optimal geometric orientation of the hydrogen bond, calculated based on the arrangement of the residues in the pocket to maximize the interaction.

The pharmacophore model created has more donor sites than acceptors, three and two respectively. This means that the chosen pocket has more acceptor sites than donors. It is, in fact, predominantly neutral but with portions tending toward red (negative). This is consistent with its amino acid composition, which includes polar and hydroxylated residues such as Ser111, Thr109, Thr120, Thr142, Ser143, and Thr144, which can act as both hydrogen bond donors and acceptors. However, the ligand must also have acceptor sites that will interface with partially positive regions of the pocket.

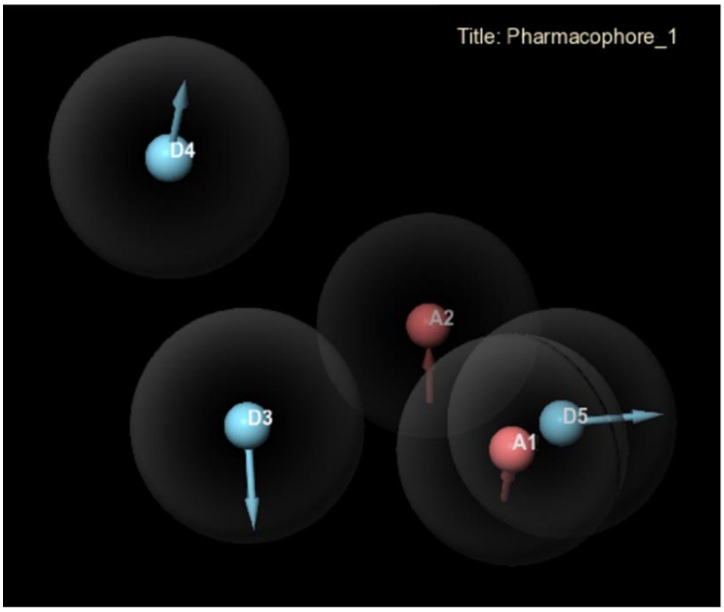


Figure 3.21: *Pharmacophore model based on the structure of the site_select_1 binding pocket, created using BioLuminate 5.8 software (developed by Schrödinger). The letter A indicates a site whose characteristic is to accept hydrogen bonds (in red) while the letter D indicates a hydrogen bond donor region (in blue). These are all ligand features.*

In addition to the pharmacophore based on the pocket actually chosen for the continuation of the project, one was also constructed relating to the siteselect_2 pocket, represented in green in Figures 3.15b and 3.19b. It has been reported here, in Figure 3.22, as it was considered a good starting point for further investigation in future projects.

This pharmacophore model shows different characteristics that the ligand must have, compared to the previous one:

- D1 and D2: donor of hydrogen bond sites;
- H3 and H4: hydrophobic regions;
- R5: aromatic ring.

This information reveals that the ligand must have hydrogen bond donor sites, i.e. hydroxyl and amino groups, and that therefore the corresponding portions of the pocket must be acceptors. This corresponds to the pocket being a region of predominantly negative potential, as seen on the MEP.

Features H3 and H4 represent three-dimensional regions of the pocket where the receptor prefers non-polar interactions, thus exhibiting hydrophobic characteristics. The fact that there are two suggests that site_select_2 has an articulated hydrophobic architecture and that the ideal ligand should be branched or have distinct hydrophobic subunits.

R5 indicates that, at that point, the site favours the presence of a planar aromatic ring in the ligand.

Due to this set of characteristics, the site_select_2 pocket is more rigid, geometrically restrictive and more negatively charged. Therefore, it has been considered ideal for constrained ligands but less suitable for the anchor of a PROTAC.

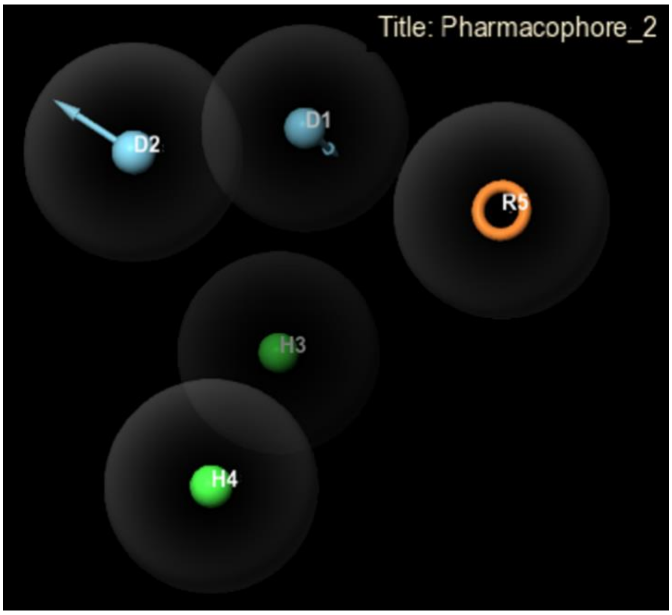


Figure 3.22: *Pharmacophore model based on the structure of the site_select_2 binding pocket, created using BioLuminate 5.8 software (developed by Schrödinger). Letters D indicate hydrogen bond donor regions (in blue), H letters refer to hydrophobic regions (in green) and R5 says that the ligand must have an aromatic ring (in orange). These are all ligand features.*

3.4 First phase of identifying potential ligands

This subchapter presents the results of the virtual screening phase and the evaluation of the properties of the compounds that were then considered, having a higher Phase Screen Score. These data supported the initial selection of ten ligands potentially suitable for application as anchors in a PROTAC molecule.

3.4.1 Virtual Screening Hits

The actual virtual screening phase was preceded by the preparation of the ligands belonging to the NCI Diversity Set, as explained in subchapters 2.4.1 and 2.4.2.

This step, in which 1,973 compounds were processed, performed using the tool Phase Ligand Screening of Maestro 14.3 software, required a computational time of about a day and a half, longer than the screening. It should be noted that one compound (NSC 84460) was discarded during the preparation phase because it contained an arsenic atom which was difficult for the software to process.

A total of 6082 structures were obtained due to the setting of 50 conformers in the process preparation phase. Figure 3.23 shows the list of the top 100, having the highest Phase Screen Score and which were subsequently analyzed.

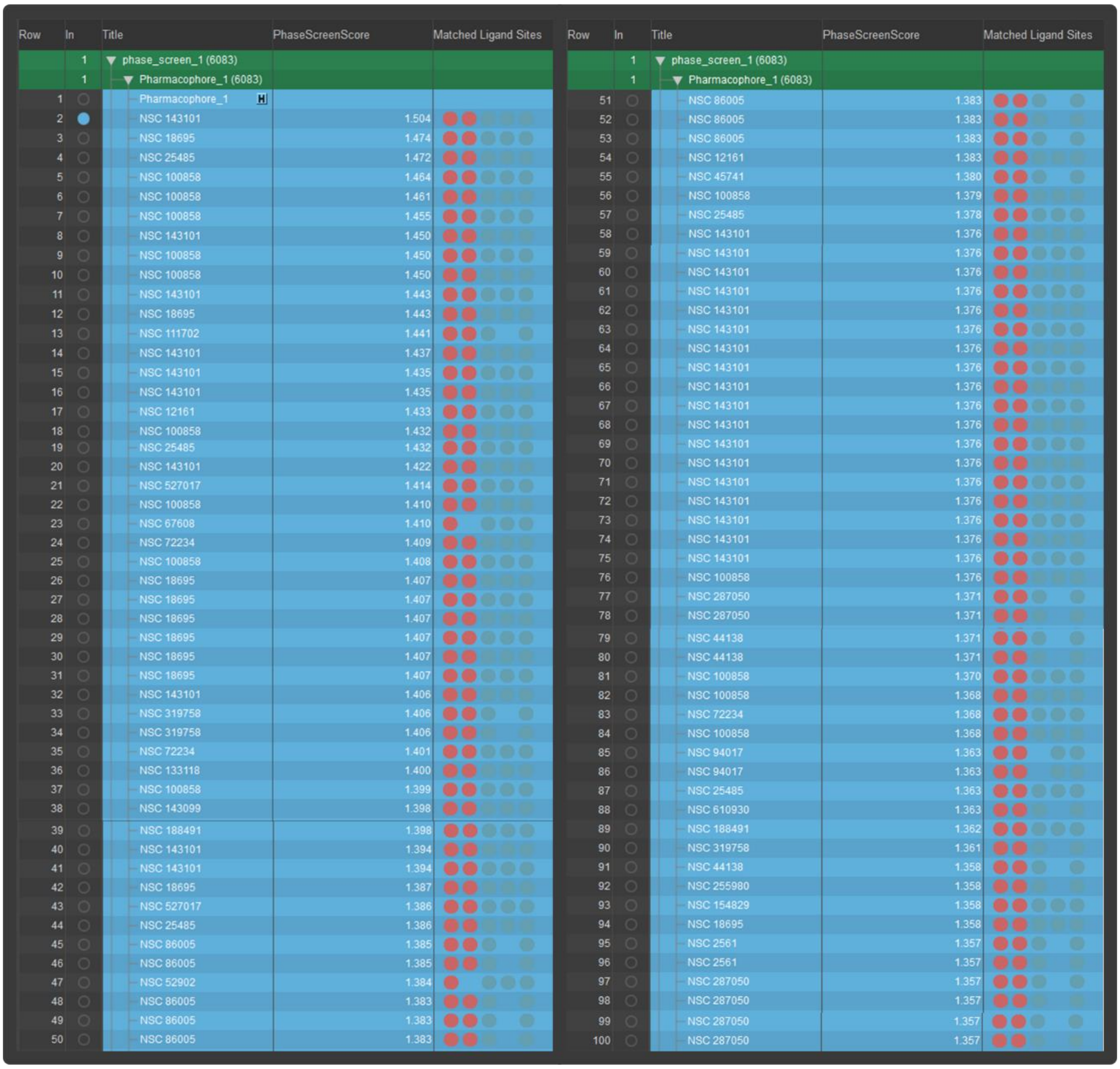
The score doesn't have a fixed theoretical range, as it depends on the number of features and the quality of the match. It's not comparable between different models, but only within the same screening. Typically, values between 1.5 and 3.5 indicate good matches, while values above 4 indicate perfect or near-perfect matches (very rare). Very low values, such as less than 1.0, indicate weak matches. In the case of the project, the obtained top hit structure has a score of 1.504 while the hundredth hit has 1.357. This indicates not excellent, but still decent matches.

Other interesting data to consider can be found in the ‘Matched Ligand Sites’ column, where, for each ligand obtained, the number of characteristics identified by the pharmacophore model that are satisfied by the structure (number of coloured dots) is displayed. It is also possible to see which of these properties are satisfied by the ligand: this is indicated by the colour and position of the dot.

- The red dots represent the characteristics of hydrogen bond acceptors (A1 and A2), which are considered the most important.
- The semi-transparent dots depict the properties of hydrogen bond donors (D3, D4 and D5).
- The empty spaces mean that the given ligand does not satisfy the specific characteristic.

It should be noted that a ligand can thus both match fewer sites and have a higher Phase Screen Score than another ligand in the screening set [92].

Different conformers of the same ligands may have obtained different scores but the number and type of features they satisfy are the same.



Row	In	Title	PhaseScreenScore	Matched Ligand Sites
	1	▼ phase_screen_1 (6083)		
	1	▼ Pharmacophore_1 (6083)		
1	○	Pharmacophore_1 H		
2	●	NSC 143101	1.504	🔴🔴⚪⚪⚪
3	○	NSC 18695	1.474	🔴🔴⚪⚪⚪
4	○	NSC 25485	1.472	🔴🔴⚪⚪⚪
5	○	NSC 100858	1.464	🔴🔴⚪⚪⚪
6	○	NSC 100858	1.461	🔴🔴⚪⚪⚪
7	○	NSC 100858	1.455	🔴🔴⚪⚪⚪
8	○	NSC 143101	1.450	🔴🔴⚪⚪⚪
9	○	NSC 100858	1.450	🔴🔴⚪⚪⚪
10	○	NSC 100858	1.450	🔴🔴⚪⚪⚪
11	○	NSC 143101	1.443	🔴🔴⚪⚪⚪
12	○	NSC 18695	1.443	🔴🔴⚪⚪⚪
13	○	NSC 111702	1.441	🔴🔴⚪ ⚪
14	○	NSC 143101	1.437	🔴🔴⚪⚪⚪
15	○	NSC 143101	1.435	🔴🔴⚪⚪⚪
16	○	NSC 143101	1.435	🔴🔴⚪⚪⚪
17	○	NSC 12161	1.433	🔴🔴⚪⚪⚪
18	○	NSC 100858	1.432	🔴🔴⚪⚪⚪
19	○	NSC 25485	1.432	🔴🔴⚪⚪⚪
20	○	NSC 143101	1.422	🔴🔴⚪⚪⚪
21	○	NSC 527017	1.414	🔴🔴⚪⚪⚪
22	○	NSC 100858	1.410	🔴🔴⚪⚪⚪
23	○	NSC 67608	1.410	🔴 ⚪⚪⚪
24	○	NSC 72234	1.409	🔴🔴⚪⚪⚪
25	○	NSC 100858	1.408	🔴🔴⚪⚪⚪
26	○	NSC 18695	1.407	🔴🔴⚪⚪⚪
27	○	NSC 18695	1.407	🔴🔴⚪⚪⚪
28	○	NSC 18695	1.407	🔴🔴⚪⚪⚪
29	○	NSC 18695	1.407	🔴🔴⚪⚪⚪
30	○	NSC 18695	1.407	🔴🔴⚪⚪⚪
31	○	NSC 18695	1.407	🔴🔴⚪⚪⚪
32	○	NSC 143101	1.406	🔴🔴⚪⚪⚪
33	○	NSC 319758	1.406	🔴🔴⚪ ⚪
34	○	NSC 319758	1.406	🔴🔴⚪ ⚪
35	○	NSC 72234	1.401	🔴🔴⚪⚪⚪
36	○	NSC 133118	1.400	🔴🔴⚪⚪⚪
37	○	NSC 100858	1.399	🔴🔴⚪⚪⚪
38	○	NSC 143099	1.398	🔴🔴⚪⚪⚪
39	○	NSC 188491	1.398	🔴🔴⚪⚪⚪
40	○	NSC 143101	1.394	🔴🔴⚪⚪⚪
41	○	NSC 143101	1.394	🔴🔴⚪⚪⚪
42	○	NSC 18695	1.387	🔴🔴⚪⚪⚪
43	○	NSC 527017	1.386	🔴🔴⚪⚪⚪
44	○	NSC 25485	1.386	🔴🔴⚪⚪⚪
45	○	NSC 86005	1.385	🔴🔴⚪ ⚪
46	○	NSC 86005	1.385	🔴🔴⚪ ⚪
47	○	NSC 52902	1.384	🔴 ⚪⚪⚪
48	○	NSC 86005	1.383	🔴🔴⚪ ⚪
49	○	NSC 86005	1.383	🔴🔴⚪ ⚪
50	○	NSC 86005	1.383	🔴🔴⚪ ⚪
51	○	NSC 86005	1.383	🔴🔴⚪ ⚪
52	○	NSC 86005	1.383	🔴🔴⚪ ⚪
53	○	NSC 86005	1.383	🔴🔴⚪ ⚪
54	○	NSC 12161	1.383	🔴🔴⚪⚪⚪
55	○	NSC 45741	1.380	🔴🔴⚪ ⚪
56	○	NSC 100858	1.379	🔴🔴⚪⚪⚪
57	○	NSC 25485	1.378	🔴🔴⚪⚪⚪
58	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
59	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
60	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
61	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
62	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
63	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
64	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
65	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
66	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
67	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
68	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
69	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
70	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
71	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
72	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
73	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
74	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
75	○	NSC 143101	1.376	🔴🔴⚪⚪⚪
76	○	NSC 100858	1.376	🔴🔴⚪⚪⚪
77	○	NSC 287050	1.371	🔴🔴⚪ ⚪
78	○	NSC 287050	1.371	🔴🔴⚪ ⚪
79	○	NSC 44138	1.371	🔴🔴⚪ ⚪
80	○	NSC 44138	1.371	🔴🔴⚪ ⚪
81	○	NSC 100858	1.370	🔴🔴⚪⚪⚪
82	○	NSC 100858	1.368	🔴🔴⚪⚪⚪
83	○	NSC 72234	1.368	🔴🔴⚪⚪⚪
84	○	NSC 100858	1.368	🔴🔴⚪⚪⚪
85	○	NSC 94017	1.363	🔴🔴 ⚪⚪
86	○	NSC 94017	1.363	🔴🔴 ⚪⚪
87	○	NSC 25485	1.363	🔴🔴⚪⚪⚪
88	○	NSC 610930	1.363	🔴🔴⚪ ⚪
89	○	NSC 188491	1.362	🔴🔴⚪⚪⚪
90	○	NSC 319758	1.361	🔴🔴⚪ ⚪
91	○	NSC 44138	1.358	🔴🔴⚪ ⚪
92	○	NSC 255980	1.358	🔴🔴⚪ ⚪
93	○	NSC 154829	1.358	🔴🔴⚪⚪⚪
94	○	NSC 18695	1.358	🔴🔴⚪⚪⚪
95	○	NSC 2561	1.357	🔴🔴⚪ ⚪
96	○	NSC 2561	1.357	🔴🔴⚪ ⚪
97	○	NSC 287050	1.357	🔴🔴⚪ ⚪
98	○	NSC 287050	1.357	🔴🔴⚪ ⚪
99	○	NSC 287050	1.357	🔴🔴⚪ ⚪
100	○	NSC 287050	1.357	🔴🔴⚪ ⚪

Figure 3.23: *List of the top 100 hits obtained at the end of the virtual screening performed with the Maestro 14.3 Phase Ligand Screening tool, developed by Schrödinger. Ligands are positioned on the rows. Compounds are listed in order of Phase Screen Score, from the highest to the lowest. The first column (called "Row") displays a numbering in ascending order indicating the decrease in this index. The "Title" column contains the compound identifiers and the pharmacophore hypothesis used in the process is highlighted in green. The red dots in the "Marched Ligand Site" column indicate how many mandatory features in the pharmacophore model and which were correctly satisfied.*

Based on the data obtained, the following observations regarding the top 100 results with the highest Phase Screen Score can be reported. All ligands satisfy at least four of the five characteristics, and it can be noted that characteristics A1, D3 and D5 are satisfied for all of them. In particular, all compounds (except NSC 67608 and NSC 52902) fulfil characteristics A1 and A2: they possess hydrogen bond acceptor sites. On the other hand, 26 structures do not correspond to the hydrogen bond donor property D4.

Overall, it can be concluded that the virtual screening process performed on compounds belonging to the NCI Diversity Set produced satisfactory results.

Although the Phase Screen Score did not produce very high values indicating an excellent match, no compound among the top 100 results obtained a score lower than 1, and in reality, lower than 1.357. Furthermore, the fact that the characteristics defined by the pharmacophore model are satisfied for most of the ligands indicates that there could be a good match between the ligands (especially those with higher scores) and the receptor, i.e. UBR1 E3 Ligase, precisely in the pocket belonging to the UBR-box domain defined site_select_1.

3.4.2 Features of Top Hit Compounds

Starting from the output compounds of the ligand screening, the first 100 compounds with the highest Phase Screen Score were evaluated. Given the high number of structures found, it was decided to perform an analysis of the essential properties in order to keep only the most promising ligands and discard the others.

First, the number of conformers of each individual compound was counted to identify the ligand "families", which resulted in 23. Consequently, an initial evaluation of the properties was conducted on 23 compounds, where the conformer with the highest score was taken for each "family". This approach made it possible to reduce the number of structures to be processed. The same compounds with different spatial conformations will have properties with different values, but they are expected to differ less than compounds belonging to different "families". In this phase, priority was therefore given to the analysis of the "families".

Table 3.4 contains data relating to the properties evaluated using Chemaxon's Playground Calculator tool for each of the 23 structures. It also includes data relating to the Phase Screen Score and the number of conformers for each compound found in the top 100 hits.

In addition to these, other data were also obtained using the software, such as the nominal mass, composition, mass spectrum, formula, IUPAC and traditional namesand the graphs representing logD and logS as a function of pH. However, they are not reported below as they were deemed not to provide added value to the analysis at this stage. For example, logD and logS were evaluated using the point value at physiological pH, a suitable condition for the project.

ID	Phase Screen Score	Number of conformers	Microspecies at pH 7.4	Molecular Weight [Da]	Intrinsic Solubility Log(S)	Log(S) at pH 7,4	Log(P)	Log(D) at pH 7.4	hERG (Activity Model) pIC_{50} in µM	hERG (Class Probability)	Prediction Class
NSC 143101	1,5	27	92.6% neutral 2.4% + 1.7% charge -1	610,5	-2,68	--	3,34	3,31	5,16	SAFE 78% TOXIC 22%	SAFE
NSC 18695	1,47	10	100% charge +2	262,3	-0,77	1,23	-5,04	-9,49	3,95	SAFE 63% TOXIC 37%	SAFE
NSC 25485	1,47	5	100% charge -1	584,7	-0,81	1,19	-2,78	-5,13	4,74	SAFE 89% TOXIC 11%	SAFE
NSC 100858	1,46	14	64.9 % neutral 34.7% charge -1 0,3 % charge +1	379,4	-0,9	--	-6,3	-6,43	4,3	SAFE 73% TOXIC 27%	SAFE
NSC 111702	1,44	1	99,9% neutral	314,3	-2,45	-2,45	-0,15	-0,15	4,02	SAFE 96% TOXIC 4%	SAFE
NSC 12161	1,43	2	92,9 neutral 7,1% charge -1	283,2	-1,58	-1,54	-2,84	-2,87	3,83	SAFE 91% TOXIC 9%	SAFE
NSC 527017	1,41	2	--	--	--	--	-2,3	-2,31	--	--	--
NSC 67608	1,41	1	97% charge +1 3% neutral	271,3	-3,1	-1,58	2,33	1,04	5,25	SAFE 58% TOXIC 42%	SAFE
NSC 72234	1,41	3	98,8% charge -1 0,7% charge -2 0,4% charge -2	422,4	-1,22	0,78	0,02	-3,46	4,65	SAFE 85% TOXIC 15%	SAFE
NSC 319758	1,41	3	99,5% charge -1 0,5% neutral	268,3	-0,68	1,32	-2,17	-4,37	3,95	SAFE 70% TOXIC 30%	SAFE
NSC 133118	1,40	1	100% neutral	268,2	-0,68	-0,68	-2,13	-2,13	3,88	SAFE 87% TOXIC 13%	SAFE
NSC 143099	1,40	1	95,6% neutral	577,5	-2,63	--	3,12	3,10	5,26	SAFE 84% TOXIC 16%	SAFE
NSC 188491	1,40	2	86,1% neutral 13,8% charge +1	339,3	-2,13	-2,06	-2,30	-2,36	4,26	SAFE 72% TOXIC 28%	SAFE
NSC 86005	1,39	8	91% neutral 6,8% + 1,6% charge -1 0,5% charge +1 0,1% charge -2	585,6	-2,90	--	-4,57	-4,59	4,50	SAFE 87% TOXIC 13%	SAFE
NSC 52902	1,38	1	100% neutral	180,2	0,66	0,66	-3,88	-3,88	3,75	SAFE 93% TOXIC 7%	SAFE

ID	Phase Screen Score	Number of conformers	Microspecies at pH 7,4	Molecular Weight [Da]	Intrinsic Solubility Log(S)	Log(S) at pH 7,4	Log(P)	Log(D) at pH 7.4	hERG (Activity Model) pIC_{50} in µM	hERG (Class Probability)	Prediction Class
NSC 45741	1,38	1	100% charge -1	298,3	-1,47	0,53	-0,63	-3,94	4,23	SAFE 95% TOXIC 5%	SAFE
NSC 287050	1,37	6	100% neutral	164,2	-0,04	-0,04	-1,89	-1,89	3,61	SAFE 94% TOXIC 6%	SAFE
NSC 44138	1,37	4	93,8% charge -2 6,2% charge -1	258,1	1,46	3,46	-3,06	-6,61	4,07	SAFE 81% TOXIC 19%	SAFE
NSC 94017	1,36	2	99,9 charge - 1 0,1% neutral	328,2	-1,94	0,06	-3,77	-3,69	4,00	SAFE 87% TOXIC 13%	SAFE
NSC 610930	1,36	1	98% neutral 1,7% + 0,3% charge -1	500,5	-5,57	-5,56	3,17	3,16	5,23	SAFE 97% TOXIC 3%	SAFE
NSC 255980	1,36	1	99,3% charge -1 0,7% neutral	401,4	-3,56	-1,41	-0,93	-3,06	4,29	SAFE 87% TOXIC 13%	SAFE
NSC 154829	1,36	1	98,5% neutral 1,%4 charge -1	340,3	-1,16	-1,16	-2,02	-2,03	4,09	SAFE 67% TOXIC 33%	SAFE
NSC 2561	1,36	2	100% neutral	240,3	-1,41	-1,41	0,07	0,07	3,71	SAFE 92% TOXIC 8%	SAFE

Table 3.4: *The columns show the properties found using the Chemaxon Playground Calculator, while each row lists each individual compound for which these properties were analyzed. Furthermore, data relating to the number of conformers of the same structure found among the first 100 hits and the score obtained in the virtual screening phase with Maestro 14.3 (the Phase Screen Score) were integrated. Boxes containing "--" indicates that the properties could not be obtained. The rows highlighted in green correspond to the compounds selected as potential ligands at the end of the virtual screening step and which were used in subsequent phases.*

Through the analysis of these data for all 23 structures, the 10 compounds highlighted in green in Table 3.4 were selected. The criteria used for the selection were those reported in Table 2.1 in subchapter 2.4.3. Although it was not possible to find compounds that fully satisfied all the required characteristics, every effort was made to identify those with the best overall profile.

As can be seen, all were classified as safe according to the hERG Classification Model. However, the percentage of safety versus toxicity varies depending on the compound. NSC 67608, NSC 154829, and NSC 319758 with a safety percentage less than or equal to 70% were rejected. The pIC_{50} parameter, calculated using the hERG Continuous Activity Model, never exceeds a value of 5.3 µM. Although the ideal criterion is to stay below 5 µM, i.e. an IC_{50} greater than 10 µM, the situation was not considered problematic.

This is in line with the Classification Model's predictions, i.e. no structures of particular concern from the cardiac safety perspective. However, an incomplete correlation between pIC_{50} and the percentage of toxicity predicted by the Classification Model is noted: a lower pIC_{50} value according to the Continuous Activity Model does not necessarily imply a greater percentage of safety.
Compounds considered potentially toxic have already been excluded, albeit conservatively, given that all were considered safe in the software's final classification. However, given equal values for other properties, the structures with lower pIC50 values were selected.

The Log(S) values, which represent the intrinsic capability of the compound to dissolve in water for structural reasons (hydrophobicity, ability to form H bonds), should not be less than -4. For this reason, the structures NSC 610930 (-5.57) and NSC 255980 (-3.56) were discarded. The first also showed a similar Log(S) value at physiological pH. The second met the criterion at the limit, but the value was still lower than that of the other structures.

One of the main criteria on which the screening of the compounds was based, was the presence of certain microspecies at physiological pH, whose distribution depends on the pKa. It was chosen to analyze this parameter because it shows the prevalent form at pH 7.4 which is the one that will most likely interact with the target.

It was decided to favor compounds with the highest possible percentage of neutral microspecies at physiological pH since MEP analysis of the binding pocket on the UBR-box domain of UBR1 E3 Ligase revealed an overall neutral potential, with only moderate charge contributions. In this context, highly charged ligands would be disadvantaged since electrostatic desolvation is not adequately compensated by site-specific Coulombic interactions. Charged microspecies would result in greater difficulty and rigidity in positioning the ligand within the pocket. Furthermore, the pharmacophore model showed mainly hydrogen bond donor/acceptor regions, with no requirement for permanent strongly cationic/anionic centers. Consequently, structures that exist predominantly in neutral form at physiological pH but retain polar groups capable of forming hydrogen bonds with the residues in the pocket were preferred.

For this reason, compounds NSC 18695, NSC 25485, NSC 100858, NSC 72234, NSC 45741, NSC 44138, and NSC 94017 were rejected. Structure NSC 527017 was excluded because insufficient information could be found for its evaluation.

Concerning the Log(P) and Log(D) parameters at physiological pH, an attempt was made to select compounds with values between 1 and 3 as far as possible to achieve the right balance between water solubility and lipophilicity. Given that it was decided to keep structures predominantly neutral in this condition, these parameters will be the same for those that are 100% neutral micro-species, while they will be almost similar if small percentages of charged micro-species (positively or negatively) are also present. In fact, log(P) refers only to the neutral form.

Therefore, those with a high difference between Log(D) and Log(P) were excluded a priori. The structures chosen that came closest to satisfying this criterion were NSC 143101 and NSC 143099, with values slightly higher than the desired range.

Unfortunately, the others have values below the lower limit of the range, but they were taken into consideration because they satisfied a number of other properties. On the other hand, structures such as NSC 67608 and NSC 610930, despite having Log(D) and Log (P) parameter values within the correct range, were not included among the candidate compounds.
In fact, both had high pIC50 values, the former at physiological pH was predominantly in the -1 charged form and the latter had too low a Log(S) value.

The Lipinski rule of five on molecular weight (less than 500 Da) could not be respected for all compounds. Due to the fact that some structures with higher weights showed satisfactory characteristics, they were nevertheless selected for the subsequent steps. An example is NSC 143101, with a molecular weight of 610.5 Da. However, it was not excluded as it obtained the highest Phase Screen Score of all (1.5) and it is also the one with the highest number of conformers in the first 100 hits of the virtual screening (27), meaning that its other spatial configurations are also promising. Furthermore, at physiological pH it presents a percentage of neutral microspecies greater than 92%.

The same consideration was also applied to structures NSC 143099 and NSC 86005, which do not completely satisfy the molecular weight criterion (and obtained lower scores than the previous one in virtual screening) but were taken into consideration for the subsequent steps as they present a high percentage of neutral micro-species at physiological pH and good values for the other parameters.

This results in a relaxation of Lipinski's rule. However, an effort was made to favour the lighter compounds.

By applying these considerations, it was possible to select 10 compounds from the 23 initially evaluated. These were considered potential candidates for use as anchors for a PROTAC that binds to UBR1 E3 Ligase in the specific site_select_1 pocket located in the UBR-box domain. They are highlighted in green in Table 3.4 and marked with a green tick in Table 2.2.

3.5 Identification of Final Candidates

Based on the results obtained in the virtual screening phase and in the selection process of potential ligands through evaluation of the pharmacokinetic properties and hERG safety profile, 10 compounds from the NCI Diversity Set were identified. As explained in subchapter 2.5.1, it was decided to expand the docking set by searching for compounds at least 95% similar in ZINC-22, resulting in a set consisting of 134 structures.

3.5.1 Docking Results

Several parameters were obtained from the molecular docking phase, performed using GNINA 1.0.1 software.

- **Affinity:** It is a fundamental result of molecular docking and represents the binding affinity of a small molecule in the predicted minimum energy state. It is an estimate of the binding free energy (ΔG) in kcal/mol. It is therefore an energy estimate obtained with the Vina-like scoring function integrated in GNINA. More negative values indicate a more favorable predicted binding (as it is an estimate of ΔG).
- **CNNscore:** It is a confidence score learned from experimental data. The CNN takes as input a voxelized 3D representation of the ligand–receptor complex and outputs this value, which ranges from 0 to 1. Values closer to 1 indicate that the CNN believes that pose is more similar to a correct pose, while values closer to 0 indicate that the pose is considered unlikely.
- **CNNaffinity [kcal/mol]:** It is defined as the affinity of the docked complex determined by the CNN, also expressed in kcal/mol. Unlike Affinity (derived from an empirical function), this parameter is a deep learning regressor trained on experimental measures of strength/affinity (on known protein-ligand complexes). It therefore integrates 3D structural information and experimental data. Like Affinity, more negative values indicate stronger predicted binding. Gnina uses this parameter to sort the poses generated by the classical function [87].

Table 3.5 collects the data relating to these three indicator parameters for each of the 134 compounds constituting the docking set.

After understanding the parameter definitions, it was determined that the ideal ligand would have a very negative affinity, a high CNN score (converging to 1), and a low CNN affinity.

In agreement with the University of Alberta laboratory, which allows experimental testing to be conducted, it was decided to identify three compounds for each of the ten “families”. Only for one “family”, containing a single member, was it obviously not possible to choose three. For this reason, a total of 28 compounds were selected. For each of the three compounds within the “family”, a priority was established with regard to the order of laboratory testing.

Therefore, the data were analyzed. The candidates and their priority were selected according to the following criterion:

- ligands with a more negative Affinity were preferred;
- with equal (or similar) Affinity, the CNNscore value was observed to understand the goodness of the prediction;
- in case of a CNN score that was too low (converging to 0), the second most negative value of the "family" was taken;
- if the Affinity value was much more negative than the others in the family, that ligand was still taken into consideration;
- if the Affinity and CNNscore values were almost equal, compounds with better CNNaffinity were selected.

Consequently, the Affinity parameter provided by Gnina was the one used preferentially for selecting candidates (it ranges between –4 and –7 kcal/mol). This is because it is considered a standard AutoDock Vina-style docking score. It is more widely described in the literature than CNNaffinity and therefore allows the results obtained to be compared more easily with other studies. Furthermore, it is more easily interpretable as an estimate of ΔG.
Note that lines in Table 3.5 with values clearly outside the acceptance range (found positive and very high) are almost certainly failed poses and should be ignored.

ID	Affinity [kcal/mol]	CNNscore	CNNaffinity [kcal/mol]
140131	**-6,563**	**0,162**	**5,080**
ZINC5085294	-6,586	0,262	5,580
ZINC5085292	-6,304	0,215	5,325
ZINC8034761	-7,204	0,071	5,289
ZINC5085293	-6,533	0,146	5,292
ZINC5085291	-5,200	0,195	5,329
ZINC195497425	-5,933	0,121	5,425
ZINC195497419	-2,449	0,594	5,889
ZINC195497422	-5,422	0,337	5,440
ZINC195497415	-5,287	0,516	5,759
ZINC101416916	-6,582	0,216	5,491
ZINC584567037	-4,341	0,444	5,556
ZINC242498472	72,818	0,002	4,082
ZINC242498473	-5,618	0,096	5,169
ZINC101416910	-5,875	0,257	5,568
ZINC242498474	-5,351	0,151	5,178
NSC 111702	**-4,909**	**0,321**	**3,711**
ZINC4994399	-3,708	0,184	3,545
ZINC4994397	-5,117	0,074	3,636
ZINC6200569	-3,521	0,345	3,680
ZINC17424869	-4,239	0,268	3,681
ZINC256824146	-3,670	0,172	3,298
ZINC256824144	-3,549	0,346	3,518
ZINC256824148	-4,183	0,198	3,752
ZINC4994398	-3,875	0,272	3,560
ZINC4994396	-3,378	0,373	3,625
ZINC1703355	-4,510	0,360	3,699
NSC 12161	**-5,334**	**0,194**	**3,428**
ZINC4018123	-4,313	0,148	3,424
ZINC4018121	31,832	0,002	2,819
ZINC8603212	-5,041	0,046	3,390
ZINC13831597	-2,778	0,252	3,506
ZINC4082270	-5,488	0,038	3,395
ZINC28090442	-4,981	0,176	3,914
ZINC13546632	-3,057	0,273	3,665
ZINC4018122	-4,852	0,223	3,575
ZINC4018120	-2,882	0,377	3,637
ZINC1531100	-5,093	0,054	3,280
ZINC3953841	-3,338	0,169	3,486
ZINC6283728	-1,573	0,626	3,698

ID	Affinity [kcal/mol]	CNNscore	CNNaffinity [kcal/mol]
NSC 133118	**-4,532**	**0,155**	**3,345**
ZINC16951333	7,292	0,293	3,798
ZINC4964037	-4,773	0,057	3,314
ZINC16951334	-3,885	0,361	3,451
ZINC166589324	-4,397	0,552	4,031
ZINC105219439	-5,050	0,124	3,543
ZINC105219446	-4,487	0,287	3,490
ZINC4964039	-4,345	0,267	3,578
ZINC4964034	-4,340	0,264	3,980
ZINC1720023	-4,340	0,349	3,521
NSC 143099	**-4,714**	**0,445**	**5,626**
NSC 188491	**-4,595**	**0,119**	**3,784**
ZINC5011744	-4,724	0,094	3,714
ZINC5011742	-5,932	0,030	3,844
ZINC13786330	-5,094	0,084	3,646
ZINC43178883	-4,342	0,109	3,559
ZINC105301382	-5,706	0,085	3,857
ZINC263611392	-5,809	0,028	3,529
ZINC5011743	-4,946	0,057	3,896
ZINC5011741	-4,275	0,049	3,679
ZINC3954489	-5,221	0,207	3,884
NSC 52902	-2,995	0,346	2,452
ZINC4722016	-4,575	0,127	2,531
ZINC4722017	-4,197	0,224	2,717
ZINC1684223	-4,328	0,227	2,657
ZINC3954052	-3,437	0,465	2,704
ZINC17313269	-4,645	0,144	2,624
ZINC105061936	-4,076	0,276	2,744
ZINC17313271	-4,239	0,475	3,114
ZINC105061931	-3,633	0,286	2,543
NSC 86005	**94,840**	**0,001**	**3,053**
ZINC195472092	18,281	0,060	4,402
ZINC195472104	2,809	0,253	4,539
ZINC102310094	-4,555	0,563	4,810
ZINC257157676	-1,805	0,423	5,217
ZINC257157677	-4,137	0,428	4,816
ZINC257157678	-4,702	0,364	4,579
ZINC257157675	-2,194	0,527	4,818
ZINC195472098	-4,519	0,553	4,786
ZINC195472108	-0,085	0,414	4,946
ZINC100133014	-0,750	0,352	4,662
ZINC100133012	-1,879	0,581	4,627
ZINC100133022	-4,422	0,577	4,842

ID	Affinity [kcal/mol]	CNNscore	CNNaffinity [kcal/mol]
ZINC100133015	-3,853	0,667	4,811
NSC 287050	**-4,376**	**0,176**	**2,479**
ZINC59065544	-3,193	0,378	2,349
ZINC3861281	-4,266	0,170	2,335
ZINC59065540	-3,257	0,487	2,547
ZINC6495395	-4,065	0,324	2,716
ZINC1532813	-2,035	0,653	2,413
ZINC3954528	-4,314	0,214	2,501
ZINC12501243	-4,364	0,242	2,604
ZINC44608692	-4,437	0,251	2,555
ZINC101184237	-4,288	0,200	2,571
ZINC3861280	-3,962	0,314	2,486
ZINC29309317	-4,367	0,183	2,391
ZINC5225021	-3,921	0,288	2,562
ZINC1532814	-3,293	0,425	2,337
ZINC3606246	-2,595	0,454	2,312
ZINC3870036	-3,395	0,200	2,120
ZINC3870035	-3,786	0,326	2,386
ZINC2043005	-4,050	0,220	2,647
ZINC15206242	-4,695	0,274	2,620
ZINC2042980	-3,561	0,398	2,510
ZINC1532815	-4,944	0,166	2,485
ZINC6490946	-3,449	0,287	2,367
ZINC44608006	-3,785	0,300	2,452
ZINC1532676	-3,827	0,180	2,537
ZINC12888359	-3,630	0,220	2,194
ZINC38282241	-3,578	0,367	2,573
ZINC5225024	-3,602	0,268	2,333
ZINC2042981	-3,949	0,295	2,556
ZINC1532816	-4,403	0,306	2,595
ZINC2047187	-3,871	0,301	2,515
ZINC4090206	-2,546	0,590	2,540
ZINC1532677	-3,789	0,253	2,276
ZINC3870034	-4,451	0,159	2,476
NSC 2561	**-5,061**	**0,155**	**2,712**
ZINC4403650	-2,632	0,477	2,978
ZINC4403651	-4,147	0,216	2,807
ZINC4403653	-1,042	0,416	2,870
ZINC36386172	-2,302	0,610	2,938
ZINC36373710	-2,507	0,466	3,405
ZINC4403652	-3,897	0,209	2,514
ZINC402451	-5,027	0,335	2,856
ZINC34570732	-2,088	0,527	3,298

ID	Affinity [kcal/mol]	CNNscore	CNNaffinity [kcal/mol]
ZINC3953805	-2,496	0,457	3,053
ZINC1319651	-3,377	0,455	2,796
ZINC44699134	-4,637	0,275	2,738
ZINC34428856	-4,979	0,234	2,836
ZINC12670933	-3,722	0,428	2,965
ZINC44699135	-4,467	0,231	2,607
ZINC40834470	-4,098	0,142	2,508
ZINC6096442	-1,692	0,432	2,964

Table 3.5: *Results obtained from the docking process using the Gnina 1.0.1 software. For each ligand in the docking set, the Affinity, CNNscore and CNNaffinity parameters are reported. The 10 compounds selected at the end of the virtual screening and preliminary property evaluation phase are highlighted in bold and from there onwards, up to the new compound in bold, the structures are part of the same family of analogues. The compounds highlighted in green have been identified as the first choice, those in blue as the second choice, and those in yellow as the third choice.*

3.5.2 ADMET Predictor v12.0 Results

ADMET properties and those related to Lipinski's rule of five were analyzed for all 134 ligands in the docking set. The results, obtained with ADMET Predictor v12.0 developed by Simulations Plus, are shown in Table 3.6. The same values or conditions were found for all ligands belonging to the same "family," and for this reason, only the ligand identifiers obtained at the end of the virtual screening phase were reported as a reference.

The only exception is the family of compound NSC 140131, whose members can be divided into two subgroups based on the results obtained. The compounds NSC 140131, ZINC5085294, ZINC5085292, ZINC8034761, ZINC5085293, ZINC5085291, ZINC195497425, ZINC195497419, ZINC195497422, ZINC195497415 obtained the results reported in Table 3.6 in the column "NSC 140131 (*)". The compounds ZINC101416916, ZINC584567037, ZINC242498472, ZINC242498473, ZINC101416910 and ZINC242498474 obtained the following values/conditions:

- MWt=594.5, HBA=13, HBD=10, RuleOf5=3, RuleOf5_Code=Hb-Mw-NO, T_PSA=230.0;
- hERG_Filter=No, hERG_pIC50=3.99, Repro_Tox=Non Toxic (45%), TOX_Risk=2.40;
- %Fa_hum-1.0=70.15, S+logD=1.59, S+logP=1.62, S+MDCK=5.09, S+Peff=0.22, S+Sw=0.08, BBB_Filter=Low (97%), LogBB=-0.69;
- CYP2D6 Inh=Yes24%), CYP3A4 Inh=No (50%), S+CL_Metab=No, S+CL_Renal=No (80%), S+CL_Uptake=No (99%).

As explained in subsection 3.5.1, the preferred criterion adopted for selecting candidates was the score obtained at the end of the docking process. Nevertheless, in the event of equal parameter results, for a more robust choice, the ADMET properties and Lipinski's rule of five were evaluated, favouring compounds with a better pharmacokinetic and overall safety profile according to the criteria listed in Table 2.4 of subchapter 2.5.3.

ID		NSC 140131 (*)	NSC 111702	NSC 12161	NSC 133118	NSC 143099	NSC 188491	NSC 52902	NSC 86005	NSC 287050	NSC 2561
Rule of five	**MWt**	610,5	314,3	283,3	268,2	578,5	339,3	180,2	585,6	164,2	240,3
	HBA	14	8	8	9	12	11	7	13	5	5
	HBD	11	3	5	4	10	7	6	6	4	3
	RuleOf5	3	0	1	0	3	2	1	3	0	0
	RuleOf5_Code	Hb Mw NO	--	Hb	--	Hb Mw NO	Hb NO	Hb	Hb Mw NO	--	--
	T_PSA	250,2	122,8	159,5	152,4	220,8	207,8	136,0	203,5	90,2	79,2
Toxicity	**hERG_Filter**	No	No (91%)	No	No (91%)	No	No	No	No	No	No (91%)
	hERG_pIC50	3,79	4,44	4,30	4,46	3,92	4,65	4,31	3,67	4,11	4,41
	Repro_Tox	NT	NT (80%)	NT (71%)	NT (90%)	T (62%)	NT (90%)	NT (64%)	T (71%)	T (52%)	T (65%)
	TOX_Risk	3,00	3,50	1,50	3,00	2,50	2,50	2,50	2,00	2,50	0,50
Permeability and absorption	**%Fa_hum-1.0**	65,36	73,22	26,46	59,63	67,69	11,31	12,29	63,06	60,80	93,80
	S+logD	1,46	-0,71	-2,18	-1,29	2,07	-2,50	-3,32	1,09	-1,92	0,06
	S+logP	1,58	-0,71	-2,18	-1,29	2,10	-2,49	-3,32	1,34	-1,92	0,06
	S+MDCK	4,83	48,53	15,79	64,27	4,88	10,72	11,65	7,31	18,86	62,07

ID		NSC 140131 (*)	NSC 111702	NSC 12161	NSC 133118	NSC 143099	NSC 188491	NSC 52902	NSC 86005	NSC 287050	NSC 2561
Permeability and absorption (continued from previous page)	**S+Peff**	0,19	0,58	0,21	0,46	0,15	0,11	0,16	0,25	0,69	1,14
	S+Sw	0,16	12,75	1,63	4,59	0,07	0,99	139,45	0,37	332,09	19,62
	BBB_Filter	Low	Low (67%)	Low (77%)	Low (46%)	Low (97%)	Low (97%)	Low (83%)	Low (97%)	Low (43%)	High (84%)
	LogBB	-0,71	-1,13	-1,22	-1,14	-0,60	-1,34	-1,16	-0,75	-0,61	-0,69
Metabolism	**CYP2D6_Inh**	No (76%)	No (97%)	No (97%)	No (97%)	Yes (33%)	No (97%)	No (97%)	No (81%)	No (97%)	No (97%)
	CYP3A4_Inh	No (47%)	No (96%)	No (96%)	No (96%)	Yes (48%)	No (96%)	No (96%)	No (96%)	No (96%)	No (96%)
	S+CL_Metab	No	No (92%)	No (98%)	No (98%)	No	No (98%)	No	No (51%)	No (98%)	No (51%)
	S+CL_Renal	No (84%)	No (99%)	Yes (95%)	Yes (45%)	No (87%)	Yes (95%)	Yes	Yes (45%)	No (97%)	No (99%)
	S+CL_Uptake	No (99%)	No (99%)	No (99%)	No (99%)	No (99%)	No (99%)	No (99%)	No	No (99%)	No (99%)

(*) Members of this "family" of compounds did not obtain the same results for all the reported properties.

Table 3.6: *Results of ADMET properties evaluated with the ADMET Predictor software, developed by SimulationPlus. They are divided into four macro-categories: Lipinski's rule of five, toxicity, permeability and absorption, and metabolism. Although the evaluation was performed on all 134 ligands, only the 10 compounds selected at the end of the virtual screening phase (one in each column) are reported, as compounds belonging to the same "family" obtained the same results for all properties. The only exception is the NSC 140131 "family". The abbreviation NT stands for Nontoxic and T stands for Toxic. "--" indicates that no conditions that would violate Lipinski's rule were detected.*

Appendix D lists which ligands belong to each of the 10 "families" of compounds.

3.5.3 Ligands selected for experimental tests

The following compounds were selected for being used as potential anchors for a PROTAC. They were evaluated as being the best candidates for binding to the UBR-box domain of UBR1 E3 Ligase in the pocket site_select_1.

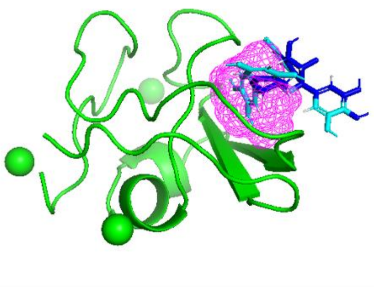
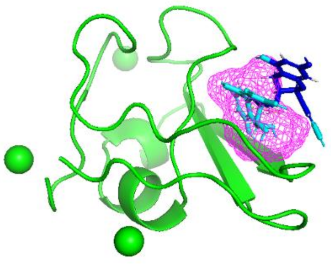
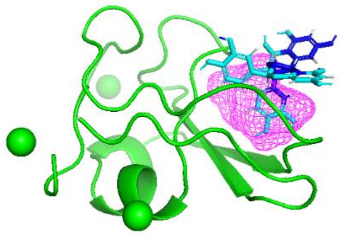
They are divided according to the "family" they belong to, highlighted in blue in the figures. For each one, the compounds selected as first, second and third choice are listed. Each structure has its own identifier below it. If the structure comes from the NCI Diversity Set, it will begin with “NSC”, while if it is one of the 95% similar compounds found in ZINC-22, it will start with the prefix “ZINC”.

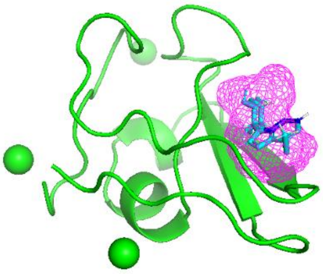
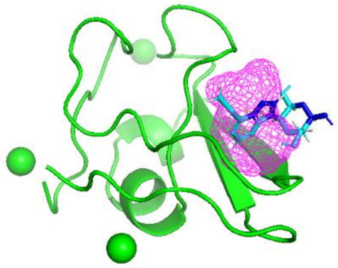
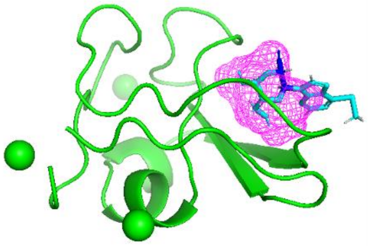
The series of figures 3.24 show where and how the ligand positions itself to bind to the UBR-box domain. During the configuration phase of the docking process, a very large box (22 Å, see Figure 2.8) was set up, which also covered other portions of the protein. Despite this, all the selected ligands are positioned in the chosen pocket, on which the pharmacophore model was built. This means that the workflow followed, which involved selecting a binding pocket, creating a pharmacophore model and using it as a hypothesis for virtual screening, yielded the desired results.

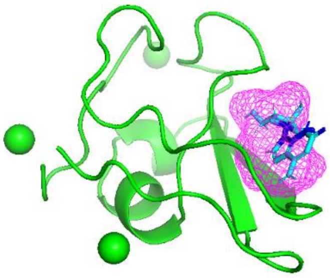
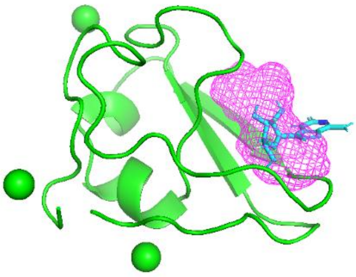
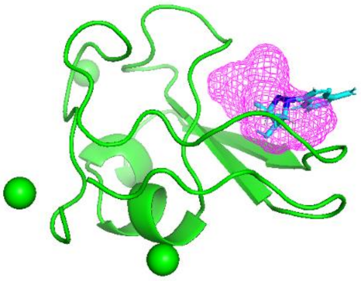
As a first choice, it was decided to identify compounds with different chemical structures, thus selecting the best compound from each group of analogues. One of the objectives of the project was to maintain high chemical diversity among the candidates, so that the experimental tests would cover different scaffolds and increase the probability of finding at least one good hit. This has helped reduce the risk of focusing entirely on a single chemical class.

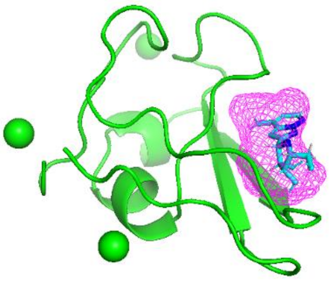
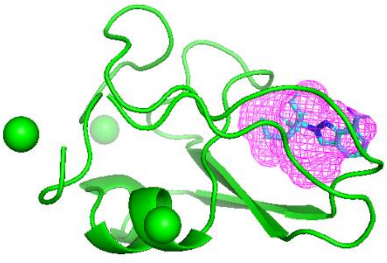
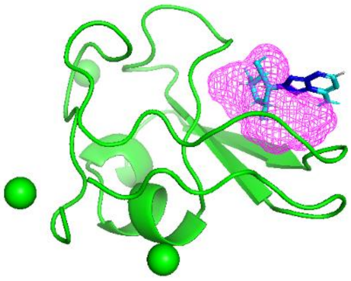
However, in order to understand whether the bond is strong or whether the activity may be lost as a result of a small change, it was decided to select the second and third choices within the same group of 95% analogues. This will facilitate structure-activity relationship (SAR) analysis, allowing effects due to the basic structure to be distinguished from those due to small peripheral changes. Furthermore, they are useful for providing “backup compounds” if the first candidates fail.

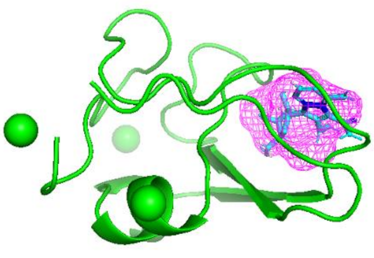
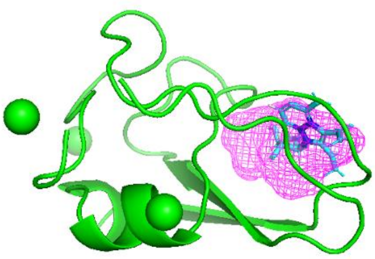
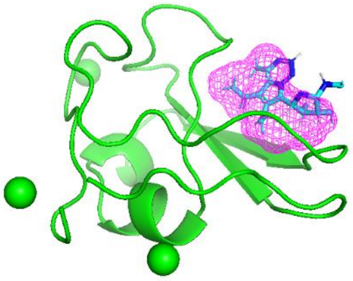
NSC 140131

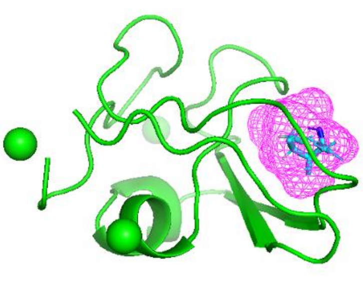
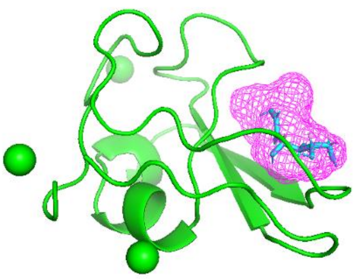
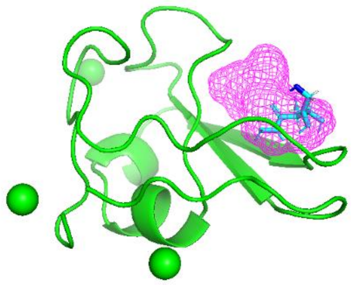
1st choice	2nd choice	3rd choice
		
ZINC8034761	ZINC5085294	ZINC101416916

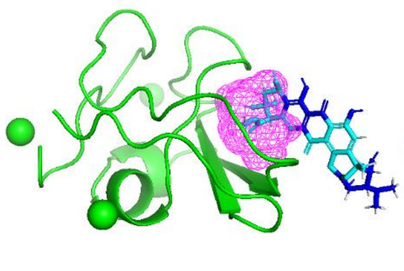
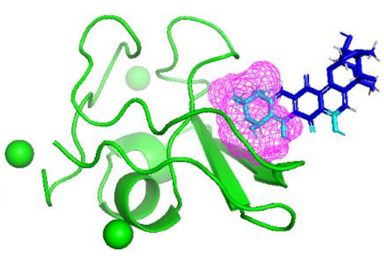
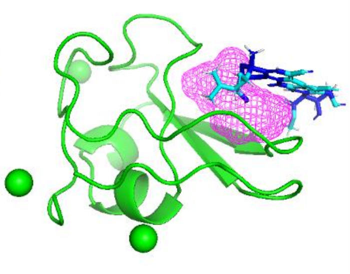
NSC 111702		
1st choice	2nd choice	3rd choice
		
ZINC4994397	NSC 111702	ZINC1703355

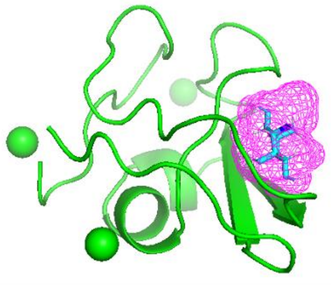
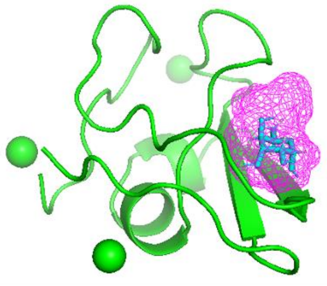
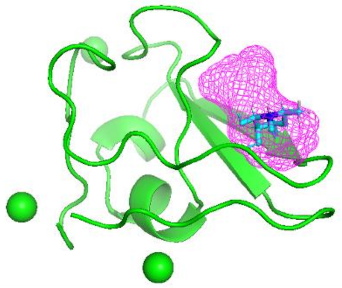
NSC 12161		
1st choice	2nd choice	3rd choice
		
ZINC4082270	NSC 12161	ZINC1531100

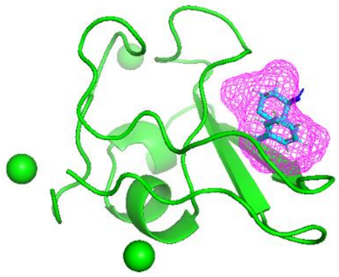
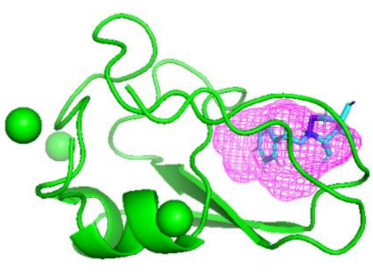
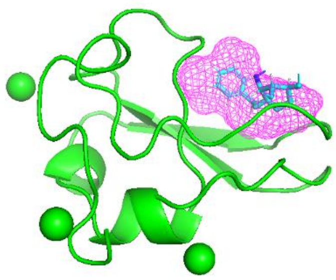
NSC 133118		
1st choice	2nd choice	3rd choice
		
ZINC105219439	ZINC4964037	ZINC105219446

NSC 188491		
1st choice	2nd choice	3rd choice
		
ZINC5011742	ZINC263611392	ZINC105301382

NSC 52902		
1st choice	2nd choice	3rd choice
		
ZINC17313269	ZINC4722016	ZINC1684223

NSC 86005		
1st choice	2nd choice	3rd choice
		
ZINC257157678	ZINC102310094	ZINC195472098

NSC 287050		
1st choice	2nd choice	3rd choice
		
ZINC1532815	ZINC15206242	ZINC3870034

NSC 2561		
1st choice	2nd choice	3rd choice
		
NSC 2561	ZINC402451	ZINC34428856

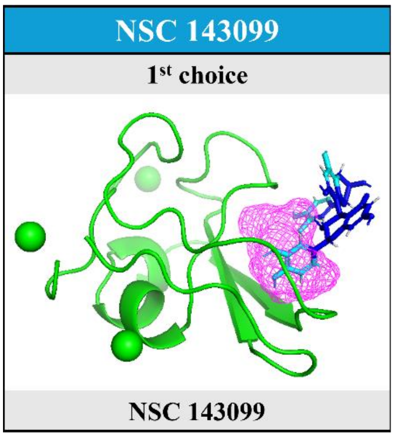


Figure 3.24: *Series of figures representing the result of docking (performed with Gnina software) between the target, i.e. the UBR-box of UBR1 E3 Ligase (in green), and the different ligands selected as candidates for experimental testing (in blue) in the site_select_1 pocket (in pink). The light blue coloured ligand residues are those 4 Å or less away from the target. The results were visualised with PyMOL. The name representing the "family" of analogues is highlighted in blue. Each structure has its identifier below it. The compounds were divided according to their priority for experimental testing.*

4. Discussion and Future Directions

Since a PROTAC that recruits UBR1 E3 Ligase and acts against MM has not yet been designed, this project can be considered a starting point for the development of this technology. The ligands selected to be used as potential anchors are 28 and are divided into 10 groups of analogues (also called "families") based on priority for testing. For the first phase of testing (first-choice candidates), 10 ligands were identified, each belonging to a different group of analogues. This decision is based on the fact that previous studies of this type have revealed a success rate of approximately 10%, therefore, at least one in ten compounds is expected to be active in experimental tests.

Although the entire UBR1 E3 Ligase protein has not yet been experimentally resolved, the computational approach used here has demonstrated the robustness of AlphaFold2's prediction. MD simulation, performed under near-physiological conditions, has revealed that the predicted structure is in thermodynamic equilibrium and is able to maintain it. Furthermore, it does not assume any unpermitted conformations. Only a few residues fall into non-permitted regions of the Ramachandran plot, but these could be more flexible loops or ends. The RMSD and secondary structure graphs show that the protein does not undergo drastic conformational changes during the simulation. However, due to limited computational resources, it was not possible to perform a simulation longer than 10 ns. In addition, the PAE shows high prediction confidence in experimentally resolved domains such as the UBR-box, while for others the uncertainty is greater. Therefore, to obtain even more robust results, the simulation time should be extended.

In the UBR-box, the structure of UBR1 predicted by AlphaFold2 was found to be very similar to the homology model, obtaining an RMSD value of 0.491 Å.

The homology model of the UBR-box domain, built with Swiss-MODEL using 3NY3 as a template structure, proved to be suitable for the project goal. In fact, the template and UBR1 sequences were found to have a very high similarity percentage and very good values of the other parameters analyzed. The obtained model also proved stability during MD simulations, both in terms of energy and secondary structure configuration, with no residues falling into prohibited regions. For this reason, it was possible to extract the most representative structure belonging to the most numerous cluster in the second half of the MD trajectory, to use it for the next steps, i.e. searching for binding pockets with the FTSite software.

Of the three pockets found, the one in which the peptide bound in the 3NY3 template structure was located, which tended to be neutral, was chosen. It showed a good balance between cavity exposure and depth and was ideal for hosting molecules with hydrophilic and hydrophobic regions. The pharmacophore model was built on it and was used as a hypothesis in virtual screening to speed up the process. It should be noted that other pockets of the UBR-box domain could be analyzed in future studies. Those found in other UBR1 domains may not be entirely reliable, as they are still only predictions.

The virtual screening was based on the NCI Diversity Set, which, despite including compounds with high chemical diversity (as evidenced by the results obtained), is small.

It should be noted that future studies could be conducted using a different, larger screening database. The ligands selected at the end of this phase were those with the best pharmacokinetic and cardiac safety profiles. Thanks to the results obtained from the docking process, performed with the Gnina software, it was possible to select the best candidates with the highest (and therefore most negative) affinity values. However, all the selected ligands, whether first, second, or third choice, positioned themselves correctly in the selected binding pocket. In addition, their overall ADMET profile also appears satisfactory for their application as anchors in a PROTAC.

The workflow employed in this project integrates several complementary techniques, such as homology modeling, MD simulations, pocket detection, pharmacophore modeling, virtual screening, docking, and ADMET prediction. However, it remains based on an entirely computational approach. Therefore, all results obtained must be interpreted as hypotheses generated in silico and not proven to be definitive of binding or biological efficacy. Indeed, each step introduces a source of uncertainty: UBR1 structural models are predicted, docking is based on approximate scoring functions and simplified representations of flexibility and solvation, and the properties of ADMET and hERG are also inferred from machine learning models trained on external datasets.

While these methods are effective for prioritizing compounds and reducing the chemical space, they cannot replace experimental validation. For this reason, the ligands selected as UBR1 E3 Ligase recruiters have been purchased and are currently being tested at the University of Alberta laboratory. They will be able to confirm or deny the affinity for UBR1 and thus verify whether the computational predictions will translate into actual biological activity.

The construction of the entire PROTAC could be based on technologies of this type already known in the literature. One example is dCBP-1. It acts as a targeted chemical degradator for the lysine acetyltransferases of the enhancers CBP and p300. The study, conducted by Vannam and colleagues, shows the use of in silico modelling as an effective guide for the synthesis of this PROTAC. It works by recruiting the ubiquitin ligase E3 CRBN which, exploiting the UPS mechanism, manages to ubiquitinate the key factors for MM survival. CRISPR-based genetic knockout data were analysed to confirm that multiple myeloma cells are most dependent on both p300 and CBP, and dCBP-1 activity was tested in several MM cell lines. This technology is therefore validated for this disease.

Schematically, dCBP-1 is composed as follows:

- The warhead is GNE-781 and it binds to the bromodomain of p300/CBP, the protein of interest;
- The anchor is a thalidomide-based ligand that recruits the E3 ligase CRBN;
- The linker, a polyethylene glycol-4 (PEG-4), is the chain that chemically connects GNE-781 to the CRBN ligand and allows for the formation of a stable ternary complex (p300/CBP–dCBP-1–CRBN) [93].

In the final synthesis of dCBP-1, the tetrahydropyran ring system was replaced with a piperidine to facilitate the bond with the linker.

If at least one of the selected compounds proves to be active in laboratory tests, this project could proceed with the construction of the entire PROTAC that acts against MM. The warhead and linker constituting the dCBP-1 technology could be used in this new PROTAC, while the compound that proved to be active could act as an anchor for UBR1 E3 Ligase.

However, it should be noted that this idea is only a hypothesis that must be supported by both validated in silico data and experimental tests. In fact, even if the same linker-warhead-POI complex works correctly in dCBP-1 technology, it is not certain that the same will happen when the new anchor is inserted. It is therefore necessary to conduct further studies to first verify whether a correct bond can be formed between the linker hypothesised here and the anchor and whether this bond negatively affects the one created between the anchor and UBR1 E3 Ligase.

The proposal to change the anchor that binds to E3 ligase arises from the fact that ligands that recruit CRBN have already been extensively studied in the literature. Furthermore, mechanisms of resistance to CRBN exist in patients with MM undergoing long-term treatment. The development of PROTACs that exploit alternative ligases, such as UBR1, could offer a therapeutic option in patients resistant to CRBN-dependent drugs or new therapeutic combinations with the simultaneous use of CRBN-PROTAC and UBR1-PROTAC.

Furthermore, this application could be used not only against MM but also against different types of cancer, by changing the warhead.

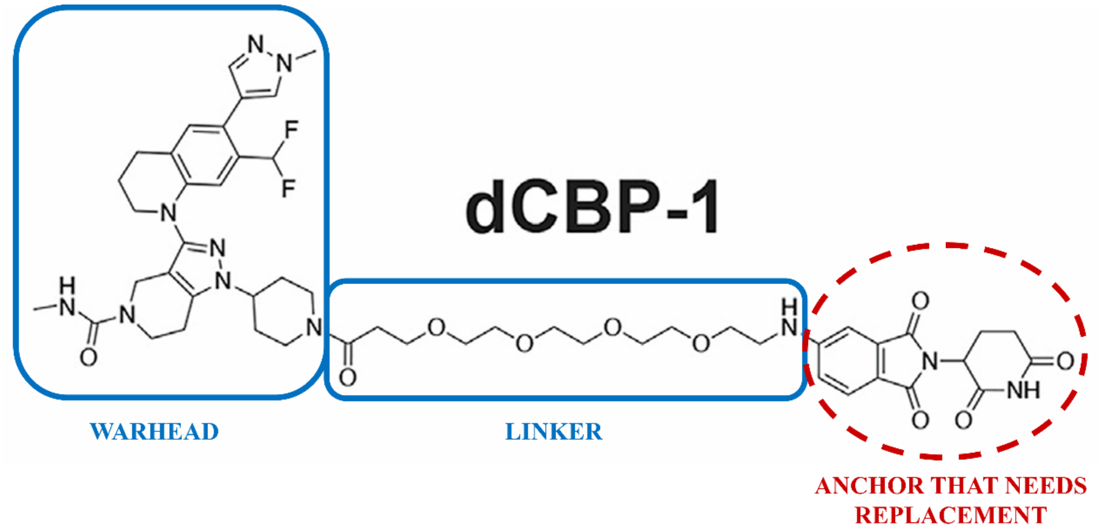


Figure 4.1: *2D structure of PROTAC dCBP-1 (CAS No. 2484739-25-3). The chemical structures of the warhead and of the linker are circled in blue. The structure of the anchor that can be replaced is circled in red.*

Appendix A

Command line for neutralising the 3NY3 structure without the initially bound peptide in Gromacs:

```
gmx pdb2gmx -f 3ny3_clean_z.pdb -o 3ny3.gro -p topol.top
      6: AMBER99SB-ILDN protein, nucleic AMBER94
      1: TIP3P

gmx editconf -f 3ny3.gro -o 3ny3_box.gro -c -d 1.0 -bt cubic

gmx solvate -cp 3ny3_box.gro -cs spc216.gro -o 3ny3_sol.gro -
p topol.top

gmx grompp -f ions.mdp -c 3ny3_box.gro -p topol.top -o ions.tpr
-maxwarn 2

gmx genion -s ions.tpr -o 3ny3_neutral.gro -p topol.top -pname
NA -nname CL −neutral
      SOL group

gmx editconf -f 3ny3_neutral.gro -o 3ny3_neutral.pdb
```

ions.mdp file:

```
; ions.mdp — parameter file used only for grompp → genion
integrator    = steep
nsteps        = 1
cutoff-scheme = Verlet
coulombtype   = PME
rcoulomb      = 1.0
vdwtype       = Cut-off
rvdw          = 1.0
pbc           = xyz
```

Appendix B

Protein sequence of HUMAN E3 ubiquitin-protein ligase UBR1 in FASTA format (canonical) from Uniprot:

```
>sp|Q8IWV7|UBR1_HUMAN E3 ubiquitin-protein ligase UBR1 OS=Homo sapiens
OX=9606 GN=UBR1 PE=1 SV=1

MADEEAGGTERMEISAELPQTPQRLASWWDQQVDFYTAFLHHLAQLVPEIYFAEMD
PDLEKQEESVQMSIFTPLEWYLFGEDPDICLEKLKHSGAFQLCGRVFKSGETTYSCRD
CAIDPTCVLCMDCFQDSVHKNHRYKMHTSTGGGFCDCGDTEAWKTGPFCVNHEPG
RAGTIKENSRCPLNEEVIVQARKIFPSVIKYVVEMTIWEEEKELPPELQIREKNERYYC
VLFNDEHHSYDHVIYSLQRALDCELAEAQLHTTAIDKEGRRAVKAGAYAACQEAKE
DIKSHSENVSQHPLHVEVLHSEIMAHQKFALRLGSWMNKIMSYSSDFRQIFCQACLR
EEPDSENPCLISRLMLWDAKLYKGARKILHELIFSSFFMEMEYKKLFAMEFVKYYKQ
LQKEYISDDHDRSISITALSVQMFTVPTLARHLIEEQNVISVITETLLEVLPEYLDRNNK
FNFQGYSQDKLGRVYAVICDLKYILISKPTIWTERLRMQFLEGFRSFLKILTCMQGME
EIRRQVGQHIEVDPDWEAAIAIQMQLKNILLMFQEWCACDEELLLVAYKECHKAVM
RCSTSFISSSKTVVQSCGHSLETKSYRVSEDLVSIHLPLSRTLAGLHVRLSRLGAVSRL
HEFVSFEDFQVEVLVEYPLRCLVLVAQVVAEMWRRNGLSLISQVFYYQDVKCREEM
YDKDIIMLQIGASLMDPNKFLLLVLQRYELAEAFNKTISTKDQDLIKQYNTLIEEMLQ
VLIYIVGERYVPGVGNVTKEEVTMREIIHLLCIEPMPHSAIAKNLPENENNETGLENVI
NKVATFKKPGVSGHGVYELKDESLKDFNMYFYHYSKTQHSKAEHMQKKRRKQEN
KDEALPPPPPPEFCPAFSKVINLLNCDIMMYILRTVFERAIDTDSNLWTEGMLQMAFH
ILALGLLEEKQQLQKAPEEEVTFDFYHKASRLGSSAMNIQMLLEKLKGIPQLEGQKD
MITWILQMFDTVKRLREKSCLIVATTSGSESIKNDEITHDKEKAERKRKAEAARLHRQ
KIMAQMSALQKNFIETHKLMYDNTSEMPGKEDSIMEEESTPAVSDYSRIALGPKRGP
SVTEKEVLTCILCQEEQEVKIENNAMVLSACVQKSTALTQHRGKPIELSGEALDPLFM
DPDLAYGTYTGSCGHVMHAVCWQKYFEAVQLSSQQRIHVDLFDLESGEYLCPLCKS
LCNTVIPIIPLQPQKINSENADALAQLLTLARWIQTVLARISGYNIRHAKGENPIPIFFN
QGMGDSTLEFHSILSFGVESSIKYSNSIKEMVILFATTIYRIGLKVPPDERDPRVPMLT
WSTCAFTIQAIENLLGDEGKPLFGALQNRQHNGLKALMQFAVAQRITCPQVLIQKHL
VRLLSVVLPNIKSEDTPCLLSIDLFHVLVGAVLAFPSLYWDDPVDLQPSSVSSSYNHL
YLFHLITMAHMLQILLTVDTGLPLAQVQEDSEEAHSASSFFAEISQYTSGSIGCDIPGW
YLWVSLKNGITPYLRCAALFFHYLLGVTPPEELHTNSAEGEYSALCSYLSLPTNLFLL
FQEYWDTVRPLLQRWCADPALLNCLKQKNTVVRYPRKRNSLIELPDDYSCLLNQAS
HFRCPRSADDERKHPVLCLFCGAILCSQNICCQEIVNGEEVGACIFHALHCGAGVCIF
LKIRECRVVLVEGKARGCAYPAPYLDEYGETDPGLKRGNPLHLSRERYRKLHLVWQ
QHCIIEEIARSQETNQMLFGFNWQLL
```

Protein sequence of neutralised 3NY3, without the bound peptide in FASTA format (canonical) found with PyMol:

```
>3ny3_neutral_

LCGRVFKVGEPTYSCRDCAVDPTCVLCMECFLGSIHRDHRYRMTTSGGGGFCDCGD
TEAWKEGPYCQKHE
```

Appendix C

Command line in GROMACS for MD simulations of homology model of the UBR-box of UBR1 E3 Ligase:

```
gmx pdb2gmx -f HM_swiss.pdb -o HM_swiss.gro -p topol.top –
ignh –heavyh
      6: AMBER99SB-ILDN protein, nucleic AMBER94
      1: TIP3P

gmx editconf –f HM_swiss.gro –o HM_swiss_box.gro –c –d 1.0 –
bt cubic

gmx   solvate   -cp   HM_swiss_box.gro   -cs   spc216.gro   -o
HM_swiss_sol.gro -p topol.top

gmx  grompp  -f  em.mdp  -c  HM_swiss_sol.gro  -p  topol.top  -o
ions.tpr -maxwarn 1

gmx  genion  -s  ions.tpr  -o  HM_swiss_ions.gro  -p  topol.top  -
pname NA -nname CL –neutral –conc 0.15
      14:  Ions  were  added  to  SOL  (with  AlphaFold  model  the
      SOL group was the 13th)

gmx  grompp  -f  em.mdp  -c  HM_swiss_ions.gro  -p  topol.top  -o
em.tpr

gmx mdrun –s em.tpr –v –deffnm em

gmx  grompp  -f  nvt.mdp  -c  em.gro  -r  em.gro  -p  topol.top  -o
nvt.tpr

gmx mdrun -deffnm nvt

gmx grompp -f npt.mdp -c nvt.gro -r nvt.gro -p topol.top -o
npt.tpr

gmx mdrun -deffnm npt

gmx grompp -f mdout.mdp -c npt.gro -r npt.gro -p topol.top -
o md.tpr –maxwarn 1

gmx mdrun -deffnm md
```

Command line in GROMACS for extracting the most representative structure of the second half of the trajectory:

```
gmx make_ndx -f md.tpr -o index.ndx
Then it was written:
     r ZN (found 3 atoms)
     1 | r ZN
     name 17 Protein_ZN
     q

gmx trjconv -s md.tpr -f md.xtc -o md_50_100ns_Pr_ZN.xtc -b
50000 -n index.ndx
group 18 (Protein_ZN) was selected (of 1017 + 3 = 1020 atoms)

gmx cluster -s md.tpr -f md_50_100ns_Pr_ZN.xtc -o clusters.xpm
-g cluster.log -dist cluster.xvg -method gromos -cl centroid.pdb
-cutoff 0,055 -n index.ndx
group 18 (Protein_ZN) was selected twice
```

em.mdp file:

```
integrator          = steep
nsteps              = 50000

emtol               = 1000
emstep              = 0.01

nstlog              = 100
nstenergy           = 100

cutoff-scheme       = Verlet
nstlist             = 10
ns-type             = Grid
pbc                 = xyz
rlist               = 1.0
coulombtype         = pme
coulomb-modifier    = Potential-shift-Verlet
rcoulomb            = 1.0
vdw-type            = cut-off
vdw-modifier        = Potential-shift-Verlet
rvdw                = 1.0

constraints         = none
```

nvt.mdp file:

```
define                  = -DPOSRES
integrator              = md
nsteps                  = 50000
dt                      = 0.002
; Output control
nstxout                 = 500
nstvout                 = 500
nstenergy               = 500
nstlog                  = 500
; Bond parameters
continuation            = no
constraint_algorithM  = lincs
constraints             = h-bonds
lincs_iter              = 1
lincs_order             = 4
; Nonbonded settings
cutoff-scheme           = Verlet
ns_type                 = grid
nstlist                 = 10
rcoulomb                = 1.0
rvdw                    = 1.0
DispCorr                = EnerPres
coulombtype             = PME
pme_order               = 4
fourierspacing          = 0.16

; Temperature coupling is on
tcoupl                  = V-rescale
tc-grps                 = Protein Non-Protein
tau_t                   = 0.1     0.1
ref_t                   = 310     310
; Pressure coupling is off
pcoupl                  = no      ; no pressure coupling in NVT
; Periodic boundary conditions
pbc                     = xyz
; Velocity generation
gen_vel                 = yes
gen_temp                = 300
gen_seed                = -1
```

npt.mdp file:

```
 define                  = -DPOSRES
integrator              = md
nsteps                  = 50000
dt                      = 0.002
; Output control
nstxout                 = 500
nstvout                 = 500
nstenergy               = 500
nstlog                  = 500
; Bond parameters
continuation            = yes       ; Restarting after NVT
constraint_algorithm = lincs
constraints             = h-bonds
lincs_iter              = 1
lincs_order             = 4

; Nonbonded settings
cutoff-scheme           = Verlet
ns_type                 = grid
nstlist                 = 10
rcoulomb                = 1.0
rvdw                    = 1.0
DispCorr                = EnerPres
; Electrostatics
coulombtype             = PME
pme_order               = 4
fourierspacing          = 0.16
; Temperature coupling is on
tcoupl                  = V-rescale
tc-grps                 = Protein Non-Protein
tau_t                   = 0.1     0.1
ref_t                   = 310     310
; Pressure coupling is on
pcoupl                  = Parrinello-Rahman
pcoupltype              = isotropic
tau_p                   = 2.0
ref_p                   = 1.0
compressibility         = 4.5e-5
refcoord_scaling        = com
; Periodic boundary conditions
pbc                     = xyz
; Velocity generation
gen_vel                 = no
```

md_simulation.mdp file:

```
; RUN CONTROL
integrator              = md          ; leap-frog integrator
dt                   = 0.002       ; 2 fs
nsteps                 = 50000000    ; 100 µs (50M · 0.002 ps)
continuation             = yes         ; continue from previous run
tinit                  = 0
; OUTPUT CONTROL
nstxout                = 500         ; coordinates every 1 ps
nstvout                = 500         ; velocities every 1 ps
nstenergy               = 500         ; energies every 1 ps
nstlog                 = 500         ; log every 1 ps
nstxout-compressed       = 1000        ; compressed trajectory every 2 ps
compressed-x-precision   = 1000
; NEIGHBOR SEARCHING
cutoff-scheme            = Verlet
nstlist                 = 10          ; update neighbor list every 20 fs
rlist                  = 1.0         ; short-range cutoff
; ELECTROSTATICS & VDW
coulombtype             = PME
rcoulomb                = 1.0
pme_order               = 4
fourierspacing           = 0.16
vdw-type                = Cut-off
rvdw                  = 1.0
DispCorr                = EnerPres    ; long-range dispersion correction
; TEMPERATURE COUPLING
tcoupl                 = V-rescale
tc-grps                = Protein Non-Protein
tau_t                 = 0.1 0.1
ref_t                 = 310 310
; PRESSURE COUPLING
pcoupl                 = Parrinello-Rahman
pcoupltype              = isotropic
tau_p                 = 2.0
ref_p                 = 1.0
compressibility          = 4.5e-5
refcoord_scaling          = com
; CONSTRAINTS
constraints              = h-bonds
constraint_algorithm      = lincs
lincs_iter               = 1
lincs_order              = 4
pbc                   = xyz   ; periodic boundary conditions
```

Commands for calculation of parameters for analysis:

```
; RMSD
gmx rms -s md.tpr -f md.xtc -o rmsd.xvg
; The backbone was selected

; Ramachandran plot
gmx rama -s md.tpr -f md.xtc -o rama.xvg

; Energy
gmx energy -f md.edr -o energy.xvg

; Secondary structure
gmx do_dssp -f md.xtc -s md.tpr -dt 100
```

Appendix D

Tables showing, for each of the 10 compounds selected in the virtual screening phase, the compounds that are 95% similar (Tanimoto coefficient) found on ZINC-22 with Cartblanche22. These compounds constitute the docking set.

Query molecule	95% similar molecule
NSC 140131_1	ZINC5085294
	ZINC5085292
	ZINC8034761
	ZINC5085293
	ZINC5085291
	ZINC195497425
	ZINC195497419
	ZINC195497422
	ZINC195497415
	ZINC101416916
	ZINC584567037
	ZINC242498472
	ZINC242498473
	ZINC101416910
	ZINC242498474

Query molecule	95% similar molecule
NSC 111702	ZINC4994399
	ZINC4994397
	ZINC6200569
	ZINC17424869
	ZINC256824146
	ZINC256824144
	ZINC256824148
	ZINC4994398
	ZINC4994396
	ZINC1703355

Query molecule	95% similar molecule
NSC 287050_1	ZINC59065544
	ZINC3861281
	ZINC59065540
	ZINC6495395
	ZINC1532813
	ZINC3954528
	ZINC12501243
	ZINC44608692
	ZINC101184237
	ZINC3861280
	ZINC29309317
	ZINC5225021
	ZINC1532814
	ZINC3606246
	ZINC3870036
	ZINC3870035
	ZINC2043005
	ZINC15206242
	ZINC2042980
	ZINC1532815
	ZINC6490946
	ZINC44608006
	ZINC1532676
	ZINC12888359
	ZINC38282241
	ZINC5225024
	ZINC2042981
	ZINC1532816
	ZINC2047187
	ZINC4090206
	ZINC1532677
	ZINC3870034

Query molecule	95% similar molecule
NSC 12161_1	ZINC4018123
	ZINC4018121
	ZINC8603212
	ZINC13831597
	ZINC4082270
	ZINC28090442
	ZINC13546632
	ZINC4018122
	ZINC4018120
	ZINC1531100
	ZINC3953841
	ZINC6283728

Query molecule	95% similar molecule
NSC 133118	ZINC16951333
	ZINC4964037
	ZINC16951334
	ZINC166589324
	ZINC105219439
	ZINC105219446
	ZINC4964039
	ZINC4964034
	ZINC1720023

Query molecule	95% similar molecule
NSC 143099	--

Query molecule	95% similar molecule
NSC 188491_1	ZINC5011744
	ZINC5011742
	ZINC13786330
	ZINC43178883
	ZINC105301382
	ZINC263611392
	ZINC5011743
	ZINC5011741
	ZINC3954489

Query molecule	95% similar molecule
NSC 52902	ZINC4722016
	ZINC4722017
	ZINC1684223
	ZINC3954052
	ZINC17313269
	ZINC105061936
	ZINC17313271
	ZINC105061931

Query molecule	95% similar molecule
NSC 86005_3	ZINC195472092
	ZINC195472104
	ZINC102310094
	ZINC257157676
	ZINC257157677
	ZINC257157678
	ZINC257157675
	ZINC195472098
	ZINC195472108
	ZINC100133014
	ZINC100133012
	ZINC100133022
	ZINC100133015

Query molecule	95% similar molecule
NSC 2561_1	ZINC4403650
	ZINC4403651
	ZINC4403653
	ZINC36386172
	ZINC36373710
	ZINC4403652
	ZINC402451
	ZINC34570732
	ZINC3953805
	ZINC1319651
	ZINC44699134
	ZINC34428856
	ZINC12670933
	ZINC44699135
	ZINC40834470
	ZINC6096442

Bibliography

[1] "Mulmy @ Seer.Cancer.Gov," *Surviellance Epidemiology and End Results Program. Multiple Myeloma Cancer facts*. 2019, [Online]. Available: https://seer.cancer.gov/statfacts/html/mulmy.html.

[2] M. A. Chapman *et al.*, "Initial genome sequencing and analysis of multiple myeloma.," *Nature*, vol. 471, no. 7339, pp. 467–472, Mar. 2011, doi: 10.1038/nature09837.

[3] P. H. Hoang et al., "Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms.," Leukemia, vol. 32, no. 11, pp. 2459–2470, Nov. 2018, doi: 10.1038/s41375-018-0103-3.

[4] M. Lionetti et al., "Molecular spectrum of BRAF, NRAS and KRAS gene mutations in plasma cell dyscrasias: implication for MEK-ERK pathway activation.," Oncotarget, vol. 6, no. 27, pp. 24205–24217, Sep. 2015, doi: 10.18632/oncotarget.4434.

[5] R. A. Kyle et al., "Long-Term Follow-up of Monoclonal Gammopathy of Undetermined Significance," N. Engl. J. Med., vol. 378, no. 3, pp. 241–249, 2018, doi: 10.1056/nejmoa1709974.

[6] L. J. M. 3rd Robert A Kyle 1, Terry M Therneau, S Vincent Rajkumar, Janice R Offord, Dirk R Larson, Matthew F Plevak, "A Long-Term Study of Prognosis in Monoclonal Gammopathy of," N. Engl. J. Med., vol. 346, no. 8, pp. 564–569, 2002.

[7] D. E. Joshua, C. Bryant, C. Dix, J. Gibson, and J. Ho, "Biology and therapy of multiple myeloma," Med. J. Aust., vol. 210, no. 8, pp. 375–380, 2019, doi: 10.5694/mja2.50129.

[8] A. L. Garfall, "New Biological Therapies for Multiple Myeloma," Annu. Rev. Med., vol. 75, pp. 13–29, 2024, doi: 10.1146/annurev-med-050522-033815.

[9] O. Sogbein, P. Paul, M. Umar, A. Chaari, V. Batuman, and R. Upadhyay, "Bortezomib in cancer therapy: Mechanisms, side effects, and future proteasome inhibitors," Life Sci., vol. 358, p. 123125, 2024, doi: https://doi.org/10.1016/j.lfs.2024.123125.

[10] A. J. Yee, "The role of carfilzomib in relapsed/refractory multiple myeloma.," Ther. Adv. Hematol., vol. 12, p. 20406207211019612, 2021, doi: 10.1177/20406207211019612.

[11] P. Moreau et al., "Oral Ixazomib, Lenalidomide, and Dexamethasone for Multiple Myeloma," N. Engl. J. Med., vol. 374, no. 17, pp. 1621–1634, 2016, doi: 10.1056/nejmoa1516282.

[12] M. Costacurta, J. He, P. E. Thompson, and J. Shortt, "Molecular Mechanisms of Cereblon-Interacting Small Molecules in Multiple Myeloma Therapy.," J. Pers. Med., vol. 11, no. 11, Nov. 2021, doi: 10.3390/jpm11111185.

[13] T. H. Patel, F. van Rhee, and S. Al Hadidi, "Cereblon E3 Ligase Modulators Mezigdomide and Iberdomide in Multiple Myeloma," Clin. Lymphoma Myeloma

Leuk., vol. 24, no. 11, pp. 762–769, 2024, doi: https://doi.org/10.1016/j.clml.2024.06.004.

[14] S. Sinha et al., "Impact of dexamethasone responsiveness on long term outcome in patients with newly diagnosed multiple myeloma.," Br. J. Haematol., vol. 148, no. 6, pp. 853–858, Mar. 2010, doi: 10.1111/j.1365-2141.2009.08023.x.

[15] P. Sonneveld et al., "Daratumumab, Bortezomib, Lenalidomide, and Dexamethasone for Multiple Myeloma," N. Engl. J. Med., vol. 390, no. 4, pp. 301–313, 2024, doi: 10.1056/nejmoa2312054.

[16] N. J. Bahlis et al., "Daratumumab plus lenalidomide and dexamethasone in relapsed/refractory multiple myeloma: extended follow-up of POLLUX, a randomized, open-label, phase 3 study," Leukemia, vol. 34, no. 7, pp. 1875–1884, 2020, doi: 10.1038/s41375-020-0711-6.

[17] L. S. Boussi, Z. M. Avigan, and J. Rosenblatt, "Immunotherapy for the treatment of multiple myeloma," Front. Immunol., vol. 13, no. October, pp. 1–11, 2022, doi: 10.3389/fimmu.2022.1027385.

[18] S. L. Lim et al., "Proteolysis targeting chimeric molecules as therapy for multiple myeloma: Efficacy, biomarker and drug combinations," Haematologica, vol. 104, no. 6, pp. 1209–1220, 2019, doi: 10.3324/haematol.2018.201483.

[19] X. Zhang et al., "Protein targeting chimeric molecules specific for bromodomain and extra-terminal motif family proteins are active against pre-clinical models of multiple myeloma.," Leukemia, vol. 32, no. 10, pp. 2224–2239, Oct. 2018, doi: 10.1038/s41375-018-0044-x.

[20] S. Su et al., "Potent and Preferential Degradation of CDK6 via Proteolysis Targeting Chimera Degraders.," J. Med. Chem., vol. 62, no. 16, pp. 7575–7582, Aug. 2019, doi: 10.1021/acs.jmedchem.9b00871.

[21] S. Yamanaka et al., "Lenalidomide derivatives and proteolysis-targeting chimeras for controlling neosubstrate degradation," Nat. Commun., vol. 14, no. 1, pp. 1–18, 2023, doi: 10.1038/s41467-023-40385-9.

[22] O. Champion et al., "BCLXL PROTAC degrader DT2216 targets secondary plasma cell leukemia addicted to BCLXL for survival.," Front. Oncol., vol. 13, p. 1196005, 2023, doi: 10.3389/fonc.2023.1196005.

[23] L. Zhao, J. Zhao, K. Zhong, A. Tong, and D. Jia, "Targeted protein degradation: mechanisms, strategies and application," Signal Transduct. Target. Ther., vol. 7, no. 1, 2022, doi: 10.1038/s41392-022-00966-4.

[24] M. H. Glickman and A. Ciechanover, "The ubiquitin-proteasome proteolytic pathway: Destruction for the sake of construction," Physiol. Rev., vol. 82, no. 2, pp. 373–428, 2002, doi: 10.1152/physrev.00027.2001.

[25] N. Chondrogianni and E. S. Gonos, Structure and function of the ubiquitin-proteasome system: Modulation of components, 1st ed., vol. 109. Elsevier Inc., 2012.

[26] D. Finley, “Recognition and processing of ubiquitin-protein conjugates by the proteasome,” Annu. Rev. Biochem., vol. 78, pp. 477–513, 2009, doi: 10.1146/annurev.biochem.78.081507.101607.

[27] J. R. Skaar and M. Pagano, “Control of cell growth by the SCF and APC/C ubiquitin ligases,” Curr. Opin. Cell Biol., vol. 21, no. 6, pp. 816–824, 2009, doi: 10.1016/j.ceb.2009.08.004.

[28] A. Hershko and A. Ciechanover, “The ubiquitin-proteasome system,” vol. 31, no. March, pp. 137–155, 2006.

[29] P. J. Teoh and W. J. Chng, “P53 abnormalities and potential therapeutic targeting in multiple myeloma,” Biomed Res. Int., vol. 2014, 2014, doi: 10.1155/2014/717919.

[30] K. M. Sakamoto, K. B. Kim, A. Kumagai, F. Mercurio, C. M. Crews, and R. J. Deshaies, “Protacs: Chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation,” Proc. Natl. Acad. Sci. U. S. A., vol. 98, no. 15, pp. 8554–8559, 2001, doi: 10.1073/pnas.141230798.

[31] X. Wang et al., “Annual review of PROTAC degraders as anticancer agents in 2022,” Eur. J. Med. Chem., vol. 267, no. January, 2024, doi: 10.1016/j.ejmech.2024.116166.

[32] M. Pettersson and C. M. Crews, “PROteolysis TArgeting Chimeras (PROTACs) — Past, present and future,” Drug Discov. Today Technol., vol. 31, pp. 15–27, 2019, doi: 10.1016/j.ddtec.2019.01.002.

[33] M. Xiao, J. Zhao, Q. Wang, J. Liu, and L. Ma, “Recent Advances of Degradation Technologies Based on PROTAC Mechanism,” Biomolecules, vol. 12, no. 9, pp. 1–16, 2022, doi: 10.3390/biom12091257.

[34] Q.-H. Chen, E. Munoz, and D. Ashong, “Insight into Recent Advances in Degrading Androgen Receptor for Castration-Resistant Prostate Cancer.,” Cancers (Basel)., vol. 16, no. 3, Feb. 2024, doi: 10.3390/cancers16030663.

[35] J. Li and J. Liu, “PROTAC: A Novel Technology for Drug Development**,” ChemistrySelect, vol. 5, no. 42, pp. 13232–13247, 2020, doi: 10.1002/slct.202003162.

[36] X. Li, W. Pu, Q. Zheng, M. Ai, S. Chen, and Y. Peng, “Proteolysis-targeting chimeras (PROTACs) in cancer therapy,” Mol. Cancer, vol. 21, no. 1, pp. 1–30, 2022, doi: 10.1186/s12943-021-01434-3.

[37] Q. Yang, J. Zhao, D. Chen, and Y. Wang, “E3 ubiquitin ligases: styles, structures and functions,” Mol. Biomed., vol. 2, no. 1, 2021, doi: 10.1186/s43556-021-00043-2.

[38] Y. Liu et al., “Expanding PROTACtable genome universe of E3 ligases,” Nat. Commun., vol. 14, no. 1, 2023, doi: 10.1038/s41467-023-42233-2.

[39] F. Eisele and D. H. Wolf, “Degradation of misfolded protein in the cytoplasm is mediated by the ubiquitin ligase Ubr1,” FEBS Lett., vol. 582, no. 30, pp. 4143–4146, 2008, doi: 10.1016/j.febslet.2008.11.015.

[40] NCBI, “Gene @ Www.Ncbi.Nlm.Nih.Gov.” 2016, [Online]. Available:

http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=3265.

[41] A. Varshavsky, “The N-end rule pathway and regulation by proteolysis,” Protein Sci., vol. 20, no. 8, pp. 1298–1345, 2011, doi: 10.1002/pro.666.

[42] “E3_detail @ hanlaboratory.com.” [Online]. Available: https://hanlaboratory.com/E3Atlas/E3_detail?E3=UBR1.

[43] A. V. Sadybekov and V. Katritch, “Computational approaches streamlining drug discovery,” Nature, vol. 616, no. 7958, pp. 673–685, 2023, doi: 10.1038/s41586-023-05905-z.

[44] H. M. Berman and S. K. Burley, “Protein Data Bank (PDB): Fifty-three years young and having a transformative impact on science and society,” Q. Rev. Biophys., vol. 58, 2025, doi: 10.1017/S0033583525000034.

[45] J. Carlsson and A. Luttens, “Structure-based virtual screening of vast chemical space as a starting point for drug discovery,” Curr. Opin. Struct. Biol., vol. 87, p. 102829, 2024, doi: 10.1016/j.sbi.2024.102829.

[46] J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.

[47] J. Abramson et al., “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” Nature, vol. 630, no. 8016, pp. 493–500, 2024, doi: 10.1038/s41586-024-07487-w.

[48] “38fcb9fba6f08719adbd762602f014b5ba4f6627 @ www.biosolveit.de.” [Online]. Available: https://www.biosolveit.de/2025/03/27/enamines-real-space-march-2025-update-now-76-billion/#:~:text=Enamine’s REAL Space March 2025,BioSolveIT.

[49] “f887789115148e1d9304d27fac9a40fbfa03cd9d @ www.biosolveit.de.” [Online]. Available: https://www.biosolveit.de/2025/03/13/chemriya-space-update-55-billion-molecules-unlocking-new-frontiers-in-drug-discovery/.

[50] B. I. Tingle et al., “ZINC-22─A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery,” J. Chem. Inf. Model., vol. 63, no. 4, pp. 1166–1176, 2023, doi: 10.1021/acs.jcim.2c01253.

[51] “national-cancer-institute-library @ scs.illinois.edu.” [Online]. Available: https://scs.illinois.edu/national-cancer-institute-library.

[52] B. J. Bender et al., “A practical guide to large-scale docking,” vol. 16, no. 10, pp. 4799–4832, 2022, doi: 10.1038/s41596-021-00597-z.A.

[53] M. S. Badar, S. Shamsi, J. Ahmed, and M. A. Alam, “Molecular Dynamics Simulations: Concept, Methods, and Applications,” Integr. Sci., vol. 5, pp. 131–151, 2022, doi: 10.1007/978-3-030-94651-7_7.

[54] X. Wu, L. Y. Xu, E. M. Li, and G. Dong, “Application of molecular dynamics simulation in biomedicine,” Chem. Biol. Drug Des., vol. 99, no. 5, pp. 789–800, 2022,

doi: 10.1111/cbdd.14038.

[55] M. Didandeh, A. H. Souderjani, and M. Asgari, “Applications of molecular dynamics simulation in nanomedicine,” Nanomedicine Technol. Appl., pp. 397–405, 2023, doi: 10.1016/B978-0-12-818627-5.00007-5.

[56] “Search @ Www.Rcsb.Org.” [Online]. Available: https://www.rcsb.org/search?q=citation.rcsb_authors:Chen, X.

[57] E. Matta-Camacho, G. Kozlov, F. F. Li, and K. Gehring, “Structural basis of substrate recognition and specificity in the N-end rule pathway,” Nat. Struct. Mol. Biol., vol. 17, no. 10, pp. 1182–1187, 2010, doi: 10.1038/nsmb.1894.

[58] K. Lindorff-Larsen et al., “Improved side-chain torsion potentials for the Amber ff99SB protein force field.,” Proteins, vol. 78, no. 8, pp. 1950–1958, Jun. 2010, doi: 10.1002/prot.22711.

[59] M. A. D. and G. G. Jack A. Tuszynski, “Introduction to Computational Drug Discovery Bioinformatics,” pp. 1–277.

[60] UniProt, “Entry @ Www.Uniprot.Org.” [Online]. Available: https://www.uniprot.org/uniprotkb/Q8IWV7/entry.

[61] F. Madeira et al., “The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024,” Nucleic Acids Res., vol. 52, no. W1, pp. W521–W525, 2024, doi: 10.1093/nar/gkae241.

[62] Z. Xia et al., “A Review of Parallel Implementations for the Smith–Waterman Algorithm,” Interdiscip. Sci. – Comput. Life Sci., vol. 14, no. 1, pp. 1–14, 2022, doi: 10.1007/s12539-021-00473-0.

[63] A. Waterhouse et al., “SWISS-MODEL: Homology modelling of protein structures and complexes,” Nucleic Acids Res., vol. 46, no. W1, pp. W296–W303, 2018, doi: 10.1093/nar/gky427.

[64] “Q8IWV7 @ alphafold.ebi.ac.uk.” [Online]. Available: https://alphafold.ebi.ac.uk/entry/Q8IWV7.

[65] G. Jones et al., “Elucidation of protein function using computational docking and hotspot analysis by ClusPro and FTMap,” Acta Crystallogr. Sect. D Struct. Biol., vol. 78, pp. 690–697, 2022, doi: 10.1107/S2059798322002741.

[66] C. H. Ngan, D. R. Hall, B. Zerbe, L. E. Grove, D. Kozakov, and S. Vajda, “FtSite: High accuracy detection of ligand binding sites on unbound protein structures,” Bioinformatics, vol. 28, no. 2, pp. 286–287, 2012, doi: 10.1093/bioinformatics/btr651.

[67] “pdb2pqr @ server.poissonboltzmann.org.” [Online]. Available: https://server.poissonboltzmann.org/pdb2pqr.

[68] “b683cca10f8f258a5d433f16f154215f664512bf @ pdb2pqr.readthedocs.io.” [Online]. Available: https://pdb2pqr.readthedocs.io/en/latest/.

[69] E. Jurrus et al., “Improvements to the APBS biomolecular solvation software suite,” Protein Sci., vol. 27, no. 1, pp. 112–128, 2018, doi: 10.1002/pro.3280.

[70] D. Giordano, C. Biancaniello, M. A. Argenio, and A. Facchiano, “Drug Design by Pharmacophore and Virtual Screening Approach,” Pharmaceuticals, vol. 15, no. 5, pp. 1–16, 2022, doi: 10.3390/ph15050646.

[71] “vialed-plated-compounds-0 @ dctd.cancer.gov.” [Online]. Available: https://dctd.cancer.gov/drug-discovery-development/reagents-materials/vialed-plated-compounds-0.

[72] V. R. Pothineni et al., “Screening of NCI-DTP library to identify new drug candidates for Borrelia burgdorferi,” J. Antibiot. (Tokyo)., vol. 70, no. 3, pp. 308–312, 2017, doi: 10.1038/ja.2016.131.

[73] “compound-2 @ scs.illinois.edu.” [Online]. Available: https://scs.illinois.edu/resources/cores-scs-research-and-service-facilities/high-throughput-screening-facility/compound-2.

[74] “2c03a5750a3821958dcc10a1f163feae6722740a @ www.schrodinger.com.” [Online]. Available: https://www.schrodinger.com/platform/products/ligprep/.

[75] S. L. Dixon, A. M. Smondyrev, E. H. Knoll, S. N. Rao, D. E. Shaw, and R. A. Friesner, “PHASE: A new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results,” J. Comput. Aided. Mol. Des., vol. 20, no. 10–11, pp. 647–671, 2006, doi: 10.1007/s10822-006-9087-6.

[76] “calculators_playground @ docs.chemaxon.com.” [Online]. Available: https://docs.chemaxon.com/display/docs/calculators_playground.md.

[77] A. Castaño and M. S. Maurer, “Protonation and pK changes in protein-ligand binding,” Q. Rev. Biophys., vol. 20, no. 2, pp. 163–178, 2015, doi: 10.1017/S0033583513000024.Protonation.

[78] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings,” Adv. Drug Deliv. Rev., vol. 23, no. 1–3, pp. 3–25, 1997, doi: 10.1016/S0169-409X(96)00423-1.

[79] W. Jorgensen and E. M.Duffy, “Prediction of drug solubility from structure,” Elsevier Sci. B.V., 2002.

[80] M. J. Waring, “Lipophilicity in drug discovery,” Expert Opin. Drug Discov., vol. 5, no. 3, pp. 235–248, 2010, doi: 10.1517/17460441003605098.

[81] “calculators_herg @ docs.chemaxon.com.” [Online]. Available: https://docs.chemaxon.com/display/docs/calculators_herg.md.

[82] F. Melnikov, L. T. Anger, and C. Hasselgren, “Toward Quantitative Models in Safety Assessment: A Case Study to Show Impact of Dose–Response Inference on hERG

Inhibition Models,” Int. J. Mol. Sci., vol. 24, no. 1, 2023, doi: 10.3390/ijms24010635.

[83] K. Patidar et al., “An in silico approach to identify high affinity small molecule targeting m-TOR inhibitors for the clinical treatment of breast cancer,” Asian Pacific J. Cancer Prev., vol. 20, no. 4, pp. 1229–1241, 2019, doi: 10.31557/APJCP.2019.20.4.1229.

[84] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” J. Chem. Inf. Model., vol. 50, no. 5, pp. 742–754, 2010, doi: 10.1021/ci100050t.

[85] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, “Molecular fingerprint similarity search in virtual screening,” Methods, vol. 71, no. C, pp. 58–63, 2015, doi: 10.1016/j.ymeth.2014.08.005.

[86] “master @ github.com.” [Online]. Available: https://github.com/gnina/gnina/tree/master.

[87] J. Cheminform et al., “GNINA 1 . 0 : molecular docking with deep learning,” J. Cheminform., pp. 1–20, 2021, doi: 10.1186/s13321-021-00522-2.

[88] M. Aminpour et al., “Computational determination of toxicity risks associated with a selection of approved drugs having demonstrated activity against COVID-19,” vol. 5, pp. 1–20, 2021.

[89] “help @ swissmodel.expasy.org.” [Online]. Available: https://swissmodel.expasy.org/docs/help?

[90] R. A. Laskowski, N. Furnham, and J. M. Thornfon, “The Ramachandran plot and protein structure validation,” pp. 62–75, 2013.

[91] B. Honig and A. Nicholls, “Classical Electrostatics in Biology and Chemistry,” vol. 268, no. 5214, pp. 1144–1149, 2015.

[92] “Ligand-Based Virtual Screening Using Phase,” 2025.

[93] R. Vannam et al., “Article Targeted degradation of the enhancer lysine acetyltransferases CBP and p300 ll ll Article Targeted degradation of the enhancer lysine acetyltransferases CBP and p300,” Cell Chem. Biol., vol. 28, no. 4, pp. 503-514.e12, 2021, doi: 10.1016/j.chembiol.2020.12.004.