Politecnico di Torino

Polytechnic of Turin

Master's Degree in Biomedical Engineering

A.a 2021/2022

Graduate session December 2022

Rationalizing the warhead of the PROteolysis TArgeting Chimera (PROTAC) against γ-Tubulin 1

Supervisors

Prof. Jack TUSZYNSKI

Prof. Maral AMINPOUR

Candidate

Fabiano ALTIERI

Abstract

PROteolysis-TArgeting-Chimera (PROTAC) technology is a recent unconventional way to target and destroy tumour cells, which exploits biological processes instead of inhibiting the function of a protein, as typical of conventional drugs.
Essentially, PROTAC drugs consist of three parts: the warhead's ligand, which binds to the Protein Of Interest (POI); the E3 ligand, which binds to E3 ligase and a linker, which links the warhead and the E3 ligand covalently.

Moreover, PROTAC drugs aim to accelerate the degradation rate of the target proteins since they have the goal of keeping POI and E3 ligase close together for as long as possible. Thus they accelerate the ubiquitination process and, consequently, the degradation mediated by proteasome.
The main technical advantage, which is impossible to achieve with inhibitory drugs, is the possibility of using any compound, even biologically inert ones, that bind to any part of the protein and is no longer just the active site.

However, rationalizing each part of the PROTAC is critical because incorrect ligand design could lead to the degradation of unwanted targets, causing much more unpredictable effects than inhibition of protein's functions. After all, in the latter case, the Protein-Protein Interactions (PPIs) might still be working, as the inhibited POI is still present inside the cell or on its membrane.

The current work is the first step in rationalizing PROTACs that, through the E3 ligase UBR1, lead to the degradation of γ-Tubulin (γT); in particular, focuses more on the warhead design (i.e. the ligand that has to bind the γT selectively).

The γT is one of the central proteins involved in Microtubule nucleation. It is particularly over-expressed in Glioblastoma Multiforme, a brain tumour for which no treatment exists.

Since no known ligands bind the γT with solid evidence, a possible solution is virtual screening (VS) of multi-billion compounds databases such as the free access ZINC20 database.

After preparing the POI appropriately via Molecular Dynamics and identifying potential binding sites, two approaches to do Virtual Screening (VS) on databases are applied: (1) conventional approach (pre-processing of databases $\rightarrow$ conformational sampling $\rightarrow$ energy minimization $\rightarrow$ pharmacophore filtering $\rightarrow$ consensus docking) and a (2) new VS approach based on DeepDock and Docking.

However, the conventional approach is very computationally expensive, even just for databases of millions of compounds, so a small dataset of compounds of biological origin has been used.

On the other hand, DeepDock is a recent package which is essentially a mix of Ligand-Based Virtual Screening (LBVS) and Structure-Based Virtual Screening (SBVS): it is indeed a Quantitative Structure-Activity Relationships (QSAR) based on Deep Neural Network (DNN) models trained on docking scores of a small subset of a large database to predict docking scores for the rest.

In previous works, it has been shown to be a very effective and promising method to reduce the size of multi-billion compounds databases by more than 99%.
In addition, it enriches it with top-ranked hits, avoiding significant loss of favourable virtual hits so that standard docking is performed using what remains.

Table of Contents

Acronyms

α-/β-T α-/β-tubulin

αT α-Tubulin

βT β-Tubulin

γT γ-Tubulin

γTuRC γ-Tubulin Ring Complex

γTuSC γ-Tubulin Small Complex

ADP Adenosine Di-Phosphate

ADMET Absorption, Distribution, Metabolism, Excretion and Toxicity

AI artificial intelligence

ATP Adenosine Tri-Phosphate

AUC Area Under the ROC Curve

BARD1 Brca1-Associated Ring Domain 1

BRCA1 Breast cancer type 1 susceptibility protein

CADD computer-aided drug design

CD Consensus Docking

CRBN Cereblon

CRL Cullin RING ligase

CUL Cullin

CUL2 Cullin 2

CS Consensus Scoring

DL Deep Learning

DNN Deep Neural Network

E1 Ubiquitin-activating enzymes

E2 Ubiquitin-conjugating enzyme

E3 Ubiquitin-Ligase enzyme

EM Energy Minimization

ESIs E3-Substrate Interations

FP FingerPrint

FPDE Full Predicted Database Enrichment

FPR False Positive Rate

GCP γ-Tubulin Complex Proteins

GDP Guanosine Di-Phosphate

GTP Guanosine Tri-Phosphate

HECT Homologous to the E6AP Carboxyl Terminus

HPC High Performances Computing

HTS High-Throughput Screening

LBVS Ligand-Based Virtual Screening

MD Molecular Dynamics

ML Machine Learning

MOE Molecular Operating Environment

MTs microtubules

MT microtubule

MTOCs MT-organizing centers

PCM Pericentriolar Matrix

PLB Propensity Ligand Binding

PME Particle-Mesh Ewald

POI Protein Of Interest

PPIs Protein-Protein Interactions

PROTAC PROteolysis-TArgeting-Chimera

PTMs post-translational modifications

QSAR Quantitative Structure-Activity Relationships

RING Really Interesting New Gene

RMSD Root-Mean-Square Deviation

ROC Receiver Operating Characteristic

SARS-CoV-2 Severe Acute Respiratory Syndrome CoronaVirus 2

SASA Solvent Accessible Surface Area

SBCD Score-based Consensus Docking

SBVS Structure-Based Virtual Screening

smPROTAC small-molecule-based PROTAC

RBR RBR (RING-between-RING)

RBX1 RING-Box protein 1

TPR True Positive Rate

UBR1 UBiquitin-protein ligase E3 component - Recognin 1

UP Ubiquitin Protein

UPS Ubiquitination Proteasome System

VBC ElonginB–ElonginC-VHL

VHL Von Hippel-Lindau

VS Virtual Screening

ZF zinc-finger

Chapter 1

Introduction

1.1 Introduction

Most of the strategies developed in drug discovery in the last century to knock out a given protein follow the occupancy-driven paradigm. Indeed, the traditional therapeutic agents are small molecules with high affinity, and their mechanism of action mainly occupies the target protein's active site on which its functions depend [1, 2].
This strategy has proved very successful, yet is limited by the number of protein targets currently considered "druggable".
Inhibitory, agonist or other site-active ligands target a fraction of the entire human proteome ($\sim 20\%$) and, therefore, cannot target whole classes of proteins known to have a major role in disease but that do not have druggable sites. Examples of these proteins are transcription factors, scaffolding proteins, proteins with active sites covered by other proteins via strong Protein-Protein Interactions (PPIs) and other non-enzymatic proteins.

Moreover, relatively high local concentrations of conventional small molecules are often required for a long administration time to ensure therapeutic efficacy, but off-target binding and side effects may follow.
To make matters worse, the cancer cells, which are affected by a significantly higher rate of mutations than healthy cells, can develop drug resistance with a higher probability than healthy cells.
Such gene mutations can lead to structural changes in the target protein's active site, making such a site no longer druggable.
Finally, another typical problem with small molecule-based strategies is the accumulation of inhibited proteins if not adequately removed via cellular cleaning systems [1–3].

In recent years another paradigm has been increasingly adopted: employ event-driven strategies.
This alternate approach exploits existing biological machinery and pathways that somehow induce the degradation of proteins or block upstream traduction from mRNA or autophagy, to reduce the concentration of the Protein Of Interest (POI).

Other strategies aim to reduce the POI, but all are affected by several weaknesses [4, 5].

Antibodies, although they have high sensitivity and specificity, have fewer cross-reactions and low costs as B cell and bioengineered/recombinant bacteria cultures produce them, have a high molecular weight, so they mainly target proteins located at the plasma membrane due to low membrane permeability.
The best way to overcome this issue is using nanocapsules; however, they require challenging manufacturing processes that can compromise the properties of antibodies.

Short-interfering RNAs (siRNA) inhibit upstream gene expression avoiding the generation of oncogenic proteins; nevertheless, they require improvements in the delivery system because they can locate in unwanted tissues.

Other popular event-driven strategies use antisense oligonucleotides or genome editing strategies such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). Still, they are not yet used officially as therapeutic agents for several complications [2, 6].

Finally, the strategy considered in the current work is to hijack Ubiquitination Proteasome System (UPS), which is one of the systems responsible in cells for cleansing proteins, through PROteolysis-TArgeting-Chimera (PROTAC) technology [1–3, 6, 7].
The PROTAC are small molecules composed of (1) a targeting ligand, termed as *warhead* which binds selectively to the POI, (2) an E3 ligase recruitment ligand and (3) a chemical linker that connects the two ligands.
The primary mechanism of action of such compounds is the recruitment of both POI and E3 ligase and, by standing close and inducing strong PPIs, accelerate the process of POI ubiquitination. When POIs have a polyubiquitin chain recognizable by the 26S proteasome, they are finally degraded.

Many studies point out that knocking out or degrading a POI causes more notable effects than inhibiting just the functions, and the possibility of the PROTAC degrading undruggable proteins caused many pharmaceutical companies to invest in this new technology.
Examples of this growing interest are the cooperation between Arvinas and Pfizer, which are running the first clinical trial of as many three PROTACs: the ARV-110 and ARV-766 lead to degradation of Androgen Receptor (AR) in men with metastatic castration-resistant prostate cancer (mCRPC), while the ARV-471 leads to degradation of Estrogen Receptor (ER) in women with locally advanced or metastatic ER+/HER2- breast cancer (mBC) [2, 5, 8–10].
All these PROTACs are currently in phase I-II as reported in https://clinicaltrials.gov.

Since they were first made over 20 years ago, several major advantages over small molecule inhibitors have emerged that have pushed the boundaries of traditional drug discovery.
The pros include [7, 11–14]:

- similar effects as using already known knockdown techniques such as CRISPR, siRNA or short hairpin RNA (shRNA);
- sub-stoichiometric concentration, even with a magnitude of pM, is sufficient to have PROTAC still catalytically active;
- no need to use ligands with high affinity;
- no priority in designing ligands that bind to the active-/allosteric-site of the POIs. Thus inert/nonfunctional ligands can also be used.

As a consequence of the latter pros, targeting of proteins considered undruggable by conventional drugs, such as transcription factors, small lacking-hydrophobic-pocket enzymes, scaffolds and regulatory proteins, is now possible, expanding tremendously the proteome that drugs can target [3, 6, 14–17].

Another advantage of PROTAC that deserves to be discussed separately from the previous list is that it can also degrade specific isoforms and subtype proteins, i.e. proteins that share high sequence and structural similarity mainly in their binding sites but which present significant structural diversity at their surface [14].

The goal of the project of which this work is the first phase is to rationalize *in silico* several PROTACs to degrade γ-Tubulin (γT) exploiting the UBR1 as E3 ligases. Removing the γT, which is an essential component of the microtubule (MT) nucleator, can disrupt the integrity of MT to prevent or at least restrict mitotic division, with the ultimate goal of destroying cells affected by glioblastoma multiforme.

Indeed, 70% of all brain tumours are gliomas. However, due to their position and their highly invasive and infiltrative nature, the treatment of these tumours by targeting the microtubule structure is challenging.

The emergence of drug-resistant tumour cells impedes several tubulin-binding agents. In addition, the inability of many drugs to cross the blood-brain barrier and the development of drug-induced neurotoxicity prevent the use of many conventional agents targeting tubulin.

However, all tubulins have a very high sequence and structural similarity over entire structures.
Moreover, the active site of γT and β-Tubulin (βT), which is namely also as G domain, is where the Guanosine Tri-Phosphate (GTP) binds to and then hydrolyzed to Guanosine Di-Phosphate (GDP), while the active site of the α-Tubulin (αT) binds only to GDP. Both active sites are highly conserved in Eukaryota from yeast to human species and highly similar even with many other proteins of the same biological system that have the G domain embedded, like the RAt Sarcoma (RAS) and other small-GTPases [18, 19].
Thus, designing selective PROTACs is far more challenging if a ligand which binds the G domain is taken into account.

Subsequently, since no other ligands bind the γT selectively with solid evidence in scientific literature, exploring databases of compounds seemed the only possible alternative.
In the current work, it was decided to perform a large Virtual Screening (VS) with a state-of-art approach: using Deep Neural Network (DNN) models to reduce the size of the ZINC20 database, which contains more than one billion[1] compounds and enrich it with top-hits. After that, the resulting top-hits will be subjected to conventional approaches to find the best potential warhead.

After properly preparing the protein models of all tubulins via Molecular Dynamics (MD), the potential binding sites on γT are compared to the relative binding site of other tubulins to have the unique binding sites by evaluating electrostatic maps and the similarity of residues around potential binding sites.

[1]Note that billion will be abbreviated as B, million as M and thousand as K

1.2 The Ubiquitination Proteasome System (UPS)

One of the ways in which cells maintain cellular protein homeostasis and regulate numerous processes, such as gene transcription, DNA pairing, cell cycle control, and apoptosis, is mediated by the UPS that engages the protein-degradation machinery proteasome [20–23].
The proteasomes are highly evolutionarily conserved among eukaryotes and degrade proteins involved in the processes mentioned above, such as cell surface receptors, growth factor receptors, transcription factors, tumour suppressor factors such as p53[2], oncogene products, spindle-bound proteins, misfolded proteins, and other intracellular and nuclear proteins that, if not degraded, most of them lead to the pathogenesis of many diseases [2, 5, 20–22, 24].

The UPS relies on Adenosine Tri-Phosphate (ATP) as "fuel" and high amount of PPIs and consists of the following steps (Fig. 1.1)
[2, 21–23, 25–31]:

1. Activation of a single UP, a small protein of 76 aminoacids, is mediated by Ubiquitin-activating enzymes (E1) in an ATP-dependent reaction in which the CYS residue on its active site forms a thioester bond, which is energetically unstable, with the C-terminus GLY76 residue of UP.
2. The C-terminus GLY76 residue of the activated UP forms a high-energy thioester bond[3] with the CYS residue of the active site of the Ubiquitin-conjugating enzyme (E2).
   In some cases, E2 binds directly to the POI, transferring the UP to it [32].
3. In the presence of the POI, the Ubiquitin-Ligase enzyme (E3), which is typically a large complex with a "clamp" structure consisting of substrate adaptors and

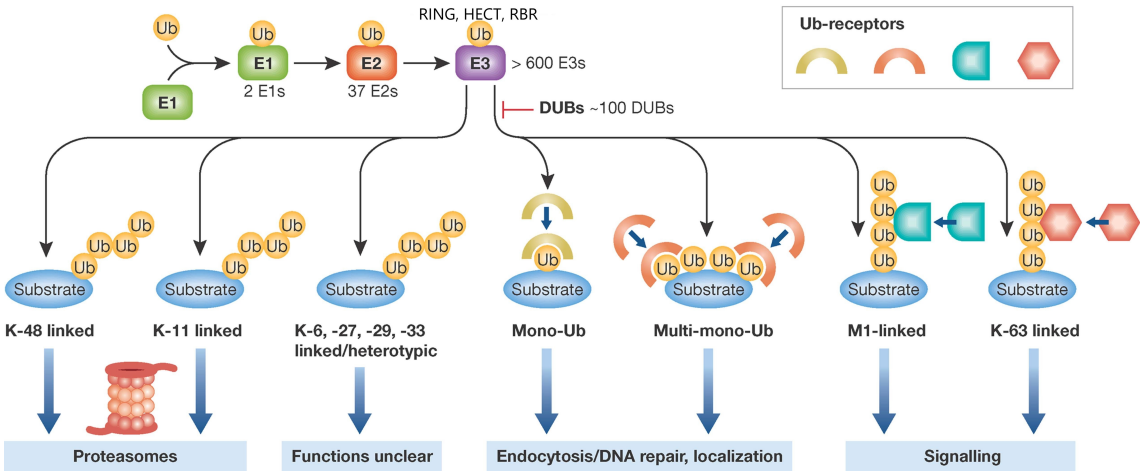


Figure 1.1 – *The steps of the ubiquitination process based on E3 class type and the different pathways based on the rearrangement of Ubiquitin Protein (UP)s. (S.Fulda, 2012)*

[2]Tumor protein p53 is a transcription factor that regulates DNA repair, cell cycle, and apoptosis. His mutation is found in more than 50% of cancers

[3]This mechanism is called *trans-thioesterification*

accessory proteins (An example can be found in Fig. 1.5), transfers the UP from the temporary complex E2-UP to the residue LYS of the POI, which forms an isopeptide bond[4] with the C-terminus residue of UP.

4. After the first ubiquitination of protein, the pathways can be different depending on the UP-site of POI, the type of E2 and E3, and even which residue of UP is involved (K6, K11, K27, K29, K33, K48, K63 and M1); in fact, the type of the mechanism described above can be mono-, multi-mono- or poly-ubiquitination (Fig. 1.1).

The different ubiquitination modes lead to many pathways that include: chromosome cohesion/segregation, DNA repair, replication, transcription, translation, subcellular localization, endocytic trafficking, G protein regulation, PPIs modifications, inflammation, glucogenesis, cell cycle and division, differentiation, cell migration, apoptosis regulation, subunit replacements, misfolded protein signalling, spermatogenesis and neurogenesis, oxygen/NO sensing, and proteasomal or lysosomal degradation [2, 22, 27–31, 33, 34].

What makes the ubiquitination process even more complex is the presence of deubiquitylases (DUBs), which remove UPs from the proteins or modify the poly-UP chains already bound [34]. However, it deserves to be investigated separately from the following work.

In the case of UPS-mediated degradation, the protein must be ubiquitinated with a poly-chain of at least four UPs linked together so that the C-terminus of the previous UP (GLY76) is bound to the LYS48 residue of the next UP (Fig. 1.2D) [2, 21, 33].

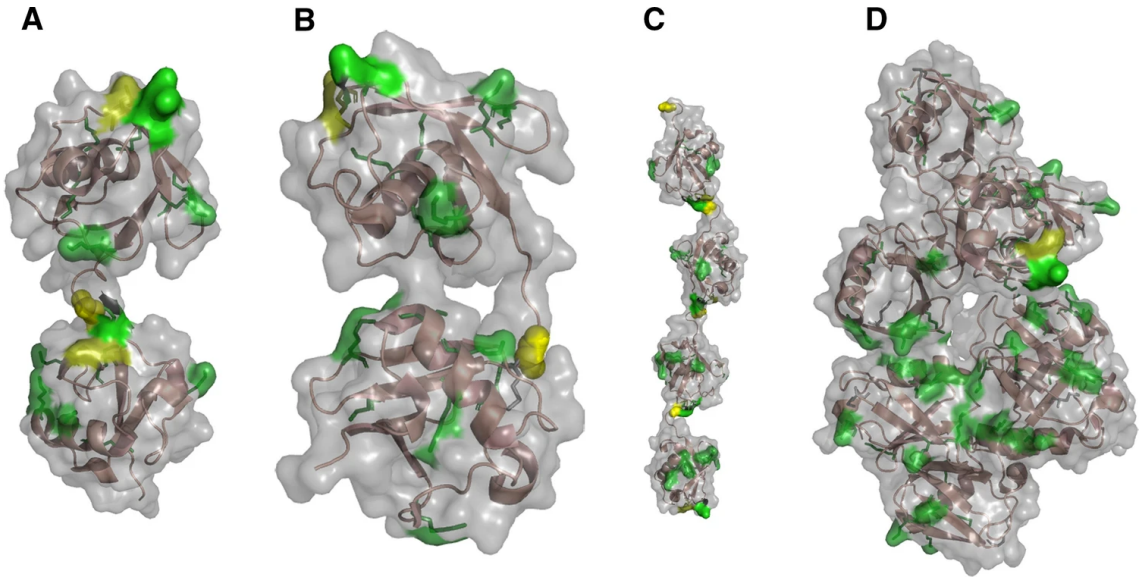


Figure 1.2 – *(A) K63-linked diubiquitin; (B) M1-linked diubiquitin; (C) K63-linked tetraubiquitin; (D) K48-linked tetraubiquitin. (A.Dòsa 2022)*

[4]A peptide bond between NH2 amino group of the side chain, like LYS, with a carboxyl group of another sidechain or C-terminus residue

The poly-ubiquitinated protein will now be recognized by the cap-like regulatory subunit of 26S proteasome, a large barrel-shaped multi-subunit protein complex consisting of six proteolytic sites (Fig. 1.3).

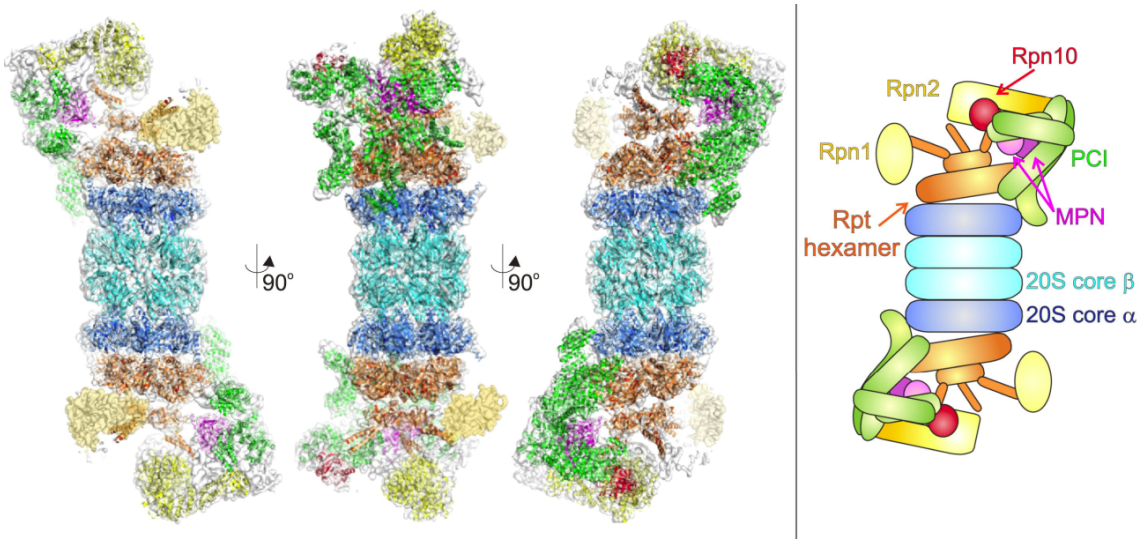


Figure 1.3 – *Molecular model and scheme of the 26S proteasome, where the main subunit domains are color coded. (P.C.A. da Fonseca, 2012)*

At this point, the UP-tagged protein is transported to the cylindrical core of the 20S, where it is hydrolyzed into oligopeptides by various enzymes and finally released from the proteasome [21–23]. On the other hand, the UP molecules dissociate from the substrate and return to the cytoplasm for re-utilization.

It is interesting noting that with a slight difference in the residue position, the LYS63-linked polyubiquitin chain leads instead to a self-digesting mechanism (i.e. autophagy) (Fig. 1.2A,C).
This big pathway difference may be derived from the significant structural difference between the two definitive polyubiquitin chains.
The LYS48-linked polyubiquitin is more compacted (Fig. 1.2D). In contrast, the LYS63-linked polyubiquitin is more elongated and flexible, behaving as an unfolded protein [33].

The role and the various type of e3 ligase in protein degradation

Currently, 3 E1s, 41 E2s and over 650 E3s are known to be encoded in the mammalian genome (https://www.uniprot.org).
The fact that so many E3 ligases exist suggests that they, in conjunction with the combined action of E2 enzymes, are highly specific for a wide range of substrates [5, 20, 22, 23].
In addition, the POIs have short peptide sequences, known as degrons, which are motifs that can influence protein degradation rates and may be present in more than one site on the same protein; however, degrons do not always correspond to sites of ubiquitination, rich of LYS residues, but they are often found close together [1,

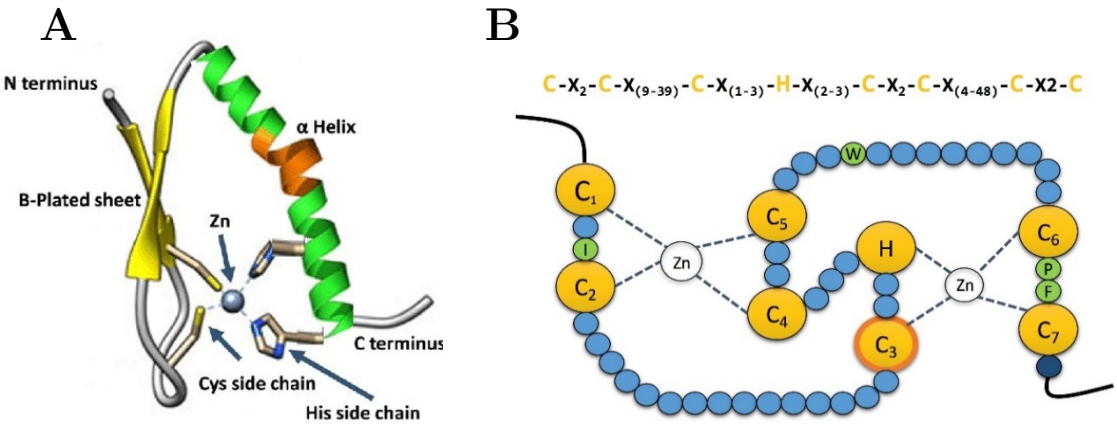


Figure 1.4 – *(A) ZF motif; (B) configuration of RING domain with 2 ZFs*

34–36].
Furthermore, degrons locate in disordered regions, but when E3 ligase binds to the protein, the degron seems to move in a more ordered region. Interestingly, degrons can be transplanted through post-translational modifications (PTMs) [37].

There are three main classes of E3 families based on their structural domains and mode of actions in transferring UP to the POI [1, 23, 26, 38, 39]:

- **Really Interesting New Gene (RING) domain**
  It is the most prevalent class in the E3 ligase family and is further divided into two subclasses: (1) the effective RING domains have a zinc-finger (ZF) motif[5] pair (Fig. 1.4B), while (2) the U-box domains adopt the same RING fold without zinc ions.
  Both domains are mainly responsible for binding E2:UP and transferring UP directly to POI, without the formation of a covalent bond [26, 38, 39].
  RING-/U-box-based E3 ligases exist as monomers[6], dimers[7], or even extensive multi-subunit assemblies such as the Cullin RING ligase (CRL).

  They all share the same features: a substrate recognition domain in C-terminus and an E2 binding RING domain in N-terminus [26, 38, 43]. E3-RING/U-box can ubiquitinate in several ways depending only on the E2 type, a condition that is not always met in other E3 types.

- **Homologous to the E6AP Carboxyl Terminus (HECT) domain**

[5]The ZF motif consists of a quartet of CYS/HIS which coordinate a Zn^{2+} ion (Fig. 1.4A). Between two secondary structures (alpha helix and beta sheet), there is an elongated loop of variable length 10÷30 aminoacids.
The ZF motif is often associated with DNA-binding domain in mammals and has been widely used in genetic engineering as a ZF nucleases: in fact, a ZF can bind to a specific DNA triplet depending on the combination of HIS and CYS residues and, when paired with other ZFs in a row and the FokI restriction endonuclease, the complex can cleave the DNA [40].

[6]The only E3 as monomer form known to have ubiquitination activity with strong evidence is the Breast cancer type 1 susceptibility protein (BRCA1) [41, 42]. See sec. 1.5 for further details

[7]In case of homodimeric RINGs/U-box, 2 E2s can simultaneously bind to two domains

E6-AP was initially identified as E3, which led to the, degradation of p53 in association with the human papillomavirus E6 oncoprotein.
The proteins with HECT domain catalyse the UP transfer to the substrate through a two-step reaction, unlike the proteins with RING domain.
The UP is first transferred from a recruiting E2:UP through contacts with the N-terminus domain to a catalytic CYS in the HECT C-terminus domain by trans-thiolation reaction, then from the HECT domain to the substrate, which is also bound with C-terminus [26, 44, 45].
Unlikely the next class and the majority ofRING, which constitute many subunits, the proteins having HECT domain have a bi-lobar architecture with a flexible hinge that allows changes in the relative orientations of the lobes during UP transfer [26, 44].
Recent studies showed that HECT proteins could interact noncovalently with a secondary UP through N-terminal contacts, promoting polyubiquitylation of the substrate [44].

- **RBR (RING-between-RING) (RBR) domain**
  RBR proteins are multidomain complex E3 and are a kind of hybrid of HECT and RING because they have at least 2 RING domains (RING1 and RING2) separated by an in-between-RING domain (IBR) and catalyze UP transfer through a two-step reaction, like HECT proteins [1, 38, 44, 46].
  Both RING1, RING2 and IBR domains each coordinate 2 Zn2+ ions.
  The RING1 and RING2 are not identical to the classic RING domain, but the former is the most RING-like among the 3 RBR subdomains because they bear many ZF-like motifs [38, 46]. Another feature is that RING1 recruits the E2:UP in a similar manner to other RING domains.
  On the other hand, the RING2 domain has a catalytic CYS site that forms a covalent bond with UP coming from RING1 via trans-thiolation, after which the UP is finally transferred to the substrate.
  As for the role of IBR, it remains still unclear, but IBR has a highly variable aminoacid sequence among different RBR-E3 ligases and does not have a CYS-based active site, unlike RING1 and RING2.
  RBR proteins can have other subunits and/or domains, determining the final structure and mechanism of action and having the highest number of E2 involved [38].

When a mutation affects one of the three domains described above, severe ubiquitination dysfunctions occur, which can lead to carcinogenicity in the cell.
It has been observed that depriving in several E3 the RING domain or altering the amino acid sequence, particularly at the CYS/HIS residues of ZFs, or chelating Zn ion, there is no binding between E3 ligase and E2:UP and therefore, the transfer of UP to the POI is no longer allowed [26, 27, 33, 47–49].

Lastly, other more complex phenomena that interfere with the E3 ligase activity occur: for example, there are specialized proteins that "wrap" around the scaffold part of the E3, inhibiting the "clamping" effect. Thus the ubiquitination of the POI from the E2-binding subunit [50].
One of these E3 inhibitor proteins is the Cand1 which has a super-helical structure

consisting of 27 tandem HEAT[8] repeats, each of 40 aminoacids in helix secondary structure, and a preference versus the E3 ligase complex constituted of Skp1-Cul1-F box (SCF) and Roc1 subunits. Cand1 "wraps" around the Cullin (CUL)-1, but other studies found that Cand1 can "wraps" all six CULs [50].

[8]Huntingtin-Elongation-A subunit-TOR

1.3 The PROteolysis TArgeting Chimera

The first PROTACs

Sakamoto's group was the first in 2001 to develop the first proof of concept of PROTAC: a bifunctional molecule constituted of ovalicin[9] (OVA), and a 10-AA phosphorylated peptide fragment (DRHDpSGLDpSM) derived from *NF-κB*[10] *inhibitor alpha* (IκBα) [1, 2, 4, 5, 11, 51].
The IκBα is a negative regulator of the NF-κB transcription factor and is polyubiquitinated by $Skp1^{\beta-TRCP}$ (the $\beta-TRCP$ is the β-transducin repeat-containing protein three which is an F-box protein, while the Skp1 is the S-phase kinase-associated protein 1, a Cullin family member and substrate recognition component) and consequently degraded by the proteasome [11, 52]. The peptide-based PROTAC developed by Sakamoto et al. has successfully triggered the Methionine AminoPeptidase-2 (MetAP-2) ubiquitination and subsequent protea somal degradation in Xenopus egg cell extracts.

Next, the first PROTAC tested in vivo was developed by Schneekloth et al. in 2004: a bifunctional molecule comprised of (1) an artificial ligand (AP21998) which binds to the mutated (F36V) immunophilin FKBP12 and (2) a 7-AA peptide ($ALAP_{OH}YIP$), which is the minimal sequence of the hypoxia-inducible factor 1α (HIF1α) [11] [2, 5, 12, 51, 53–55].
The hydroxylated proline in the peptide is recognized by the Von Hippel-Lindau (VHL) tumor suppressor protein, which is the substrate recognition component of the E3 ligase $CRL2^{VHL}$ complex constituted of Cullin 2 (CUL2) as scaffold, RING-Box protein 1 (RBX1)[12] and ElonginB–ElonginC pair which, together with VHL, forms the VBC (Fig. 1.5) [2, 4, 5, 12, 51–57].

From these pioneering studies, many laboratories developed new peptide-based PROTACs having different POIs as targets, like huntingtin, tau, AKT, death-associated protein kinase 1 (DAPK1), scaffolding protein PSD-95, X-protein of the hepatitis B virus, and so on [1, 4, 5].
Despite the success of early experimental studies of these early PROTACs, several limitations led to them no longer being synthesized in many laboratories for two main reasons [1, 2, 4, 5, 11, 58]:

- the presence of peptides that leads to low cell permeability; indeed, in the first works, microinjections were employed to deliver the PROTACs into living cells to overcome this obstacle;

[9] An angiogenesis inhibitor which can covalently bind to the active site of methionine aminopeptidase-2 (MetAP-2)

[10] Nuclear factor kappa-light-chain-enhancer of activated B cells

[11] The HIF1α is a factor inducing erythropoietin under hypoxic conditions, whereas in normoxic conditions, the intracellular concentration of HIF1α is instead maintained at low levels by hydroxylation of specific proline residues, which leads to the UPS

[12] RBX1 has a variant RING domain at its C-terminal region: the domain is characterized by the canonical region containing 2 Zn^{2+} ions plus an extended part containing a third zinc ion coordinated by three cysteines and one histidine.

- the micromolar-range potency;
- the large size of the peptide-based PROTACs that leads to being easily recognized by the immune system for which it produces antibodies.

To overcome these limitations, researchers switched the peptide binders for POI and E3 ligase with small molecules developing the small-molecule-based PROTAC (smPROTAC).
However, these early PROTACs represented a significant opportunity to induce selective degradation of traditionally undruggable proteins, thus conceiving the possibility of rationalizing drugs based on event-driven pharmacology and no longer on occupancy-driven pharmacology.

The modern PROTACs

Nowadays, smPROTACs are constituted of a warhead (the POI binder), a linker of variable nature and an E3 ligand [1–4, 6].
The warheads and E3 ligands can be small molecules ad-hoc assembled or be already known, such as inhibitors, agonists or other commercially available compounds.
Using small molecules rather than peptidic binders has several advantages [4, 12, 51]:

- strong, specific, and biophysically validated binding affinities to their targeted E3 (i.e. shows the effective degradation of the POIs);
- acceptable physicochemical profile that can increase cell permeability, decrease the molecular weight, improve the lipophilicity, solubility;
- lack of reactive groups or metabolic hotspots;
- lack of pan-assay interference compounds (PAINS)[13] alerts;
- well-characterized structural information of their binding modes.

But the most important advantage over traditional drugs, which make up more than 98% of the pharmaceutical market, is the possibility of employing small molecules that (1) do not have a strong binding affinity, which is required instead by traditional drugs mainly to keep the administration concentration low and (2) even bind to inert pockets[14] [1, 3, 5].
The latter point is the key to their success: theoretically, targeting any protein pocket can significantly increase the fraction of the proteome that drugs can target, i.e. the *drug target space* [3, 6, 15–17].

Examples of proteins now potentially hijackable by PROTACs but considered "undruggable" by traditional drugs with corresponding reasons are given below:

- The scaffolding proteins, also called as proteic glues, such as CUL and WD40, constitute a very large portion of the proteome, but they lack catalytic activity and instead interact with other proteins via PPIs in a high phosphorylation-dependent manner [59, 60];

[13] In High-Throughput Screening (HTS), the PAINS are the false positives

[14] I.e. Sites that have shown not to change the protein's function or cell viability when bounded to the inhibitors or agonist

- The large multicomponent protein complexes are based principally on PPIs, and if there is an inhibition of one subunit, it may not be entirely deleterious to the complex function [6]. However, some evidence suggests that if the degradation of one subunit occurs, it may destabilize the entire complex leading to the degradation of the remaining subunits through the natural protein quality control machinery, a process known as bystander ubiquitination;
- The RAt Sarcoma (RAS) proteins, which are part of the small GTPase family and often associated with cancers due to their critical roles in transmitting signals within cells, lack of deep hydrophobic pockets and the only one theoretically targetable is its active GTP-site, but it generally binds with the GTP with such a high affinity (with a magnitude of picomolar) that no known drugs can win the competition [6, 61];
- The Transcription Factors (TFs) are DNA-binding proteins responsible for gene regulation, and precisely because they bind to DNA that is negatively charged, it makes challenging the use of ligands that simultaneously interact with TFs and cross the cell membrane; moreover, some TFs are protein complexes that have a large PPIs.

It is important to note that the currently developed smPROTACs involve a few dozen E3 ligases.
Still, since there are more than 400 human known E3 ligases, 4K known E3-Substrate Interations (ESIs) and 1.8M predicted ESIs with high confidence (data available on the rich-information databases E3net,UbiBrowser2.0, UbiNet2.0 and DregPred) [1, 2, 5, 37, 62, 63], the landscape in the designing PROTACs by exploiting specific E3 ligase is tremendously immense.
Moreover, some E3 ligases are expressed in certain disease conditions; thus there is the potential to develop disease-specific PROTACs.

Next, smPROTACs exploiting two E3 ligases that are the most commonly studied will then be discussed.

VHL-based PROTACs

Underlying the studies of peptide-based PROTACs in which the E3 VHL ligase was recruited, it was evident that strong PPIs were formed at the interface between the ElonginB–ElonginC-VHL (VBC) sub-complex and the human HIF-1α.
Therefore, Ciulli's team developed and synthesised through a combination of *in silico* methods and fragment-based screening guided by co-crystal structures, several ligands that interfere with the PPIs between VHL and HIF-1α with nanomolar binding affinity, opening a new door to the VHL-based PROTACs technology [64–67].

Shortly after, the same team developed the first VHL-based PROTAC, namely MZ1, which targets the Bromodomain- and Extraterminal domain (BET) proteins (BRD2, BRD3, and BRD4) as POIs through the linking of the (1) JQ1, a known BET inhibitor, and (2) VH032, a potent and specific VHL ligand, via a (3) three-unit PEG linker.
MZ1 showed to have effects with nanomolar affinity prolonged effects (up to 24h) with very low concentration due to multiple rounds of POI ubiquitylation.
Interestingly, another aspect that emerged from MZ1 studies that makes PROTAC a

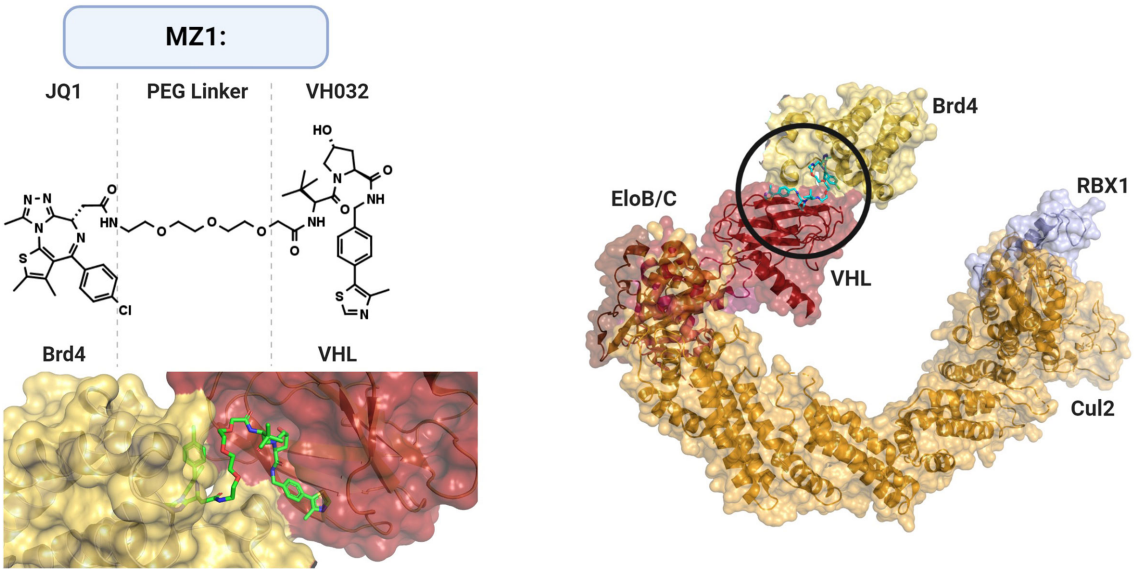


Figure 1.5 – *Model of the crystallography of $CRL2^{VHL}$ bound to MZ1 and $BRD4^{BD2}$ (N.Scholz, 2020)*

very feasible and powerful technology is that degrades selectively the BRD4 over all other BET bromodomains proteins, although the high degree of structural homology [14].
Lastly, the BRD4 is a critical protein overexpressed in human cancer cells and, if suppressed by knockdown or UPS-degradation, causes terminal differentiation and apoptosis in acute myeloid leukaemia cells [2, 5, 12, 51, 68].

Interestingly, the crystal structure of the ternary complex between VHL, BRD4, and MZ1 showed that PROTAC binds by folding back on itself into a cavity formed by BRD4 and VHL, precisely at the extended PPIs, as predicted, resulting in high stability and cooperativity of the ternary complex (Fig. 1.5) [57, 69].

Other two successful VHL-based PROTACs have as POIs the Estrogen-Related Receptor-α (ERRα), which incorporate the thiazolidinedione-based ligand, and the Receptor-Interacting Rerine/Threonine-protein Kinase-2 (RIPK2), which integrate the vandetanib [2, 3, 5, 12].

CRBN-based PROTACs

The (1) Cereblon (CRBN) is the substrate recognition of a E3 ligase complex composed of (2) scaffold protein CUL4A, (3) RBX1[15] and (4) DDB1[16].

The first ligand discovered in 2010 that binds to CRBN was the thalidomide, a potent phthalimide immunomodulatory (IMiDs) firstly discovered in 1957 and primarily used as a drug against insomnia but later found to have antiangiogenic and anti-inflammatory effects, which caused the birth of 8K÷12K children with severe deformities [3, 12, 52, 70, 71].

[15]Note: it is the same subunit in $CRL2^{CRBN}$ E3 ligase complex

[16]DDB1 is made up of three WD40 β-propeller domains (BPA, BPB and BPC)

Due to its teratogenic effects, derivatives such as lenalidomide and pomalidomide have been developed as CRBN ligands in the PROTACs and shown to successfully induce the ubiquitination and subsequent degradation of BET, TGF-β and BTK proteins and Ikaros transcription factors [1, 5, 12].
For example, PROTAC ARV-825 uses the inhibitor OTX015 as a warhead, pomalidomide as the E3 ligand and a flexible linker of PEG_{4units} and showed a significant reduction of BRD4 levels even with picomolar concentrations [4, 72].

As one might expect, differences in E3 ligase recruiting could lead to several effects: Bondeson et al. developed CRBN- and VHL-recruiting PROTACs using as warhead the kinase inhibitor foretinib, a molecule capable of binding to more than 100 substrate proteins.
The team found out that (1) by changing the linker or the E3-recruiting moiety, the binding profile of the compounds also changed significantly and (2) the CRBN-recruiting PROTAC effectively degraded 14 POIs over 52 bound proteins, while the VHL-recruiting PROTAC degraded 9 POIs over 62. Furthermore, between the two PROTACs, 6 POIs were the same kinases [3].

CRBN-based PROTACs to degrade microtubules

An attempt to design PROTACs that target the tubulin family is made recently by Gasic et al. [13].
Since α-/β-tubulin (α-/β-T) exists as dimeric form and it is continually subjected to an exchange between soluble and polymerized in microtubule forms[17], their degradation may occur only in the soluble form [13, 73–75].

Monomethyl auristatin E (MMAE, vedotin) or combretastatin A-4 (CA4) were used as warheads for the α-/β-T, whereas the pomalidomide was used as ligand for the $CRL4^{CRBN}$ E3 ligase.
Interestingly, the designing of the linker was done in a computational approach (protein-protein docking through the Rosetta framework) since its properties can significantly affect the overall PROTAC physicochemical properties as discussed below (See sec 1.3). The optimal linker lengths were found by clustering the best poses by interface scores and then observing via visualization tool the distance between the warhead and E3 ligand.
However, Western blotting and *in vivo* assays showed that degradation did not occur because there were no or weak engagements of CRBN with MAEE- and CA4-based PROTACs.

However, the MAEE-based PROTAC preserved the destabilizing property typical of the original MAEE inhibitor, which caused an increased concentration of soluble α-/β-T and decreased MTs; on the other hand, the CA4 completely lost the destabilizing ability.

Although the failure in the degradation of tubulin dimer, these results are significant and may have many possible interpretations; the authors claim that is unlikely to be because of a thermodynamic instability of the ternary complex $CRL4^{CRBN}$ -

[17] This mechanism is termed *Microtubule dynamic instability*, see sec. 1.5 for further details

PROTAC - α-/β-T since the simulations showed that there were no steric or electrostatic clashes which disfavour the complex formation [13].
Another hypothesis could be the absence of adequate PPIs between the CRL4^{CRBN} and α-/β-T, which do not trigger the movement of the "clamp" typical of E3 ligase required to transfer the UP to the POI from E2.

Another recent work worth noting focused on rationalizing a PROTAC having the p-Tau protein instead of microtubule components as POI [76].
Tau is well known to be associated with Alzheimer's Disease but is an important MT-stabilizing protein that has the critical role of maintaining the stability of the long MTs along the axons [76, 77].
When Tau is hyperphosphorylated, which occurs for unclear reasons, it dissociates from the MT surface causing the dismantling of MT, thus the collapse of the axons with subsequent neuron death.

Jampalli et al. rationalized a PROTAC in a basic computational manner: they performed a VS based on pharmacophore modelling and filtering, then a run of docking (See sec. 1.7.1 1.7.4) to find the warhead and exploit the CRL4^{CRBN} as an E3 ligase.
However, they did not experimentally test the designed PROTAC and did not even report a significant conclusion showing computational results. Moreover, tau degradation could worsen the collapsing of the axonal MTs.

Considerations for a good rationalizing of PROTACs

Generally in small-molecule drug discovery, the gold standard approach is synthesizing millions of compounds and performing HTS. Still, despite large libraries of compounds, this method can explore only a small area of chemical space.
But in PROTACs discovery, things are worse because the chemical space is tremendously bigger than small-molecule designing, in addition to a greater amount of resources needed.
Furthermore, the distinction between compounds that disrupt E3 ligase-POI interactions and those that inhibit POI or E3 ligase ubiquitination activity does not provide straightforward readouts via conventional *in vivo* assays and HTS [12].

Maintaining a chemical/experimental approach, the best ways to design PROTACs without wasting a significant amount of money and time, are two [78]:

- the *Fragment-Based Drug Design (FBDD)* approach is a gold standard method to find new binding pockets, but instead of starting with a fragment with low affinity, the starting point should be an E3 binder and warhead already known to then expand into bifunctional PROTACs;
- the *Structure-Guided Drug Design (SGDD)* approach consists of observing typical interactions in the cocrystal structures of E3 ligases with fragment peptides of substrate proteins. However, even if suitable ligands are found, there is a need to synthesize dozens, if not hundreds, of PROTACs by varying only the length and/or chemical composition of the linker.

But in the last 5÷7 years, it has become increasingly common practice to design PROTACs first with *in silico* methods, allowing to estimate binding energy, predict the poses of ternary complex E3-PROTAC-POI and kinetics of association and dissociation based on respective binary and ternary affinities [6].
Some of these computational protocols for the design of PROTACs are included in the Rosetta framework and MOE and MATLAB-SimBiology software, with the first two that perform a series of docking runs. Also, running MD is a viable option if high computational resources are available, such as High Performances Computing (HPC) [6, 79–83].
However, these methods are relatively new and require further optimization since the degree of freedom is remarkably higher than the POI+ligand situation.

For example, Drummond et al. developed a series of 6 useful tools on the MOE platform called *PROTAC modelling tools*, which help the designing of the PROTACs at the first stages. Still, it is worth mentioning the tool *Method 1* that guides the linker's PROTAC designing in a stepwise manner [81, 82]. The linker, which will be discussed better below, is a crucial component of the PROTACs.
Nevertheless, Drummond et al. reported that the best hit rate, defined as the proportion of ternary complexes poses similar to true crystallography models with a Root-Mean-Square Deviation (RMSD) < 10 Å, was very high only for the test structure, while it remained very low for different systems (0÷40%), even after a significant improvement of the tools [82, 83].

Regardless of the method used, in the rational design of small molecule-based PROTACs, one must take into account several important aspects that can determine the effectiveness and success of the PROTAC itself:

- **The right choice of E3 and its ligand**
  Even with the same POI, exploiting different E3 ligases can have different effects, including absence of degradation [3, 13, 84].
  Before rationalising PROTACs, it is strongly recommended to test whether E3 leads to POI degradation in any way, even under unrealistic *in vitro* conditions, such as a high concentration of UP, E1, E2, E3 and proteasome 26S in the presence of POI. This type of experiment is quite simple to perform (Western blotting).
  Only when some degradation occurs can ligand designing begin if not available.
- **The Linker**
  The designing of the ligand plays a crucial role since it can significantly affect the physicochemical properties and bioactivity of PROTACs, and, even when the warheads and E3 ligands have been optimized as best as possible, failures can occur with high rates [13, 14].
  There are no general rules in the designing of linkers. Indeed, in many works, there is a tendency to take an iterative trial & error approach, which may be very resource-consuming, such as synthesising entire PROTACs by changing slightly only the linker's length and chemistry and then employing HTS methods.
  Moreover, based on 600 PROTACs available online (PROTAC-DB, 65% of the linkers are alkyl and PEG[18], which are often used as a starting point for later designing a better linker, 15% are modified PEG, 7% are alkynes, 6% are triazoles, and 4% are saturated heterocycrationalizingiperazine and piperidine [14, 85].
  Thus, there are mainly two complementary factors: the length and the chemistry. The length's design may be more intuitive than chemistry because generating the models via crystallography (which is the best approach), MD or docking protein-protein of the complex POI-warhead + ligand-E3 (with no linker between the two moieties), can give some vital information that would make it easier the design of the length [13, 14]. In general shorter PROTAC linker facilitates favourable PPIs, while a longer linker could generate steric clashes.
  Regarding the chemistry, it was observed that rigid rather than flexible linkers, such as the use of phenyl or macrocyclic groups, can significantly affect potency and selectivity [6, 14, 78]
- **The PPIs between E3 and POI**
  The degradation potency of a given PROTAC correlates better with its ability to form a stable ternary complex (E3:PROTAC:POI) than with the binding affinity of the warhead and/or E3 ligand.
  This happens because the PPIs between the E3 and POI could most likely compensate when the ligands have a low binding affinity (high free energy ΔG). Indeed, several studies have reported similar comparisons between weak and strong ligands in terms of effective degradation [3, 6].
  Moreover, although there is no strong evidence in the literature, the few published crystallographies of ternary complexes E3:PROTAC:POI, like the ones

[18]PEG are very used because of flexibility, simplicity in the synthesis and tunability of length and lateral groups influencing easily the important physicochemical and biological properties

developed by Ciulli's and Crew's teams, suggest that exploiting the native PPIs is the optimal key mechanism which triggers the ubiquitination process with high probability and this may occur in biological systems with specific combinations of E3:POI [23, 64, 69, 72].
This hypothesis is also based on the fact that there are more than 400 known E3s, which suggests that each protein has a specific set of E3s. It should be remembered that E3s can also induce post-translational modifications, as discussed in Sec. 1.2, that do not lead to degradation, so only a few of these are those that can catalyze the addition of 48LYS-linked poly-UP chains to protein targets.
However, given the large number of subunits in most E3 ligase complexes, it is possible that even by changing only one subunit, particularly the substrate-receptor, the ubiquitin activity changes significantly, as in the case of E3 with CUL as a scaffold.

Briefly, a generic computational workflow of a PROTAC design is:

1. make the crystallography of the POI+ligand and E3+ligand if possible (together should be perfect), otherwise use structures already available on Protein Data Bank or predicted by AlphaFold;
2. run a series of docking with the ligands of interest to check out the most exposed group;
3. predict the PPIs via protein-protein docking (there are several databases which have already explored both known and predicted ESIs, such as the database mentioned before);
4. model the linker based on the results of POI+ligand and E3+ligand and prepare the PROTAC;
5. perform another series of protein-protein docking to evaluate the PPIs with the PROTAC bound and estimate the binding energy by comparing with that of the one POI:E3 situation.

To conclude, it is worth mentioning that it is also possible having highly functional linkers by employing click chemistry or even photoswitchable linkers based on azobenzene chemistry.
The last linker type allows high spatiotemporal control on the PROTACs.
However, these two topics deserve further study separately.

1.4 The E3 ligase: UBR1 and degrons

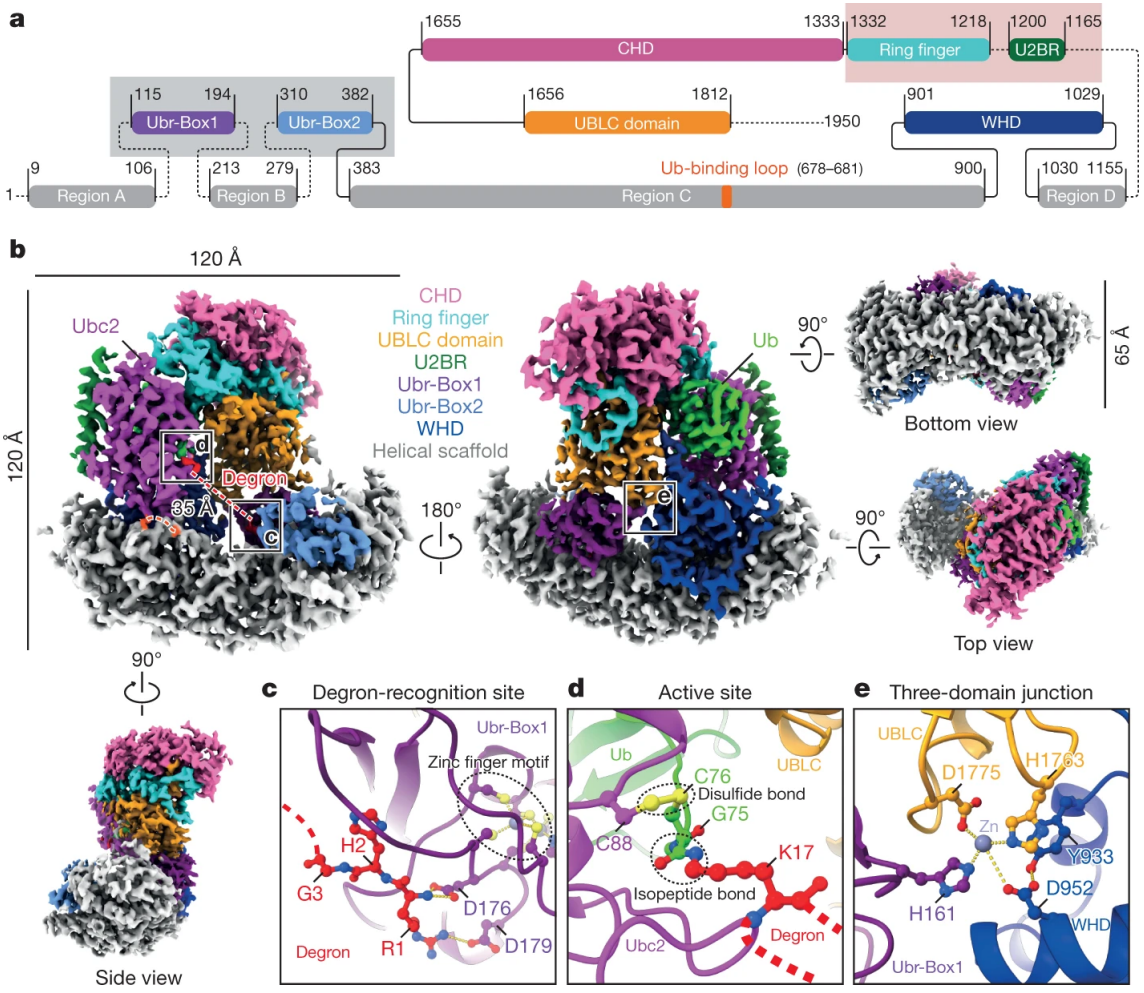


Figure 1.6 – *(A) Domain diagram of Ubr1, the dotted lines represent unresolved linkers and regions; grey box = substrate-recruiting domains; pink box = Ubc2-recruiting domains. (B) Cryo-EM maps of the initiation complex (color code as previous). (C) Molecular interactions between Degron and Ubr-Box1; dotted lines = H-bonds and electrostatic interactions. (D) Molecular structure of the active site. (E) Molecular interactions between UBLC, Ubr-Box1 and WHD domains with metal coordination bonds and hydrogen bonds. (M.Pan, 2021)*

The E3 ligase which is exploited for the targeting of γT is the UBiquitin-protein ligase E3 component - Recognin 1 (UBR1), although there is no strong evidence of their interactions in the scientific literature (https://thebiogrid.org/113134/summary/homo-sapiens/tubg1.html, see below for further details)[19].
It has been asked to use anyway the UBR1 since Fahlman's team is an affiliate of Tuszynski's team exploits this protein to ubiquitinate other proteins they are working on.

[19]Only one report mentions a possible interaction between γT and another protein of UBR family, the UBR5, but is cited only in Fig.2 on the study which shows weak binding interaction. However, the results are from high-throughput mass spectrometry in UBR5 immunoprecipitates; thus, further separate studies are needed [86].

The main reason is due to a high number of copies of *UBR* family genes[20] and high similarity in the corresponding protein structures: in this way, if one copy of *UBR1* gene mutates, the other UBRs may replace it [87].
But if two of these UBRs (UBR1 and UBR2 or two UBR1 copies) are somehow mutated or knock-out, defects or diseases can occur with higher probability, such as the congenital disorder called Johanson-Blizzard syndrome (JBS) or early embryonic lethality in mice, respectively [87].

The crystallography of entire UBR1 is available on Protein Databank (PDB: 7MEX (initiation complex) and 7MEY (elongation complex)) and, like many other E3s, the overall UBR1 is like a "clamp" structure.
It is constituted of a helical scaffold subunit with three main domains: Ubr-box1, Ubr-box2 and a Winged Helical Domain (WHD) [88].
Interestingly, fig. 1.6 shows the UBR1 crystallography in an elongation phase, in which it is catalyzing the transfer of UP from E2 (not shown) to the first UP already bound to the LYS residue of the degron, which coincides with the active UP-site of the POI (Fig. 1.6D).

Some evidence report that one of what the half-life of the protein depends on the most is the exposure of degrons to the cytosolic environment [34, 89].
Several "rules" were proposed to classify the pathways in which E3s recognize the POIs via degron [34, 35, 89]:

- **N-end rule** controls the protein's half-life throughout its N-terminal residue, referred to as the N-degron. This class is further divided into two subclasses:
  - type 1 N-degrons, which consist mainly of strong positively charged residues, such as arginine, lysine and histidine, sorted by decreasing direction by the strength of interaction. If the first residue is ARG, it seems that degradation pathways prevail; instead, other residues in N-terminus with post-translational modifications, such as acetylation, deamidation, arginylation, leucylation and formulation, can lead to totally different pathway
  - type 2 N-degrons, which consist of bulky hydrophobic residues, such as proline. They locate mostly in the second/third residue from the N-terminal residue.
- **C-end rule** has only recently been proposed because C-degrons were only recently discovered in 2018, effectively complicating what was known about the N-end rule: it is structurally similar to the N-degrons, and they tend to bind to the CRL2 E3 recognin subunit.

It is important to note that internal degrons exist too, but their roles and how they affect the protein functions are still unclear since it is not easy to interact with them precisely because it is too deeply embedded [34, 90].

[20] A gene has a maximum of two chromosomal copies (one derived from mother, while the other from father), but gene duplication can occur, thus generating more than two copies, termed as paralogs. Over time, the new copies can mutate differently and independently (termed as *subfunctionalization*) or one can preserve itself while the other is freer to mutate (termed as *neofunctionalization*) (For a better understanding of these concepts, see the following detailed thread).

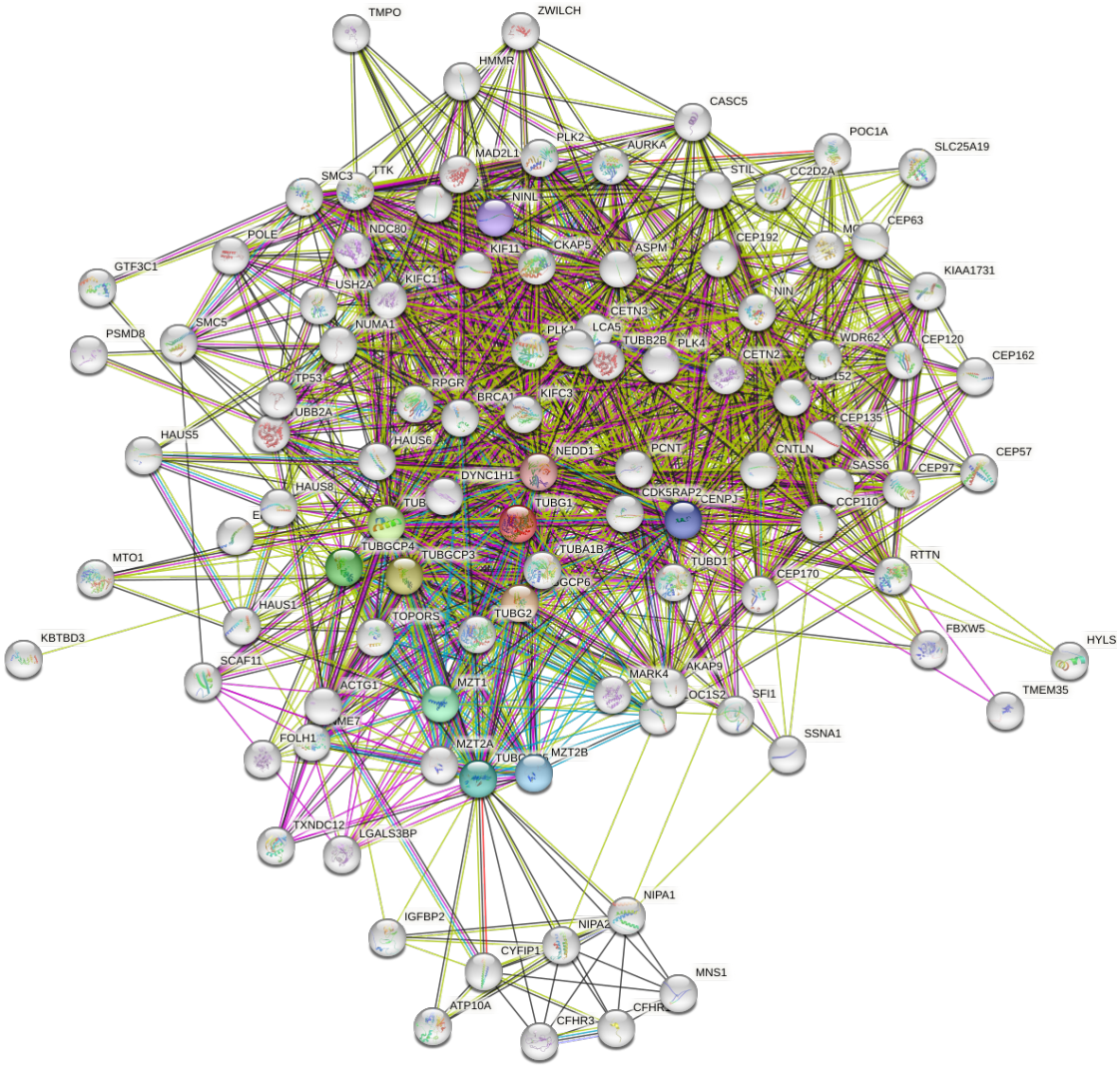


Figure 1.7 – *Result search STRING-DB using TUBG1 as keyword and score search > 0.4)*

Moreover, other proteins, such as cochaperones, including the Hsp90 family, can be protected by degrons, prolonging the protein half-life.

The 1 type N-degrons are mainly recognized by the UBR E3 family: UBR1 and UBR2 ($\sim$ 200 KDa), UBR4 ($\sim$ 570 KDa) and UBR5 ($\sim$ 300 KDa). In particular, such degrons bind selectively to UBR-box, a ZF domain of $\sim$ 70 aminoacids with a primary pocket negatively charged, which binds directly to N-degrons (the highest binding affinity is with ARG residue), and a secondary pocket that is more prone to establish hydrophobic interactions depending on the residue's chemistry.

To test the potency and selectivity of N-degron-recognin of UBR-box1 and UBR-box2, Munoz-Escobar et at. have iteratively changed the second and third residues of a degron by substitution of 20 aminoacids, leaving ARG residue as the first residue. They observed that there are higher chances of UP covalent binding with small-/medium hydrophobic residues such as GLY, ALA, VAL and TYR as the second residue but not with clashing hydrophobic residues' side-chains, such as PRO and LEU, which occlude the pocket, depleting essential the H-bonds and electrostatic

interactions needed to completing the ubiquitination process.

Lastly, UBR1- and UBR2-box domains, which include the pocket of degron-recognition, share a 77% identity, further motivating the strategy of using UBRs so there is a lower risk that mutations may render PROTAC ineffective.

It could be interesting to design a ligand that mimics the N-degron to maximise the success of PROTAC's bioactivity.

It is worth noting that not even several rich-data databases of both known and predicted ESIs report that there is evidence of their interactions of any type by using as a keyword the γT (E3net and UbiNet2.0).

Interestingly, besides the BRCA1, which is the only experimentally confirmed E3 that ubiquitinates γT, UbiBrowser2.0 predicts 20 E3s binding to γT with a high confidence score, but none of them is the UBR1.

A powerful data-miner (STRING-DB)is used to search all possible interactions instead by looking at keywords and other types of information, but this also resulted in negative results (Fig. 1.7). All known interactions with γT with a combined score (aggregation of several scores such as confirmed experimental results, text extraction and so on) are reported in the Appendix A.1.

rquires two arguments. The second argument will be used for the pdf section name, while the first argument will be displayed with (La)TeX

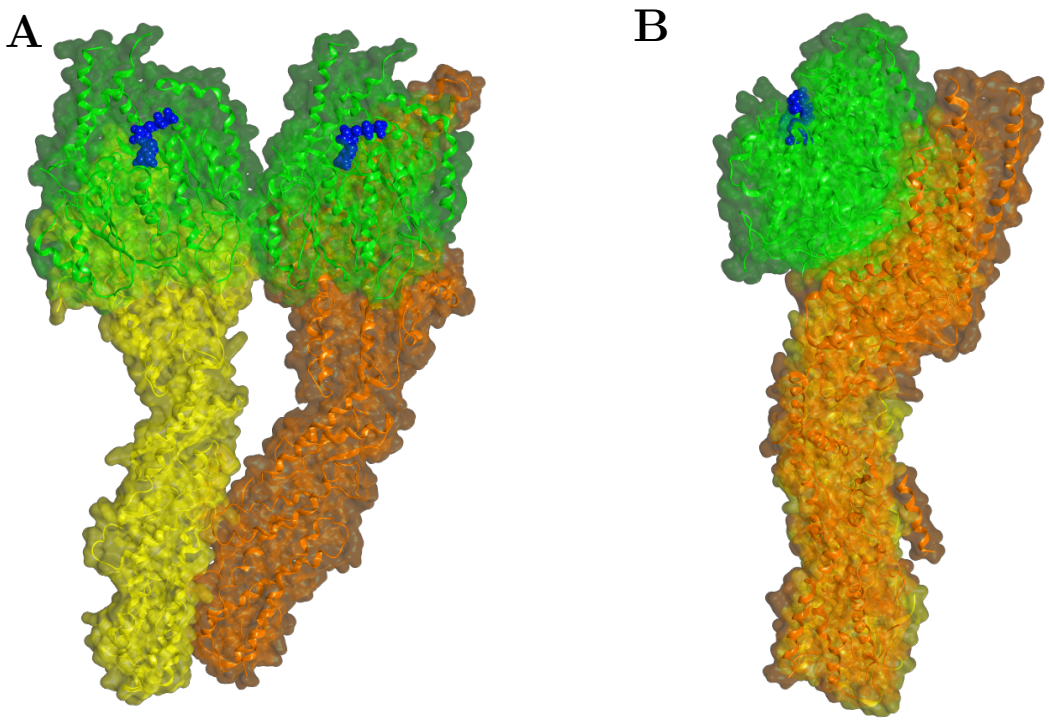


Figure 1.8 – *The γT-Small Complex crystallographic model (PDB: 6V6S); (A) frontal view, (B) lateral view. Yellow = GCP2; orange = GCP3; Green = γT; blue = GDP. Models realized in MOE*

1.5 The POI: γ-Tubulin and its role in the Microtubule nucleation

In humans, there are two genes that encode for γT: *TUBG1*, which is ubiquitous in every cell type, and *TUBG2*, whose express corresponding protein mostly in the brain [91–95].
The two genes express the relative proteins with 97% sequence identity (difference of 10 AA) and 99% in sequence similarity.
It is important to note that, especially in the older work, it had not been possible to distinguish the two tubulins because the anti-*gamma*-tubulin marker, used in almost all the published papers, is not selective toward either tubulin, so in the present work, when said γT, it refers to both [94].

In mouse knockout studies, several important results were observed [94, 96, 97]:

- *TUBG1* knockout: the embryos showed to survive until the morula/blastocyst stage, from which the division arrests because the microtubules (MTs) of the mitotic spindles are deformed and abnormal, although their preserved stability. This causes unaligned and abnormally condensed chromosomes and suggests the critical role of γT-1 in mitosis. Interestingly, the proteins expressed from *TUBG2* were absent on the centrosome's surface, suggesting that γT-2 is not involved directly in the cell cycle and the assembly of γ-Tubulin Ring Complex

(γTuRC).

- It has been shown that a reduction of 50% of γT-1 levels in mice with *TUBG1*-knockdown via siRNA led the cells to apoptosis after three days, during which mitosis operated adequately, probably due to high availability of γT-1 in the cytoplasm, after which it was exhausted.
- *TUBG2* knockout: the embryos showed to be still viable and fertile, although they exhibited some defects, including slight G2/M delay and abnormalities in circadian rhythm.
  However, there are not many studies on the functions of γT-2 for the reasons discussed above, so it is still not well understood in which mechanisms it is involved, despite the almost identical similarity to γT-1

The γT-1 is a 451 AA globular protein highly conserved among all eukaryotes but may vary among species in terms of protein expression levels.
For example, in mammalian cells, 80% of γT is cytoplasmic and not associated with the centrosomes; furthermore, the γT represents less than 1% of the total tubulin content in the cell [91, 92, 96, 98, 99].

Generally, when extracted, γT can be found mainly in the form of a tetrameric complex termed as γ-Tubulin Small Complex (γTuSC), which consists of two γT monomers (both γT-1 and γT-2 were found) and two different paralogs of γ-Tubulin Complex Proteins (GCP), which mostly are a pair of GCP2 and GCP3, but also GCP4-6 exist, even in a lower concentration.
The tetramer form allows it to be more stable against degradation and maintains

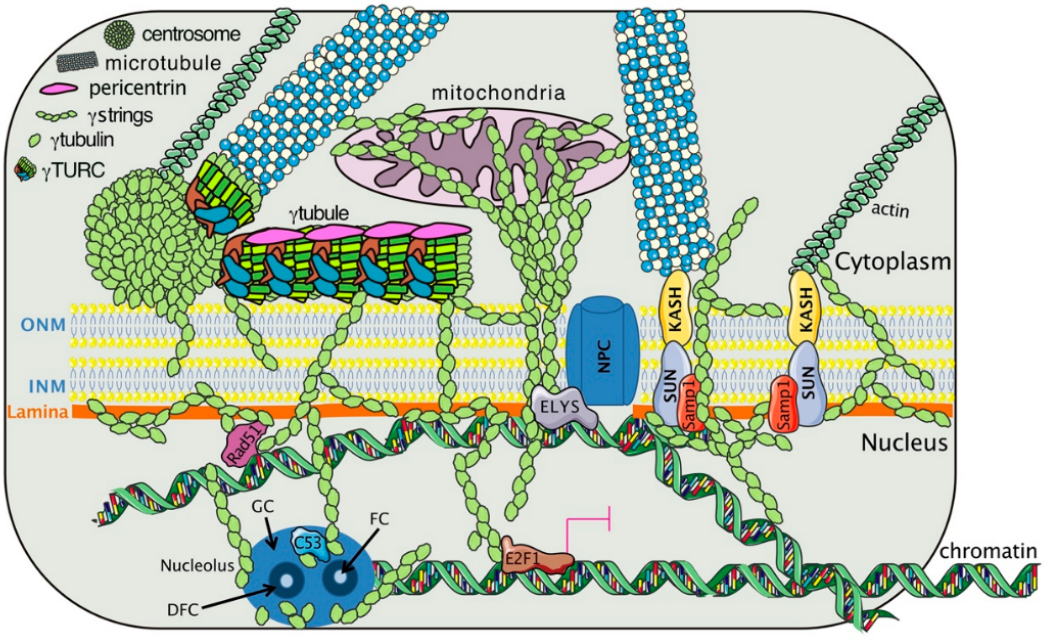


Figure 1.9 – *Hypothetical distribution of γTs among γ-Tubules, γ-Strings, centrosome, cytoskeletal components (MTs and actin), chromatin, nuclear membrane (outer and internal nuclear membrane (ONM and INM respectively), mitochondria, nuclear lamina and nuclear pore complex (NPC). ELYS=Embryonic Large molecule derived from Yolk Sac; FC=fibrillar center; GC=granular center; DFC=dense fibrillar component; E2F1=E2 promoter binding factor 1. (M.Corvaisier, 2020)]*

high solubility compared to the γT alone (Fig. 1.8) [95, 100–102].

Interestingly, by suppressing the γT, the GCP2-3 levels highly reduce, whereas if the opposite is the case, the levels are not reduced as much as in the first case. This difference may be explained by the exposition of their degrons which leads to the degradation: in fact, the UP-site and degrons of GCP2-3 are located at the interface with γT, while the γT-degron is predicted[21] to be internal [101].

The tetrameric complexes are not dispersed in the cytoplasm as if they were solute, but they are organized into further complex rearrangements, termed as γ-String. Moreover, γTuRCs already assembled and aggregated in a row have also been identified, and they are termed as γ-Tubule (Fig. 1.9) [97, 103–105].

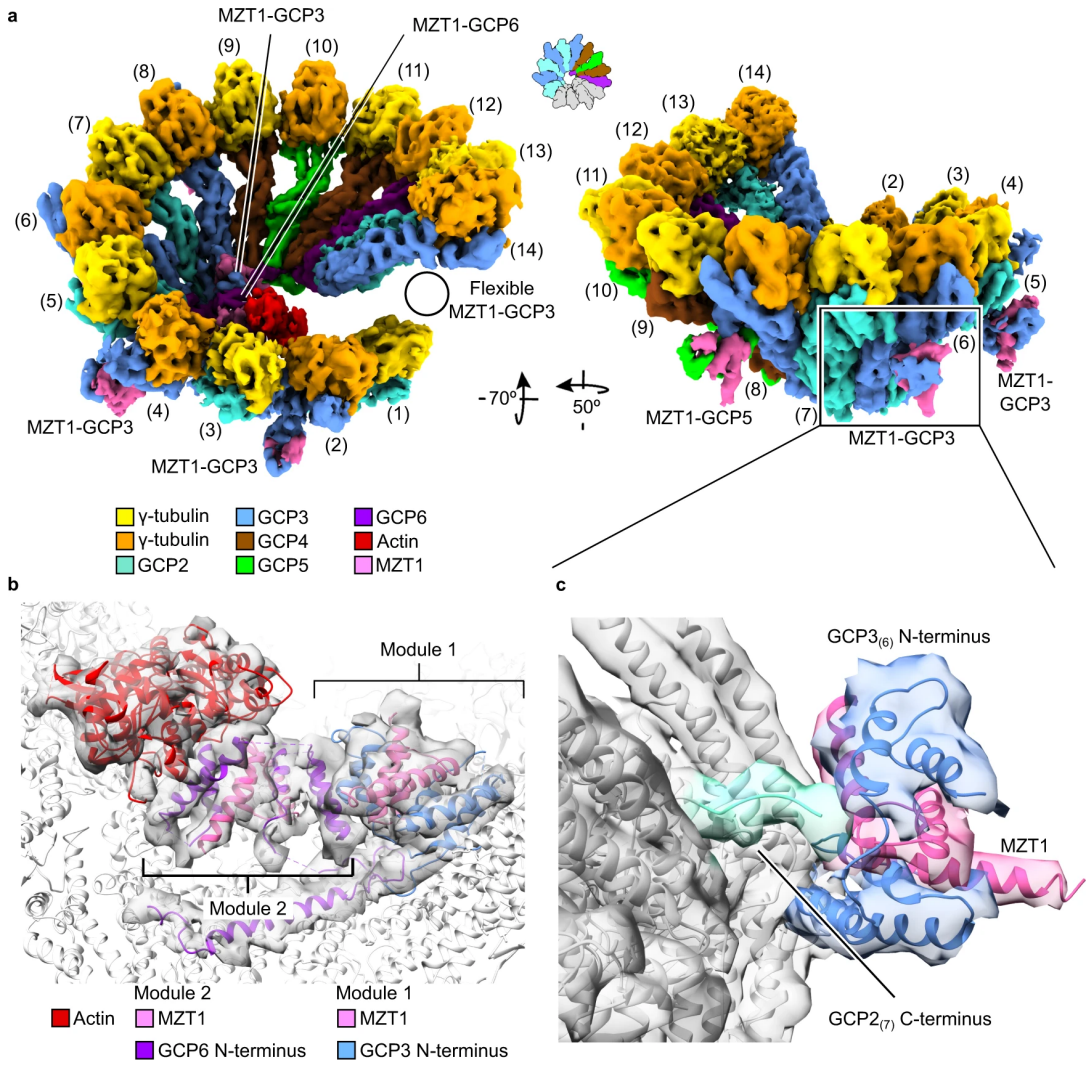


Figure 1.10 – *(A) Views on the cryo-EM reconstruction of the recombinant human γTuRC; (B) structure and composition of the lumenal bridge; (C) Cryo-EM density of the MZT1-GCP3 module associated with GCP3. (PDB 6X0U; M.Wurtz, 2022)*

[21]The prediction are made via an online tool. See Sec. 4 for further explanation

The first high-level structure is static, while the latter is highly dynamic and varies in length as if they were other MTs: both will be classified as meshwork.
In some ways, these meshworks constitute an additional component of the cytoskeleton since they may provide further mechanical influence on the cell interior and signal transduction, other than as a reservoir to efficiently and quickly assemble the γTuRC [97].

How the single components (γTuSC and γTuRC) come off from the meshwork is still unclear, but several studies both in human and plant cells revealed that the γ-Tubules and γ-Strings are significantly abundant in the cell's cytoplasm during interphase, in particular before the S/G2 phase (Fig. 1.15), after which their levels substantially decrease [92, 102, 106, 107].

This coincides with the fact that many γTs[22] migrate toward the of spindle poles' surface (I.e. Centrosomes, which are that have duplicated; see Sec. 1.6) forming the MT-organizing centers (MTOCs) of centrosome [92, 102, 106].
In this way, a high level of γTuRCs start to nucleate high levels of new MTs efficiently and quickly, in particular, asters, which orient toward cell membrane and kinetochore MTs which indeed grow toward the single kinetochores located in chromosome's centromeres [97, 102, 108–110].

It is important to note that, even though the amount of γT generally decreases at the beginning of mitosis, no UPs are bound to these previously mentioned structures, suggesting the idea that the γTuSC or γTuRC can be disassembled from the meshwork just as they can be reassembled after cell division is completed [92].

The γ-Tubulin Ring Complex

The γTuRC (2.2 MDa) has an overall shape of an asymmetric left-handed spiral and consists of five γTuSC composed of 2 γT + GCP2-3 each, one γTuSC of 2 γT + GCP4-5 and one γTuSC of 2 γT + GCP5-6 + additional proteins with roles not fully elucidated[23], primarily located in lumenal bridge such as one molecule of actin[24] and two similar α-helical structural modules that include the Mitotic-spindle organizing protein 1 (MTZ1 or MOZART2A/2B)[25] (Fig. 1.10) [100, 111, 112].

There are several models on how γTuRC is assembled, and the closest to the experiments and on which many different works agree is the one proposed by Wurtz's team in a very recent work which was able to extract entire recombinant human γTuRC from *E.coli* (Fig. 1.13) [100, 102, 111, 112]:

[22]Note: it is not clear if these migrating γTs are in γTuSC form (probably from γ-Strings), in γTuRC form (probably from γ-Tubules) or in both

[23]Problably not only as structural integrity.

[24]Actin is not required in γTuRC assembling or structural support, but if mutated or suppressed in human cells, defects in MT nucleation and chromosome alignment were observed [111]

[25]More modules of MTZ1 bind to GCPs both from inside the ring and outside; they probably act as a regulatory mechanism with a potential role during γTuRC recruitment to the centrosome [111].

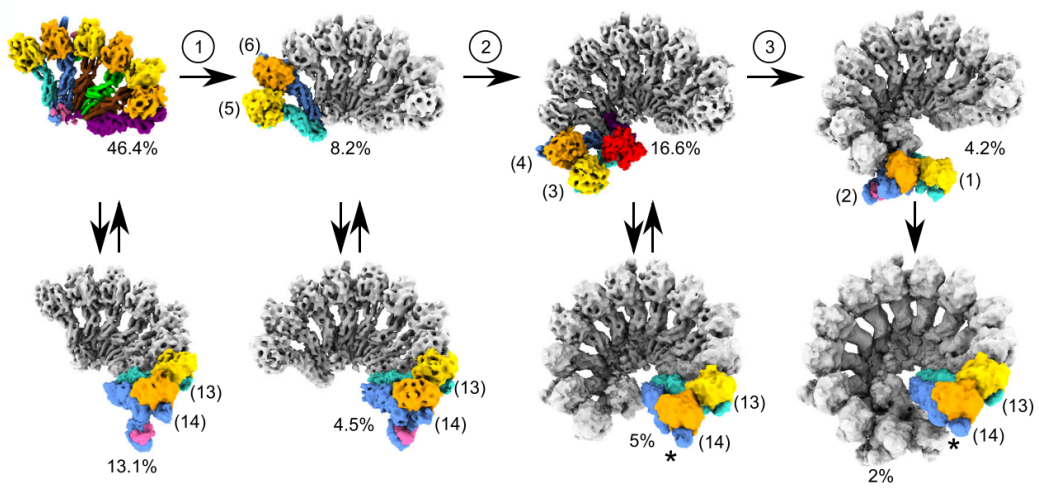


Figure 1.11 – *Cryo-EM reconstructions of γTuRC subcomplexes intermediates arranged into a possible pathway to the complete form and alternative assembly pathways and intersections between different. The percentage is given for homogeneous sets of particles representing distinct assembly states. Each step is coloured, while the already present are grey. (M.Wurtz, 2022)*

1. six spokes[26] of positions (7,8,9,10,11,12) integrate into one subcomplex "core";
2. the core expands by recruiting γTuSC (spokes of GCP2-3) in stepwise additions, while the one unit of actin binds by interposing itself in the luminal bridge and MTZ1s bind to several GCPs (Fig. 1.13);
3. during the assembly, the spokes in positions (13,14), termed as locker, can associate and dissociate from the core many times, maybe affecting the stability or assembly rate, but when the last γTuSC (spokes in positions (1,2)) is embedded, it binds directly to the locker, stopping to dissociate from the core and closing the ring.

The Microtubules

Once the γTuRCs are assembled, independently from where they are located, they act as a mould for MT assembly by binding with αT of the α-/β-T dimers and promoting their lateral interactions.

The MT is one of the cytoskeletal and nucleoskeletal components in a cell, together with actin and intermediate filaments, such as keratins, lamins, vimentins, and desmins.

They (all cytoskeletal/nucleoskeletal components) play critical roles in several cellular functions: providing form and mechanical support; organizing the genome during cell growth; assisting in cell movement and transport and controlling PPIs. Both actin filaments and microtubules can dynamically assemble and disassemble polar filaments, whereas intermediate filaments form static structures [92].

In the case of MTs, these structures are highly dynamic, and their functions strictly depend on tubulin abundance, which is regulated through balanced protein synthesis

[26]A spoke is defined as a heterodimer composed of γT + one of the paralogs of GCP2-6s

and degradation, a phenomenon termed as *dynamic instability*.

The MT building block is the α-/β-T heterodimer, which constitutes over 2.5% of the total protein content in a cell. It assembles head-to-tail to form a polar protofilament with αT exposed on the minus end and βT on the plus end [73, 74, 96].
13 protofilaments assemble into a cylinder, forming the microtubule with the minus end generally paired with γTuRC.

Ten and nine genes in humans exist for αT and βT respectively, and some of these isoforms are ubiquitous (*TUBB4B, TUBB5, TUBB6*), while others are not (*TUBB1* in hematopoietic stem cells and leukocytes where there are over 50% of all βT; *TUBB2A* in the brain in high levels; *TUBB3* in both central and peripheral nervous systems; *TUBB4A* in the brain where represent 46% of all βT in the brain; *TUBB8* in oocytes) [113–115].
All tubulins have a GTP binding site, but the βT is also a GTPase which hydrolyzes the GTP into GDP[27], while the αT "hosts" the nucleotide only.

There are currently two opposing models for the relationship between GTP and conformational change in the single αT and βT structure [73, 116, 117]:

- *allosteric model* supposes that GTP binds to unpolymerized α-/β-T dimers dispersed in the cytosol, inducing straighter conformation along the dimer, which allows higher lateral interactions with other dimers. Meanwhile, other dimers spontaneously bind on the top and bottom of the dimer itself.
  Several electron microscopy-based experiments on large multi-protofilament assemblies have shown that with bound GTPs, the structure is straighter than in the case where there is GDP instead, which tends to depolymerize the MT because not sufficiently energetic to overcome the energetic barrier;
- *lattice model* supposes that αT and βT dimers adopt a curved conformation independently from GTP binding. Instead, GTP strengthens the MT lattice by improving the lateral PPIs; thus, the 13 protofilaments are already polymerized and bound together.
  This hypothesis is supported by the SAXS[28] and crystal structure of α-/β-T bound to colchicine, which locates at the α-/β-T interface, and a stathmin-like domain which is showed still a curved conformation, despite the bound GTP.
  Additionally, the presence of bound GTP or GDP did not affect the affinity of colchicine in the interface.

However, both models are not enough to explain the MT assembly. One of the several reasons is the lacking considerations of Mg^{2+} ion, which affect the αT-βT interactions at their interface significantly.

The Microtubule nucleation and the influence of GTP/GDP

Another model of how MT assembly occurs *in vivo* involves the γT.

[27] I.e. Hydrolysis consists in losing the third phosphate group.

[28] Small-Angle XR Scattering

The GTP binds to its active site of γT, which is part of a highly conserved domain among all tubulins in eukaryotic cells, inducing the initiation of MT nucleation activity.
Still, other proteins are involved in MT nucleation, such as CDK5RAP2, XMPA215, TPX2, CLASP1 and augmin, other than MZT1 and actin; additionally, also the phosphorylation via kinases and post-translational modifications in the proximity of degrons via E3 ligases can affect the MT nucleation by inducing further structural changes on the ring complex [91, 98, 105, 116–119].

Nevertheless, the presence of GTP or GDP does not affect the overall structure of γT significantly in terms of RMSD, as reported by Rice et al. and further confirmed in the current work, although the high similarity with all other tubulins [117]. Moreover, strong evidence from an experiment conducted on yeast[29] by Gombos et al. proved that GTP has a critical role in MT nucleation and structural organization, instead of γTuRC assembling through immunostaining and mutations of residues part of binding GTP-site [116].
Indeed, the following characteristics were observed in the case of mutations of the GTP-binding-site:

- monopolarity of the spindle and disorganized centrosomal MTs (cMTs) in 20÷50 of yeast cells;
- lower number of MT content than wt-*TUB4*;
- cMTs longer and less dynamic (hyperstable) than wt-*TUB4*.

In a nutshell, the GTP binding in γT seems to somehow affect PPIs with αT, which are directly bound to γT: this alteration would impact the entire MT structure.

Additionally, binding affinity was also found: the K_d of wild-type γT for the GTP is 45 ± 12 nM, while the K_d for the GDP is 206 ± 76 nM demonstrating that GDP is easy enough to replace with new GTP nucleotides, but at the same time it is challenging to design a ligand that has a lower binding affinity than GTP.

Interestingly, MT nucleation via γTuRC can take place not only in centrosome but also in other MTOC locations, such as mitochondria, Golgi apparatus, endoplasmatic reticulum, nuclear envelope, plasma-membrane associated sites, pre-existing MTs, and chromatin, thus giving γT a key role in the architecture of the MT network [91, 93, 97, 100].

It was widely believed until recently that for proper MT nucleation to occur, only γTuRC was sufficient. Still, it was observed that MTs formed the same even in their absence, although the significantly reduced amount of MT and its kinetics.
Wherever a recent study showed that another protein is essential to MT nucleation: XMPA215 [120]. Its C-terminus binds to the γTuRC while the two of the five TOG domains (TOG1-2) to the α-/β-T dimers, synergically regulating the MT nucleation (Fig. 1.12).

It is well known that experiments *in vitro* showed that MTs could assemble spontaneously from the high concentrations of αT and βT, but this rarely occurs *in vivo*, where the αT and βT concentrations are limited [119].

[29] Human γT and analogous γT in *Saccharomyces cerevisia* encoded in *TUB4* gene have 39.20% of identity and 60.60% of similarity. Scores re-evaluated on a tool online.

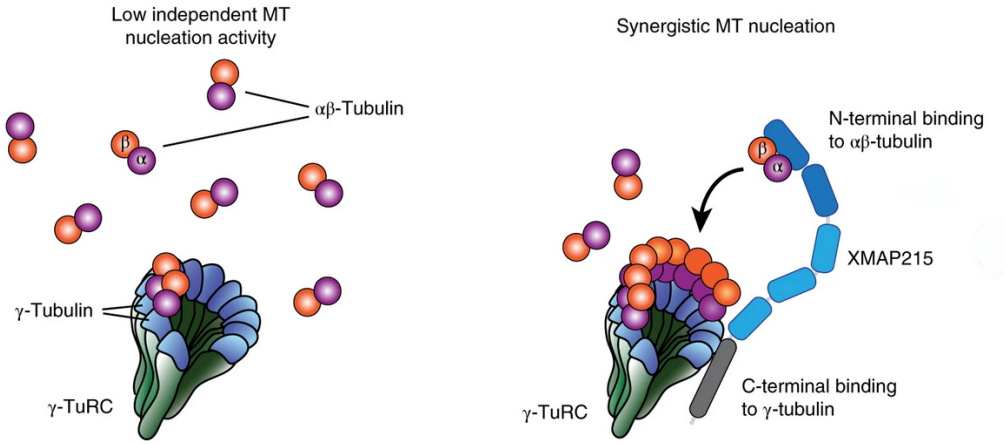


Figure 1.12 – *Schematic representation for how XMAP215 and γTuRC could together promote MT nucleation based on the absence and presence of XMPA215.*

Instead, in the presence of XMAP215, the number of MTs increased in a concentration-manner, without affecting the MT growth speed and with the typical concentration of αT and βT *in vivo*. On the other hand, with the same concentration of αT and βT but without XMAP215, the MT was completely abrogated.

The only known E3 ligase of γT

Lastly, the only experimentally E3 confirmed that ubiquitin the γT is the BRCA1, which is a RING E3 type and the only known E3 able to ubiquitinate as a monomer. Nevertheless, its mechanisms are not fully understood because several studies observed that E3 mono-ubiquitinates the γT, whereas others reported apparent contrasting results in terms of γT degradation [41, 98, 121–124].
Briefly [42, 48, 49, 98, 125–128]:

- The overexpression of BRCA1 leads to arrested growth or apoptosis.
- The overexpression of BRCA1 fragment (504÷803) which corresponds to the γT-binding site leads to apoptosis;
- The suppression of BRCA1 activity by knockdown via siRNA or knockout leads to centrosome amplification.
- If 1 or more residues of the RING domain are mutated or truncated, even without hindering its heterodimerization with Brca1-Associated Ring Domain 1 (BARD1), there is the abolition of E2 UbcH5c binding, hence lack of ubiquitination ligase function, leading to centrosome amplification and a 2-fold increase in MT content compared to control experiment and no γT ubiquitination was experimentally observed.
- Using BRCA1 fragment (1÷500 AA), which includes RING domain and excludes γT-binding domain, in complex with BARD1, weak ubiquitination of γT was observed: the MT content decreased by 40% compared with over 83% for full-length protein. This result suggests that the BRCA1 fragment can also ubiquitinate and consequently degrade γT, lowering the nucleation rate of MT.

- The reversible and external inhibition via BIF[30] on the recognin site located in tandem BRCA1 C Terminus (tBRCT) domain induced a 1.4-fold increase in MT content compared to control.

Although BRCA1 alone can ubiquitinate, albeit by transferring only one UP, a possible explanation for γT degradation observed *in vivo* is that there are many other proteins involved, which increase the ubiquitination activity of BRCA1 (I.e. transferring more than one UP) and consequentially γT degradation.
Currently, BARD1, Obg-Like ATPase 1 (OLA1) and Receptor of Activated protein Kinase C 1 (RACK1) are known proteins that bind to BRCA1 affecting its E3 activity, but other unknown proteins may be involved (Fig. 1.13)

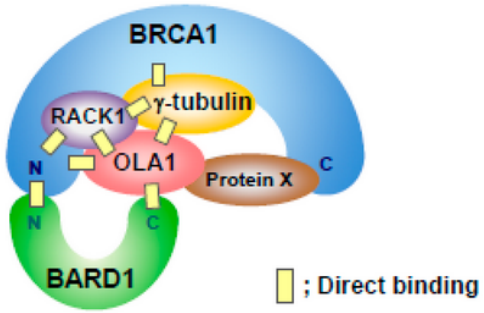


Figure 1.13 – *E3 ligase BRCA1:BARD1 complex with known subunits interacting with γT. (K.Otsuka, 2020)*

The known ligands

- Colchicine, combretastatin A-4 and paclitaxel were all known to bind in approximately the same pocket of βT. Since the latter protein and γT are structurally almost identical and share a high similarity, a study tested the effects of these compounds on recombinant human γT by the fluorescence spectroscopy assay and observed that colchicine and combretastatin A-4 bind to γT at the same binding site on βT but not paclitaxel.
  The K_d of colchicine was estimated to be 13.9±0.4μM ($\Delta G \simeq -6.90 \pm 0.03$ kcal/mol at 310K). The authors reported that this value is virtually identical to their experimental results of colchicine binding to αT and βT, suggesting an equal affinity of colchicine for both γT and βT [96].
  Docking simulations via Autodock on γT centroids of several clusters estimated that colchicine gave a range of blinding energy $-8.7 \div -7.9$ kcal/mol [96]. It is important to note that Autodock has no optimized algorithms and can return a significant standard deviation (±2.5 kcal/mol).
  On the other hand, the estimated energy of colchicine and derivates ranges $-8.0 \div -8.6$ kcal/mol, demonstrating comparable results in binding energy to colchicine.
  However, because of its toxicity, colchicine didn't pass the clinical trials as a βT inhibitor; authors suggested that a more targeted γT inhibitor might be able to have a lower dosage for the therapeutic window. Furthermore, the authors speculated that a possible reason for colchicine toxicity is the inhibition of γT itself.

[30]BIF peptide is typically used, and it is an RNA helicase A fragment that binds to the 1650÷1800 AA sequence of BRCA1

Regarding the combretastatin A-4, K_d was not calculated; however, another study reported that combretastatin A-4 binds to βT with a $K_d = 0.40 \pm 0.06 \mu M$, thus such compound has a higher affinity toward βT than colchicine [129].
Docking simulation via Autodock returned as best binding energy, a value of -5.9 kcal/mol, while most binding poses were deep inside the hypothesized colchicine binding pocket.
Authors suggested that this site could be a good pocket on which to base designing targeted derivates [96].
Generally, further work is still required to experimentally confirm the binding location to γT. Nevertheless, it is important to note that γT is far less abundant than βT, thus comparable binding affinity. At the same time, the hypothetical compound binds both proteins. Theoretically, there is a high probability that the compound will bind more frequently with βT rather than γT.

- Citral has to be found that binds to γT; however, studies focused on plant γT, which showed by immunostaining disrupted MTs. Interestingly, these effects are more pronounced during mitotic phases [130].
  After 2 h of treatment with 1 μM citral, 50% of the cells in interphase remained with normal MTs, while less than 10% of cells in different stages of mitosis preserved MTs. However, several studies tested the citral toxicity on human HeLa cells and rat embryonal fibroblast cells did not result in similar effects: partial MT disruptions were observed unless of higher dose and/or higher time administration were given [130, 131] A possible reason may be due to the higher complexity with which animal γT is wired.
- Citral dimethyl acetal (CDA), a derivative of citral, has been shown that increase E2F activity without affecting microtubules but trigger an accumulation of cells in the G2-M phase, showing a cytotoxic effect in cells with a nonfunctional RB1 pathway.
  It has been demonstrated *in silico* and *in vitro* experiments that CDA prevents GTP binding to γT. Inhibition of the MT nucleating activity of γT has been proposed to specifically target malignant tumours without affecting healthy cells [103].

  In a study, luciferase reporter assays showed that U2OS cells treated with citral showed an increased baseline luciferase activity caused by the endogenous activity of E2F in a concentration manner. Additionally, citral treatment also increased RB1 protein levels.
  This suggests that citral is an inhibitor of the nuclear activity of γT since the effect is similar to that caused by the reduction of γT protein levels via shRNA [132].
- A resveratrol[31] derivative, the 3,4,4'-Trimethoxystilbene (3,4,4'-TMS) has been observed that has binding affinity versus γT by 5.5-fold compared to αT and βT and ability to alter MT polymerization dynamics in cancer cells triggering

[31] 3,5,4'-trihydroxy-trans-stilbene

multipolar spindles and mitotic arrest leading to apoptosis due to mitotic catastrophe [133].
Computational analysis, Traversi et al. reported that 3,4,4'-TMS binds to the same combretastatin e colchicine binding sites and disrupts PPIs between two γT by getting in the way.
However, the authors used crystallography (PDB: 3CB2) in which the two γT interface in a "wrong" way: observing the interfaces in another model such as the complete γTuRC (PDB: 6V6S), the γT are not directed toward each other (PDB: 3CB2), but toward the corresponding GCPs. For this reason, this compound is not being considered until further investigation is done.

- 9'-bromonoscapine is a derivate of noscapine, a phthalide isoquinoline alkaloid, a non-addictive opioid with an antitumor activity that easily crosses the blood-brain barrier. It has been observed in βT that it binds in the proximity of the colchicine binding site [134].
  Docking and MD simulations showed that it also binds with high affinity into a pocket located at the binding interface between two adjacent γT [135].
  However, the same previous problem was also encountered here (the same crystallography models, PDB: 3CB2).
- Gatastatin G2, a derivative of glaziovianin A (AG1) and the next version of Gatastatin, is derived from a screening of a collection of colchicine ligands by replacing several chemical groups, such as the methoxyl group in O7 with the phenylmethoxy group and methoxyl group in O6 with a propargyloxy group [136, 137].
  The new molecule resulted in better selectivity and higher binding affinity to AG1 towards γT than αT and βT and effectively inhibited the MT nucleation during G2/M phase.
  Altered and shorter spindle formation, which leads to misaligned chromosomes and the complete inhibition of γT-dependent centrosome-induced aster formation, were observed.
  Investigations made on the previous version of gatastain reported through docking and dynamic simulations that the ligand binds in the GTP-binding site of γT [118]. Additionally, the gatastatin binding changed the γT surface that could prohibit the interactions with αT, inducing tubulin polymerization arrest.
  Nevertheless, there is not much agreement in the experimental results. It is not so sure that it binds the GTP binding site specifically to γT and not to others such as αT and βT, since the GTP binding site is highly conserved and similar among all tubulins.

Disfunctions and the Glioblastoma Multiforme

Dysfunction and critical mutations in γT, αT, βT, GCP2-6 or luminar bridge proteins can lead to alterations of MT nucleation, which changes the overall organization of the cytoskeleton and cellular physiology causing tumorigenesis and/or brain deformations [95, 98, 124, 136, 138–141].
For example, mutations in gT can cause indirect effects on the MT architecture to the extent that kinesis-5 and kinesin-14, proteins known to "walk" on MT as railroads

in both the direction, are not able to perform their roles anymore, as if the steps, being now different, does not allow the proper movement of the "feet" [102].

In most cancers, such as glioblastoma, astrocytoma, lung, ovarian and prostate cancer, γT and/or GCP2-3 are often found to be overexpressed: high amount of γTuRC leads to a higher amount of MTs which induce centrosome amplification (See Sec. 1.6), thereby, causing increased invasiveness in cancer cells.
Moreover, another study reported that the γT is co-distributed with βT3, which is also overexpressed. Similar to what happens with the Tau protein, in glioblastoma, as in many other brain tumours such as medulloblastoma, many γTs incorporate into insoluble aggregates [93].

One of the diseases in which mutations in *TUBG1-2* are involved is the Glioblastoma Multiforme which constitutes 16% of all malignant primary brain tumours and 54% of all gliomas, with a survival rate of 28.4% after one year from diagnosis, 3.4% after 5 years.
The main issue of this type of tumour is that there is currently no effective pharmacological therapy, and nowadays, the gold standard, which remains unchanged from 2004, is radiotherapy alone or radiotherapy with concomitant temozolomide (TMZ) chemotherapy followed by six cycles of adjuvant TMZ treatment [142]. In previous studies in which the transcriptional profile of a wide range of brain tumours have been compared, genes encoding for cytoskeleton-related proteins (including γT) are expressed at higher levels in grade IV tumours compared to normal brain, and lower grade tumours (grade I and II) [96, 143].

No works have been found reporting that specific mutations in *TUBG1-2* genes, rather it is more likely to be a set of factors leading to tumour formation rather than individual mutations.

Ivanova et al. have found 5 *TUBG1* variants which express the relative mutated proteins TYR92CYS, SER259LEU, THR331PRO, and LEU387PRO: these mutations cause cortical abnormalities, termed as Malformations of Cortical Development (MDC) [95].
Experiments on mice reported that the MDCs are due to disrupting neuronal migration, probably caused by defects in MT architecture and/or its dynamics. At the same time, there is no evidence of alteration or mislocation of the centrosome, which, in the case of anomalies, is associated with tumours.

1.6 Centrosome and MT-organizing centers

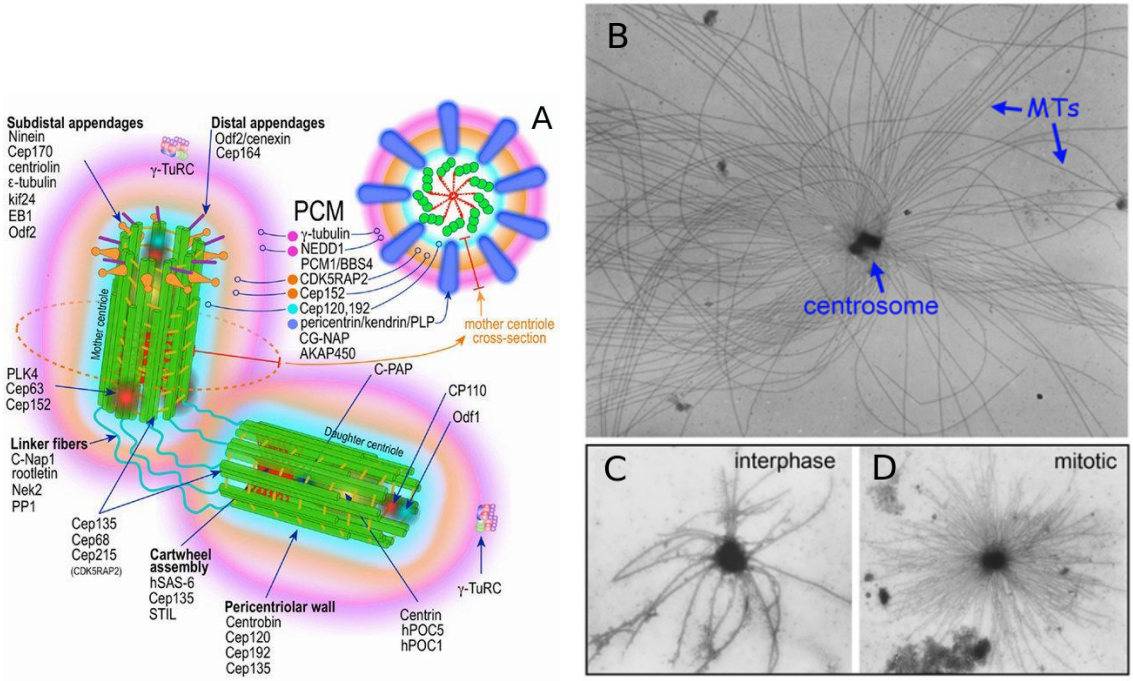


Figure 1.14 – *(A) Architectural features of the vertebrate centrosome after the centriole disengagement and before the procentriole nucleation [G.Pihan, 2013]. (B-C) Electron microscopy of isolated centrosome with MTs in vitro: most of MTs start from the centrosome (B) and the organelle changes states during cell cycle (C) [R.Kuriyama, 1981].*

The centrosome is a membrane-free organelle and contains hundreds of proteins, most of which have important functions in the cell cycle, especially in G2/M progression, in which the centrosome controls the number, polarity and distribution of MTs [109, 144].

Among the proteins that constitute the centrosome, there are the two centrioles called the mother and daughter centriole (each composed of 9 sets of MTs, which in turn are composed of a triplet of protofilaments composed of α-/β-T dimers) and the proteins of the Pericentriolar Matrix (PCM), which surrounds the centrioles. It does not have a well-defined structure but is instead an amorphous and fibrous matrix. In the outermost layer of the PCM, there is a high concentration of γTuRC that act as anchors to the overlying MTs (Fig.1.14 A-B).

Centrosomes are the main MTOCs in cells: MTs are part of the cytoskeleton and perform among the most important basic functions such as polarity, ordered transport of vesicles powered by motor proteins, motility, maintenance of cell shape, and cell division [91, 109, 110, 144].

MTs are intrinsically dynamic, as they stochastically oscillate between periods of growth and depolymerization in a process known as "dynamic microtubule instability" and this phenomenon is most pronunced during the mitosis phase [91] (Fig.1.14 C-D).

For each cell cycle, centrosome duplication is divided into several phases beginning with interphase, the first phase of mitosis[48, 98, 109, 110, 124, 145]:

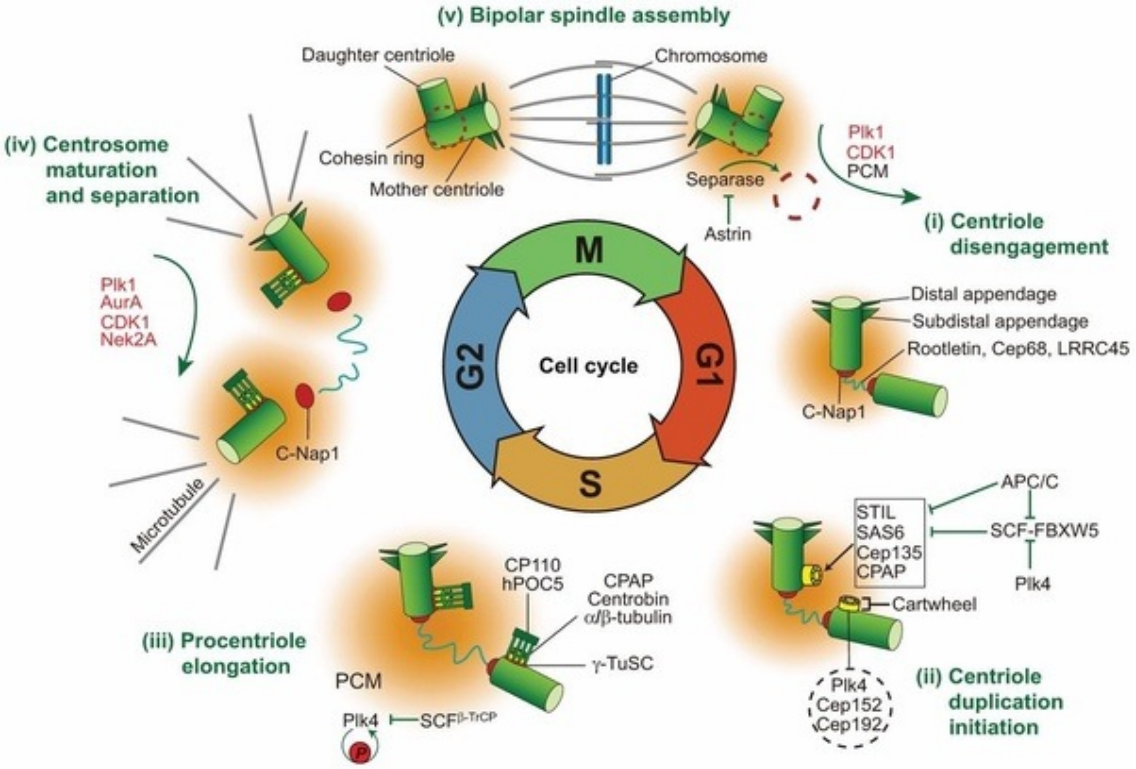


Figure 1.15 – *The different phases of centrosome duplication during cell cycle [G.Wang, 2014]*

phase G1) The mother and daughter centrioles physically separate but remain connected by a linker and become two mother centrioles; at the end of the G1 phase, following the phosphorylation of several centrosomal proteins, there is the formation of the procentriole in the sidewall of each mother centriole (Fig.1.15-i-ii).

phase S) Daughter centrioles are generated from each procentriole (Fig.1.15-iii).

phase G2) The linker between the parent centrioles is degraded and new PCM is forming (Fig.1.15-iv).

phase M) The cell has 2 mature centrosomes that localize next to nuclei. At the beginning of mitosis, the 2 centrosomes head towards the opposite poles of the cell, making up the spindle poles. From the surface of each spindle pole, MTs grow up several MTs: some of them are called asters and are mostly directed toward the cell membrane, where they will then anchor to it to properly guide the positioning and orientation of the mitotic spindle apparatus and subsequently to govern cell division[32]; while others MTs are directed toward to single kinetochores[33], ensuring proper segregation of chromosomes and triggering the metaphase (Fig.1.15-v,Fig.1.16).

[32]Although astral MTs are not required for the progression of mitosis, they are still required to ensure the fidelity of the process-determination of cell geometry.

[33]Mechanism known as *merotely*

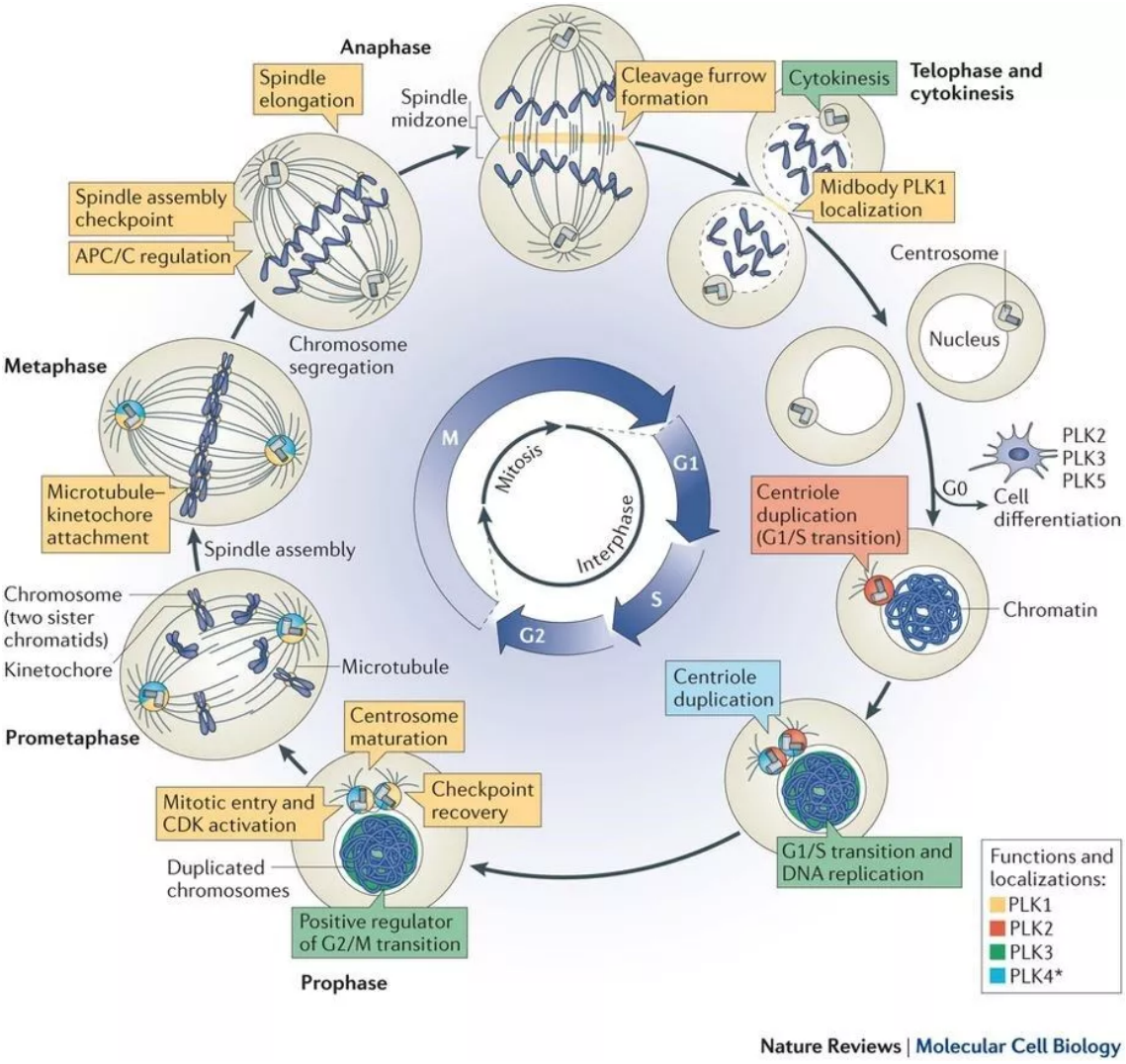


Figure 1.16 – *The different phases of the cell cycle viewed from the entire cell [S.Zitouni, 2014]*

In cancer cells, most common in the breast, prostate, bladder, colon, and brain, there is a strong correlation between centrosome instabilities and altered levels of centrosome-associated proteins [138, 145]. For example, centrosome amplification, in which more than two centrosomes are formed within the same cell, is associated with defective spindle formation, chromosome instabilities, aneuploidy, telomere shortening, chromosome breakage, and abnormal karyotypes and its frequency often increases during tumor growth and metastasis [138, 139, 145–148]. Moreover, centrosome amplification leads also to the remodeling of cytoskeleton and cell-cell adhesion due to the formation of dense microtubular arrays that promote cell polarization during directional migration, significantly increasing the invasiveness and thus drug resistance in cancer cells [91, 124, 139, 149].

During the interphase, the multiple centrosomes remain next to the nucleus. Still, as the cell enters in the prophase, the first stage of mitosis, these centrosomes start scattering across the cytoplasm. After that, with the loss of the nuclear membrane in the prometaphase, the multiple mitotic spindle poles are formed and split aberrantly

from the chromosome. Although chromosome missegregation occasionally leads to apoptosis, these mechanisms are the primary cause of heterogeneity among tumor cells. To prevent apoptosis during the metaphase, the multiple centrosomes can be clustered as if they were two normal centrosomes so as to form the 2 pseudo spindle poles at opposite poles of the cell interior (Fig.1.17) [139, 145].

The main cause of centrosome amplification is that the mother and daughter centrioles decouple themselves before reaching mitosis (phase M), so from the procentrioles of each mother centriole, additional daughter centrioles will be formed [91, 124, 149].

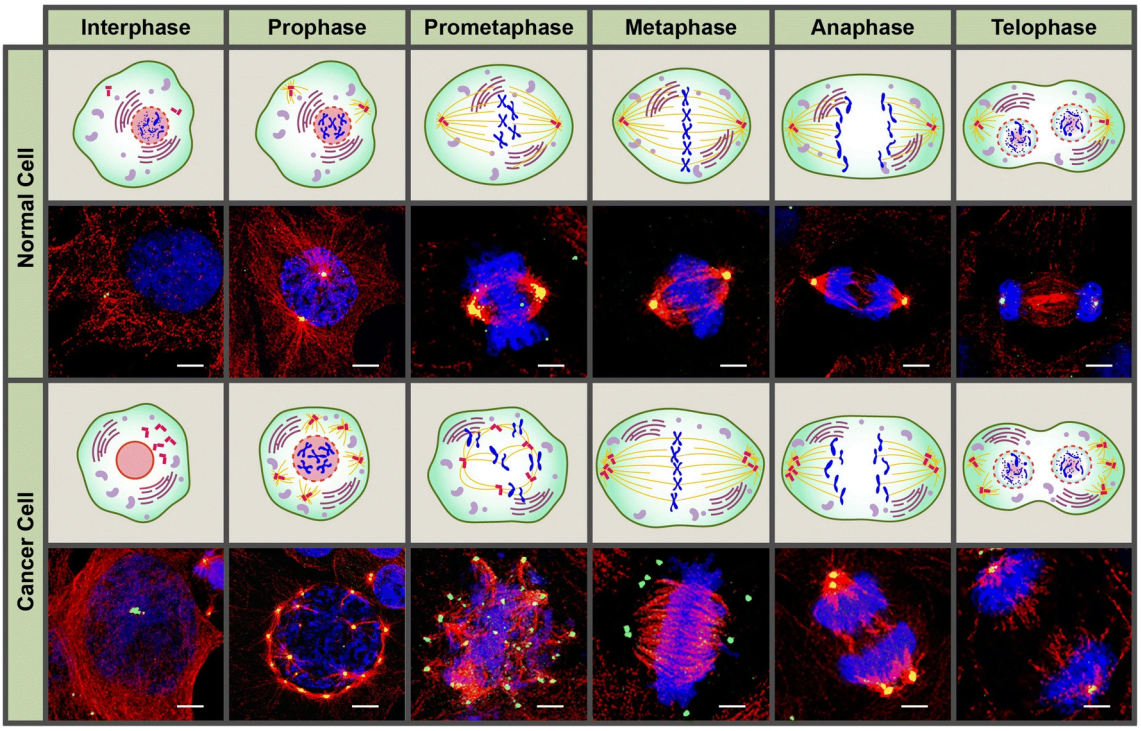


Figure 1.17 – *Confocal immunomicrographs and graphic illustration of a normal and cancer cell with extra centrosomes across the different phases of the cell cycle. Green: γT; red: αT; blue: DNA. Scale bar = 5 μm [K.Mittal, 2020].*

1.7 Computer-Aided Drug Design

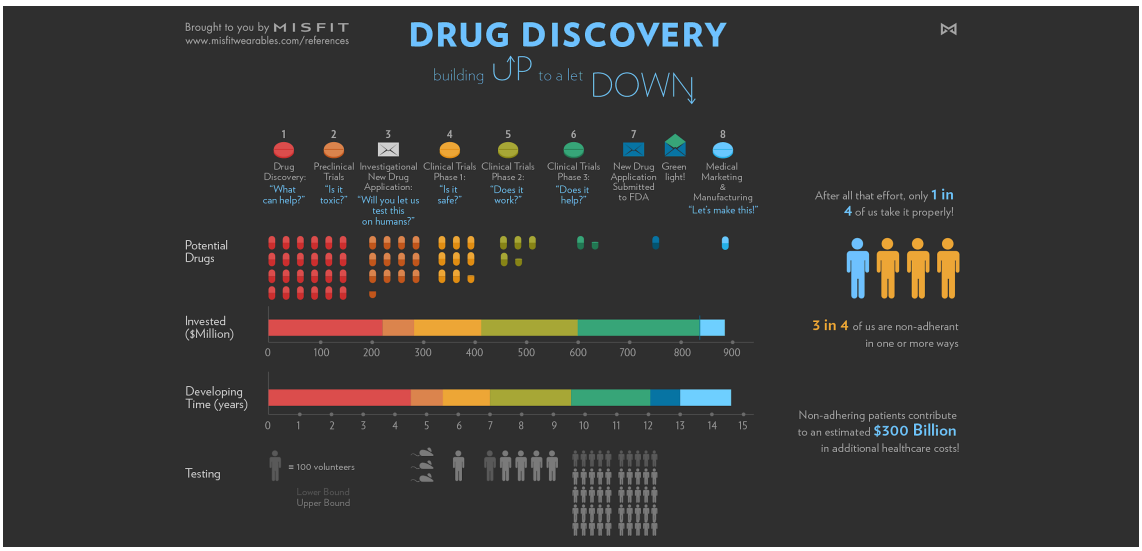


Figure 1.18 – *Timeline and money invested in Drug Discovery.*

Although the rapid developments in the last 20 years in the combinatorial chemistry fields and HTS technologies that allowed to synthesize and screen vast libraries of compounds against a molecular target in a very short time, a significant amount of challenges still occurs, wasting a considerable amount of money and time.
This happens because chemical approaches like HTS are a "brute-force" way to explore the bioactivity of compounds, thus low rate of success in finding compounds against a specific protein target with acceptable biological properties, described briefly as Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) [150, 151].

Generally speaking, developing therapeutic drugs is highly expensive (with a range of 0.5÷2 billion dollars) and time-consuming (with a range of 10÷20 years of development depending on the drug target) (Fig.1.18) [151–155].

Approximately 50% of this process includes drug discovery and optimization and pre-clinical testing (*in vitro, in vivo* and in animal experiments). On the other hand, the other half of drug development involves instead predominantly clinical trials, regulatory filings and post-marketing surveillance: the latter part is generally the most time-consuming because of long observations of patients, while the former is the most expensive, mainly because of needed reagents to synthesize the compounds and animal models [153, 155].

But in the last decade, along with new progress to replicate more realistic and personalized[34] biological systems such as bioreactors and multi organ-on-chip, the computer-aided drug design (CADD), or as *in silico* methods, has become a true essential step in early-stage drug discovery both at academic and industrial/pharmaceutical levels.
In fact, for example, the VS, which filter out from a massive database of compounds,

[34]In these systems, drugs can be tested directly on patients' cells.

whose low binding affinity, detrimental ADMET, and/or other properties are predicted to make them ineffective; in this way, it is more efficient to find potential hit candidates [150, 151].

After this early-stage, the filtered databases will subsequently be subjected to HTS, biological assays and/or other typical laboratory experiments, thus allowing more focus on the optimization of a few successful compounds and, most importantly, saving a significant amount of money and time [150–152].

A clear and recent example was the research behind the work on vaccines development against the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) that intensively involved computational immunology, reducing the research time to months instead of years typically required to develop vaccines [156–158].
Indeed, mainly through crystallography of SARS-CoV-2, which became immediately available after the identification of the virus, it was possible to find virtually the epitopes of the Spike[35] protein, predict mutations and how virus behaviour changes, predict the PPIs between Spike proteins and candidate neutralizing antibodies to be used as a vaccine, and so on [156, 158].

There are two main categories of VS [150, 151, 160]:

- Ligand-Based Virtual Screening (LBVS), which correlates somehow the structure/physicochemical properties of a ligand (ex. hydrophobicity) with known biological activity (Ex. ADMET, binding affinities and/or inhibitory concentration IC_{50})[36];
- Structure-Based Virtual Screening (SBVS) which retrieve information from the structure of the protein, usually obtained from several forms of imaging such as X-ray crystallography or modelling such as Homology Modeling.

[35]The membrane glycoprotein expressed on SARS-CoV-2 that binds with high affinity to the cellular receptor angiotensin-converting enzyme 2 (ACE2), highly expressed on human airway epithelial cells, for which translocation to endosomes occurs and then the viral genome is released [156, 159].

[36]Common LBVS methods are similarity and substructure searching, Quantitative Structure-Activity Relationships (QSAR), and pharmacophore 3D-shape matching

1.7.1 The Docking

The SBVS approach is more robust than the LBVS approach because it takes into account further details like features of POI.
The most common SBVS method of which it is mainly used in current work, is the docking: it estimates the binding free energy ΔG between different conformations[37] of a given ligand and a binding site of the protein target in the most diverse and reasonable ways.
It then ranks the definitive poses based on docking score which is an estimation of the binding affinity[38].

Despite the significant advantages of CADD, in particular VS when a high number of compounds are involved, this is not without challenges: the most critical issue of this type of research is the generation of a considerable number of false positive[39], defined as False Positive Rate (FPR) (Eqn. 1.1) [150, 151, 161, 162].

$$FPR = \frac{FP}{FP + TN} \tag{1.1}$$

where FP = False Positive, TN = True Negative.

For example, the most evident problem is in the docking approach itself, although it is one of the most SBVS used today.
In general, many docking protocols rely on 2 main components [151]:

- search algorithms: they explore iteratively different conformations of ligands at the binding site of the POI based on the translational and rotational degrees of freedom[40];
- scoring functions: they are mathematical equations to estimate the energy/force of non-covalent interactions between a ligand and a POI; additionally, it is the one on which the final predicted results of docking depend.

However, it is possible to obtain different results from different software using the same input or, on the other hand, it could return poses different from others although the similar docking score with no way to distinguish which one is correct [151, 162, 163]. One of the several reasons is the different nature itself of underlying the scoring function; there are three main types of scoring functions commonly used in docking programs, of which there may also be hybrids[41], plus another more recent one that will be discussed directly in the dedicated subsection 1.7.5 [150, 151, 160, 162, 163]:

[37] The top ligand conformations after docking are called as *poses*

[38] Note that high affinity means with low values of binding ΔG; conversely, low binding affinity, high ΔG values.

[39] I.e. compounds with the valuable score but disagree with the experimental binding affinities tested *in vivo* o *in vitro*

[40] Three macro types of search algorithms exist: systematic (changing the degrees of freedom gradually), stochastic (Monte Carlo and Genetic Algorithms) and deterministic (Energy Minimization (EM) and/or MD)

[41] The docking software Vina is knowledge-based + empirical, while DOCK6 and AutoDock4 is force-field + empirical [164–166]

- *Force Field-based*: the ΔG is obtained by an estimation of potential energy, which is a sum of intermolecular terms, such as electrostatic/coulombic and van der Waals (Leonard-Jones potential) interactions, hydrogen bondings and intramolecular terms such as stretching/bending/torsional bindings (Eqn. 1.2 is the standard equation of potential energy). The coefficients are obtained from experimental data. Sometimes, contributions such as solvation and entropy are also considered at the price of higher computational costs.
  Docking MOE's tool fall into this category [167].

$$
\begin{aligned}
V(r_1, r_2, \cdots, r_n) = & \sum_{bonds} \frac{1}{2} k_l [l - l_0]^2 + \sum_{angles} \frac{1}{2} k_\theta [\theta - \theta_0]^2 + \\
& \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\varphi - \delta)] + \sum_{\substack{improp \\ dihedrals}} \frac{1}{2} k_\zeta [\zeta]^2 + \\
& \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left\{ \frac{q_i q_j}{(4\pi\varepsilon_0 \varepsilon_r r_{ij})} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}^{12}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\}
\end{aligned}
\tag{1.2}
$$

  - Bond term: k_l = Force constant (kcal/mol Å); l_0= equilibrium distance (Å).
  - Angle term: k_θ = Force constant (kcal/mol rad^2); θ_0 = Equilibrium bond angle (rad).
  - Dihedral/torsional term: V_n = Energy barrier for the rotation around a given bond (kcal/mol); φ = Dihedral angle (rad); δ =Phase offset (rad); $n = [0 : 2\pi]$.
  - Electrostatic/Coulomb: q = charge; ε = dielectric; r_{ij} = Interatomic distance (Å).
  - Van Der Waals term: ϵ_{ij} = Strength of the interaction (kcal/mol); σ_{ij} = Van der Waals radius (Å).

- *Knowledge-based*: the ΔG is a sum of the statistical score (based on observations performed on large databases of protein-ligand complexes with known 3D structures) of all individual intermolecular interactions of atom pairs or functional groups that occur more frequently.
- *Empirical*: the ΔG is estimated as a weighted sum of interaction terms like hydrogen and ionic bonding, hydrophobic contacts, desolvation and entropic effects generated by fitting experimental binding affinity data for a training set of protein-ligand complexes via linear regression.
  Glide falls into this category [168].

To overcome the problem of disagreement in docking score across the different scoring functions, there are other possibilities to improve the quality of results. For example, using more accurate scoring functions that consider further terms for entropy and solvation is the most explicit way. Still, it is not very easy to obtain further experimental parameters.

Otherwise, another possibility is running MD or docking software that integrate quantum calculations like the new GROMACS package CP2K (a MD software), but those methods are computationally expensive even with a small number of compounds, therefore, should only be used towards the end of the VS, when few candidate compounds are found [150, 162, 169].

1.7.2 The Consensus Scoring and Docking, Score-based Consensus Docking and DockBox package

Other relatively straightforward strategies that exploit the disagreement of the results generated by different scoring functions or poses, especially if they come from different docking software, are Consensus Docking (CD) and Consensus Scoring (CS), which aim to increase the accuracy of a given dataset by filtering out false positives by aggregating the results of different used docking programs [151, 162, 163].

CD takes the best poses of different chosen docking programs and keeps those most similar based on a given RMSD (Eqn. 1.3) cut-off, which is usually 2 Å[162, 163].
Although this technique is not very used in VS, it has shown that predicted poses are more likely to correspond to native binding poses [151, 163].

$$RMSD(A,B) = \frac{1}{N}\sum_{i=1}^{n}(\vec{\boldsymbol{p}}_{A,i} - \vec{\boldsymbol{p}}_{B,i})^2 \tag{1.3}$$

where A and B are two structures (proteins or ligands), N is the total number of atoms[42] and $\vec{p_i}$ is the 3D position of atom i-esim.

On the other hand, CS is far faster than the former and other VS strategies since it is just a sort of aggregation of scores already calculated, and it can rather be done in several ways [162]:

- voting system based on counting the number of functions that reported a value within a predetermined threshold;
- averaging of ranks or binding affinities of different scoring functions;
- combining the different scores using fitted coefficients (regression).

J.Preto and F.Gentile introduced another simple algorithm that integrates both CD and CS, called Score-based Consensus Docking (SBCD).

It is integrated into the open-source package DockBox (repo available on GitHub), which is a relatively outdated and unsupported package that aggregates the results of one or more among seven available programs and/or uses their scoring functions to perform SBCD (AutoDock4, AutoDock Vina, DOCK6, DSX (only scoring function), Molecular Operating Environment (MOE), Glide (only docking), and Genetic Optimisation for Ligand Docking (GOLD)) [162, 164–167].
It is important to note that the user has to install manually, wherever possible, the desired docking programs and older version of Amber ($16 \leq x \leq 18$), required to minimize the poses energetically which will be rescored (in case it is set up).

When DockBox runs, it first starts individual dockings through the selected and installed programs, independently from each other, based on a configuration file in which settings are saved for each docking program chosen.
When poses are generated and energetically minimized (in case it is set up), then the second step is an evaluation of all generated poses based on chosen modalities [162, 163]:

[42] Note that the two structures must have the same number of atoms.

- CS: it takes the pose with the best definitive score computed as a linear combination of the scores calculated by single docking programs (Eqn. 1.4).

$$S = \sum_{i}^{n} \alpha_i \frac{s_i - \mu_i}{\sigma_i} \tag{1.4}$$

  where α_i = weight; μ_i and σ_i = mean and standard deviation, respectively, of the scores on all poses associated with the scoring function i
- CD: it performs the Houston method by selecting the best poses that are similar to each other (usually RMSD < 2 Å).
- SBCD: it firstly takes all generated poses[43] and then performs the scoring with all available scoring functions; after that, the top-ranked poses are subject to Houston Consensus Docking (usually RMSD < 2 Å).

In the end, DockBox returns the pose with the lowest RMSD whether it is successful during CD step.

Preto et al. employed a combination of Vina, Autodock4, DOCK6 and DSX and showed that the best minimal combination to perform the SBCD algorithm is Vina and DOCK6. Individually, DOCK produced a success rate of 68.7% while Vina was 41.7%.

Additionally, SBCD performed better than CD modality: 86.2% of success rate versus 69.3%, respectively (with Vina and DOCK6).

In conclusion, two key factors are important to consider during a VS campaign with DockBox:

- since it runs independent docking processes, it is also computationally expensive, thus this method is not very feasible when the database size is >1M, and even worse when the size is >1B;
- MOE can pre-generate (although not very differently) different 3D conformations as poses before the docking on its own, while the other software, including Vina, DOCK6 and Autodock, do not.
  Such programs take as input the ligand without pre-processing, thus, they need 3D conformations already processed as input files. For further details, see sec. 1.7.3.

[43] In the case of Consensus Scoring, it takes only the scores *already* calculated by the single docking program

1.7.3 The Conformational sampling

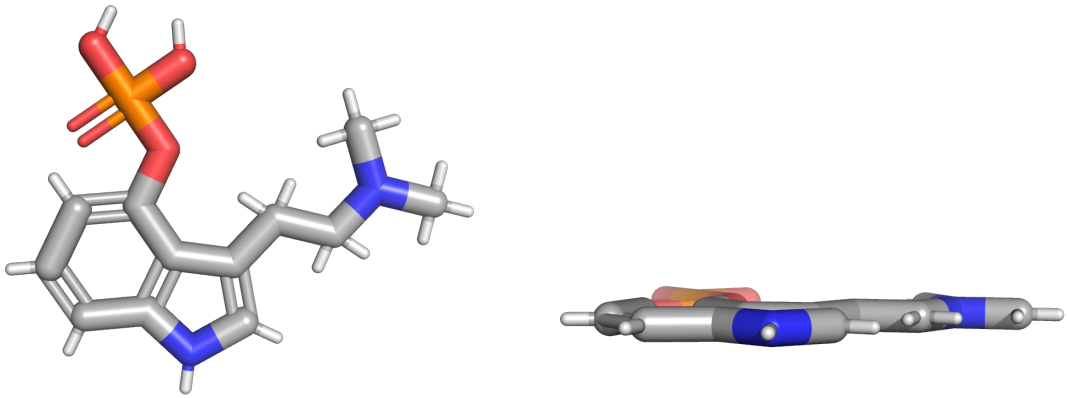


Figure 1.19 – *Example of a planar 3D conformation. Prepared with PyMol*

Generally, before running any VS methods that involve the three-dimensionality of ligands, such as the docking, it is good practice to do a conformational sampling for each of them mainly for two reasons.
First, it is well known that conformations significantly affect the activity, reactivity, and many other physicochemical properties [170].
Secondly, a large number of ligands contained in online databases, like the free-access ZINC15/20, are usually downloaded as SMILE strings because of lower memory, and faster download time.
However, such type of information does not contain information regarding spatial dimensionality; many programs like OpenBabel or MOE[44] convert the SMILE format easily into standard 3D file formats like .sdf or .mol2, but actually as a planar 3D conformation (i.e., one of the three dimension columns is all zero, as shown in the Fig.1.19).

Thus, running EM with whatever software is used (Amber, GROMACS, EM MOE's tool or others) directly from such conformations is technically wrong, even though the EM returns a 3D conformation that is no longer planar.

The reason is that starting from a planar conformation and then running the EM, it returns the nearest local minimum of the potential energy function. At the same time, there could be a local minimum with potential energy far lower than the one found from planar conformation (Fig.1.20A).

Therefore, running a conformational sampling with a stochastic algorithm is strongly suggested to generate different 3D conformations of the same ligand from which EM is then processed, thus collecting low energy local minima, which are often a good representative of the entire conformation space (Fig.1.20B).

In this way, better quality results are expected compared with the conversion-EM-docking procedure because more low-energy conformations of the same ligand can be used in the docking step, possibly leading to lower (thus better) binding scores.

[44]The conversion is automatic in the case of using the Graphical User Interface (GUI), but this is not technically feasible when over millions of compounds are involved; nevertheless, it is possible to convert manually through proper internal functions (see D

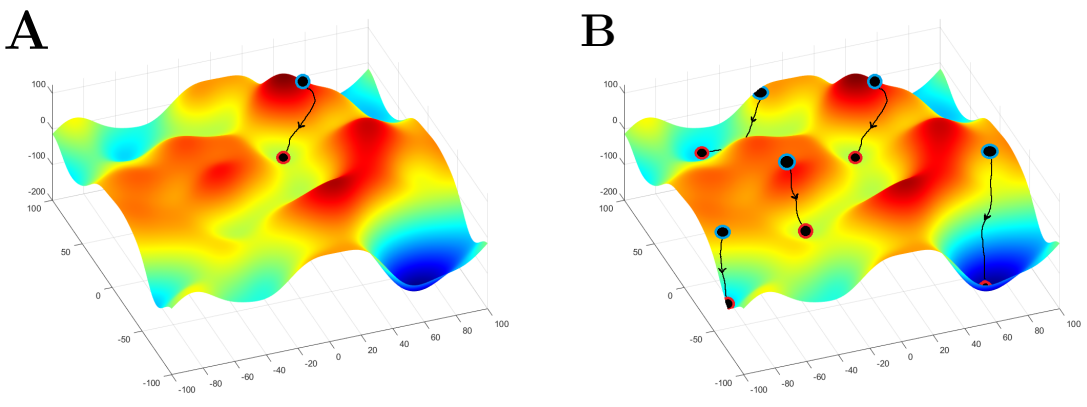


Figure 1.20 – *Example of a small fraction of Phase Space or Potential Energy Surface: (A) situation of running EM from planar 3D conformation obtained after format conversion from SMILE string; (B) situation of running EM after conformational sampling.*
Circles with blue border: Start points; circles with red border: End points after EM. Images prepared with Mysimlabs, a MATLAB package available online.

After acquiring the 3D conformations, in the case of large amounts of ligands, it is still not feasible running the docking directly because the docking itself is computationally expensive when a high number of ligands is involved.
Thus, reducing the size of such databases by filtering out somehow ligands that do not fit in a potential binding site is strongly recommended.

1.7.4 The Pharmacophore Filtering

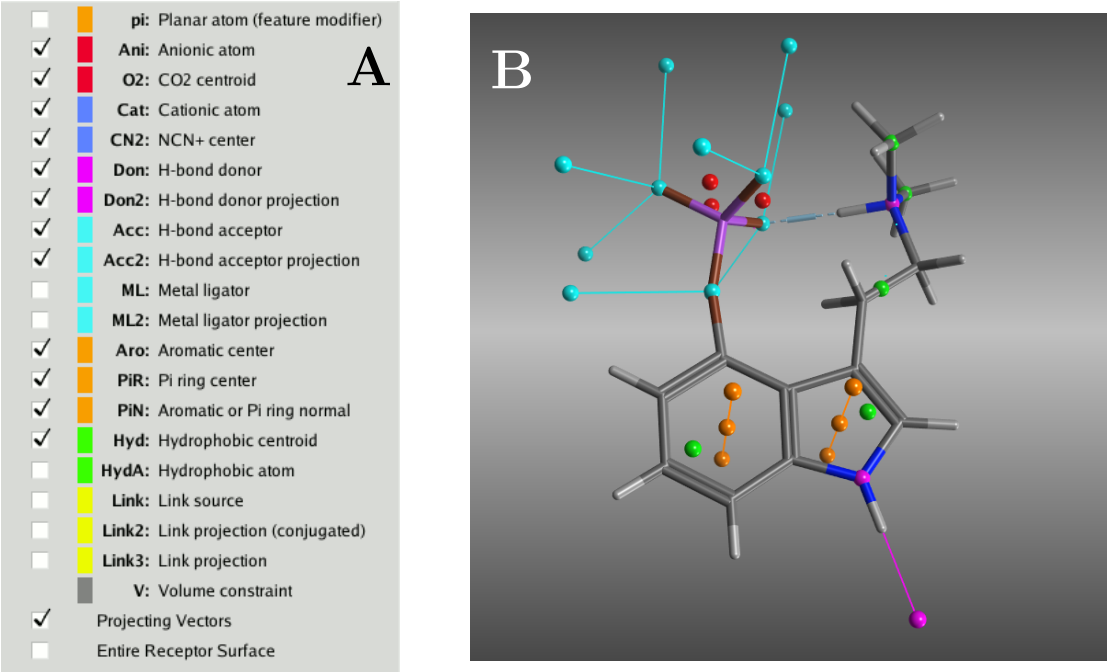


Figure 1.21 – *(A) Example of features available on Pharmacophore Editor MOE's tool; additional features available are Volume Excluded, Ligand shape and Volume Occupied. (B) How the pharmacophore features are shown on a generic compound: from this, it is possible to build a pharmacophore ligand-based model.*

A common technique in VS is pharmacophore filtering, which is supposed to remove as many false positives as possible, thus, also reducing the size of a given database and increasing the recall, i.e. the True Positive Rate (TPR) (Eqn. 1.6) [161].
This technique is widely used due to its simplicity, which enables the filtering of databases in a very short time, on the order of minutes for hundreds of thousands of compounds, when docking takes several hours/days. Moreover, it can be used anytime, before and after the docking.
Essentially, a pharmacophore model constitutes a three-dimensional ensemble of chemical and physical features manually added (Fig.1.21A) based on available information [150, 161, 171]. During a pharmacophore search, a ligand is declared successful if it has all or part[45] of the features satisfied. Or conversely, in the case of decoy-pharmacophore models, the ligand must not meet the feature constraints [151].

There are two main ways to build pharmacophore models, depending mainly on available data.
In case. Of the most valuable and straightforward way, if crystallography of ligand-protein complex is available, the non-covalent interactions, especially H-bonds, and

[45] In *Pharmacophore Editor* of MOE's tool, if the 'Partial Match' search is set, the ligand passes the filtering search if it satisfies a defined number of features

how the pose fills the site, remark important information that could be exploited. In this situation, it is technically satisfactory to building a pharmacophore ligand-based model by just taking into account features of this ligand (Fig 1.21A), or also features representing the aminoacids, thus a hybrid ligand-structure-based pharmacophore model (Fig. 2.2B).

Otherwise, if the complex protein-ligand crystallography is unavailable, another possibility is building pharmacophore structure-based models, i.e. models based on target protein only (Fig. 3.26A÷D).
Therefore, this method is firmly protein structure-dependent: it is not very efficient for those proteins or precisely for dynamically unstable candidate binding sites, so the spatial change of surrounding residues may make the features no longer applicable.

1.7.5 The Machine-Learning methods and DeepDock package

The approaches previously described are not designed to explore multi-billion compound databases. In fact, a generic docking run usually does not exceed 0.1 billion compounds [151, 160, 172].
Therefore a faster and more efficient method is mandatory if there is not even a basic knowledge or features derived from known compounds such that they can be exploited to reduce the size of such databases.

There is another less common scoring function type which is based on Machine Learning (ML) or Deep Learning (DL)[46], but in the last few years, the number of publications has exponentially increased, although there are few drugs developed in this way [151, 172, 174–176].
The main reasons for this trend are that some open-source libraries and frameworks are fairly well-established and of general use, such as TensorFlow and PyTorch libraries and, when possible, the access to supercomputers, also known as HPC clusters, with increasingly powerful processors are more accessible to universities and industries thanks to conspicuous government fundings and grants [160, 176–178].
Indeed, these systems have at least a thousand nodes with CPU and GPU units with high performances and storage systems with the possibility to save permanently or temporally an amount of data with orders of magnitude of even petabytes.

The "classical" score functions described earlier (1.7.1) are substantially all linear combinations of several coefficients and variables. On the contrary, the ML-based scoring function use ML methods that are mainly non-linear, such as Support Vector Machine (SVM)[47], Random Forest (RF), DNN, Convolutional Neural Network (CNN) and Graph Neural Network (GNN) [151, 160].

Many ML methods are widely used in QSAR analysis which, as discussed previously, aim to determine the relationship between structural/physicochemical properties of compounds under evaluation to their biological activity.

However, to be considered a good predictor, ML-based scoring functions require a large amount of data to prepare training, testing and validation sets needed to train and subsequently test the models [151, 175].
Nevertheless, this is not always possible: sometimes pharmaceutical industries are reluctant to publish specific data, such as unpatented ligands because they know that important information can be derived from such data with good accuracy with the potential of developing new drugs so that they can patent and profit.
Furthermore, if data is available but too numerous to be managed by personal computers or a single mainframe, HPC clusters are indispensable.

In the current work, a DL-based scoring function called DeepDock, developed by F.Gentile et al., is employed to reduce the size of the ZINC database (>1 billion compounds) iteratively by predicting the score of unprocessed ligands that bind

[46] ML is a more generic term than DL, but the latter is technically a ML with more hyperparameters, such as higher number of hidden layers, more complex architecture and so on, usually used for process high amount of data or when the data has high dimensionality. For more details, see the general comparative review of H.Alaskar and T.Saba [173]

[47] Linear and non-linear SVM classifiers both exist

selectively to γT [172, 176].

Additionally, the DNN models of DeepDock seem to enrich reduced databases with top-hits by considering the unfavourable hits from which it draws and penalizes certain chemical groups and incentivizing the good ones instead.
In contrast, in the case of generic docking, such hits are processed and then rejected, thus needlessly wasting a high amount of energy required to run docking software on HPC clusters.

Hence, to avoid generating bias in DNN models, it is not reasonable to perform on such databases pharmacophore filtering or filtering based on precomputed physicochemical parameters and drug-like criteria, such as molecular weight, volume, octanol-water partition coefficient, polar surface area, number of rotatable bonds and hydrogen bond donors and acceptors and others [172].

F.Gentile et al. have tested the effectiveness of DeepDock on 12 POIs belonging to 4 main drug-target families[48] and on SARS-CoV-2 main protease (Mpro) [172, 176, 179]. They were able to reduce a >1B database to a predicted and enriched database with a size range of 250K $\div$ 1M compounds and a recall (Eqn. 1.6) $\sim$ 0.9 after 9 $\div$ 11 iterations with datasets of $\div$ 1M each.
Additionally, A.Ton and F.Gentile further experimentally tested the final top-hits of DeepDock on Mpro and identified that 15% of them were active against the target, with the established IC_{50} values ranging from 8 to 251 μM [176, 179].

There are four key factors to consider that affect somehow the DeepDock performances:

- the training of DNN models strongly depends on the used docking program. Indeed, they correlate the docking score generated by the docking program with the 2D descriptors of the ligands, thus meaning that biases are unavoidably present among datasets if only one docking program is used.
  Due to time constraints, only *Docking* MOE's tool is used in the current work, but the ideal approach to overcome this issue is performing Consensus Docking and Scoring for each dataset before training the DNN models.
- as a consequence of the previous point, the DNN models are strongly dependent on the binding site of the POI.
  This is both an advantage and a disadvantage because it means that the models can only be used on a specific binding site since they are trained on the docking scores between ligands and that binding site.
  Thus, if there are multiple binding sites, multiple independent runs of the training of the DNN models must be done, meaning higher computational costs and challenges in handling large amounts of data.
- The size of training size significantly affects the performances of DeepDock.
  F.Gentile et al. showed that after 4 iterations with an original database of >1B, the predicted and reduced database were $\sim$ 58M and $\sim$ 108M for two training sets of 700K and 350K respectively. Moreover, the reduction in the case of 350K training set was similar to 700K only at 8^{th} iteration (Fig.1.22) [172, 176].
- The recall (Eqn. 1.6) must be set cautiously: if too high (that might seem

[48] Nuclear receptors, kinases, G protein-coupled receptors and ion channels

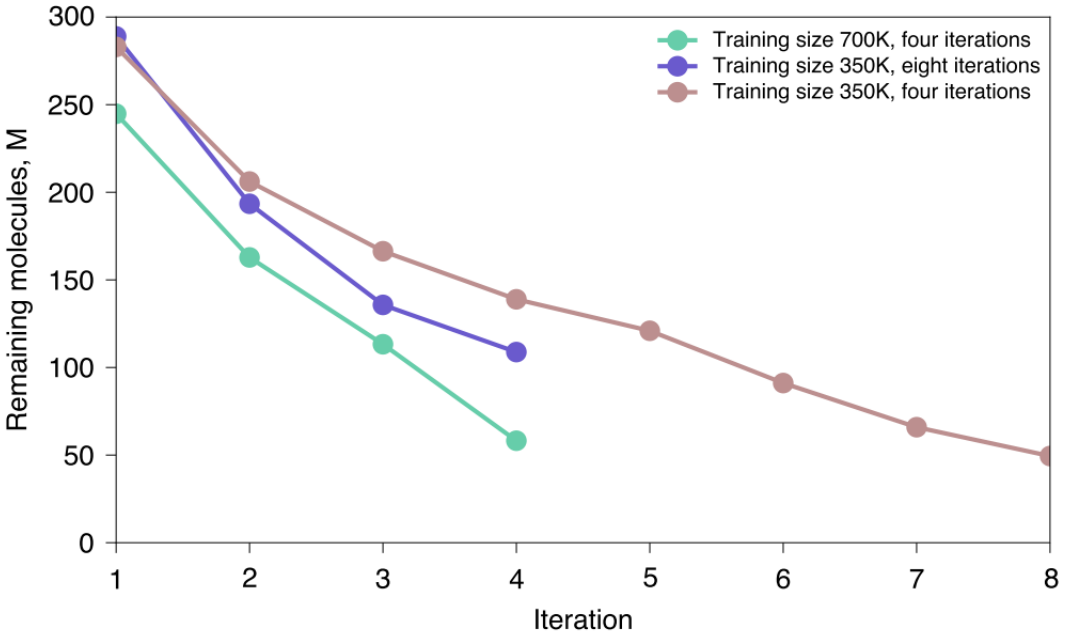


Figure 1.22 – *Effect of varying the training size and number of iterations on the number of remaining molecules for screening ZINC20 against the dimerization site of androgen receptor (PDB: 1R4I) (F.Gentile, 2022)*

beneficial), the size of predicted databases could still be too high; while on the other hand, if too low, a large portion of virtual hits may be discarded.

More technically speaking, a model of DeepDock is a binary classifier in the form of feedforward (multilayer perceptions) trained on a 1,024-bit circular Morgan fingerprint.
Given a cutoff, ligands with a score below that cutoff are labelled as virtual-hits (i.e. positive samples), while ligands with a score above the cutoff are labelled as non-hits (i.e. negative samples). This is the reason for the engagement of the binary classifier.

In a nutshell, DeepDock's protocol is described as follows [172, 176]:

1. ligand-based QSAR descriptors (such as molecular fingerprints) are computed for all the ligands of a given database;
2. the datasets (training, test and validation sets[49]) are prepared by randomly extracting a given number of ligands from the database;
3. conventional docking is performed by using the previously extracted datasets;
4. the docking score of each ligand is extracted and associated with its corresponding 2D molecular descriptor;
5. the DNN models are trained: the labels "virtual-hits" and "non-hits" generated for the given dataset are correlated somehow with the indexes of Morgan Fingerprints;

[49]The test and validation sets are generated only in the first iteration and will be the same throughout all iterations.

6. the DNN models predict docking scores of unprocessed ligands contained in the original database;
7. the training set is generated again by randomly extracting from the predicted database of the previous step;
8. the iteration ends and is repeated from point 3 until a given number of iterations is reached or when the size of the predicted database converges.

A proper DD training set should effectively reflect the database's chemical diversity. In fact, as in all ML methods, not only in CADD, the preparation of the datasets makes the difference in the outcomes, so they need to be prepared very carefully to minimize as much as possible bias.

There are several evaluation metrics to keep track progress of the DNN model performances throughout several iterations, but two, in particular, are those that describe well the overall progress: the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) [151, 160, 173].
The ROC curve tracks TPR and FPR and represents the performance of a neural network model, while the AUC, that is a value within the range of $0 \div 1$, is an estimation of the probability that a result is true.
Thus, generally, one has to have a model with AUC as high as possible and is considered accurate when AUC > 0.9. On the contrary, a poor model with an AUC=0.5 is considered random [151, 173].

Another parameter, which is used in Gentile's works, is the Full Predicted Database Enrichment (FPDE) (Eqn. 1.8): like AUC, it evaluates overall DD performances.

$$recall = \frac{TP}{TP + FN} \tag{1.5}$$

$$precision = \frac{TP}{TP + FP} \tag{1.6}$$

$$random\ precision = \frac{TP|_{database}}{total\ molecules|_{database}} \tag{1.7}$$

$$FPDE = \frac{precision}{random\ precision} \tag{1.8}$$

where TP = True Positive, FN = False Negative, FP = False Positive

Chapter 2

Materials & Methods

The following work covers a large number of seemingly unconnected parts; therefore, the following picture briefly describes what was done in the following work.

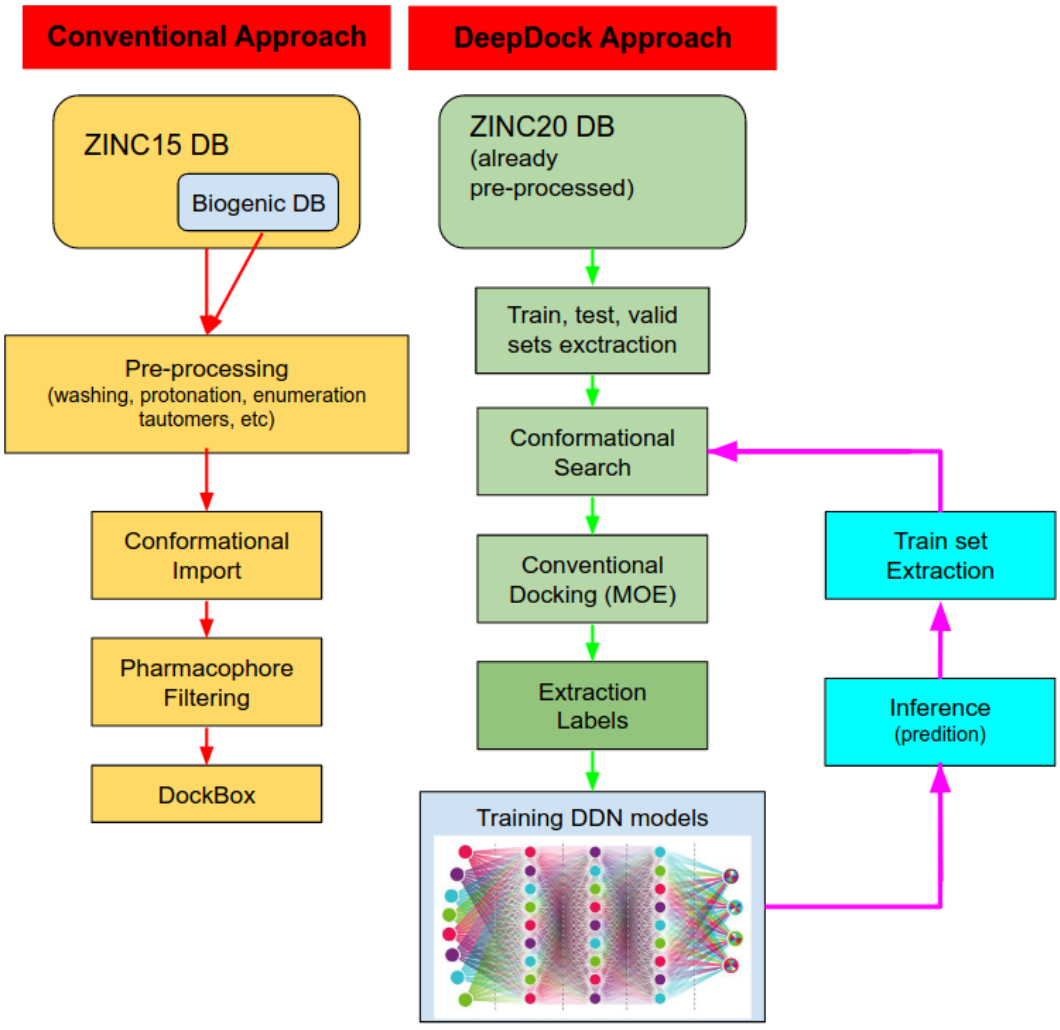


Figure 2.1 – *A synthetic scheme of the current work*

2.1 Identification of Binding Sites on Protein of Interest

2.1.1 Preparation of target proteins

The αT, βT and γT PDB models were kindly granted by P.Vottero, who processed[1] them earlier during her master's thesis (available on PoliTo master thesis database). The original structures of tubulins are obtained from Protein DataBank and the PDB codes are:

- 3CB2 for γT (resolution: 2.71 Å);
- 5EYP for αT and βT (resolution: 1.90 Å)[2].

γT is further processed by adding missing aminoacids in the C-terminus (GLY-THR-GLN-GLU-GLN) and N-terminus (MET). Using this model as a template, the model of the subtype encoded by *TUBG2* gene is built by the Homology Modeling tool integrated into MOE. All proteins used are then assigned the hydrogen atoms, partial charges and protonation state for the aminoacids through proper MOE tools.

As discussed previously (See Sec 1.5Several studies claim that the presence of GDP or GTP has a role beyond the function of the αT and βT as individual proteins, without significantly affecting the structure [73, 116, 117]
Therefore, to verify if the absence or the presence of such nucleotides significantly alters the overall structure of γT-1 and γT-2, a MD of 100 ns, clustering, RMSD and Solvent Accessible Surface Area (SASA) calculation are done to compare the differences among γT-only, γT+GDP and γT+GTP.

Briefly, the models used in the following work are:

- γT-1 only;
- γT-1 + GDP;
- γT-1 + GTP;
- γT-2 only;
- γT-2 + GDP;
- γT-2 + GTP;
- βT-2a + GDP;
- βT-2b + GDP;
- βT-3 + GDP;
- βT-4a + GDP;

[1]The structure were prepared through *Structure Preparation* MOE's tool. The major fixes are (1) residues with alternate locations were corrected by using the highest occupancy one; (2) missing backbone atoms in the protein chain C- or N- termini were deleted, and the terminus was capped; (3) missing residues inside the chain were corrected by building a loop and (4) inconsistencies between the residue name and its structure or missing atoms were corrected

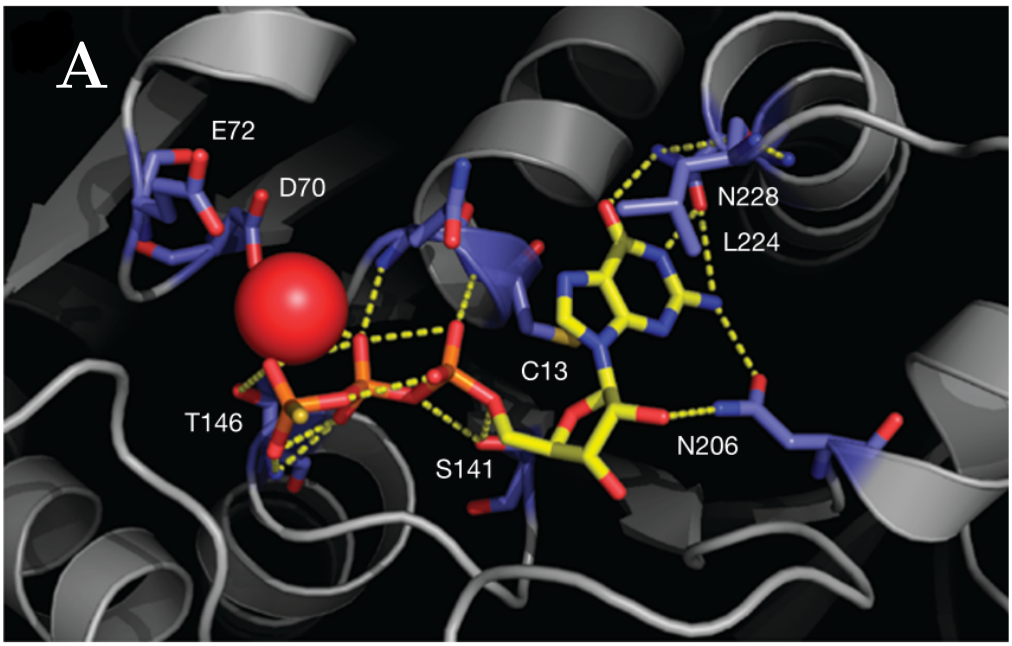
[2]The αT present in the original crystallography is expressed by gene TUBA1B and only that subtype was used. On the other hand, the βT present in the original crystallography is expressed by gene TUBB2B and it was used as a template in Homology Modeling to build the other βT subtype models for the genes TUBB2A, TUBB3, TUBB4A, TUBB4B, TUBB5, TUBB6, TUBB8 by using the FASTA sequences downloaded from the UniProt KnowledgeBased Database).

- βT-4b + GDP;
- βT-5 + GDP;
- βT-6 + GDP;
- βT-8 + GDP;

The αT and βT-1 are not considered because the former was discovered only toward the end of the work that the wrong structure was taken (it is actually βT-8, thus **any details regarding αT must be ignored in the current work!!**).
The βT-1 has been forgotten instead.

2.1.2 Building pharmacophore models based on available cristallography to build γT+GTP

A γT-GTP complex crystallography is not available online; thus, by relying on how the GDP is spatially and chemically configured into the active site of γT in crystallography (PDB: 3CB2), two pharmacophores complex[3]-based models are built via *Pharmacophore Editor* MOE's tool[4].
Several strict features (Radius=1Å) at guanosine nucleobase are taken into account, as it is common in both GDP and GTP molecules, while one less stringent feature (Radius=3Å) is considered at 145THR acceptor residue (Fig.2.2B). SER, THR and, to a lesser extent, TYR are known to be the aminoacids mainly involved in the breaking of the bond between 2^{th} and 3^{th} phosphate group in the phosphorylation processes [180].



[3] γT-1 and γT-2

[4] Scheme: Unified

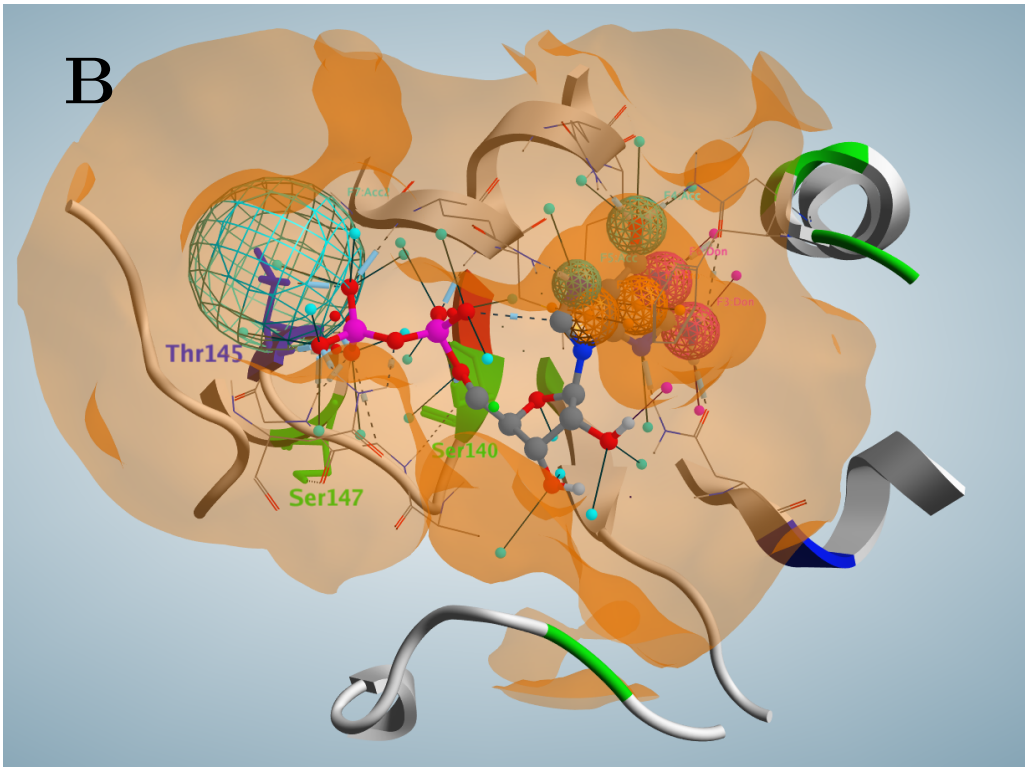


Figure 2.2 – *(A) Homology model of the GTP-binding pocket of Tub4 (yeast γT) based on the crystal structure of the human γT (PDB ID: 1Z5V). Red sphere: Mg^{+2} ion; yellow dashed lines: probable hydrogen bonds [L.Gombos, 2013]. (B) dashed spheres: pharmacophore features; in fuchsia/red small spheres: the phosphate groups; orange: molecular surface of γT.*
See Fig. 1.21A for features details. Realized with MOE

Moreover, Gombos et al. reported that by mutating only the THR localized in front of the GTP/GDP in γT of yeast, which shares a high number of residues with human γT (Fig.2.2A), the K_D is significantly affected: the K_{D_GTP} increases 3.6-fold (45 nM in case of wild-type and 164 nM in case of T146A); while the K_{D_GDP} increase 17-folds (206 nM in case of wild-type and 3.6x10^3 nM in case of T146A). These results strongly suggest the importance of this residue in the GTPase activity of γT [116]. Interestingly, there is Mg^{2+} ion between ASP70, THR146 and phosphate groups of nucleotide, which is not present in crystallography used in the following work [116].

2.1.3 Docking of GTP and Molecular Dynamics

The codes to prepare the data, automatize the simulations, retrieve the results and post-process them are reported in the appendix B, while files .mdp, which set up parameters for each simulation step are reported in appendix C.

Docking

Several Docking of GTP with the γT models are prepared on MOE with the following settings:

- Placement method: Triangle Matcher; Score: London dG; Timeout: 100000 s; number of returned poses: 100000; number of passed placement poses: 5000.
- Refinement method: Induced fit + free sidechain; Score (rescoring): GBVI/WSA dG; number of passed refinement poses: 10.
- pharmacophore models added for both γT-1 and γT-2.

Molecular Dynamics

The simulations are executed via GROMACS software (version 2021.4) mainly on the Narval cluster.

For each model, a generic simulation is set up with the following system settings: (1) non-bonded interactions and Particle-Mesh Ewald (PME) are calculated through GPUs since these types of computations require high computing power and rely on matrix calculations[5], for which GPUs are optimized.
As for the bonded interactions, these calculations are done via CPUs instead since they require lower computing power.

After changing the parameters iteratively, the best SLURM parameters to run individual MD simulations are 4 NVidia A100 GPUs, 12 CPUs per task and 4 tasks (total 48 cores requested, for further technical information see on the Narval cluster documentation webpage).

CHARMM36 is used as forcefield in GROMACS instead of the default forcefield already available (*charmm36-jul2021.ff* available on the CHARMM website).

Each simulation is divided into 4 steps:

1. Energy Minimization (first order only);
2. Canonical ensemble (NVT) with T=310K and no velocities assignment from Maxwell-Boltzmann distribution[6], t= 100 ps and position restraints for protein and ligand if any;
3. Isothermal–isobaric ensemble (NPT) with p=1bar, Berendsen pressure coupling and isotropic pressure, t= 100 ps and position restraints for protein and ligand if any;
4. MD with t= 100 ns and no position restrains.

[5]Indeed, the standard equation 1.2 relies mainly on the 3D coordinates of atoms

[6]The velocity for each atom is assigned starting from T=0K minimizing unnatural behaviours of the protein models

The solvent is explicit and the CHARMM36 TIP3P water model is used (available on the charmm36 forcefield directory).

Electrostatic interactions are calculated via PME which calculates the short-range electrostatic interactions in real space, while the long-range in Fourier space, as discussed in the (REF CHAPTER WHERE ELECTROSTATICS THEORY IS DISCUSSED).
Van der Walls potentials are switched from a radius of 1 nm and then switched off after a radius of 1.2 nm.

Post-processing of target proteins

RMSD, SASA and clustering are executed with internal GROMACS tools.
2 types of RMSD are calculated for each model: the first is calculated on the backbone, and the first frame as a reference to figure out the time (ns) after which equilibrium of the system is reached; while the second is also calculated on the backbone but on all frames as a reference, thus obtaining a RMSD matrix, which is necessary for the clustering step.
In the end, based on the start equilibrium time, clustering is done considering *Gromos* as clustering method and a cutoff of 0.15 nm; then, for each cluster found, the structure with the lowest RMSD is taken as a centroid.

The centroid of the most populated cluster is considered; if there are several clusters with a similar population, all centroids are considered and further investigation is made.

2.1.4 Identification of candidate binding site

After preparing the models and considering the centroid of the most populated cluster, finding candidate binding sites is trivial since the virtual screening of multi-billion compound databases against any and/or random pockets of the centroid is unsustainable computationally speaking. Even with more than 10 possible binding sites or docking over 1 million ligands, the virtual screening of such databases is still particularly computationally intensive.

There are two common ways of identifying binding sites in a protein [160]:

- blind docking: it is avoided in the current work because it is highly computationally expensive since it is just docking the ligands over the entire POI trying to find a favourable ligand binding site;
- pocket prediction algorithms: each algorithm find candidate binding site in different ways depending on the used programs[7], but the most common is based on the spatial composition of amino acids or using the chemical probe

This section aims to find one or a few candidate binding sites on γT-1, which is much unique as possible among all tubulins.

The strategy to reduce the number of candidate binding sites proposed in this work is evaluating electrostatic maps qualitatively, considering the Propensity Ligand Binding (PLB) score and the number of dummy points and calculating the residue similarity around the same pocket of all analyzed tubulins.

Exploring the centroid's surface via *Site Finder*

Site Finder MOE's tool is used to find all possible binding sites of the centroid of γT-1, which is the focus of this work, with the non-default options:

- Probe Radius 1: 1.4
- Probe Radius 2: 1.8
- Isolated Donor/Acceptor: 0.9
- Connection Distance: 1.5
- Minimum site size radius: 2
- Radius:1.5

Only candidate binding sites with PLB score > 1 are considered. Consequently, dummy points around a candidate binding site are created[8]. Dummy points indicate where possible atoms of the ligand may locate in the pocket; moreover, the function creates 2 types of dummy points: (1) the red sphere is associated with a hydrophilic dummy, thus N or O atoms; (2) white sphere is associated with a hydrophobic dummy instead, likelihood C atom.

Analysis of electrostatic maps

For each candidate binding site of each tubulin previously found, electrostatic maps are calculated through APBS PYMOL's plugin (METTERE CITAZIONE). After

[7]Example of available software that detect binding pockets: AlphaSpace, FTMap, MDpocket, Fpocket, SiteMap, MOE SiteFinder

[8]Option available in the bottom window of *Site Finder*.

that, each electrostatic map of the candidate binding site of γT-1 is compared to the similar pocket of other tubulins, in order to compare qualitatively.

Similarity of candidate binding sites

For each candidate binding site previously found, the similarity of residues around such sites is evaluated among all tubulin models via *Sequences Identities And Similarities* webserver tool available on http://imed.med.ucm.es/Tools/sias.html. The tool evaluates the similarity based on the physicochemical properties of all aminoacids, each assigned relative score compared to other aminoacids [181].

It is important to note that the tool is supposed to evaluate the similarity of residues of entire structures. Thus the *Similarity Result* is considered to avoid the addition of penalties based on gaps and evaluate the similarity of given smaller aminoacid sequences. All options in *Similarity amino acid grouping* section are selected except for *small* subsection.

2.2 Virtual screening

2.2.1 The ZINC databases and file preparation

Since there are no known ligands with solid evidence of binding on γT in scientific literature as discussed previously (See Sec. 1.5), exploring multi-billion compounds databases seems the only possible option, although it is not very feasible. Three ZINC open-access databases are downloaded and used in the current work:

- the biogenic type database includes 156'695 molecules of biological origin;
- the entire ZINC15 database consists of 885'492'702 ligands instead, including the biogenic database;
- the entire ZINC20 database consists of 1'006'650'596 ligands and it is updated to early March 2021. It is already processed (stereochemistry and tautomer exploration and protonation at pH 7.4) by A.Cherkasov and F.Gentile [176, 182].

The first two databases are subjected to the following pre-processes through MOE's SD tools which can be used in HPC clusters since they could contain duplicates and/or raw information. All pre-processing codes are reported in the appendix D. The preprocessing is indexed into the following steps:

1. preparation step: removal of duplicates based on SMILE string and enantiomers[9];
2. 1°washing: disconnect common metal salts (convert to ionic notation), remove minor components (salts, miscellaneous adducts) and mark atom-centred chirality;
3. 1°filtering: eliminate molecules containing non-organic elements such as transition metals and atoms with more than 4 bonds and allow compounds having only the elements C, H, N, O, S, P, F, Cl, Br and I;
4. 2°washing: by imposing pH=7, numerate tautomers and protomers (1000 tautomers/molecule);
5. 2°filtering: removing compounds with reactive groups (metals, phospho-, N/O/S-N/O/S single bonds, thiols, acyl halides, Michael Acceptors, azides, esters, etc.) and taking only the tautomer most populated (C>80%) generated in the previous step;
6. sorting and report unique compounds.

[9]The enantiomers are re-generated in the next step, in this way, it is possible to have the cleanest possible database

2.2.2 Conformation Sampling

The downloaded databases from the ZINC server contain only the SMILE string, with no information about its 3D structure.

As extensively discussed above (section 1.7.3), conformational sampling is required before running docking.

Docking MOE's tool has an algorithm in which the generation of some conformations is executed before docking itself (point 1 on sec. 2.2.5).
However, it is a systematic algorithm, while stochastic followed by a rapid EM is desired to improve the high variability of conformations.

Two internal conformational sampling MOE tools are available and both are used in the current work based on the explored database [167]:

- *Conformational Import.* This function is used in the conventional approach (See the scheme on page 53), and it is optimized to process a large number of ligands. Briefly, the algorithm follows several steps[10]:
  1. *Washing & Filtering if any.* Here, it is possible using other 2D QSAR descriptors for further filtering. But, since there is no knowledge about what the candidate ligand may be, this option is set off.
  2. *Fragmentation.* Instead of sampling the entire ligand, as typical of other conformational sampling algorithms, the ligand is split into overlapping fragments.
  3. *Fragment conformations generation.* Before running a conformational sampling of individual fragments, a lookup for each fragment is made on a fragment database where the previously sampled fragments[11] are saved temporally. The lookup look for identical or at least its same carbon skeleton. In the case the search fails, stochastic conformational search is then performed, and the result is stored in the fragment database, which could be used at other time.
  4. *Conformational Assembly.* Rigid body superposition is used to position the generated fragment conformations; if there are bad VdW contacts, then the overall conformation will be rejected.
  5. *Strain Energy Estimation.* Intramolecular energies, also called strain energy, of overall conformation is the sum of strain energy of subset fragments that overlap by 3 or fewer atoms. If the overall strain energy is higher than the default value of 4 kcal/mol, the conformation will be rejected.

  The options for *Conformational Import* are left as default, except for:
  - Maximum number of conformations for ligand: 100;
  - Refine Conformations (i.e.: EM) yes;
  - Write Filtered Compounds: no;
  - Database packed.

[10] See MOE documentation for further information, reported in the section *Conformation Import*

[11] Each fragment is sampled independently from others

- *Conformational Search.* The following type of conformational sampling is used in the second approach, in which the DeepDock package is involved (See the scheme on page 53).
  Briefly, the algorithm follows several steps depending on the chosen method: *Stochastic Search* method[12] is selected because it is a hybrid Monte Carlo algorithm allowing the detection of most of the local minima of ligands, including ring conformations and invertible tetrahedral centers due precisely to random search, significantly reducing the biases typically present in *Systematic Search* method. The workflow is described as follows:

  1. *Initialize.* Set the conformer list C to empty and identify all rotation bonds, including rings and invertible stereo centers;
  2. *Generate.* Single conformation is generated by randomly rotating all bonds (including ring bonds) and randomly inverting invertible tetrahedral centers;
  3. *Minimize.* EM of the conformation until the RMS gradient falls below the MSD Gradient test parameter;
  4. *Save.* The conformation is discarded if the current conformation has a strain energy greater than the Strain Cutoff parameter or if the RMSD between the current conformation and the one present in the conformer list C is less than the RMSD Limit parameter;
  5. *Termination.* Repeat the algorithm from Step 2 or terminate the algorithm if the total number of iterations exceeds the Iteration Limit parameter or if no novel conformation was observed in the Rejection Limit parameter consecutive iterations.
  6. *Descriptors.* Sort the conformation list C by strain energy and calculate energetic and shape-based descriptors for each conformation.

  The options for *Conformational Search* are left as default, except for:

  – Allow amide rotational bond
  – Conformation Limit: 50 (these conformations will represent the poses which will be given as input in the MOE docking function)
  – Method: Stochastic

[12] See MOE documentation for further information, reported in the section *Generating and Analyzing Conformations*

2.2.3 Pharmacophore Filtering

Pharmacophore Filtering is an intermediate step of the conventional approach (See the scheme on page 53).
This step is not done in the "DeepDock" approach to avoid generating biases in the datasets because Pharmacophore Filtering overlooks the less favourable conformations/ligands, which represent instead useful information during the training of the DNN models.

For each definitive candidate binding site (Sec.2.1.4) of the centroid of γT-1 of the most populated cluster, pharmacophore protein-based models are built through *Pharmacophore Editor* MOE's tool, helped by occupancy of residues, dummy points distribution and PLB scores created in 2.1.4.
After that, the conformations generated via *Conformational Import* are filtered via *Pharmacophore Search* MOE's tool with the default option except for:

- Results: Conformations
- Hits: First per Molecule

2.2.4 Consensus Docking - DockBox

This is the final step of the conventional approach (See the scheme on page 53).

3 docking software are used: MOE, DOCK6 and Vina, since the latter two showed to be the best minimal combination (See on sec. 1.7.2) [162].
All three are considered both in the generation of poses and scoring function; their settings are saved in the *config.ini*, a configuration file (See appendix E for further details).

The forked[13] version of DockBox is available online and it contains several fixes in *vina.py* python script. Additionally, a script to prepare the slurm job files in a more efficient and automated way and generic-use is written and is available online. In particular, the script *01_prepareCompoundsProteinsSites.sh* prepares the SLURM job files by proving one or more proteins (.pdb format), the configuration file (.ini format) and the ligands (.mol2 format).
If the file with the center's coordinates of the binding site is not provided, blind docking is performed on the best 50 potential binding sites found through the *siteFinder* MOE's tool. However, since the candidate binding sites are previously found (Sec. 2.1.4), the centers are calculated as the average of the 3D position of the dummies around a specific binding site.

Note that SBCD is automatically selected, with a RMSD cutoff = 2 Å, during the extraction of the best poses to have higher quality in the results (See sec. 1.7.2).

[13] I.e. Copied the original code to possibly modify without compromising the original one. Here the original repo.

2.2.5 Docking

This step is done in the DeepDock approach (See the scheme on page 53) to prepare the datasets required to train DNN models.

For each potential binding site (Sec.2.1.4) of the centroid of γT-1 of the most populated cluster, the ligands contained in the datasets are first docked.

Docking scripts required to be executed on HPC clusters are locally prepared through the option *Batch* inside the *Docking* MOE's tool.

Briefly, the docking algorithm follows the following steps [167]:

1. *Conformational Analysis.* In this work, this step is skipped since the conformational database that is the input for the next step has already been calculated. Otherwise, if not supplied, conformations can be generated by Conformation Import via a fragment-based approach or from a single 3D conformer by applying a collection of preferred torsion angles to the rotatable bonds.
2. *Pharmacophore Pre-filtering and Pre-placement.* In this work, this step is also skipped to avoid generating biases. Generally, the poses can be constrained to fit a previously created or ad hoc pharmacophore query.
3. *Placement.* Several placements called poses are generated from the pool of ligand conformations with a placement method selected from several available[14].
4. *Initial Scoring.* The previously generated poses are scored with a scoring method selected from several available[15]. Typically, scoring functions emphasize favourable hydrophobic, ionic and hydrogen bond contacts.
5. *Refinement.* The poses can be submitted to energy minimization.
6. *Pharmacophore Constraints.* The user may provide a pharmacophore to constrain the final poses. In this work, this step is skipped.
7. *Final Scoring.* The final poses can be rescored using one of the available scoring schemes.
8. *Final Filtering.* The top-scoring poses are subject to optional duplicate removal.

. The docking is done with the following settings:

- Placement Method: Proxy Triangle;
- Scoring (after placement): London dG;
- Maximum number of poses after scoring: 20;
- Refinement (EM): Induced Fit with Free Side Chains option activated[16] and a RMSD=6Å[17];
- Rescoring (after EM: GBVI/WSA dG;

[14]See MOE documentation for further information, reported in the section *Docking - Placement Methodology*

[15]See MOE documentation for further information, reported in the section *Docking - Computational Background*

[16]The atoms of side chains of the aminoacids around the binding site are free to move without constraints, except for the atoms of the back-bone

[17]I.e. the cutoff distance used to decide which receptor atoms to include in the energy minimization

- Maximum number of Poses after rescoring: 1.

The *Proxy Triangle* is chosen as the placement method because it is optimized to manage many conformations and/or ligands. Indeed, the technique pre-superposes all conformers before being placed into the binding site, and the scoring considers the atom representatives rather than all of the ligand atoms, thus saving computational time [167].

Regarding the scoring and rescoring, both calculate the free energy ΔG of the binding ligand from given poses but in different ways[167].:

- *London dG* (Empirical-based scoring function)

$$\Delta G = c + E_{flex} + \sum_{h-bonds} c_{HB} f_{HB} + \sum_{m-lig} c_M f_M + \sum_{atomsi} \Delta D_i \qquad (2.1)$$

  - c: average gain/loss of rotational and translational entropy;
  - E_{flex}: energy due to the loss of flexibility of the ligand (calculated from ligand topology only);
  - f_{HB}: geometric imperfections of hydrogen bonds, value ∈ [0,1];
  - c_{HB}: energy of an ideal hydrogen bond;
  - f_M:1 geometric imperfections of metal ligations, value ∈ [0,1];
  - c_M:energy of an ideal metal ligation;
  - D_i: desolvation energy of atom i and its equation is the following:

$$\Delta D_i = c_i R_i^3 \left\{ \iiint_{u \notin A \cup B} |u|^{-6} du - \iiint_{u \notin B} |u|^{-6} du \right\}$$

  - A and B are the protein and/or ligand volumes with atom i belonging to volume B;
  - R_i: solvation radius of atom i (taken as the OPLS-AA van der Waals sigma parameter plus 0.5 Å);
  - c_i desolvation coefficient of atom i.

  The coefficients c, c_{HB}, c_i were fitted from approximately 400 X-ray crystal structures of protein-ligand complexes with available experimental pKi data. Atoms are categorized into about a dozen atom types for the assignment of the ci coefficients. The triple integrals are approximated using Generalized Born integral formulas.

- *GBVI/WSA dG Scoring* (Forcefield-based scoring function)

$$\Delta G = c + \alpha \left[\frac{2}{3} (\Delta E_{Coul} + \Delta E_{sol}) + \Delta E_{VdW} + \beta \Delta SA_{weighted} \right]$$

  - c: average gain/loss of rotational and translational entropy;
  - α, β: constants which were determined during training (along with c and are forcefield-dependent);
  - E_{Coul}: coulombic electrostatic term which is calculated using currently loaded charges, using a constant dielectric of $\epsilon = 1$;
  - E_{sol} solvation electrostatic term, which is calculated using the GB/VI solvation mode;
  - E_{VdW}: Van der Waals contribution to binding;
  - $SA_{weighted}$: surface area, weighted by exposure. This weighting scheme penalizes exposed surface area.

  This scoring algorithm has been trained using the MMFF94x and AMBER99 forcefield on the 99 protein-ligand complexes of the SIE training set. Moreover, since the binding usually takes place in the presence of water, the desolvation energies of the ligand and the protein are sometimes taken into account using implicit solvation methods such as GBSA or PBSA [167].

2.2.6 Training the DNN models - DeepDock

Morgan fingerprints, known as circular fingerprints, are already calculated (Morgan Algorithm) for each ligand in the ZINC20 database as binary with radius 2 and size of 1'024 bits and available online [176, 183].
The files are also ready to be used and read by DeepDock: each line contains ZINC ID and the indexes of 1[18].

As discussed in sec.1.7.5, the number of iterations and size of datasets heavily affect the performances of DeepDock and its outcomes: thus, 1M as size for datasets and a maximum of 11 iterations are chosen.
Regarding the recall and the number of hyperparameters (i.e. number of hidden layers and neurons, dropout frequencies, over-sampling of minority class and class weights), it has been set to 0.9 and 24, respectively.
These parameters should be sufficient to reduce a database of 1B ÷ 1.5B compounds [172, 176].

The original DeepDock package is forked and modified to be adapted in the current work. The forked DeepDock repo is available online on the personal Git-Hub.

New bash scripts to prepare the datasets in an automatic way are available online on the personal Git-Hub. Further details about the automatic preparation of datasets are extensively discussed in appendix F.

Note that it is assumed that the user has installed MOE version 2022.2[19] on the HPC cluster with *Slurm Workload Manager* as job scheduler and GPUs with at least 14GB for job.
Moreover, with 24 as the number of hyperparameters and 100 files that were split from the original database, DeepDock generates for each iteration 24 DNN models to be independently trained (thus 24 jobs) and 100 inference jobs (i.e. prediction phase). Thus 2'400 jobs-per-iteration in total will be sent to the cluster.
Despite the high computational demands, each job takes from 15 minutes (first iterations) to 1 hour (last iterations).

[18] The indexes are the position in which 1 is present in a binary string.
Ex. for a binary string 100101, the indexes are 0,3 and 5, considering the first number as index 0.

[19] Version 2020 should still work well.

Chapter 3

Results

3.1 The preparation of models

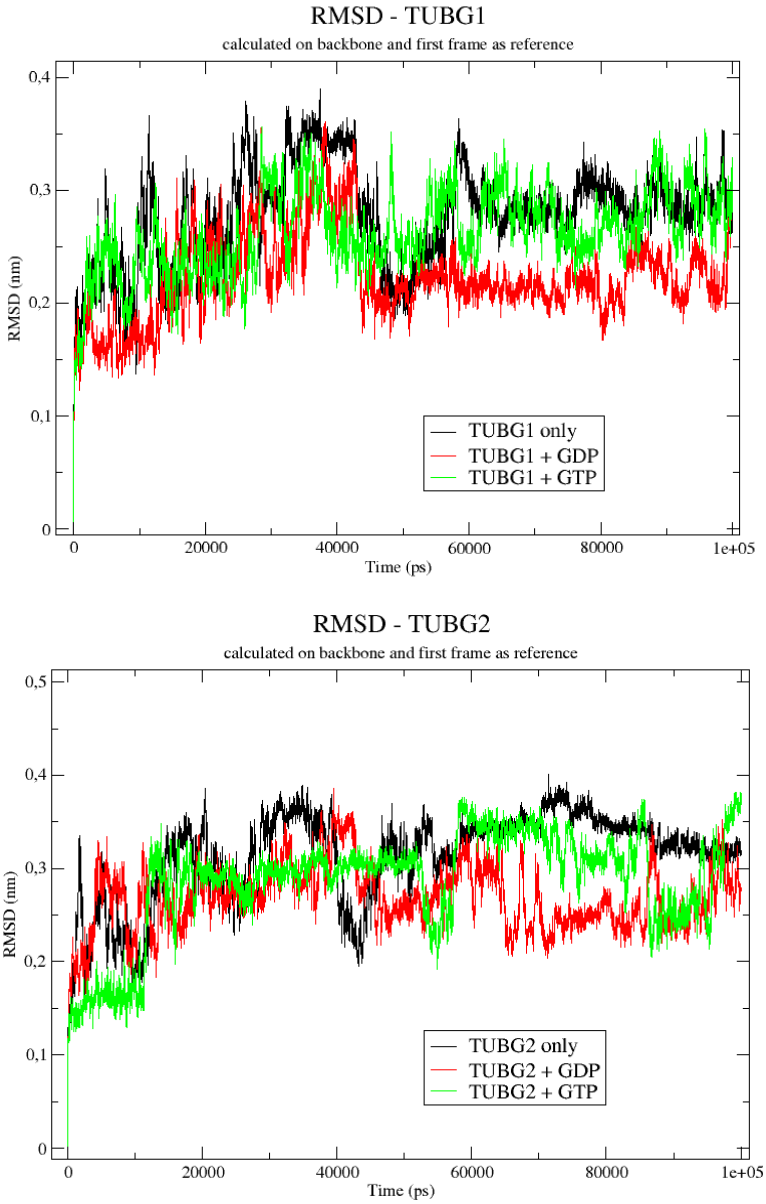


Figure 3.1 – *RMSD of γT1-2 after MD simulation*

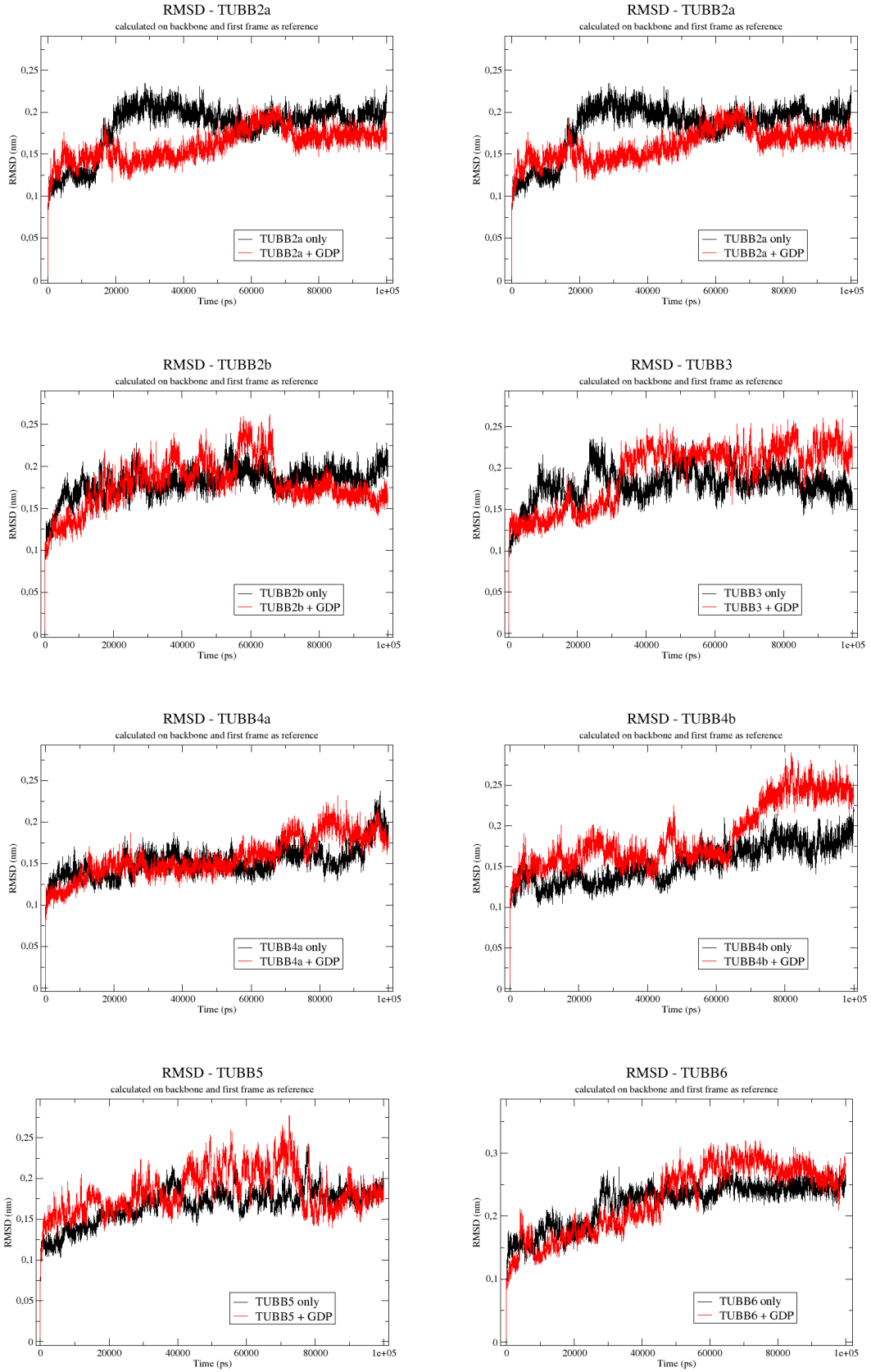


Figure 3.2 – *RMSD of βTs after MD simulation*

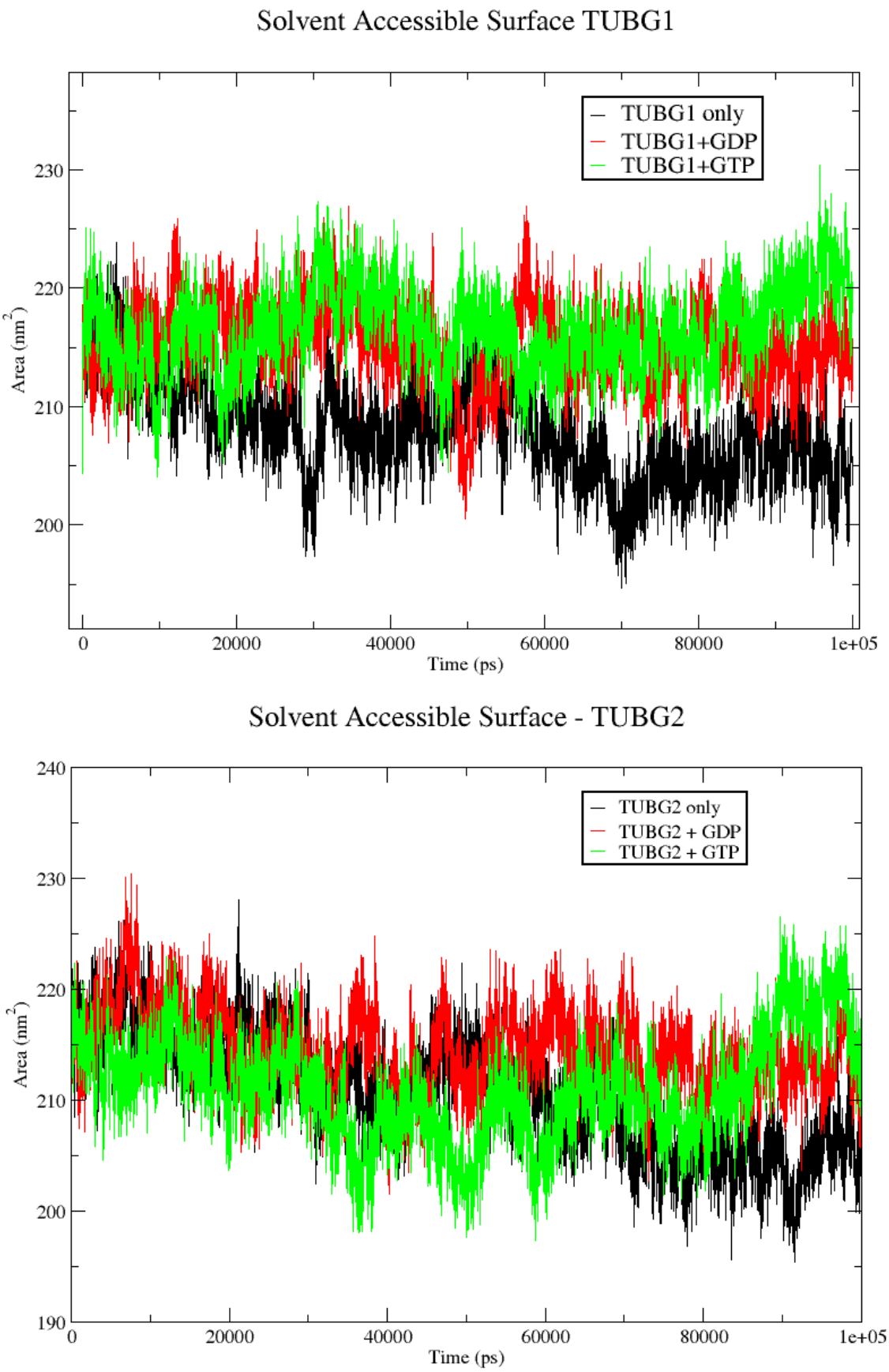


Figure 3.3 – *SASA calculation. Black = protein only; red = protein+GDP; green = protein+GTP*

All the models after MD of 100 ns showed to be very structurally stable and their RMSD does not differ significantly even in the presence of GDP or GTP: γT1-2 do not exceed 0.4 nm (Fig. 3.1), while all other tubulins do not exceed 0.25 nm (Fig. 3.2).
The little difference between γTs and others is caused by the added C-terminus in γT1-2, which causes excessive movement during the MD.

Further confirmation of low influence of nucleotides on the γT structures is provided by SASA calculation (Fig. 3.3).
Nevertheless, a slight decrease of ~ 10 Å^2 is observed in the case of protein only: this may happen because the models are from a crystallography in which GDP is bound, thus its absence causes small rearrangement in the surface area during MD simulation.

For each model post MD, the time from which the system reaches equilibrium is considered to cluster the frames[1] after that time:

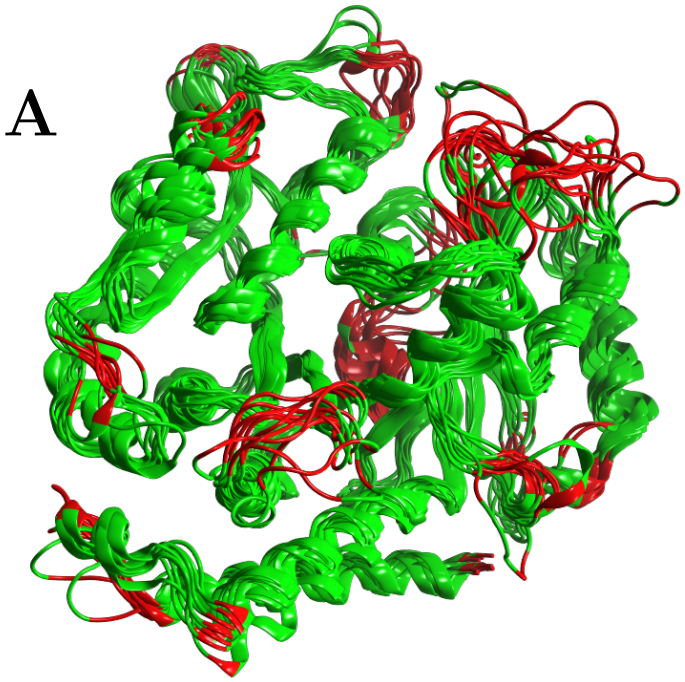
- TUBG1: 65 ns $\rightarrow$ 5 clusters (3162+234+101+3+1);
- TUBG2: 60 ns $\rightarrow$ 5 clusters (3762+181+44+12+1);
- TUBB2A: 30 ns $\rightarrow$ 3 clusters (6991+6+4);
- TUBB2B: 57 ns $\rightarrow$ 1 clusters (4301);
- TUBB3: 40 ns $\rightarrow$ 3 clusters (5974+24+3);
- TUBB4A: 30 ns $\rightarrow$ 3 clusters (6915+71+15);
- TUBB4B: 64 ns $\rightarrow$ 2 clusters (3597+4);
- TUBB5: 50 ns $\rightarrow$ 5 clusters (4959+23+12+5+2);
- TUBB6: 35 ns $\rightarrow$ 5 clusters (6478+15+6+1+1);
- TUBB8: 60 ns $\rightarrow$ 4 clusters (1640+1324+33+4);

The most populated clusters are taken into account for each tubulin to take the centroid[2] which represents the protein with the highest probability to be found in such phase state. An exception is made for TUBB8, of which there are two clusters almost equally populated; thus, two centroids are considered instead.

All extracted centroids have similar RMSD values and lower than 3 Å, confirming that γT is structurally almost identical to other tubulins studied here as reported from several studies (Fig. 3.4).
By calculating the RMSD between single residues, it is possible to find the most diverse spots to help identify potential active sites (Fig. 3.5).

[1] A MD simulation is not a continue-time simulation, but each frame represents a single state of space phase.

[2] The centroid is the frame belonging to a certain cluster with the lowest RMSD.



B

RMSD = 2.139 A

	1	2	3	4	5	6	7	8	9	10	11
1: TUBG1	0.00	2.24	2.77	2.84	2.85	2.47	2.74	2.68	2.55	2.79	2.76
2: TUBG2	2.24	0.00	2.71	2.70	2.79	2.49	2.57	2.69	2.40	2.65	2.60
3: TUBB2A	2.77	2.71	0.00	1.71	1.61	1.42	1.71	1.73	1.99	1.89	2.09
4: TUBB2B	2.84	2.70	1.71	0.00	1.58	1.68	1.78	2.12	1.92	1.94	2.12
5: TUBB3	2.85	2.79	1.61	1.58	0.00	1.60	1.47	1.91	1.85	1.77	2.01
6: TUBB4A	2.47	2.49	1.42	1.68	1.60	0.00	1.52	1.40	1.73	1.82	1.95
7: TUBB4B	2.74	2.57	1.71	1.78	1.47	1.52	0.00	1.87	1.68	1.79	1.90
8: TUBB5	2.68	2.69	1.73	2.12	1.91	1.40	1.87	0.00	1.98	2.06	2.09
9: TUBB6	2.55	2.40	1.99	1.92	1.85	1.73	1.68	1.98	0.00	1.91	1.97
10: TUBB8_1	2.79	2.65	1.89	1.94	1.77	1.82	1.79	2.06	1.91	0.00	1.00
11: TUBB8_2	2.76	2.60	2.09	2.12	2.01	1.95	1.90	2.09	1.97	1.00	0.00

4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0

Figure 3.4 – *(A) superposition, after sequence alignment, of centroids TUBG1, TUBG2, TUBB2A, TUBB2B, TUBB3, TUBB4A, TUBB4B, TUBB5, TUBB6, TUBB8_1 and TUBB8_2; green = $RMSD_{mean}$ < 2.139 Å, red = $RMSD_{mean}$ > 2.139 Å(B) RMSD between centroids. C-terminus is deleted because it highly affects the RMSD value and is highly motile.*

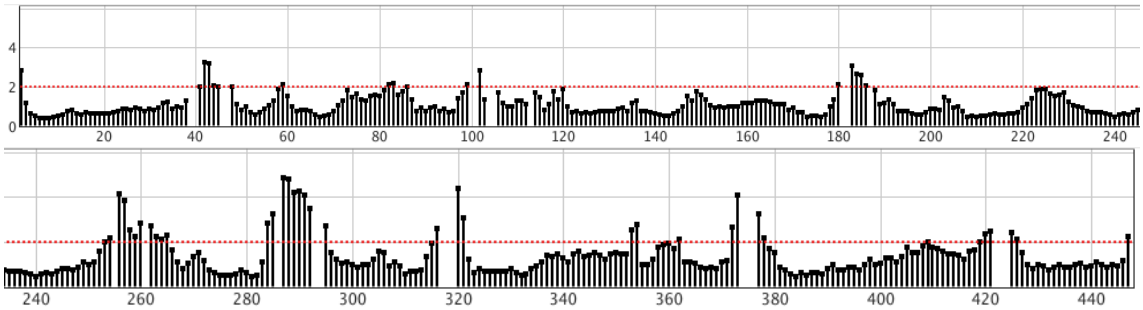


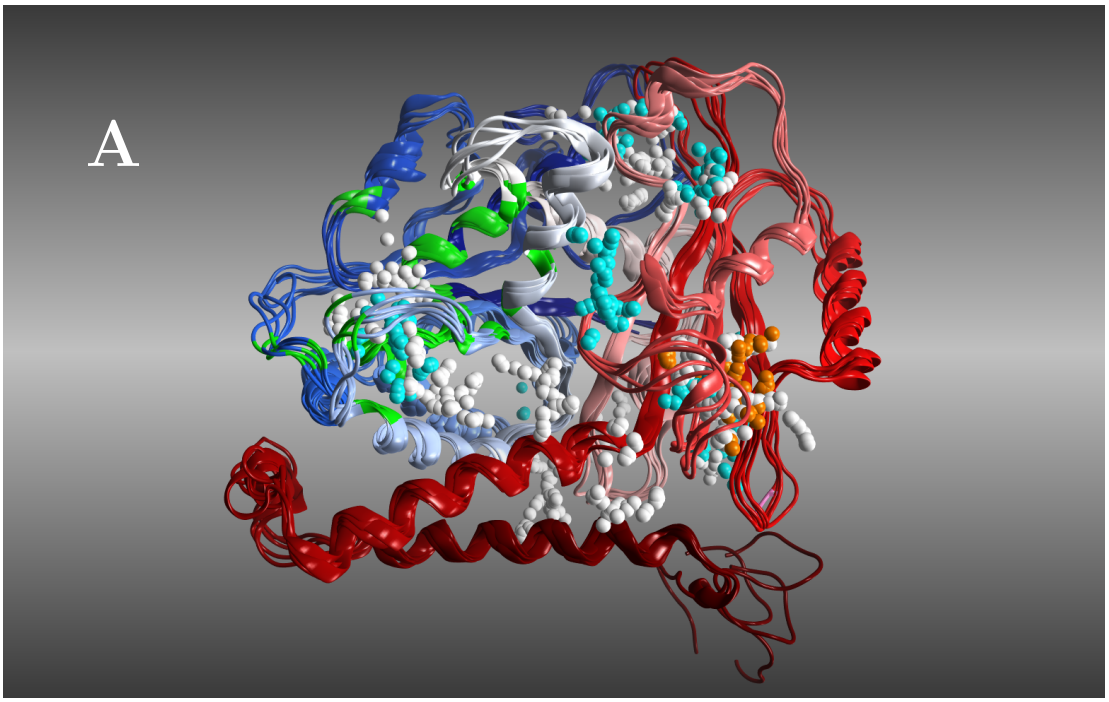
Figure 3.5 – *RMSD of single residues between all tubulins.*
Red dashed line: $RMSD_{mean}$ = *2.139 Å*

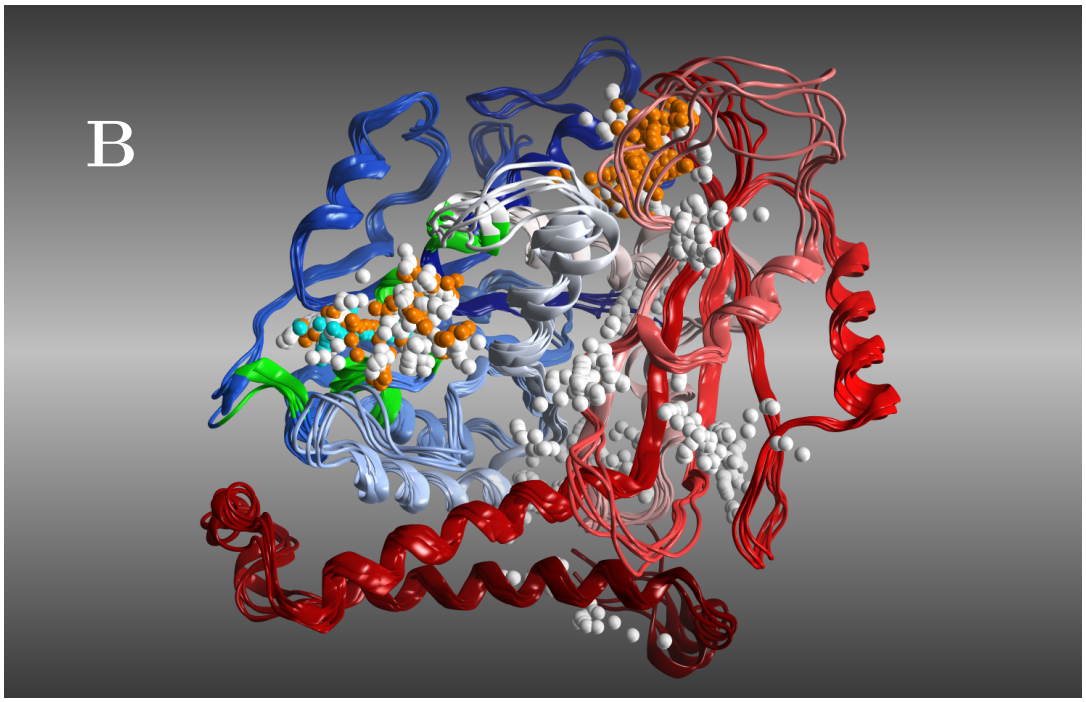
3.2 Finding the potential binding sites

Site Finder MOE's tool is used to find all possible binding sites by giving PLB scores to each found point for all tubulins. After that, dummies are generated in correspondence with these points with a PLB score > 1 (Fig 3.6).

Although the high amount of dummies points in the GTP binding site, the such pocket must be ignored for the reasons discussed in Sec. 1.5.

Along with the information regarding the high RMSD values of single residues between the tubulins (Fig. 3.5) and giving priority to those dummies that are more numerous and have higher scores in a certain candidate pocket of the γT1, eight candidate binding sites are found (Fig. 3.7-14).





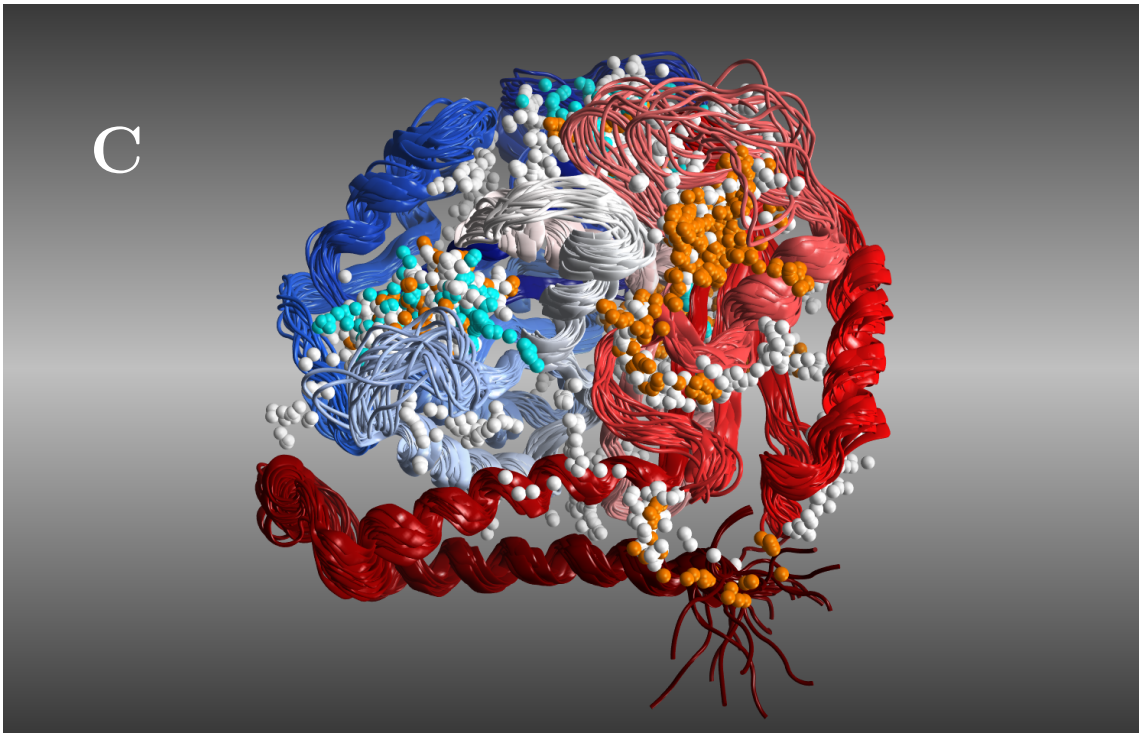


Figure 3.6 – *(A) TUBG1, (B) TUBG2, (C) TUBBs. white sphere = dummy with* $PLB > 1$*; cyan sphere = dummy with* $3 < PLB < 4$*; orange sphere = dummy with* $PLB > 4$*; green ribbon = GTP binding site; blue to red ribbon = N-terminus to C-terminus*

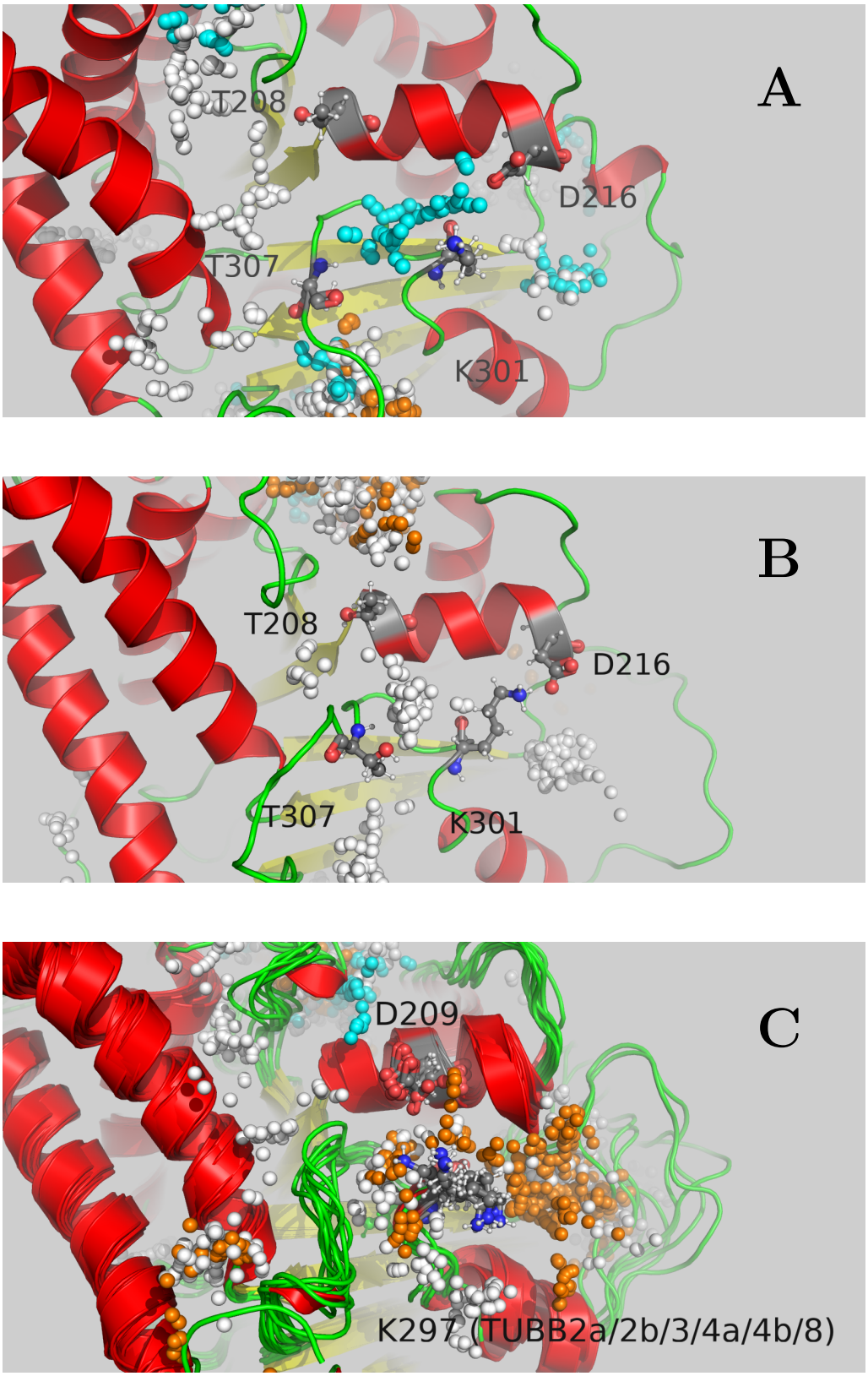


Figure 3.7 – *(1) potential binding site. (A) $\gamma T1$; (B) $\gamma T2$; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

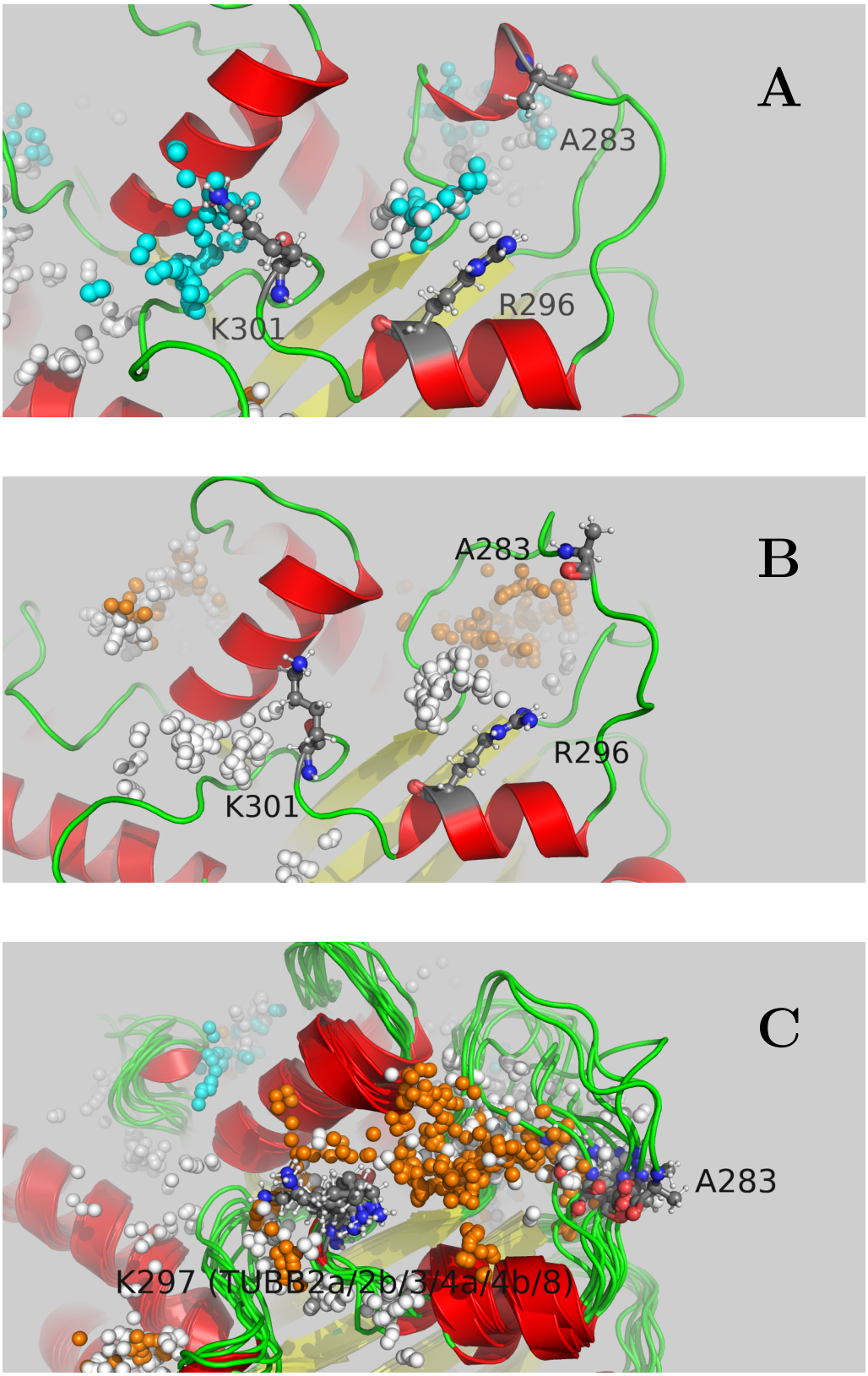


Figure 3.8 – *(2) potential binding site. (A) γT1; (B) γT2; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

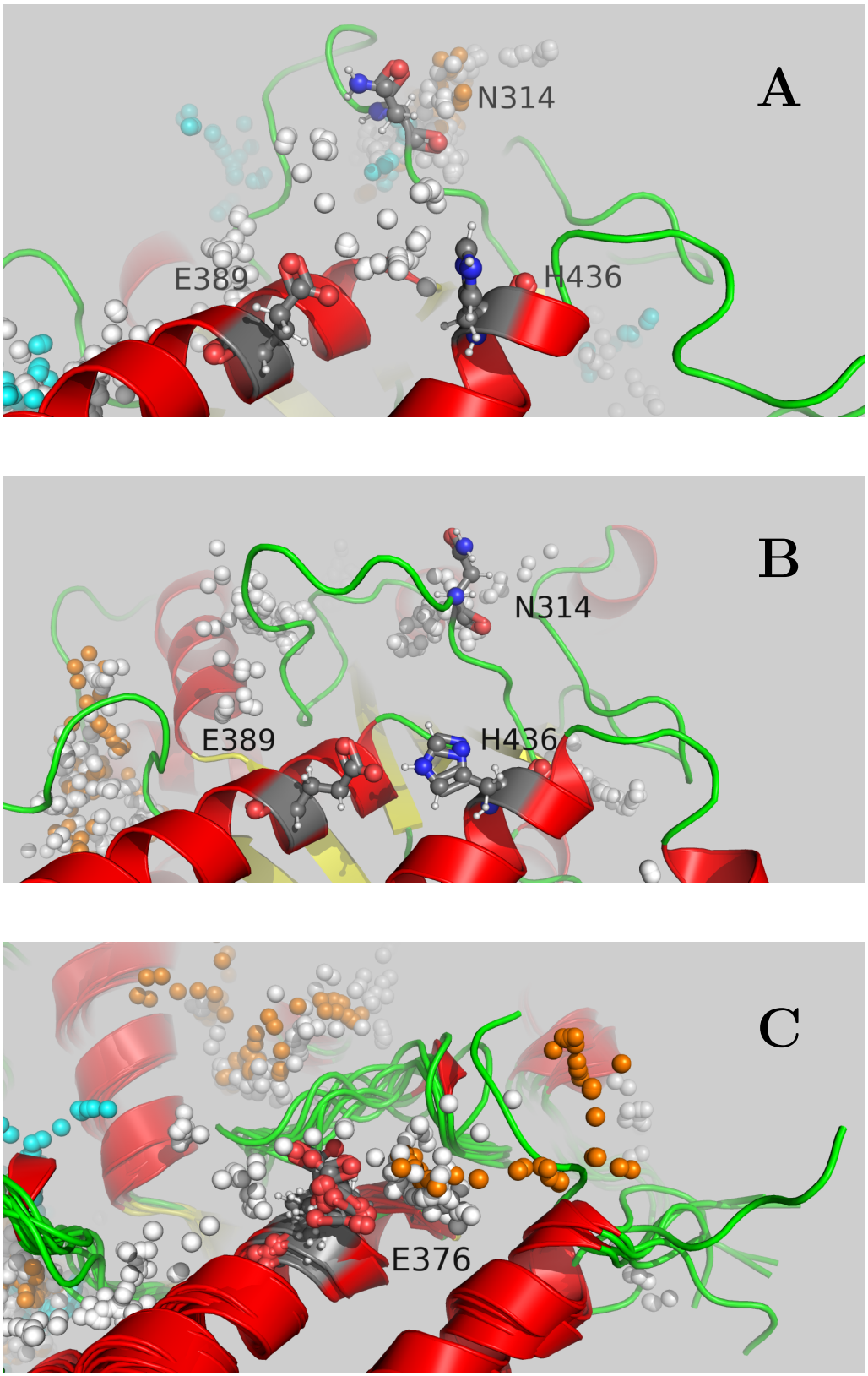


Figure 3.9 – *(3) potential binding site. (A) γT1; (B) γT2; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

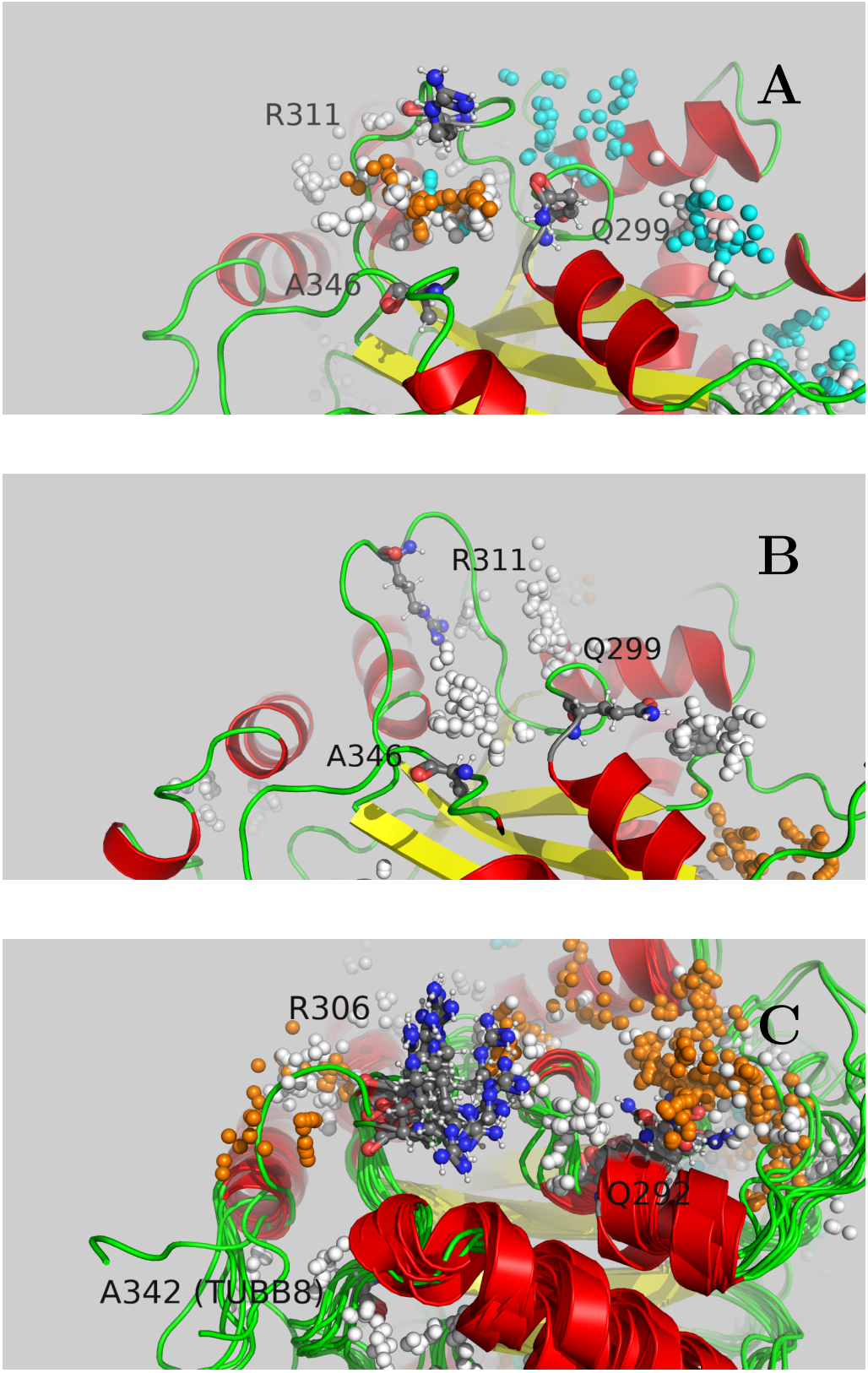


Figure 3.10 – *(4) potential binding site. (A) $\gamma T1$; (B) $\gamma T2$; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

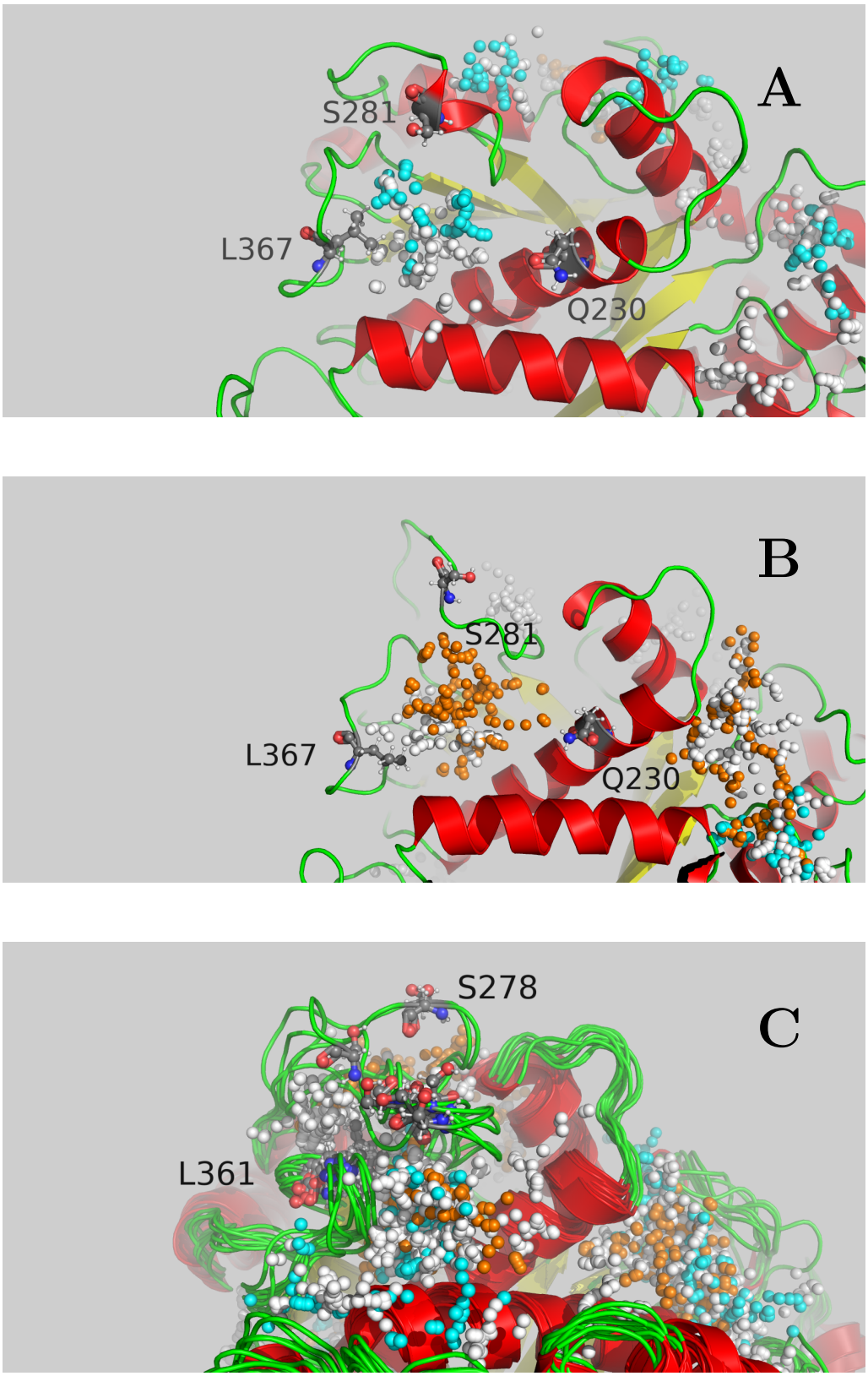


Figure 3.11 – *(5) potential binding site. (A) γT1; (B) γT2; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

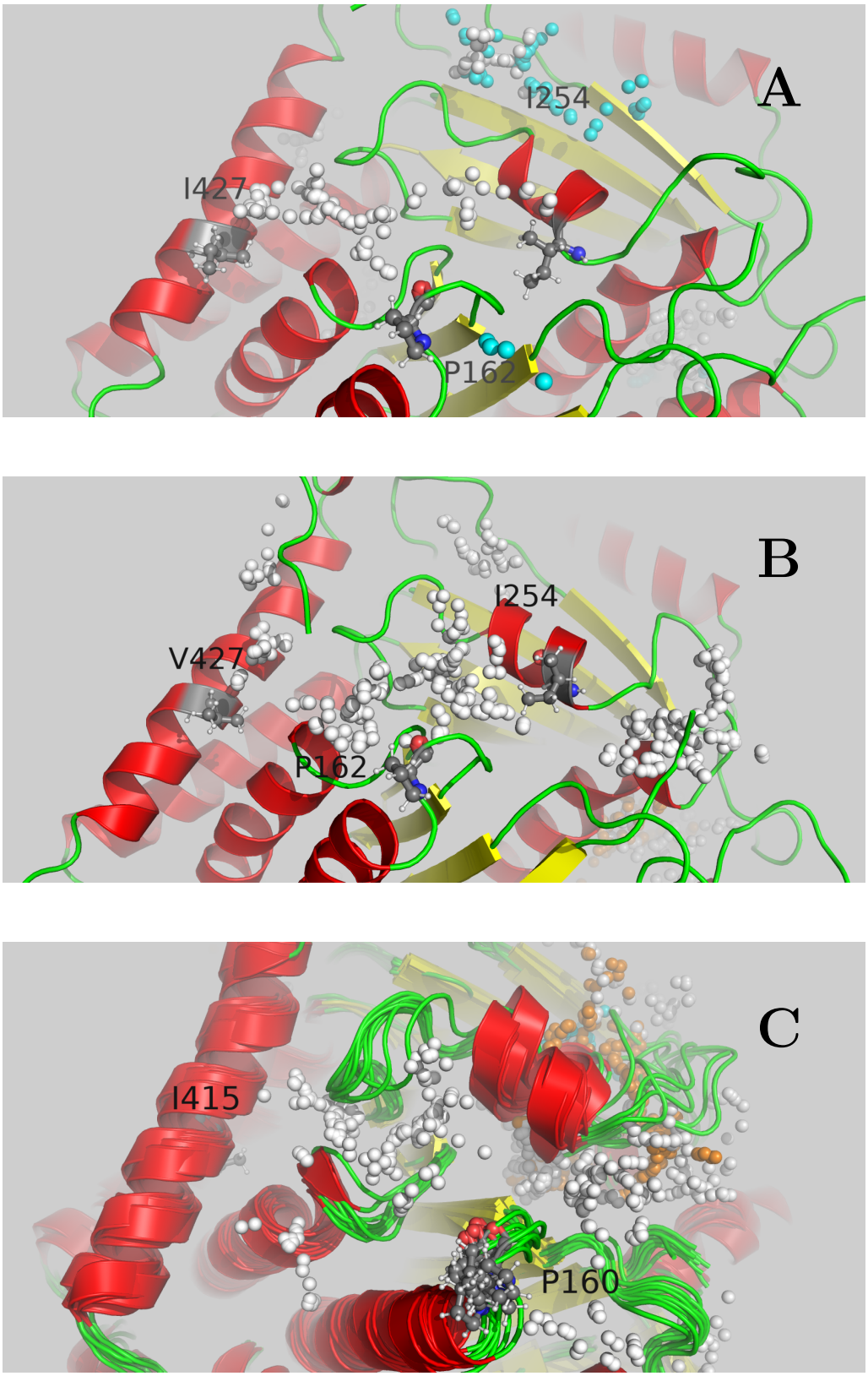


Figure 3.12 – *(6) potential binding site. (A) $\gamma T1$; (B) $\gamma T2$; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

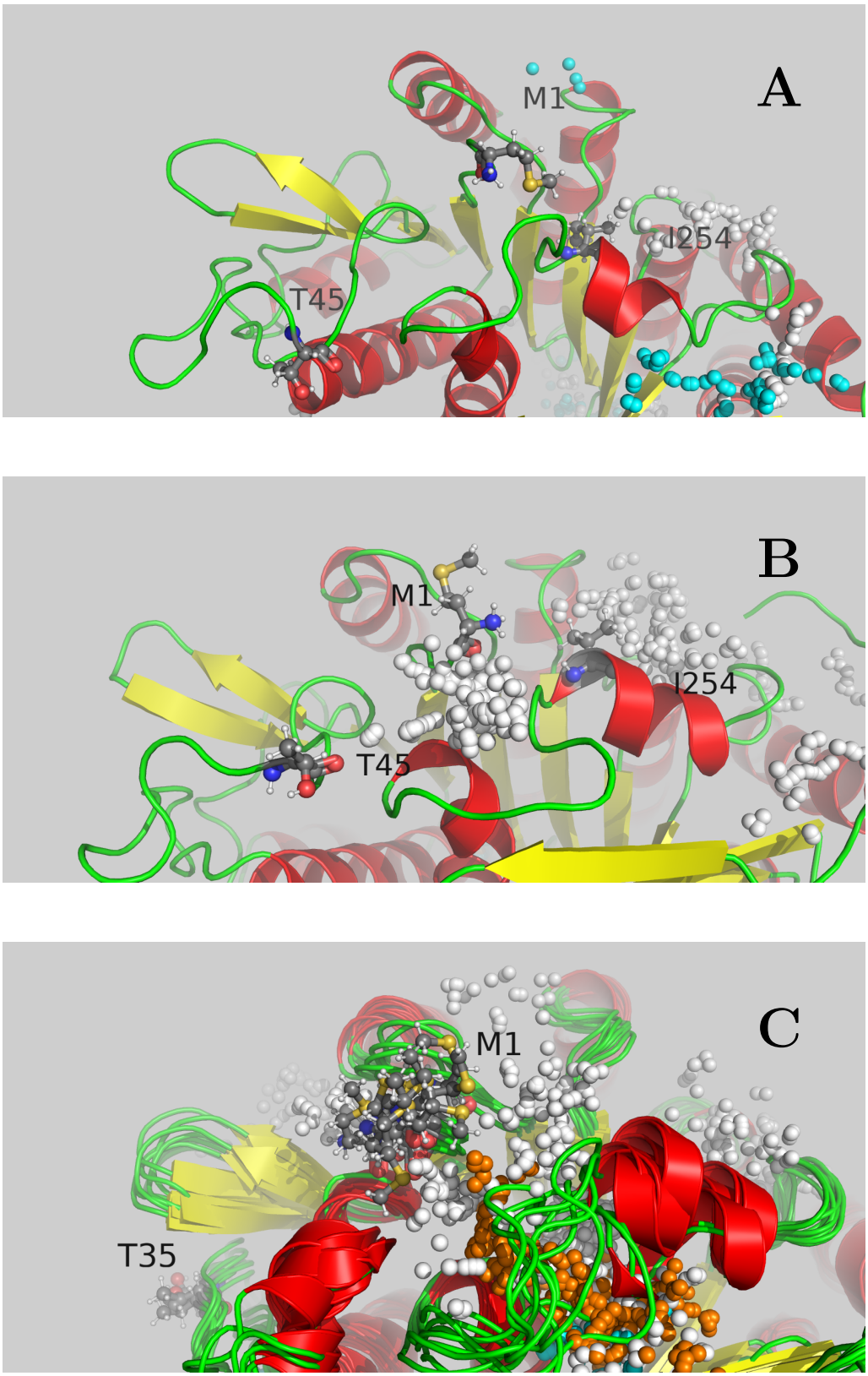


Figure 3.13 – *(7) potential binding site. (A) γT1; (B) γT2; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

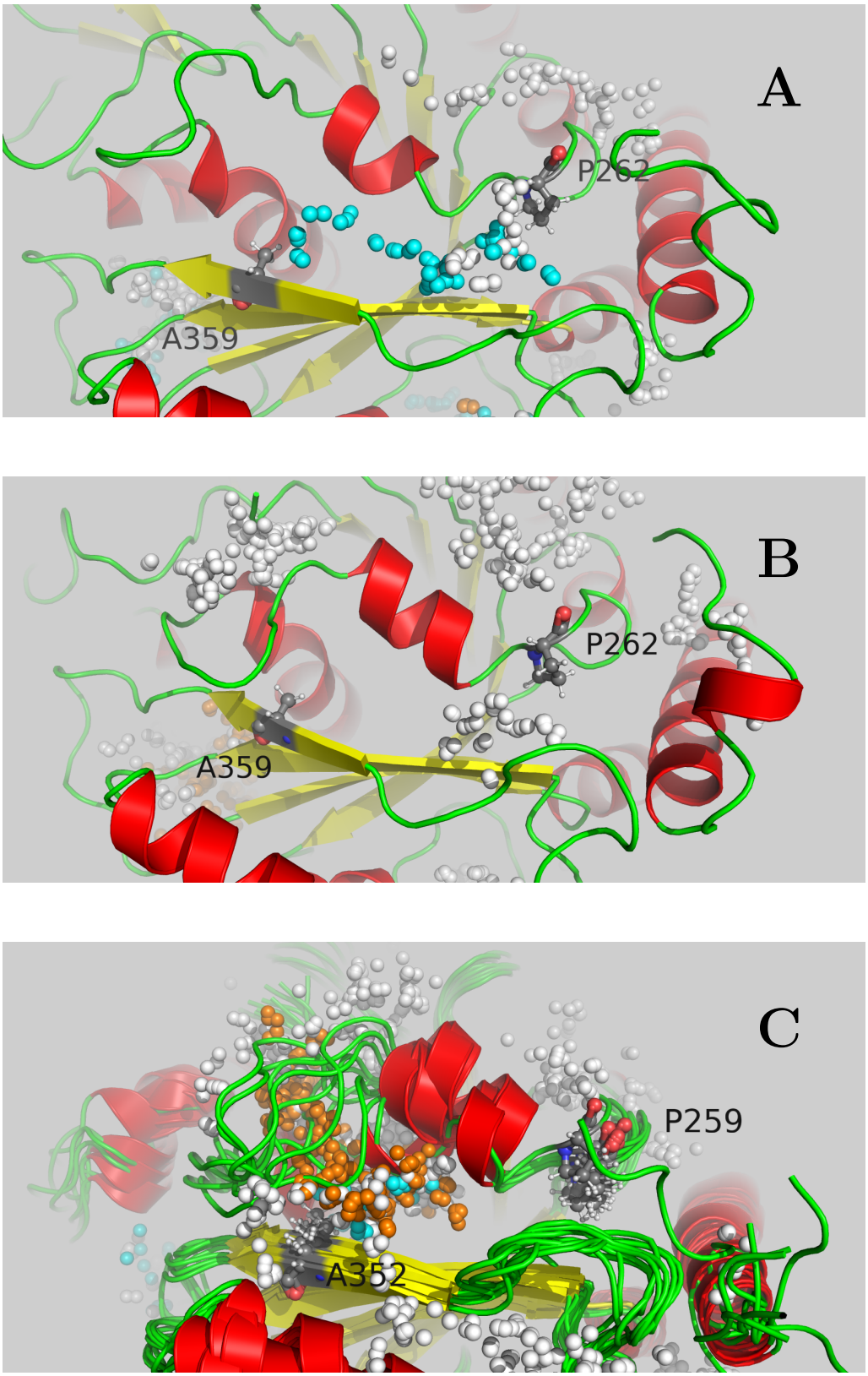


Figure 3.14 – *(8) potential binding site. (A) $\gamma T1$; (B) $\gamma T2$; (C) βTs (centroids all superposed). White sphere = dummy with $PLB > 1$; cyan sphere = dummy with $3 < PLB < 4$; orange sphere = dummy with $PLB > 4$.*

TUBG1	100%									
TUBG2	97.05%	100%								
TUBB2a	56.25%	53.12%	100%							
TUBB2b	56.25%	53.12%	100%	100%						
TUBB3	59.37%	56.25%	96.87%	96.87%	100%					
TUBB4a	59.37%	56.25%	96.87%	96.87%	100%	100%				
TUBB4b	59.37%	56.25%	96.87%	96.87%	100%	100%	100%			
TUBB5	59.37%	56.25%	96.87%	96.87%	100%	100%	100%	100%		
TUBB6	56.25%	53.12%	90.62%	90.62%	93.75%	93.75%	93.75%	93.75%	100%	
TUBB8	62.5%	59.37%	93.75%	93.75%	96.87%	96.87%	96.87%	96.87%	90.62%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.15 – *Residue Similarity of the potential binding site (1)*

TUBG1	100%									
TUBG2	100%	100%								
TUBB2a	52.94%	52.94%	100%							
TUBB2b	52.94%	52.94%	100%	100%						
TUBB3	52.94%	52.94%	94.28%	94.28%	100%					
TUBB4a	55.88%	55.88%	97.14%	97.14%	97.14%	100%				
TUBB4b	55.88%	55.88%	97.14%	97.14%	97.14%	100%	100%			
TUBB5	55.88%	55.88%	97.14%	97.14%	97.14%	100%	100%	100%		
TUBB6	52.94%	52.94%	88.57%	88.57%	94.28%	91.42%	91.42%	91.42%	100%	
TUBB8	55.88%	55.88%	97.14%	97.14%	97.14%	100%	100%	100%	91.42%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.16 – *Residue Similarity of the potential binding site (2)*

TUBG1	100%									
TUBG2	92.5%	100%								
TUBB2a	27.5%	35%	100%							
TUBB2b	27.5%	35%	100%	100%						
TUBB3	27.5%	35%	100%	100%	100%					
TUBB4a	27.5%	35%	100%	100%	100%	100%				
TUBB4b	27.5%	35%	100%	100%	100%	100%	100%			
TUBB5	27.5%	35%	100%	100%	100%	100%	100%	100%		
TUBB6	25%	32.5%	97.5%	97.5%	97.5%	97.5%	97.5%	97.5%	100%	
TUBB8	27.5%	35%	100%	100%	100%	100%	100%	100%	97.5%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.17 – *Residue Similarity of the potential binding site (3)*

TUBG1	100%									
TUBG2	100%	100%								
TUBB2a	46.87%	46.87%	100%							
TUBB2b	46.87%	46.87%	100%	100%						
TUBB3	50%	50%	96.87%	96.87%	100%					
TUBB4a	50%	50%	96.87%	96.87%	100%	100%				
TUBB4b	50%	50%	96.87%	96.87%	100%	100%	100%			
TUBB5	50%	50%	96.87%	96.87%	100%	100%	100%	100%		
TUBB6	46.87%	46.87%	93.75%	93.75%	96.87%	96.87%	96.87%	96.87%	100%	
TUBB8	53.12%	53.12%	93.75%	93.75%	96.87%	96.87%	96.87%	96.87%	93.75%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.18 – *Residue Similarity of the potential binding site (4)*

TUBG1	100%									
TUBG2	100%	100%								
TUBB2a	64.15%	64.15%	100%							
TUBB2b	64.15%	64.15%	100%	100%						
TUBB3	60.37%	60.37%	94.33%	94.33%	100%					
TUBB4a	62.26%	62.26%	98.11%	98.11%	92.45%	100%				
TUBB4b	64.15%	64.15%	100%	100%	94.33%	98.11%	100%			
TUBB5	62.26%	62.26%	96.22%	96.22%	94.33%	98.11%	96.22%	100%		
TUBB6	56.6%	56.6%	90.56%	90.56%	88.67%	92.45%	90.56%	90.56%	100%	
TUBB8	62.26%	62.26%	98.11%	98.11%	96.22%	96.22%	98.11%	94.33%	88.67%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.19 – *Residue Similarity of the potential binding site (5)*

TUBG1	100%									
TUBG2	97.95%	100%								
TUBB2a	63.41%	63.41%	100%							
TUBB2b	63.41%	63.41%	100%	100%						
TUBB3	63.41%	63.41%	100%	100%	100%					
TUBB4a	63.41%	63.41%	100%	100%	100%	100%				
TUBB4b	63.41%	63.41%	100%	100%	100%	100%	100%			
TUBB5	63.41%	63.41%	100%	100%	100%	100%	100%	100%		
TUBB6	65.85%	65.85%	92.68%	92.68%	92.68%	92.68%	92.68%	92.68%	100%	
TUBB8	65.85%	65.85%	97.56%	97.56%	97.56%	97.56%	97.56%	97.56%	95.12%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.20 – *Residue Similarity of the potential binding site (6)*

TUBG1	100%									
TUBG2	100%	100%								
TUBB2a	65.21%	65.21%	100%							
TUBB2b	65.21%	65.21%	100%	100%						
TUBB3	65.21%	65.21%	100%	100%	100%					
TUBB4a	65.21%	65.21%	100%	100%	100%	100%				
TUBB4b	65.21%	65.21%	100%	100%	100%	100%	100%			
TUBB5	65.21%	65.21%	100%	100%	100%	100%	100%	100%		
TUBB6	65.21%	65.21%	100%	100%	100%	100%	100%	100%	100%	
TUBB8	65.21%	65.21%	100%	100%	100%	100%	100%	100%	100%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.21 – *Residue Similarity of the potential binding site (7)*

TUBG1	100%									
TUBG2	100%	100%								
TUBB2a	65.38%	65.38%	100%							
TUBB2b	65.38%	65.38%	100%	100%						
TUBB3	61.53%	61.53%	96.15%	96.15%	100%					
TUBB4a	65.38%	65.38%	100%	100%	96.15%	100%				
TUBB4b	65.38%	65.38%	100%	100%	96.15%	100%	100%			
TUBB5	61.53%	61.53%	96.15%	96.15%	100%	96.15%	96.15%	100%		
TUBB6	61.53%	61.53%	96.15%	96.15%	100%	96.15%	96.15%	100%	100%	
TUBB8	65.38%	65.38%	100%	100%	96.15%	100%	100%	96.15%	96.15%	100%
	TUBG1	**TUBG2**	**TUBB2a**	**TUBB2b**	**TUBB3**	**TUBB4a**	**TUBB4b**	**TUBB5**	**TUBB6**	**TUBB8**

Figure 3.22 – *Residue Similarity of the potential binding site (8)*

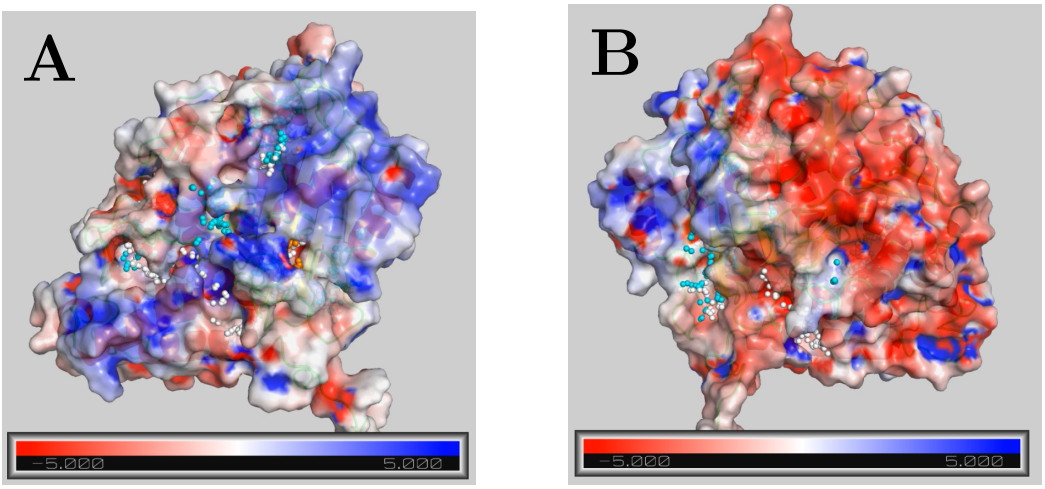


Figure 3.23 – *APBS - electrostatic maps of γT1; (A) front; (B) back.*

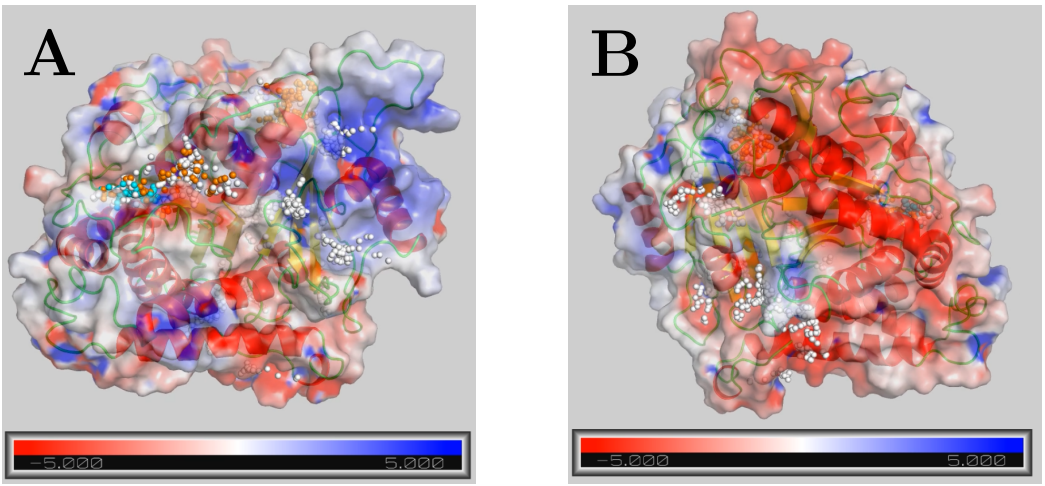


Figure 3.24 – *APBS - electrostatic maps of γT2; (A) front; (B) back.*

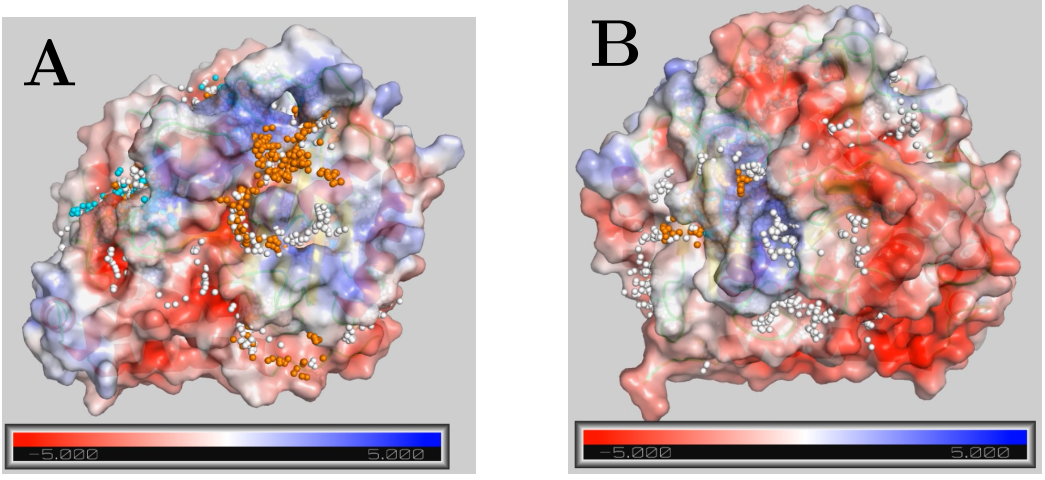


Figure 3.25 – *APBS - electrostatic maps of βTs (superposition); (A) front; (B) back.*

Since doing VS on eight potential binding sites is computationally and labour costly, other two tools are used to reduce the number of such pockets:

- the similarity of residues in a candidate pocket is calculated with the online tool SIAS (Sequence Identity And Similarity) (Fig. 3.15-22);
- electrostatic maps are calculated through the PyMol plugin APBS (Adaptive Poisson-Boltzmann Solver) for centroids of all clusters[3] (Fig. 3.23-25).

Given a large number of detailed images of electrostatic maps, everything has been compacted into videos, and they are available on the personal student account in UofA google drive.

By aggregating the similarity of residues, electrostatic maps and distribution and score of the dummies, careful remarks are made for each potential binding site:

Site 1) There is a slightly higher number of dummies in γT1 compared to γT2, but all their PLB scores are > 1 (Fig. 3.7A-B), although the similarity of 97.5% (Fig. 3.15). Electrostatic maps of γT1-2 show to be more positively charged than βT, where there are also negative areas. The similarity between γT1-2 and all βTs is lower than 60%, except for βT8.
This site is taken into account.

Site 2) This site is quite similar to the first for the same reasons (βTs show to have little positive area and at the same time show negative areas). Additionally, the similarity is lower than 60% for all βTs.
This site is taken into account.

Site 3) Although the very low amount of dummies in γT1 or their absence in γT2, the similarity between γT1-2 and all βTs significantly differs: 27.5%. Curiously, the similarity of this site is almost identical among the glsbTs (Fig. 3-17). However, the electrostatic maps are quite similar.
This site is taken into account.

Site 4) No space resembles a pocket in all βTs (further confirmed by calculating APBS on all centroids). Moreover, there are no dummies in βTs, while instead there are many in γT1, with some having PLB score >4. The similarity further confirmed the high difference between γT1-2 and all βTs (<50% except for βT8).
This site is taken into account.

Site 5) The distribution of dummies and electrostatic maps are similar among all tubulins and similarity is higher than 60%.
This site is not considered.

Site 6) The similarity is higher than 60%, while the dummies are few in γT1, while there are plenty in γT2, suggesting that this site could be a selective site for γT2. However, electrostatic maps are very similar and negatively charged.
This site is not considered.

Site 7) γT1 lacks totally the dummies and its electrostatic maps do not show possible space for a pocket. Additionally, this site shows the worst similarity among all tubulins. Even though it was said that pockets with PLB less than 1 would be

[3]The reason is to evaluate better how the charge in the potential pocket changes.

ignored, it is worth noting that there are no dummies around the N-terminus, which may be a target from E3 UBR1. Additionally, the electrostatic map of γT2 shows a possible pocket, since is strongly negatively charged, whereas the βT has weak positive areas. This site is not considered.

Site 8) Although the high score dummies in γT1 compared to γT2, there is little difference in electrostatic maps among all tubulins. This site is not considered.

Interestingly, the candidate binding sites taken into account (1,2,3,4) are sites all located in one "face" of the γT. This may suggest that this "face" are those that distinguish specific PPIs from other tubulins.

3.3 The conventional approach

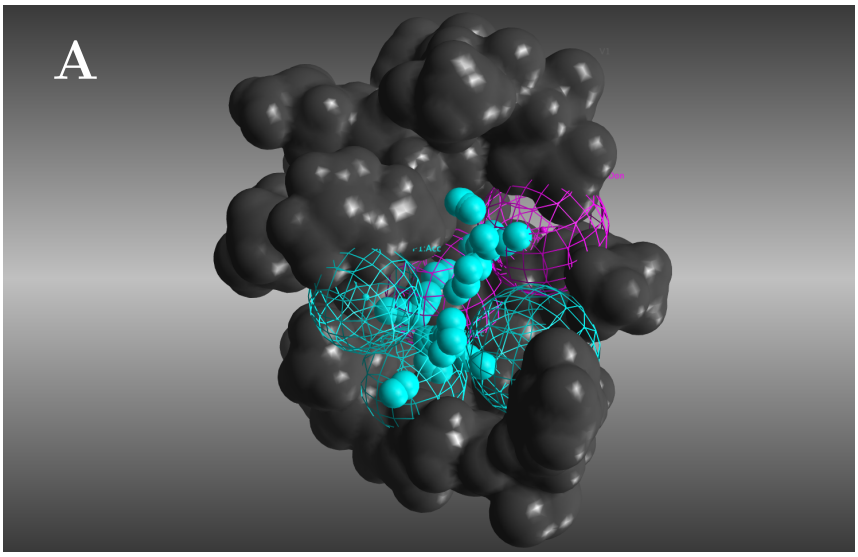
The preparation of ZINC15 databases has a slight reduction in size (Table 3.1).

The ligands of both databases are now sampled via Conformational Import. Unluckily, given the huge size of the entire ZINC15 database, the Conformational Import has not been completed for that database.
From here on, only biogenic DB will be further processed and taken as a comparison for the DeepDock approach.

The pharmacophore models are built iteratively to leave at least a handful of ligands, but in candidate binding sites 1 and 4, no ligands have the filtering successfully.

	biogenic DB	ZINC15 DB
start	156'695	885'492'702
raw data cleaning	103'693	557'709'097
1°washing	103'693	557'709'097
1°filtering	103'687	557'550'236
2°washing	178'527	874'930'148
2°filtering	152'021	821'677'638
sorting and uniq	152'021	821'677'657
Conform. Import	3'025'159	UNCOMPLETED
Ph4 Filtering Site 1	0	X
Ph4 Filtering Site 2	490	X
Ph4 Filtering Site 3	2'626	X
Ph4 Filtering Site 4	0	X

Table 3.1: Result pre-processing ZINC15 databases and pharmacophore filtering (biogenic DB is already included in ZINC15 DB)



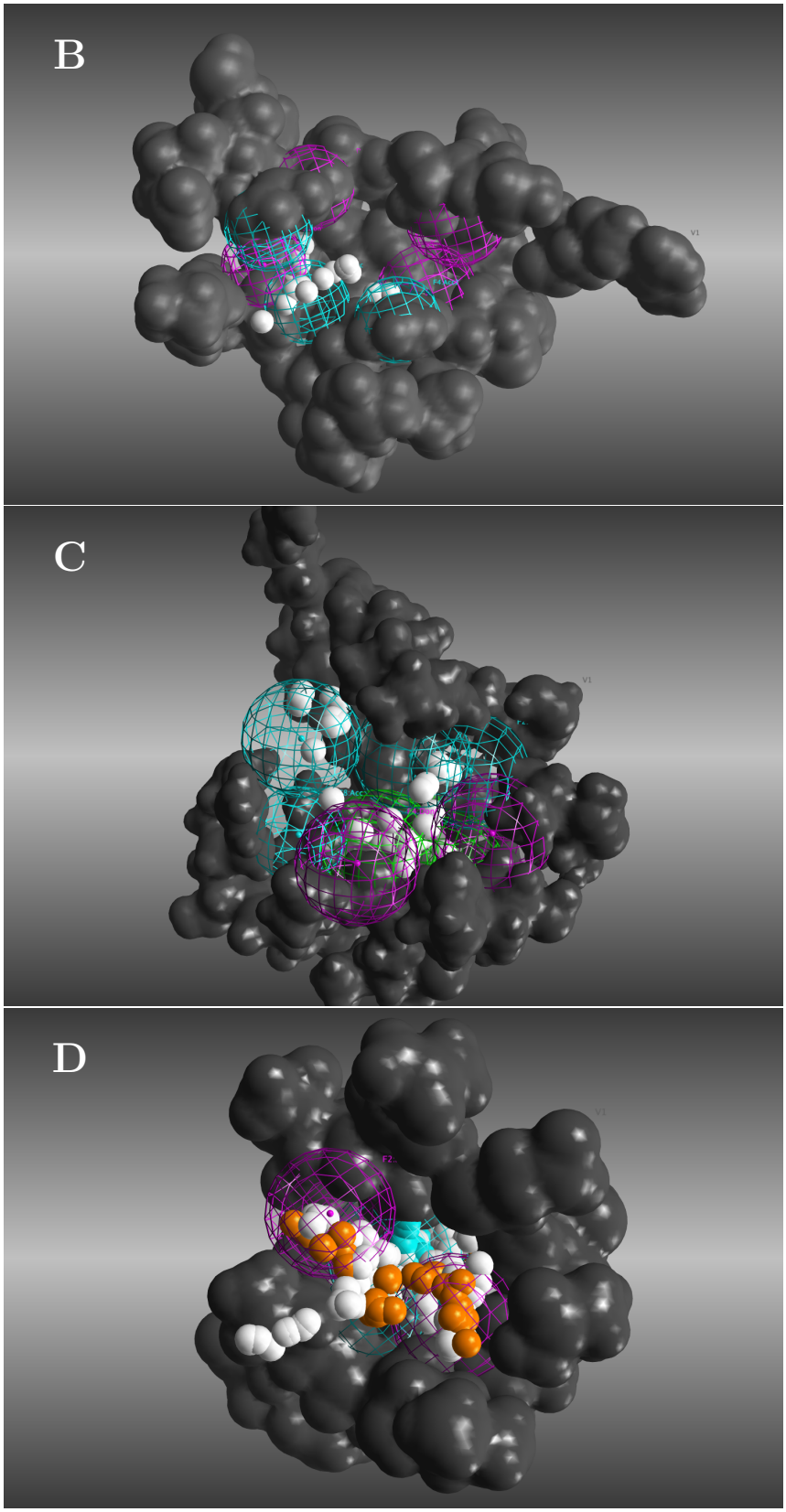


Figure 3.26 – *The pharmacophore protein-based models (ph4) built via Pharmacophore Editor MOE's tool: (A) ph4 for site 1; (B) ph4 for site 2; (C) ph4 for site 3; (D) ph4 for site 4. Dark-grey volume: volume excluded feature, dashed cyan sphere: acceptor feature; dashed magenta sphere: donor; white* ($1 < PLB < 3$)*, orange* ($3 < PLB < 4$) *and cyan* ($PLB > 4$) *small filled sphere are dummies generated in 2.1.4 section.*

3.4 The results of DockBox and DeepDock

For each iteration of the DNN training, the size of the training set is always approximately 1M, while the recall is always set to 0.9.
The predicted number from the test set given from the best model approximately always coincides with the effective predicted number of the reduced database from the database of the previous iteration (for the first iteration, this coincides with the original database of > 1B). Therefore, this proves the solidity of DNN models even when datasets that are not training sets, which change iteratively and is enriched with top-hit, are not used.

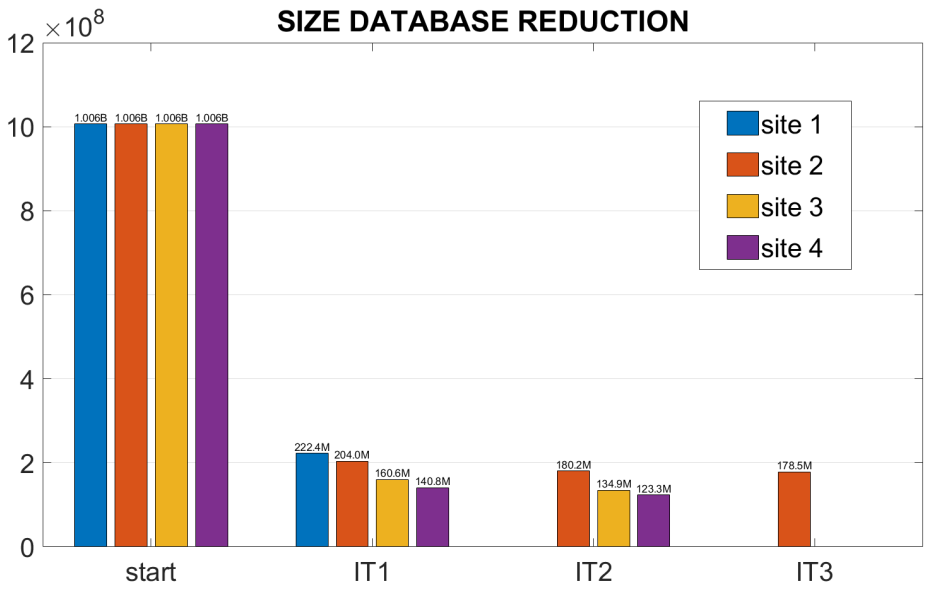


Figure 3.27 – *Reduction of the original database after several iterations of DNN training.*

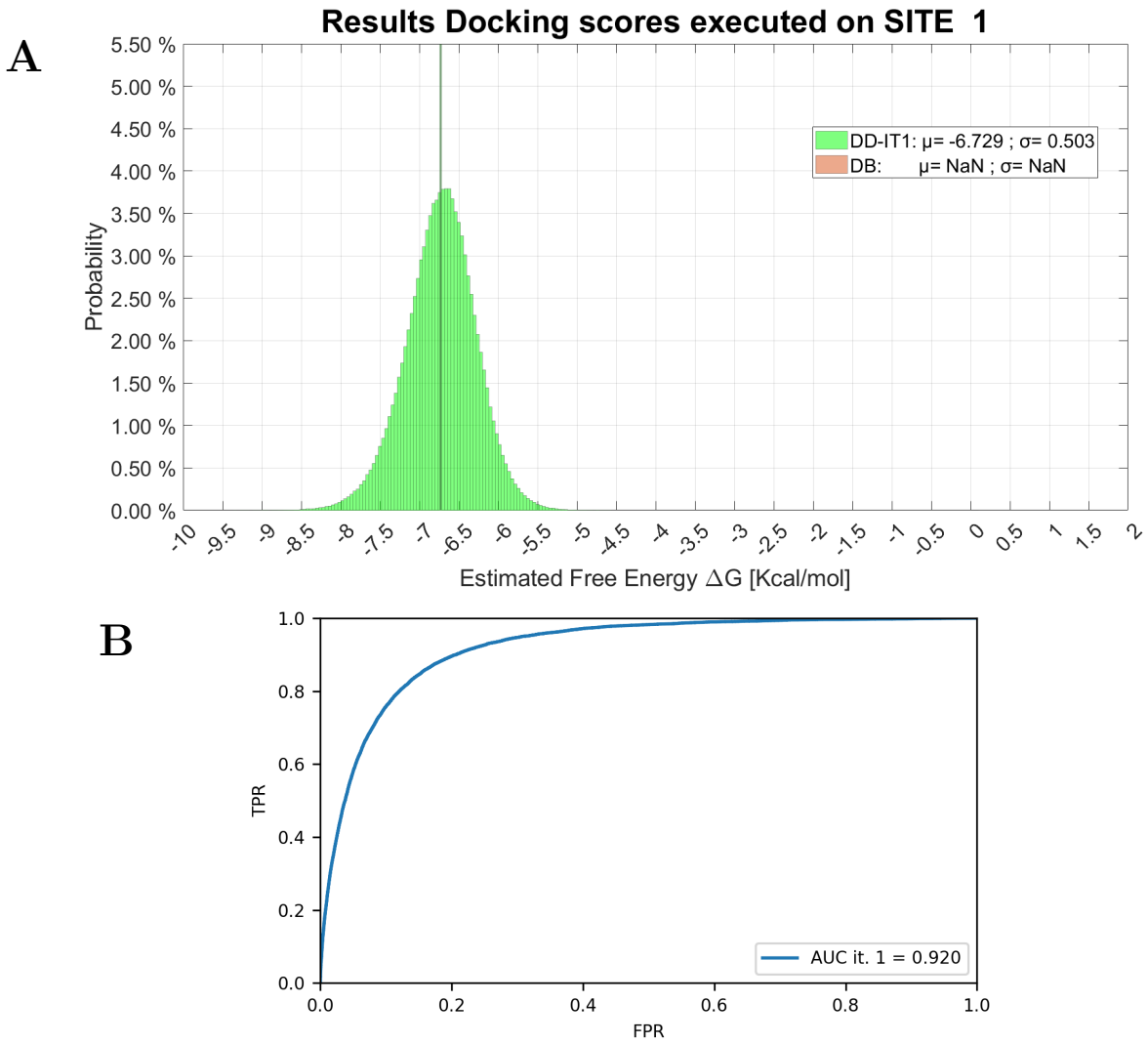


Figure 3.28 – *(A) Comparison of results between the two approaches in the site 1. DB= Conformational Search + Pharmacophore Filtering + DockBox (SBCD on MOE, VINA, DOCK6 is used). DD= DL-based Scoring approach (Conventional Docking on Training sets generated by DeepDock models for each iteration). Data prepared with MATLAB. (B) AUC Curve between the different iterations (DD approach). Prepared with plot_progress.py python tool available online.*

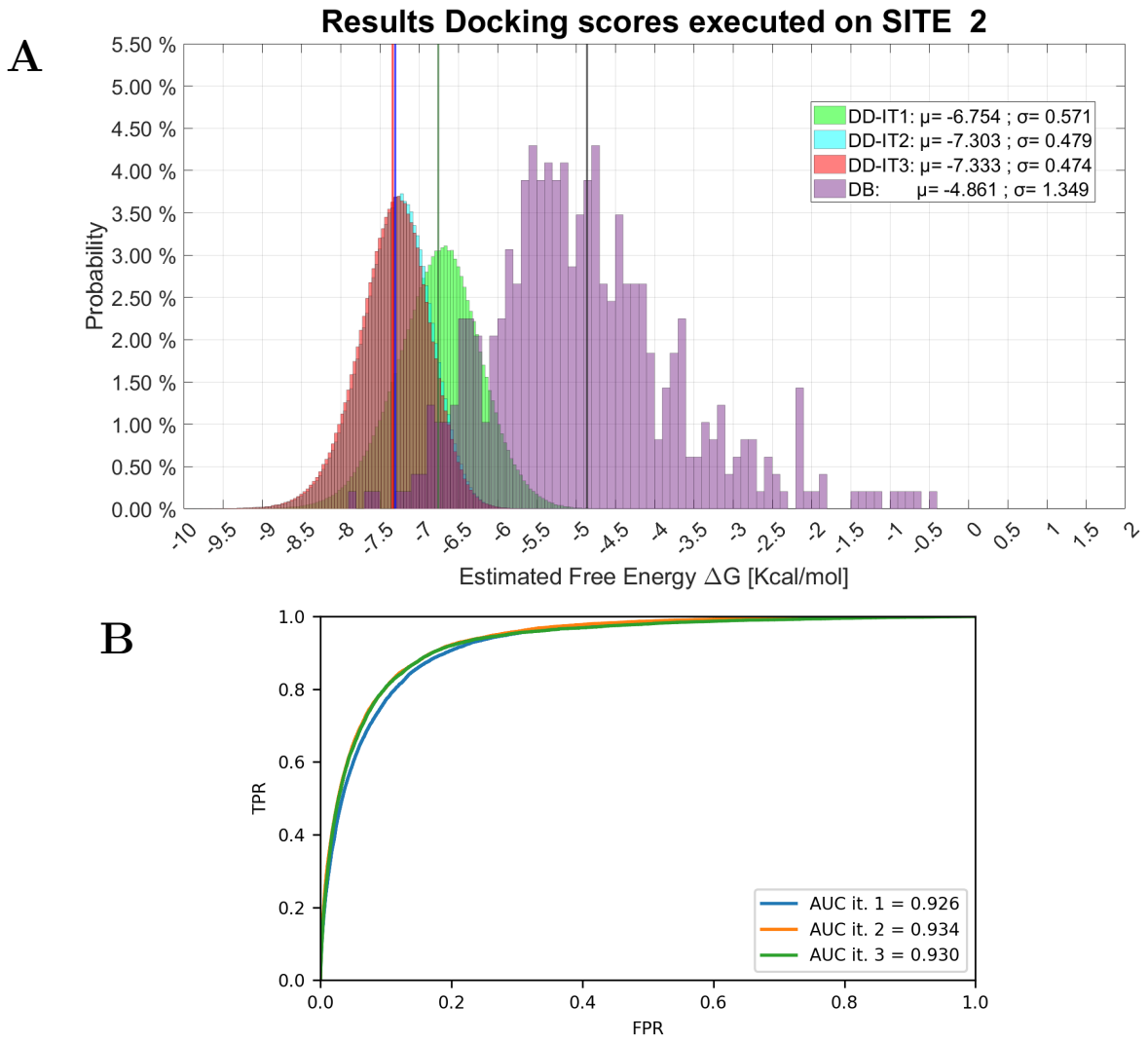


Figure 3.29 – *(A) Comparison of results between the two approaches in the site 2. DB= Conformational Search + Pharmacophore Filtering + DockBox (SBCD on MOE, VINA, DOCK6 is used). DD= DL-based Scoring approach (Conventional Docking on Training sets generated by DeepDock models for each iteration). Data prepared with MATLAB. (B) AUC Curve between the different iterations (DD approach). Prepared with plot_progress.py python tool available online.*

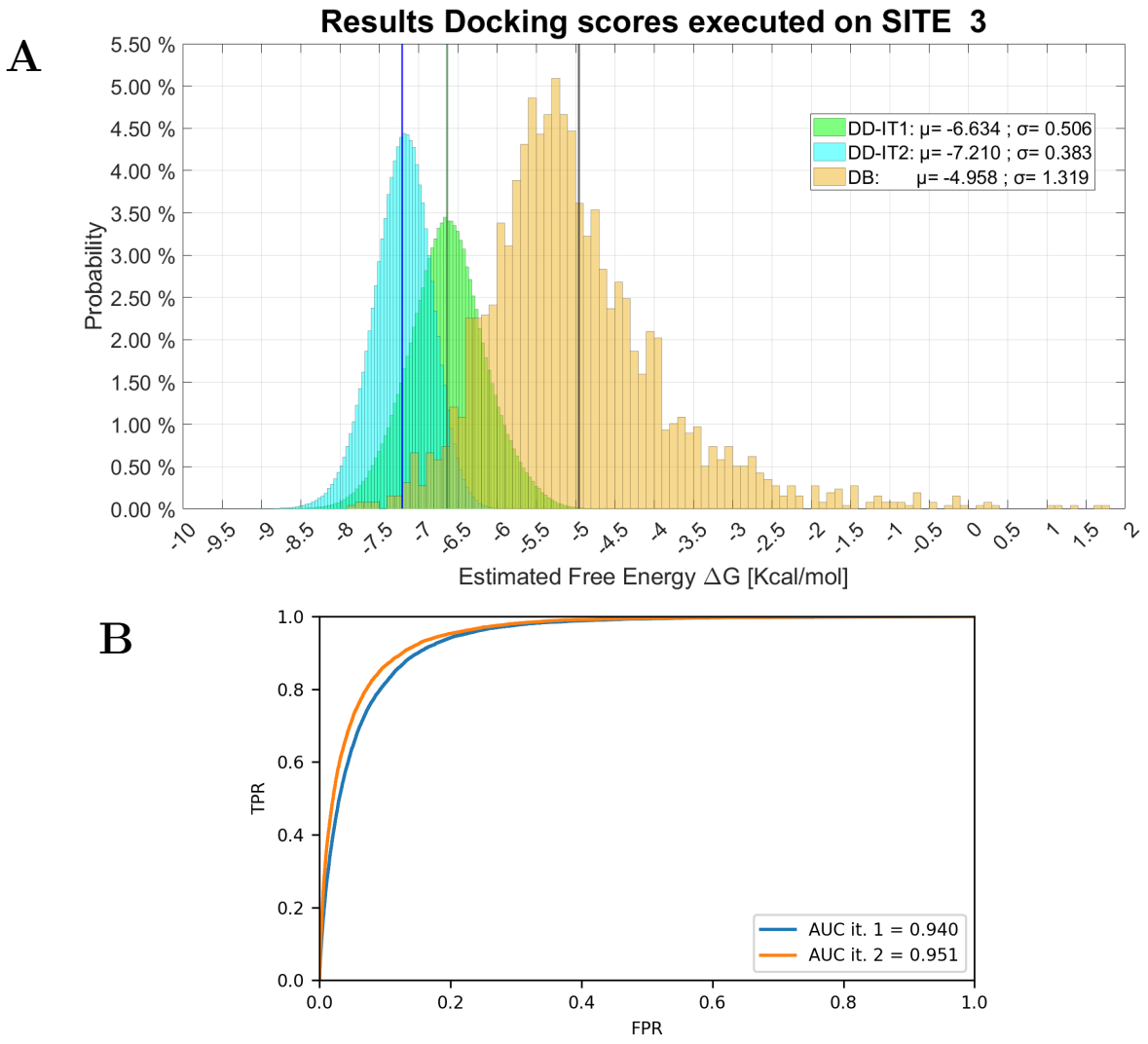


Figure 3.30 – *(A) Comparison of results between the two approaches in the site 3. DB= Conformational Search + Pharmacophore Filtering + DockBox (SBCD on MOE, VINA, DOCK6 is used). DD= DL-based Scoring approach (Conventional Docking on Training sets generated by DeepDock models for each iteration). Data prepared with MATLAB. (B) AUC Curve between the different iterations (DD approach). Prepared with plot_progress.py python tool available online.*

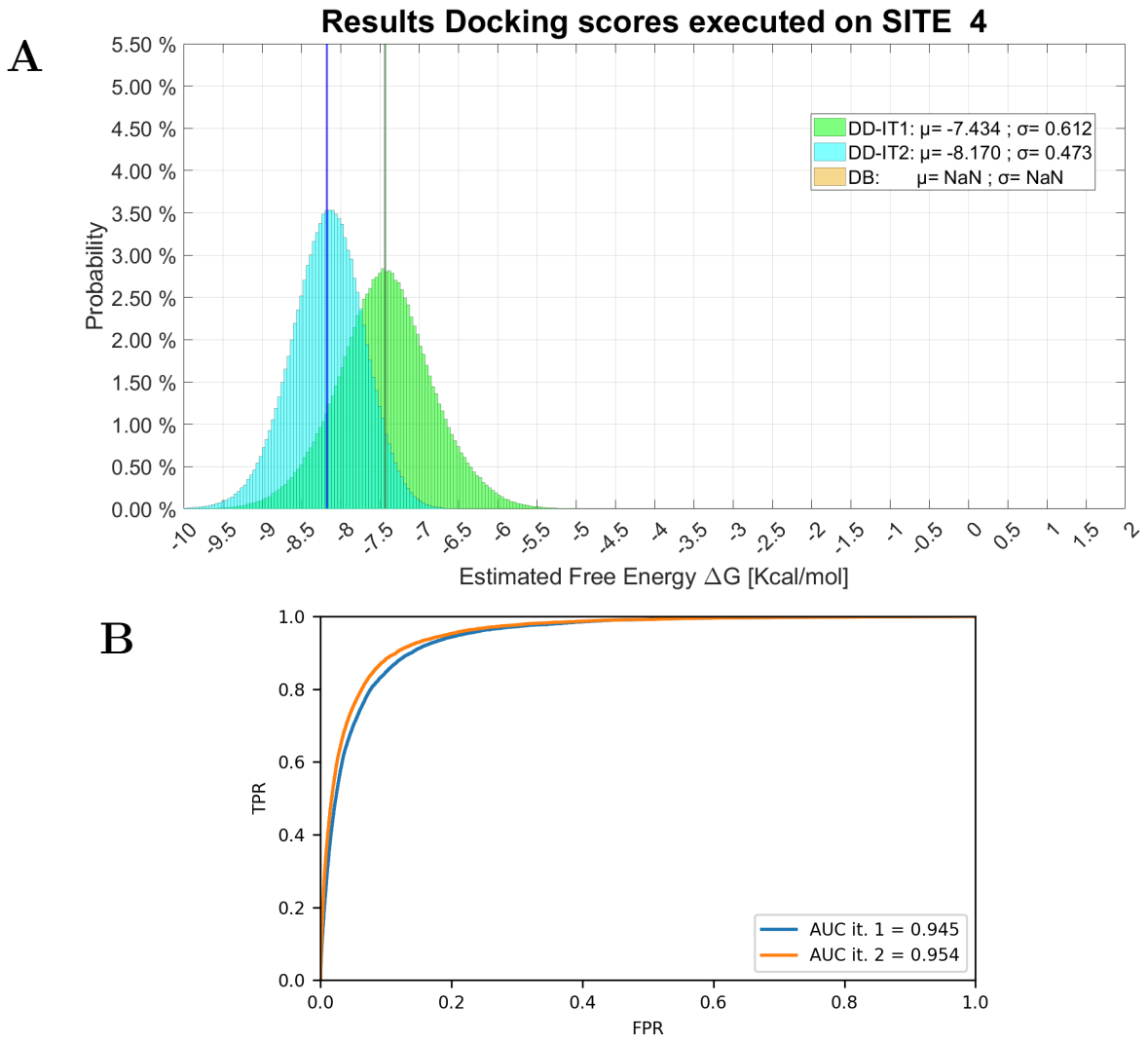


Figure 3.31 – *(A) Comparison of results between the two approaches in the site 4. DB= Conformational Search + Pharmacophore Filtering + DockBox (SBCD on MOE, VINA, DOCK6 is used). DD= DL-based Scoring approach (Conventional Docking on Training sets generated by DeepDock models for each iteration). Data prepared with MATLAB. (B) AUC Curve between the different iterations (DD approach). Prepared with plot_progress.py python tool available online.*

Chapter 4

Discussion

To design one or more PROTACs that target the γT through the UBR1 E3 ligase, a rational step-by-step approach is taken: the first step, the one made in the current work, is to find a binder that is selective only toward γT versus other tubulins.
In scientific literature, ligands that bind selectively such POI with solid evidence are not found. The only promising selective ligand is the Gatastatin G2; however, it binds specifically to the GTP binding site, which is believed that it is not a desirable site to target because the such pocket is highly conserved among eukaryotes. Still, more importantly, it is very similar to the GTP-binding site in αT and βT.

For these reasons, a VS is made to explore the multi-billion ligands databases available online.
However, conventional approaches are not sustainable regarding computational cost and time given the huge size of such databases, which, importantly, are growing more and more.

Thus, new techniques to reduce the databases prioritizing certain details without significantly losing others are essential.
Additionally, conventional approaches are also wasteful processes: once the top-hit ligands are found, all the rest is much thrown out, so in case a researcher/drug designer wants to try again using a new database later, she/he has to repeat the whole procedure all over again.

DeepDock overcomes these important issues by doing two simple things: reducing a database size by tending to leave top-hits quickly, especially if GPUs are used, and training DNN models, which can be used later when the new and updated databases are available.

Unfortunately, the work was revealed to be far more complex than expected and the results reported here are not complete for several reasons:

- it was intended to use state-of-the-art VS techniques. Thus much time was "spent" figuring out how DockBox and DeepDock work y reading their codes;
- many tools took time to learn how to use (PyMol, MOE, APBS, LaTex, Rosetta, AutoDockGPU and so on);
- thanks to ComputeCanada Summer School, there was a great opportunity to learn how to manage ComputeCanada HPC clusters as optimally as possible and with caution;

- the preparation of datasets required to train DNN models is relatively time-consuming, and continuous data integrity checking was also done to keep the dataset's size and any forms of bias low.

However, by what it has been able to do so far, several considerations can be made regarding the better performances of DNN approach compared to the conventional approach even when the maximum iteration reached is only 2 over 11 for the second binding site and 2 over 11 for the third and fourth binding site:

- a substantial portion ((1)= 77.89%, (2)= 79.72%, (3)= 84.04% and (4)= 86.0%) of the original database is removed at the first step by wiping off those unfavourable compounds easy to predict, as happened in the works of Gentile et al. (Fig. 3.31)[172, 176, 179];
- the rate of reduction decreases differently depending on the target as expected (Fig. 3.27);
- models appear to improve iteratively as the AUC curve increases (Fig. 3.29-31B);
- the distribution of docking scores in case of DNN approach significantly shift toward lower binding energy (Fig. 3.27-30A), while the distribution of docking scores of conventional approach shows to be not very optimal although the pharmacophore filtering application (Fig. 3.28-29A).
  However, it is important to note the different sizes of considered datasets: the training sets have 1M each, while the datasets of the conventional approach are 500÷2600, so it has a statistical significance that is not very valid as the former.

The aim of this work is worth to be finished; thus, even after graduation, it is going to continue since it has been allowed to extend access to Compute Canada Clusters.

After completing all iterations or when the database size converges, the few top-hits will be subjected to the conventional approach to select the final best ligands based on an intensive procedure of consensus docking and scoring.

The next step is trying to optimize the top-hits via combinatorial search, which allows the addition or remotion of single atoms or chemical groups based on the pocket chemistry and space. This process uses highly efficient algorithms with shape-based directional descriptors to screen hundred-thousands of fragments within seconds on standard hardware to create optimized suggestions [184].

If successful, the most exposed groups of the ligands will be considered linker attachment points, and the step regarding the γT will be finally completed.

Thereafter, the project will proceed to the following steps:

2) preparation of UBR1 E3 models through MD with the same protocols used to prepare the tubulins in the current work;
3) intensive consensus docking of E3+ligand[1] embedded in its known pocket;
4) poses analysis and checking the most exposed group where attach the linker covalently;
5) running conventional protein-protein docking via Rosetta framework and/or MOE platform to estimate the minimum and maximum linker's length;

[1]The ligand is provided by Dr.Richard Fahlman

6) designing the PROTACs with variable linkers and attachment points (if found more at the end of step 1 or 4)
7) perform Molecular Dynamics with POI:PROTAC:E3.

Improving the resources and changing docking program

Despite the effectiveness of DeepDock, some important factors should be set differently to minimize the general computational and time costs, especially regarding preparing the datasets required to train the DNN models.

For example, the *Conformational Search* and *Docking* MOE's tool has been revealed to be very computationally expensive (More than 500 core/years are dedicated only to prepare the datasets, Fig. 4.1A) and relatively slow, since its algorithms are configured to run on CPU systems. But recent updates suggest that these algorithms will soon integrate libraries to run on alternative computations units such as GPUs.

Figure 4.1 – *(A) CPU usage. A core year is an equivalent of running computations on a CPU core constantly for a period of one year. (B) GPU usage. A GPU year is an equivalent of running computations on one GPU unit for a period of one year.*

On the other hand, there are plenty of alternative docking programs, free and open source or commercial, with updated and efficient algorithms that can be run on GPU systems, such as AutoDockGPU. Nevertheless, they require little time to learn how to use it efficiently.

In comparison, DeepDock seems to be very efficient in performing high amounts of calculations by running on GPU (Fig. 4.1B).

The potential binding sites and the predicted degrons

Further investigations are made regarding the potential binding sites since it has been found while exploring known possible degrons in γT, a recent tool called DegPred which aims to predict the position of the degrons by giving the protein [37].

The tool is based on a BERT-based deep learning approach and the models are trained on a total of 303 degrons extracted from ELM motif database and literature with experimental results. Unlike other predictive algorithms available, which are often based on motifs only, rather than on the function itself, DegPred is able to integrate further information such as PTMs, phosphorylation sites, intrinsically disordered regions, molecular recognition features, solvent accessibility and rich LYS UP-sites [37].

The tool returns several information like (1) basic information such as known E3 ligase, (2) predicted degron sequence and relative scores, (3) disordered region scores and (4) ELM motif in correspondence of high score predicted degrons.

By providing the γT-1, two possible degrons are predicted (Fig. 4.3), but they locate quite far from the potential binding sites found in the current work (Fig. 4.2). Further experimental analysis is required to confirm the influence of these predicted degrons on the ubiquitination activity.

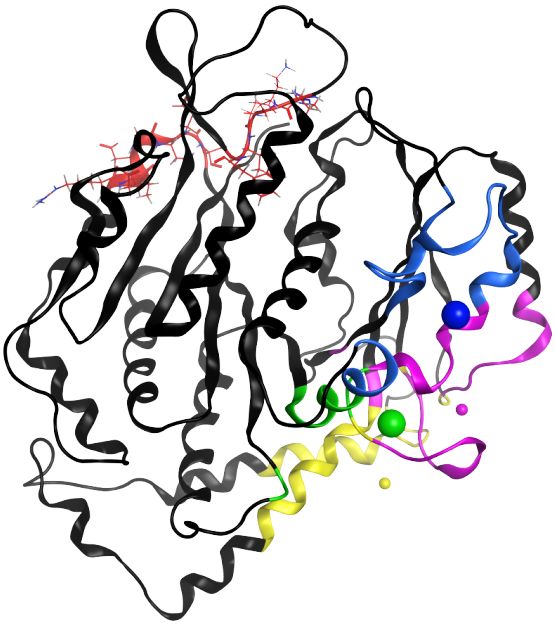


Figure 4.2 – *Model of γT1. Red ribbon = predicted degrons; blue, yellow, magenta, green ribbons = potential binding sites; spheres = center of dummies*

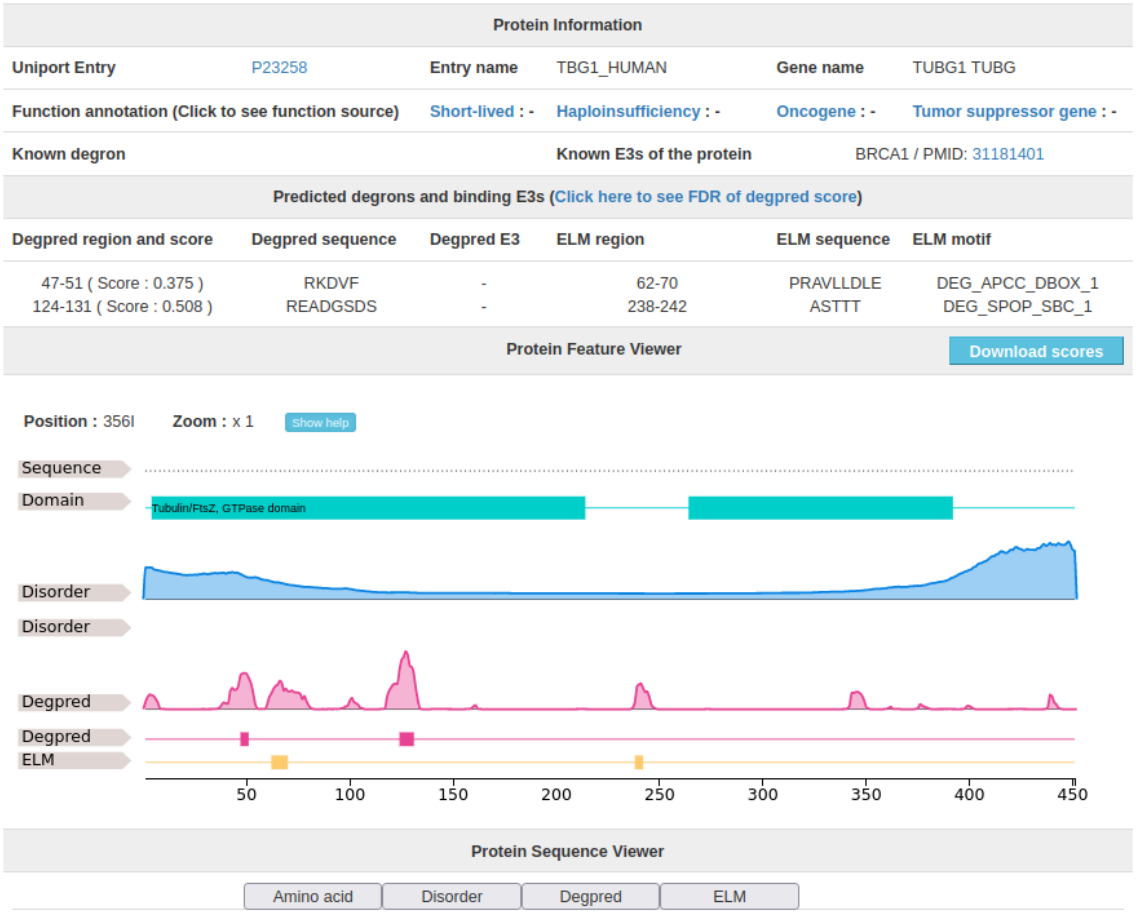


Figure 4.3 – *The results of DegPred degron predictive tool.*

Appendix A

All known and predicted interactions γ-Tubulin

node1	node2	coexpression	expDetInt	DBAnn	autoTextMining	cScore
TUBG1	PSMD8	0.108	0.27	0	0.538	0.673
TUBG1	AURKA	0.255	0.27	0	0.464	0.683
TUBG1	TUBGCP6	0.197	0.987	0.6	0.821	0.999
TUBG1	TUBG2	0.158	0.682	0.8	0.812	0.942
TUBG1	STIL	0.077	0	0	0.4	0.423
TUBG1	AKAP9	0.045	0.27	0	0.242	0.425
TUBG1	TPX2	0.231	0	0	0.316	0.452
TUBG1	TUBA1B	0.137	0.368	0	0.558	0.486
TUBG1	NIN	0	0	0	0.54	0.54
TUBG1	TOPORS	0	0	0.54	0	0.54
TUBG1	ASPM	0.133	0.282	0	0.334	0.55
TUBG1	CETN3	0.064	0.155	0	0.481	0.553
TUBG1	CEP152	0	0.094	0	0.536	0.561
TUBG1	TTK	0.151	0.164	0	0.45	0.576
TUBG1	NDC80	0.118	0.128	0	0.501	0.583
TUBG1	BLOC1S2	0.089	0	0.54	0.093	0.586
TUBG1	KIF11	0.208	0.289	0	0.379	0.619
TUBG1	HAUS8	0.051	0	0	0.63	0.633
TUBG1	ERC2	0	0.056	0	0.645	0.65
TUBG1	WDR62	0.076	0.27	0	0.53	0.655
TUBG1	CKAP5	0.165	0.257	0	0.5	0.663
TUBG1	CETN2	0.064	0.155	0	0.621	0.674
TUBG1	HAUS5	0.14	0	0	0.643	0.679
TUBG1	TXNDC12	0.053	0.683	0	0	0.686
TUBG1	LGALS3BP	0	0.685	0	0.063	0.692
TUBG1	KIFC3	0.131	0.524	0	0.341	0.703
TUBG1	DYNC1H1	0.046	0.184	0	0.656	0.709
TUBG1	HAUS1	0.072	0	0	0.708	0.718
TUBG1	PCNT	0.059	0.213	0	0.654	0.721
TUBG1	CNTLN	0	0	0	0.725	0.725
TUBG1	SSNA1	0.065	0	0	0.731	0.738
TUBG1	CEP162	0	0	0	0.74	0.74
TUBG1	MARK4	0	0.486	0.54	0	0.753
TUBG1	MAD2L1	0.194	0.631	0	0.277	0.766
TUBG1	NUMA1	0.072	0	0	0.764	0.771
TUBG1	KIFC1	0.187	0.524	0	0.497	0.788
TUBG1	SASS6	0.085	0	0	0.789	0.798
TUBG1	CEP135	0.064	0	0	0.8	0.805
TUBG1	CCP110	0	0.27	0	0.749	0.809
TUBG1	PLK4	0.095	0.14	0	0.778	0.812

TUBG1	ACTG1	0.096	0.82	0	0.29	0.874
TUBG1	CDK5RAP2	0	0.661	0	0.654	0.877
TUBG1	HAUS6	0.171	0.166	0	0.841	0.881
TUBG1	NME7	0.062	0.713	0.6	0.062	0.885
TUBG1	MZT2A	0	0.279	0.6	0.643	0.888
TUBG1	RPGR	0	0.948	0	0.062	0.949
TUBG1	BRCA1	0.118	0.874	0.54	0.189	0.953
TUBG1	PLK1	0.162	0.345	0.9	0.399	0.962
TUBG1	NEDD1	0.133	0.411	0.8	0.726	0.968
TUBG1	NINL	0.052	0.549	0.9	0.422	0.972
TUBG1	CENPJ	0.066	0.748	0.54	0.788	0.974
TUBG1	MZT2B	0.218	0.87	0.6	0.825	0.991
TUBG1	MZT1	0.065	0.944	0.6	0.846	0.996
TUBG1	TUBGCP5	0.062	0.983	0.6	0.537	0.996
TUBG1	TUBGCP4	0.122	0.986	0.6	0.759	0.998
TUBG1	TUBGCP2	0.327	0.985	0.8	0.787	0.999
TUBG1	TUBGCP3	0.206	0.993	0.8	0.74	0.999

Table A.1: Results of https://string-db.org/ using TUBG1 as keyword and level score > 0.4

coexpression = correlation expression in same sample;
expDetInt = found the interactions experimentally;
DBann = database annotations;
autoTextMining = found in the same study both keywords;
cScore = combined Score;

Appendix B

Codes to automatize the MD

Prepare the data

```
#!/bin/bash

read -p "TUBG1 | TUBG2 | TUBA | TUBB? " answer1
if [ $answer1 = "TUBG1" ]
then
    risp1="1_TUBG1"
elif [ $answer1 = "TUBG2" ]
then
    risp1="2_TUBG2"
elif [ $answer1 = "TUBA" ]
then
    risp1="3_TUBA1"
elif [ $answer1 = "TUBB" ]
then
    risp1="4_TUBB"
    read -p "what tubulin (put number)? [ 1) TUBB2a, 2) TUBB2b, 3)
   TUBB3, 4) TUBB4a, 5) TUBB4b, 6) TUBB5, 7) TUBB6, 8) TUBB8 ]: " tubb
else
    exit 1
fi


read -p "empty or GDP or GTP ? " answer2
if [ $answer2 = "empty" ]
then
    risp2="1_empty"
elif [ $answer2 = "GDP" ]
then
    risp2="2_GDP"
elif [ $answer2 = "GTP" ]
then
    risp2="3_GTP"
else
    exit 1
fi

if [[ $tubb = [12345678] ]]
then
    case $tubb in
        1) rispT=1_TUBB2a ;;
        2) rispT=2_TUBB2b ;;
        3) rispT=3_TUBB3  ;;
```

```
42          4) rispT=4_TUBB4a ;;
43          5) rispT=5_TUBB4b ;;
44          6) rispT=6_TUBB5  ;;
45          7) rispT=7_TUBB6  ;;
46          8) rispT=8_TUBB8  ;;
47      esac
48      directory=$risp1/$rispT/$risp2
49  else
50      directory=$risp1/$risp2
51  fi
52
53  sorgente=`pwd`
54  cd $directory
55
56  gmx pdb2gmx -f 1*.pdb -o 2_start.gro -ignh <<eof
57      1
58      1
59  eof
60  gmx editconf -f 2_start.gro -o 3_box -c -d 1.2 -bt dodecahedron
61  gmx solvate -cp 3_box.gro -o 4_solvate.gro -p topol.top
62  gmx grompp -f $sorgente/fileMDP/em.mdp -c 4_solvate.gro -o 5_solvate.
      tpr -p topol.top -maxwarn 2
63  gmx genion -s 5_solvate.tpr -neutral -conc 0.15 -o 6_ions.gro -p topol.
      top -pname SOD -nname CLA
64  gmx make_ndx -f 6_ions.gro
65
66  ##transfer files on cluster - NB: dirs are named in the same way in the
       cluster too
67  dirRemote="fabiano@narval.computecanada.ca:/home/fabiano/scratch/
      fabiano/1_gromacsSimulations/$directory"
68  scp -r charmm36-jul2021.ff/ 6_ions.gro topol* posre* index.ndx
      $dirRemote/.
```

Automate the MD (on cluster)

```
#!/bin/bash

#SBATCH --account=def-jtus
#SBATCH --nodes=1
#SBATCH --gres=gpu:a100:4       ##request 1 GPU as generic resource
#SBATCH --cpus-per-task=12       ## number of OpenMP threads per MPI process
#SBATCH --mem-per-cpu=1500M
#SBATCH --time 0-20:00:00            # time limit (D-HH:MM:ss)
#SBATCH --ntasks=4
#SBATCH --mail-user=fab.alt@protonmail.com
#SBATCH --mail-type=BEGIN
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL

module purge
## load the necessary modules for gromacs
module load StdEnv/2020 gcc/9.3.0 cuda/11.4 openmpi/4.0.3 gromacs/2021.4

export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK

#FIRST ARGUMENT:     $1= 1_TUBG1 | 2_TUBG2 | 3_TUBA1 | 4_TUBB
#SECOND ARGUMENT:    $2= 1_empty | 2_GDP | 3_GTP
#THIRD ARGUMENT (not mandatory):     $3= 1_TUBB2a | 2_TUBB2b | 3_TUBB3 | 4_TUBB4a | 5_TUBB4b | 6_TUBB5 | 7_TUBB6 | 8_TUBB8
sorgente=`pwd`
if [[ -n $3 ]]
then
        directory=$1/$3/$2
else
        directory=$1/$2
fi

cd $directory

########################################## EM
gmx grompp -f $sorgente/fileMDP/em.mdp -c 6_ions.gro -o 7_em.tpr -p topol.top -maxwarn 2
gmx mdrun -ntmpi 1 -ntomp $SLURM_CPUS_PER_TASK -deffnm 8_em -nb gpu -bonded cpu -s 7_em.tpr
########################################## NVT
gmx grompp -f $sorgente/fileMDP/nvt.mdp -c 8_em.gro -p topol.top -o 9_nvt.tpr -r 8_em.gro -maxwarn 2
gmx mdrun -gputasks 0123 -ntmpi 4 -ntomp $SLURM_CPUS_PER_TASK -bonded cpu -nb gpu -pme gpu -npme 1 -deffnm 10_nvt -s 9_nvt.tpr
########################################## NPT
gmx grompp -f $sorgente/fileMDP/npt.mdp -c 10_nvt.gro -p topol.top -o 11_npt.tpr -r 10_nvt.gro -maxwarn 2
gmx mdrun -gputasks 0123 -ntmpi 4 -ntomp $SLURM_CPUS_PER_TASK -bonded cpu -nb gpu -pme gpu -npme 1 -deffnm 12_npt -s 11_npt.tpr
########################################## MD
gmx grompp -f $sorgente/fileMDP/md.mdp -c 12_npt.gro -p topol.top -o 13_md.tpr -r 12_npt.gro -maxwarn 2
```

```
gmx mdrun -gputasks 0123 -ntmpi 4 -ntomp $SLURM_CPUS_PER_TASK -bonded cpu -nb gpu -pme gpu -npme 1 -deffnm 14_md -s 13_md.tpr

########################################## fix trj
gmx trjconv -s 13_md.tpr -f 14_md.xtc -o 15_md_adj.xtc -pbc mol -ur compact -center <<eof
 1
 0
eof
# save the first frame
gmx trjconv -f 15_md_adj.xtc -s 13_md.tpr -dump 0 -o 15_md_0.gro <<eof
 0
eof

################################## full RMSD, save on analysis dir
mkdir analysis
if [ $2 = '1_empty' ]
then
    gmx rms -s 13_md.tpr -f 15_md_adj.xtc -fit rot+trans -n index.ndx -m analysis/rmsdMatrix_entire.xpm -o analysis/rmsd_entire.xvg <<eof
 1
 4
eof
elif [ $1 = '3_TUBA1' ]
then
    gmx rms -s 13_md.tpr -f 15_md_adj.xtc -fit rot+trans -n index.ndx -m analysis/rmsdMatrix_entire.xpm -o analysis/rmsd_entire.xvg <<eof
 26
 4
eof
else
    gmx rms -s 13_md.tpr -f 15_md_adj.xtc -fit rot+trans -n index.ndx -m analysis/rmsdMatrix_entire.xpm -o analysis/rmsd_entire.xvg <<eof
 19
 4
eof
fi
```

Retrieve the results from cluster and run post-processing

```
#!/bin/bash

sorgenteRemote=/home/fabiano/scratch/fabiano/1_gromacsSimulations

# first arg: 1_TUBG1 | 2_TUBG2
# second arg: 1_empty | 2_GDP | 3_GTP

#FIRST ARGUMENT:     $1= 1_TUBG1 | 2_TUBG2 | 3_TUBA1 | 4_TUBB
#SECOND ARGUMENT:    $2= 1_empty | 2_GDP | 3_GTP
#THIRD ARGUMENT (not mandatory):     $3= 1_TUBB2a | 2_TUBB2b | 3_TUBB3 | 4_TUBB4a | 5_TUBB4b | 6_TUBB5 | 7_TUBB6 | 8_TUBB8

if [[ -n $3 ]]
then
        directory=$1/$3/$2
else
        directory=$1/$2
fi

mkdir $directory/RESULTs
mkdir $directory/RESULTs/analysis
scp fabiano@narval.computecanada.ca:$sorgenteRemote/$directory/1[35]_md* $directory/RESULTs
scp fabiano@narval.computecanada.ca:$sorgenteRemote/$directory/analysis/rmsd* $directory/RESULTs/analysis

cd $directory/RESULTs/analysis
##NOTE: after looking full RMSD, check the start time in which there is equilibrium along RMSD curve
read "start time X EQL? " startTime
gmx xpm2ps -f rmsdMatrix_"$startTime"ns.xpm -o plot_rmsdMatrix_"$startTime"ns.eps

mkdir rmsd
mv rmsd[_M]* plot* rmsd/

gmx sasa -s ../13_md.tpr -f ../15_md_adj.xtc -o sasa.xvg -odg sasa_estimatedSolvationFreeEnergy.xvg -tv sasa_totalVolum_density.xvg
mkdir sasa
mv sasa[_.]* sasa/

mkdir cluster
cd cluster
gmx cluster -method gromos -f ../../15_md_adj.xtc -s ../../13_md.tpr -dm ../rmsd/rmsdMatrix_"$startTime"ns.xpm -n ../../../index.ndx -o rmsd-clust_"$startTime"ns.xpm -sz rmsd-size_"$startTime"ns.xvg -cl -b $startTime -tu ns -cutoff 0.15
gmx xpm2ps -f rmsd-clust_"$startTime"ns.xpm -o plot_rmsd-clust_"$startTime"ns.eps
```

Appendix C

Parameters for each simulation step

Energy Minimization

```
title                = Minimization  ; Title of run
; Parameters describing what to do, when to stop and what to save
integrator           = steep         ; Algorithm (steep = steepest descent
    minimization)
emtol                = 10.0          ; Stop minimization when the maximum
    force < 10.0 kJ/mol
emstep               = 0.01          ; Energy step size
nsteps               = 100000        ; Maximum number of (minimization)
    steps to perform

; NEIGHBORSEARCHING PARAMETERS
cutoff-scheme   = Verlet
ns_type              = grid          ; Method to determine neighbor list (
    simple, grid)
rlist                = 1.2           ; Cut-off for making neighbor list (
    short range forces)
nstlist              = 1             ; Frequency to update the neighbor list
     and long range forces
comm-mode            = Linear
nstcomm              = 100

; Periodic boundary conditions
pbc                  = xyz           ; 3-D PBC
periodic_molecules  = no

; Allowed energy error due to the Verlet buffer in kJ/mol/ps per atom,
; a value of -1 means: use rlist
verlet-buffer-tolerance     = 0.005

;------------------------------------------------------------
; BONDS
;------------------------------------------------------------
constraints          = h-bonds
; Type of constraint algorithm
constraint-algorithm     = lincs
; Highest order in the expansion of the constraint coupling matrix
lincs-order          = 4

;------------------------------------------------------------
; ELECTROSTATICS
;------------------------------------------------------------

```

```
; grid spacing of 0.1 nm and cubic interpolation the electrostatic
    forces have an accuracy of 2-3*10-4.

coulombtype     = P3M-AD        ; PP Particle-Mesh algorithm with
    analytical derivative for long range electrostatic interactions.
    The method and code is identical to SPME (i.e. coulomb-type:PME),
    except that the influence function is optimized for the grid. This
    gives a slight increase in accuracy.
rcoulomb        = 1.2

; EWALD/PME/PPPM parameters
fourierspacing  = 0.1           ; grid spacing for FFT
pme_order       = 4             ; interpolation order (4=cubic)
ewald_rtol      = 1e-06         ; The relative strength of the Ewald-
    shifted direct potential
ewald-rtol-lj   = 0.001
lj-pme-comb-rule= Geometric
ewald_geometry  = 3d
epsilon_surface = 0
implicit_solvent= No

;-----------------------------------------------------------
; VAN DER WAALS (LJ)
;-----------------------------------------------------------
vdwtype         = cutoff        ; switch is deprecated, replaced by
    using vdwtype=Cut-off with vdw-modifier=Potential-switch
vdw-modifier    = Potential-switch
rvdw-switch     = 1.0           ; inner radium (nm)
rvdw            = 1.2           ; outer radium (nm)

; Apply long range dispersion corrections for Energy and Pressure
DispCorr        = No
; Extension of the potential lookup tables beyond the cut-off
table-extension = 1
```

Canonical ensemble (NVT)

```
title                    = Protein-ligand complex NVT equilibration
; Parameters describing what to do, when to stop and what to save
integrator               = md            ; leap-frog integrator
dt                       = 0.002         ; 2 fs
nsteps                   = 50000         ; 2 * 50000 = 100 ps
comm-mode                = None
nstcomm                  = 100
continuation             = no            ; first dynamics run
; Output control
nstenergy                = 500           ; save energies every 1.0 ps
nstlog                   = 500           ; update log file every 1.0 ps
nstxout-compressed       = 500           ; save coordinates every 1.0 ps

; Periodic boundary conditions
pbc                      = xyz           ; 3-D PBC
periodic_molecules       = no

;---------------------------------------------------------------------
; NEIGHBORSEARCHING PARAMETERS
;---------------------------------------------------------------------
cutoff-scheme            = Verlet
ns_type                  = grid
rlist                    = 1.2
nstlist                  = 20                  ; largely irrelevant with Verlet
; Allowed energy error due to the Verlet buffer in kJ/mol/ps per atom,
    a value of -1 means: use rlist
verlet-buffer-tolerance  = 0.005

;---------------------------------------------------------------------
; BONDS
;---------------------------------------------------------------------
constraint_algorithm     = lincs      ; holonomic constraints
constraints              = h-bonds    ; bonds to H are constrained
lincs_iter               = 1          ; accuracy of LINCS
lincs_order              = 4          ; also related to accuracy

;---------------------------------------------------------------------
; ELECTROSTATICS
;---------------------------------------------------------------------
coulombtype              = P3M-AD
coulomb-modifier         = None
rcoulomb-switch          = 1.1
rcoulomb                 = 1.2
; Relative dielectric constant for the medium and the reaction field
epsilon_r                = 1     ; (default) the relative dielectric constant
    . A value of 0 means infinity.
epsilon_rf               = 0     ; The relative dielectric constant of the
    reaction field. This is only used with reaction-field
    electrostatics

; EWALD/PME/PPPM parameters
```

```
48 fourierspacing  = 0.1            ; grid spacing for FFT
49 pme_order       = 4              ; interpolation order (4=cubic)
50 ewald_rtol      = 1e-06          ; The relative strength of the Ewald-
       shifted direct potential
51 ewald-rtol-lj   = 0.001
52 lj-pme-comb-rule= Geometric
53 ewald_geometry  = 3d
54 epsilon_surface = 0
55 implicit_solvent= No
56
57 ;-----------------------------------------------------------
58 ; VAN DER WAALS (LJ)
59 ;-----------------------------------------------------------
60 vdwtype          = cutoff
61 vdw-modifier     = Potential-switch
62 rvdw-switch      = 1.0            ; inner radium (nm)
63 rvdw             = 1.2            ; outer radium (nm)
64
65 ; Apply long range dispersion corrections for Energy and Pressure
66 DispCorr         = No
67 ; Extension of the potential lookup tables beyond the cut-off
68 table-extension = 1
69
70 ;-----------------------------------------------------------
71 ; NVT parameters
72 ;-----------------------------------------------------------
73 ; Temperature coupling
74 tcoupl           = V-rescale              ; modified Berendsen thermostat
75 tc-grps          = non-Water Water        ; two coupling groups - more
       accurate
76 tau_t            = 0.1    0.1             ; time constant, in ps
77 ref_t            = 310    310             ; reference temperature, one
       for each group, in K
78 ; Pressure coupling
79 pcoupl           = no                     ; no pressure coupling in NVT
80 ; Velocity generation
81 gen_vel         = no                     ; no assign velocities from
       Maxwell distribution, start from T=0K instead
```

Isothermal–isobaric ensemble (NPT)

```
title                    = Protein-ligand complex NPT equilibration
; Parameters describing what to do, when to stop and what to save
integrator               = md                ; leap-frog integrator
dt                       = 0.002             ; 2 fs
nsteps                   = 50000             ; 2 * 50000 = 100 ps
comm-mode                = None
nstcomm                  = 100
continuation             = yes               ; velocities from NVT run

; Output control
nstenergy                = 500               ; save energies every 1.0 ps
nstlog                   = 500               ; update log file every 1.0 ps
nstxout-compressed       = 500               ; save coordinates every 1.0 ps

; Periodic boundary conditions
pbc                      = xyz               ; 3-D PBC
periodic_molecules       = no

;============================================================
; NEIGHBORSEARCHING PARAMETERS
;============================================================
cutoff-scheme            = Verlet
ns_type                  = grid
rlist                    = 1.2
nstlist                  = 20
verlet-buffer-tolerance  = 0.005

;============================================================
; BONDS
;============================================================
constraint_algorithm     = lincs
constraints              = h-bonds
lincs_iter               = 1
lincs_order              = 4

;============================================================
; ELECTROSTATICS
;============================================================
coulombtype              = P3M-AD
coulomb-modifier         = None
rcoulomb-switch          = 1.1
rcoulomb                 = 1.2
; Relative dielectric constant for the medium and the reaction field
epsilon_r                = 1
epsilon_rf               = 0
; EWALD/PME/PPPM parameters
fourierspacing           = 0.1
pme_order                = 4
ewald_rtol               = 1e-05             ; INCREASED by 10-fold compared to NVT
    run
ewald-rtol-lj            = 0.001
```

```
lj-pme-comb-rule= Geometric
ewald_geometry  = 3d
epsilon_surface = 0
implicit_solvent= No

;----------------------------------------------------------
; VAN DER WAALS (LJ)
;----------------------------------------------------------
vdwtype         = cutoff
vdw-modifier    = Potential-switch
rvdw-switch     = 1.0           ; inner radium (nm)
rvdw            = 1.2           ; outer radium (nm)

; Apply long range dispersion corrections for Energy and Pressure
DispCorr        = No
; Extension of the potential lookup tables beyond the cut-off
table-extension = 1

;----------------------------------------------------------
; NVT parameters
;----------------------------------------------------------
; Temperature coupling
tcoupl          = V-rescale
tc-grps         = non-Water Water
tau_t           = 0.1    0.1
ref_t           = 310    310
; Pressure coupling
pcoupl          = Berendsen                    ; pressure coupling is on
    for NPT
pcoupltype      = isotropic                    ; uniform scaling of box
    vectors
tau_p           = 2.0                          ; time constant, in ps
ref_p           = 1.0                          ; reference pressure, in
    bar
compressibility = 4.5e-5                       ; isothermal
    compressibility of water, bar^-1
refcoord_scaling= com

; Velocity generation
gen_vel         = no
```

Molecular Dynamics Parameters

```
title           = Protein-ligand complex MD equilibration
; Parameters describing what to do, when to stop and what to save
integrator      = md            ; leap-frog integrator
dt              = 0.002         ; 2 fs
nsteps          = 50000000      ; 2 * 50000000 = 100000 ps (100 ns)
comm-mode       = None
nstcomm         = 100
continuation    = yes           ; velocities from NPT run

; Output control
nstenergy       = 20000         ; save energies every 50.0 ps
nstlog          = 200000        ; update log file every 500.0 ps
compressed-x-precision  = 1000
; Output frequency and precision for .xtc file
nstxout-compressed  = 5000      ; save coordinates every 10.0 ps in
    compressed file to reduce file size

; Periodic boundary conditions
pbc                 = xyz       ; 3-D PBC
periodic_molecules  = no

;-------------------------------------------------------------------
; NEIGHBORSEARCHING PARAMETERS
;-------------------------------------------------------------------
cutoff-scheme   = Verlet
ns_type         = grid
rlist           = 1.2
nstlist         = 20
verlet-buffer-tolerance  = 0.005

;-------------------------------------------------------------------
; BONDS
;-------------------------------------------------------------------
constraint_algorithm    = lincs
constraints             = h-bonds
lincs_iter              = 1
lincs_order             = 4

;-------------------------------------------------------------------
; ELECTROSTATICS
;-------------------------------------------------------------------
coulombtype         = P3M-AD
coulomb-modifier    = None
rcoulomb-switch     = 1.1
rcoulomb            = 1.2
; Relative dielectric constant for the medium and the reaction field
epsilon_r           = 1
epsilon_rf          = 0
; EWALD/PME/PPPM parameters
fourierspacing  = 0.1
pme_order       = 4
```

```
ewald_rtol      = 1e-0&         ; DECREASED BY 10-fold compared to NPT
    run
ewald-rtol-lj   = 0.001
lj-pme-comb-rule= Geometric
ewald_geometry  = 3d
epsilon_surface = 0
implicit_solvent= No

;----------------------------------------------------------
; VAN DER WAALS (LJ)
;----------------------------------------------------------
vdwtype         = cutoff
vdw-modifier    = Potential-switch
rvdw-switch     = 1.0           ; inner radium (nm)
rvdw            = 1.2           ; outer radium (nm)

; Apply long range dispersion corrections for Energy and Pressure
DispCorr        = No
; Extension of the potential lookup tables beyond the cut-off
table-extension = 1

;----------------------------------------------------------
; NVT parameters
;----------------------------------------------------------
; Temperature coupling
tcoupl          = V-rescale
tc-grps         = non-Water Water
tau_t           = 0.1   0.1
ref_t           = 310   310
; Pressure coupling
pcoupl          = Parrinello-Rahman         ; pressure coupling is on
    for NPT, model adapted to md
pcoupltype      = isotropic                 ; uniform scaling of box
    vectors
tau_p           = 2.0                       ; time constant, in ps
ref_p           = 1.0                       ; reference pressure, in
    bar
compressibility = 4.5e-5                    ; isothermal
    compressibility of water, bar^-1

; Velocity generation
gen_vel         = no
```

Appendix D

Pre-processing of Databases with SD MOE's tools

```
sdsort -unique -molcmp [smiles|tautomers]
sdwash -smi input.smi -o output.smi -strict -compfield lost_frag -ylide
    -chiral -salts -wedge @$MOE/lib/sdabbrev.txt
sdfilter -smi input.smi -o output.smi -smarts '[#T]' 0 -smarts '[!D0!D1
    !D2!D3!D4]' 0 -elements C,H,N,O,S,P,F,Cl,Br,I
sdwash -smi 4_SD_1filter.smi -pH 7 -enumsize 1000 -o 5_SD_2wash.smi -
    addH
sdfilter -smi input.smi -o output.smi -nonreactive -numfield '%C' 80+
sdsort -smi input.smi -o output.smi -unique

#CONVERSION SMILE FORMAT 2 MOE DATABSE BINARY FILE
moebatch -exec "db_Open['output.mdb','create']"          #note that
    INPUT.mdb file does not exist yet
moebatch -exec "db_ImportASCII [ascii_file:'input.smi',db_file:'output.
    mdb',names:['mol','ZINCID'],types:['molecule','char'],titles:0]"
```

Appendix E

Prepare the files to run DockBox package

The configuration file *config.ini*

```
    [DOCKING]
program = dock,moe,vina
rescoring = yes
minimize = yes
cleanup = yes

    [RESCORING]
program = dock,moe,vina

    [AUTODOCK]
ga_run=20
spacing=0.3

    [VINA]
num_modes=20
cpu=1
energy_range=3

    [DOCK]
attractive_exponent=6
extra_margin=2.0
grid_spacing=0.3
maximum_sphere_radius=4.0
max_orientations=10000
minimum_sphere_radius=1.4
nposes=20
num_scored_conformers=5000
probe_radius=1.4
repulsive_exponent=12

    [MOE]
gtest=0.01
maxpose=20
placement=Proxy Triangle
placement_maxpose=100
placement_nsample=20
remaxpose=20
rescoring=GBVI/WSA dG
scoring=London dG
```

Appendix F

Preparation of Datasets for DeepDock package

The codes to automate the preparation of datasets with minimal loss in the number of ligands are too long to be reported here. Still, they are available on personal repository.

It is important to note that the automatic preparation is supposed to be placed side by side with MOE (version 2022.2) since, apart from initiating the conformational search and docking, some functions inside the codes, like the recognition of completion of chunks, find specific text strings in slurm files *.out* generated by internal MOE's tools (See the comments like *confSearch is complete* or *if docking is complete. ready to postprocess* inside the script 03_phase1_C_checkingDockingPostprocess.sh.), but they can be easily modified according to the program chosen.

Input files required:

- morgan FingerPrint (FP) files (available online);
- SMILE files (available online);
- *logs.txt* file: the file must have specific information for each line in precise order as described below:
  1. file path of main directory;
     Ex: /home/<username>/scratch/DeepDock);
  2. name project directory where everything is stored;
     Ex: *gammaTubulin* is the name of the project, the path is
     /home/<username>/scratch/DeepDock/gammaTubulin
  3. file path of morgan FP directory
     Ex: /home/<username>/scratch/DeepDock/morganFP_original
  4. file path of SMILE directory
     Ex: /home/<username>/scratch/DeepDock/smileDir
  5. number of total ligands in the datasets (i.e., M ligands for training set + M ligands for test set + M ligands for the validation set → 3*M)
- *runConfS.sh* file: it is generated by *Batch* option inside the *Conformational Search* MOE's tool and it is saved into the following path:
  /home/<username>/scratch/DeepDock/csearchMainCode
  Settings are as described in sec. 2.2.2
- *runDock.sh* e *rec.moe* files: it is generated by *Batch* option inside the *Docking* MOE's tool and saved into the following path:
  /home/<username>/scratch/DeepDock/dataXdockingStepBasedOnSite/dockSite*
  where * is the number ID of binding site under study. Settings are as described in sec. 2.2.5

Briefly, each code does the following:

1. *01_phase1_A_prep.sh*
   Given the iteration N, it returns the updated size of the entire database (if N=1, this coincides with the original database) predicted from previous iteration N-1. Then, it extracts randomly from the database of N-1 iteration (or original database if N=1) M ligands (both SMILE and morgan FP formats).
2. *02_phase1_B_preProcess_conformationalSearch*
   It splits the selected and previously generated dataset in blocks of L ligands, then converts them to the appropriate MOE database format (.mdb) and initiates the conformational search (stochastic method).
3. *03_phase1_C_checkingDockingPostprocess.sh*
   It searches any possible errors[1] among the blocks of the selected dataset and tries to fix it based on the situation, with the possibility to split the blocks further in smaller chunks or restarting[2]/resume[3] where the error occurs.
   Additionally, it checks after the completion of conformational search or docking, if the new block contains at least 98% of the ligands contained in the original block[4]. If the conformational search is complete, it starts the docking.
   If the docking is complete, it starts the postprocessing.
   For each successful block, the postprocessing compresses all data leaving out the .sdf file where best poses are saved and a .log file in which the total number of successful and missing ligands are reported.
4. *04_phase2.sh*
   It extracts the score information saved in .sdf file as a label and generates the scripts which start the training of several deep neural network models.
5. *05_phase3_evaluation.sh*
   Check the best deep neural network model based on recall, precision, AUC and number of predicted ligands in the test set to be used as a predictor of the score of ligands contained in the original database.
   Iteration N is finished. At this point, restart from point 1.
6. *06_plotTrend.sh*
   Track and plot the trend of the DeepDock's results between two selected iterations.
7. *07_phaseEND.sh*

[1]Two macro types of error exist: SLURM errors (i.e., failures/unpredicted interruptions of MPU jobs or time expiration in running the job) which are independent of the employed program, and MOE errors (ex: corruption of the original file).

[2]option available in both conformational search and docking. Useful when the cluster is operatively unstable, thus with the risk of starting over in the case of conformational search

[3]Available only in docking.

[4]For unknown reasons, sometimes MOE return a job considered completed even when the block is less than half the size of the original block.

Bibliography

[1] P. M. Cromm and C. M. Crews, «Targeted Protein Degradation: from Chemical Biology to Drug Discovery», *Cell Chemical Biology*, vol. 24, no. 9, pp. 1181–1190, Sep. 2017, ISSN: 2451-9456. DOI: 10.1016/J.CHEMBIOL.2017.05.024. [Online]. Available: http://www.cell.com/article/S2451945617301873/fulltext%20http://www.cell.com/article/S2451945617301873/abstract%20https://www.cell.com/cell-chemical-biology/abstract/S2451-9456(17)30187-3 (cit. on pp. 1, 2, 6–8, 10–12, 14).

[2] A. C. Lai and C. M. Crews, «Induced protein degradation: an emerging drug discovery paradigm», *Nature Reviews Drug Discovery 2016 16:2*, vol. 16, no. 2, pp. 101–114, Nov. 2016, ISSN: 1474-1784. DOI: 10.1038/nrd.2016.211. [Online]. Available: https://www.nature.com/articles/nrd.2016.211 (cit. on pp. 1, 2, 4, 5, 10–13).

[3] D. P. Bondeson *et al.*, «Catalytic in vivo protein knockdown by small-molecule PROTACs», *Nature Chemical Biology 2015 11:8*, vol. 11, no. 8, pp. 611–617, Jun. 2015, ISSN: 1552-4469. DOI: 10.1038/nchembio.1858. [Online]. Available: https://www.nature.com/articles/nchembio.1858 (cit. on pp. 1–3, 11, 13, 14, 17).

[4] Y. Zou, D. Ma, and Y. Wang, «The PROTAC technology in drug development», *Cell Biochemistry and Function*, vol. 37, no. 1, pp. 21–30, Jan. 2019, ISSN: 10990844. DOI: 10.1002/CBF.3369 (cit. on pp. 1, 10, 11, 14).

[5] S.-M. Qi *et al.*, «PROTAC: An Effective Targeted Protein Degradation Strategy for Cancer Therapy», *Frontiers in Pharmacology*, vol. 12, May 2021, ISSN: 1663-9812. DOI: 10.3389/fphar.2021.692574. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphar.2021.692574/full (cit. on pp. 1, 2, 4, 6, 10–14).

[6] M. J. Bond, C. M. Crews, and M. J. Bond, «Proteolysis targeting chimeras (PROTACs) come of age : entering the third decade of targeted protein degradation», pp. 725–742, 2021. DOI: 10.1039/d1cb00011j (cit. on pp. 2, 3, 11, 12, 16, 17).

[7] C. M. Crews, «Inducing Protein Degradation as a Therapeutic Strategy», *Journal of Medicinal Chemistry*, vol. 61, no. 2, pp. 403–404, Jan. 2018, ISSN: 15204804. DOI: 10.1021/ACS.JMEDCHEM.7B01333 (cit. on p. 2).

[8] T. Buchanan, «Arvinas, Pfizer Team Up on PROTACs», *Cancer Discovery*, vol. 8, no. 4, pp. 377–378, Apr. 2018, ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-NB2018-015. [Online]. Available: https://cancerdiscovery.aacrjournals.org/content/8/4/377.2%20https://cancerdiscovery.aacrjournals.org/content/8/4/377.2.abstract (cit. on p. 2).

[9] A. Bouchie, M. Allison, S. Webb, and L. Defrancesco, «Nature Biotechnology's academic spinouts of 2013», *Nature Biotechnology 2014 32:3*, vol. 32, no. 3, pp. 229–238, Mar. 2014, ISSN: 1546-1696. DOI: 10.1038/nbt.2846. [Online]. Available: https://www.nature.com/articles/nbt.2846 (cit. on p. 2).

[10] H. Xie, J. Liu, D. M. A. Glison, and J. B. Fleming, «The clinical advances of proteolysis targeting chimeras in oncology», *Exploration of Targeted Anti-tumor Therapy*, vol. 2, no. 6, pp. 511–521, Dec. 2021, ISSN: 2692-3114. DOI: 10.37349/ETAT.2021.00061. [Online]. Available: https://www.explorationpub.com/Journals/etat/Article/100261 (cit. on p. 2).

[11] K. M. Sakamoto *et al.*, «Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation», *Proceedings of the National Academy of Sciences*, vol. 98, no. 15, pp. 8554–8559, Jul. 2001, ISSN: 0027-8424. DOI: 10.1073/PNAS.141230798. [Online]. Available: https://www.pnas.org/content/98/15/8554%20https://www.pnas.org/content/98/15/8554.abstract (cit. on pp. 2, 10).

[12] T. Ishida and A. Ciulli, «E3 Ligase Ligands for PROTACs: How They Were Found and How to Discover New Ones», *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, vol. 26, no. 4, pp. 484–502, Apr. 2021, ISSN: 2472-5552. DOI: 10.1177/2472555220965528. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2472555220965528 (cit. on pp. 2, 10, 11, 13, 14, 16).

[13] I. Gasic *et al.*, «Tubulin Resists Degradation by Cereblon-Recruiting PROTACs», *Cells*, vol. 9, no. 5, p. 1083, Apr. 2020, ISSN: 2073-4409. DOI: 10.3390/cells9051083. [Online]. Available: https://www.mdpi.com/2073-4409/9/5/1083/htm%20https://www.mdpi.com/2073-4409/9/5/1083 (cit. on pp. 2, 14, 15, 17).

[14] R. I. Troup, C. Fallan, and M. G. J. Baud, «Current strategies for the design of PROTAC linkers: a critical review», *Exploration of Targeted Anti-tumor Therapy*, vol. 1, no. 5, pp. 273–312, Oct. 2020, ISSN: 2692-3114. DOI: 10.37349/etat.2020.00018. [Online]. Available: https://www.explorationpub.com/Journals/etat/Article/100218 (cit. on pp. 2, 3, 13, 17).

[15] C. V. Dang, E. P. Reddy, K. M. Shokat, and L. Soucek, «Drugging the 'undruggable' cancer targets», *Nature Reviews Cancer 2017 17:8*, vol. 17, no. 8, pp. 502–508, Jun. 2017, ISSN: 1474-1768. DOI: 10.1038/NRC.2017.36. [Online]. Available: https://www-nature-com.ezproxy.biblio.polito.it/articles/nrc.2017.36 (cit. on pp. 3, 11).

[16] S. D. Edmondson, B. Yang, and C. Fallan, «Proteolysis targeting chimeras (PROTACs)in 'beyond rule-of-five' chemical space: Recent progress and future challenges», *Bioorganic and Medicinal Chemistry Letters*, vol. 29, no. 13, pp. 1555–1564, 2019, ISSN: 14643405. DOI: 10.1016/j.bmcl.2019.04.030. [Online]. Available: https://doi.org/10.1016/j.bmcl.2019.04.030 (cit. on pp. 3, 11).

[17] G. Ermondi *et al.*, «PROTACs and Building Blocks: The 2D Chemical Space in Very Early Drug Discovery», *Molecules 2021, Vol. 26, Page 672*, vol. 26, no. 3, p. 672, Jan. 2021, ISSN: 14203049. DOI: 10.3390/MOLECULES26030672. [Online]. Available: https://www.mdpi.com/1420-3049/26/3/672/htm%20https://www.mdpi.com/1420-3049/26/3/672 (cit. on pp. 3, 11).

[18] L. Goitre, E. Trapani, L. Trabalzini, and S. F. Retta, «The Ras Superfamily of Small GTPases: The Unlocked Secrets», in 2014, pp. 1–18. DOI: 10.1007/978-1-62703-791-4_1. [Online]. Available: http://link.springer.com/10.1007/978-1-62703-791-4_1 (cit. on p. 3).

[19] Y. Takai, T. Sasaki, and T. Matozaki, «Small GTP-Binding Proteins», *Physiological Reviews*, vol. 81, no. 1, pp. 153–208, Jan. 2001, ISSN: 0031-9333. DOI: 10.1152/physrev.2001.81.1.153. [Online]. Available: https://www.physiology.org/doi/10.1152/physrev.2001.81.1.153 (cit. on p. 3).

[20] D. Nandi, P. Tahiliani, A. Kumar, and D. Chandu, «The ubiquitin-proteasome system», *Journal of Biosciences 2006 31:1*, vol. 31, no. 1, pp. 137–155, Mar. 2006, ISSN: 0973-7138. DOI: 10.1007/BF02705243. [Online]. Available: https://link.springer.com/article/10.1007/BF02705243 (cit. on pp. 4, 6).

[21] D. Finley, «Recognition and Processing of Ubiquitin-Protein Conjugates by the Proteasome», *http://dx.doi.org/10.1146/annurev.biochem.78.081507.101607*, vol. 78, pp. 477–513, Jun. 2009, ISSN: 00664154. DOI: 10.1146/ANNUREV.BIOCHEM.78.081507.101607. [Online]. Available: https://www.annualreviews.org/doi/abs/10.1146/annurev.biochem.78.081507.101607 (cit. on pp. 4–6).

[22] S. Alessandra, «Progettazione , sintesi ed attività di inibitori del Proteasoma a base peptidica», 2014. [Online]. Available: https://iris.unife.it/retrieve/handle/11392/2389430/123560/713.pdf (cit. on pp. 4–6).

[23] C. E. Berndsen and C. Wolberger, «New insights into ubiquitin E3 ligase mechanism», *Nature Structural & Molecular Biology 2014 21:4*, vol. 21, no. 4, pp. 301–307, Apr. 2014, ISSN: 1545-9985. DOI: 10.1038/nsmb.2780. [Online]. Available: https://www.nature.com/articles/nsmb.2780 (cit. on pp. 4, 6, 7, 18).

[24] P. Bhattacharjee, M. Mazumdar, D. Guha, and G. Sa, «Ubiquitin-proteasome system in the hallmarks of cancer», in *Role of Proteases in Cellular Dysfunction*, Springer, New York, NY, Jan. 2014, pp. 159–186, ISBN: 9781461490999. DOI: 10.1007/978-1-4614-9099-9_9. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4614-9099-9_9 (cit. on p. 4).

[25] B. A. Schulman and J. Wade Harper, «Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways», *Nature Reviews Molecular Cell Biology 2009 10:5*, vol. 10, no. 5, pp. 319–331, Apr. 2009, ISSN: 1471-0080. DOI: 10.1038/nrm2673. [Online]. Available: https://www.nature.com/articles/nrm2673 (cit. on p. 4).

[26] P. K. Jackson *et al.*, «The lore of the RINGs: Substrate recognition and catalysis by ubiquitin ligases», *Trends in Cell Biology*, vol. 10, no. 10, pp. 429–439, 2000, ISSN: 09628924. DOI: 10.1016/S0962-8924(00)01834-1 (cit. on pp. 4, 7, 8).

[27] S. Fulda, K. Rajalingam, and I. Dikic, «Ubiquitylation in immune disorders and cancer: from molecular mechanisms to therapeutic implications», *EMBO Molecular Medicine*, vol. 4, no. 7, pp. 545–556, Jul. 2012, ISSN: 1757-4684. DOI: 10.1002/EMMM.201100707. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/emmm.201100707%20https://onlinelibrary.wiley.com/doi/abs/10.1002/emmm.201100707%20https://www.embopress.org/doi/10.1002/emmm.201100707 (cit. on pp. 4, 5, 8).

[28] F. Ikeda and I. Dikic, *Atypical ubiquitin chains: New molecular signals. 'Protein Modifications: Beyond the Usual Suspects' Review Series*, Jun. 2008. DOI: 10.1038/embor.2008.93. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1038/embor.2008.93%20https://onlinelibrary.wiley.com/doi/abs/10.1038/embor.2008.93%20https://www.embopress.org/doi/abs/10.1038/embor.2008.93 (cit. on pp. 4, 5).

[29] M. Miranda and A. Sorkin, «Regulation of Receptors and Transporters by Ubiquitination: New Insights into Surprisingly Similar Mechanisms», *Molecular Interventions*, vol. 7, no. 3, p. 157, Jun. 2007, ISSN: 1534-0384. DOI: 10.1124/MI.7.3.7. [Online]. Available: http://triggered.clockss.org/ServeContent?url=http%5C%3A%5C%2F%5C%2Fmolinterv.aspetjournals.org%5C%2Fcontent%5C%2F7%5C%2F3%5C%2F157.full%20http://triggered.clockss.org/ServeContent?url=http%5C%3A%5C%2F%5C%2Fmolinterv.aspetjournals.org%5C%2Fcontent%5C%2F7%5C%2F3%5C%2F157.abstract (cit. on pp. 4, 5).

[30] D. Komander, «The emerging complexity of protein ubiquitination», *Biochemical Society Transactions*, vol. 37, no. 5, pp. 937–953, Oct. 2009, ISSN: 0300-5127. DOI: 10.1042/BST0370937. [Online]. Available: https://portlandpress.com/biochemsoctrans/article/37/5/937/64415/The-emerging-complexity-of-protein-ubiquitination (cit. on pp. 4, 5).

[31] D. Komander and M. Rape, «The Ubiquitin Code», *Annual Review of Biochemistry*, vol. 81, no. 1, pp. 203–229, Jul. 2012, ISSN: 0066-4154. DOI: 10.1146/annurev-biochem-060310-170328. [Online]. Available: https://www.annualreviews.org/doi/abs/10.1146/annurev-biochem-060310-170328 (cit. on pp. 4, 5).

[32] Y. Ye and M. Rape, «Building ubiquitin chains: E2 enzymes at work», *Nature Reviews Molecular Cell Biology 2009 10:11*, vol. 10, no. 11, pp. 755–764, Nov. 2009, ISSN: 1471-0080. DOI: 10.1038/nrm2780. [Online]. Available: https://www.nature.com/articles/nrm2780 (cit. on p. 4).

[33] A. Dósa and T. Csizmadia, «The role of K63-linked polyubiquitin in several types of autophagy», *Biologia Futura*, vol. 73, no. 2, pp. 137–148, Jun. 2022, ISSN: 26768607. DOI: 10.1007/S42977-022-00117-4/FIGURES/4. [Online]. Available: https://link.springer.com/article/10.1007/s42977-022-00117-4 (cit. on pp. 5, 6, 8).

[34] A. Varshavsky, «N-degron and C-degron pathways of protein degradation», *Proceedings of the National Academy of Sciences*, vol. 116, no. 2, pp. 358–366, Jan. 2019, ISSN: 0027-8424. DOI: 10.1073/pnas.1816596116. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1816596116 (cit. on pp. 5, 6, 20).

[35] A. Varshavsky, *The N-end rule pathway and regulation by proteolysis*, Aug. 2011. DOI: 10.1002/pro.666 (cit. on pp. 7, 20).

[36] C. N. Okoye, P. J. E. Rowling, L. S. Itzhaki, and C. Lindon, «Counting Degrons: Lessons From Multivalent Substrates for Targeted Protein Degradation», *Frontiers in Physiology*, vol. 13, p. 1292, Jul. 2022, ISSN: 1664-042X. DOI: 10.3389/fphys.2022.913063. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2022.913063/full (cit. on p. 7).

[37] C. Hou *et al.*, «Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning», *BMC Biology*, vol. 20, no. 1, p. 162, Dec. 2022, ISSN: 1741-7007. DOI: 10.1186/s12915-022-01364-6. [Online]. Available: https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-022-01364-6 (cit. on pp. 7, 12, 100).

[38] K. K. Dove and R. E. Klevit, «RING-Between-RING E3 Ligases: Emerging Themes amid the Variations», *Journal of Molecular Biology*, vol. 429, no. 22, pp. 3363–3375, 2017, ISSN: 10898638. DOI: 10.1016/j.jmb.2017.08.008. [Online]. Available: https://doi.org/10.1016/j.jmb.2017.08.008 (cit. on pp. 7, 8).

[39] S. Lipkowitz and A. M. Weissman, «RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis», *Nature Reviews Cancer 2011 11:9*, vol. 11, no. 9, pp. 629–643, Aug. 2011, ISSN: 1474-1768. DOI: 10.1038/nrc3120. [Online]. Available: https://www.nature.com/articles/nrc3120 (cit. on p. 7).

[40] J. Bitinaite, D. A. Wah, A. K. Aggarwal, and I. Schildkraut, «FokI dimerization is required for DNA cleavage», *Proceedings of the National Academy of Sciences*, vol. 95, no. 18, pp. 10 570–10 575, Sep. 1998, ISSN: 0027-8424. DOI: 10.1073/pnas.95.18.10570. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.95.18.10570 (cit. on p. 7).

[41] L. Semmler, C. Reiter-Brennan, and A. Klein, «BRCA1 and Breast Cancer: a Review of the Underlying Mechanisms Resulting in the Tissue-Specific Tumorigenesis in Mutation Carriers», *Journal of Breast Cancer*, vol. 22, no. 1, pp. 1–14, Mar. 2019, ISSN: 1738-6756. DOI: 10.4048/JBC.2019.22.E6. [Online]. Available: https://doi.org/10.4048/jbc.2019.22.e6 (cit. on pp. 7, 30).

[42] S. R. Witus, M. D. Stewart, and R. E. Klevit, «The BRCA1/BARD1 ubiquitin ligase and its substrates», *The Biochemical journal*, vol. 478, no. 18, pp. 3467–3483, Sep. 2021, ISSN: 14708728. DOI: 10.1042/BCJ20200864/229865. [Online]. Available: /biochemj/article/doi/10.1042/BCJ20200864/229865/The-BRCA1-BARD1-ubiquitin-ligase-and-its%20Https://portlandpress.com/biochemj/article/doi/10.1042/BCJ20200864/229865/The-BRCA1-BARD1-ubiquitin-ligase-and-its (cit. on pp. 7, 30).

[43] J. N. Pruneda *et al.*, «Structure of an E3:E2 Ub Complex Reveals an Allosteric Mechanism Shared among RING/U-box Ligases», *Molecular Cell*, vol. 47, no. 6, pp. 933–942, Sep. 2012, ISSN: 1097-2765. DOI: 10.1016/J.MOLCEL.2012.07.001 (cit. on p. 7).

[44] J. D. Wright, P. D. Mace, and C. L. Day, «Noncovalent Ubiquitin Interactions Regulate the Catalytic Activity of Ubiquitin Writers», *Trends in Biochemical Sciences*, vol. 41, no. 11, pp. 924–937, Nov. 2016, ISSN: 09680004. DOI: 10.1016/j.tibs.2016.08.003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968000416301037 (cit. on p. 8).

[45] A. Mishra and N. R. Jana, «Regulation of turnover of tumor suppressor p53 and cell growth by E6-AP, a ubiquitin protein ligase mutated in Angelman mental retardation syndrome», *Cellular and molecular life sciences : CMLS*, vol. 65, no. 4, pp. 656–666, Feb. 2008, ISSN: 1420-682X. DOI: 10.1007/S00018-007-7476-1. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18193166/ (cit. on p. 8).

[46] I. Marín, J. I. Lucas, A. C. Gradilla, and A. Ferrús, «Parkin and relatives: the RBR family of ubiquitin ligases.», *Physiological Genomics*, vol. 17, no. 3, pp. 253–263, May 2004, ISSN: 1094-8341. DOI: 10.1152/PHYSIOLGENOMICS.00226.2003. [Online]. Available: http://europepmc.org/article/MED/15152079 (cit. on p. 8).

[47] C. Garcia-Barcena, N. Osinalde, J. Ramirez, and U. Mayor, «How to Inactivate Human Ubiquitin E3 Ligases by Mutation», *Frontiers in Cell and Developmental Biology*, vol. 8, p. 39, Feb. 2020, ISSN: 2296634X. DOI: 10.3389/FCELL.2020.00039/BIBTEX (cit. on p. 8).

[48] S. Sankaran, L. M. Starita, A. M. Simons, and J. D. Parvin, «Identification of domains of BRCA1 critical for the ubiquitin-dependent inhibition of centrosome function», *Cancer Research*, vol. 66, no. 8, pp. 4100–4107, 2006, ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-05-4430 (cit. on pp. 8, 30, 35).

[49] L. M. Starita *et al.*, «BRCA1-Dependent Ubiquitination of γ-Tubulin Regulates Centrosome Number», *Molecular and Cellular Biology*, vol. 24, no. 19, pp. 8457–8466, 2004, ISSN: 0270-7306. DOI: 10.1128/mcb.24.19.8457-8466.2004 (cit. on pp. 8, 30).

[50] S. J. Goldenberg *et al.*, «Structure of the Cand1-Cul1-Roc1 Complex Reveals Regulatory Mechanisms for the Assembly of the Multisubunit Cullin-Dependent Ubiquitin Ligases», *Cell*, vol. 119, no. 4, pp. 517–528, Nov. 2004, ISSN: 00928674. DOI: 10.1016/j.cell.2004.10.019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867404009535 (cit. on pp. 8, 9).

[51] S. Gu *et al.*, «PROTACs: An Emerging Targeting Technique for Protein Degradation in Drug Discovery», *BioEssays*, vol. 40, no. 4, p. 1 700 247, Apr. 2018, ISSN: 02659247. DOI: 10.1002/bies.201700247. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/bies.201700247 (cit. on pp. 10, 11, 13).

[52] E. Bulatov and A. Ciulli, «Targeting Cullin–RING E3 ubiquitin ligases for drug discovery: structure, assembly and small-molecule modulation», *Biochemical Journal*, vol. 467, no. 3, pp. 365–386, May 2015, ISSN: 0264-6021. DOI: 10.1042/BJ20141450. [Online]. Available: /biochemj/article/467/3/365/48501/Targeting-Cullin-RING-E3-ubiquitin-ligases-for%20https://portlandpress.com/biochemj/article/467/3/365/48501/Targeting-Cullin-RING-E3-ubiquitin-ligases-for (cit. on pp. 10, 13).

[53] J. S. Schneekloth *et al.*, «Chemical Genetic Control of Protein Levels: Selective in Vivo Targeted Degradation», *Journal of the American Chemical Society*, vol. 126, no. 12, pp. 3748–3754, Mar. 2004, ISSN: 00027863. DOI: 10.1021/JA039025Z (cit. on p. 10).

[54] J. H. Min *et al.*, «Structure of an HIF-1α-pVHL complex: Hydroxyproline recognition in signaling», *Science*, vol. 296, no. 5574, pp. 1886–1889, Jun. 2002, ISSN: 00368075. DOI: 10.1126/SCIENCE.1073440/SUPPL_FILE/1073440S4_THUMB.GIF. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1073440 (cit. on p. 10).

[55] W. C. Hon *et al.*, «Structural basis for the recognition of hydroxyproline in HIF-1α by pVHL», *Nature*, vol. 417, no. 6892, pp. 975–978, Jun. 2002, ISSN: 00280836. DOI: 10.1038/nature00767. [Online]. Available: https://www.nature.com/articles/nature00767 (cit. on p. 10).

[56] C. E. Stebbins, W. G. Kaelin, and N. P. Pavletich, «Structure of the VHL-elonginC-elonginB complex: Implications for VHL tumor suppressor function», *Science*, vol. 284, no. 5413, pp. 455–461, Apr. 1999, ISSN: 00368075. DOI: 10.1126/SCIENCE.284.5413.455/ASSET/4E4114D4-CF4A-4559-A369-A971949FB21C/ASSETS/GRAPHIC/SE1497425006.JPEG. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.284.5413.455 (cit. on p. 10).

[57] T. A. Cardote, M. S. Gadd, and A. Ciulli, «Crystal Structure of the Cul2-Rbx1-EloBC-VHL Ubiquitin Ligase Complex», *Structure*, vol. 25, no. 6, 901–911.e3, Jun. 2017, ISSN: 18784186. DOI: 10.1016/J.STR.2017.04.009/ATTACHMENT/174D96BF-6C26-4B07-939A-0CD49492D531/MMC1.PDF. [Online]. Available: http://www.cell.com/article/S0969212617301272/fulltext%20http://www.cell.com/article/S0969212617301272/abstract%20https://www.cell.com/structure/abstract/S0969-2126(17)30127-2 (cit. on pp. 10, 13).

[58] K. M. Sakamoto *et al.*, «Development of Protacs to target cancer-promoting proteins for ubiquitination and degradation», *Molecular & cellular proteomics : MCP*, vol. 2, no. 12, pp. 1350–1358, 2003, ISSN: 1535-9476. DOI: 10.1074/MCP.T300009-MCP200. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/14525958/ (cit. on p. 10).

[59] K. Oh and G.-S. Yi, «Prediction of scaffold proteins based on protein interaction and domain architectures», *BMC Bioinformatics*, vol. 17, no. S6, p. 220, Jul. 2016, ISSN: 1471-2105. DOI: 10.1186/s12859-016-1079-5. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1079-5 (cit. on p. 11).

[60] J. Hu *et al.*, «Systematic Prediction of Scaffold Proteins Reveals New Design Principles in Scaffold-Mediated Signal Transduction», *PLOS Computational Biology*, vol. 11, no. 9, A. Rzhetsky, Ed., e1004508, Sep. 2015, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004508. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1004508 (cit. on p. 11).

[61] S. Lu *et al.*, «Drugging Ras GTPase: a comprehensive mechanistic and signaling structural view», *Chemical Society Reviews*, vol. 45, no. 18, pp. 4929–4952, 2016, ISSN: 0306-0012. DOI: 10.1039/C5CS00911A. [Online]. Available: http://xlink.rsc.org/?DOI=C5CS00911A (cit. on p. 12).

[62] Z. Li *et al.*, «UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase–substrate interactions», *Database*, vol. 2021, Mar. 2021, ISSN: 1758-0463. DOI: 10.1093/database/baab010. [Online]. Available: https://academic.oup.com/database/article/doi/10.1093/database/baab010/6162640 (cit. on p. 12).

[63] Y. Li *et al.*, «An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network», *Nature Communications*, vol. 8, no. 1, p. 347, Dec. 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-00299-9. [Online]. Available: http://www.nature.com/articles/s41467-017-00299-9 (cit. on p. 12).

[64] D. L. Buckley *et al.*, «Small-Molecule Inhibitors of the Interaction between the E3 Ligase VHL and HIF1α», *Angewandte Chemie International Edition*, vol. 51, no. 46, pp. 11 463–11 467, Nov. 2012, ISSN: 1521-3773. DOI: 10.1002/ANIE.201206231. [Online]. Available: https://onlinelibrary-wiley-com.ezproxy.biblio.polito.it/doi/full/10.1002/anie.201206231%20https://onlinelibrary-wiley-com.ezproxy.biblio.polito.it/doi/abs/10.1002/anie.201206231%20https://onlinelibrary-wiley-com.ezproxy.biblio.polito.it/doi/10.1002/anie.2012 (cit. on pp. 12, 18).

[65] P. Soares *et al.*, «Group-Based Optimization of Potent and Cell-Active Inhibitors of the von Hippel-Lindau (VHL) E3 Ubiquitin Ligase: Structure-Activity Relationships Leading to the Chemical Probe (2S,4R)-1-((S)-2-(1-Cyanocyclopropanecarboxamido)-3,3-dimethylbutanoyl)-4-hydr», *Journal of Medicinal Chemistry*, vol. 61, no. 2, pp. 599–618, Jan. 2018, ISSN: 15204804. DOI: 10.1021/ACS.JMEDCHEM.7B00675/SUPPL_FILE/JM7B00675_SI_001.PDF. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.7b00675 (cit. on p. 12).

[66] C. Galdeano *et al.*, «Structure-guided design and optimization of small molecules targeting the protein-protein interaction between the von hippel-lindau (VHL) E3 ubiquitin ligase and the hypoxia inducible factor (HIF) alpha subunit with in vitro nanomolar affinities», *Journal of Medicinal Chemistry*, vol. 57, no. 20, pp. 8657–8663, Oct. 2014, ISSN: 15204804. DOI: 10.1021/JM5011258/SUPPL_FILE/JM5011258_SI_001.PDF. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/jm5011258 (cit. on p. 12).

[67] J. Frost *et al.*, «Potent and selective chemical probe of hypoxic signalling downstream of HIF-α hydroxylation via VHL inhibition», *Nature Communications 2016 7:1*, vol. 7, no. 1, pp. 1–12, Nov. 2016, ISSN: 2041-1723. DOI: 10.1038/ncomms13312. [Online]. Available: https://www.nature.com/articles/ncomms13312 (cit. on p. 12).

[68] M. Zengerle, K. H. Chan, and A. Ciulli, «Selective Small Molecule Induced Degradation of the BET Bromodomain Protein BRD4», *ACS Chemical Biology*, vol. 10, no. 8, pp. 1770–1777, Aug. 2015, ISSN: 15548937. DOI: 10.1021/ACSCHEMBIO.5B00216/SUPPL_FILE/CB5B00216_SI_003.AVI. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acschembio.5b00216 (cit. on p. 13).

[69] M. S. Gadd *et al.*, «Structural basis of PROTAC cooperative recognition for selective protein degradation», *Nature Chemical Biology 2017 13:5*, vol. 13, no. 5, pp. 514–521, Mar. 2017, ISSN: 1552-4469. DOI: 10.1038/nchembio.2329. [Online]. Available: https://www.nature.com/articles/nchembio.2329 (cit. on pp. 13, 18).

[70] T. W. Kowalski *et al.*, «CRL4-Cereblon complex in Thalidomide Embryopathy: a translational investigation», *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–13, Jan. 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-57512-x. [Online]. Available: https://www.nature.com/articles/s41598-020-57512-x (cit. on p. 13).

[71] M. E. Matyskiela *et al.*, «Crystal structure of the SALL4–pomalidomide–cereblon–DDB1 complex», *Nature Structural & Molecular Biology 2020 27:4*, vol. 27, no. 4, pp. 319–322, Apr. 2020, ISSN: 1545-9985. DOI: 10.1038/s41594-020-0405-9. [Online]. Available: https://www.nature.com/articles/s41594-020-0405-9 (cit. on p. 13).

[72] J. Lu *et al.*, «Hijacking the E3 Ubiquitin Ligase Cereblon to Efficiently Target BRD4», *Chemistry & Biology*, vol. 22, no. 6, pp. 755–763, Jun. 2015, ISSN: 1074-5521. DOI: 10.1016/J.CHEMBIOL.2015.05.009 (cit. on pp. 14, 18).

[73] P. Binarová and J. Tuszynski, «Tubulin: Structure, Functions and Roles in Disease», *Cells 2019, Vol. 8, Page 1294*, vol. 8, no. 10, p. 1294, Oct. 2019, ISSN: 2073-4409. DOI: 10.3390/CELLS8101294. [Online]. Available: https://www.mdpi.com/2073-4409/8/10/1294/htm%20https://www.mdpi.com/2073-4409/8/10/1294 (cit. on pp. 14, 28, 54).

[74] A. Roll-Mecak, «The Tubulin Code in Microtubule Dynamics and Information Encoding», *Developmental Cell*, vol. 54, no. 1, pp. 7–20, Jul. 2020, ISSN: 1534-5807. DOI: 10.1016/J.DEVCEL.2020.06.008. [Online]. Available: http://www.cell.com/article/S1534580720304585/fulltext%20http://www.cell.com/article/S1534580720304585/abstract%20https://www.cell.com/developmental-cell/abstract/S1534-5807(20)30458-5 (cit. on pp. 14, 28).

[75] M. A. Kristensson, «The Game of Tubulins», *Cells*, vol. 10, no. 4, p. 745, Mar. 2021, ISSN: 2073-4409. DOI: 10.3390/cells10040745. [Online]. Available: https://www.mdpi.com/2073-4409/10/4/745 (cit. on p. 14).

[76] J. A. Pradeepkiran and P. H. Reddy, «Phosphorylated tau targeted small-molecule PROTACs for the treatment of Alzheimer's disease and tauopathies», *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1867, no. 8, p. 166 162, Aug. 2021, ISSN: 09254439. DOI: 10.1016/j.bbadis.2021.166162. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925443921000958 (cit. on p. 15).

[77] K. Iqbal, F. Liu, and C.-X. Gong, «Tau and neurodegenerative disease: the story so far», *Nature Reviews Neurology*, vol. 12, no. 1, pp. 15–27, Jan. 2016, ISSN: 1759-4758. DOI: 10.1038/nrneurol.2015.225. [Online]. Available: http://www.nature.com/articles/nrneurol.2015.225 (cit. on p. 15).

[78] A. Bricelj *et al.*, «E3 Ligase Ligands in Successful PROTACs: An Overview of Syntheses and Linker Attachment Points», *Frontiers in Chemistry*, vol. 9, Jul. 2021, ISSN: 2296-2646. DOI: 10.3389/fchem.2021.707317. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fchem.2021.707317/full (cit. on pp. 16, 17).

[79] X. Zhou *et al.*, «Recurrence patterns in patients with high-grade glioma following temozolomide-based chemoradiotherapy», *Molecular and Clinical Oncology*, vol. 5, no. 2, pp. 289–294, Aug. 2016, ISSN: 2049-9450. DOI: 10.3892/MCO.2016.936. [Online]. Available: http://www.spandidos-publications.com/10.3892/mco.2016.936/abstract%20https://www.spandidos-publications.com/10.3892/mco.2016.936 (cit. on p. 16).

[80] N. Bai *et al.*, «Rationalizing PROTAC-Mediated Ternary Complex Formation Using Rosetta», *J. Chem. Inf. Model*, vol. 61, p. 2021, 2021. DOI: 10.1021/acs.jcim.0c01451. [Online]. Available: https://dx.doi.org/10.1021/acs.jcim.0c01451 (cit. on p. 16).

[81] M. L. Drummond and C. I. Williams, «In Silico Modeling of PROTAC-Mediated Ternary Complexes: Validation and Application», *Journal of Chemical Information and Modeling*, vol. 59, no. 4, pp. 1634–1644, Apr. 2019, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00872. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.8b00872 (cit. on p. 16).

[82] M. L. Drummond, A. Henry, H. Li, and C. I. Williams, «Improved Accuracy for Modeling PROTAC-Mediated Ternary Complex Formation and Targeted Protein Degradation via New In Silico Methodologies», *Journal of Chemical Information and Modeling*, vol. 60, no. 10, pp. 5234–5254, Oct. 2020, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00897. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.0c00897 (cit. on p. 16).

[83] J. Liao *et al.*, «In Silico Modeling and Scoring of PROTAC-Mediated Ternary Complex Poses», *Journal of Medicinal Chemistry*, vol. 65, no. 8, pp. 6116–6132, Apr. 2022, ISSN: 15204804. DOI: 10.1021/acs.jmedchem.1c02155. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jmedchem.1c02155 (cit. on p. 16).

[84] B. E. Smith *et al.*, «Differential PROTAC substrate specificity dictated by orientation of recruited E3 ligase», *Nature Communications*, vol. 10, no. 1, p. 131, Dec. 2019, ISSN: 2041-1723. DOI: 10.1038/s41467-018-08027-7. [Online]. Available: http://www.nature.com/articles/s41467-018-08027-7 (cit. on p. 17).

[85] G. Weng *et al.*, «PROTAC-DB: an online database of PROTACs», *Nucleic Acids Research*, vol. 49, no. D1, pp. D1381–D1387, Jan. 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa807. [Online]. Available: https://academic.oup.com/nar/article/49/D1/D1381/5917660 (cit. on p. 17).

[86] K. Saurabh *et al.*, «UBR-box containing protein, UBR5, is over-expressed in human lung adenocarcinoma and is a potential therapeutic target», *BMC Cancer*, vol. 20, no. 1, pp. 1–12, Aug. 2020, ISSN: 14712407. DOI: 10.1186/S12885-020-07322-1/FIGURES/5. [Online]. Available: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-07322-1 (cit. on p. 19).

[87] J. Muñoz-Escobar *et al.*, «Bound Waters Mediate Binding of Diverse Substrates to a Ubiquitin Ligase», *Structure*, vol. 25, no. 5, 719–729.e3, May 2017, ISSN: 09692126. DOI: 10.1016/j.str.2017.03.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0969212617300643 (cit. on p. 20).

[88] M. Pan *et al.*, «Structural insights into Ubr1-mediated N-degron polyubiquitination», *Nature*, vol. 600, no. 7888, pp. 334–338, Nov. 2021, ISSN: 14764687. DOI: 10.1038/s41586-021-04097-8. [Online]. Available: https://www.nature.com/articles/s41586-021-04097-8 (cit. on p. 20).

[89] A. Shemorry, C.-S. Hwang, and A. Varshavsky, «Control of Protein Quality and Stoichiometries by N-Terminal Acetylation and the N-End Rule Pathway», *Molecular Cell*, vol. 50, no. 4, pp. 540–551, May 2013, ISSN: 10972765. DOI: 10.1016/j.molcel.2013.03.018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1097276513002220 (cit. on p. 20).

[90] A. Chatr-aryamontri, A. van der Sloot, and M. Tyers, «At Long Last, a C-Terminal Bookend for the Ubiquitin Code», *Molecular Cell*, vol. 70, no. 4, pp. 568–571, May 2018, ISSN: 10972765. DOI: 10.1016/j.molcel.2018.05.006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1097276518303563 (cit. on p. 20).

[91] P. Dráber and E. Dráberová, «Dysregulation of Microtubule Nucleating Proteins in Cancer Cells», *Cancers*, vol. 13, no. 22, p. 5638, Nov. 2021, ISSN: 2072-6694. DOI: 10.3390/cancers13225638. [Online]. Available: https://www.mdpi.com/2072-6694/13/22/5638/htm%20https://www.mdpi.com/2072-6694/13/22/5638 (cit. on pp. 23, 24, 29, 35, 37, 38).

[92] L. Lindström and M. Alvarado-Kristensson, «Characterization of gamma-tubulin filaments in mammalian cells», *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1865, no. 1, pp. 158–171, Jan. 2018, ISSN: 01674889. DOI: 10.1016/j.bbamcr.2017.10.008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167488917302835 (cit. on pp. 23, 24, 26, 27).

[93] V. Caracciolo *et al.*, «Differential expression and cellular distribution of γ-tubulin and βIII-tubulin in medulloblastomas and human medulloblastoma cell lines», *Journal of Cellular Physiology*, vol. 223, no. 2, pp. 519–529, May 2010, ISSN: 00219541. DOI: 10.1002/jcp.22077 (cit. on pp. 23, 29, 34).

[94] T. Ohashi, T. Yamamoto, Y. Yamanashi, and M. Ohsugi, «Human TUBG2 gene is expressed as two splice variant mRNA and involved in cell growth», *FEBS Letters*, vol. 590, no. 8, pp. 1053–1063, Apr. 2016, ISSN: 18733468. DOI: 10.1002/1873-3468.12163. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/1873-3468.12163 (cit. on p. 23).

[95] E. L. Ivanova *et al.*, «TUBG1 missense variants underlying cortical malformations disrupt neuronal locomotion and microtubule dynamics but not neurogenesis», *Nature Communications*, vol. 10, no. 1, pp. 1–18, Dec. 2019, ISSN: 20411723. DOI: 10.1038/s41467-019-10081-8 (cit. on pp. 23, 25, 33, 34).

[96] D. E. Friesen *et al.*, «Discovery of Small Molecule Inhibitors that Interact with γ-Tubulin», *Chemical Biology and Drug Design*, vol. 79, no. 5, pp. 639–652, 2012, ISSN: 17470277. DOI: 10.1111/j.1747-0285.2012.01340.x (cit. on pp. 23, 24, 28, 31, 32, 34).

[97] C. A. Rosselló *et al.*, *γ-Tubulin–γ-Tubulin Interactions as the Basis for the Formation of a Meshwork*, Oct. 2018. DOI: 10.3390/ijms19103245 (cit. on pp. 23, 25, 26, 29).

[98] S. Sankaran *et al.*, «Centrosomal Microtubule Nucleation Activity Is Inhibited by BRCA1-Dependent Ubiquitination», *Molecular and Cellular Biology*, vol. 25, no. 19, pp. 8656–8668, 2005, ISSN: 0270-7306. DOI: 10.1128/mcb.25.19.8656-8668.2005 (cit. on pp. 24, 29, 30, 33, 35).

[99] M. Moudjou, N. Bordes, M. Paintrand, and M. Bornens, «gamma-Tubulin in mammalian cells: the centrosomal and the cytosolic forms», *Journal of Cell Science*, vol. 109, no. 4, pp. 875–887, Apr. 1996, ISSN: 0021-9533. DOI: 10.1242/JCS.109.4.875. [Online]. Available: https://journals.biologists.com/jcs/article/109/4/875/24800/gamma-Tubulin-in-mammalian-cells-the-centrosomal (cit. on p. 24).

[100] M. Wieczorek *et al.*, «Asymmetric Molecular Architecture of the Human γ-Tubulin Ring Complex», *Cell*, vol. 180, no. 1, 165–175.e16, Jan. 2020, ISSN: 00928674. DOI: 10.1016/j.cell.2019.12.007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867419313698%20https://doi.org/10.1016/j.cell.2019.12.007%20https://reader.elsevier.com/reader/sd/pii/S0092867419313698?token=34C8D542873B36DE0BE3936917A817D7F7B8B8193F10312756835268089D400EA8733AA00EC756C5C95 (cit. on pp. 25, 26, 29).

[101] C. Yin, E. S. W. Lui, T. Jiang, and R. Z. Qi, «Proteolysis of γ-tubulin small complex proteins is mediated by the ubiquitin-proteasome system», *FEBS Letters*, vol. 595, no. 15, pp. 1987–1996, Aug. 2021, ISSN: 0014-5793. DOI: 10.1002/1873-3468.14146. [Online]. Available: https://onlinelibrary-wiley-com.login.ezproxy.library.ualberta.ca/doi/full/10.1002/1873-3468.14146%20https://onlinelibrary-wiley-com.login.ezproxy.library.ualberta.ca/doi/abs/10.1002/1873-3468.14146%20https://febs-onlinelibrary-wiley-com.login.ezproxy.library (cit. on p. 25).

[102] B. R. Oakley, V. Paolillo, and Y. Zheng, «γ-tubulin complexes in microtubule nucleation and beyond», *Molecular Biology of the Cell*, vol. 26, no. 17, D. G. Drubin, Ed., pp. 2957–2962, Sep. 2015, ISSN: 19394586. DOI: 10.1091/mbc.E14-11-1514. [Online]. Available: https://www.molbiolcell.org/doi/10.1091/mbc.E14-11-1514 (cit. on pp. 25, 26, 34).

[103] L. Lindström *et al.*, «Therapeutic targeting of nuclear g-tubulin in RB1-negative tumors», *Molecular Cancer Research*, vol. 13, no. 7, pp. 1073–1082, Jul. 2015, ISSN: 15573125. DOI: 10.1158/1541-7786.MCR-15-0063-T (cit. on pp. 25, 32).

[104] M. Corvaisier and M. Alvarado-Kristensson, «Non-Canonical Functions of the Gamma-Tubulin Meshwork in the Regulation of the Nuclear Architecture», *Cancers*, vol. 12, no. 11, p. 3102, Oct. 2020, ISSN: 2072-6694. DOI: 10.3390/cancers12113102. [Online]. Available: https://www.mdpi.com/2072-6694/12/11/3102 (cit. on p. 25).

[105] L. Lindström *et al.*, «The GTPase domain of gamma-tubulin is required for normal mitochondrial function and spatial organization», *Communications Biology*, vol. 1, no. 1, p. 37, Dec. 2018, ISSN: 2399-3642. DOI: 10.1038/s42003-018-0037-3. [Online]. Available: http://www.nature.com/articles/s42003-018-0037-3 (cit. on pp. 25, 29).

[106] E. Zupa *et al.*, «The cryo-EM structure of a γ-TuSC elucidates architecture and regulation of minimal microtubule nucleation systems», *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–12, Nov. 2020, ISSN: 2041-1723. DOI: 10.1038/s41467-020-19456-8. [Online]. Available: https://www.nature.com/articles/s41467-020-19456-8 (cit. on p. 26).

[107] E. Różańska *et al.*, «Expression of both Arabidopsis γ-tubulin genes is essential for development of a functional syncytium induced by Heterodera schachtii», *Plant Cell Reports*, vol. 37, no. 9, pp. 1279–1292, Sep. 2018, ISSN: 07217714. DOI: 10.1007/S00299-018-2312-7/FIGURES/5. [Online]. Available: https://link.springer.com/article/10.1007/s00299-018-2312-7 (cit. on p. 26).

[108] J. M. Kollman, A. Merdes, L. Mourey, and D. A. Agard, «Microtubule nucleation by γ-tubulin complexes», *Nature Reviews Molecular Cell Biology 2011 12:11*, vol. 12, no. 11, pp. 709–721, Oct. 2011, ISSN: 1471-0080. DOI: 10.1038/nrm3209. [Online]. Available: https://www.nature.com/articles/nrm3209 (cit. on p. 26).

[109] H. Fujita, Y. Yoshino, and N. Chiba, «Regulation of the centrosome cycle», *https://doi.org/10.1080/23723556.2015.10* vol. 3, no. 2, Mar. 2016, ISSN: 23723556. DOI: 10.1080/23723556.2015.1075643. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/23723556.2015.1075643 (cit. on pp. 26, 35).

[110] R. E. Uzbekov and T. Avidor-Reiss, «Principal Postulates of Centrosomal Biology. Version 2020», *Cells 2020, Vol. 9, Page 2156*, vol. 9, no. 10, p. 2156, Sep. 2020, ISSN: 20734409. DOI: 10.3390/CELLS9102156. [Online]. Available: https://www.mdpi.com/2073-4409/9/10/2156/htm%20https://www.mdpi.com/2073-4409/9/10/2156 (cit. on pp. 26, 35).

[111] M. Würtz *et al.*, «Modular assembly of the principal microtubule nucleator γ-TuRC», *Nature Communications 2022 13:1*, vol. 13, no. 1, pp. 1–16, Jan. 2022, ISSN: 2041-1723. DOI: 10.1038/s41467-022-28079-0. [Online]. Available: https://www.nature.com/articles/s41467-022-28079-0 (cit. on p. 26).

[112] M. Würtz *et al.*, «Reconstitution of the recombinant human γ-tubulin ring complex», *Open Biology*, vol. 11, no. 2, Feb. 2021, ISSN: 20462441. DOI: 10.1098/RSOB.200325. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsob.200325 (cit. on p. 26).

[113] L. J. Leandro-García *et al.*, «Tumoral and tissue-specific expression of the major human β-tubulin isotypes», *Cytoskeleton*, vol. 67, no. 4, pp. 214–223, Apr. 2010, ISSN: 19493584. DOI: 10.1002/cm.20436. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/cm.20436 (cit. on p. 28).

[114] B. Hořejší *et al.*, «Nuclear γ-tubulin associates with nucleoli and interacts with tumor suppressor protein C53», *Journal of Cellular Physiology*, vol. 227, no. 1, pp. 367–382, Jan. 2012, ISSN: 00219541. DOI: 10.1002/jcp.22772. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/jcp.22772 (cit. on p. 28).

[115] C. D. Katsetos, A. Legido, E. Perentes, and S. J. Mörk, *Class III β-tubulin isotype: A key cytoskeletal protein at the crossroads of developmental neurobiology and tumor neuropathology*, Dec. 2003. DOI: 10.1177/088307380301801205. [Online]. Available: http://journals.sagepub.com/doi/10.1177/088307380301801205 (cit. on p. 28).

[116] L. Gombos *et al.*, «GTP regulates the microtubule nucleation activity of γ-tubulin», *Nature Cell Biology*, vol. 15, no. 11, pp. 1317–1327, Nov. 2013, ISSN: 1465-7392. DOI: 10.1038/ncb2863. [Online]. Available: http://www.nature.com/articles/ncb2863 (cit. on pp. 28, 29, 54, 56).

[117] L. M. Rice, E. A. Montabana, and D. A. Agard, «The lattice as allosteric effector: Structural studies of $\alpha\beta$- and γ-tubulin clarify the role of GTP in microtubule assembly», *Proceedings of the National Academy of Sciences*, vol. 105, no. 14, pp. 5378–5383, Apr. 2008, ISSN: 0027-8424. DOI: 10.1073/pnas.0801155105. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.0801155105 (cit. on pp. 28, 29, 54).

[118] A. Rayevsky *et al.*, «In silico mechanistic model of microtubule assembly inhibition by selective chromone derivatives», *Journal of Molecular Structure*, vol. 1241, p. 130 633, Oct. 2021, ISSN: 00222860. DOI: 10.1016/j.molstruc.2021.130633. [Online]. Available: https://doi.org/10.1016/j.molstruc.2021.130633%20https://linkinghub.elsevier.com/retrieve/pii/S0022286021007663 (cit. on pp. 29, 33).

[119] K. Tsuchiya and G. Goshima, «Microtubule-associated proteins promote microtubule generation in the absence of γ-tubulin in human colon cancer cells», *Journal of Cell Biology*, vol. 220, no. 12, Dec. 2021, ISSN: 0021-9525. DOI: 10.1083/jcb.202104114. [Online]. Available: https://doi.org/10.1083/jcb.202104114%20https://rupress.org/jcb/article/220/12/e202104114/212818/Microtubule-associated-proteins-promote (cit. on p. 29).

[120] A. Thawani, R. S. Kadzik, and S. Petry, «XMAP215 is a microtubule nucleation factor that functions synergistically with the γ-tubulin ring complex», *Nature Cell Biology 2018 20:5*, vol. 20, no. 5, pp. 575–585, Apr. 2018, ISSN: 1476-4679. DOI: 10.1038/s41556-018-0091-6. [Online]. Available: https://www.nature.com/articles/s41556-018-0091-6 (cit. on p. 29).

[121] S. R. Witus *et al.*, «BRCA1/BARD1 site-specific ubiquitylation of nucleosomal H2A is directed by BARD1», *Nature Structural & Molecular Biology 2021 28:3*, vol. 28, no. 3, pp. 268–277, Feb. 2021, ISSN: 1545-9985. DOI: 10.1038/s41594-020-00556-4. [Online]. Available: https://www.nature.com/articles/s41594-020-00556-4 (cit. on p. 30).

[122] Y. Yoshino *et al.*, «BRCA1-Interacting Protein OLA1 Requires Interaction with BARD1 to Regulate Centrosome Number», *Molecular Cancer Research*, vol. 16, no. 10, pp. 1499–1511, Oct. 2018, ISSN: 1541-7786. DOI: 10.1158/1541-7786.MCR-18-0269. [Online]. Available: https://mcr.aacrjournals.org/content/16/10/1499%20https://mcr.aacrjournals.org/content/16/10/1499.abstract (cit. on p. 30).

[123] Y. Yoshino *et al.*, «RACK1 regulates centriole duplication by controlling localization of BRCA1 to the centrosome in mammary tissue-derived cells», *Oncogene 2019 38:16*, vol. 38, no. 16, pp. 3077–3092, Jan. 2019, ISSN: 1476-5594. DOI: 10.1038/s41388-018-0647-8. [Online]. Available: https://www.nature.com/articles/s41388-018-0647-8 (cit. on p. 30).

[124] Y. Yoshino *et al.*, *Dysregulation of the centrosome induced by BRCA1 deficiency contributes to tissue-specific carcinogenesis*, 2021. DOI: 10.1111/cas.14859 (cit. on pp. 30, 33, 35, 37, 38).

[125] S. L. Clark *et al.*, *Structure-function of the tumor suppressor BRCA1*, Apr. 2012. DOI: 10.5936/csbj.201204005 (cit. on p. 30).

[126] K. Yamane, E. Katayama, and T. Tsuruo, «The BRCT Regions of Tumor Suppressor BRCA1 and of XRCC1 Show DNA End Binding Activity with a Multimerizing Feature», *Biochemical and Biophysical Research Communications*, vol. 279, no. 2, pp. 678–684, Dec. 2000, ISSN: 0006-291X. DOI: 10.1006/BBRC.2000.3983 (cit. on p. 30).

[127] Z. Y. Pessetto, Y. Yan, T. Bessho, and A. Natarajan, «Inhibition of BRCT(BRCA1)-phosphoprotein interaction enhances the cytotoxic effect of olaparib in breast cancer cells: A proof of concept study for synthetic lethal therapeutic option», *Breast Cancer Research and Treatment*, vol. 134, no. 2, pp. 511–517, May 2012, ISSN: 15737217. DOI: 10.1007/S10549-012-2079-4/FIGURES/4. [Online]. Available: https://link.springer.com/article/10.1007/s10549-012-2079-4 (cit. on p. 30).

[128] Y. Miki *et al.*, «A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1», *Science*, vol. 266, no. 5182, pp. 66–71, 1994, ISSN: 00368075. DOI: 10.1126/SCIENCE.7545954. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.7545954 (cit. on p. 30).

[129] D. W. Siemann, D. J. Chaplin, and P. A. Walicke, «A review and update of the current status of the vasculature-disabling agent combretastatin-A4 phosphate (CA4P)», *Expert Opinion on Investigational Drugs*, vol. 18, no. 2, pp. 189–197, Feb. 2009, ISSN: 1354-3784. DOI: 10.1517/13543780802691068. [Online]. Available: http://www.tandfonline.com/doi/full/10.1517/13543780802691068 (cit. on p. 32).

[130] D. Chaimovitsh *et al.*, «The relative effect of citral on mitotic microtubules in wheat roots and BY2 cells», *Plant Biology*, vol. 14, no. 2, pp. 354–364, Mar. 2012, ISSN: 14358603. DOI: 10.1111/j.1438-8677.2011.00511.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1438-8677.2011.00511.x (cit. on p. 32).

[131] D. Chaimovitsh *et al.*, «Microtubules are an intracellular target of the plant terpene citral», *The Plant Journal*, vol. 61, no. 3, pp. 399–408, Feb. 2010, ISSN: 09607412. DOI: 10.1111/j.1365-313X.2009.04063.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2009.04063.x (cit. on p. 32).

[132] Å. Ehlén *et al.*, «Tumors with Nonfunctional Retinoblastoma Protein Are Killed by Reduced γ-Tubulin Levels», *Journal of Biological Chemistry*, vol. 287, no. 21, pp. 17 241–17 247, May 2012, ISSN: 00219258. DOI: 10.1074/jbc.M112.357038. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021925820502469 (cit. on p. 32).

[133] G. Traversi *et al.*, «A novel resveratrol derivative induces mitotic arrest, centrosome fragmentation and cancer cell death by inhibiting γ-tubulin», *Cell Division*, vol. 14, no. 1, pp. 1–12, Apr. 2019, ISSN: 17471028. DOI: 10.1186/S13008-019-0046-8/TABLES/1. [Online]. Available: https://celldiv.biomedcentral.com/articles/10.1186/s13008-019-0046-8 (cit. on p. 33).

[134] P. K. Naik, S. Santoshi, A. Rai, and H. C. Joshi, «Molecular modelling and competition binding study of Br-noscapine and colchicine provide insight into noscapinoid–tubulin binding site», *Journal of Molecular Graphics and Modelling*, vol. 29, no. 7, pp. 947–955, Jun. 2011, ISSN: 10933263. DOI: 10.1016/j.jmgm.2011.03.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1093326311000490 (cit. on p. 33).

[135] C. Suri and P. Naik, «Combined molecular dynamics and continuum solvent approaches (MM-PBSA/GBSA) to predict noscapinoid binding to γ-tubulin dimer», *SAR and QSAR in Environmental Research*, vol. 26, no. 6, pp. 507–519, Jun. 2015, ISSN: 1062-936X. DOI: 10.1080/1062936X.2015.1070200. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/1062936X.2015.1070200 (cit. on p. 33).

[136] T. Chinen *et al.*, «The γ-tubulin-specific inhibitor gatastatin reveals temporal requirements of microtubule nucleation during the cell cycle», *Nature Communications*, vol. 6, no. 1, pp. 2–11, Dec. 2015, ISSN: 20411723. DOI: 10.1038/ncomms9722. [Online]. Available: http://www.nature.com/articles/ncomms9722 (cit. on p. 33).

[137] K. Shintani *et al.*, «Structure Optimization of Gatastatin for the Development of γ-Tubulin-Specific Inhibitor», *ACS Medicinal Chemistry Letters*, vol. 11, no. 6, pp. 1125–1129, Jun. 2020, ISSN: 1948-5875. DOI: 10.1021/acsmedchemlett.9b00526. [Online]. Available: https://pubs.acs.org/doi/10.1021/acsmedchemlett.9b00526 (cit. on p. 33).

[138] E. A. Nigg, «Centrosome aberrations: cause or consequence of cancer progression?», *Nature Reviews Cancer 2002 2:11*, vol. 2, no. 11, pp. 815–825, Nov. 2002, ISSN: 1474-1768. DOI: 10.1038/nrc924. [Online]. Available: https://www.nature.com/articles/nrc924 (cit. on pp. 33, 37).

[139] S. A. Godinho *et al.*, «Oncogene-like induction of cellular invasion from centrosome amplification», *Nature*, vol. 510, no. 7503, pp. 167–171, 2014, ISSN: 14764687. DOI: 10.1038/nature13277 (cit. on pp. 33, 37, 38).

[140] C. D. Katsetos *et al.*, «Class III β-Tubulin and γ-Tubulin are Co-expressed and Form Complexes in Human Glioblastoma Cells», *Neurochemical Research*, vol. 32, no. 8, pp. 1387–1398, Aug. 2007, ISSN: 0364-3190. DOI: 10.1007/s11064-007-9321-1. [Online]. Available: http://link.springer.com/10.1007/s11064-007-9321-1 (cit. on p. 33).

[141] R. Shen *et al.*, «A novel TUBG1 mutation with neurodevelopmental disorder caused by malformations of cortical development», *BioMed Research International*, vol. 2021, M. Tang, Ed., pp. 1–8, Feb. 2021, ISSN: 2314-6141. DOI: 10.1155/2021/6644274. [Online]. Available: https://www.hindawi.com/journals/bmri/2021/6644274/ (cit. on p. 33).

[142] N. Scholz, K. M. Kurian, F. A. Siebzehnrubl, and J. D. Licchesi, «Targeting the Ubiquitin System in Glioblastoma», *Frontiers in Oncology*, vol. 10, Nov. 2020, ISSN: 2234943X. DOI: 10.3389/fonc.2020.574011. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2020.574011/full (cit. on p. 34).

[143] D. S. Rickman *et al.*, «Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis», *Cancer Research*, vol. 61, no. 18, pp. 6885–6891, 2001, ISSN: 00085472. [Online]. Available: https://aacrjournals.org/cancerres/article/61/18/6885/508046/Distinctive-Molecular-Profiles-of-High-Grade-and (cit. on p. 34).

[144] P. L. Chavali, M. Pütz, and F. Gergely, «Small organelle, big responsibility: the role of centrosomes in development and disease», *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1650, 2014, ISSN: 14712970. DOI: 10.1098/RSTB.2013.0468. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2013.0468 (cit. on p. 35).

[145] K. Mittal *et al.*, «Centrosome amplification: a quantifiable cancer cell trait with prognostic value in solid malignancies», *Cancer and Metastasis Reviews 2020 40:1*, vol. 40, no. 1, pp. 319–339, Oct. 2020, ISSN: 1573-7233. DOI: 10.1007/S10555-020-09937-Z. [Online]. Available: https://link.springer.com/article/10.1007/s10555-020-09937-z (cit. on pp. 35, 37, 38).

[146] C. D. Katsetos *et al.*, «Altered cellular distribution and subcellular sorting of gamma-tubulin in diffuse astrocytic gliomas and human glioblastoma cell lines.», *Journal of Neuropathology and Experimental Neurology*, vol. 65, no. 5, pp. 465–477, May 2006, ISSN: 0022-3069. DOI: 10.1097/01.JNEN.0000229235.20995.6E. [Online]. Available: https://europepmc.org/article/MED/16772870 (cit. on p. 37).

[147] W. L. Lingle *et al.*, «Centrosome hypertrophy in human breast tumors: Implications for genomic stability and cell polarity», vol. 95, no. 6, pp. 2950–2955, Mar. 1998. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.95.6.2950%20https://www.pnas.org/content/95/6/2950%20https://www.pnas.org/content/95/6/2950.abstract (cit. on p. 37).

[148] W. L. Lingle *et al.*, «Centrosome amplification drives chromosomal instability in breast tumor development», *Proceedings of the National Academy of Sciences*, vol. 99, no. 4, pp. 1978–1983, Feb. 2002, ISSN: 0027-8424. DOI: 10.1073/PNAS.032479999. [Online]. Available: https://www.pnas.org/content/99/4/1978%20https://www.pnas.org/content/99/4/1978.abstract (cit. on p. 37).

[149] A. B. D'Assoro *et al.*, «Amplified Centrosomes in Breast Cancer: A Potential Indicator of Tumor Aggressiveness», *Breast Cancer Research and Treatment 2002 75:1*, vol. 75, no. 1, pp. 25–34, Sep. 2002, ISSN: 1573-7217. DOI: 10.1023/A:1016550619925. [Online]. Available: https://link.springer.com/article/10.1023/A:1016550619925 (cit. on pp. 37, 38).

[150] A. Lavecchia and C. Giovanni, «Virtual Screening Strategies in Drug Discovery: A Critical Review», *Current Medicinal Chemistry*, vol. 20, no. 23, pp. 2839–2860, Jun. 2013, ISSN: 09298673. DOI: 10.2174/09298673113209990001 (cit. on pp. 39–42, 47).

[151] E. H. B. Maia *et al.*, «Structure-Based Virtual Screening: From Classical to Artificial Intelligence», *Frontiers in Chemistry*, vol. 8, p. 343, Apr. 2020, ISSN: 22962646. DOI: 10.3389/FCHEM.2020.00343/BIBTEX (cit. on pp. 39–41, 43, 47, 49, 52).

[152] B. B.-W. GmbH. «Virtual database screening saves time and money». (2008), [Online]. Available: https://www.gesundheitsindustrie-bw.de/en/article/press-release/virtual-database-screening-saves-time-and-money. (accessed: 16.11.2022) (cit. on pp. 39, 40).

[153] V. Prasad and S. Mailankody, «Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval», *JAMA Internal Medicine*, vol. 177, no. 11, pp. 1569–1575, Nov. 2017, ISSN: 2168-6106. DOI: 10.1001/JAMAINTERNMED.2017.3601. [Online]. Available: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2653012 (cit. on p. 39).

[154] S. Morgan *et al.*, «The cost of drug development: A systematic review», *Health Policy*, vol. 100, no. 1, pp. 4–17, Apr. 2011, ISSN: 0168-8510. DOI: 10.1016/J.HEALTHPOL.2010.12.002 (cit. on p. 39).

[155] S. M. Paul *et al.*, «How to improve R&D productivity: the pharmaceutical industry's grand challenge», *Nature Reviews Drug Discovery 2010 9:3*, vol. 9, no. 3, pp. 203–214, Feb. 2010, ISSN: 1474-1784. DOI: 10.1038/nrd3078. [Online]. Available: https://www.nature.com/articles/nrd3078 (cit. on p. 39).

[156] N. Vabret *et al.*, «Immunology of COVID-19: Current State of the Science», *Immunity*, vol. 52, no. 6, pp. 910–941, Jun. 2020, ISSN: 1074-7613. DOI: 10.1016/J.IMMUNI.2020.05.002. [Online]. Available: http://www.cell.com/article/S1074761320301837/fulltext%20http://www.cell.com/article/S1074761320301837/abstract%20https://www.cell.com/immunity/abstract/S1074-7613(20)30183-7 (cit. on p. 40).

[157] S. O. Abd Albagi *et al.*, «A multiple peptides vaccine against COVID-19 designed from the nucleocapsid phosphoprotein (N) and Spike Glycoprotein (S) via the immunoinformatics approach», *Informatics in Medicine Unlocked*, vol. 21, p. 100 476, Jan. 2020, ISSN: 2352-9148. DOI: 10.1016/J.IMU.2020.100476 (cit. on p. 40).

[158] C. Arnold, «How computational immunology changed the face of COVID-19 vaccine development», *Nature Medicine*, Jul. 2020, ISSN: 1078-8956. DOI: 10.1038/D41591-020-00027-9 (cit. on p. 40).

[159] A. M. Almehdi *et al.*, «SARS-CoV-2 spike protein: pathogenesis, vaccines, and potential therapies», *Infection*, vol. 49, no. 5, pp. 855–876, Oct. 2021, ISSN: 14390973. DOI: 10.1007/S15010-021-01677-8/FIGURES/4. [Online]. Available: https://link.springer.com/article/10.1007/s15010-021-01677-8 (cit. on p. 40).

[160] C. Yang, E. A. Chen, and Y. Zhang, «Protein-Ligand Docking in the Machine-Learning Era», *Molecules 2022, Vol. 27, Page 4568*, vol. 27, no. 14, p. 4568, Jul. 2022, ISSN: 1420-3049. DOI: 10.3390/MOLECULES27144568. [Online]. Available: https://www.mdpi.com/1420-3049/27/14/4568/htm%20https://www.mdpi.com/1420-3049/27/14/4568 (cit. on pp. 40, 41, 49, 52, 59).

[161] M. L. Peach and M. C. Nicklaus, «Combining docking with pharmacophore filtering for improved virtual screening», *Journal of Cheminformatics*, vol. 1, no. 1, pp. 1–15, May 2009, ISSN: 17582946. DOI: 10.1186/1758-2946-1-6/FIGURES/11. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-1-6 (cit. on pp. 41, 47).

[162] J. Preto and F. Gentile, «Assessing and improving the performance of consensus docking strategies using the DockBox package», *Journal of Computer-Aided Molecular Design*, vol. 33, no. 9, pp. 817–829, Sep. 2019, ISSN: 15734951. DOI: 10.1007/S10822-019-00227-7/FIGURES/4. [Online]. Available: https://link.springer.com/article/10.1007/s10822-019-00227-7 (cit. on pp. 41–43, 64).

[163] D. R. Houston and M. D. Walkinshaw, «Consensus docking: Improving the reliability of docking in a virtual screening context», *Journal of Chemical Information and Modeling*, vol. 53, no. 2, pp. 384–390, Feb. 2013, ISSN: 1549960X. DOI: 10.1021/ci300399w. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/ci300399w (cit. on pp. 41, 43).

[164] O. Trott and A. J. Olson, «AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading», *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, Jan. 2010, ISSN: 1096-987X. DOI: 10.1002/JCC.21334. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.21334%20https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334%20https://onlinelibrary.wiley.com/doi/10.1002/jcc.21334 (cit. on pp. 41, 43).

[165] W. J. Allen *et al.*, «DOCK 6: Impact of new features and current docking performance», *Journal of Computational Chemistry*, vol. 36, no. 15, pp. 1132–1156, Jun. 2015, ISSN: 01928651. DOI: 10.1002/jcc.23905. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/jcc.23905 (cit. on pp. 41, 43).

[166] G. M. Morris *et al.*, «AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility», *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009, ISSN: 1096-987X. DOI: 10.1002/JCC.21256. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jcc.21256%20https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21256%20https://onlinelibrary.wiley.com/doi/10.1002/jcc.21256 (cit. on pp. 41, 43).

[167] C. C. G. ULC, *Molecular operating environment (moe), 2022.02*, 2022 (cit. on pp. 42, 43, 62, 65, 66).

[168] R. A. Friesner *et al.*, «Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy», *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004, ISSN: 0022-2623. DOI: 10.1021/jm0306430. [Online]. Available: https://pubs.acs.org/doi/10.1021/jm0306430 (cit. on p. 42).

[169] T. D. Kühne *et al.*, «CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations», *The Journal of Chemical Physics*, vol. 152, no. 19, p. 194 103, May 2020, ISSN: 0021-9606. DOI: 10.1063/5.0007045. [Online]. Available: https://aip.scitation.org/doi/10.1063/5.0007045%20http://aip.scitation.org/doi/10.1063/5.0007045 (cit. on p. 42).

[170] D. K. Agrafiotis *et al.*, «Conformational sampling of bioactive molecules: A comparative study», *Journal of Chemical Information and Modeling*, vol. 47, no. 3, pp. 1067–1086, 2007, ISSN: 1549960X. DOI: 10.1021/CI6005454/ASSET/IMAGES/LARGE/CI6005454F00006.JPEG. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/ci6005454 (cit. on p. 45).

[171] K.-H. Kim, N. D. Kim, and B.-L. Seong, «Pharmacophore-based virtual screening: a review of recent applications», *Expert Opinion on Drug Discovery*, vol. 5, no. 3, pp. 205–222, Mar. 2010, ISSN: 1746-0441. DOI: 10.1517/17460441003592072. [Online]. Available: http://www.tandfonline.com/doi/full/10.1517/17460441003592072 (cit. on p. 47).

[172] F. Gentile *et al.*, «Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery», *ACS Central Science*, vol. 6, no. 6, pp. 939–949, Jun. 2020, ISSN: 23747951. DOI: 10.1021/ACSCENTSCI.0C00229/ASSET/IMAGES/MEDIUM/OC0C00229_M005.GIF. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acscentsci.0c00229 (cit. on pp. 49–51, 67, 98).

[173] H. Alaskar and T. Saba, «Machine Learning and Deep Learning: A Comparative Review», pp. 143–150, 2021. DOI: 10.1007/978-981-33-6307-6_15. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-33-6307-6_15 (cit. on pp. 49, 52).

[174] F. Ren *et al.*, «AlphaFold Accelerates Artificial Intelligence Powered Drug Discovery: Efficient Discovery of a Novel Cyclin-dependent Kinase 20 (CDK20) Small Molecule Inhibitor», Jan. 2022. DOI: 10.48550/arxiv.2201.09647. arXiv: 2201.09647. [Online]. Available: https://arxiv.org/abs/2201.09647v2 (cit. on p. 49).

[175] I. A. Guedes *et al.*, «New machine learning and physics-based scoring functions for drug discovery», *Scientific Reports*, vol. 11, no. 1, p. 3198, Dec. 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-82410-1. [Online]. Available: http://www.nature.com/articles/s41598-021-82410-1 (cit. on p. 49).

[176] F. Gentile *et al.*, «Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking», *Nature Protocols 2022 17:3*, vol. 17, no. 3, pp. 672–697, Feb. 2022, ISSN: 1750-2799. DOI: 10.1038/s41596-021-00659-2. [Online]. Available: https://www.nature.com/articles/s41596-021-00659-2 (cit. on pp. 49–51, 61, 67, 98).

[177] «About lumi». (2022), [Online]. Available: https://www.lumi-supercomputer.eu/about-lumi/. (accessed: 21.11.2022) (cit. on p. 49).

[178] O. of Science - U.S.A. «About lumi». (2022), [Online]. Available: https://www.energy.gov/science/articles/fact-sheet-inflation-reduction-act-supporting-future-doe-science. (accessed: 21.11.2022) (cit. on p. 49).

[179] A.-T. Ton *et al.*, «Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds», *Molecular Informatics*, vol. 39, no. 8, p. 2 000 028, Aug. 2020, ISSN: 1868-1743. DOI: 10.1002/minf.202000028. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/minf.202000028 (cit. on pp. 50, 98).

[180] J. V. Olsen *et al.*, «Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks», *Cell*, vol. 127, no. 3, pp. 635–648, Nov. 2006, ISSN: 0092-8674. DOI: 10.1016/J.CELL.2006.09.026 (cit. on p. 55).

[181] J. D. Stephenson and S. J. Freeland, «Unearthing the root of amino acid similarity», *Journal of Molecular Evolution*, vol. 77, no. 4, pp. 159–169, Oct. 2013, ISSN: 00222844. DOI: 10.1007/s00239-013-9565-0 (cit. on p. 60).

[182] J. J. Irwin *et al.*, «ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery», *Journal of Chemical Information and Modeling*, vol. 60, no. 12, pp. 6065–6073, Dec. 2020, ISSN: 1549960X. DOI: 10.1021/ACS.JCIM.0C00675/ASSET/IMAGES/LARGE/CI0C00675_0007.JPEG. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c00675 (cit. on p. 61).

[183] D. Rogers and M. Hahn, «Extended-connectivity fingerprints», *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, May 2010, ISSN: 1549960X. DOI: 10.1021/CI100050T/ASSET/IMAGES/LARGE/CI-2010-00050T_0017.JPEG. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/ci100050t (cit. on p. 67).

[184] P. Penner *et al.*, «Shape-Based Descriptors for Efficient Structure-Based Fragment Growing», *Journal of Chemical Information and Modeling*, vol. 60, no. 12, pp. 6269–6281, Dec. 2020, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00920. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.0c00920 (cit. on p. 98).