UNIVERSITAT DE BARCELONA

Development and application of computational methods to address current challenges in drug discovery

Álvaro Serrano Morrás

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

DEVELOPMENT AND APPLICATION OF COMPUTATIONAL METHODS TO ADDRESS CURRENT CHALLENGES IN DRUG DISCOVERY

ÁLVARO SERRANO MORRÁS
2024

PROGRAMA DE DOCTORAT EN BIOMEDICINA

DIRECTORS: Dr. CARLES GALDEANO i Dr. XAVIER BARRIL ALONSO

TUTOR: Dr. CARLES GALDEANO

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

DEPARTAMENT DE FARMÀCIA I TECNOLOGIA FARMACÈUTICA, I FISICOQUÍMICA

PROGRAMA DE DOCTORAT EN BIOMEDICINA

DEVELOPMENT AND APPLICATION OF COMPUTATIONAL METHODS TO ADDRESS CURRENT CHALLENGES IN DRUG DISCOVERY

Memòria presentada per Álvaro Serrano Morrás per optar al títol de doctor per la universitat de Barcelona

Xavier Barril Alonso
Director de tesi

Carles Galdeano Cantador
Director de tesi i tutor

Álvaro Serrano Morrás
Doctorand

ÁLVARO SERRANO MORRÁS
2024

“Lo mejor es enemigo de lo bueno”

- Voltaire

AKNOWLEDGEMENTS

Well, it is about time I wrote this damned thesis. It has been a long journey and I would not have been able to succeed by myself. I will start with the obvious: the people who supervised me directly. First of all, thank you Xavi for welcoming me in your group and guiding me through the projects with the freedom I needed; to fail and learn, but ensuring I did not get lost. I also want to thank Maciej, who guided me during the masters and whom I have taken as a reference on how to supervise people through positive reinforcement. Finally I'd like to thank Jordi who has adopted me in his group this last period and given me very priceless advice and support, both academically and personally.

Then I really need to thank all my colleagues from the lab, past and present, where I've had the best horizontal collaboration, scientific stimulus and support one can ask. I have been here too long, so this is an extensive list, but many thanks to the previous generation Morena, Moira, Maciej, Salvo, Dylan, Arnau, Alex, Marina, and Guillermo of course for teaching me system administration. I'd like to specially thank Juan and Serena for their personal support, and for being real friends inside and outside the university. I will dearly remember the meetings with Miriam during the covid lockdown and all the work we did together; I learnt a lot from you. I'd also like to thank Bego for challenging me with all the scientific discussions for your crazy projects and their technical difficulties, but most importantly for our talks and being so caring and affectionate with me.

I am infinitely grateful to Andrea. We started this together and we are finishing it together. In spite of all the summers invested needlessly, difficulties and dereliction, we have supported each other and managed to pull through.

After the covid lockdown, the group mostly renewed and a new generation of amazing people started. I cherish the interactions and the good environment we have been able to re-establish. Many many thanks to Roger, Patricia, Varbina, Anna, Adrián, Blanca, Alejandra, Jordi Pou, and Salo for the last years, for the coffees and beers, and for listening to me when I needed to vent. You are great and you'll do amazingly in the future, I genuinely hope we meet again wherever we end up.

I would like to thank the colleagues from Curu's group, starting from him; the most patient and stalwart scientist ever, who has also helped me through my bureaucracy incompetence. Thanks to Marina, Renato, Özge, and especially thanks to Dani and Beste, who started with me and who pushed me to better understand the theory and with whom I shared my day-to-day in the lab, even when I was alone in "my office".

I'd also like to acknowledge the people who have come to the group, even if temporarily, always with a thirst for knowledge and good disposition: Torben, Ania, Corentin, Ernestas, Helena, Vincenzo, and Julian. Every one of you (and the ones I have surely missed) has helped me broaden my view on science and challenged my *status quo* making me grow.

Something I really appreciated during the PhD is being able to transmit some of the knowledge I acquired to students. I cannot

express how much I learned while supervising Marc and Carme through their TFG and TFM. Thank you for your goodwill, your motivation, and the patience you had with me.

I am grateful to Julien, who accepted me for a research stay in his group. I learnt so much and had an incredible time with all the people in room 234. Special thanks to Audrius who truly made me feel a part of the group.

While academic support is important, completing this journey would have not been possible without my friends and family (which I feel it is, in fact, two ways of saying the same thing). Thanks to ‘Sopapos’; Yasmine, Laia, Guille, Danae, Mariana, Diego, Jaume and Marc, the best people in the world, to David, to Cesc and Erola. Special thanks to my adoptive family; ‘mis suegros favoritos’ Hugo y Ceci. Thanks to Alberto and Miriam, and many thanks to Jeremy.

I am coming to an end here, with those who have always been there for me, since the beginning. Mom, you are my academic, but also life reference. Mushu, my stress reliever and cuddle machine.

Last but not least, thank you Ani. Thank you for being my companion, for understanding me and for supporting me when I was impossible to understand. The best part of this thesis, and of the last eleven years of my life, has been growing by your side.

ABSTRACT

The recent advances in drug discovery have introduced many opportunities to solve previously unattainable medical needs. These opportunities challenge the *status quo* of the methodologies currently in use. Computational tools can help alleviate these challenges, and have been steadily growing in reliability and importance in drug discovery. The main motivation of this thesis is to develop and apply structure-based computational methodologies to address some of the main challenges in drug discovery. In particular, I have focused on expanding the available target space, explored new approaches to exploit ultra-large chemical libraries for virtual screening and developed new implementations of steered molecular dynamics to evaluate protein-ligand complexes.

E3 ligases are the main recognition component of the ubiquitin-mediated degradation machinery of the cell. Despite their crucial role for targeted protein degradation applications, only a few have been targeted. To expand the available E3 toolbox, I have performed a ligandability assessment of a set of 23 E3 ligases using mixed organic solvent molecular dynamics, and validated the predicted pockets both retrospectively and prospectively. Moreover, I evaluated the potential conformational impact of the binding of a FBW7 allosteric modulator found during the prospective validation.

Ultra-large on-demand chemical libraries offer the opportunity to revolutionize *in silico* hit identification, providing a diversity and chemotype depth unprecedented. However, current virtual screening technologies are not yet adapted to perform under such library sizes. I have developed a pipeline to efficiently explore the ultra-large on-demand chemical libraries from the bottom-up to improve hit identification campaigns. We have validated the approach finding BRD4 inhibitors by growing scaffolds selected from three different scenarios, obtaining a high ratio of diverse and potent hits.

Finally, I have assessed the use of Dynamic Undocking as a fast tool for relative binding free energy calculations and activity cliff identification, delimiting its domain of applicability. I have developed and validated OpenDUck, a python library to facilitate the creation of new applications within the Dynamic Undocking framework. Finally, I have evaluated the use of the Crooks fluctuation theorem for free energy calculations from non-equilibrium simulations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ABF	Adaptive biasing force
AC	Activity cliffs
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
AFE	Alchemical free energy
AI	Artificial intelligence
AM1-BCC	Austin Model 1 with Bond Charge Correction
AMBER	Assisted model building with energy refinement
ASB9	Ankyrin repeat and SOCS-box containing protein 9
AUTAC	Autophagy-targeting chimera
BACE1	Beta-secretase 1
BD1	Bromodomain 1
BRD4	Bromodomain-containing protein 4
BTB	Broad-complex, Tramtrack and Bric brac domain
BTRC	Beta-transducing repeat containing protein
CADD	Computer-aided drug discovery
CDC20	Cell division cycle protein 20
CDK2	Cyclin-dependent kinase 2
CFT	Crooks fluctuation theorem
CGI	Crooks Gaussian intersection
CK1α	Casein kinase 1α
CLI	Command line interface
cMD	conventional MD
C-MYC	Myc proto-oncogene protein
CPU	Central processing unit
CRBN	Cereblon protein
CUL1	Cullin 1
CV	Collective Variable
Cryo-EM	Cryogenic electron microscopy
DEL	DNA-encoded library
DL	Deep learning
DNA	Desoxyribonucleic acid

DSF	Differential scanning fluorimetry
DUBTAC	Deubiquitination-targeting chimera
DUck	Dynamic Undocking
ECFP	Extended connectivity figerprint
EMBL	European molecular biology laboratory
FAIR	Findable, Accessible, Interoperable and Reusable
FBW7	F-box and WD repeat domain containing protein 7)
FEP	Free energy perturbations
FES	Free energy surface
FP	Fluorescence polarization
GAFF	General AMBER force field
GPCR	G-protein coupled receptors
GID4	Glucose-induced degradation protein 4
GPU	Graphic processing unit
HAC	Heavy atom count
HMR	Hydrogen mass repartitioning
HPC	High performing computers
HSP90	Heat shock protein 90
HTS	High-throughput screening
HTVS	High-throughput virtual screening
IQTC	Institut de química teòrica
ITC	Isothermal titration calorimetry
JE	Jarzynski equation
KCTD5	Potassium channel tetramerization domain 5
KDE	Kernel density estimation
KEAP1	KELCH-like ECH-associated protein 1
KLHDC2	KELCH domain-containing protein 2
LBDD	Ligand-based drug discovery
MACCS	Molecular access system

MAE	Mean absolute error
MCC	Matthew correlation coefficient
MCS	Maximum common substructure
MD	Molecular dynamics
MDMIX	Molecular dynamics with mixed solvents
ML	Machine learning
MM/GBSA	Molecular Mechanics and Generalized Born Surface Area
MOA	Mechanism of action
MST	Microscale thermophoresis
NPT	Isothermal-isobaric ensemble
NVT	Canonical ensemble
PAINS	Pan-assay interference compounds
PBC	Periodic boundary conditions
PC	Principal component
PCA	Principal component analysis
PDB	Protein Data Bank
PIP	Proximity-induced pharmacology
PME	Particle mesh Ewald
PMEMD	Particle mesh Ewald MD
PMF	Potential mean force
PPI	Protein-Protein interaction
PRKN	Parkin
PROTAC	Proteolysis-targeting chimera
QSAR	Quantitative structure-activity relationship
RAMD	Random accelerated molecular dynamic
RBFE	Relative binding free energy
RBX1	Ring-box 1
REAL Space	Readily accessible fragment space

RES	Red española de supercomputacion
RESP	Restrained electrostatic potential
RMSD	Root mean square distance
RMSF	Root mean square fluctuations
RNA	Ribonucleic acid
RNF	RING finger protein
SAR	Structure-activity relationship
SBDD	Structure-based drug discovery
SCF	SKP, Culin, F-box containing complex
SKP1	S-phase kinase associated protein 1
SMARTS	SMILES arbitrary target specification
SMD	Steered Molecular Dynamics
SMILES	Simplified molecular input line entry system
SMIRNOFF	SMIRKS native open force field
SPR	Surface plasmon resonance
SVL	Scientific Vector Language
TFLOPS	Tera floating point operation per second
TI	Thermodynamic integration
TPD	Targeted protein degradation
TR-FRET	Time-resolved Förster resonance energy transfer
TRIM	Tripartite motif-containing protein
tSNE	t-distributed stochastic neighbour embedding
UPS	Ubiquitin-proteasome system
VHL	Von Hippel-Lindau
VS	Virtual screening
WDR	Tryptophan-aspartic acid repeat domain

STATEMENT OF CONTRIBUTORS

Most of the projects expounded in this thesis are the result of the collaborative efforts of a multidisciplinary team of scientists, to whom I am grateful. During the explanation of each section I have highlighted the specific contributions of each collaborator. However, here I shall state all of these external contributions in a joint manner.

Regarding the project on the ligandability assessment of the E3 ligases, Miriam Martinez Cartró and I did together the analysis on the MDMix trajectories and the selection of ligandable pockets. She then did the virtual screening, SAR and biophysical validation on the FBW7 G+H pockets which resulted in the A5 ligand. Together with Roger Castaño-Muñiz and Andrea Bertran-Mostazo they corroborated the binding site with SPR, ITC, MST and FP. Finally Andrea has evaluated the phenotypic effect of A5's family of compounds and their mechanism of action through more biophysical and cellular experiments.

The *bottom-up exploration of the chemical space* was a very collaborative project mainly involving Marina Miñarro, Andrea Bertran-Mostazo and me. Marina Miñarro performed the initial fragment virtual screening to obtain the computational scaffolds. Arnau Comajuncosa performed the chemical checker clustering and diversity analysis of the compounds. Andrea Bertran-Mostazo performed all the biophysical validation of the compounds, including DSF, SPR, TR-FRET and all the experimental set up and optimization that implies. Finally, David Ruiz spearheaded the crystallization efforts in the context of a collaboration with Maria Garcia-Alai.

Finally, I started the development of the OpenDUck python library from the foundation code written by Mihaela Sminova, Simon Bray and Anthony Bradley. Moreover, Simon Bray has kept contributing, finding bugs and assisting on the implementation of specific features

CHAPTER 1: INTRODUCTION

1. INTRODUCTION

1.1 A general overview of drug development

The recent sanitary worldwide crisis caused by the emergence of the SARS-CoV-2 virus has highlighted the importance of drug discovery for our society and the necessity to quickly develop and deploy novel potent drugs upon the emergence of new and more resistant pathogens.[1] The need is equally pressing in other therapeutic areas, where scientific advances, particularly in the 'omic' fields, have expanded the possibility of personalized and high precision treatments for the numerous unmet medical needs.[2] At the same time, the hindrances to developing new therapies have never been higher, with ever-increasing demands for safety, efficacy and availability by the regulatory agencies, the health systems and the broader society. Drug discovery is quickly evolving to reconcile and solve these diverse and challenging needs.

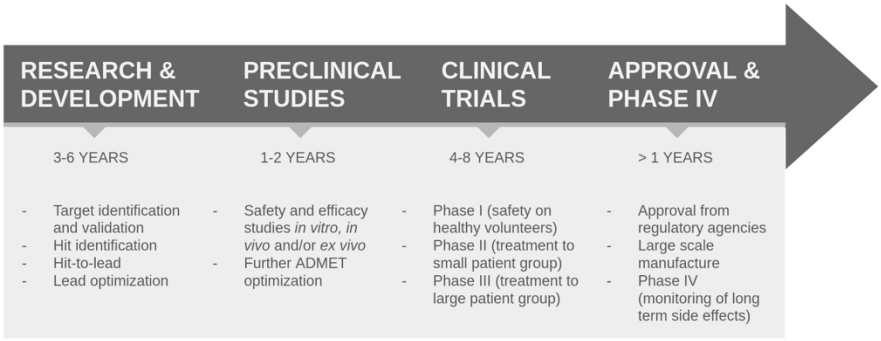


Figure 1.1. Schema of the drug discovery process. The different stages, their duration and the research activities encased within them are highlighted.

The development of a new medicine usually starts with the identification of a promising target (a dysfunctional macromolecule,

signaling pathway or molecular mechanism linked to a disease).[3] To be considered therapeutic targets, external agents must be able to bind and change their function, thus restoring or improving health. These early stages are usually tested on cellular and animal models, which are sometimes too different from, and do not translate to, human biology.[4] The process continues with the search for chemical entities that bind and/or modulate the chosen therapeutic target. A wide range of drug modalities exists nowadays to achieve this effect. They are broadly classified into chemical entities and biological entities. The first category includes traditional small molecules, but also peptides and other entities with a well-defined chemical structure. The second category is better defined by the production method and includes proteins, nucleic acids, viruses, etcetera.

In the small-molecule drug modality, hit identification often relies on high throughput screening (HTS) of large libraries of compounds using scalable biophysical or functional assays.[5] The chemical space around the best hits is then explored more in-depth in the hit-to-lead phase. Structure-activity relationship (SAR) studies are used to guide the evolution of a hit into a lead.[6] Leads are not exclusively optimized in terms of potency, but also binding mechanism,[7] pharmacokinetic properties,[8] safety[9] and patentability. The discovery phase ends with a potent lead which will be developed into a drug candidate during the preclinical studies.

To nominate a molecule as a drug candidate, its efficacy and safety are evaluated through *in vitro*, *in vivo* and/or *ex vivo* assays. Additionally, their pharmacokinetics, so-called ADMET (absorption, distribution, metabolism, excretion, toxicity) profile, becomes a high

priority.[10] In a simplified form, a therapeutic window is established from the preclinical results, and the candidate enters clinical trials.

There are four phases in the clinical trials, numbered I to IV.[11] Phase I involves treating a small number of healthy volunteers to assess the safety and tolerability of the candidate. In phase II, patients from different hospitals are treated to determine whether or not the molecule yields the desired therapeutic effect. Secondary, but not negligible, obstacles also arise at this stage (e.g. synthesis scalability and administration).[12] Phase III is the most expensive part of the whole drug development process. In this step, the efficacy of the treatment is evaluated in a large group of patients that should be representative of the population variability. Upon successful completion of the phase III trial, the candidate is submitted to regulatory agencies for approval for general commercialization. Most agencies require post-marketing surveillance of rare or long-term effects not detected in previous studies, known as phase IV.

Developing a new drug is a very arduous, expensive and risky undertaking. The complete process, from a validated target to an approved drug takes 10 to 15 years and has an average cost of 2.8 billion dollars.[13] Moreover, ninety percent of candidates entering clinical trials fail, despite the efforts in the discovery phase.[14] Albeit clinical trials are the most expensive part of drug development, overall investment in the discovery phase accounts for about 40% of the cost.[15] More importantly, the strategies, procedures and decisions taken in the early stages of drug discovery determine the targets, mechanisms of action, characteristics and, indeed, the fate of the specific molecules that will reach the clinical stage. Therefore,

most time-saving, cost-saving and transformative opportunities reside in the earlier discovery and preclinical stages.

1.2 Current challenges in drug discovery

In the last decades the field has seen huge advances, which have both opened new opportunities to enhance the drug discovery process and exposed new challenges. Whilst the field is filled with innovation, here I will highlight three current challenges regarding target-based drug discovery that are particularly relevant to the thesis.

1.2.1 Overcoming resistance to drug therapies

Streamlining of drug discovery has increased the amount of new approved drugs that enter the market each year.[16] In parallel to the increased repertoire of drugs at our disposal, there has been an increase in the report of drug resistance, mainly to antibiotics, antivirals and to cancer therapies. Despite having different etiologies, the three kinds of resistance to treatment demand multiple drugs acting through different targets, with different mechanisms of action or with different resistance profiles for a given disease or infection.[17]

It has been several decades since the first warnings from the scientific community about an imminent crisis of multiresistant pathogens.[18] Whilst the origins of antibiotic resistance predate humans[19], the excessive, and sometimes inappropriate, use of antibiotic treatments and disinfectants has generated a fitness pressure in pathogens, triggering the appearance of new drug-resistance adaptations.[20] Such adaptations then can spread quickly

through horizontal transmission of genes. Resistance mechanisms can be grouped into three categories: limiting in-cell presence of the drug (limiting the intake or active efflux), inactivating a drug and modifying the drug target. Understanding the underlying molecular bases of such drug-resistance adaptations is critical to select the most appropriate targets for the new generation antibiotics and antivirals.[21]

The appearance of chemotherapy resistance in cancer seems to be triggered by similar mechanisms. Most chemotherapy strategies rely on the higher reproduction and metabolism of cancer cells, and their resulting reliance on specific genes to target and kill those cells. However, the dysfunctioning DNA quality control of such cells allows for the introduction of mutations at a high pace. If by chance, a given mutation allows for the resistance to the therapy they are subjected to, the presence of mutated cells will increase. This treatment-induced selection is similar to the fitness pressure that triggers antibiotic resistant pathogens.[22] In addition, cancers embody a high heterogeneity which allows some cells to avoid treatment through alternative mechanisms such as dormancy or tissue migration.[23]

Strategies to deal with drug resistance depend on cocktails of drugs, each targeting different target proteins or genes responsible for both the pathology and the resistance. Thus, we are in dire need of new targets and drugs that circumvent the resistance of cancer cell lines and pathogens.

1.2.2 Expanding the target space

Out of the approximately 20,000 protein-coding genes in the human genoma,[24] very few are used for drug discovery. However, as stated

before, only druggable targets are considered therapeutically interesting. Druggability is conceived as the ability to bind small molecules capable of modulating the activity of the disease. Only 3,000 human protein-coding genes have been considered druggable.[25]

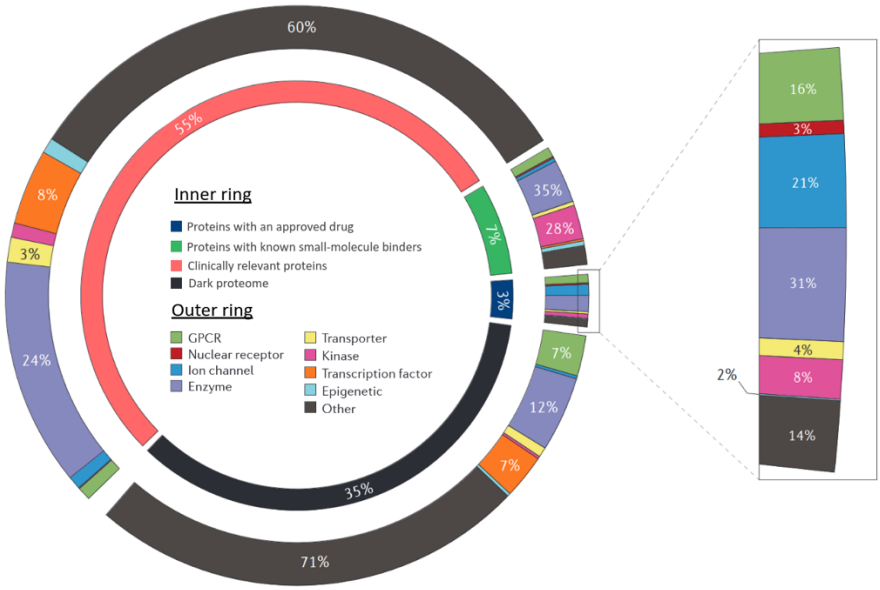


Figure 1.2. Druggability pie chart of the human proteome. The outer circle shows the representation of protein families of the known human proteome. The inner circle shows the clinical status of those proteins. To the right, the protein families of the proteins with an approved drug are highlighted. Adapted from [26]

Furthermore, most of them fall into a very limited, and privileged, selection of protein families (Figure 1.2). These privileged targets (e.g. kinases, GPCR, ion channels) are the 'comfort zone' of drug discovery. Their mechanisms of action, catalytic sites and predisposition to bind certain chemical families have been thoroughly studied. However, selectivity has been a major limitation of these target families, which accentuates the need of alternative strategies.[27,28]

Although most drugs acting on privileged targets use orthosteric binding sites (i.e. the regions where endogenous ligands and substrates bind), alternative (non-orthosteric) binding sites are very common.[29] Binding to non-orthosteric sites does not necessarily imply having a modulatory effect, however the ubiquity of these sites[30] suggest many new therapeutic opportunities (i.e. obtaining the therapeutic effect through external factors with new drug modalities such as protein-protein interaction modulators).[31] Drugs that modulate a target's activity through non-orthosteric binding sites are referred to as allosteric drugs. The modulating mechanisms of allosteric drugs are very diverse, but usually entail an information transfer through the allosteric site to the active site. Alternatively, they might also have a more direct effect on other aspects of the target (e.g. formation of protein-protein complexes) that ultimately also compromise the target's activity.[32] Thus, allosteric drugs do not compete with natural substrates and are generally more resilient to resistance-inducing mutations in the orthosteric site. Moreover, contrary to orthosteric binding sites, allosteric sites are not as highly conserved, which enables a higher selectivity for allosteric drugs.[33]

Despite allosteric drugs providing extra therapeutic opportunities to common targets there is still a large majority of unused genes in drug discovery. Most of these genes are still deemed undruggable due to difficulties in designing binders (e.g. flat contact surfaces, disordered regions).[34] Due to their clinical relevance, some of these targets have continued being the focus of unrelenting research efforts. Recently KRAS, one of the most frequently mutated oncogene proteins which was deemed undruggable, had its first inhibitor approved by the FDA.[35] This milestone emphasizes the lack of appropriate tools to drug the undruggable targets we have had until now, as opposed to

their fundamental *undruggability*.[36] Further research will prove that undruggable targets were really 'yet to be drugged'.[34]

Some of the undrugged targets still have an unknown structure or biological relevance. These are part of the so-called 'dark' proteome.[37] Several computational efforts such as AlphaFold,[38] have helped shrink the structural unknowns from 26 to 10% of the predicted human proteome. However, the pool of proteins with unknown function is practically untapped.[38] These numbers are even higher when referring to viral and bacterial proteomes. Moreover, on top of the unknown proteome genes, structural isoforms and proteins from alternative splicing and post-translational modifications remain poorly characterized.[39] Insights into the structural and functional properties of these elusive targets can enhance our understanding of disease mechanisms and allow for new therapeutic targets.

Finally, the surge in sequencing capabilities has highlighted the large variability in our proteome. This structural variability affects the patient's response to a given treatment both through direct and indirect mechanisms. On the one hand, drug-target interactions vary for each variant. This can potentially be exploited through the optimization of drugs' affinity towards specific genetic variants. On the other hand, in a more indirect manner, genetic variability might also affect the pharmacokinetic profile of a drug. Thus, ADMET parameters can potentially be highly personalized to improve the patient's response to treatment. Overall, understanding target variability holds the potential of maximizing treatment outcomes while minimizing adverse effects, ultimately leading to more successful therapeutic interventions.[40]

1.2.4 Translation from pre-clinical data to clinical results

Despite the efforts to increase the pool of therapeutically available targets in the early discovery stages, many of them fail to translate their promising results in preclinical setting to the clinical trials. More than 90% of drugs entering phase I do not reach the market.[41] Meta-analyses of clinical trial data show that most of those failures come from a lack of clinical efficacy (40-50%), high toxicity (30%) and low bioavailability (10-15%).[14] This raises the question whether certain aspects of drug development are being overlooked. Firstly, true target validation is still very challenging. Many *in vitro* and animal disease models fail to correctly represent diseases; due to their multifactorial nature, physiological and cell compensatory mechanisms, patient variability, etc.[42] New advances like the *organs-on-a-chip*[43] and proteomics[44] aim to better model the disease and the drug effect. On another note, during drug optimization, lead compounds are extensively optimized in terms of potency and drug-like properties, however other aspects such as tissue exposure or binding kinetics are still lagging behind.[45] Secondly, wrong toxicity predictions (both on-target and off-target), are the result of the simplification of the testing models used. Moreover, the accumulation of drug in vital organs or blood cells is one of the major factors for toxicity, against which there are still not well-developed strategies.[46] However, the growing AI strategies are promising prospects in this direction.[47]

1.2.5 New drug modalities

The expansion of the target space and the discrepancies between research and clinical results has emphasized the need for a deeper understanding of the biological pathways, disease models, and the

possibility to interrogate those new challenging targets with novel chemical modalities beyond traditional small molecules. The last years have represented an inflection point for many new chemical modalities (e.g. RNA-based approaches, next generation peptide, proximity-induced pharmacology). After several decades in basic research, many are demonstrating their effectiveness by reaching clinical stages.[48]

RNA-based drugs represent another novel disruptive therapeutic modality. This modality involves the silencing of disease-causing genes through complementary oligonucleotides and small interfering RNAs.[49] Despite their *long* history, their implementation has been hindered by challenges in their intracellular delivery, stability and triggering of the immune response. A crucial milestone with a lipid formulation of siRNA has expedited the approval of new drugs (e.g. patisiran[50], givosiran[51]) and solidifies the prospects of this new modality.

Mimicking native protein-protein interactions, peptide drugs are often very selective and potent. At the moment, more than 60 peptides have been approved and more than 140 are in clinical trials.[52] However, peptides have generally poor plasma stability and membrane permeability, which hinders their application against intracellular targets. To address it, a new generation of peptides (e.g. cyclic peptides, bicyclic and cysteine-rich scaffolds) and peptidomimetics (e.g. foldamers, peptide macrocycles) and peptide conjugates (e.g. conjugation with lipids, antibodies or small molecules) have emerged.[53]

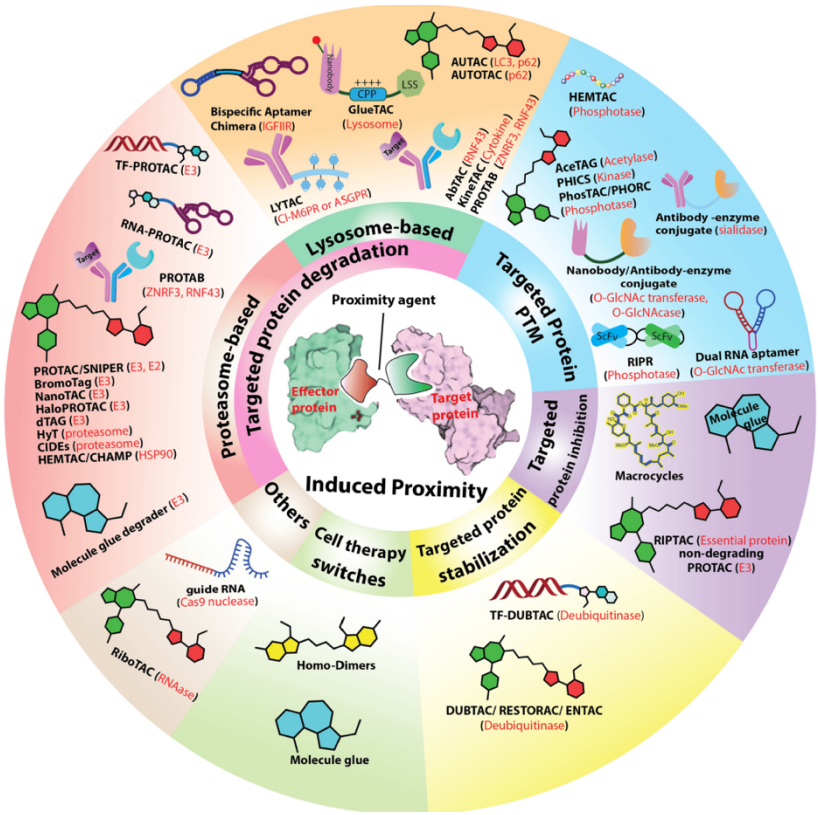


Figure 1.3. Mechanisms and modalities of proximity-induced pharmacology. In the inner circle, the mechanisms are highlighted and in the outer ring there are examples of the respective modalities. Red words in parenthesis correspond to the effector of the PIP modality, whilst in black is represented the name given to the proximity inducing small-molecule. Extracted from [54]

Interestingly, on top of their alternative mechanisms, these new chemical modalities defy the traditional definition of druglikeness (i.e. rule of five). Druglikeness was conceived as a set of property raanges that would ensure a drug's bioavailability and success in the clinics. However, molecules beyond the druggability ranges have proven *in vivo* efficacy and biodistribution.[55] *In silico* methods have given insights in understanding and quantifying underlying mechanisms that allow it (e.g., chameleonicity, NH-π interactions).[56]

The small-molecule modality is also venturing into new spaces and mechanisms of action, such as beyond rule of 5 compounds,[57] covalent targeting,[58] and Proximity-induced pharmacology (PIP).[54] PIP is based on hijacking native cell machinery to introduce post translational modifications or change cell location to specific targets of interest. The first PIP strategies were pioneered in the early 2000s, with the first peptide-based protein degraders.[59] These degraders induce the ubiquitination of the proteins of interest by enabling or inducing their contacts to the ubiquitin-proteasome system. Nowadays, the field has exploded with tens of protein degraders being advanced into clinical trials.[60] Following the lead of targeted-protein degradation (TPD), a plethora of other proximity induced mechanisms has emerged (Figure 1.3). As relevant examples, AUTACs (autophagy-targeting chimeras) induce intracellular degradation independently of the proteasome system,[61] PhosTACs (phosphorylation-targeting chimeras) control protein phosphorylation which is highly relevant in the activation of enzymes and transcription factors,[62] DUBTACs stabilize protein stabilization by deubiquitination of the targets[63], etc. Overall, PIP strategies have made available an unprecedented spectrum of biological mechanisms to develop new drugs.[64]

1.3 Computer-aided drug discovery

The use of computation to assist drug discovery (CADD) has been present for decades and has been through several cycles of hype and disillusionment.[65] However, over the years, computational technologies have become an integral part of the drug discovery process. Enabled by recent advances in other fields such as structural biology (e.g. cryo-EM), combinatorial chemistry, GPU and

cloud computing or machine learning algorithms, the quality of the computational models and their predictions has improved. Nowadays, computational approaches in drug discovery have become a key driving force for drug discovery.[66] CADD aims mainly (but not exclusively) to ease the experimental burden of the early discovery and preclinical phases. CADD approaches have been classified in a variety of manners: depending on the nature of data employed, depending on the complexity of the models employed or the drug discovery phase they are involved with.

Using the first criteria we can distinguish two main methodologies: ligand-based (LBDD) and structure-based drug discovery (SBDD). They are differentiated by the need of information about the target's structure. In the first iterations of CADD, LBDD was predominant due to the lack of good sources of structural information available. Quantitative-structure activity relationship (QSAR) studies were the flagship of ligand-based approaches.[67] QSAR models establish statistical relationships between descriptors obtained from the ligands' structure and experimentally measured properties. Despite the many improvements in the chemical descriptors and regression algorithms, finally QSAR failed to deliver on the inflated expectations it created.[68] The main reasons being activity cliffs, which break the assumption of a continuous structure-activity landscape, and the limited scope of the models.[69] Nowadays, a new era of LBDD has started with machine learning and other artificial intelligence approaches. Still, most studies seem to indicate that improved performance is incremental rather than transformative, as the fundamental limitations still apply.[70] Thanks to the simpler nature of their inputs, ligand-based methods are generally very quick, but their reliance on pre-existing data has limited their application. Currently,

ligand-based AI implementations are being used extensively for ADMET predictions.[71]

On the other hand, SBDD approaches require structural data from the therapeutic targets. The limited amount and quality of structures coupled with the higher computational cost of the methods made the introduction of structure-based methods slower than the ligand counterpart. However, the above-mentioned advances in structural biology have improved dramatically the quality and completeness of the receptor structures available.[72] Furthermore, structure availability surged with the release of AlphaFold and AI predicted models.[38] Leveraging high-quality structural data, structure-based techniques mostly employ physical-chemistry principles to extract relevant information (e.g. conformational dynamics, binding free energies or catalytic barriers). Thus, they can be applied to a broader spectrum of steps in drug discovery: from druggability assessments for target validation, to virtual screening for hit discovery or binding free energy calculations during hit-to-lead. Structure-based methods are generally slower and computationally more demanding than ligand-based, but they can provide insights on more complex problems. In this thesis I focus on the development of structure-based methods and pipelines to undertake challenges from different stages of drug discovery.

During target validation, an accurate evaluation of the molecular mechanism of the disease is crucial to avoid the preclinical to clinical pharmacology mismatches. Molecular dynamics (MD) has been extensively used to assess the effect of pathological mutations and to propose the key components to restore their native function.[73–76] Moreover, ligandability or druggability assessment of the potential therapeutic targets have drifted from experimental research to

computational tools almost entirely.[77] In particular, mixed-solvent MD approaches have shown very good capabilities for predicting ligandable regions in targets and even anticipate the appearance of druggable cryptic sites.[78–81] These methods use mixtures of organic solvents as proxies to evaluate target-ligand interactions. Due to the ease, reliability and fast convergence of these simulations, mixed-solvent MD methods have all but replaced the experimental counterpart (i.e. multiple solvent crystal structures).[82]

The following step of the drug discovery process, i.e. hit identification, is the recipient of the most prominent SBDD pipeline: virtual screening (VS). High-throughput VS (HTVS) mirrors the very costly HTS experimental campaigns, by testing large datasets of compounds against a given target. In contrast to HTS where compounds are evaluated through simple assays, structure-based VS rely on molecular docking programs to generate binding pose hypotheses and assess their affinity through a scoring function. Usually, a small subset of compounds is chosen after VS. Since its inception, there have been an ever-growing number of VS papers reporting new methods and active compounds.[83] Despite the high reported hit ratios, VS campaigns still yield many false positives, which is accentuated by the low amount of ligands typically chosen to test experimentally in academia. Moreover, the success of VS campaigns is highly dependent on the expertise of the researchers behind them.[84] To address this issue, VS platforms are incorporating more rigorous calculations, such as free energy calculations, to postprocess docked ligand poses.[85] However, the computational cost of the most accurate free energy calculation methods is incompatible with the needed throughput of VS. In light of these considerations, there is a need to develop post-docking ranking

algorithms with a reasonable tradeoff between speed and accuracy.[86]

On the other hand, the amount of molecules to consider reduces significantly on the hit-to-lead and lead-optimization phases. Hence, intensive calculations become much more adequate.[87] During the hit-to-lead, the affinity of hits is optimized. Thus, it requires a deep understanding of the underlying binding free energy and how it changes across a series of analogues. Free energy methods can be broadly classified in two groups: those introducing geometrical transitions (describing different positional, orientational or conformational states) and those introducing alchemical transformations (bridging potential energy functions through non-physical intermediates). In both cases, accurate predictions become increasingly challenging as the difference between the initial and end states increases. For this reason, predicting absolute binding free energies by either association/dissociation simulations (geometrical transition) or by creation/disappearance of the ligand (alchemical transformation) remains too computationally intensive. Short transitions, as done in the Dynamic Undocking method (see chapter 5), and small transformations, as in relative binding free energy (RBFE) calculations between highly similar compounds carry less uncertainty.[88] When introducing larger modifications it might be necessary to plan networks of smaller transformations (e.g. RBFE perturbation networks) to reduce the error through path redundancy.[89,90]

Binding affinity is not the sole determinant of a drug's success. In the past decade, the role of protein-ligand kinetics for drug research has gained recognition. Recent studies have postulated that drug-target

residence time ($1/k_{off}$) might be a better predictor of *in vivo* efficacy than binding affinity.[91,92] The lack of efficacy is one of the major attrition factors in drug discovery.[93] A number of examples have been presented where drugs with short binding pocket occupancies show worse pharmacological properties compared to drugs with similar potency but longer residence times. This motivates the assessment of scaffolds with the desired binding kinetics profiles to complement affinity-directed hit-to-lead campaigns. Studies in this direction have increasingly included computation of kinetic rate constants using new simulation techniques (e.g. Brownian dynamics[94], metadynamics[95], Markov state modeling[96], random accelerated MD[97] or steered MD[98]). As noted above, the magnitude of the transition makes these calculations very challenging and computationally demanding.

1.4 Current challenges in CADD

Overall, the presence of CADD methodologies is well established across the different stages of drug discovery. Nevertheless, the introduction of the big data paradigm has brought new challenges and opportunities that need to be addressed.

1.4.1 Adapting to the era of big data

The exponential increase of data generation has affected all areas of drug development, from a genomic understanding of diseases, to molecular pathways, safety concerns, clinical trial outcomes, etc. However, its transformation into knowledge is not trivial. Accompanying the big data, algorithms such as machine learning (ML) and deep learning (DL) have also been heavily promoted. In SBDD, the AlphaFold breakthrough towards solving the long-

standing problem of protein folding marked the start of the trend of knowledge-based approaches to solve most scientific investigations.[99] Still, AlphaFold was not directly trained with X-ray diffraction patterns, electron microscopy images or magnetic resonance spectra, but from curated three dimensional models stored in the PDB database.[100] This example emphasizes the role of data quality and curation and the development of tools to extract information from data to capitalize on the big data trends (Figure 1.4).[101]

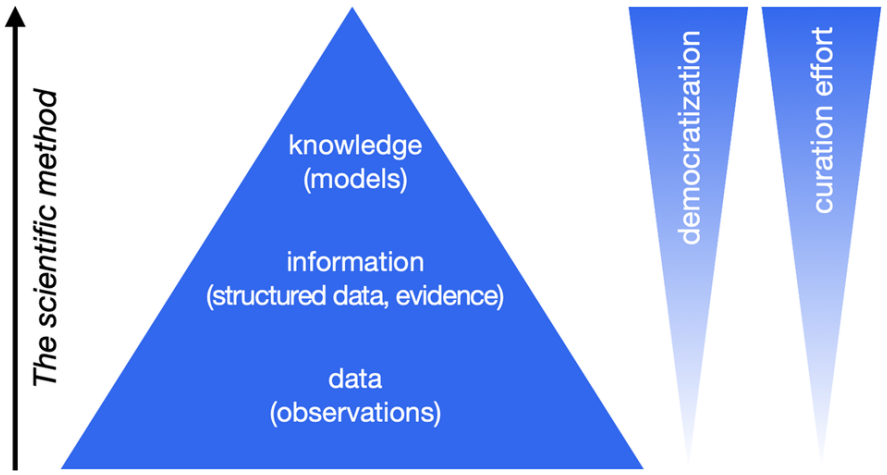


Figure 1.4. Data-to-information-to-knowledge hierarchical process and its correspondence with curation and data democratization. Bottom-up, following the scientific method process data is transformed into information which supports the creation of knowledge. This is accompanied by a curation effort which allows for the divulgation and democratization of the knowledge generated. Extracted from [101]

Similar to what happened with QSAR, more complex models, which inevitably become black boxes, are believed to have the top predictive performance.[102] In opposition, several authors have highlighted the need to pursue the use of interpretable models[103,104] and to keep realistic expectations in order to safely integrate knowledge-based algorithms to the CADD repertoire.

The advent of big data is not only confined to AI applications. The size and length of molecular dynamic simulations has also grown exponentially. Through intensive utilization of supercomputers, molecular modeling has scaled up three-fold every decade, even surpassing the Moore's law ratio (Figure 1.5). Nowadays, it is standard to see μs-long simulations of multi-protein complexes, and the timescales only become larger when incorporating multiscale approaches.[105] These simulations aim to bridge the experimental and computational time-scales, and can provide insights on molecular mechanisms of slow and complex processes (e.g. folding of chromatin fibers,[106] virion assembling[107]). Still, most life science problems are proposed with a more reductionist approach, limited to single macromolecule systems. Moreover, advances in the application of statistical thermodynamics to enhance the sampling have proven to already reproduce experimental time-scales. Thus, there is a relevant demand for new algorithms and pipelines to fully profit from the new scope of the simulations, outside of brute-forcing.[108]

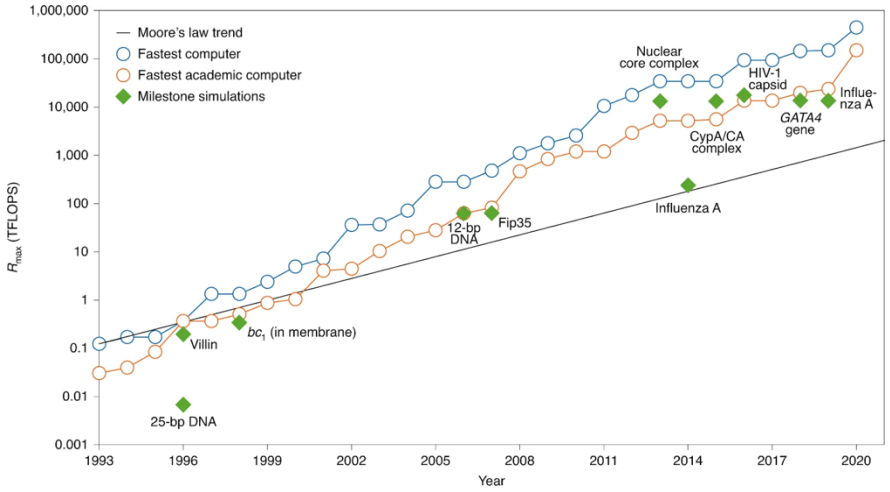


Figure 1.5 Yearly evolution of computational power and milestone simulations. The computational power, expressed in TFLOPS (i.e. trillion floating point operations per second), of the fastest overall and academic computers is represented in blue and orange respectively. The highlighted milestone simulations are the following: 25-bp DNA (5 ns ~21.000 atoms)[109], Vilin (1 µs ~12,000 atoms)[110], bc_1 membrane complex (1 ns ~91,000 atoms)[111], 12-bp DNA (1.2 µs ~16,000 atoms)[109], Nuclear complex (1 µs ~15.5 million atoms)[112], influenza A virus (1 µs >1 million atoms)[113], CYpA/CA complex (100ns ~25.6 million atoms)[114], HIV-1 solvated capsid (1 µs ~64 million atoms)[115], GATA4 gene (1ns ~ 1 billion atoms)[106], and H1N1 influenza virus (120 ns ~160 million atoms)[116]. Extracted from[117].

Big data and ML are also having an impact on the exploration of chemical space. Generative approaches aim to enable efficient exploration of the vast number of possible molecules (estimated at 10^{40}-10^{60} for drug-like molecules,[118–120] identifying those meeting the desired criteria. Like the *De Novo* methods in the 1990s, generative models may suffer from difficult transition to the bench owing to poor synthetic feasibility of the predicted compounds.[121] Exploration of large chemical spaces is also possible by other means. Experimentally, DNA-encoded libraries (DELs) offer the possibility of screening up to 10^{12} compounds.[122] Finally, ultra-large on-demand collections are also attracting a lot of attention in the field of virtual

screening. These collections are the successors of combinatorial chemistry research from the 20th century.[123] Initially, these collections were dismissed due to past failures and diversity concerns yet, pioneered by Enamine in 2015, a new wave of large, diverse and synthetically available collections emerged.[124] Today, they contain more than 10^{11} compounds and will reach the trillion scale very soon, growing towards the theoretical accessible chemical space. These chemical collections have the potential to revolutionize the drug discovery process, delivering hits that are closer to the final drug, even for the most difficult targets.[125] The size of such collections has surpassed the throughput of traditional SBDD pipelines. Thus, new efficient strategies to navigate and evaluate the chemical space are highly needed.[126]

1.4.2 Data and software democratization

The application of big data to CADD is limited by the fact that most of the (positive and negative) activity, toxicity, synthesizability, etc. data is still proprietary. This has obvious justification from pharmaceutical companies, which have held historical records of such data. Nevertheless, the improvement of data hungry algorithms such as ML is also beneficial to such companies. In that direction, some companies are starting to release clinical data,[127–129] but sharing of preclinical data is so far lagging behind. Solely sharing big amounts of data is far from enough to help democratize drug discovery. For relatively 'simple' kinds of data (e.g. gene sequences), standardization is often enough.[130] However, more complex observations need new data formats and routines to extract the underlying information (e.g. mmCIF/PDBx to share crystallographic information).

The availability of such algorithms gatekeeps the proper utilization of the data and the generation of knowledge.[131] Open-source research started in the software industry with the development of operating systems (i.e Linux) alternative to Microsoft's. Now, it has also expanded to the biological and pharmaceutical research fields. Initiatives like OpenMM[132] and OpenForcefield[133] have played a crucial role in advancing computational chemistry by providing researchers with powerful tools for simulating molecular dynamics and optimizing force fields. The FAIR (findable, accessible, interoperable and reusable) guidelines are the consensus framework established for the proper functioning of Open Science.[134]

The open-driven projects have extended to whole drug discovery campaigns. Open-science for drug discovery is not a new concept but, owing to the urgency created by the global pandemic, new collaborative efforts to accelerate drug discovery have resulted in big initiatives such as the Moonshot consortium.[135]

CHAPTER 2: OBJECTIVES

2. OBJECTIVES

2.1 General objective

The recent advances in drug discovery have introduced many opportunities to solve previously unattainable medical needs. Nonetheless, these opportunities challenge the *status quo* of the methodologies currently in use. Computational tools can help alleviate these challenges, and have been steadily growing in reliability and importance in drug discovery. This PhD thesis aim to contribute to these efforts by developing computational tools and pipelines to address current challenges in drug discovery such as expanding the target space, improving hit discovery or evaluating protein-ligand complexes.

2.2. Specific objectives

Objective 1. Proximity induced pharmacology, and more specifically TPD strategies, have emerged as a new drug modality allowing to modulate targets previously considered undruggable. In particular, E3 ligases are a key component in the ubiquitin-proteasome system to trigger specific protein degradation. However, very few E3 ligases have been exploited for TPD purposes. The aim of this project is to expand the available toolbox of E3 ligases, evaluating their ligandability through computational means. In addition, another aim of this work is elucidate the mechanism of action of previously found FBW7 allosteric modulators discovered during the prospective validation of the ligandability study.

Objective 2. HTVS has become an established methodology to Identifying new hits for drug discovery programs. Moreover, the appearance of ultra-large on-demand collections, that expand the available chemical space, offer the possibility of finding hits in a higher ratio and with higher potency. Yet, these libraries have outgrown the throughput of traditional HTVS strategies. Thus, there is a need to develop effective strategies and computational tools to navigate and evaluate those spaces. We aimed to develop a pipeline to efficiently navigate the ultra-large chemical collections and to select the best hits within them.

Objective 3. Evaluating the binding free energy of protein-ligand complexes in a high-throughput manner is still a highly demanded, but unresolved, ambition in computational chemistry. The current gold-standard protocols are highly accurate but have a high cost, thus making them inadequate for tasks requiring high-throughput. In this regard, I aim to delimitate the applicability of DUck, a fast SMD protocol, to evaluate relative binding free energies and activity cliff prediction. Moreover, to spread the use of the method and to facilitate the development of new applications for drug research, I aim to develop an open-source implementation of Duck and, in parallel, to evaluate the use of the Crooks fluctuation theorem for the calculation of free energies from non-equilibrium simulations.

CHAPTER 3:
METHODS AND PROTOCOLS

3. METHODS AND PROTOCOLS

All computational methods rely on models that strive to approximate reality as closely as possible. Usually, the choice of the method depends on the complexity of the prediction, throughput, or even economic needs. Despite each having its own pitfalls, these approximations are very powerful and let us explore the chemical, conformational and energetic landscape of our systems of interest. By incorporating methods of increasing complexity in our protocols we can mitigate the errors introduced by our models and make reliable predictions. In this first section, I will expound an overview of the methods used in this thesis, followed by an explanation of the specific protocols implemented to generate the predictions on target ligandability, ligands' binding poses and free energies, and their mechanism of action.

3.1 Methods

The methods are presented in three sections: chemoinformatics, molecular docking and molecular dynamics. This shows the typical hierarchical order of protocols implemented in this thesis, where simple and cheap methods are followed by more complex and costly ones.

3.1.1 Chemoinformatics

The first step to modeling the chemical reality is to find a representation of molecules that fulfills our specific needs. Generally, molecular representations aim to be standardized, easily interpreted by a computer, compact, and ideally human readable. To

date, several representations have been proposed and are still actively used: 2D or 3D molecular graphs, SMILES, fingerprints or bitstrings, among others. These representations serve as fundamental tools for storing, analyzing, and interpreting chemical information.

The use of molecular descriptors enables the manipulation and analysis of chemical structural information. Lists of such descriptors, stored in either binary or continuous variables, are called fingerprints. These descriptors are stored as numerical values that characterize properties of molecules, either intrinsic, like physicochemical or topological properties, or external, like biological activity or toxicity. Fingerprints are one of the main tools in chemoinformatics, and are used extensively to compare molecules (molecular similarity) and make predictions (QSAR models, machine learning, etc).[136] The most commonly used fingerprints are MACCSKeys,[137] Atom Pairs,[138] Extended connectivity (ECFP),[139] and Morgan (or circular) fingerprints.[140]

Molecular similarity is used in many applications crucial to this thesis, namely navigating chemical libraries, clustering or as a defining characteristic of activity cliffs. There are many similarity indexes, with Tanimoto being the most frequently used. Tanimoto is calculated as the ratio of shared features to the total number of features. One reason for the popularity of the Tanimoto coefficient for similarity searching is that it includes a degree of size normalization. We have also used indexes like the Dice coefficient, cosine or euclidean distances. Whilst Tanimoto is the most commonly used coefficient in chemoinformatics, the other similarity indexes have their specific usage (e.g. cosine distance is most optimal in sparse data).[136]

As one of the most human readable and compact representations, SMILES is a text-based notation system for representing chemical structures using ASCII strings.[141] When canonicalized, SMILES can be used as a universal identifier for specific chemical entities and that has made them very widely used in virtual compound databases like the ones used for virtual screening in this work. As an alternative to similarity, we might want to navigate molecule databases looking for specific substructures within them. SMARTS is a language derived from SMILES that allows to specify substructures using rules, similar to wildcards or regular expressions.[142] SMARTS are also very suitable to define general scaffolds shared by a group of molecules (i.e. maximum common substructure (MCS)).

3.1.1.1 Fragment spaces and SpaceMACs

In-silico molecular databases have been growing exponentially and, despite its high compactness, the SMILES format is showing its limitations. They lack the compactness needed to effectively store and search billion-sized libraries through current algorithms. On the other hand, graph-structured databases are able to exploit the combinatorial nature of the ligands in such databases (i.e. many compounds are synthesized using different combinations of the same building blocks and reactions).[124] The graph nature of combinatorial fragment spaces, where nodes are building blocks and edges reactions, allows for a very compact format, where molecules are only enumerated upon extraction. This strategy also opens the door for many possibilities for intellectual property generation without limiting their distribution.[143]

The compartmentalized structure of graphs has allowed for the development of very efficient tools to navigate the fragment spaces,

both using similarity[144] and substructure searches.[145] In the work presented here, we have extensively used SpaceMACS to perform substructure searches in make-on-demand combinatorial fragment spaces. This algorithm was published in 2022, but we used an early development implementation that was generously shared by the group of Prof. Matthias Rarey in 2021.

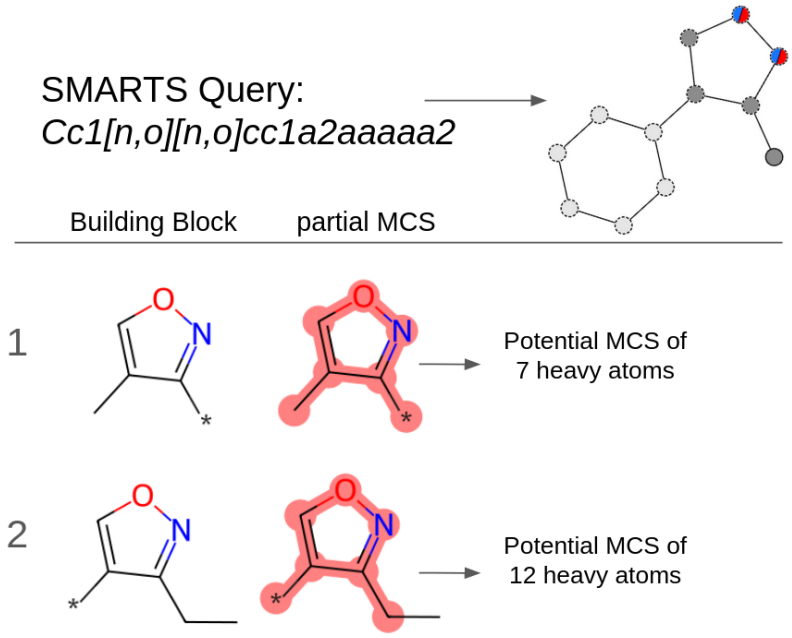


Figure 3.1. Example of the tree prioritization of SpaceMACS. Given a query SMARTS, two similar building blocks can have different potential MCS coverage depending on the position of the linker atom (represented with an asterisk in the 2D representation of the building blocks).

When using SMARTS for substructure search within fragment spaces, a first MCS with the individual building blocks is calculated. If this initial MCS matches the complete pattern, then the search becomes trivial as all the ligands derived from that building block will be a match. On the other hand, if the MCS covers only partially a given building block it is given a score according to its potential to complete the MCS. Atoms in the partial MCS are weighted

depending on their participation in a building block atom or linker using the RIMACS algorithm.[146] These combinatorial trees are disentangled depending on their potential coverage of the MCS (figure 3.1). Finally, the resulting matches are enumerated, filtered, and sorted based on the size of the MCS and the actual size of the ligands.

3.1.2 Molecular Docking

Molecular docking aims to describe the interaction between two molecules. These binding partners usually involve a protein (receptor) that can bind molecules of different nature (ligands), namely small-molecules, peptides or proteins, and nucleic acids. In this thesis, I have focused on the exploration of the binding of small molecules to proteins of interest.

To accurately model the tridimensional shape of both protein and organic molecules, we employ molecular mechanics. The molecules involved are treated as collections of balls (atoms) connected by springs (bonds), and their interactions are described using empirical force fields. These force fields assign parameters to different types of atoms and bonds, as well as to non-bonded interactions such as Van der Waals forces or electrostatic interactions (see Section 3.1.3). Molecular docking programs have two main routines: (i) pose generation, in which various conformations or orientations of the ligands are generated with respect to the receptor, and (ii) pose scoring using specific scoring functions, which determines the energy associated for each binding pose and uses these energy terms to rank the most favorable conformations and ligands.

Due to the high throughput expected for docking programs, they do not explore the conformational landscape of the protein-ligand complex effectively. To alleviate this shortcoming, different algorithms have been implemented,[147] however they are still very inefficient without further guidelines. The usual strategies to effectively guide docking calculations are the limitation of the region of the receptor where the small molecules bind (i.e. cavity definition) and the enforcement of specific interactions between the receptor and the ligands (i.e. pharmacophore definition).

Even the most efficient conformational sampling would be meaningless without a way to rank the most promising conformations. Thus, molecular docking programs use scoring functions to differentiate the stable binding modes, observable experimentally, from the transient ones. On top of that, these scoring functions should also distinguish between good and bad binders. There are three main groups of scoring functions: theoretical, empirical, and a hybrid group of theoretical functions adjusted to fit empirical data. Theoretical functions estimate the free energy using the bonded (energy strain on the ligands' conformation) and non-bonded terms (usually the intermolecular interactions between the ligand and the protein) from the force field.[148] Empirical scoring functions use statistical modeling, representing the binding modes through different descriptors, to reproduce known structures.[149] Finally, theoretical functions can be reweighted to fit experimental results trying to harness the power of both modes, correcting the imperfections on the force field but keeping a physical meaning to the score.

3.1.2.1 rDock

All the docking procedures presented in this thesis have been performed with the rDock software package. rDock was originally developed by RiboTargets (subsequently Vernalis (R&D) Ltd) between the years 1998 and 2006.[150] In 2012, the program was released as open-source, and its maintenance and further development has been led by the Barril's group since then with a second publication in 2014.[151]

To generate the binding poses, rDock uses a concatenation of three stages of a genetic algorithm followed by two minimization algorithms: a low temperature Monte Carlo and a Nelder-Mead's Simplex.[152] The genetic algorithm applies *'mutations'* following a uniform distribution across the molecule's *'chromosome'*, which is defined by the position (center of mass), orientation (Euler angles for the principal axes) and conformation (dihedrals). New generations from each mutated line are populated until there is no gain in score. The scoring function we use in rDock is a theoretical function, using physical properties taken from a modified version of Sybyl 5.2 force field, re-weighted to fit an experimental dataset that was prepared during the original implementation (equation 3.1).

$$S^{total} = S^{inter} + S^{intra} + S^{site} + S^{restraint} \quad (3.1)$$

$$S^{inter} = W^{inter}_{vdw}\cdot S^{inter}_{vdw} + W^{inter}_{polar}\cdot S^{inter}_{polar} + W^{inter}_{repul}\cdot S^{inter}_{repul} + W^{inter}_{arom}\cdot S^{inter}_{arom} + W_{rot}\cdot S_{rot} + W_{constant} \quad (3.2.a)$$

$$S^{intra} = W^{intra}_{vdw}\cdot S^{intra}_{vdw} + W^{inter}_{polar}\cdot S^{inter}_{polar} + W^{intra}_{repul}\cdot S^{intra}_{repul} + W^{intra}_{dihedral}\cdot S^{intra}_{dihedral} \quad (3.2.b)$$

$$S^{site} = W^{site}_{vdw}\cdot S^{site}_{vdw} + W^{site}_{polar}\cdot S^{site}_{polar} + W^{site}_{repul}\cdot S^{site}_{repul} + W^{site}_{dihedral}\cdot S^{site}_{dihedral} \quad (3.2.c)$$

$$S^{restraint} = W_{cavity}\cdot S_{cavity} + W_{tether}\cdot S_{tether} + W_{nmr}\cdot S_{nmr} + W_{ph4}\cdot S_{ph4} \quad (3.3.d)$$

Both the genetic algorithm and the scoring function can receive additional restrictions to guide the generation and ranking of ligand binding poses. The cavity definition can be done using two procedures: the two spheres method and the reference ligand method. The first method uses two different-sized spheres to define a cavity that is only accessible by the small spheres but not the large spheres. The second method selects the volume of a given radius around a binding mode of a reference ligand. Penalties are applied to ligands occupying areas outside the cavity. In addition, information on the required molecular interactions can be included as pharmacophoric restrictions. These are given as regions in the space, which require specific groups (i.e. coordinates, tolerance radius and the nature of the group such as H-bond acceptor, donor, a cation, etc.). These pharmacophoric requirements will influence both the binding pose initialization during the genetic algorithm and the scoring function. Finally, spatial and conformational restrictions can be applied to the ligand in what we call tethered docking. Each one of the ligands in our library is aligned to a reference compound using their MCS. Then, the selected atoms remain fixed in place during the genetic algorithm mutations.

3.1.3 Molecular Dynamics

Molecular dynamics (MD) enables us to describe the motion of molecules using physical principles from classical mechanics to simulate time-dependent behavior of molecules at the atomic level. Newton's second law relates the acceleration (*a*) of a body with a given mass (*m*) to the force (*F*) acting on the body.

$$F = m \cdot a \tag{3.3}$$

As the acceleration is the second derivative of the positions (*r*) to time we can write simple equations of motion, where v is the velocity.

$$\frac{dr}{dt} = v \qquad (3.4.a)$$

$$\frac{dv}{dt} = \frac{F}{m} \qquad (3.4.b)$$

The motion equations can be formulated in many different ways. The Hamiltonian formulation, equivalent to the equations above, depends on a function called Hamiltonian (*H*) that describes the total energy of a system given the positions (*r*) and its momentum (*p*).

$$\frac{dr}{dt} = \frac{\delta H}{\delta p} \qquad (3.5.a)$$

$$\frac{dp}{dt} = \frac{\delta H}{\delta r} \qquad (3.5.b)$$

The Hamiltonian consists of two main components: the kinetic energy (*T*), which depends on momentum (*p*), and the potential energy (*V*), which depends on their positions (*r*).

$$H(r,p) = T(p) + V(r) \qquad (3.6)$$

Whilst we can calculate the kinetic energy directly from the momenta (equation 3.7), the potential energy (equation 3.8) is approximated using force fields with parameters obtained experimentally or fitted from quantum-mechanics calculations.[153] Several force fields have been developed for MD simulations of biomolecular systems such as AMBER,[154] CHARMM,[155] GROMOS,[156] OPLS.[157] Due to their large chemical diversity, small molecules usually require specific parametrization, or the use of generalizable force fields such as

GAFF[158] or the force fields from the OpenFF initiative.[133] There are two terms in the potential energy function; the bonded terms (bonds, angles, dihedrals) and the non-bonded terms (van der Waals and electrostatic).

$$T = \frac{1}{2}\sum_{i}^{N}\frac{|p_i|^2}{m_i} \tag{3.7}$$

$$V = V_{\text{bonded}} + V_{\text{non-bonded}} \tag{3.8.a}$$

$$V_{\text{bonded}} = \sum_{\text{bonds}}\frac{K_r}{2}(r - r_0)^2 + \sum_{\text{angles}}\frac{K_\theta}{2}(\theta - \theta_0)^2 + \sum_{\text{dihedrals}}\frac{V_k}{2}[1 + \cos(n\theta - \gamma)] \tag{3.8.b}$$

$$V_{\text{non-bonded}} = \sum_{\text{pairs}}\left[4 \in \left(\left(\frac{\delta}{r}\right)^{12} - \left(\frac{\delta}{r}\right)^{6}\right)\right] + \sum_{\text{pairs}}\frac{q_i q_j}{4\pi \in_0 r_{ij}} \tag{3.8.c}$$

The equations of motion are typically integrated numerically using iterative algorithms (in steps of 1 to 4 ps), being the Verlet algorithm the most commonly used.[159] Despite the apparent deterministic nature of the formulas, the simulations we perform have several approximations that introduce stochastic variability (e.g. Thermal Fluctuations).

In MD, we often reduce the systems to their minimal expression (e.g. an isolated single protein in an aqueous environment), and then, use periodic boundary conditions (PBC) to emulate the large and complex *in vivo* conditions of our systems. The PBC simplify the system by defining limits (i.e. a simulation box) where molecules and interactions are propagated in all directions, hence simulating an *infinite* system. These approximations can create artifacts, specially

with long-range interactions or in non-symmetrical systems (membrane bilayers with osmotic gradients), but are crucial in providing conditions that closely resemble those found in biological contexts.

To connect our individual observations in the atomistic simulations to the macroscopic world, we use statistical mechanics. By setting different conditions in our simulations (known as ensembles), we can derive thermodynamic properties from the system. In this thesis I have used two ensembles, (i) the canonical ensemble (NVT), where the atoms (N), the volume (V) and the temperature (T) are constant, and (ii) the isothermal-isobaric (NPT) ensemble, where the volume is freed but the pressure (P) is kept constant.[160] In both ensembles, the system is allowed to exchange heat with the *outer space*. The temperature is kept constant by a thermostat, such as the Langevin thermostat, where the particles' movement (i.e. heat) is modulated using a stochastic force to increase it or a dampening effect to decrease it. In the NPT ensemble, we also require a barostat (e.g. Nosé-Hoover, Berendsen) to control the pressure of the system.

3.1.3.1 MDMix

Molecular dynamics (MD) simulations are a powerful tool that enables us to model the conformational space of molecules through their movement and can even obtain a dynamic equilibrium of the intra- and inter-molecular interactions of a given system. With enough conformational sampling (i.e. multiple binding and unbinding events) it is possible to extract a free energy associated with such interactions. Ideally, we would use MD to assess the binding free energy of our compounds during a virtual screening campaign, even without knowledge of the binding mode or even the binding site. Yet,

the simulation time to obtain such sampling in classical MD is beyond our reach.

Small organic solvents (e.g. ethanol, pyridine, acetamide), however, are smaller, diffuse much faster and can be used as a proxy to assess regions in the protein surface prone to interact with small molecules.[78] Information from the solvent/protein interaction surfaces can then be extrapolated to what would happen with bigger ligands. This concept has been used experimentally to identify ligand-binding sites through NMR and crystallography.[161,162] *In-silico* equivalent of such experiments are MD simulations with mixed organic/aqueous solvents (MDMix). pyMDMix is an open source python module designed to set up and analyze MDMix simulations developed by the Barril's group.[163] Other similar approaches (SILCS,[80] MixMD,[164] EXPRORER[165]) have also been developed and shown good results in binding site identification and ligand design.

The standard MDMix protocol involves solvating a prepared system in a truncated octahedron of pre-equilibrated aqueous/organic mixed solvents. Then a minimization in 5000 steps is performed using *sander*, followed by four heating steps from 100K to 300K in intervals of 50K in a total of 800 ps in the NVT ensemble. Before production, a density equilibration step of 1ns at 1 bar in the NPT ensemble with the Berendsen Barostat is performed. Production is separated in 1ns steps using the NVT. All simulations from the equilibration onwards are performed using the Langevin thermostat (with 4 collisions per ps) and with a timestep of 2fs, using the CUDA implementation of PMEMD in AMBER.[166] All protein non-hydrogen atoms are restrained with a soft harmonic potential of 0.01 kcal mol^{-1} $Å^{-2}$ to facilitate grid analysis.

Production simulation steps are superimposed on a reference structure and the density of the selected organic solvent groups is recounted using a grid format. Using the Boltzmann relation (equation 3.9) we can convert the observed density (N_i) into free energy by comparing it to the expected distribution (N_0) (i.e. the uniform distribution of densities given the organic solvent concentration). A volume correction is applied due to the high concentration of organic solvent with respect to standard state conditions (comparing the simulated volume (V_{sim}) to the expected volume at 1M).

$$\Delta G_{\mathrm{Bind}} = -RT\left(ln\frac{N_i}{N_0} + ln\frac{V_{sim}}{V_{1M}}\right) \qquad (3.9)$$

The energy grids generated can be visualized to help the identification of binding hotspots from the most energetic points. Alternatively to using the energy grids visually, the energy associated to such hotspots can be obtained by doing an exponential average of the top energy points (0.02 percentile) within a radius of the point.

$$\Delta G_{\mathrm{hotspot}} = -RTln\langle e^{\frac{\Delta G_i}{-RT}}\rangle \qquad (3.10)$$

3.1.3.2 Biased molecular dynamics

Limitations in sampling cannot always be addressed by using simplified systems as a proxy, as in MDMix. Often, the data obtained from simplified cases cannot be generalized to the systems of interest. In these situations, we bias the MD simulation to improve the sampling. These biases induce the appearance of rare events by introducing external forces along a collective variable (CV) (e.g.

metadynamics,[167] umbrella sampling,[168] steered molecular dynamics (SMD)[169]) or by employing altered statistical ensembles (e.g. parallel tempering,[170] replica exchange,[171] scaled MD[172]). CV are measurements taken from a system that aim to describe complex motions or processes of interest in a low-dimensional representation. Methods relying on CV benefit highly from previous knowledge of the possible states of the system through the process. On the other hand, when the methods are more exploratory the dependency on CV is a huge limitation, and altering the statistical ensembles becomes more suitable. Altogether, these methods help to characterize the thermodynamics of the studied processes by reconstructing their associated free-energy surfaces (FES).

The use of biased MD methods is not trivial and usually is computationally demanding. Thus, some of these methods forfeit the reconstruction of FES of the process, steering away from equilibrium in order to achieve a higher throughput. In 1997, Jarzynski proposed a very general relation between equilibrium free energies and non-equilibrium work.[173] His theorem states that the work (W) performed on a system during non-equilibrium simulations is larger than or equal to the free energy difference (ΔF) between the equilibrium end points of the simulated process (equation 3.11).

$$\langle W \rangle \geq \Delta F \qquad (3.11)$$

In an ensemble of non-equilibrium simulations, different work values are obtained leading to a statistical distribution of works. This happens due to the applied bias magnifying the stochastic thermal fluctuations of the system, which appear in the form of dissipative work (W_{diss}). This dissipative work becomes negligible in the case of a quasistatic process (i.e. a process that takes an infinite amount of

time to complete) and grows the faster the change happens. We can remove the bias by averaging over the works' exponential, which weights higher low dissipative work trajectories (eq. 3.12a) and obtains good estimations of equilibrium free energies (eq. 3.13.b).

$$\langle e^{-W(k_\beta T)^{-1}} \rangle = e^{\Delta F(k_\beta T)^{-1}} \qquad (3.12.a)$$

$$\Delta F = -k_\beta T ln \left[\frac{1}{N} \sum_{i}^{N} e^{-W(k_\beta T)^{-1}} \right] \qquad (3.12.b)$$

In addition, there are alternative approaches for obtaining free energy differences from non-equilibrium simulations (e.g. Crooks fluctuation theorem, Bennett's acceptance ratio). The Crooks fluctuation theorem relates the equilibrium free energy difference to the work distributions of the forward and reverse process.[174] With Bennett's acceptance ratio, we obtain the free energy difference by sampling distributions of simulations biased towards the process end-points.

In this work, I have used two out-of-equilibrium protocols during the realization of this work: Dynamic undocking (DUck) and Random Accelerated Molecular Dynamics (RAMD).

3.1.3.2.1 Dynamic Undocking

Certain H-bonds present strong opposition to structural perturbations and act as kinetic traps during ligand dissociation, hindering the transition between H-bonds to water-mediated bridges.[175] These H-bonds are generally buried in the receptor cavity shielded from solvation. Despite the local nature of the structural stability of water-shielded H-bonds, it has been a very effective descriptor to predict global properties.[176–178]

Dynamic Undocking, is a particular implementation of SMD where the work needed to break the main hydrogen bond in a protein-ligand complex is evaluated.[176] In the final geometry, the ligand is in a quasi-bound state and we calculate the minimum work needed to reach this state (W_{QB}) (Figure 3.2.b). The W_{QB} describes the steepness of the first barrier during the ligands' dissociation.

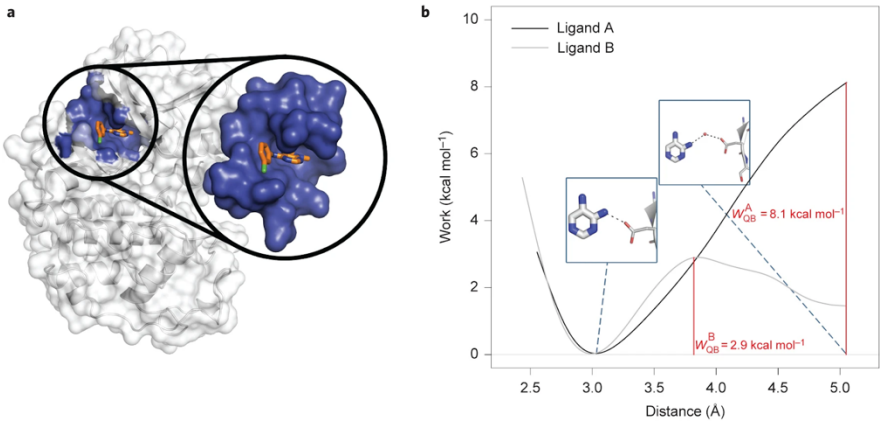


Figure 3.2. DUck chunk and work profile example. a) Example of a chunk containing only the local environment around a key H-bond. b) Representative work profiles of strong (ligand A) and weak (ligand B) binders obtained from DUck simulations. It shows how end points during the simulation are water-mediated interactions.

The standard DUck protocol starts with the definition of the H-bond to study. A subset of the protein receptor around the H-bond is selected, referred to as the chunk from now onwards (Figure 3.2.a). The chunk is created to reduce the computational cost of the simulations whilst maintaining the molecular environment of the interaction. Using a set of scientific vector language scripts (SVL, a programmatic language implemented in MOE[179]), the chunk and ligand are parametrized using AMBER ff99 and parm@frosst[180] respectively, obtaining the ligand's charges with AM1-BCC. The system is then solvated in a cuboid TIP3P water box, spanning 18 Å further than the last protein atom in each direction, and neutralized

using Na^+ or Cl^- when needed. Due to the disconnected nature of the chunk, protein heavy atoms will always maintain a harmonic restraint with a force constant of 1 kcal mol^{-1} $Å^{-2}$. The rupture of the key hydrogen bond during non-steered simulations is prevented with a flat bottom restraining potential for distances beyond 3 Å. The system is minimized for 1000 steps, gradually heated from 100 to 300 K for 400 ps in four steps of 50 K each in the NVT ensemble, and then the pressure brought to 1 bar during a 1ns simulation in the NPT ensemble. All equilibration and simulation steps are done using a Langevin thermostat with a collision frequency of 4 ps^{-1}. At this stage, we execute iterations (SMD cycles) of 0.5 ns non-steered simulations followed by two SMD simulations from the same restart but at different temperatures (300 and 325K). Each SMD consists of steering the H-bond from 2.5 to 5 Å at a constant speed of 5 Å ns^{-1} for a total of 500 ps with a spring constant of 50 kcal mol^{-1} $Å^{-2}$. A stopping criterion is set between SMD cycles using a minimum W_{QB} threshold (figure 3.3).

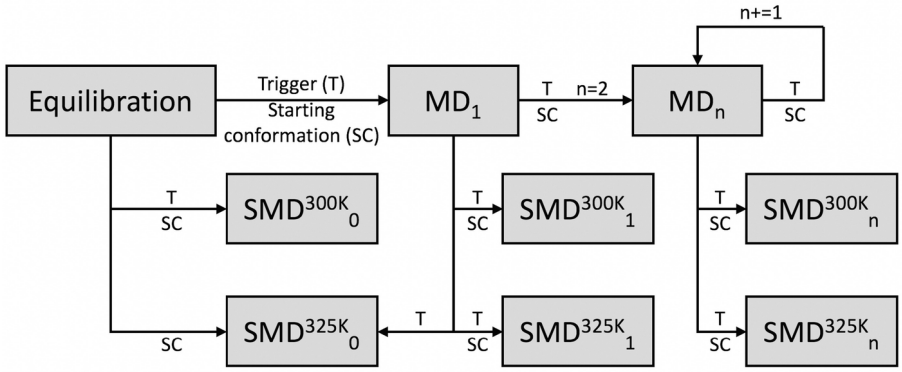


Figure 3.3. Scheme of DUck workflow. Extracted from Majewski et al. 2018.[181]

3.1.3.2.2 Random Accelerated Molecular Dynamics

RAMD is an out-of-the-equilibrium method to compare residence times of a series of ligands bound to the same protein receptor. It applies small randomly oriented external forces to expel a ligand from its binding site. Every couple of simulation steps the position displacement of the ligand's center of mass is checked. If the ligand has not been displaced enough, the orientation of the force is randomized again. Through iterations of this process, eventually the ligand dissociates. The τRAMD protocol uses RAMD to calculate relative residence times (τ) of protein-ligand complexes.[97] τ values are estimated from the distribution of simulation steps it takes for the ligand to dissociate. Such distributions are generated by bootstrapping a small pool of RAMD simulations. It is not possible to recover absolute τ from these simulations. Given the random nature of the applied bias, ligands are free to explore different dissociation pathways without the definition of CVs. Therefore, τRAMD is a very useful method to obtain insights on the dissociation pathways of ligand protein complexes.[182]

3.2 Protocols

Using the methods and models described above, I have compiled specific protocols to address the objectives set for each project. The protocols detailed below were all executed using resources from the Red Española de Supercomputación (RES) and in the Institut de Química Teòrica (IQTC).

3.2.1 Druggability assessment of E3 ligases

3.2.1.1 Mixed solvent molecular dynamic simulations

3.2.1.1.1 Compilation of E3 ligases dataset

We built a comprehensive database of all available E3 ligases with X-ray structure, including structural and functional data. Ubihub[183] was used to obtain the full list of E3 ligases and to classify them based on their involvement in protein degradation (UPS score). Then, the E3 ligases with available crystal structure were downloaded from the PDB database including relevant data such as the crystal quality (e.g. resolution, Free), coverage, and the presence of ligands. More precise data such as domains and their function was obtained from Uniprot[184] and PROSITE.[185] A representative PDB structure was chosen for each E3 ligase using the best resolution, coverage and the presence of the degron recognition domains (i.e. substrate recognition sites) as the selection criteria. A selection of the top E3 ligases was made, ensuring the presence of E3 ligases with known binders for retrospective validation and novel, undrugged E3 ligases for a prospective validation. A total of 23 structures for E3 ligases were selected.

3.2.1.1.2 Structure preparation of the E3 ligases for pyMDMix

The selected structures were then used as initial structures for pyMDMix. The protein preparation was performed with MOE 2016.[179] The adaptor proteins present in ASB9, CRBN, DCAF15, FBXO44, FBXW7, SKP2, SOCS2 and VHL crystal structures were kept in the system preparation in order to preserve the native E3 conformation. The systems were solvated using an octahedral solvation box of a pre-equilibrated 1:4 ethanol-water mixture. The

solvation box used for TRIM28 RBCC domain (PDB code 6QAJ) was changed to a prism box to optimize the analysis given the shape of the protein system. The amber ff14SB force field was used to parametrize CRBN, TRIM28 , PRKN and RNF4. The coordinated Zn^{2+} ions were parametrized using the semiempirical mode of the MCPB.py protocol[186] included in Ambertools 2017.[187]

3.2.1.1.3 Mixed-solvent molecular dynamic simulations and pocket identification

The standard pyMDMix[163] pipeline was carried out on the selected E3 ligases. The binding hotspots obtained for each of them were used to perform the pocket identification through clustering. Three 50 ns-long replicas were simulated using soft restraints with a force constant of 0.01 kcal $mol^{-1}Å^{-2}$ on the protein heavy atoms. A hierarchical clustering of the high energetic regions in the energy grids (the 0.02 percentile with the best free energy) was performed to extract a point representation of the binding hotspots.

Druggable pockets were predicted from the binding hotspots using a clustering algorithm.[78] Firstly, using the protein atoms' Van der Waals radii as an exclusion space, the distance between each hotspot was calculated using the A* pathfinding algorithm.[188] A surface distance matrix was generated using those distances. In the E3 ligases that showed significant conformational rearrangements, we used different representative frames to perform the surface distance calculations. Then, starting from the hotspots with highest energy, a graph was created with the other hotspots within 6Å, trimming redundant points (distance less than 1Å). A minimum graph size of 4 was set to ensure a relevant hotspot density in the pocket. Points in the graph were then removed from the pool and the process

repeated with the next highest hotspot. After all points were either isolated or assigned to graphs, a graph grouping step was done with an 8Å threshold. Finally, a manual filtering was done to ensure a good pocket geometry and balance between polar and hydrophobic hotspots.

3.2.1.2 Classical molecular dynamic simulations of the SCF^{FBXW7} complex

3.2.1.2.1 SCF^{FBXW7} complex and A5 ligand preparation

We used conventional MD (cMD) to study the behavior of the SCF^{FBXW7} complex in the presence and absence of A5, as a starting point to create hypotheses of the changes in c-Myc degradation. A5 is a FBXW7 ligand discovered by Míriam Martínez-Cartró during the prospective validation of the E3 ligases ligandability study. We simulated three variations of the system: apo, holo and holo-without-ligand. This was done to eliminate the effect of the starting conformation of the holo independently of the presence of the ligand. The SCF^{FBXW7} complex is typically formed in nature by FBXW7-SKP1-CUL1-RBX1-UBE2D1. However, we hypothesized that the effect of A5 modulation will only reach FBXW7 and SKP1. Thus, we cut from CUL1 onwards (i.e. only simulating FBXW7-SKP1-CUL1), also removing the interface between RBX1 and CUL1 to reduce artifacts during the simulation. Due to the lack of resolved structures of the SCF^{FBXW7} super-assembly we used the SKP1-CUL1 from the PDB entry 7B5L.[189] This structural model comes from CryoEM densities and portrays the SCF^{RBR} E3-E3 super-assembly (NEDD8, CUL1, RBX1, SKP1, SKP2, CKSHS1-CyclinA, CDK2-p27, UBE2L3, Ub, ARIH1). For the apo complex, FBXW7 coordinates were extracted from the PDB entry 2OVP, remodeling clashing side

chains in the FBOX-SKP1 interface with Modeller.[190] We used the conformation of FBXW7 and SKP1 employed for the virtual screening campaign that yielded A5 ligand to model the holo and the holo-without-ligand simulations. This conformation was selected from the pyMDMix simulations that aimed to identify druggable binding sites. In this case, the interface between SKP1 and CUL1 was remodeled using 7B5L as a template in Modeller.

Using Rosetta[191] we modeled a 28 residue-long loop in CUL1 (from S55 to V84) and a 10 residue-long loop (from H1065 to D1084). Both loop reconstructions were done using the kinematic closure (KIC) algorithm with 5000 iterations. A KIC loop refinement was executed independently for each loop, doing 2500 iterations for the CUL1 loop and 1500 for the SKP1 loop. Protonation states, and capping of all proteins was done in MOE 2022.[179] Partial charges for A5 were derived using the RESP protocol at HF/6-31G(d) level, calculated with Gaussian16.[192] GAFF2[193] parameters (bond lengths and angles) were adjusted from the ones observed after a DFT optimization at the b3lyp/6-31G* level with empirical dispersion. Parameters for the proteins were obtained from ff14SB. All systems were solvated using a TIP3P truncated octahedral box with 23 Å of buffer space from the solute to the periodic boundaries for a total of approximately 330000 atoms per system. Hydrogen mass repartitioning (HMR)[194] was performed on the resulting topology.

3.2.1.2.2 Holo and Holo-without-ligand SCF^{FBXW7} simulations

The minimization was done in four stages, each consisting of 3000 steps of steepest descent and then conjugate gradient until convergence. First, the position of water molecules was minimized, keeping everything else restrained using a harmonic potential with

force constant of 1.0 kcal $mol^{-1}Å^{-2}$. In the three subsequent stages SKP1, CUL1 and FBXW7 were unrestrained sequentially. The Holo complex had a fifth step unrestraining the ligand. The furthermost region of CUL1 that had been truncated remained constrained during the whole process (including equilibration and production).

The systems were heated in the NVT ensemble from 100 to 300 K in four steps, increasing the temperature 50 K on each step. The first step was 1 ns long and the rest 400 ps. A 2 ns-long step in the NPT ensemble was performed to equilibrate the system at 1 bar. During equilibration and all posterior simulations, Langevin thermostat (with a collision constant of 4 ps^{-1}) was used for temperature control, coupled with a Berendsen barostat in the NPT ensemble. For each system, 6 replicas of 1.5 μs were simulated, restarting the velocities from the equilibrated system. All production simulations were done using a 4 fs integration step due to using SHAKE restraints on non water hydrogens and HMR. All simulations were performed with the AMBER22 CUDA version of PME.[166]

A normal mode analysis was performed using the dried trajectories, simplified to only the protein alpha carbons and trimmed to represent 40 ps per simulation frame. The analysis of the moments of inertia and their projections were performed using pcasuite.[195]

3.2.1.3 Scaffold expansion of FBXW7 modulators

3.2.1.3.1 Construction of the A5 analogs library

An A5 scaffold-focused sublibrary from Enamine's REALSpace v2021-04 was created using SpaceMACS. The SMARTS pattern (‘*CCN1[CH2][CH2][CH2]c2[cH]c(C)[cH][cH]c21’*) was derived from the most stable (lowest RMSF) region of A5 from the SCF^{FBXW7}

simulations, blocking the ring substitution at unwanted positions. The subset execution mode was activated, allowing only for exact MCS matches in the output, limiting the output molecules' size to 20-35 heavy atoms. We filtered out molecules outside Lipinski standards,[119] with a logS <= -5 or raising a PAINS flag.[196] The resulting compound library was prepared using ChemAxon for tautomer and protonation states at pH 7,[197] and Corina (v4.4.0)[198] to generate stereoisomers, ring conformations and 3D conformations. The 400,765 ligands recovered in the scaffold search generated 694,564 conformers.

3.2.1.3.2 Virtual Screening of A5 analogs

The molecules were docked tethering the MCS to the equilibrated A5 binding pose from the most populated k-means cluster in the SCF^{FBXW7} simulations and to the original A5 docking pose, each with its FBXW7 conformation.The rDock cavity was generated using the reference ligand method with a radius of 6Å, using the corresponding A5 pose for each case. Two pharmacophoric restraints (a polar one next to N365 and an aromatic one where A5's benzene ring is positioned) were added to reproduce the virtual screening conditions used to discover A5.[199] Each conformer was docked through 25 iterations. From poses tethered to both reference conformations, a consensus score was generated and used in the following steps. The conformer with the highest SCORE.TOTAL was chosen for each molecule. The docked molecules were grouped using hierarchical clustering based on Tanimoto similarity using MACCSKeys with a threshold of 0.7. The representatives of each cluster, chosen by SCORE.INTER.norm, were selected for DUck calculations.

OpenDUck was used to set-up the dynamic undocking for the 2360 cluster representatives. The receptor chunk was defined with a radius of 10Å from the N365 amide nitrogen (ND2) in the side chain, which was used afterwards to define the steering vector. The atom types were obtained from ff14SB and GAFF2[193] for the protein and the system was solvated with TIP3P waters with a buffer space of 10Å. HMR was performed and the simulation steps increased to 4 fs. Each ligand ran for 5 SMD cycles with a W_{QB} threshold of 6 kcal mol^{-1}. From the top ligands (with W_{QB} larger than 10 kcal mol^{-1}), after a visual inspection, we selected 15 compounds prioritizing diversity in the linker region and the terminal group, to be synthesized by Enamine.

3.2.2 Bottom-up exploration of the chemical space

The bottom-up exploration of the chemical space pipeline consists of a hierarchy of increasingly sophisticated methods. The computational pipeline was built in a modular manner with two main independent sections (Figure 3.4). The first module, the exhaustive search of the low molecular weight chemical space, was performed by Marina Miñarro, using a fragment-based virtual screening.[200] The second module of the computational pipeline was the scaffold growing, which I took upon. The scaffold growing section is based on searching all the ligands with a certain scaffold within the ultra-large chemical libraries and then performing a tethered-based virtual screening, maintaining the original scaffold's binding mode.

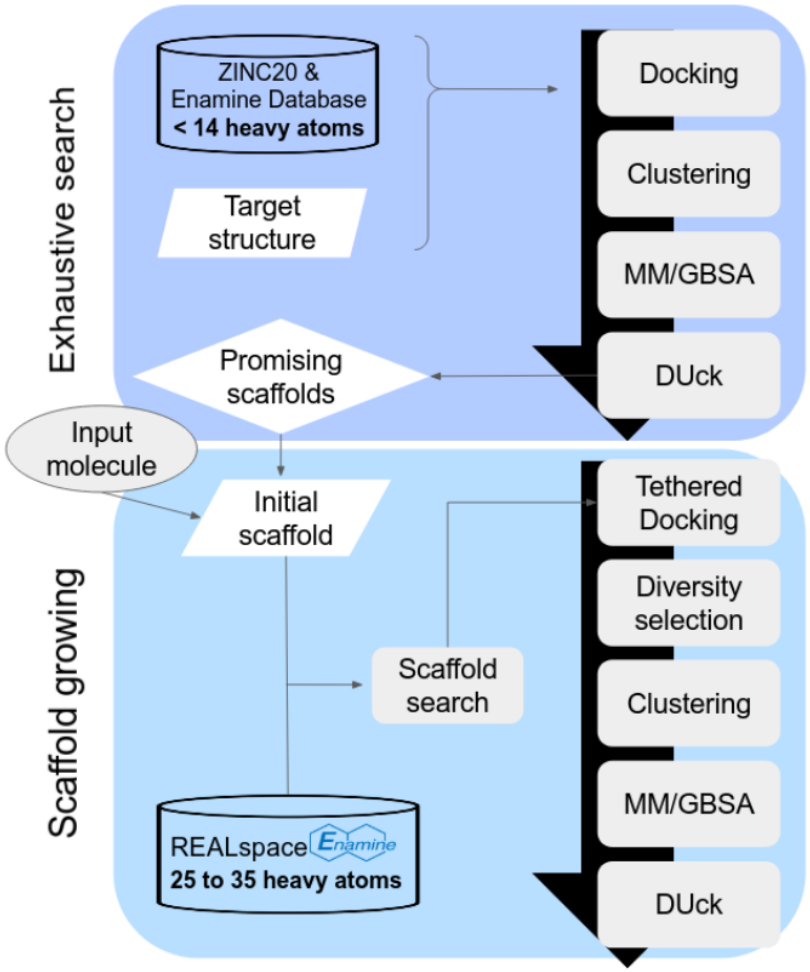


Figure 3.4. Computational pipeline of the bottom-up exploration of the chemical space.

3.2.2.1 Scaffold-focused libraries construction

We used an early developmental version of SpaceMACS (v0.9.5)[145] to perform the substructure search on the Enamine REALSpace (v04-2021) database. SpaceMACS queries were set up to generate up to 20 million compounds with the desired substructure and with 25 to 35 heavy atoms. The ligands were filtered to select the molecules fulfilling the Lipinski rule of five and having less than 8 rotatable bounds. Compounds that contain known PAINS motifs

were also discarded using MOE 2020.[179] We selected the top 10 million drug-like molecules to be the scaffold-focused libraries. Protonation states and tautomers (ChemAxon v0.21.0),[197] stereoisomers and three-dimensional structures (Corina v4.4.0)[198] were generated for each scaffold-focused library. With this protocol, we generated an average of 3.5 conformers per compound.

3.2.2.2 BRD4 receptor preparation

The PDB entry 4LR6 was used to obtain the coordinates of the BRD4 receptor. The structure was prepared using MOE 2020[179] by removing the ligand, protonating of protein residues and capping of terminal groups (i.e. acetylating the N-terminal amino groups and N-methylation of the C-terminal groups). A previous work in the group had determined the presence of a constituent water network present in multiple bromodomains.[201] We decided to maintain the whole water network (residues 302, 305, 311, 322, 327, 331 and 332 from the PDB id 4LR6) both in the docking calculations and posterior filtering steps.

3.2.2.3 Tethered docking of the scaffold-focused libraries

We docked the scaffold-focused libraries with rDock. The docking cavity was defined by the reference ligand method using the co-crystallized ligand in the 4LR6 PDB structure with a radius of 7 Å. The prepared library was docked with two pharmacophoric restraints, one hydrogen bond acceptor at a distance of 2 Å from Nδ of N140, and one hydrophobic spot at 2.5 Å of the crystallographic water network center of mass (Figure 3.5).[201] The acceptor point was defined with a 0.5 Å radius and the hydrophobic point with a 1 Å radius.

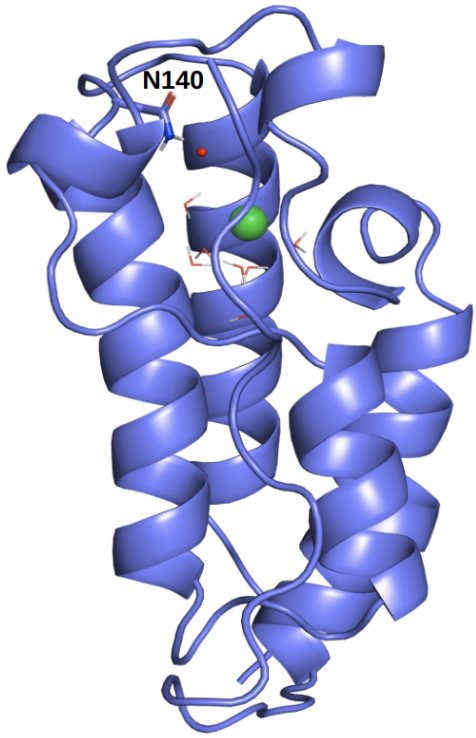


Figure 3.5. Structure of the Bromodomain-containing protein BRD4 (BD1). Highlighted in sticks, the residue N140 reported as the main interaction. The two spheres represent the pharmacophoric restraints used for docking, red and green for polar and hydrophobic respectively.

Before docking, ligands were aligned using their maximum common substructure to each reference compound. Atoms corresponding to the MCS were tethered with a minimum translation and rotation allowance (0.01 Å per internal iteration). A high throughput virtual screening (HTVS) filtering protocol was added to speed-up the docking. Compounds ran a maximum of 25 iterations; with the HTVS protocol stopping after the first and tenth iterations if the desired intermolecular and pharmacophoric scores were not met. Despite the hydrophobic pharmacophoric restriction imposed at the bottom of BRD4 cavity, rDock still favored the presence of certain polar groups close to the water network. Those poses were discarded

before selecting the best structure for each compound. The intramolecular scoring function (SCORE.INTRA) was used to determine the best conformer and binding pose per ligand. Then the ligands were sorted by the intermolecular scoring function (SCORE.INTER) to be clustered.

3.2.2.4 Ligands similarity-based clustering

The resulting list of compounds from each scaffold-library underwent a first selection based on compound diversity. We clustered the compounds using a hierarchical algorithm based on MACCSkeys-based Tanimoto similarity with a threshold of 0.95. This high similarity threshold reduced the compounds to test for the next clustering to approximately $2 \cdot 10^5$ compounds per library. The cluster representatives were chosen using the one with the best SCORE.INTER.

The resulting representatives were then re-clustered by Arnau Comajuncosa using the A1-A5 Chemical Checker *signaturizers*.[202] These include information related to 2D and 3D topological fingerprints, scaffolds, structural keys and broad physicochemical properties. We clustered them with the Sklearn Kmeans algorithm[136] setting a fixed number of 1000 clusters. The compound closest to the cluster centroid was selected as the final representative of each cluster. Three different measurements were employed for the clustering evaluation. Firstly, a comparison between the distance distributions of same-cluster and different-cluster compounds with the corresponding ROC. Secondly, a comparison between the average standard deviation for each feature of same-cluster and different-cluster compounds. Finally, description of several physico-chemical properties (molecular weight, logP, QED, number of H-

bond donors and acceptors, polar surface area, number of rotatable bonds and aromatic rings) of each cluster representative. This clustering step was performed for all but one scaffold-focused library. The comp 5-derived library was already reduced to 946 representatives during the hierarchical clustering.

3.2.2.5 MM/GBSA & Dynamic Undocking as post-docking filters

The final representatives were evaluated using MM/GBSA and dynamic undocking. For the MM/GBSA calculations, we used the whole BRD4 structure with the water network and the ligands' binding pose coming from rDock. The ligand files were modified adding all the non-polar hydrogens (which are eliminated by rDock) using Open babel.[203] The ΔG_{bind} and $\Delta G_{Solvation}$ were obtained using the Schrödinger Prime MM/GBSA. In parallel, the standard dynamic undocking protocol was performed, pulling from the ND2 (Nδ) of N140. The chunk was manually prepared using MOE 2020,[179] selecting the residues within 6 Å of the asparagine (corresponding to the residue mask :81-89, 91-94, 97, 101, 104, 105, 131, 135-137, 139-140, 144-146, 149) and water molecules 302, 305, 311, 322, 327, 331, and 332 (numbering according to PDB structure 4LR6). The compounds of each library were simulated for five SMD cycles, using the W_{QB} obtained by the fragment/ligand, where the scaffold was derived as the W_{QB} threshold.

Compounds were ranked using a consensus score from MM/GBSA's $\Delta G_{Solvation}$ and DUck. The top 10 representatives per scaffold were selected to be synthesized by Enamine.

3.2.2.6 Experimental methods

The synthesized compounds were tested experimentally by Andrea Bertran-Mostazo. As a first screening protocol, compounds were tested using differential scanning fluorimetry (DSF) and surface plasmon resonance (SPR) at 10 μM of ligand's concentration, by duplicate. Positive hits, defined by having a signal (i.e. ΔTm in DSF and RU in SPR) outside the control's error were then tested in a dose response manner with SPR and time-resolved Förster resonance energy transfer (TR-FRET). TR-FRET titrations ranged from 0.1nM to 10 μM and were performed by duplicate. In parallel, in the context of a collaboration with Dr. Maria Garcia at the EMBL-Hamburg, crystallization soaking experiments were performed by Dr. David Ruiz.

3.2.3 Steered molecular dynamics in drug discovery

3.2.3.1 Prediction of activity cliffs

3.2.3.1.1 Construction of the activity cliff datasets

Three datasets were prepared for ligands binding to HSP90, CDK2 and BACE1 proteins, with the aim of detecting activity cliffs in those datasets. The HSP90 ligands, their binding mode and the receptor conformation they bind to were generously shared by Kokh D.[97,204] The datasets for CDK2 and BACE1 were compiled by extracting the ligands with a resolved crystal structure in the PDB.[100] The CDK2 dataset was expanded using compounds from ChEMBL with a known activity.[205]

We generated the 3D structure of ChEMBL compounds and prepared them using LigPrep from the Schrödinger suit.[206] Their

binding poses were obtained with tethered docking calculations using rDock.[151] We employed the MCS with the most similar compound from the PDB to select the tethered atoms. This was done under the assumption that small structural differences between the reference and docked compounds would not cause major shifts in binding mode. Nonetheless, those ligands that showed drastic changes in the binding mode during the MD sampling were removed from the series. The ligands with a disruption of the binding mode adopt a suboptimal position where the key hydrogen bond is very likely to dissociate. Therefore, the first indicator of a shift in the binding mode is an abnormally low W_{QB}. Finally, the ligands underwent a manual inspection to correct wrong protonation states and other artifacts from the preparation process.

In order to ensure the quality of the reference experimental data, only ligands with reliable potency values were selected. As such, we filtered out the ligands without an exact potency value (e.g. having '<', '>' or '~' symbols) in the PDBbind[207], Binding DB[208] or in the Binding MOADB.[209] With the purpose of broadening the access to as many compounds as possible, the potency parameter selected was the IC_{50} instead of others present in fewer compounds (i.e. K_d or k_i). We are aware of the lack of consistency in the IC_{50} values between different assays and experimental conditions, and the effect it has on the validation of our predictions. The datasets were constructed with both analogous compound pairs with similar potency (non cliff) and those with high potency difference (cliffs). This was done to enable an analysis of the method's performance in a realistic situation. Activity cliffs are defined by a high disparity in potency and a high structural similarity. While most authors agree on 100-fold as the potency difference threshold (i.e. -2.73 kcal mol^{-1} @ 300K),

structural similarity is defined in a variety of manners in the literature. Fingerprint-based numerical similarity allows the study of a wider variety of changes, thus we used MACCskey fingerprints coupled with the Tanimoto coefficient,[210] aiming to maximize the available cliffs.

Altogether, the HSP90 dataset consisted of 93 HSP90 inhibitors (13 in the loop confirmation and 80 in the helical conformation), resulting in 207 pairs (18 activity cliffs and 89 non-cliff pairs). The CDK2 series consisted of 497 active inhibitors, forming 1164 pairs (121 activity cliffs and 1043 non-cliff pairs), with the series very evenly populated. The 50 BACE-1 inhibitors formed 82 pairs with only 3 activity cliffs.

3.2.3.1.2 Dynamic Undocking simulations of HSP90, CDK2 & BACE1 protein-ligand complexes

DUck standard pipeline was used for each protein-ligand complex. Proteins were prepared using MOE 2016,[179] correcting missing side chains, fixing protonation states and adding terminal capping. The chunk selection was done using the residues within 8 Å from the receptor's atom involved in the hydrogen bond, and then manually trimming out the external residues that do not interact with the binding pocket. To account for the plasticity of HSP90 binding site, receptor coordinates were extracted from different PDB structures: 2VCI, 5J9X, 5OD7, and 5OCI for resorcinol loop-binder, resorcinol helix-binder, quinazoline, and imidazole compounds, respectively. In the case of CDK2, the PDB 1H1S was used to generate the protein conformation and the H-bond to study was formed with L83's backbone in the hinge. The 1TQF PDB entry was used to model the BACE1 structure where H-bonds with D93's carboxylic oxygen were steered. Each ligand was submitted to a DUck run with 60 SMD

cycles. All the simulations were performed with the AMBER16 CUDA version of PME.[166] Quasi-bond free energies (ΔG_{QB}) were calculated using the Jarzynski equality using the normalized work from each replica.

3.2.3.1.3 τ-RAMD simulations on HSP90 ligands and, the selected CDK2 series and BACE1 representatives

The protocol to perform the τ-RAMD simulations was reproduced from Kokh et al.[97,204] All HSP90 ligands were tested, while only testing 2QZL from BACE-1 and 10 CDK2 inhibitors (CHEMBL409731, CHEMBL1234085, CHEMBL112136, CHEMBL333382, CHEMBL411426, CHEMBL271842, CHEMBL419534, CHEMBL316809, 2R3K and 2C68). Using the complete (not chunked) receptors, the systems were solvated using a truncated octahedral TIP3P box with sodium chlorine to obtain charge neutrality. GAFF and ff14SB were used to obtain the atom types and parameters for protein and ligands respectively. HSP90 ligands' partial charges were provided by Dr. Daria Kokh, which had been calculated with the RESP procedure from *ab initio* calculations at the Hartree-Fock level with the 6-31G*(1d) basis set in GAMESS.[211] HSP90 and BACE1 ligands' parameters were obtained from the DUck preparation (i.e. SMIRNOFF99Frosst force field and AM1-BCC derived partial charges). The systems were minimized in AMBER14 with 500 steps of steepest descent and 1000 steps of conjugate gradient. Then they were gradually heated to 300K in 1 ns using harmonic restraints on the heavy atoms with a force constant of 50 kcal $mol^{-1}Å^{-2}$. The restraints were decreased in 5 steps during equilibration (each of 1 ns with force constants of 50, 10, 5, 2, and 0.5 kcal $mol^{-1}Å^{-2}$). Then a 10 ns sampling step was done without restraints. Equilibration and production were simulated at 300K

using a Langevin thermostat (with a collision frequency of 4 ps^{-1}) and a Berendsen barostat at 1 bar, using 2 fs steps with hydrogen atoms constrained using the SHAKE algorithm.[212] The equilibrated systems were transferred to NAMD format and re-heated in steps of 10K without restraints. A second equilibration was performed in NAMD[213] in the NPT ensemble using a Langevin thermostat and Nosé-Hoover langevin pressure control at 300K and 1 bar. For each ligand, two snapshots of the equilibration were extracted as starting points for the RAMD simulations.

We used the τ-RAMD tcl wrapper to perform random accelerated trajectories with a force of 16 kcal $mol^{-1}Å^{-1}$. The center of mass (CoM) of the ligand was checked every 100 fs. If the CoM of the ligand moved less than 0.025 Å, the direction of the force vector was updated. Simulations were stopped when the ligand's center of mass moved further than 30 Å from its original position. A range of 15-50 dissociation trajectories were calculated depending on the τ convergence (given by the bootstrapping procedure expounded in the paper). In addition to the ligand's center of mass distance progression, the H-bond studied in DUck was also recorded.

3.2.3.1.4 MM/GBSA and rDock single point calculations

In order to perform baseline comparisons of DUck with other quick methods, we optimized rDock scores and performed single point MM/GBSA calculations. For the rDock scores, the ligand coordinates (i.e. the crystal structures for HSP90 and BACE1 and the tethered docking poses for CDK2) were minimized with the minimise.prm rDock protocol. Then SCORE.INTER was recorded as the docking descriptor. In parallel, the original poses were transformed to Maestro format using the *structcat* command from the Schrodinger

suite.[206] The MM/GBSA calculations were performed using the prime package in Schrödinger prime suite, taking into account the ligand strain.[214] The reported ΔG_{Bind} was recorded as the MM/GBSA descriptor. Differences in SCORE.INTER and ΔG_{Bind} were used to predict activity cliffs, in the same way as ΔG_{QB} in DUck.

3.2.3.1.5 Prediction performance evaluation

Areas under the ROC curve were calculated from the thresholds following the rank order from each scoring method. For static evaluations, the $\Delta\Delta G_{QB}$ threshold was set as the relation of the differences in ΔG_{QB} and the absolute value, in order to use it regardless of the overall magnitude of the values. We selected 0.5 as the threshold (i.e. pairs of compounds where the $\Delta\Delta G_{QB}$ was higher than half the minimum ΔG_{QB}). Due to the imbalanced nature of the data set (15% cliffs vs 85% non-cliff pairs) performance was evaluated using the Matthews correlation coefficient (MCC).[215]

$$MCC = \frac{(TPxTN) - (FPxFN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.13)$$

3.2.3.2 Development and validation of the OpenDUck python library

3.2.3.2.1 Dynamic Undocking simulations of the SERAPhiC and Iridium datasets

The prepared systems from SERAPhiC and Iridium datasets were obtained from Majewski et al. 2019.[181] Hydrogen bonds in complexes were identified using Chimera's FindHBond function.[216] Then each bond was treated as a core for the individual simulation system defining the center point for the chunk (with a radius of 7 Å).

A total of 469 H-bond interactions (345 from Iridium and 124 from SERAPhiC) were selected from the 105 protein-ligand systems.

Table 3.1. Simulation conditions run on the Iridium and SERAPhiC datasets used to validate the openDUck implementation.

Simulation Engine	Ligand force field	HMR	Water model	Box buffer space (Å)	Ionic Strength (M)	SMD cycles
OpenMM	SMIRNOFF	No	TIP3P	10	0.1	20
AMBER	SMIRNOFF	No	TIP3P	10	0.1	20
AMBER	SMIRNOFF	Yes	TIP3P	10	0.1	20
AMBER	GAFF2	Yes	TIP3P	10	0.1	20
AMBER	GAFF2	Yes	TIP3P	10	0.1	500
AMBER	GAFF2	No	TIP3P	10	0.1	20
AMBER	GAFF2	Yes	SPCE	10	0.1	20
AMBER	GAFF2	Yes	TIP4Pew	10	0.1	20
AMBER	GAFF2	Yes	TIP3P	18	0.1	20
AMBER	GAFF2	Yes	TIP3P	10	0	20

Each chunk-ligand system was prepared using different conditions on OpenDUck, varying the ligand simulation engine (OpenMM, AMBER), force field (GAFF2, SMIRNOFF), water models (SPCE, TIP3P, TIP4Pew), solvation box size, ionic strength, applying HMR and the number of replicas. All conditions are summarized in Table 3.1. Contrary to normal DUck executions, no W_{QB} threshold was specified. Systems where no minimum was found were discarded, as normalization was not possible. The proposed default conditions: AMBER+GAFF2+HMR+TIP3P+10 buffer+0.1M ionic strength were simulated twice using different equilibration seeds, to assess the impact of equilibration in the W_{QB} and ΔG_{QB} obtained and the

convergence. ΔG_{QB} was calculated using the Jarzynski equality bootstrapping 25 iterations resampling half the SMD cycles with reposition. In the case mentioned above with extended simulation replicas, the bootstrapped ΔG_{QB} were calculated using different subsample sizes sequentially from 20 to 500 in steps of 10 SMD cycles (i.e. 20, 30, 40, [..], 490, 500).

3.2.3.3 Crooks Gaussian Intersection to speed

3.2.3.3.1 CRBN-CK1α and HSP90 system preparation

The structure models of Cereblon (CRBN) and Casein Kinase isoform 1 alpha (CK1α) were recreated from Miñarro et al. 2023.[217] The CRBN, CK1α and lenalidomide coordinates were obtained from the crystallographic data in the PDB entry 5FQD,[218] removing the adaptor protein DDB1. Mutants of CK1α (i.e. I35G, I37E, N39G and G40N) were created using the mutagenesis tool in PyMOL.[219] Amber ff14SB and GAFF2 force fields were used to assign parameters to proteins and lenalidomide respectively. Partial charges for lenalidomide were derived using the RESP protocol at HF/6-31G(d) level calculated with Gaussian09.[192] The Zn^{2+} coordination with CRBN was represented using the out-of-center dummy model.[220] Each system was solvated on a truncated octahedral box of TIP3P[221] water molecules with 12 Å buffer margin. Counterions were added to attain charge neutrality.

The structure model of HSP90 was obtained from the PDB entry 2VCI. Protein preparation protocols including capping of terminal amino acids, reconstruction of missing side chains, and selection of protonation states for residues were done using MOE 2022-2.[179] The initial coordinates of the ligands were extracted from their respective PDB entries (2BSM, 2UWD, 2VCI, 6J64, 5NYI, 6ELN, 6ELO, 6F1N,

2YKI) and aligned to 2VCI. The protein-ligand topologies were generated using openDUck as the interface. Amber ff14SB and GAFF2 force fields were used to assign parameters to HSP90 and the ligands respectively. Partial charges were obtained using the AM1-BCC method with *antechamber*. Each system was solvated on a prismatic box of TIP3P water molecules with 15 Å buffer margin. Counterions were added to attain charge neutrality.

3.2.3.3.2 Equilibrium and steered molecular dynamics simulations

Each system was minimized in two stages. First, the position of water molecules was minimized with a combination of 3500 steps of steepest descent and 6500 steps of conjugate gradient, while keeping the protein and the ligands restrained using a harmonic potential with force constant of 5.0 kcal mol^{-1} $Å^{-2}$. Secondly, side chains and water molecules were minimized using 4500 steps of steepest descent, followed by 7500 steps of conjugate gradient.The ligand and the Zn^{2+} cation in the CRBN-CK1α systems were restrained using a harmonic potential using the same force constant.

The systems were then heated in the NVT ensemble moving from 100 to 298 K in three phases of 250 ps (100-150 K, 150-250 K and 250-298 K), keeping the restricted atoms from the last minimization step. To equilibrate the systems at 1 bar, a 1 ns-long step in the NPT ensemble was performed. During equilibration and all posterior simulations, Langevin thermostat (with a collision constant of 3 ps^{-1}) was used for temperature control, coupled with a Berendsen barostat in the NPT ensemble.

For the CRBN-CK1α systems, three H-bond were studied (I29-N542, N31-W591, T30-H548) and the distance between heavy atoms involved in the interaction was assigned as the CV. In HSP90 only the H-bonds formed between D93 and the ligands were studied. Due to the symmetric nature of aspartate, the center of mass of the carboxylic acid was used to define the CV. We performed 100 independent replicas for each H-bond. Half of the replicas (50) followed the protocol from Miñarro et al.[217] Firstly, velocities were reinitialized and the system was equilibrated for 10 ns using a flat-bottom restraining schema for the H-bonds, keeping them between 2.5 and 3.5 Å, with a harmonic force constant of 60 kcal mol^{-1} $Å^{-2}$. Secondly, a 1 ns SMD stimulation was performed forcing the interaction to 2.5 Å using the same force constant. Starting from the end position of the last step, a SMD simulation is performed steering the hydrogen bond up to 5 Å, using a stiff spring constant (500 kcal mol^{-1} $Å^{-2}$). For the other half of the replicas, we added a fourth stage where the coordinates are kept but velocities reinitialized. In this fourth step, the H-bond is brought back from 5 to 2.5 Å at the same speed. The CV distances were adjusted in HSP90 to incorporate the effect of using the center of mass of the aspartate instead of one of the two oxygen (3 to 5.5 Å). Four different speeds (5, 2.5, 1 and 0.5 Å ns^{-1}) were tested on the two last steps (also named “Forward”, “Reverse” SMD steps) starting from the same respective end point of the second step. Force and work integration steps were adjusted to be the same for each speed (Figure 3.6).

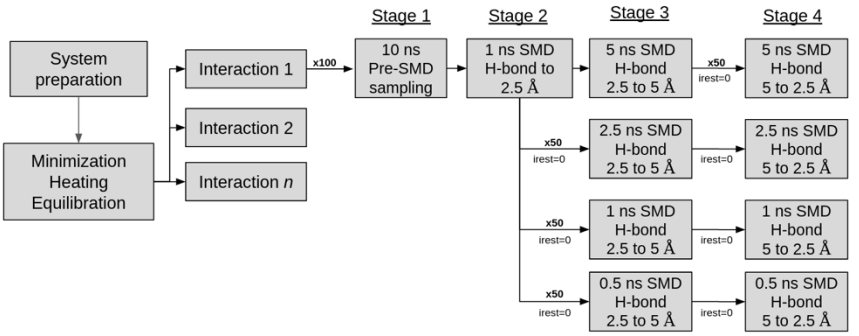


Figure 3.6. Workflow diagram of the simulation steps performed on the CRBN-CK1α system. After minimization, each interaction is simulated separately, performing 100 replicas of the three first production stages. We simulated the last stage on half (50) of the replicas. The two last SMD stages (3 & 4) are performed at four different speeds (50 replicas per speed). "irest=0" signals the velocities are reassigned from the restart file.

An additional harmonic restraining schema was added to all heavy atoms in HSP90 simulations to avoid the pocket collapsing during the forward SMD steps, which caused artifacts in the reverse SMD processes. All the simulations mentioned above were performed using SHAKE[212] restraining schema to hydrogens with a time step of 2 fs. The simulation engine was the CUDA accelerated implementation of PMEMD in Amber22.[222] Finally, PMF were calculated using the Jarzynski equality (eq. 3.12.b). Error estimations of the PMF profiles were obtained by bootstrapping 10 times subsamples of 25 replicas with replacement at each distance point for the set of work values.

CHAPTER 4: RESULTS

EXPANDING THE E3 LIGASES TOOLBOX

4. EXPANDING THE E3 LIGASES TOOLBOX

4.1 Background

Targeted protein degradation (TPD) is a as a new drug modality that has changed how we confront targeting *undruggable* proteins. Degraders differentiate from traditional small molecules in several ways. TPD allows us to use small molecule binders without relevant phenotypical activity and convert them into potent degraders. This also enables targeting other binding pockets outside the orthosteric cavity, increasing the druggable proteome and also circumventing resistances caused by mutations in the canonical binding site.[223] Unlike traditional small molecule inhibitors, protein degraders can induce sustained pharmacological responses even after being metabolized.[224] The reduction of protein levels can also avoid the upregulation of alternative signaling pathways and decrease side effects.[223]

In eukaryotic organisms, protein degradation can happen via two independent, but interconnected, pathways; the proteasome Ubiquitin-Proteasome system (UPS) and the lysosomal system.[225] Generally, the UPS is responsible for the elimination of short-lived and soluble misfolded proteins. On the other hand, lysosomes eliminate insoluble protein aggregates, and even intracellular parasites. The most common TPD applications (i.e. PROTACS, bivalent, and molecular glues, monovalent) hijack the cell's ubiquitination machinery to signal selected proteins of interest for degradation through the UPS.[226] This process occurs through a cascade of enzymatic reactions. Initially, ubiquitin is activated by an

E1 enzyme. The enzyme is then transferred to a E2 conjugating enzyme. Then ubiquitin is transferred to the substrate, previously captured by the E3 ligase, through an exposed lysine of the substrate protein. This process is repeated several times, resulting in chains of ubiquitins which depending on the chaining pattern will result in different substrate protein fates, one of them being proteasome-dependent degradation (Figure 4.1).

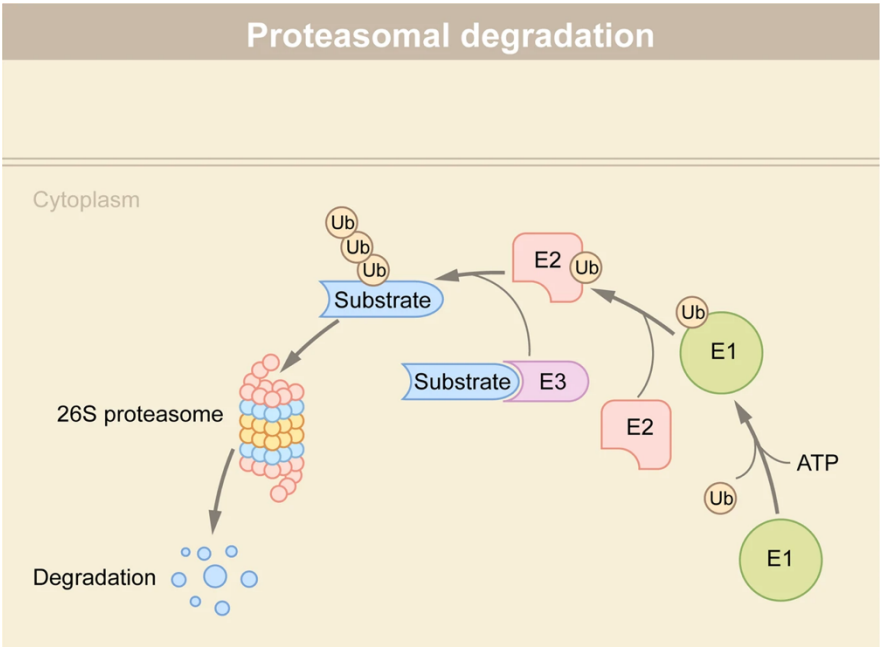


Figure 4.1. Protein degradation via the UPS. Extracted from [224]

E3 ubiquitin ligases confer substrate specificity to the UPS. E3 ligases recognize specific regions of their substrates named degron, which usually are associated with post-translational modifications (PTM). Whilst there are only around 40 known E2 ligases, there are more than 600 described E3 in humans. E3 can be classified into three different types based on their ubiquitin transfer mechanism: really interesting new genes (RING), ring-between-ring (RBR), and homologous to E6AP C-terminus (HECT). RING E3's bind the

substrate whilst the ubiquitin is transferred directly by the E2. RBR and HECT form a E3-ubiquitin intermediate before transferring it to the substrate. [227]

Recruitment by an E3 ligase is a mandatory step for the action of PROTAC and molecular glues. Today, only a handful of E3 ligases have been exploited for TPD strategies. The major hurdle for exploiting these E3 ligases is the lack of corresponding small-molecule binders for them that could be incorporated into the degrader structure. Considering that E3 ligases represent a large gene family with more than 600 members, there is a clear mismatch between the number of members of the family and the number of E3 that have been successfully engaged and harnessed by small-molecules degraders. This reflects more on the novelty of the field rather than the intrinsic undruggability of the E3s.

The two most exploited E3 ligases both in academic and pharma settings have been VHL and cereblon (CRBN). Currently, only molecular glues targeting CRBN are currently approved.[228] Although VHL and CRBN have routinely demonstrated their potential degrading dozens of targets following a 'plug-and-play' approach, they still suffer from several limitations and bottlenecks. On top of CRBN and VHL, other E3 ligases are starting to be exploited with small molecules (e.g. IAP, MDM2, RNF114, RNF4, DCAF15, DCAF16, AhR, GID4), or peptides (e.g. KEAP1, BTRC). Diversifying the available E3 toolbox allows to exploit their diverse temporal and spatial expression profiles in the pursuit of more selective treatments.

E3 ligases specificity to degrade substrates make this protein family highly appealing for drug discovery therapies aiming to inhibit their

action. In particular, to avoid the deletion of transcription factors, E3 inhibitors decrease side effects associated with proteasome inhibitors.

4.2 Study of E3 ligandability through mixed solvent molecular dynamics

4.2.1 Selection of relevant E3 ligases

Together with Dra. Míriam Martínez-Cartró, a former PhD student, we compiled a comprehensive database of all available E3 ligases with X-ray structure including structural (i.e. crystal quality, coverage) and functional (i.e small molecules, predicted relevance for protein degradation) data. We obtained the structural information from PDB and the functional annotations from Ubihub,[183] Uniprot[184] and PROSITE.[185] Using this information, a representative subset of 23 E3 ligases was selected, based on protein families and structural motifs. We chose both untargeted E3 ligases and E3 ligases with known small-molecule binders to allow for retrospective cross validation. The selected E3 ligases are shown in Table 4.1. Due to the low coverage of the domains of interest, two different structures of TRIM28 were selected. This selection was done before the publication of AlphaFold,[38] which would have streamlined the selection. However, the relative position of multiple domains still remains a challenge and could have been a source of noise in our study.

Table 4.1. List of selected E3 ligases to study with MDMix.

E3 ligase	Simulated structural motif	PDB entry	Coverage	E3 family
ASB9	ASB + SOCS-BOX	3ZKJ	88%	SOCS-BOX
BTRC	WDR + F-BOX	6M90	66 %	F-BOX
CDC20	WDR	4GGC	63 %	APC
CRBN	CULT	5FQD	85 %	DCAF
COP1	WDR	5HQG	44 %	Misc.
DCAF1	WDR	4CC9	20 %	DCAF
DCAF15	WDR	6UD7	71 %	DCAF
FBO44	FBA + F-BOX	3WSO	99 %	F-BOX
FBW7	WDR + F-BOX	2OVP	62 %	F-BOX
GID4	GID/CTLH	6CCU	55 %	Misc.
KCTD5	BTB	3DRX	76 %	BTB
KEAP1	KELCH	6TYM	46 %	BTB
KLHDC2	KELCH	6DO3	78 %	VHL-BOX
PRKN	RING1-IBR-RING2	5C1Z	83 %	RBR
RNF4	ZF-RING	4PPE	35 %	Misc.
RNF43	Extracellular domain	4KNG	19 %	Misc.
SKP2	LR + F-BOX	2AST	77 %	F-BOX
SOCS2	SH2 + SOCS-BOX	6I5N	81 %	SOCS-BOX
SPOP	MATH + BTB	3HQI	78 %	BTB
SPSB1	SPRY	2JK9	74 %	SOCS-BOX
TRIM21	SPRY	2IWG	38 %	TRIM
TRIM28	PHD-BRD RBCC	2RO1 6QAJ	23 % 33 %	TRIM
VHL	SOCS-BOX	4W9H	74 %	VHL-BOX

4.2.2 Identification of ligandable pockets

4.2.2.1 Mixed solvent molecular dynamics and hotspot generation

We set up a pipeline to discover and evaluate possible binding sites on the surface of the selected E3 ligases. To assess ligandability, we used the energy profiles of ethanol in a 1:4 mixture with water, separating polar (OH) and hydrophobic (CT) contributions using the MDMix protocol. For the MDMix simulations we conserved adaptor proteins present in ASB9, CRBN, DCAF15, FBO44, FBW7, SKP2, SOCS2 and VHL (e.g. SKP1, DDB1, Elongins, etc). Previous work has shown that the lack of such adaptor proteins can alter the E3s' conformation, on top of introducing noise in the form of unreachable protein-protein binding pockets. We clustered the most energetic points on the E3s surfaces using surface distance as the descriptor. To ensure the ligandability of the pockets, we set a minimum of 4 hotspots per cluster, all within a maximum volume of 500 $Å^3$ to be fulfilled. The resulting pockets were manually trimmed considering their geometry and requiring a balance between polar and hydrophobic hotspots. Finally, we ranked the resulting pockets using the sum of energies derived from their pockets. This energy value cannot be directly taken as a theoretical maximum binding affinity for small molecules in the pocket, as the interactions described are limited to those of a alcohol group and an aliphatic chain from the ethanol. Moreover, it assumes an additive relationship between them, which might not necessarily be fulfilled. Nonetheless, this energy value gives a good insight into the probability of small-molecules to bind, and can be used to design the pharmacophoric features a possible ligand would need.

4.2.2.2 Retrospective validation: E3 ligases with known binders

To validate retrospectively the pocket predictions from the clustered MDMix energies, we focused on the crystallized E3 ligases with reported non-covalent small-molecules binders (BTRC, CDC20, CRBN, DCAF1, DCAF15, GID4, KEAP1 and VHL). VHL degron recognition is driven by the substrates' hydroxyproline. This chemical moiety has been extensively used for the development of PROTAC molecules. Remarkably, one of the selected pockets with better energy profiles (VHL-pocket E, -6.26 kcal/mol) corresponds with the hydroxyproline binding pocket. The two main hotspots of the Pocket E match perfectly the hydroxyl and the tert-butyl binding surface, characteristic of the VHL inhibitors reported (Figure 4.2 A).[229] Due to the very mild protein restraints used during the mixed-solvent MD simulations and induced by the hydrophobicity of ethanol, cryptic binding sites appeared in some proteins. Interestingly, in the case of the VHL simulations, which we selected as druggable (VHL-Pocket C, -9.6 kcal/mol), corresponds to a previously described VHL allosteric binding site[230] (Figure 4.2 B). We observed an even bigger conformational change in GID4's degron recognition site entering the β-barrel. A pocket (GID4-Pocket D) in that region was discarded due to being much bigger (circa. 1000 $Å^3$) than the volume thresholds we established (Figure 4.2 C). Nonetheless, two small-molecules have been reported to bind in that pocket (PDBs 7SLZ and 8V1P). It is important to adjust the energies found in *transient* pockets as they might be underestimated due to the partial occupancy possible.

CRBN is known to bind immunomodulatory drugs (IMiDs), which act as molecular glues promoting the degradation of non-native protein substrates (neo-substrates).[231] CRBN neo-substrates are zinc fingers (ZF) of transcription factors, and bind to the so called CULT domain. Buried on CRBN's ZF binding pocket of the CULT domain, the IMiDs position their glutarimide moiety, whereas the phthalimide region remains more solvent exposed (or exposed to the substrate). The CRBN predicted pocket G (-5.65 kcal/mol) matches the IMiDs' binding side. The glutarimide site being the most energetic, with the hotspots matching the polar interactions. The phthalimide moiety was not described as accurately due to being completely water exposed in our simulations (Figure 4.2 D).

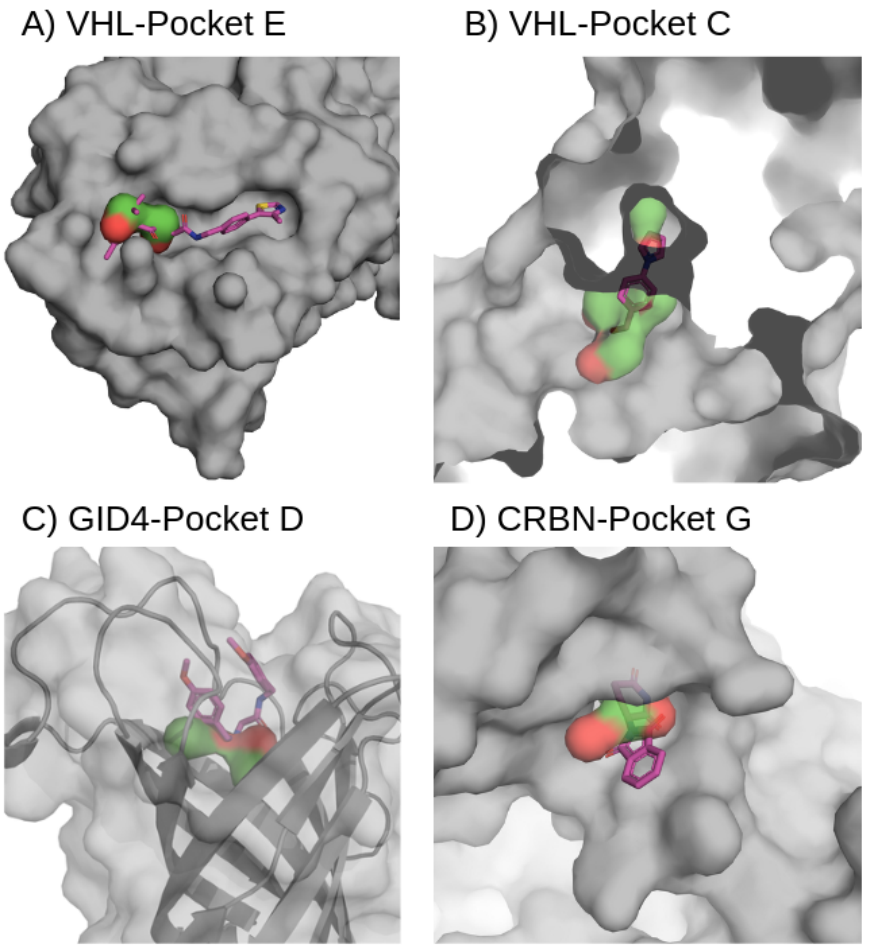


Figure 4.2. Representation of E3 with the crystallized small molecules matching the predicted druggable pockets. Polar and hydrophobic hotspots are shown in red and green respectively.

Similarly to IMiDs, indisulam has been described as a molecular glue inducing the degradation of neo-substrates through the E3 ligase DCAF15.[232] In spite of the indisulam binding site appearing during the hotspot clustering, this pocket did not fulfill the druggability conditions we proposed. Both examples show the need of their neo-substrates to create the appropriate binding site for molecular glues.

KEAP1, BTRC and DCAF1 degron recognition sites are situated on top of a WDR domain. On one hand, KEAP1 small-molecule binders inhibit the interaction with its natural substrate NRF2.[233] There are several chemical families of KEAP1 inhibitors which bind to different regions to the degron recognition site. KEAP1-pocket B (-9.61 kcal/mol) covers hotspots interacting with S602 and W334 and entering the WDR domain. However, the rest of the cavity, which is used by some inhibitor families, was not predicted in the MDMix simulations. An example of a KEAP1 inhibitor covering both regions, extracted from the PDB 5NFU, can be seen in Figure 4.3 A. On the other hand, BTRC binders have been reported to enhance the interaction with its natural substrates (e.g. β-Catenin, CDC25, Wee1).[234,235] This case is similar to molecular glues, where the ligands' interacting site is not reconstructed correctly due to the need of a second protein interacting partner. Although BTRC-Pocket C (-6.42 kcal/mol) do not reproduce completely the interactions between BTRC and the ligand nor substrate, the most energetic hydrophobic hotspot aligns with the ligand's trifluoromethyl moiety (Figure 4.3 B) Similarly to BTRC, CDC20 inhibitor's binding pocket is only predicted partially, matching the trifluoromethyl group of Apcin[236] (Figure 4.3 C). Finally, in the DCAF1 degron recognition site, the predicted pocket A missed most of the interactions with the ligand (reported in the PDB entry 7SSE).[237] In the holo structure of DCAF1 (7SSE) we can appreciate a big conformational change of a loop that was capped on our simulations; this might explain the mismatch (Figure 4.3 D).

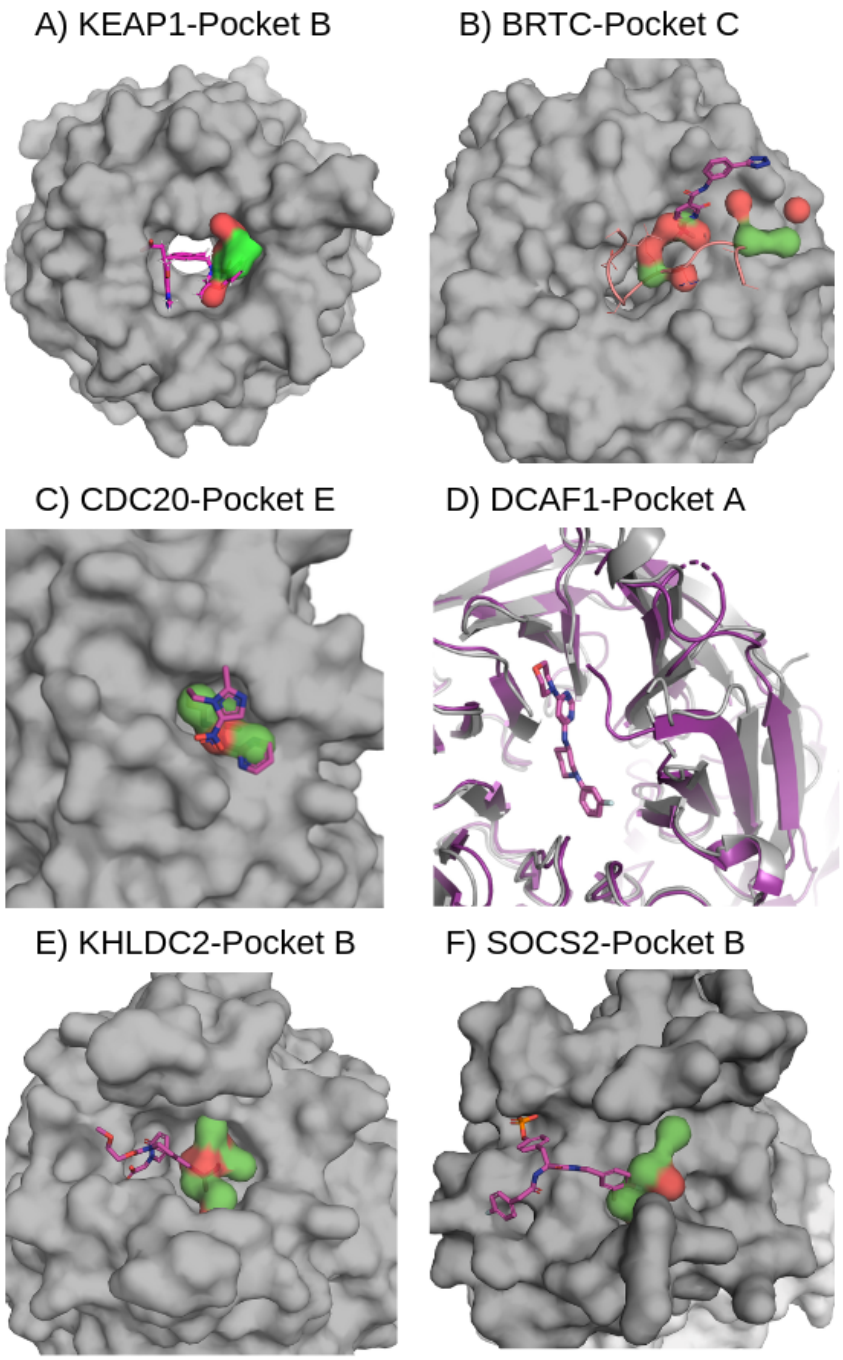


Figure 4.3. Representation of E3 with the crystallized small molecules. Polar and hydrophobic hotspots are shown in red and green respectively. DCAF1 (D) from the 7SSE structure (pink) is aligned to the simulated structure (grey).

Recently, Hickey et al.[238] presented a ligand bound on the KLHDC2 degron recognition site. The predicted KLHDC2-Pocket B (-10.42 kcal/mol) in that region, misses most of the ligands interaction. However, it matches the natural substrates diglycine C-End degron peptide interactions,[238] particularly the ones forming H-bonds with R189 and K147 from KHLDC2 (Figure 4.4.E). Ciulli's group has recently reported a covalent SOCS2 ligand.[239] The ligand overlaps slightly with the degron recognition site, but we were not able to find hotspots reproducing its interactions (Figure 4.3.F). Despite covalent ligands having high potency due to the bonded part, we assumed that the non-reactive region has certain complementarity, which we should have been able to catch.

4.2.2.3 Ligandability of degron recognition sites

Degron recognition sites can be called the main orthosteric regions where E3s interact with their substrates. Thus, they are prime regions to target for drug discovery. As mentioned above, VHL, KEAP1, BTRC, GID4 and DCAF1 are examples of systems where ligands binding to the degron recognition sites have been developed. CRBN is an exception, since IMiDs induce the recognition of the neo-substrates through a different region. Despite being one of the most exploited E3 ligases, few is known about CRBN native substrates, which are reported to interact with the Lon N-terminal domain.[240] Interestingly, we predicted three ligandable pockets (pockets B, E and J, with associated energies of -12.04, -6.18 and -5.65 kcal/mol respectively) through the MDMix protocol in this region.

MDMix hotspots matching PRKN and TRIM21 degron recognition sites were also correctly clustered and identified as druggable pockets. We reported two possible druggable pockets for PRKN

(pocket A and C). PRKN-Pocket A (-9.2 kcal/mol) covers the area between the degron recognition region and the ubiquitin-like domain. This last domain has been related to PRKN modulation. The pocket C (-8.43 kcal/mol) matches directly the degron binding site, albeit only partially. On the case of TRIM21, the interactions with the natural substrate IG-γ,[241] were recovered on the pocket C, in particular the interaction with TRIM21 N451 (Figure 4.4 A). In the case of SOCS2, we discarded the hotspot cluster that appeared in the degron recognition site due to its flat surface (Figure 4.4 B). Considering that all the degron recognition sites are protein-protein interaction sites, it is surprising that this was not a more common case.

SKP2 and COP1 recognise proteins that act as a scaffold to degrade the target protein. In these cases we cannot talk directly about degron recognition sites, however the modulation of their protein-protein interactions might still be of interest. SKP2 degrades p27 using CSK1B as intermediate. The predicted pocket D (-10.08 kcal/mol) for SKP2 overlaps with CSK1B's peptide, but it extends further in the opposite direction (Figure 4.4 C). COP1 recognises TRIB1 which then recognises substrates, like AKT, for degradation.[242] COP1-Pocket D (-9.29 kcal/mol) mirrors TRIB1 peptide interactions accurately (Figure 4.4 D). Similarly to what we observed in SOCS2, SPSB1 substrate binding site is a very flat region and the energy clusters found there were discarded during the pocket druglikeness classification.

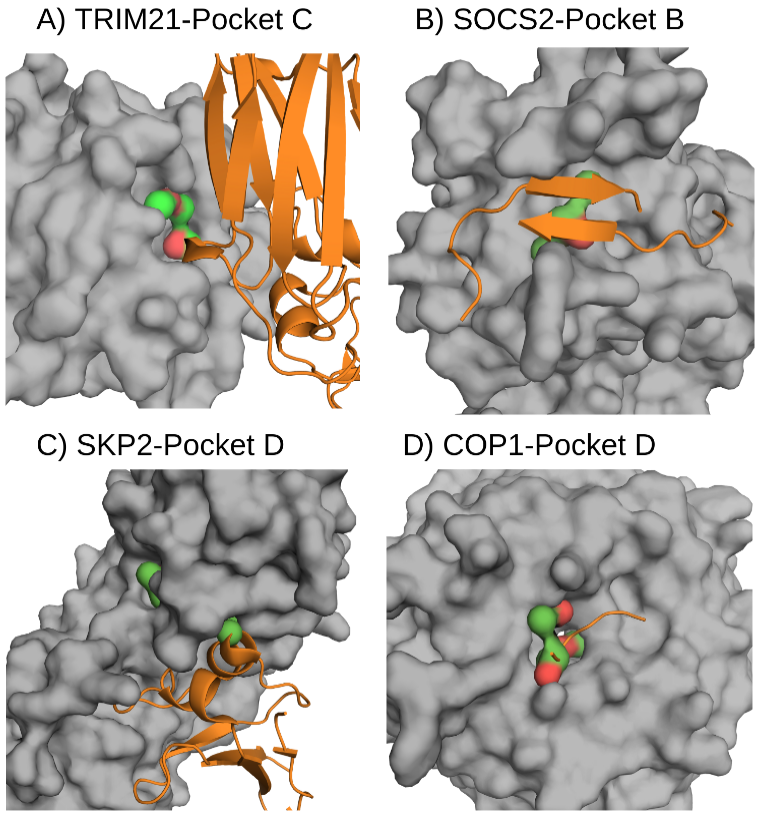


Figure 4.4. Representation of E3 ligases with the crystallized native substrates. Polar and hydrophobic hotspots are shown in red and green respectively, native substrates are shown in orange and cartoon representation.

In summary, despite being protein-protein interaction sites, we were able to find druggable pockets in the degron recognition sites of most of the E3 ligases (Table 4.2). While some of those degron recognition sites have been exploited already for drug discovery, there is a significant opportunity for small-molecules to modulate the action of new E3 ligases either through their degron recognition site or alternative, allosteric binding pockets.

Table 4.2. Summary of the orthosteric pockets predicted with MDMix

E3 ligase	Known degron recognition site	Substrate or substrate adaptor crystallized	Ligand crystallized in degron site	Pockets in degron or close by	Degron site predicted as ligandable
ASB9	Yes	Yes	No	B,C,M	No
BTRC	Yes	Yes	No	C	Yes
CDC20	Yes	Yes	No	No	No
CRBN	Predicted	No	No	B, J, E	Yes
COP1	Yes	Yes	Yes	D	Yes
DCAF1	Yes	Yes	Yes	A	Yes
DCAF15	No	-	-	-	-
FBO44	No	-	-	-	-
FBW7	Yes	Yes	No	Yes	Yes
GID4	Yes	Yes	Yes	D	No
KCTD5	No	-	-	-	-
KEAP1	Yes	Yes	Yes	B	Yes
KLHDC2	Yes	Yes	Yes	B	Yes
PRKN	Yes	No	No	D, F	Yes
RNF4	Yes	No	No	No	No
RNF43	No	-	-	-	-
SKP2	Yes	Yes	No	D	Yes
SOCS2	Yes	Yes	Yes	B, D	No
SPOP	Yes	Yes	No	No	No
SPSB1	Yes	Yes	No	G	No
TRIM21	Yes	Yes	No	C	Yes
TRIM28	Yes	No	No	J, O, I	No
VHL	Yes	Yes	Yes	E	Yes

4.2.2.4 MDMix predicts novel allosteric pockets

MDMix serves as a valuable tool for uncovering previously unexplored protein cavities. A significant portion (66%) of the predicted pockets were identified as novel and allosteric, not being described before. As mentioned before in the cases of VHL allosteric site and GID4, the mild restraints during the simulations allowed for conformational changes induced by the lipophilicity of the organic solvent. In a similar manner we observed the opening of cryptic sites in DCAF1 and RNF34. DCAF1-Pocket J (-5.34 kcal/mol) is situated close to the protein surface. On the other hand, RNF43 pockets opened in a deep cavity, which despite fulfilling the ligandability assessment, seemed unreachable for molecules much bigger than ethanol. Other than opening cryptic pockets, organic solvents can induce conformational changes in pre-existing pockets.[243]

The selected E3 ligases within the BTB family were KCTD5, SPOP, KEAP1 and GID4. KCTD5 and SPOP lack known small-molecule binders. The pentameric form of KCTD5 was investigated and revealed three potential binding pockets. Notably, KCTD5-Pocket E (-6.98 kcal/mol) and KCTD5-Pocket A1 (-6.46 kcal/mol) were found to be closely situated on the protein surface and of sufficient size to fit drug-like molecules (Figure 4.5 A). Additionally, KCTD5-Pocket H (-4.65 kcal/mol) was identified as a smaller fragment-sized pocket containing highly energetic hotspots, but in a less solvent accessible region. In a different structural context, for the SPOP E3 ligase, albeit from the same family, we predicted a highly energetic pocket. The SPOP-Pocket C (-7.63 kcal/mol) was located in a hinge region, behind the degron recognition site, which could introduce allosteric effects if altered (Figure 4.5 B). The last E3 ligase we studied of the

BTB family was GID4, which, as mentioned above, has a reported ligand. Two cavities were predicted to be druggable in GID4, outside the degron recognition site: pockets A and B (with the sum of hotspot energies of -6.47 and -0.03 kcal/mol respectively). Both pockets seem to be driven by hydrophobic interactions, which make envisage problems when designing small molecule binders.

In the previous sections, we saw difficulties reproducing the ligandability of the degron binding sites for the SOCS-BOX family, due to their surface geometries. Thus, they are a prime example on why allosteric pockets might be interesting. SOCS2-Pocket A (-8.71 kcal/mol), very close to the interface with Elongin C, opened during simulation and we hypothesize that it could be an interesting spot for modulation of the complex (Figure 4.5 C). SPSB1 presented very energetic hotspots, especially in the SPSB1-Pocket B (-6.92 kcal/mol). Similarly to the degron binding site, SPSB1-Pocket C (-6.66 kcal/mol) covered a flatter protein surface. However, in this case we detected possible buried polar hotspots that could be used as anchor points. ASB9 also presented two potential pockets: ASB9-Pocket H (-4.21 kcal/mol), which was small but had a high energetic hotspot, and ASB9-Pocket G (-4.61 kcal/mol), which also has the possibility of getting extended to the creatine kinase binding site.[244]

Despite the structural diversity of the TRIM family, many feature RING-finger, B1-box, B2-box and coiled-coil domains.[245] TRIM21 yielded potential allosteric sites, albeit each with their own shortcomings. TRIM21-Pocket D (-5.27 kcal/mol) yielded a high energetic profile but was rather small (i.e. fragment-sized). On the other hand, TRIM21-Pocket A (-6.73 kcal/mol) did not show good hotspots associated to a high energy interaction, but it can be

extended to other areas with a high density of possible interactions. In the separate simulations focusing on the RBCC domain and PHD-BRD domain of TRIM28, which exhibit ubiquitin E3 ligase activity and SUMO E3 ligase activity, respectively, distinctive characteristics were observed. Within this domain, two interesting allosteric pockets were discovered, with TRIM28 PHD-BRD-Pocket C (-7.28 kcal/mol) emerging as the most promising for ligand binding due to its deep pocket and favorable ligand efficiency (Figure 4.5 D). Allosteric pockets within the TRIM28 RBCC domain (pockets B and K) were situated in the central region of the dimer in the coiled-coil domain. The low coverage of the RBCC domain, hindered the assessment of these pockets.

WDR is the structural motif most represented in our chosen E3 ligases. From the E3 representing that family we found possible druggable pockets outside the degron recognition site in CDC20, FBO44, BTRC and FBW7. CDC20-Pocket A (-5.96 kcal/mol), despite being flat, was selected as a good candidate due to the high *hotspot efficiency* (-1.49 kcal/mol per hotspot) and the possibilities to extend it to a more buried region. The other selected E3s with the WDR motif are classified within the F-BOX family. Out of those, we found FBO44 to have only single a hydrophobic driven cluster of hotspots, but within a good surface geometry. On the other side of the WDR, BTRC-Pocket A (-9.7 kcal/mol) shows a hydrophobic core of hotspots with good polar interactions. In a similar relative position (i.e. *underneath* the WDR domain in the hinge next to the F-Box domain), FBW7-Pocket G (-8.76 kcal/mol) has two high energy polar hotspots and good geometry (Figure 4.5 E). In addition, FBW7-Pocket D (-5.31 kcal/mol) was found near the SKP1 adaptor protein.

Finally, the KELCH domain is represented by KEAP1 and KLHDC2 in our selection. Both E3 have a known, and well predicted by MDMix, degron recognition site in the same region of the KELCH domain.

However, possible allosteric sites appeared in different regions around the β-propeller in both E3 ligases. KEAP1-Pocket D (-10.14 kcal/mol) is situated between two β-sheets with a polar hotspot as the driving interaction. KLHDC2-Pocket E (-8.5 kcal/mol) shows two very potent hotspots and a good surface geometry. Moreover, the pocket is in the interface with a linker domain which could affect the conformational dynamics or the region, similarly to what we hypothesized with FBW7-Pocket G.

In summary, we have shown how the clustered interaction energies from mixed-solvent molecular dynamics are able to reproduce known binding sites (native substrate recognition sites and small molecule binding pockets). This information can be employed to optimize current small molecule binders. Additionally, we have found new possible pockets in non-orthosteric regions or even induced by the presence of organic molecules (Table 4.3). These hotspots might serve to design anchor points for future PROTAC development, but some also appear to be in key regions with potential for allosteric modulation of the UPS.

The good agreement between known ligand binding pockets and those predicted with MDMix is a good indication on the accuracy of the method. However, despite never using structures with ligands bound, it could be argued that the analysis is biased (e.g. shape filtering, minimum amount of pockets, etc). Also, our description of those pockets might not provide enough new information (e.g. in the form of new hotspots to grow molecules to). For these reasons, we chose the new, allosteric ocket of FBW7 to perform a prospective validation and discover and develop new binders.

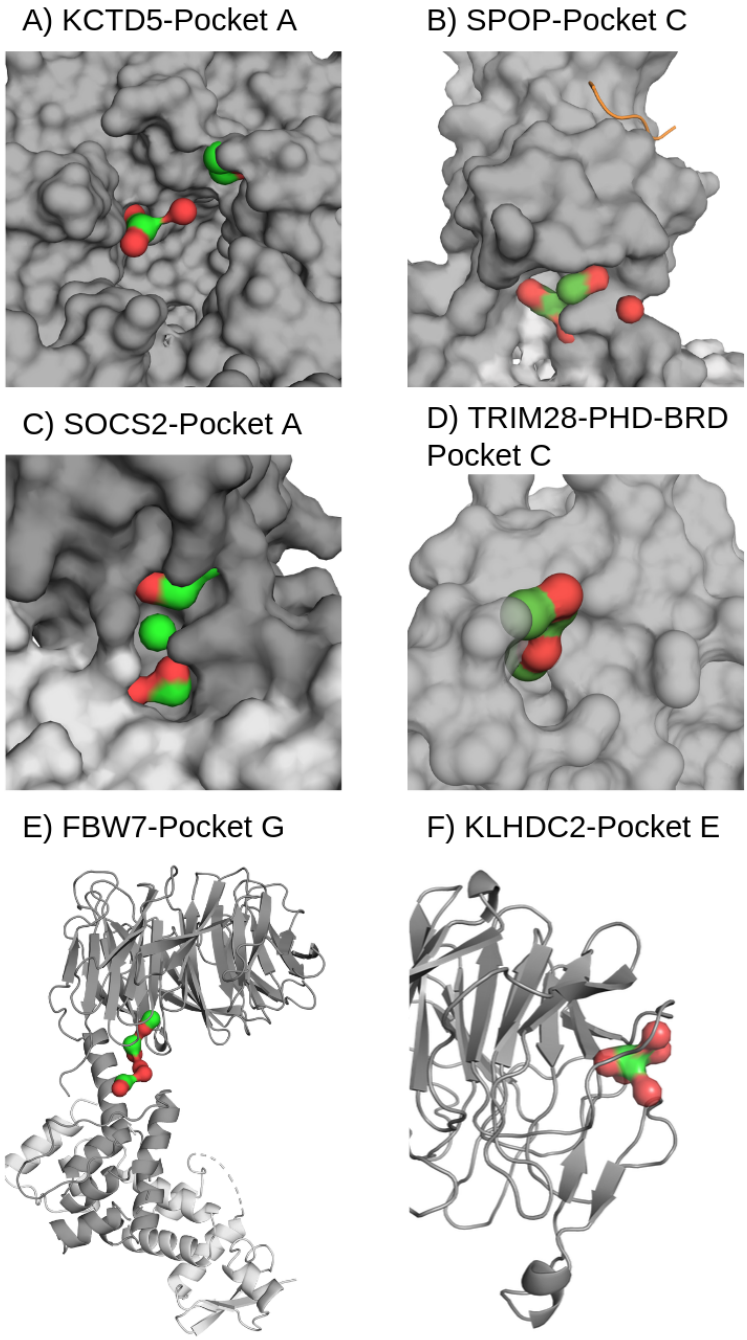


Figure 4.5. Structure of E3 ligases with predicted new allosteric pockets. Polar and hydrophobic hotspots are shown in red and green respectively. Adaptor proteins are colored in light tone while E3 are in a darker gray.

Table 4.3. Summary of predicted allosteric pockets with MDMix

E3 ligase	Known allosteric pocket with crystallized ligand	Pocket predicted matching known allosteric	Predicted new non-orthosteric pockets
ASB9	No	-	Yes
BTRC	No	-	Yes
CDC20	Yes	Yes	Yes
COP1	No	-	No
CRBN	Yes	Yes	Yes
DCAF1	No	-	Yes
DCAF15	Yes	Yes	Yes
FBO44	No	-	Yes
FBW7	No	-	Yes
GID4	No	-	Yes
KCTD5	No	-	Yes
KLHDC2	No	-	Yes
PRKN	No	-	No
RNF4	Yes (covalent)	No	No
RNF43	No	-	Yes
SKP2	Yes	No	No
SOCS2	No	-	Yes
SPOP	No	-	Yes
SPSB1	No	-	Yes
TRIM21	No	-	Yes
TRIM28	No	-	Yes
VHL	Yes	Yes	Yes

4.3 Prospective validation: identification of FBW7 small molecule binders

FBW7 is very commonly deregulated in cancer, with mutated forms found in approximately 6% of all reported cancer samples.[246] Most of its native substrates are key oncoproteins such as Cyclin-E, c-Myc, Notch, Jun, mTOR, Mcl-1 or AURKA.[247] Other FBW7 substrates such as DISC1 have been related to neuropsychiatric disorders. Given its relevance and the promising properties of the pockets encountered with MDMix, we decided to choose FBW7 for the prospective validation of the predicted pockets. As mentioned above, we reported three possible ligandable pockets. On the one hand, Pocket B was found on top of the WD40 domain, close to the degron recognition site. (Figure 4.6 A). The degron pattern recognised by FBW7 is a double phosphorylation motif associated with MAPK activity (Figure 4.6 B).[248,249] The three arginines (R465, R479, R505) that interact with substrate's phosphates negative charges are the main mutations found in cancer (Figure 4.6 C).[247] This presents both problems and opportunities when exploiting pocket B for drug Discovery purposes. In the wild-type conditions, we would expect that negatively charged compounds could bind this pocket, the associated limitations (e.g. permeability, pharmacodynamics). Moreover, we would expect those compounds to lose potency in the mutated FBW7 forms. On the other hand, if pocket B is preserved in the mutated variants, it could be a good anchoring point for PROTACs to recover function specifically in cancer cells.

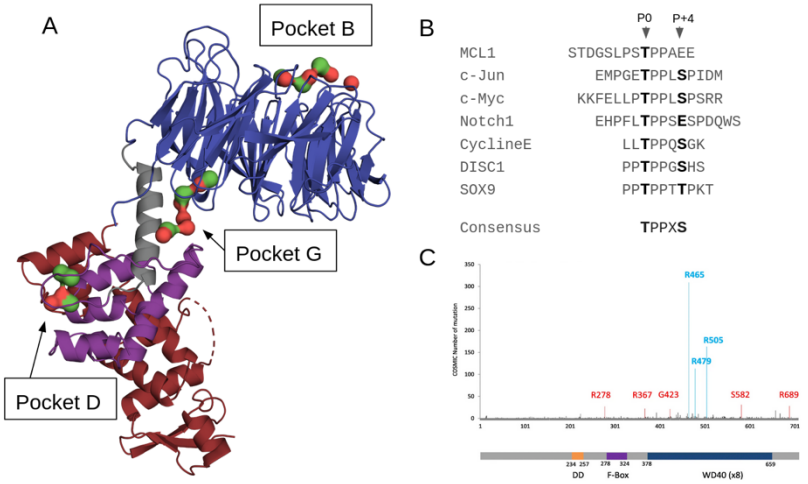


Figure 4.6. Predicted ligandable pockets, consensus degron motif and prevalent mutations in degron recognition site in FBW7. A) Cartoon representation of FBW7-SKP1 colored by domains (WD40 blue, linker gray, F-box purple and SKP1 red) with predicted pockets. Polar and hydrophobic hotspots are shown in red and green respectively. B) Degron sequence of some FBW7 native substrates' degrons, phosphorylated amino acids are marked in bold. C) FBW7 mutations and their prevalence in cancer from the COSMIC dataset, extracted from [247].

On the other hand, we found two pockets far from the degron recognition site with potential allosteric effects (Pocket D and G). Pocket D appeared in the interface between the F-Box and SKP1 and pocket G in the linker region between the WD40 and the F-FBOX domain (Figure 4.6 A). From these two pockets, the one with highest energy per hotspot was pocket G. In addition, the flexibility between the WDR and linked domains has been reported to play an important role in substrate ubiquitination in this family of UPS.[250]

Using pocket G hotspots as pharmacophoric restraints, Dra. Míriam Martínez performed a HTVS, followed by several biophysical techniques. In summary, she selected 41 compounds from the virtual screening and tested them through differential scanning fluorimetry (DSF) and surface plasmon resonance (SPR). The best hits from the

first screening were *expanded* through a SAR by catalog and validated with SPR. The most potent compounds were then submitted to a cell viability experiment in HEK293 cells, looking for a possible phenotypical effect when targeting Pocket G. Interestingly, we identified A5_MMC17 (in short A5), an analog of MMC17 (Figure 4.7), which was able to induce cell death in a concentration-dependent manner. A5 displayed a K_d of 40-60 μM (Figure 4.7), which was then confirmed by isothermal titration calorimetry (ITC). The binding site of MMC17 and its analogs was determined by Dra. Míriam Martínez and Roger Castaño with single-point FBW7 mutants, generated to disrupt the hotspots found in pocket G. The mutations that yielded a better signal were A677I and N679W, both introducing a bulkier residue blocking the entrance of the ligand to the pocket. A5 showed a 3 fold K_d decrease in SPR and a ~15 fold decrease in microscale thermophoresis (MST) (Figure 4.7). Moreover, Andrea Bertran-Mostazo demonstrated that the ligands do not compete with the substrate, by doing a fluorescence polarization displacing two different peptides of FBW7, DISC1 and c-Myc peptides (Figure 4.7). In collaboration with Dra. Bing Hao we tried to obtain the X-ray structures of A5 bound to FBW7. Even though no crystal structures with the ligand bound were resolved, we obtained apo structures of FBW7 with a sulphate molecule from the crystallization buffer bound to Pocket G.

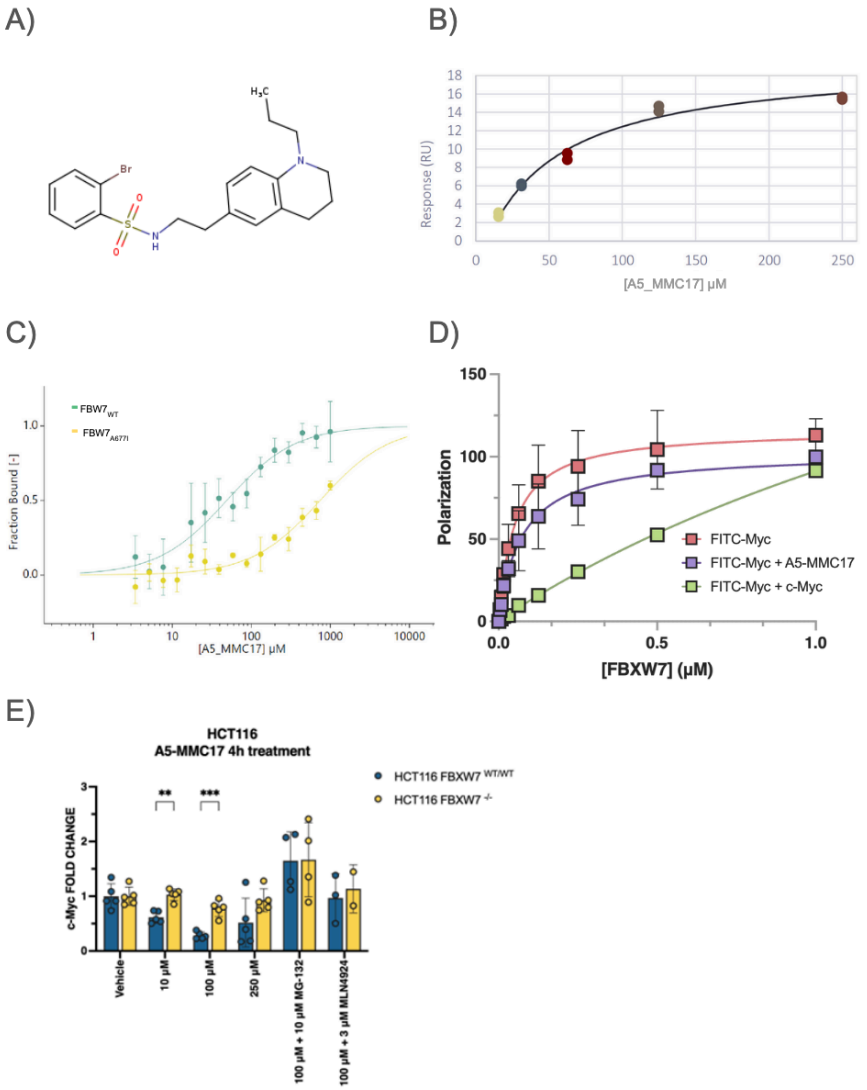


Figure 4.7. Experimental results of A5 binding to FBW7. A) The structure of A5 with the benzene-sulfonamide moiety at the left and the tetrahydroquinoline moiety at the right. B) SPR responses (RU) at the steady state for A5 at 5 concentrations, per duplicate. C) MST results of the FBW7 wildtype (green) and A677I mutant (yellow). D) Polarization changes in FP of the completely labeled c-Myc peptide (red) against A5 (purple) and against the unlabeled c-Myc(green). E) c-Myc protein quantification through Western blot in HCT116 FBW7 wildtype (blue) and knowckout (yellow).

In conjunction to this thesis, Andrea Bertran-Mostazo performed preliminary cell viability assays using A5 and other members of the family on HEK293 cells, which showed a dose dependent decrease in cells. When analyzing the concentration of possible FBW7 substrates related to cell cycle/survival, she found a dose dependent decrease in the levels of c-Myc. This effect was corroborated to be proteasome-dependent and FBW7-dependent, through the use of proteasome and neddylation inhibitors and FBW7 knock-out cells.

4.3.1 Simulating the SCF^{FBW7} complex

FBW7 is part of the SKP1, Cullin, F-Box containing (SCF) family of UPS. The E3 ligases in this group bind to SKP1 through their F-Box domain, which is connected to the E2 through the Cullin. The FBW7 degron recognition site is situated in the WD40 domain, far from the E2 (approximately 50-70Å). Lui & Nussinov hypothesized that modulating the linker motions of the SCF^{FBW7}, bringing closer the WDR domain to the E2, might facilitate the ubiquitin transfer, resulting in higher degradation profiles.[250] Seeing that the Pocket G, where A5 binds, is situated in the linker domain and it induces the degradation of c-Myc in a dose-response manner, I performed a series of long MD simulations to assess if A5 modulates the SCF^{FBW7} conformation dynamics.

Despite there not being resolved crystal structures of the complete SCF^{FBW7} complex, Schulman's group reported a SCF-RBR super-assembly with the SKP2 and ARIH1 E3 ligases, using cryogenic electron microscopy (cryo-EM).[189] SKP2 forms the SCF^{SKP2}, which is identical to SCF^{FBW7} only changing the SKP3 E3 ligase (Figure 4.8). To build the initial model, I used SKP1-CUL1 from the 7B5L PDB entry and aligned it with the FBW7 structure employed in the HTVS

campaign done by Míriam (i.e. a specific frame from the MDMix simulations starting from the 2OVP PDB entry, where pocket G becomes slightly wider). We discarded the terminal part of CUL1, the NEDD8, the RBX1 (E2 adaptor) and the E2 to limit the simulation box size. We expected the ligand to only have an effect on FBW7 and SKP1 at most, but not as far as RBX1 and the E2. Moreover, the construct used by Schulman is very atypical with two E3s. Under the assumption that the E2 position will not be affected by A5, we used the distance between FBW7's WD40 to the end of CUL1 as a proxy for the distance between the substrate and the E2 (UBE2L).

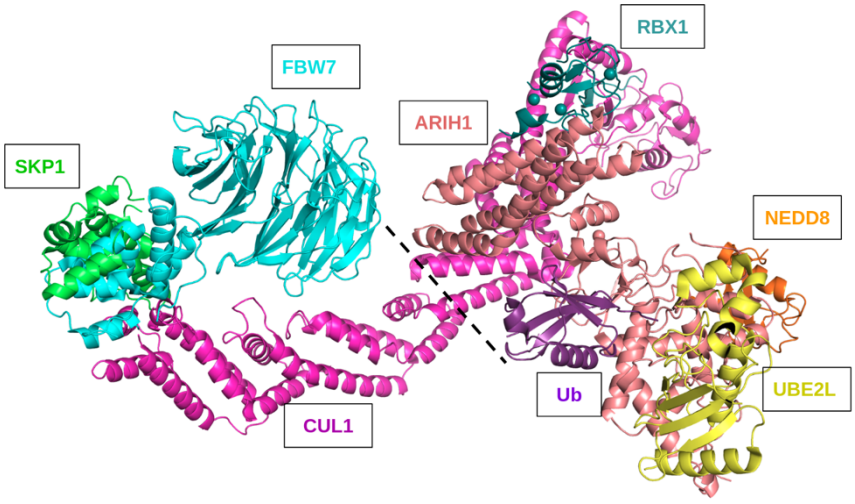


Figure 4.8. Structure of the SCF^{FBW7} complex model created from 7B5L and 2OVP. FBW7 (cyan) and SKP1 (green) come from a simulated frame of 2OVP and have been aligned to 7B5L's SKP1. CUL1 (pink), RBX1 (turquoise), ARIH1 (pastel), Ub (violet), NEDD8 (orange) and UBE2L (yellow) are part of the SCF^{SKP2}-RBR super-assembly from 7B5L. The black slashed-line marks the part where CUL1 was cut for the simulations.

I simulated 6 replicas of 1.5 μs each for the three following conformations: apo, holo and holo without the ligand. While for the apo SCF^{FBW7} conformation I used the 2OVP FBW7 coordinates, for the holo I used the MDMix frame. The holo without ligand simulations

were done to eliminate the possible confounding effect of the initial FBW7 conformation.

4.3.2 Conformational analysis of the SCF^{FBW7}

A pool of the apo, holo and holo-without-ligand simulations was decomposed in their moments of inertia using a PCA of their CA coordinates. We found four main components covering 92% of the variability (Figure 4.9). From the projection of the moments' limits we can elucidate an approximation of the motions stored in each component. PC1 matches with the motions between the WD40 and the E2 described by Nussinov,[250] albeit with different distance magnitudes. As stated before, I used the end of the CUL1 as a proxy for the position of the E2. In PC2, there is a *side-to-side* orientation change of the WD40. Measurements shown in the Figure 4.9 correspond to the furthest WD40 residue to the plane formed by CUL1. In the PC3, we see a change in the angle between the linker and the WD40. This angle change creates a shortening in the WD40-to-CUL1 distance as in PC1, moreover it is also present in PC1. Finally, PC4 shows a *breathing* of the WD40 domain with a change in diameter from 43 to 50 Å. The three first components describe motions that affect the presentation of the substrate to the E2 and all involve the WD40. Thus it is plausible that the ligand, binding to pocket G, might have an effect on those motions.

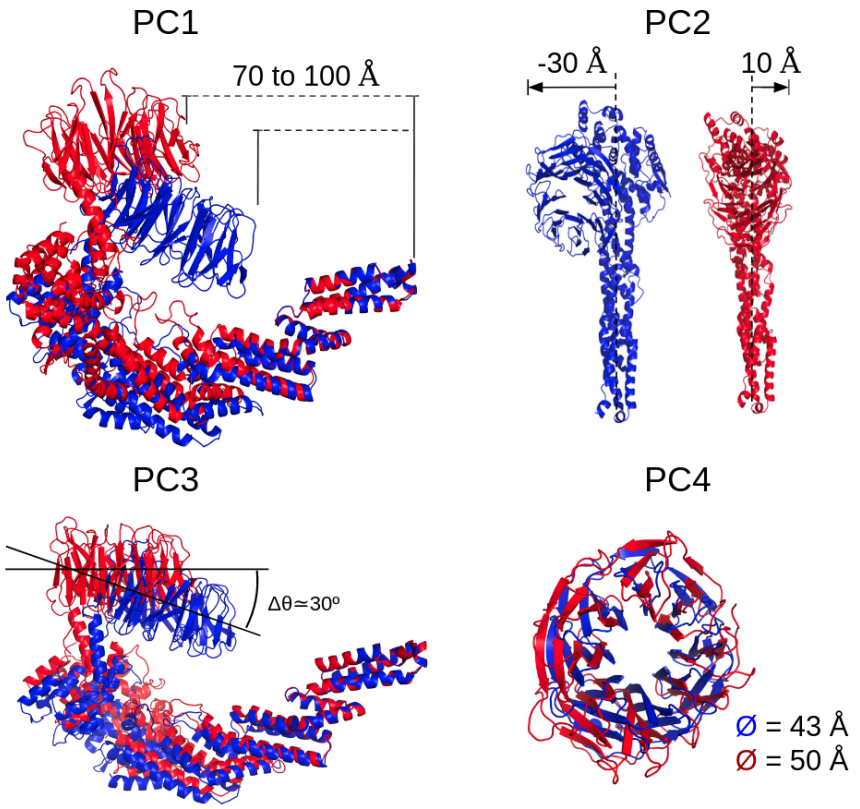


Figure 4.9. Projection of the four first principal moments of SCF^{FBW7} simulations. Each moment of inertia is represented by the coordinates' projection of its two extreme values (blue and red respectively). Measurements shown in the projections are simplifications of the variability in each component.

I projected the principal components of the motions described before separated by conformation (holo, apo and holo-without-ligand) (Figure 4.10). The distribution for the first component shows a gradient from apo to holo-without-ligand to holo, which indicates that the initial holo confirmation has an impact on the motions. However, the fact that holo-without-ligand is shifted towards apo also highlights the effect of the ligand in such motion. A clear separation can be seen in the third component, where the WD40-to-linker angle changes. Pocket G is in the hinge between the linker and WD40, and

the presence of A5 modulates the hinge motions. The differentiation of holo and the ligandless conformations agree with the hypothesis of A5 modulating FBW7 through changes in the distance and presentation of the substrate to the E2.

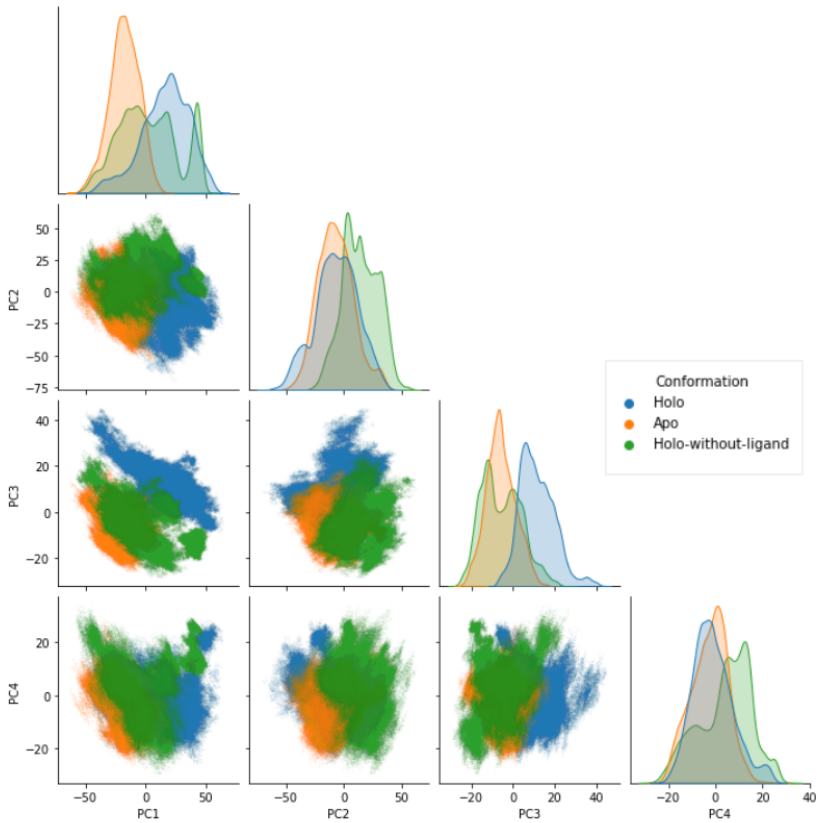


Figure 4.10. Principal component distributions of the three simulated SCF^{FBW7} conformations. Holo, apo and holo-without-ligand are represented in blue, orange and green respectively.

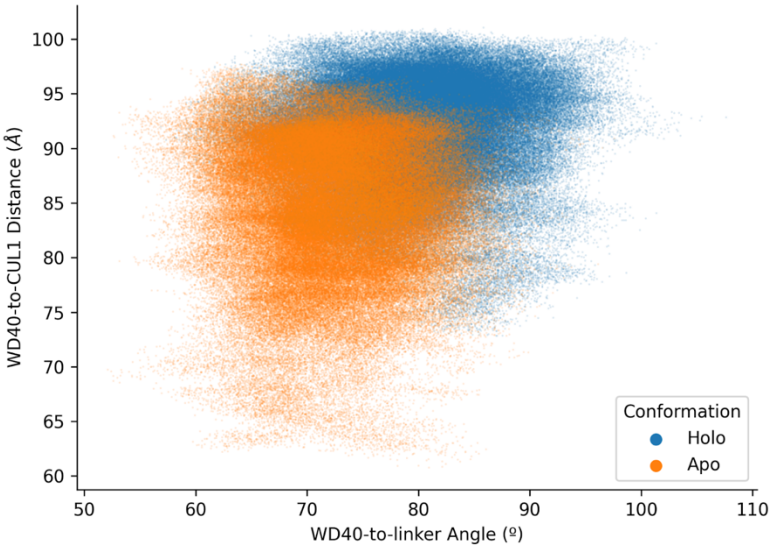


Figure 4.11. Comparison of distance and angle in apo and holo SCF^{FBW7} simulations. Y-axis and X-axis correspond to the distances and angles represented in Figure 4.9-PC1 and Figure 4.9-PC3 respectively.

Despite the complexity of the motions described by each component, we can assume that PC1 and PC3 are the ones being more influential in the distance between the degron recognition site in the WD40 and the CUL1. When comparing the simplified measurements I extracted from both components between apo and holo we can see a clear separation. (Figure 4.11). Apo conformations show a shorter distance between the WD40 and the CUL1 and a smaller angle with the linker, while holo conformations show the opposite. This contradicts the hypothesis of A5 promoting a conformation where the substrate is presented or stabilized closer to the E2 for its ubiquitination. Nonetheless, the simulations evidentiate a conformational change of the SCF^{FBW7} induced by A5.

With these results we discarded the increased proximity to the E2 as the modulating effect of A5 that induces the degradation of c-Myc.

However, due to the long and intrinsically disordered nature of c-Myc degrons, a shorter distance might not be necessary. We hypothesized that a more stable conformation, despite being *extended,* might favor the presentation and ubiquitination of c-Myc lysines by the E2. Recent results show that compounds from the A5 family might also induce the degradation of c-JUN, another natural substrate of FBW7 which shares many structural features such as long intrinsically disordered regions.

4.3.3 Optimization of FBW7 modulators through a SAR by catalog

Considering the experimental results obtained by Andrea Bertran-Mostazo, we decided to try optimizing the binding affinity of the A5 family towards FBW7. During the more than 12 μs of accumulated simulation time, A5 didn't unbind the pocket G once. However, the binding mode obtained by docking was not stable. Early in the equilibration A5 (in all replicas) the tetrahydroquinoline moiety changes from a T shaped Π-Π interaction with W365 to a displaced stacking. In some cases it fluctuates back to the original interaction, but with a different conformation for the rest of the molecule. Mainly, the H-bond used to steer during the DUck simulations on the initial virtual screening oscillates between N635, N633 and N679 on the loop at the other side of the pocket. I clustered the A5 conformations and obtained 6 different clusters, which we can group into two main binding poses: clusters 0 and 2 on one hand and clusters 1 and 5 on the other hand. In both binding poses, the shared feature is the orientation of the tetrahydroquinoline, which within each cluster stays stable (Figure 4.12). The benzenesulfonamide moiety, however, is unstable and quickly changes conformation and

orientation. This can be seen when measuring the root mean square fluctuations by atom (Figure 4.13).

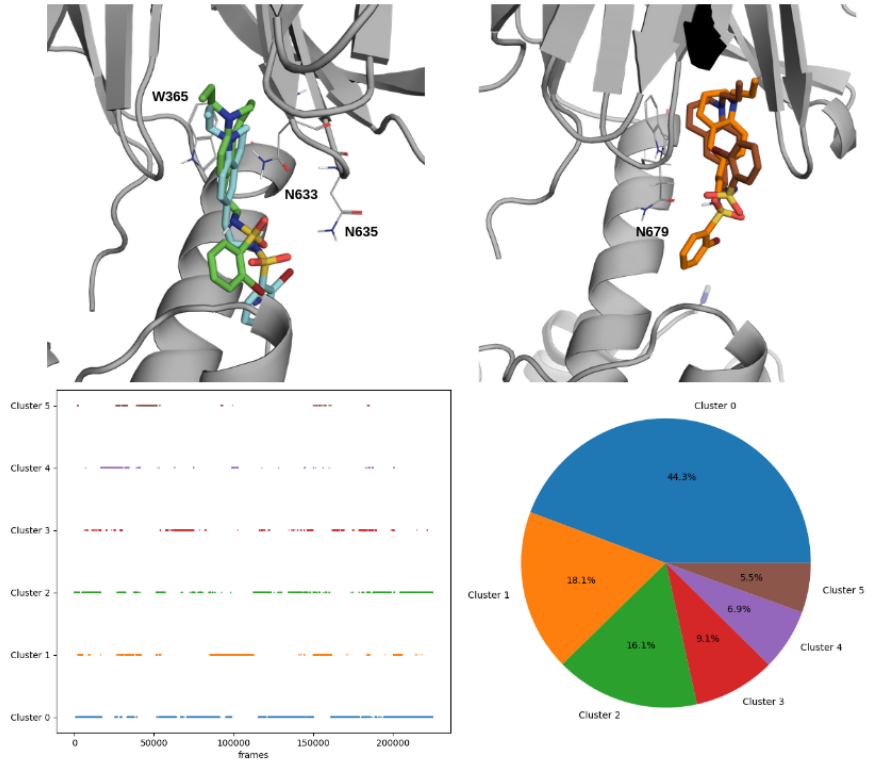


Figure 4.12. A5 binding mode clustering during the SCF^{FBW7} simulations. Clusters 0 and 2 (blue and green respectively) interact with W365 in a displaced Π-Π stacking (top-left). Clusters 1 and 6 (orange and brown) displace the tetrahydroquinoline moiety to a T shaped Π-Π interaction and detach from N635 to bind N679 (top-right). Both cluster groups are present in all replicas and have a high exchange ratio, however, the group formed by clusters 0 and 2 make are predominant (bottom).

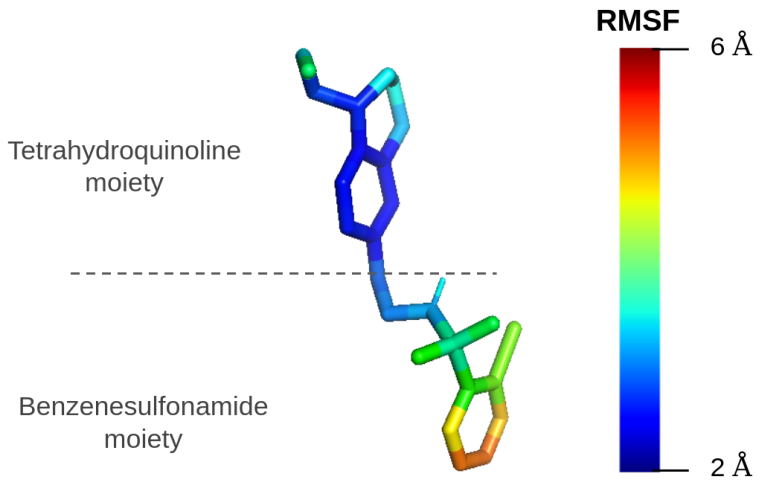


Figure 4.13. A5's root mean square fluctuations (RMSF) by atom during the MD simulations of the SCF^{FBW7} complex.

Based on the stable binding mode of the tetrahydroquinoline moiety, I looked for analogs to find alternatives for the benzenesulfonamide moiety. I performed a substructure search in Enamine's RealSpace with the intention of finding groups that could link the two aromatic rings while doing H-Bonds with the asparagines at the sides of the pocket. The quinoline-to-linker substitution position was maintained. Finally, different aromatic substituents after the linker were also enriched in the dataset. A total of $3.5x10^6$ compounds matched the structural requirements. After filtering for Lipinski, PAINS and solubility, the number of compounds to dock reduced down to $4x10^5$. I performed two parallel tethered docking procedures using the (i) original docking and (ii) cluster 0 A5's binding poses as reference respectively. The molecules were ranked through a consensus score from both tethered dockings using SCORE.INTER.norm. I used the scoring normalized by heavy atoms as there was a clear overestimation of big ligands in the docking score, as reflected on

the correlation between SCORE.INTER and SCORE.VDW (Figure 4.14)

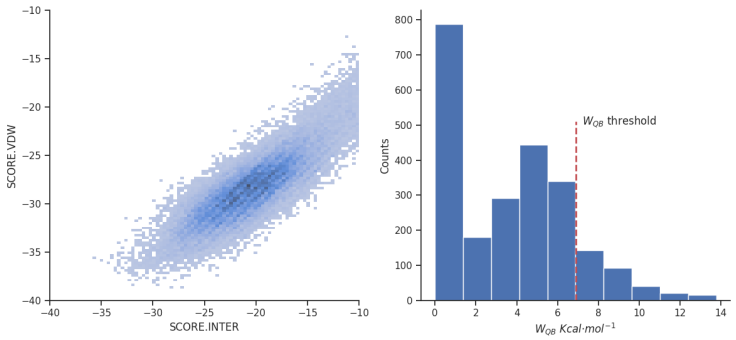


Figure 4.14. Virtual screening scores of the A5 analogs. On the left a scatter plot comparing SCORE.INTER and SCORE.VDW from the binding poses obtained from both receptor conformations. On the right, a distribution of W_{QB} values highlighting the threshold used.

A diverse subset of compounds was selected using hierarchical clustering based on Tanimoto coefficient. Dynamic undocking was performed for the 2360 representatives using openDUck with a W_{QB} threshold of 6 kcal mol^{-1} (Figure 4.14). W_{QB} and SCORE.INTER are orthogonal descriptors (Figure 4.15). Andrea and I selected compounds with a W_{QB} higher than 10 kcal mol^{-1} to manually inspect them. We prioritized diversity in the linker region (i.e. sulfonamides, amides, ureas and other linkers with the ability to perform an H-bond with N365) and the terminal group (bigger aromatic groups and different substitutions of the benzene). A final selection of 15 compounds was requested to Enamine, from which 11 were successfully synthesized.

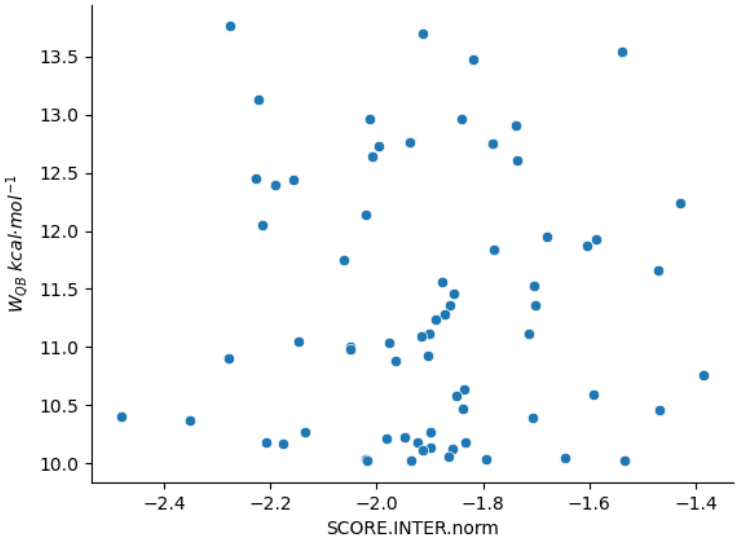


Figure 4.15. Relationship between W_{QB} and SCORE.INTER.norm in A5 analoges.

Andrea Bertran-Mostazo tested the 11 A5 analogues (named ABM1-11) by SPR (Figure 4.16) and in cell assays. ABM7, the closest compound to A5, did not show an improvement in the K_d (Table 4.4). The same happened with amide linkers. On the other hand, ABM8 shows a comparable K_d to A5. ABM8 linker is an urea directly attached to the terminal aromatic group, which makes it a very rigid ligand. This might help stabilizing the movement while maintaining the N365 H-bond. The other compound with an urea (ABM3) has higher flexibility and worse K_d.

None of the A5 analogs, even ABM8 which has a comparable K_d, induced a significant degradation of c-Myc in the cell assays. Thus, the lab efforts have centered on finding a more holistic understanding of the changes induced by A5 that might explain the UPS-dependent degradation of c-Myc (i.e. alternative E3 ligases,

intermediate substrates upstream or completely different degradation mechanisms).

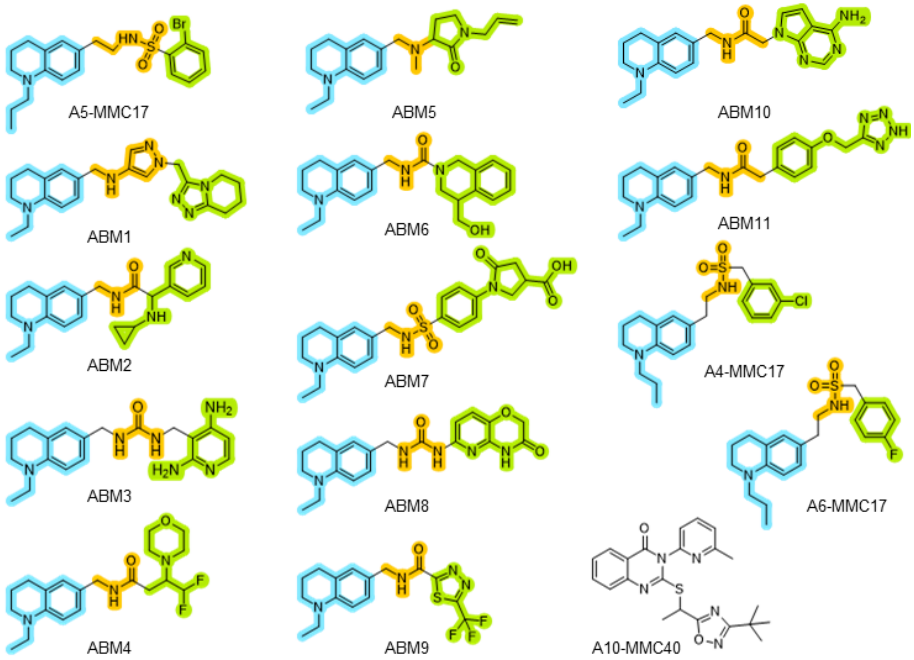


Figure 4.16. Structure and k_d of A5 (A5-MMC17) and its analogs. The tetrahydroquinoline moiety (blue), linker (orange) and terminal groups (green) are highlighted.

Table 4.4 Summary of the SPR results for the A5 analogs. K_d are represented as the average ± standard deviation of two independent replicates

Compound	K_d (μM)
A5-MMC17	67,9 ± 7,7
A6-MMC17	60,1 ± 4,8
A4-MMC17	139,6 ± 0,2
A10-MMC40	410,4 ± 8,8
ABM1	150,2 ± 0,8
ABM2	1006,6 ± 14,8
ABM3	204,9 ± 1,3
ABM4	2165,0 ± 4,2
ABM5	598,6 ± 0,4
ABM6	165,3 ± 3,0
ABM7	148,5 ± 31,0
ABM8	40,4 ± 23,2
ABM9	457,0 ± 131,7
ABM10	546,4 ± 4,8
ABM11	193,8 ± 1,8

CHAPTER 5: RESULTS

BOTTOM-UP EXPLORATION OF THE CHEMICAL SPACE

5. BOTTOM-UP EXPLORATION OF THE CHEMICAL SPACE

5.1 Background

The past years have seen major advances in the scale and success of virtual screening (VS) campaigns.[251–254] Traditionally, the scope of VS was limited to compounds either from physical in-house collections or in-stock collections available for purchase from vendors. Generally, the bigger these collections are, the higher chances of finding better hits in them; assuming that they maintain a healthy chemical diversity.[255] Thus, these collections have been steadily growing, reaching the tens of millions (e.g. MolPort, Enamine, ZINC).[256] This figure pales in comparison with the emergent on-demand chemical collections. Enamine was the first vendor that offered large on-demand libraries for drug discovery with their REAL database which comprised approximately 10^8 compounds in 2014. Today, the biggest commercially available on-demand libraries contain the order of hundreds of billions of compounds (10^{11}) and are expected to reach the trillion (10^{12}) very soon, and continue growing towards the theoretical accessible chemical space. The key success factor for these ultra-large libraries has been the high delivery rate, and speed in the synthesis of compounds, as well as their relatively low cost. Moreover, the limitations in enumeration and distribution of such libraries through the traditional strategies has driven the appearance of innovative alternatives.[257] Nowadays, most of these libraries are stored in graphs made of building blocks and reactions (i.e. the so-called

fragment spaces), and compounds are only enumerated upon extraction (section 3.1.1.1).

Despite the potential of these chemical collections to revolutionize the drug discovery process, the computational cost to navigate, enumerate, prepare, dock and analyze these compounds is prohibitive. Nonetheless, some authors have done exhaustive docking of the complete libraries and obtained new very potent binders.[252–254,258] These campaigns were possible thanks to the usage of high performing computers (HPC) and cloud computing facilities (e.g. AWS, Google Cloud). However, brute-forcing the docking calculations of the ultra-large libraries is not scalable nor sustainable with the current growth of the libraries. Thus, new strategies have been proposed to efficiently find the best binders without exploring the complete library.

One of the most explored strategies has been focused on using machine and active learning (ML/AL) methods to facilitate ultra-large VS.[259–262] The workflows are based on docking a small subset of the chemical library and to train an ML model to predict the docking score. Then, using that model as criterion they dock a new set of molecules and retrain the ML model. Despite ML predictions being faster than docking calculations, the iterative nature of the algorithms incorporates a heavy computational cost. Moreover, many of the ML models require the enumeration of molecules. Alternatively, generative methods that directly use the library's building blocks and reactions avoid the enumeration problem.[263,264] However, these algorithms still frequently propose molecules impossible to synthesize.[265]

Another strategy that exploits the building blocks and reactions graph format of the ultra-large chemical libraries is the synthon-based virtual screening. Here, the building blocks (or synthons) are prepared and docked independently. Then for the synthons with best docking scores, a second synthon is attached and minimized. The first synthon-based strategy on ultra-large VS was V-SYNTHES, where they found nanomolar hits for the cannabinoid receptors CB1 and CB2.[266] Similar strategies have appeared recently.[267,268] These strategies are very similar to what was traditionally understood as fragment growing, albeit constrained to the building blocks and reactions within the fragment space libraries. Despite their success, these approaches are constrained by the available building blocks for the initial *graph expansion* and the vector of growth that building blocks allow. Moreover, these strategies require proprietary knowledge about the composition of the database and each collection has to be explored independently. Perhaps more importantly, the selection of compounds rely exclusively on docking scoring functions, that present severe limitations, which result on Thus, it is still necessary to find a strategy capable of navigating the chemical space, capitalizing on the graph structure of the ultra-large libraries without being constrained by the building blocks.

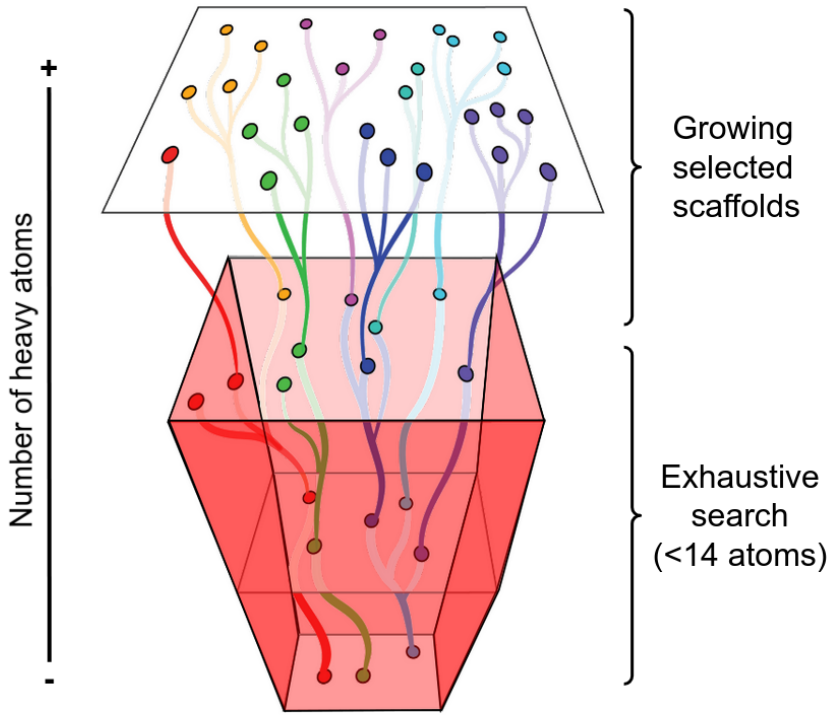


Figure 5.1. Schema of the bottom-up exploration of the chemical space. The low molecular weight space, represented at the bottom with the red boxes has fewer possible molecules and scaffolds. The scaffolds are represented by the colored dots. The scaffolds can be followed to the higher molecular weight space through the colored lines, growing them to drug-like molecules.

To address this challenge, we have developed a novel strategy that explores the chemical universe from the bottom up (Figure 5.1). First, performing a systematic exploration of the low molecular weight chemical space, but rapidly focusing on the most promising areas by growing the selected scaffolds into bigger molecules from the collections. We consider that an exhaustive search of the low molecular weight space will always be manageable, as even the theoretical space of under 14 heavy atoms peaks at 10^9 molecules.[269] Constraining the exploration around promising scaffolds allows us then to comprehensively enumerate scaffold-focused libraries and evaluate the complete available chemical

series. In contrast to other published strategies, we apply a hierarchy of increasingly sophisticated computational methods, which aims to maximize the success of the selected compounds. Docking is believed to be more proficient in ranking diverse chemotypes rather than differentiating closely related derivates of the same scaffold.[270] Thus, using more accurate, while still high-throughput, methods such as DUck allows us to explore scaffold-focused areas of the chemical space.

5.2 Validating the pipeline through three scenarios

In order to validate separately the different stages of the bottom-up chemical space exploration protocol we devised a validation based on three scenarios. The first scenario covers the whole protocol, from an initial druggability assessment to a fragment screening and then the scaffold expansion on the chosen fragments. The second scenario implies using fragments with a known binding mode (i.e. experimentally proved) and growing them, without knowing which are the optimal growing vectors. The final scenario is deconstructing scaffolds from drug candidates and growing them, having in mind which should be the optimal decorations and growing vectors. To this end, we chose the Bromodomain-containing protein 4 (BRD4), more specifically its bromodomain 1 (BD1). BRD4 is a member of the BET family and has been thoroughly studied, having ligands that fulfill the validation scenarios we devised. Moreover, the group had experience working on this protein and we had some experiments set up beforehand. Finally, we selected five fragments from the fragment-based virtual screening performed by Marina Miñarro,[200] three fragments from the PDB (PDB ids: 4LZS, 6ZED and 6ZF9) and three drug candidates (ABBV-075, IBET-151 and JQ1).

5.3 Exhaustive search of the fragment space

The exhaustive search of low molecular weight compounds for the third scenario was performed by Marina Miñarro, a lab colleague. She performed a fragment VS of the compounds with less than 14 heavy atoms from Enamine's Real database and ZINC20. A total of more than 1 billion compounds were docked using rDock with pharmacophoric restraints (Figure 3.5) and a HTVS filter. To determine the score thresholds for docking and DUck, she used a set of known BRD4 binders in ChEMBL and decoys, maximizing the enrichment (Figure 5.2). Finally she selected five diverse fragments with stable binding modes and good work profiles as the first scenario for the fragment growing module.

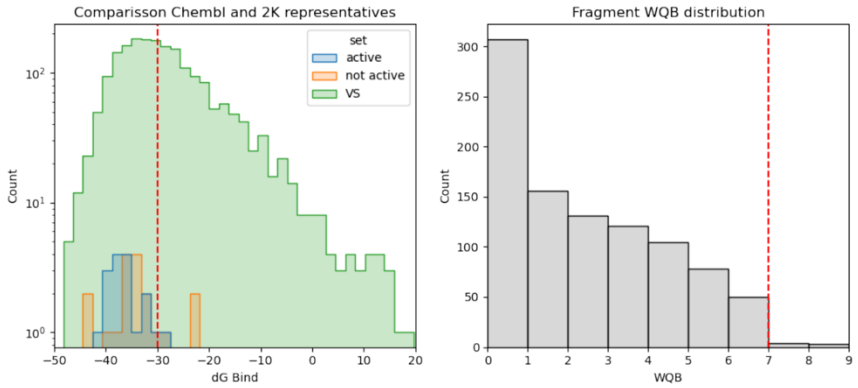


Figure 5.2. Results for the fragment virtual screening. a) Distribution of ΔGbind for the active ChEMBL set, the non-active ChEMBL set and the fragments obtained in the VS. b) Distribution of WQB values for the VS fragments. The red dotted lines represent the selected threshold values. Extracted from [200].

5.4 Scaffold Growing

5.4.1 Navigating ultra-large libraries

I performed the scaffold growing from the eleven scaffolds from the three. The main idea of the bottom-up pipeline is to perform a comprehensive search at the fragment space level, and then deepen on particular chemical space regions. But the definition of chemical space needs not be limited around an individual building block. Instead, for each scaffold, we derived a SMARTS pattern that aimed to capture the essential components of the scaffold, treating the interacting atoms as invariant, but allowing for significant changes in other parts of the structure (Figure 5.3). For each scaffold, a SMARTS pattern was derived with the intention of representing the information known for each scenario (Table 5.1).

Table 5.1. SMARTS queries for the construction of the scaffold-focused libraries with SpaceMACS

System		SMARTS query
Drug Candidates	ABBV-075	[CH3]a1aa([R])a2aa[nH]a2c1=O
	IBET151	a1[nh0,o;D2][nh0,o;D2]a([CH3])a1(a1aaaaa1)
	JQ1	[CH3]a1[nh0;D2][nh0;D2]a(C)a1c2scaa2C
Experimental Fragment Hits	4LZS	a1a([*][#7;D2;h1])[nH]a(C)a1C([*])=O
	6ZED	[CH3]a1[nh0;D2][nh0;D2]a2[nH]aaa2a1[*]
	6ZF9	[CH3][*]a1[nh0;D2][nh0;D2]a2saa([R])a12
Virtual Fragment Hits	Comp1	[nH0]1saac1C(=O)
	Comp2	O=c1[nD2]c([CH3])[nD2]cc1
	Comp4	o1[cH][cH]cc1C=O
	Comp5	n1nc([D1])sc1[NH]
	Comp6	[C,Cl;D1]c1aaa2aa[nD2;H0]a2[nD2;H0]1

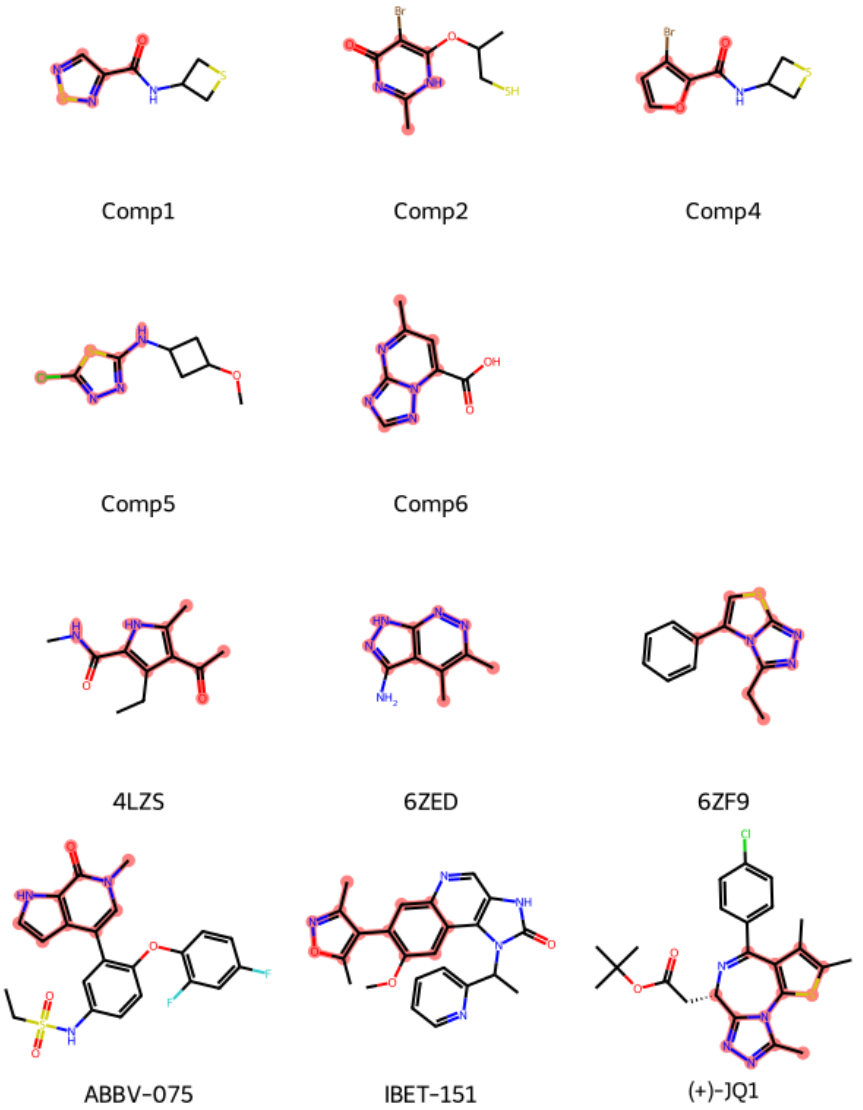


Figure 5.3. Structure of the reference compounds with the selected substructure highlighted.

The main advantage of using SMARTS, rather than finding the most similar synthon, is that we can guide the desired alternatives and growth vectors, thus fine tuning the output. Using the SMARTS as a query for SpaceMACS, we aimed to retrieve as many compounds as

possible from the REALSpace. We set a self-imposed maximum of 20M per query, which we deemed a reasonable balance between exploration, exploitation and throughput, due to hardware constraints. We only exceed this number in the case of the virtual fragments Comp1 and Comp4, due to dividing the searches by the output's heavy atom count (HAC) (i.e. the first search limiting the HAC within the 25-30 range and the second between 30-35 heavy atoms). This was done because the Comp4 and Comp1 scaffolds are very well represented in the collection, and the 20M compounds subset covered a smaller region. For the scaffolds where the 20 million threshold was not reached, we considered that the whole space around the scaffold was explored.

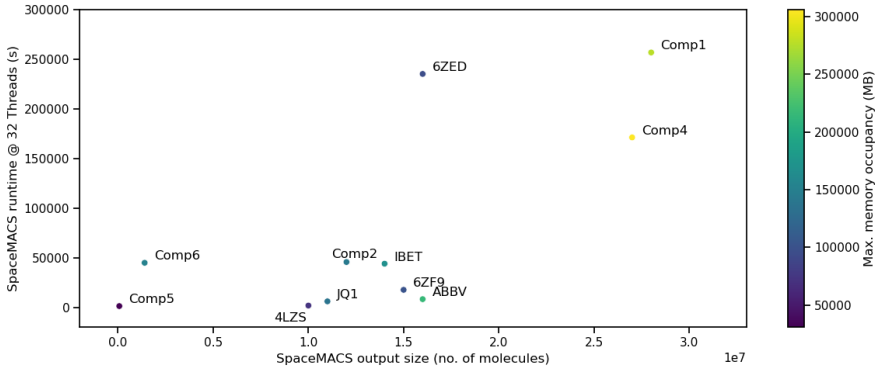


Figure 5.4. Relationship between SpaceMACS's quantity of output molecules and its computational cost. The colorbar refers to the maximum RAM memory allocated during the query and the y-axis corresponds to the wallclock time of the job, including writing the output.

Interestingly, the computational cost of the searches was very stable in spite of the matches found, but spiked for the Comp1, Comp4 and 6ZED scaffolds (Figure 5.4). I hypothesize that these scaffolds match with many starting building blocks. Thus, SpaceMACS needs to enumerate many reaction trees, which is the slowest step. This is consistent with the Comp1 and 4 scaffolds yielding more than 20

million compounds during the search. The memory requirements for the calculation are more consistent with the output size, however they were very high (i.e. between ~50 to 300 GB). New versions of SpaceMACS seem to have improved the memory efficiency of the search. Finally, after filtering for drug-likeness, the generation of 3D structures resulted in approximately 3.5 conformers per ligand. The resulting eleven scaffold-focused libraries add a total of approximately 250 million conformers to dock (Table 5.2).

Table 5.2 Summary of the computational results for the fragment growing pipeline

Scenario / Scaffold	Virtual Fragment hits (full pipeline)					Experimental Fragments			Drug Scaffolds		
	Comp1	Comp2	Comp4	Comp5	Comp6	4LZS	6ZED	6ZF9	ABBV-075	IBET-151	JQ1
SpaceMACS output	$2.8 \cdot 10^7$	$1.2 \cdot 10^7$	$2.7 \cdot 10^7$	$7 \cdot 10^4$	$1.4 \cdot 10^6$	10^7	$1.6 \cdot 10^7$	$1.5 \cdot 10^7$	$1.6 \cdot 10^7$	$1.4 \cdot 10^7$	$1.1 \cdot 10^7$
Prepared drug-like conformers	$1.6 \cdot 10^7$	$2.7 \cdot 10^7$	$1.6 \cdot 10^7$	$3 \cdot 10^5$	$2.8 \cdot 10^6$	$3.3 \cdot 10^7$	$3.6 \cdot 10^7$	$2.7 \cdot 10^7$	$2.9 \cdot 10^7$	$2.8 \cdot 10^7$	$2.6 \cdot 10^7$
Docking output	$2.7 \cdot 10^6$	$3.8 \cdot 10^6$	$2.2 \cdot 10^5$	$2.4 \cdot 10^3$	$5.3 \cdot 10^5$	10^6	$8.5 \cdot 10^5$	$7.3 \cdot 10^5$	$1.6 \cdot 10^6$	$2.8 \cdot 10^6$	$5 \cdot 10^5$
W_{QB} threshold ($kcal \cdot mol^{-1}$)	8	9	8	8	7	10	9	6	9	7	9
DUck output	13	41	14	13	12	72	11	15	23	16	13
Compounds synthesized	5	6	4	8	9	10	6	8	10	10	9

5.4.2 Virtual screening

5.4.2.1 Tethered docking

The compounds were docked to the BRD4-BD1 cavity with the same conditions as for the fragment screening (i.e. receptor, pharmacophoric restraints, HTVS filter), but tethering in place the atoms matching the scaffold defined by the SMARTS pattern. From here onwards, the pipeline works under the assumption that the binding mode of the main scaffold remains unaltered when the molecules grow in size (Figure 5.5). Thanks to the HTVS filter, molecules where the growing vectors clashed with the receptor were discarded early, reducing significantly the computational cost. The filter sets a minimum SCORE.INTER and SCORE.RESTR.PHARMA which become stricter through the rDock iterations. This influenced the resulting distribution of rDock scores, as seen in Figure 5.6. Interestingly, the distributions of SCORE.INTER are near-Gaussian for most compounds, with the averages oscillating between -25 and -15. However, ligands derived from some scaffolds (e.g. 4LZS, ABBV) had much worse scoring and thus, only the left tail of the distribution was kept. SCORE.RESTR.PHARMA is a sum of the penalty scores given by not fulfilling the pharmacophoric restrictions. These seemed to depend mostly on how well the initial tethered binding modes fulfill the pharmacophores (Figure 5.5).

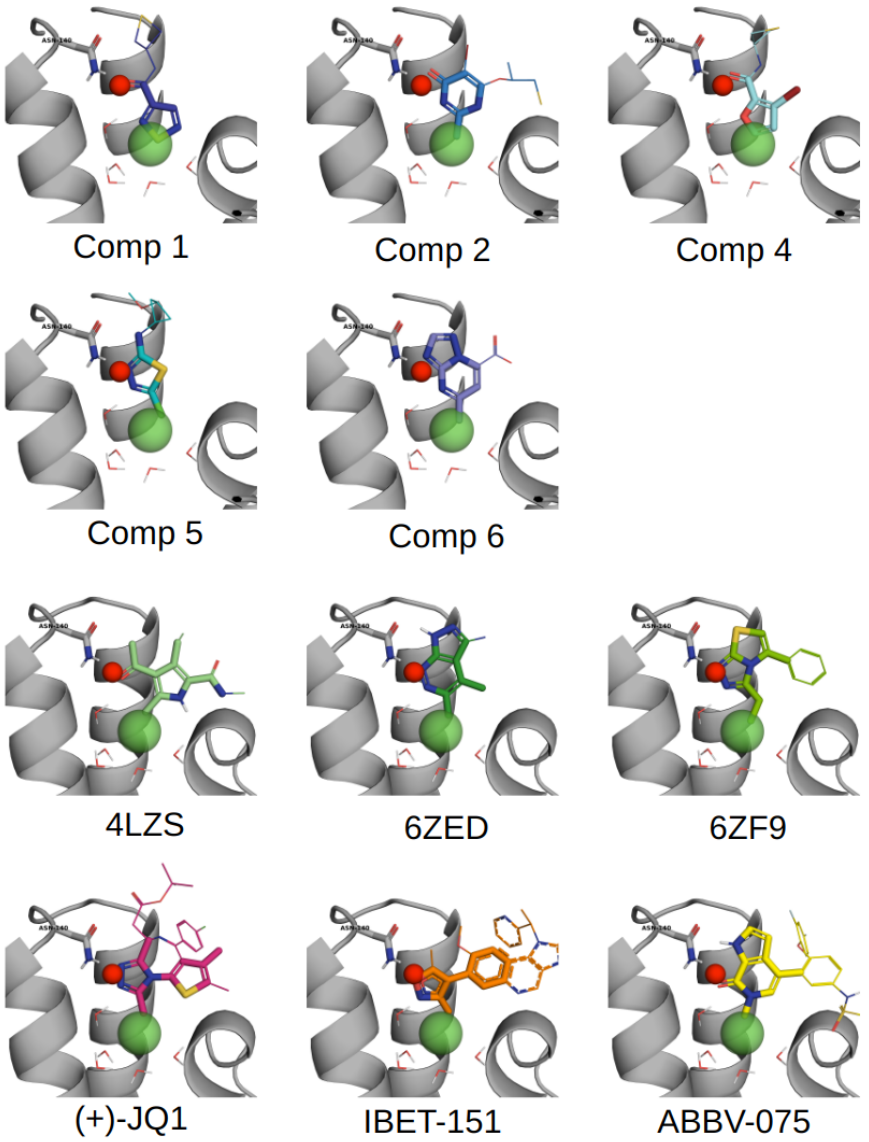


Figure 5.5. Representation of the reference scaffolds' binding mode. The structure of the reference virtual hits, experimental fragments and drug candidates is represented in lines. The scaffolds used as query and tethered on their analogs are represented in sticks. The H-bond acceptor and hydrophobic pharmacophoric restraints are represented as red and orange spheres respectively.

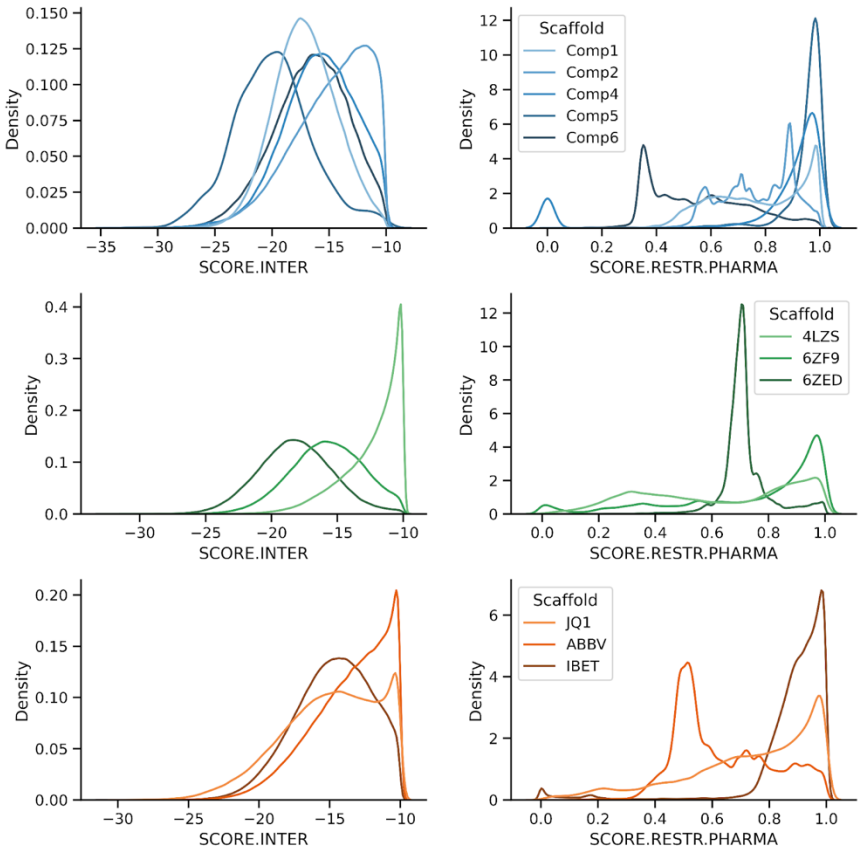


Figure 5.6. Distribution of rDock scores for the best pose and conformation of each compound separated by scaffold. The HTVS filter limited the output binding poses with SCORE.INTER and SCORE.RESTR.PHARMA lower than -10 and 1 respectively.

rDock was able to find binding poses fulfilling the threshold conditions for a total of approximately 13 million compounds (i.e. discarding approximately 95% of the input compounds) (Table 5.2). However, for the ligands with the desired scores, we obtained an average of 18 conformations out of the 25 maximum possible, as specified in the HTVS. This shows that most input compounds were discarded in the first iterations (because of clashing conformations due to the tethering) and the rest were thoroughly explored. In addition, I removed the binding poses that allocated a polar group

close to the hydrophobic pharmacophoric restraint, as rDock is unable to properly describe the nature of the water network (i.e. which does not allow for new polar interactions).

5.4.2.2 Post docking filters

Despite the high removal of compounds by the tethered docking, we are still unable to evaluate 13 million compounds using DUck or MM/GBSA. Thus we needed to rely on a clustering step to select a diverse subset of compounds to continue the pipeline. I performed an initial hierarchical clustering using a Tanimoto similarity threshold of 0.95 and reduced each scaffold focused library to approximately 2.10^5 compounds. Due to the high similarity between the compounds, MACCSkeys fingerprints showed their limitations differentiating them. Thus, we collaborated with P. Aloy's group, to implement a second clustering using more complex fingerprints (i.e. the A1 to A5 Chemical Checker *signaturizers*) to reduce the library size to 1000 compounds (Figure 5.7). This work was done by Arnau Comajuncosa. The resulting clusters were composed of 1 to 600 molecules, with maximal probability around 250 molecules. The distributions shown in Figure 5.7 correspond to the IBET-151 scaffold focused library, but the equivalent on the other clusters were very similar. The homogeneity of the clusters was confirmed by the small range of similarity standard deviations (Figure 5.7.B). The library coverage of the cluster centers can be appreciated using a dimensionality reduction such as the tSNE. tSNE shows better performance when operating on highly similar elements than other decompositions. Finally, as a sanitization, we corroborated that the physicochemical properties of the cluster representatives properly represents the library diversity (Figure 5.7.D). This is true for all but

for the number of aromatic rings, as the cluster centers are usually covered by simpler molecules.

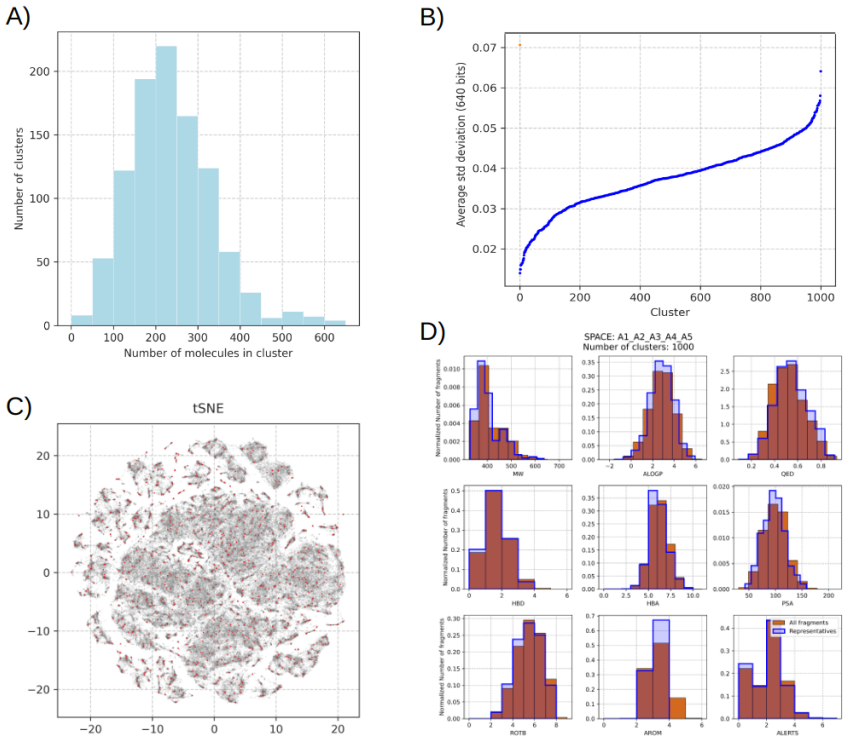


Figure 5.7. Example of assessment of the clustering quality using IBET scaffold-focused library. A) Shows the distribution of compounds per cluster. B) corresponds to the standard deviation of similarities within each clustered. C) Is a tSNE dimensionality reduction with the red dots representing the cluster representatives. D) Shows the overlap in physicochemical properties of the representatives and the rest of the library.

The cluster representatives were assessed in parallel using MM/GBSA and DUck. Both approaches are quick and yield quite orthogonal results compared to docking scores (Figure 5.8). The $\Delta G_{Solvation}$ calculated from MM/GBSA was used as the first filter to eliminate compounds with high solvation penalties. This calculation aims to compensate for the tendency of the rDock scoring function

to underestimate solvation penalties in favor of poses with a higher amount of H-bonds.

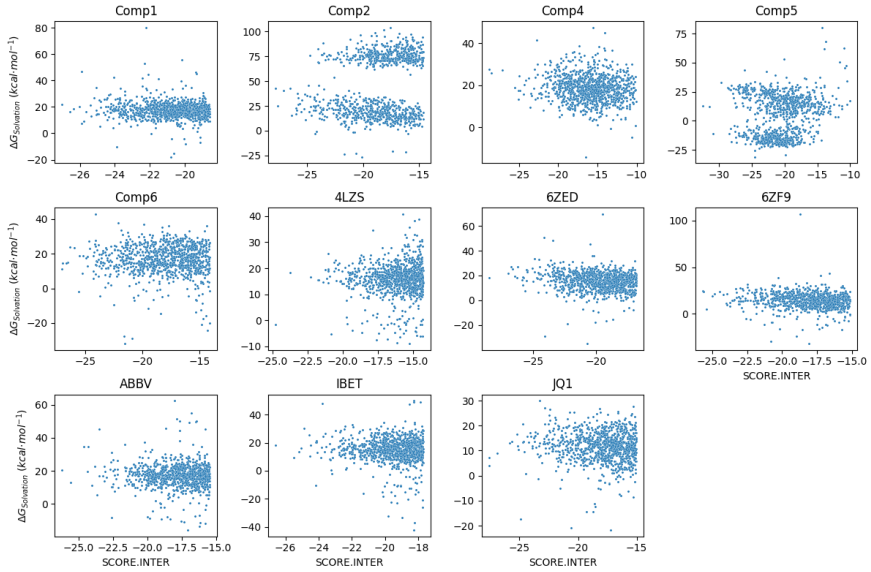


Figure 5.8. Comparison between $\Delta G_{Solvation}$ calculated with MM/GBSA and the SCORE.INTER from rDock for each scaffold-focused library.

Then the ΔG_{bind} with the ligands' strain taken into account was used to rank the compounds in consensus with DUck's W_{QB}. The distributions of the calculated ΔG_{bind} were very similar between the different scaffolds (Figure 5.9) ranging between -60 and 0 kcal/mol for all but the Comp2 library. The bimodal distribution of ΔG_{bind} matched the one observed in Figure 5.8, and it is related to the presence of positive charges in the ligands. Intriguingly, the binding modes of the two distributions also occupy different regions of the BRD4 cavity (Figure 5.10). The green binding poses in Figure 5.10 correspond to the ones with good predicted ΔG_{bind}, and mostly match the area occupied by drugs like (+)-JQ1, IBET-151 and ABBV-075. This reinforces the idea that we might have a better enrichment of

the right growing vectors when using more sophisticated methods than docking.

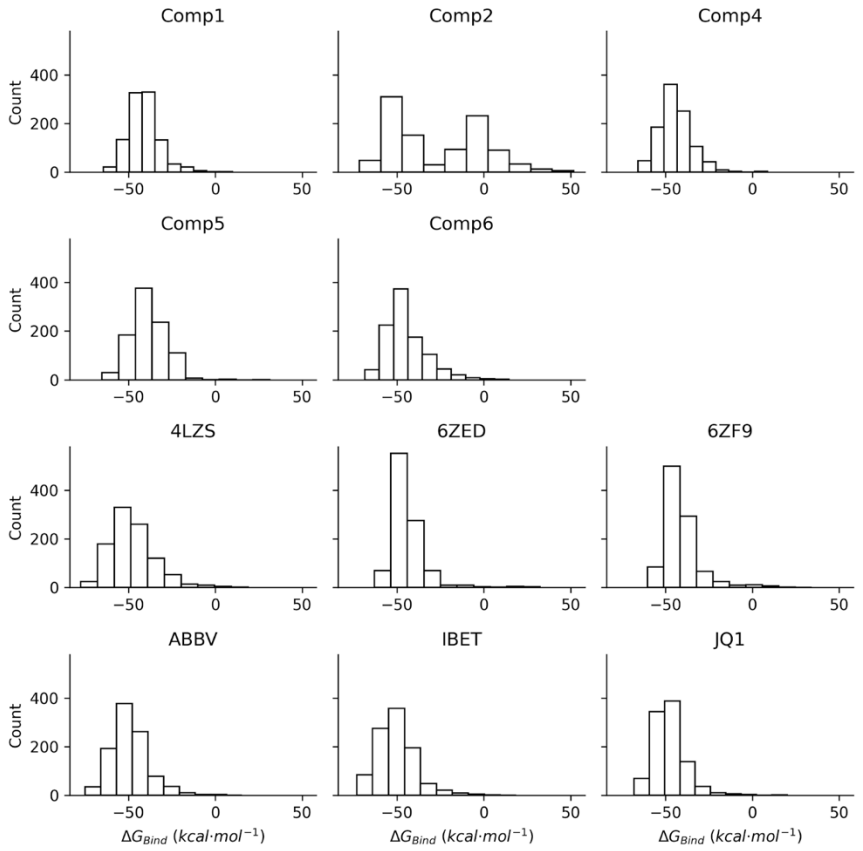


Figure 5.9. Distribution of ΔG_{bind} calculated with MM/GBSA. The ΔG_{bind} correspond to single-point calculations taking into account the ligands' strain.

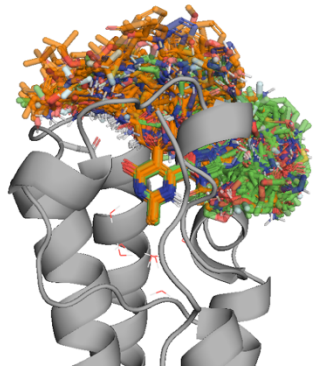


Figure 5.10. Binding mode of Comp2 cluster representatives separated by their calculated ΔG_{bind}. The compounds with a ΔG_{bind} higher than -25 kcal/mol are illustrated in orange and the compounds with a ΔG_{bind} lower than -25 kcal/mol are represented in green.

The DUck simulations were performed using the W_{QB} thresholds set by the reference compounds of each library and with five iterations (Table 5.2). Between 85% and 97% of the compounds did not reach the minimum W_{QB} and were filtered-out (Figure 5.11). Despite how similar the ΔG_{bind} distributions were in MM/GBSA, DUck's W_{QB} distributions are very different between scaffold-focused libraries. As an example, the compounds derived from 4LZS yielded overall higher W_{QB}, being higher than 10 kcal/mol for 72 compounds. On the other hand, scaffolds like 6ZF9 yielded very few compounds with W_{QB} higher than 6 kcal/mol.

Finally, I ranked the compounds that surpassed the W_{QB} threshold set for their series using the consensus ranking from the MM/GBSA ΔG_{bind} and W_{QB} values. We then selected the top 10 compounds to be synthesized by Enamine. They were able to synthesize 85 of the requested compounds (Table 5.2), which corresponds to approximately 80% of the compounds, with a 90% purity and within 4 weeks.

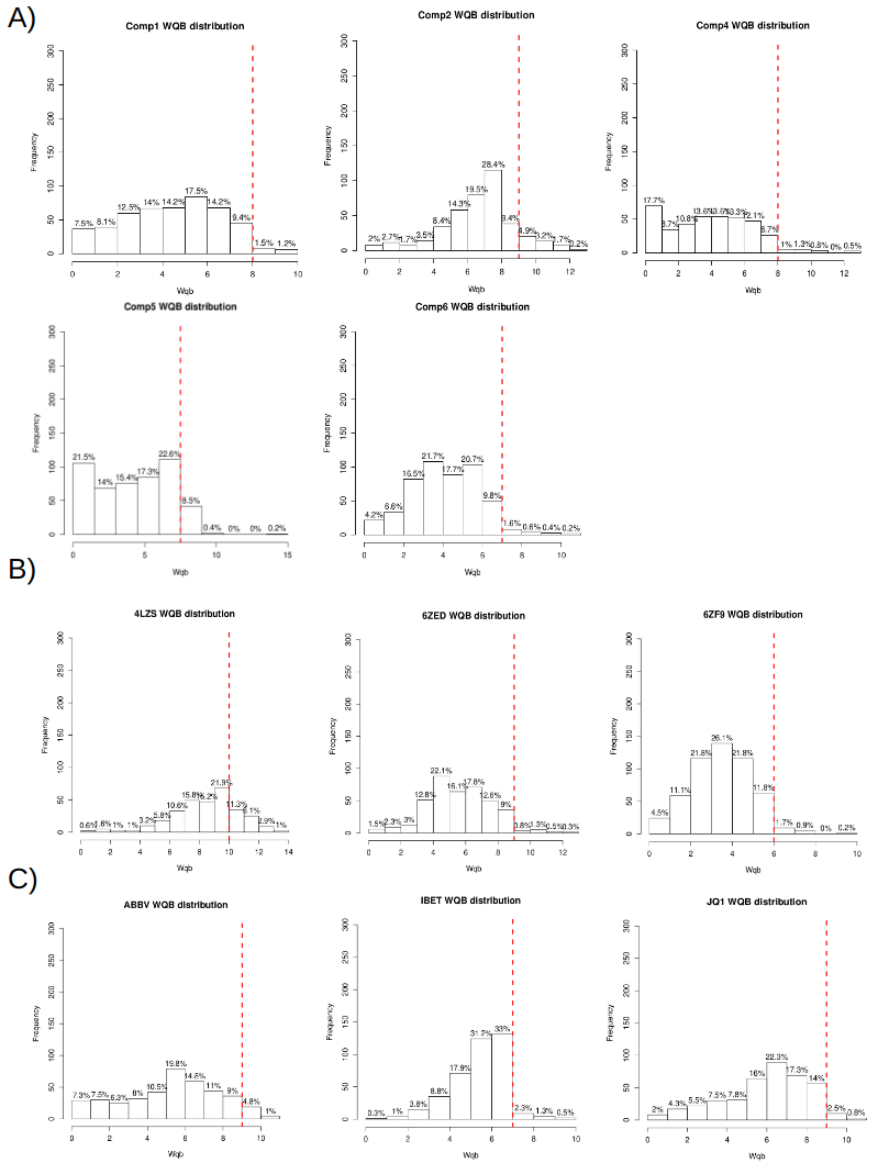


Figure 5.11. Distribution of W_{QB} values for the cluster representatives. A, B and C correspond to hits derived from the virtual fragments, experimental fragments and drug scaffolds' respectively. The red dotted lines represent the selected W_{QB} threshold values.

5.4.3 Future scalability of the pipeline

Excluding fragment and drug scaffolds employed for validation, the complete virtual screening process entailed preparation and docking of 4 million fragments and 10 million drug-like compounds per follow-up scaffold. This represents a mere 0.03% of the entire enamine REAL collection. Importantly, part of the computational time saved compared to a brute force approach, can be used to apply higher levels of theory, thus ensuring that selected compounds have better chances of being active.

As the on-demand collections are set to rapidly increase their size, it is interesting to speculate about the sustainability of this approach. We consider the exhaustive search of the fragment space a fixed cost, as it is not expected to increase beyond 1 billion.[269] The number of follow-up compounds may increase with larger collections, but only if they increase in depth rather than breadth. Comparisons of on-demand collections from different vendors indicate that there is very little overlap between them, which suggests that future growth will mostly stem from an expansion of the covered chemical space rather than increasing the density of currently explored regions. Therefore, a moderate increase in chemotype depth is expected (e.g. 2-fold for every 10-fold increase in overall collection size). This ensures that exploring exhaustively the areas around certain scaffolds remains achievable within the expected growth of computational resources.

5.5 Experimental validation: BRD4 binders

From the 85 synthesized compounds,t 32 (38%) correspond to the virtual fragments scenario, 24 (28%) to the experimental fragments

and 29 (34%) to the libraries from drug scaffolds. Andrea Bertran-Mostazo, a fellow PhD candidate in the group, evaluated the compounds through biophysical assays. Firstly, she performed a single-dose differential scaffold fluorimetry (DSF) experiment at three ligand concentrations (50μM, 10μM and 1μM). The shifts in melting temperature observed were comparable to our positive control (+)-JQ1 at the lowest concentrations (i.e. around 2 ºC), however, the thermal shifts did not increase as much as the control when increasing the concentrations (i.e at 10 and 50 μM). The most likely reason being the low solubility of our compounds compared to the control. She then performed a single dose screening by SPR, in order to obtain an orthogonal validation of the hits. To allow direct comparison of the results obtained by DSF, she performed the SPR screening at 10μM.

Both techniques are orthogonal in the manner of directly detecting binding, thus we used the results from both to classify the compounds as hits. We determined as a hit any molecule that yielded a result significantly differentiable from the negative control (i.e. having the 95% confidence interval of the compounds and the negative control not overlapping). We obtained an overall 29.4% hit ratio (Figure 5.12). Out of the three validation scenarios, the experimental fragments yielded the best hit ratio (39%) followed by the virtual fragment hits (31%) and finally the drug scaffolds (20%). Out of the two experiments, DSF yielded a higher amount of hits for the experimental and drug scaffolds, whereas the computational scaffolds showed higher responses in SPR.

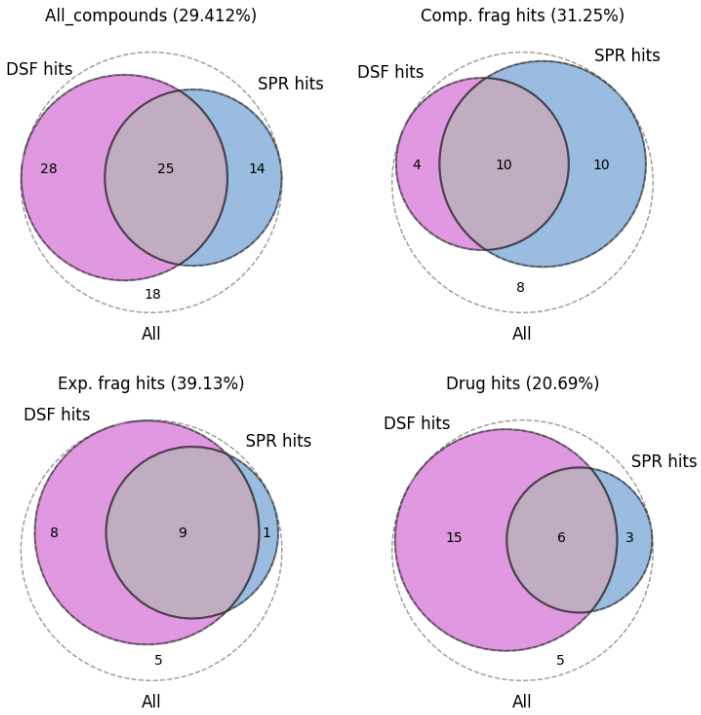


Figure 5.12. Venn diagrams of the single dose screening with DSF and SPR hit ratios. Both experiments were performed at 10 μM ligand concentration. The hits were defined by the non-overlapping between the blank measurement and each compound's measurement taking into account a margin of two standard deviations.

Andrea then performed competitive TR-FRET experiments, titrating each compound from 0.1 nM to 10 μM. In these experiments, the compounds compete with a labeled reference ligand (peptide with acetylated lysines), disrupting the FRET between the dyes. Thus, indirectly reporting the binding. In TR-FRET most of the hits yielded IC_{50} values between 10 μM and 1 nM (Figure 5.13). Furthermore, 16 of those compounds yielded comparable IC50s than (+)-JQ1, a drug candidate in phase III clinical trials. As seen in Figure 5.13, the

compounds from the computationally elucidated scaffolds showed overall better IC50s than the other scaffolds. These two sets of experiments match the trend seen in other virtual screening campaigns using ultra-large libraries of finding more and highly potent hits.

In parallel to the biophysical experiments, Dr. David Ruiz performed soaking crystallization experiments with the compounds, in the context of a collaboration with Dr. Maria Garcia-Alay at the EMBL-Hamburg. He was able to obtain crystals for 3 compounds bound to BRD4-BD1 and resolve their structure (Figure 5.14). From the three compounds crystallized, two come from the JQ1 derived scaffold (BC-07C and BC-17C) and one from the Comp5 scaffold (BC-17A). One of the main working hypotheses of the bottom-up pipeline we have proposed is that the scaffold's binding pose is maintained through the family. Thus, these results should validate the applicability of the assumption to the rest of the compounds.

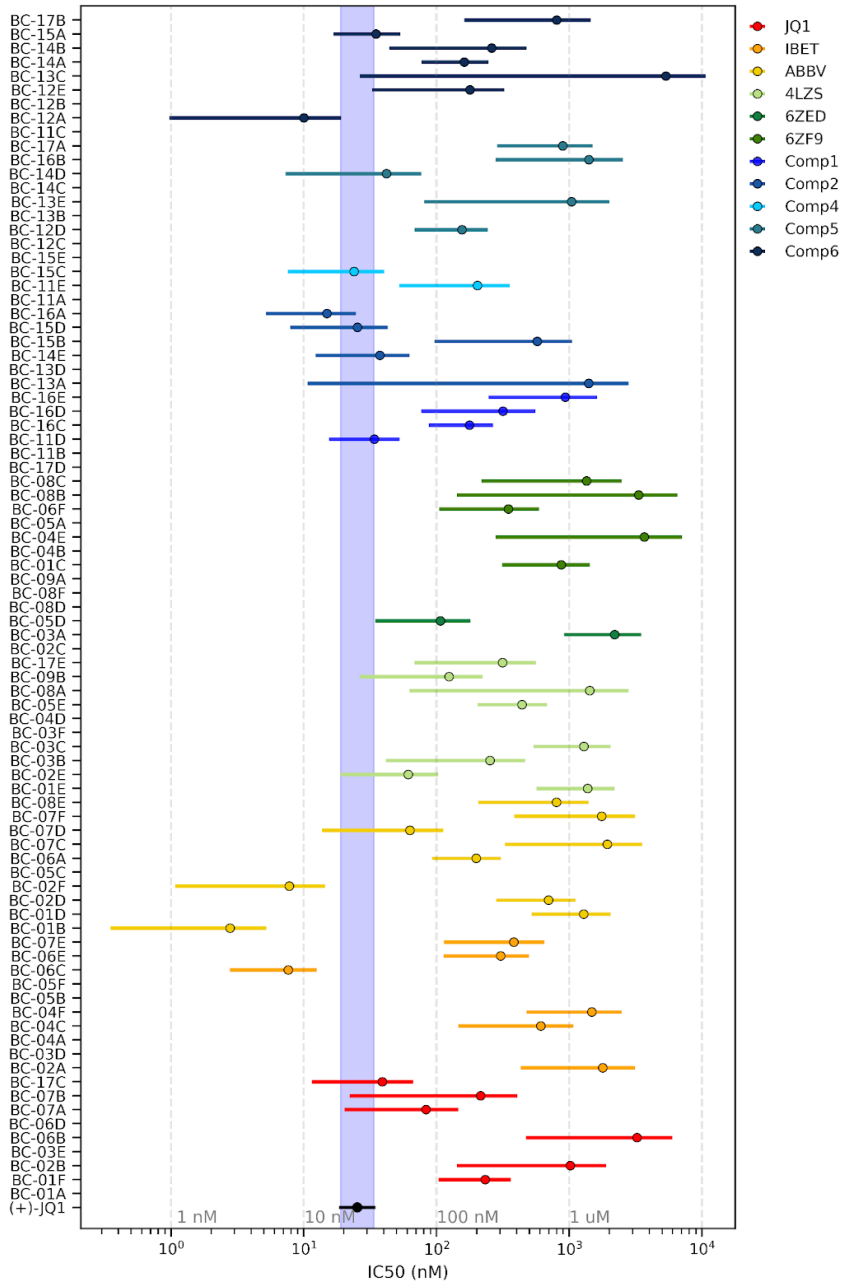


Figure 5.13. IC50 values measured from the TR-FRET experiment at 1.5h. Compounds were titrated (from 0.1 nM to 10 µM) by duplicate. Each dot corresponds to the average IC50 measured and the error bars to the 95% confidence interval. The blue band corresponds to the positive control (+)-JQ1 with its 95% confidence interval.

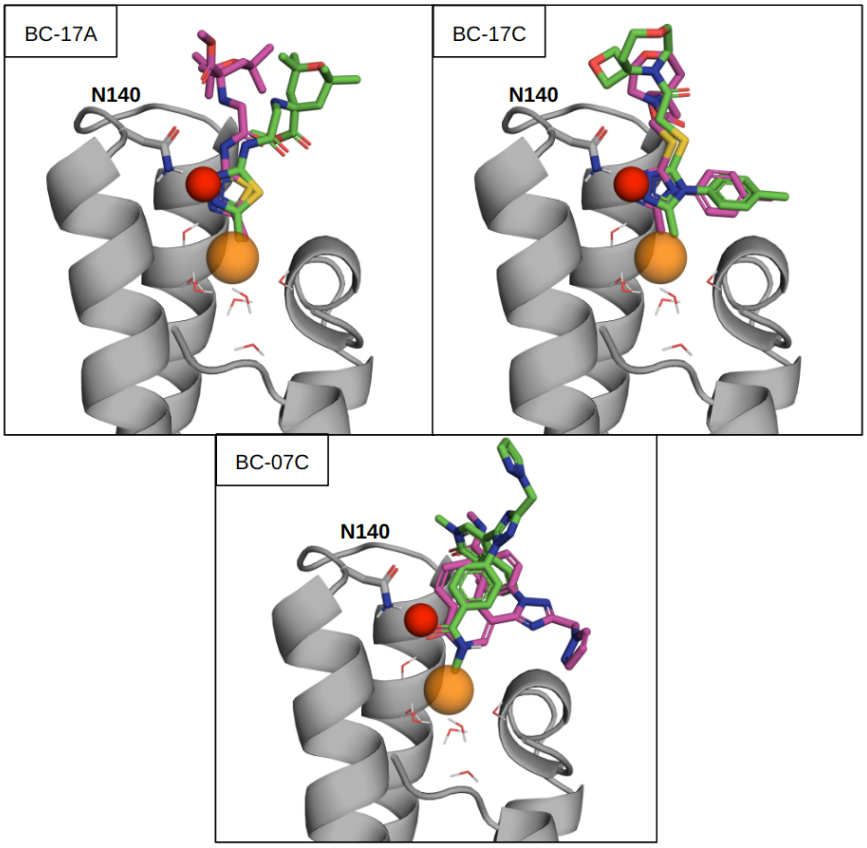


Figure 5.14. Resolved X-ray crystallographic BRD4-ligand structures. The structures of BRD4-BD1 (grey) bound to the compounds BC-07C (JQ1 scaffold), BC-17A (Comp5 scaffold) and BC-17C (JQ1 scaffold) were obtained through soaking crystallization experiments. The diffraction maps were modeled into 3D structures by Dr. Lucas De Felipe. The depicted ligands' binding poses correspond to the docking experiments (pink) and the resolved structure (green). The red and orange spheres correspond to the H-bond acceptor and hydrophobic pharmacophore employed in docking. The residue N140 is highlighted as it participates in the interaction studied in DUck.

All the scoring and post-docking filtering methods rely on the accuracy of the binding poses generated from the tethered docking. Thus, validating that the docking poses are recovered in the crystal structure is highly relevant for the pipeline. The compounds BC-17A and BC-17C fulfill this condition (Figure 5.14), and the tethered

scaffold matches the one reported by crystallography. On the other hand, the compound BC-07C showed a completely different binding mode than the one predicted obtained by docking. Yet, the crystallographic binding pose fulfills the two pharmacophoric points from docking and would pass the post docking filters, albeit not with the best scores.

5.6 Chemical diversity analysis

On top of obtaining a set of potent BRD4 binders we wanted to assess the chemical novelty and diversity of these compounds compared to known BRD4 binders. Arnau Comajuncosa compared the top 10 compounds (i.e. with the lowest IC50s reported by TR-FRET) against all the BRD4 binders from ChEMBL and a random subset of molecules from the Chemical Checker Universe,[202] as the background (Figure 5.15). Similarly to the clustering shown in section 5.4.2.2, he employed the Chemical Checker A fingerprints. As seen in Figure 5.15.A, the hits from the bottom-up pipeline do not occupy the space of known BRD4 binders.

Moreover, the selected hits are more diverse between themselves than the rest of the BRD4 binders (Figure 5.15.B). By providing hits from different areas of the chemical space, we increase the probability of finding a scaffold with desirable properties that we do not optimize for, such as permeability, metabolism, etcetera.

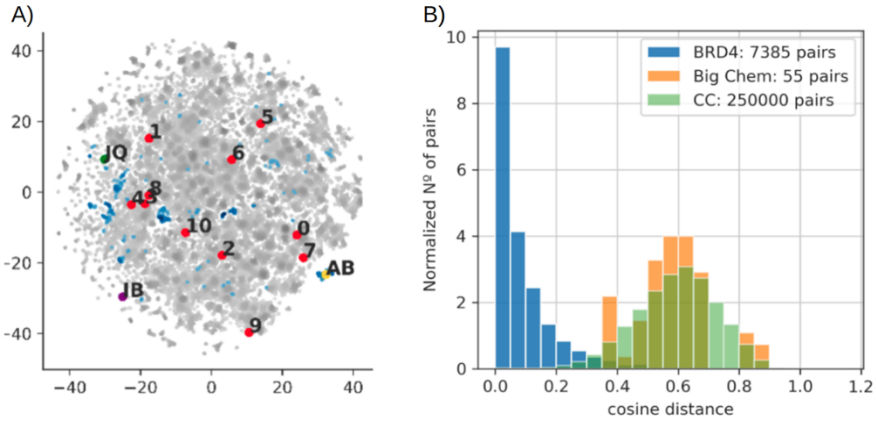


Figure 5.15. Chemical diversity of the most potent hits compared to the known BRD4 binders. A) A tSNE decomposition of the Chemical Checker A fingerprints from a background of randomly selected compounds from the Chemical Checker Chemical space (grey), BRD4 binders from ChEMBL and the top 10 compounds from the pipeline (red). B) Histogram of the cosine distances between the compounds in each group.

CHAPTER 6: RESULTS

STEERED MOLECULAR DYNAMICS FOR DRUG DISCOVERY

6. STEERED MOLECULAR DYNAMICS FOR DRUG DISCOVERY

6.1 Background

Predicting protein-ligand binding affinities is of high interest for drug discovery, but still remains one of the major challenges of computational chemistry. The binding free energy (ΔG_{bind}) of non-covalent complexes relates directly with the thermodynamic equilibrium constant (K_d), and indirectly with the most common functional observables (e.g. K_i, IC_{50}, EC_{50}, efficacy) that are often taken as a proxy of binding affinity. The importance of ΔG_{bind} has motivated the development of a wide spectrum of computational approaches to predict this property. Empirical (or trained) methods, also known as quantitative structure-activity relationship (QSAR) methods, rely on pre-existing data and have seen an important resurgence with the emergence of machine learning (ML) algorithms.[271] These methods often assume that, provided the core scaffold is unaltered, chemical modifications have an additive effect on the ΔG_{bind}. In this context, structural changes that create unexpected differences in the ΔG_{bind}, referred to as activity cliffs (AC) in the structure-property landscape, are often reported as outliers that are particularly challenging to predict.[272,273]

As mentioned, AC are discontinuities in the structure-activity landscape which have historically been problematic for QSAR models. They highlight crucial structural motifs that heavily influence the activity of certain molecules to a target, generally related to non-additive effects on the protein-ligand interactions. AC are defined by

a high disparity in potency and a high structural similarity. While most authors agree on 100-fold as the potency difference threshold, structural similarity is defined in a variety of manners in the literature (e.g. matched molecular pairs, fingerprint-based similarity).[274]

There are several structure-based methods to calculate ΔG_{bind} that, in principle, should be able to correctly predict these outliers because they do not rely on pre-existing experimental data. These predictive methods can be broadly classified as end-point (e.g. MM/GBSA, MM/PBSA),[275] path-sampling (e.g. metadynamics, ABF)[276] and alchemical methods (e.g. TI, FEP).[277] However, their predictive power is oftentimes hindered by the need to balance accuracy with computational throughput. The calculation of relative binding free energies ($\Delta\Delta G_{bind}$) offers a useful shortcut when investigating congeneric series of compounds that explore a similar configurational space.[278] In this context, alchemical free energy calculation methods (AFE) have become the golden standard for computer-aided drug design.[279] Under optimal conditions AFE are reliably able to predict ΔG_{bind} values within 1 kcal/mol of the experimental value, although their accuracy decreases for challenging transformations (many atoms, topological changes or net charge).[280] Transformation uncertainties can be assessed[89] and reduced using multiple pairwise transformations per ligand but this increases the overall cost, which can become prohibitive for large series of compounds. A rapid alternative to this method would be highly desirable, particularly in the hit-to-lead and early optimization stages of drug discovery, where thousands of analogues may need to be considered. Even if not as accurate, it would be a useful addition to the drug design toolbox as long as it can predict activity

cliffs (i.e. pairs of highly similar compounds with large discrepancies in potency), thus helping identify non-obvious leaps in potency.

6.2 The quasi-bound state as a quick predictor of ΔΔG in activity cliffs

In the first section of this chapter I investigate the potential of Dynamic Undocking (DUck) as a method to calculate $\Delta\Delta G_{bind}$. DUck is a particular implementation of Steered Molecular Dynamics (SMD) which evaluates the mechanical stability of protein-ligand hydrogen bonds (Figure 6.1). The computed magnitude is the free energy required to break the hydrogen bond (from equilibrium to 5 Å distance), reaching a so-called quasi-bound state (ΔG_{QB}). The performance of the method was unexpected because it is a local property with no correspondence to any macroscopic observable (Figure 6.1). Nonetheless, later works showed that: (i) most protein-ligand complexes form mechanically robust interactions,[181] (ii) Correct binding modes are generally also the most robust ones,[177] and most surprisingly, (iii) ΔG_{QB} quantitatively predicted the relative binding affinity of the ternary complex between Cereblon, Lenalidomide and various mutant forms of Caseine Kinase 1α.[217] But the question remains if ΔG_{QB} can rank-order congeneric series of protein-ligand complexes and, more importantly, why should this out-of-equilibrium property inform at all about thermodynamic properties. In particular, I investigated the presence of AC in congeneric series of HSP90, CDK2 and BACE1 binders.

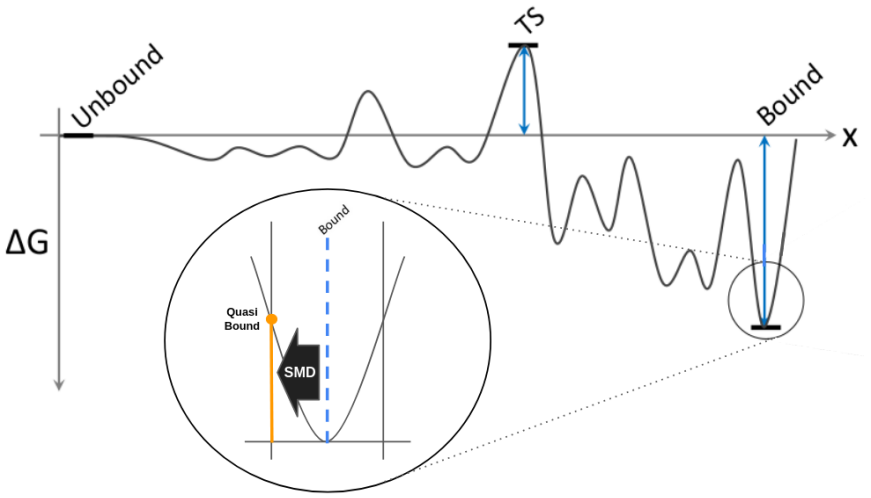


Figure 6.1. Schematic representation of the free energy landscape of a protein-ligand dissociation. Typically this landscape has one or more barriers that create transition states until the ligand becomes unbound. The scope of Dynamic Undocking is constrained to measuring the work needed to perturbate the system to the quasi bound state across the first of such barriers.

6.2.1 Affinity and kinetics predictions in HSP90

I started by analyzing a HSP90 dataset[97,204] that includes both thermodynamic and kinetic experimental data. HSP90 is a molecular chaperone that assists the folding of proteins, particularly the ones involved in cell growth. It has been historically targeted to its ATP binding site, as inhibiting it can lead to a diminished tumor growth. It has been repeatedly demonstrated that the robustness of the essential interaction with Asp93 is an excellent predictor of binding affinity for HSP90 inhibitors.[176] This makes HSP90 an excellent test case for our approach. Furthermore, there have been other attempts reported in the literature attempting to rationalize the kinetic behavior of HSP90 ligands.[175]

In this dataset, there are 93 inhibitors that were paired by similarity and grouped by the receptor conformation they bind to, (13 in the loop conformation and 80 in the helical conformation, which introduces opening of a cryptic site[281]). The resulting 207 pairs showed a proportion 1:10 of AC, with only 18 pairs with affinity changes over the 100-fold threshold and thus considered AC. Using DUck, I evaluated the robustness of each HSP90-inhibitor complex. The ΔG_{QB} values range from 4 to 29 kcal/mol, but are clearly skewed towards high values (Figure 6.2.A), and match the top end of the distribution observed previously for diverse sets of protein-ligand complexes.[181] Plotting the difference of ΔG_{QB} for each pair of compounds ($\Delta\Delta G_{QB}$) vs. the experimental relative binding free energies ($\Delta\Delta G_{bind}$) a trend was clear, where the largest changes in binding affinity also involve a large $\Delta\Delta G_{QB}$ value (Figure 6.2.B). While there is not a straight correlation, the method is surprisingly good at detecting AC: out of the 207 pairs, setting a cut-off at $\Delta\Delta G_{QB}$ < -5 kcal/mol, the method identifies 29 possible AC, 13 (45%) of which turned out to be correct. Conveniently, the number of false negatives is very low (5, 2%), resulting in a Matthew correlation coefficient (MCC) of 0.52. For comparison, the well-known MM/GBSA scoring scheme has no predictive power (MCC=-0.13).

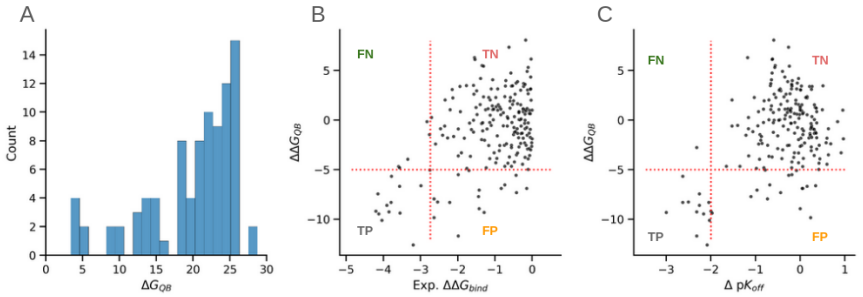


Figure 6.2. Prediction of activity cliffs from the HSP90 RAMD dataset. A) Distribution of ΔG_{QB} values obtained from the Dynamic Undocking calculations in kcal/mol. Comparison between $\Delta\Delta G_{QB}$ vs $\Delta\Delta G_{bind}$ (B) and $\Delta\Delta G_{QB}$ vs ΔpK_{off} (C) for the 207 compound pairs. A $\Delta\Delta G_{QB}$ threshold of 5 kcal/mol is set for predicting the activity cliffs, which are defined by a $\Delta\Delta G_{bind}$ of 2.73 kcal/mol. Both thresholds are represented by dotted lines. Prediction outcomes False Negative, True Negative, False Positive and True Positive are defined in the sections divided by those thresholds.

To better understand the predictive power of the method, I then investigated its performance in predicting changes in kinetic off rates, which should be more directly related to the calculated magnitude. As shown in Figure 6.2.C predictions of cliffs based on koff show a better correspondence (MMC of 0.64) than based on K_d.

6.2.1.1 One step dissociation links a local property to a global process

Many false positive predictions in the HSP90 dataset involved specific ligands that are potent binders but present relatively low ΔG_{QB}. I analyzed their dissociation pathways, comparing them with the true positive (high potency, high ΔG_{QB}). Additionally, I reproduced the simulations from Koh et al. [97], performing τRAMD simulations on the HSP90 dataset. On top of measuring the residence times, I analyzed the dissociation pathways, following main interaction studied during DUck and the movement of the ligands' center of mass. While most (88, 95%) ligands follow a one-

step dissociation (bound-unbound) with short, undetectable transition states (Figure 6.3.A), some (6, 5%) showed a delay between the break of the main H-bond and the full dissociation (Figure 6.3.B). The ligands following the delayed dissociation were involved in all the false AC predictions.

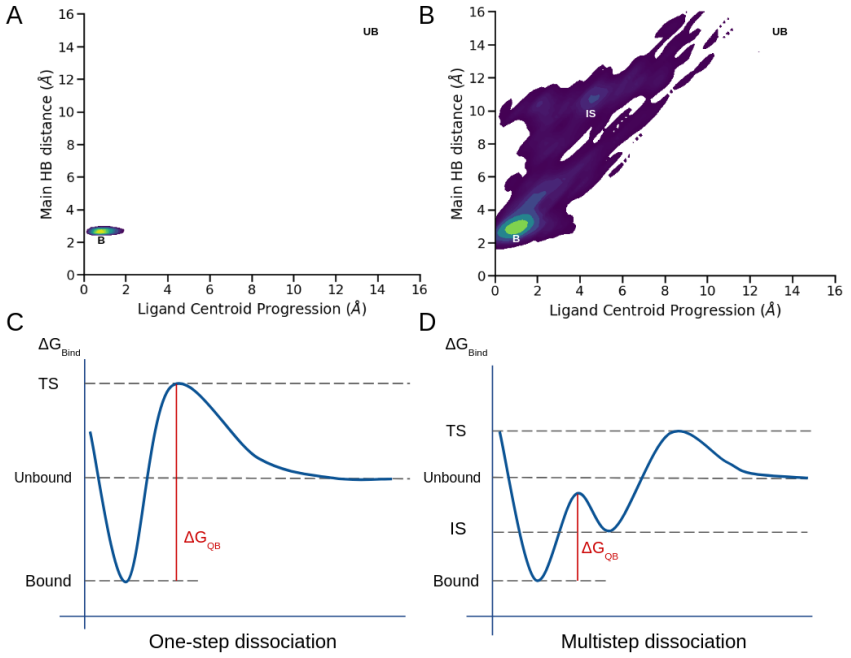


Figure 6.3. Schema of the different dissociation mechanisms seen in τRAMD simulations of HSP90. A&B) Density representation of the ligand centroid movement (x axis) and the distance between the main interaction atoms (y axis) for 1g (A) and for 2226756 (B). C&D) Schematic representation of a one-step dissociation (C) and a multistep dissociation (D) free energy profiles corresponding to what is seen in A and B respectively.

Based on these results, we hypothesize that the underlying mechanism has a dramatic effect on the performance of DUck: if the unbinding mechanism does not involve metastable intermediate states, the quasi-bound state is a good approximation of the unbinding transition state, hence the ΔG_{QB} is a good predictor of the

$k_{off.}$ Conversely, according to the rate determining step approximation, the existence of a metastable transition state that exchanges with the bound state faster than with the unbound state, would make the first barrier in the dissociation process (i.e. the one measured by DUck) irrelevant from a thermodynamic perspective (Figure 6.3.D).

6.2.2 Trends observed in HSP90 are also present in CDK2.

To further explore the correspondence between DUck's predictability and one-step dissociation, I repeated the analysis on a Cyclin Dependent Kinase 2 (CDK2) inhibitor dataset. CDK2 is a well studied kinase involved in the regulation of the cell cycle with more than 8000 activity entries in ChEMBL, of which 35% are IC50. Moreover the interaction pattern of the CDK2-inhibitors complex is well studied, being the three H-bonds with backbone atoms from the hinge region reported as the most relevant for binding (Figure 6.4.A). I evaluated the work needed to break the three H-bonds independently to assess cooperative effects and identify whether one of them could act as a predictor of the strength of the remaining H-bonds. In general, the H-bonds made with the backbone nitrogen of L83 (L83-N) yielded the highest W_{QB} values (14.9 ± 4.1 kcal/mol, Figure 6.4.B). Additionally, L83-N showed the highest correlation to the other two, resulting in the breakage of the neighboring H-bonds when the interaction with L83-N was steered. This effect is reflected in the correlation of the W_{QB} obtained when steering the H-bonds independently (r^2 of 0.6 and 0.4 between the L83-N interaction and E81-O and L83-O respectively). In contrast, steering from E81-O and L83-O resulted

in pivoting of the ligands, yielding significantly lower correlation between the W_{QB} (r^2 of 0.1).

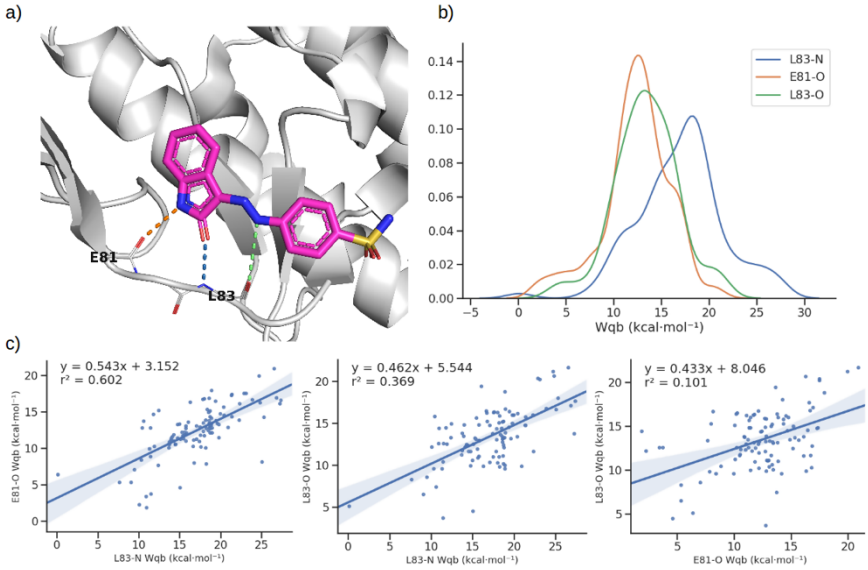


Figure 6.4. Analysis of the three hinge H-bonds of the CDK2-inhibitors complexes. (a) Representation of hydrogen bonds from the CDK2 hinge region, with the ligand from the pdb 1FVT, used as reference for the Serie B. (b) W_{QB} distribution of the hydrogen bond associated with the leucine's 83 nitrogen (blue), oxygen (green) and the glutamic acid's 81 backbone oxygen (orange). (c) Correlation between the W_{QB} of the different hinge hydrogen bonds. From left to right they represent the correlation of L83-N against E81-O, L83-N against L83-O and E81-O with L83-N.

I subdivided the CDK2 AC dataset into four distinct congeneric series (CDK2-A, CDK2-B, CDK2-C & CDK2-D), based on the similarity of their core to the scaffold of the inhibitors in co-crystallized structures available in the PDB with PDB ids 2R3K, 1FVT, 2VTO and 1JSV respectively. This minimized the structural differences between pairs inside each set and isolated the main location of the chemical changes involved in each cliff. However, due to the fingerprint-similarity used, changes on multiple regions of the ligands could still occur (Figure 6.5). On the one hand, in series A all ligands

established only the H-bonds with L83-O and L83-N. On the other hand, H-bond with E81-O was only present in series B and some ligands from series C and D (Figure 6.5). I obtained a total of 1164 pairs from combinating the 497 CDK2 active ligands. 121 of those pairs being AC. For each series, a representative pair of compounds was selected to run τRAMD, in order to explore *a-priori* the dissociation model.

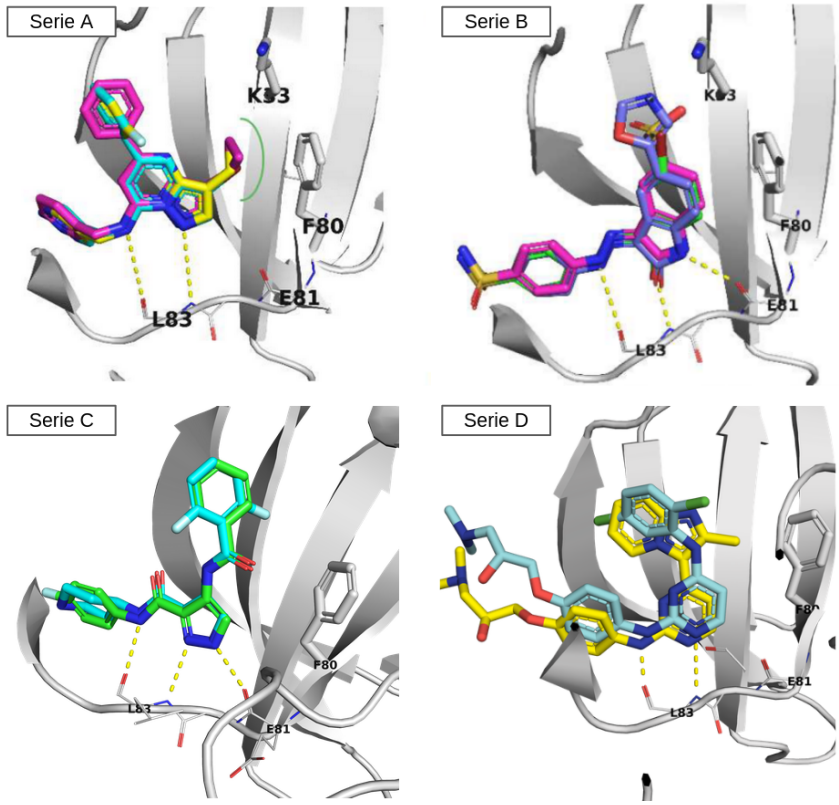


Figure 6.5. Structural representatives of the CDK2 activity cliff chemical series. On the serie A, the ligands from 2R3K, 2R3M and 2R3N PDB entries are shown as yellow, magenta and cyan respectively. For serie B, the AC made from CHEMBL269827 (purple) and ChEMBL412091 (magenta) ligands is shown. Ligands from 2VTO (green) and 2VTQ (cyan) PDB entries show the serie C scaffold. For CDK2 serie D, 1H01 (light blue) and 1OIR (yellow) are represented.

Inhibitors from series A and C showed metastable intermediate states, where the ligand no longer formed the H-bond with the hinge region, but preserved the rest of interactions (i.e. similar to figure 6.3.B). By contrast, the representatives from series B and D presented a one-step dissociation where, owing to the formation of a third H-bond with the hinge region or a more rigid chemical structure, rupture of the H-bonds resulted in immediate expulsion of the ligands from the the binding site. Following the hypothesis generated in the HSP90 case, we predicted that DUck should be more predictive for compounds in series B and D than for compounds in series A and C.

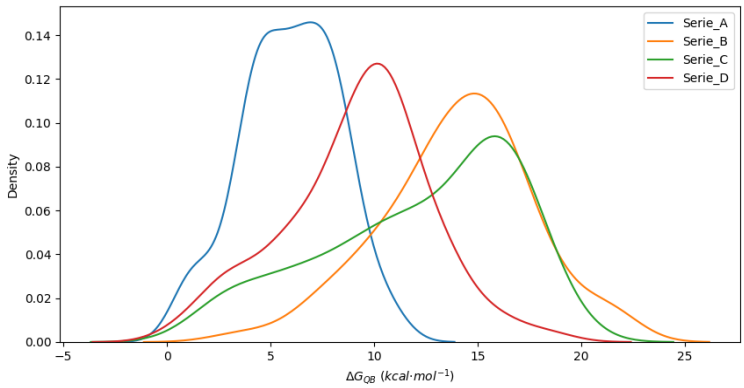


Figure 6.6. Distribution of ΔG_{QB} values for the four CDK2 congeneric series.

The complete set of ligands was then explored with DUck. Despite the high potency of the ligands in the dataset, the CDK2 series ΔG_{QB} distributions show lower values than with HSP90 (Figure 6.6). Compounds from the serie A have predominantly labile interaction (5.9 ± 2.3 kcal/mol), followed by serie D (9.2 ± 3.5 kcal/mol). Compounds in series B and C present robust interactions similar to HSP90 (14.0 ± 3.6 kcal/mol and 12.2 ± 4.6 kcal/mol respectively). A scan of $\Delta\Delta G_{QB}$ thresholds was performed through a receiver

operator curve. Finally, I settled for a relative threshold where the $\Delta\Delta G_{QB}$ needs to be higher than half the lowest ΔG_{QB} in the pair. Based on this threshold, the AC predictions of the serie B surpassed the HSP90 results, with a MMC of 0.6 and recovering 78% of the activity cliffs. The MMC for compounds belonging to series D was 0.281, while the MMC for compounds in series A and C was close to 0 (i.e. 0.075 and 0.065, respectively) The bad results from serie C confirms that a steep quasi-bond barrier is not a sufficient estimator of DUck's prediction confidence. A one-step dissociation mechanism is also needed. On the opposite spectrum, the lack of metastable states in serie D with a wider barrier can still yield predictions, albeit worse. Moreover, not all compounds in serie D form the third H-bond, associated with the rigidity of the dissociation, which might alter the dissociation mechanism within the series.

I studied the performance of a standard rescoring method in predicting AC to have a baseline comparison for DUck. Specifically, I used the MM/GBSA approach on every ligand-protein complex to calculate the ΔG_{bind} for the ligand. MM/GBSA is a well established method when rescoring compounds that come from virtual screening campaigns,[86,282,283] and therefore constitutes a solid reference to benchmark the performance of DUck. Using the predicted $\Delta\Delta G_{bind}$ within the same A, B ,C and D series, I classified the pairs in activity cliffs and non-cliffs. The MMC for MM/GBSA predictions was smaller than DUck for all cases (i.e. -0.287, 0.163, -0.04 and 0.08 respectively) (Table 6.1). The trend in difficulty by scaffold series is shared between DUck and MM/GBSA. Which could be due to differences in the ligands' flexibility giving more degrees of freedom during the dynamics and the correctness of the initial binding mode. As a baseline null-hypothesis I also used rDock's SCORE.INTER

from the tethered docking and the ligand's molecular weight to predict the formation of activity cliffs. The predictive capabilities of these two last descriptors are reported in Table 6.1.

Table 6.1. Matthew Correlation Coefficients of the activity cliff predictions using different descriptors

Systems	DUck	MM/GBSA	rDock	Molecular Weight
HSP90	0.527	-0.121	-0.133	-0.006
CDK2-Serie A	0.075	-0.287	-0.226	0.090
CDK2-Serie B	0.616	0.163	0.136	0.077
CDK2-Serie C	0.065	-0.04	-0.187	0.0214
CDK2-Serie D	0.281	-0.08	0.092	0.0788
BACE1	0.556	0.25	-0.16	-0.0363

The bad results from serie C confirms that a steep quasi-bond barrier is not a sufficient estimator of DUck's prediction confidence. A one-step dissociation mechanism is also needed. On the opposite spectrum, the lack of metastable states in serie D with a wider barrier can still yield predictions, albeit worse. Moreover, not all compounds in serie D form the third H-bond, associated with the rigidity of the dissociation, which might alter the dissociation mechanism within the series. The trend in difficulty by scaffold series is shared between DUck and MM/GBSA. Which could be due to differences in the ligands' flexibility giving more degrees of freedom during the dynamics and the correctness of the initial binding mode. s smaller than DUck for all cases, being close to 0 (random predictions) in most of them. Moreover, the trend on *'easier'* systems is no longer detected.

6.2.3 Defining *a priori* the confidence in the predictions

Considering the aforedescribed results, we propose the following protocol to evaluate the applicability of structural stability prediction of activity cliff through dynamic undocking: First, it is necessary to establish whether the ligand-protein complex displays at least one robust interaction. Our results indicate that comparing changes in robust interactions are a better predictors of AC (i.e. HSP90, CDK2 seriesB) than comparing labile interactions. Thus, testing a representative compound for the chemical series to ensure a high ΔG_{QB} profile is desirable. Second, a one step dissociation profile is essential for the applicability of DUck to the prediction of ACs. Therefore, a dissociation analysis of the representative compound should be performed to evaluate the unbinding mechanism. We propose τRAMD as it provides an enhanced sampling of the dissociation without defining a specific pathway as a CV, but alternative protocols such as metadynamics could be equally valid. If those two conditions are met, we expect a high confidence in the AC predictions through their differences in structural stability.

6.2.3.2 Testing the *a priori* conditions in BACE1

To confirm this protocol, I performed a prospective validation following the above mentioned steps. Due to recent insights in structural stability that point at crystallized ligands having higher structural stability,[177] I tested activity cliff prediction excluding ligands without a crystal structure on a new system. I tested it on BACE1, a pharmacologically relevant target with a less defined key interaction than kinases. From the set of 50 ligands from the PDBind, forming 82 pairs, the proportion of cliffs to ligands was lower than in CDK2 due to the high scaffold diversity within the set. The location, number

and size of the changes also has a strong impact in that regard (i.e. having overall smaller similarities) Despite the structural diversity, the main interaction groups were maintained across all ligands; an amide nitrogen with the backbone oxygen of G291 (Figure 6.7.A) Thus, we hypothesized that the dissociation mechanism is shared across the series, and the conclusions extracted with the reference compound will be applicable to the whole BACE1 dataset.

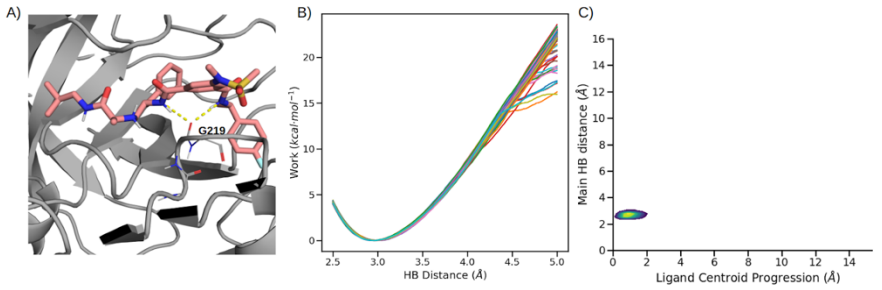


Figure 6.7. Representation of the *a priori* conditions fulfilled by the ligand of 2QZL. A) A cartoon representation of BACE1 (grey) bound to the 2QZL ligand (pink). The H-bonds with G219 oxygen are highlighted in yellow. B) The work profiles of the H-bond during DUck. C) The density plot of the H-bond distance and the ligand's center of mass movement during the RAMD simulations. The density only appears at 2-3 Å because the simulations stop when the ligand exits the cavity.

Following the proposed protocol, I selected a representative ligand, in this case the ligand in the crystallographic structure with PDBid: 2QLZ . Using the dissociation trajectories obtained with τRAMD, I established the one-step dissociation mechanism, together with a high residence time (Figure 6.7.C). BACE1 has a loop closing the binding pocket, which opens during the unbinding event. This change conformation has slow kinetics, but does not seem to interfere with the ligand dissociation pathway creating meaningful metastable states. During the creation of the DUck chunk, however, this loop is removed due to the heavily restrained nature of the

simulations. The ΔG_{QB} profile of 2QZL ligand showed a very narrow distribution and a high free energy barrier of 18.4 kcal·mol^{-1} (Figure 6.7.B). Therefore, as the system fulfilled the two *a priori* conditions of our protocol, it was deemed as of high confidence for the prediction of AC using DUck.

BACE1 ligands present a bimodal ΔG_{QB} distribution (Figure 6.8). The majority of compounds yield ΔG_{QB} values between to 5-10 kcal·mol^{-1} and a smaller peak at very high quasi-bond free energies (15-20 kcal·mol^{-1}). Despite the lack of correlation between the experimental K_i values and the ΔG_{QB}, the AC predictions yielded a high accuracy (MMC = 0.556). This can be attributed to most cliffs being associated to the 2P4J ligand, with a ΔG_{QB} of 20.59 kcal·mol^{-1} and a K_i of 1.1 nM.

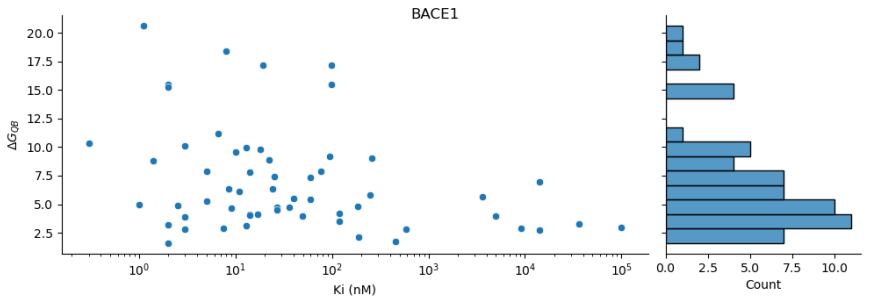


Figure 6.8. Relationship between K_i and predicted ΔG_{QB} of BACE1 inhibitors.

Overall, the results presented show that, despite being a very local property of protein-ligand complexes, ΔG_{QB} can be used to confidently predict a global property such as the ΔG_{bind}, at least with big potency changes as in AC. Ligands with a steep and single dissociation energy barrier fall under these criteria. Whilst the steepness of the barrier is given by the ΔG_{QB} itself, it is harder to ensure a one-step dissociation mechanism. Here τRAMD quickly

showed a clear signal between one-step dissociation ligands and ligands with a complex dissociation. Finally, these conditions seem to be extrapolable to the rest of the series given sufficient chemical similarity, defining a clear applicability domain.

6.3 OpenDuck: A python-based open-source implementation of Dynamic Undocking

As discussed in the previous section, assessment of the structural stability has shown promising applicability in the field of relative free energy calculations. Furthermore, it has also proven useful in other applications, such as in covalent docking,[284] binding mode predictions[177] and ternary complex formation in the context of molecular glues.[217] However, a factor limiting the widespread use of the concept and methodology is the lack of automatic protocols to set up, execute and analyze the simulations. The original DUck implementation was written as a set of Scientific Vector Language (svl) scripts. SVL is a scripting language exclusive to the Molecular Operational Environment software package (MOE),[179] a program only available under a purchasable license. Moreover, DUck was implemented only in AMBER which until the 22 was also licensed even for academics. The reliance on two expensive licensed programs and the rigidity of the implementation has difficulted its maintenance, distribution and the development of new implementations of the methodology. Thus, with the success of structural stability as a guiding principle in drug discovery, and the prospect of new applications coming up, a push towards an open source, flexible and accessible implementation of DUck became imperative.

6.3.1 Development of the OpenDuck python library

The development of OpenDUck follows the lead of other open source initiatives on computational chemistry such as OpenMM,[132] OpenForcefield[133] or BioSimSpace.[285] The underlying idea was to create a python-based implementation of DUck that surpassed the limitations imposed by the reliance on licensed software, while keeping or extending its performance and capabilities.

As such, OpenDuck is a python library that leverages libraries from OpenMM and OpenForcefield initiatives, and others. On the one hand, OpenMM is a high-performance library for molecular simulation that enables users to simulate molecular dynamics in a highly customizable manner. On the other hand, OpenForceFields provides a toolkit for the parametrization of molecules using various force fields (e.g. their own open force field, but also amber or machine learning force fields). OpenDuck can be used both as a stand-alone software and as a python toolkit, allowing for a more basic usage, based on the executable and, simultaneously, enabling advanced users to develop their own applications of structural stability. OpenDUck is stored and openly distributed through github https://github.com/CBDD/openduck), which also facilitates the involvement of other members of the community. In this regard, an initial proof of concept of the present implementation of OpenDUck has been developed as an academia-industry collaboration. The academia effort, by the group of Frank Von Delft and Brian Marsden from the university of Oxford and the group of Xavier Barril from the University of Barcelona. From industry, Anthony Bradley led the effort from XChem and Simon Bray from Gain Therapeutics. The initial implementation of OpenDuck

included the OpenMM implementation of the custom harmonic forces to emulate AMBER's SMD implementation. I have been responsible for further developments described hereafter, with support for testing and coding review from Simon Bray.

Mimicking DUck's main schema (Figure 3.3), OpenDUck consists of three main steps: a) the system preparation, b) the production of SMD trajectories and c) the processing of those trajectories to obtain the W_{QB} and free energy profiles (Figure 6.9). All these steps can be executed independently from OpenDUck's command line interface (CLI) and are also integrated in the full pipeline execution modes. The CLI is wrapped in a single executable (*openduck.py*) with different subcommands to allow independent execution of each step. OpenDuck supports two input modes, either as option flags given directly on the CLI, or using a single configuration file written in YAML.

During the preparation of the system for Dynamic Undocking we can distinguish three main steps, first the chunking of the protein, then the parametrization of the ligands and finally solvation box construction and receptor and ligand parameterization. The receptor *chunking* refers to the process of choosing a representative subset of the cavity (a chunk), big enough to conserve the environment of the protein-ligand interaction to study. The receptor *chunking* is a crucial step to reduce the computational cost of the simulations. From the user-specified interaction atom, the atoms outside a cutoff radius are removed. Afterwards, the residue selection is sanitized with a very simple heuristic by closing single amino acid breaks so the cappings do not overlap and removing unwanted crystallization buffer molecules in the receptor file.

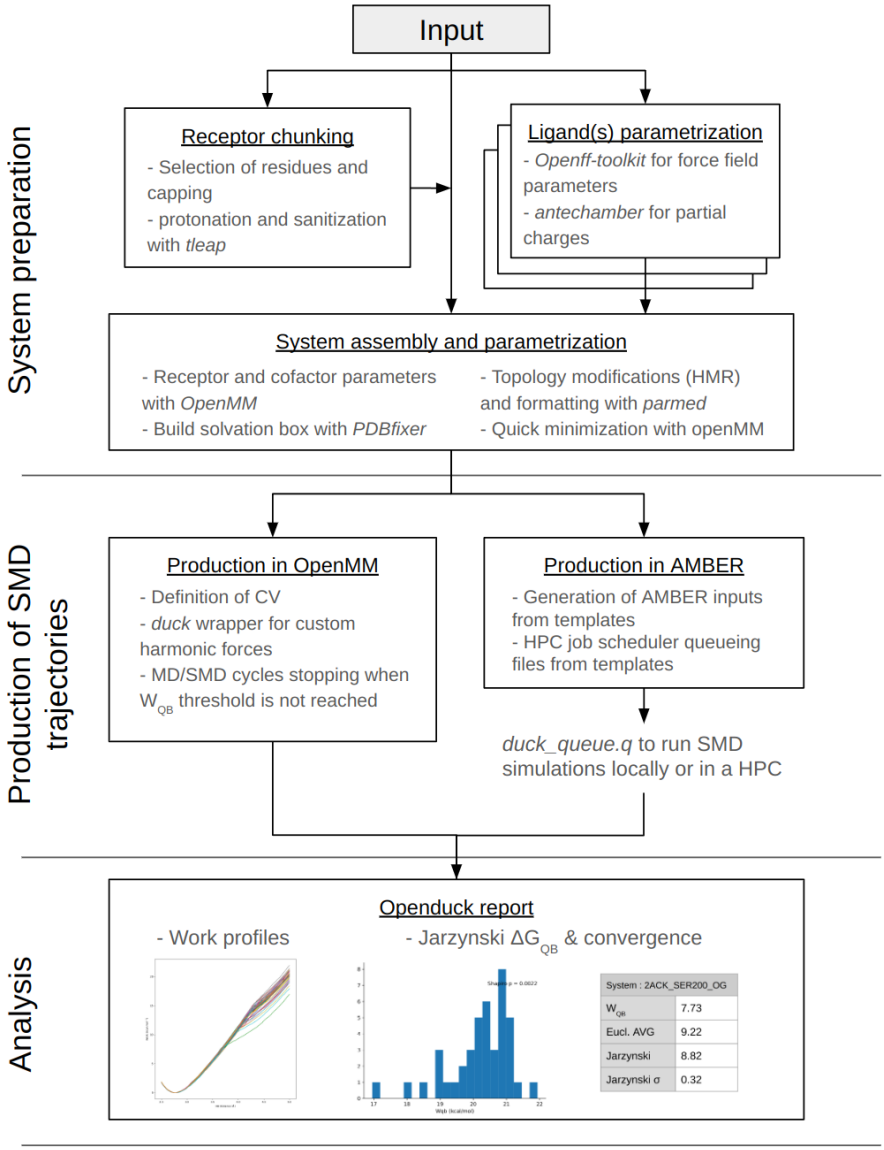


Figure 6.9. Workflow of the OpenDuck executable.

Then, the capping schema is established and the chunk is protonated with the *tleap* software from the AmberTools software package.[222] Once the receptor is prepared, the ligand and structural waters are added and the complex is solvated with the desired water model (e.g. adding extra dummy atoms for the dipole charge of

TIP4P). The parameters for each of the elements in the system (receptor, ligand, solvation, cofactors) are added stepwise to the complex topology.

Whilst the parameters of protein, ions and waters follow standard classical force field parameterization procedures, the parametrization of small molecules residues requires an *ad-hoc* functionality that can be easily adjusted for different needs. Currently, OpenDuck supports the calculation of partial charges using the semi empirical fitting AM1-BCC from the the *sqm* tool from the Ambertools package. The bonded terms in the topology (bonds, angles, dihedrals) are assigned using the openff-toolkit python library, which includes support for GAFF, GAFF2, the different versions of the open forcefield (SMIRNOFF, Parsley and Sage) and new force fields that use machine learning potentials (e.g. ESPALOMA[286]). Once charges and force field parameters are assigned, the system is transferred to a *parmed* topology object which handles the interoperability between MD engines. Topology-wise modifications like HMR are also supported at this stage leveraging the *parmed* python library.[287]

At this point, the execution of OpenDUck changes based on the simulation engine. On the one hand, if the simulations are to be done using the AMBER MD engine, the input files and job execution scripts system are created using a templating module of inputs and job scheduler queueing files. The templating system easily allows for the modification of relevant simulation parameters (e.g. steering speed, trajectories saved, wrapping, friction coefficient, etc.) and to adapt the execution of the calculations to an HPC facility or a local computing environment. On the other hand, if the user wishes to use

openMM, the production simulations are integrated directly into the OpenDuck library.

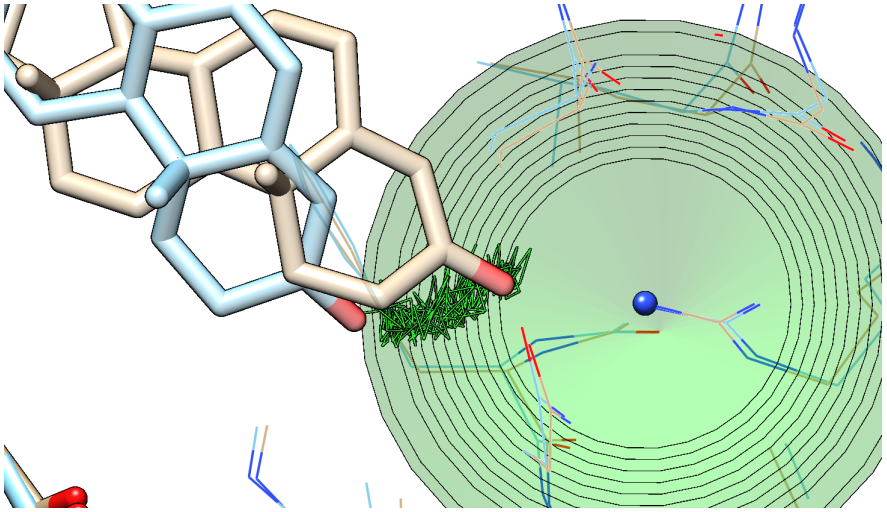
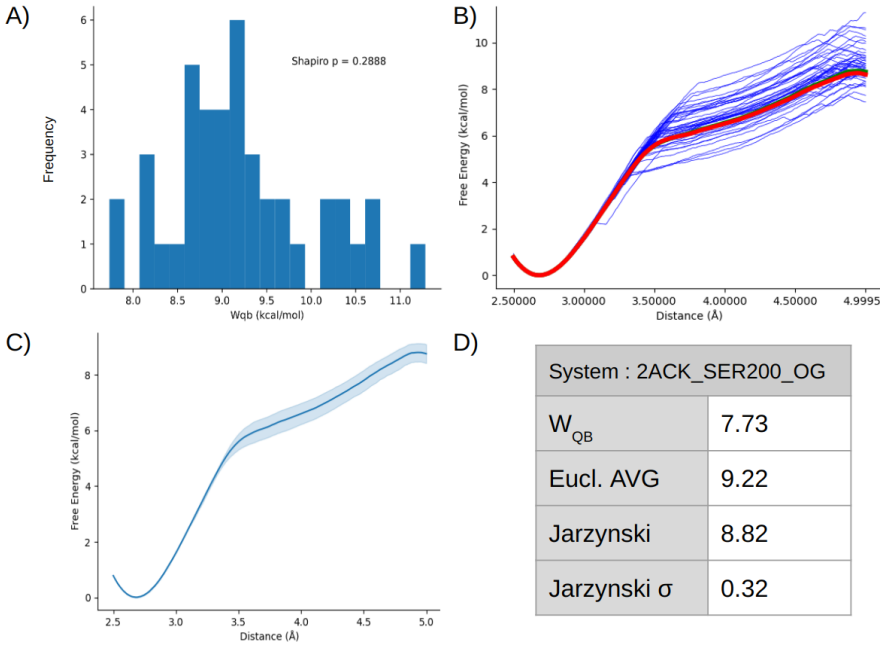


Figure 6.10. Representation of the updating schema for the steered vector in OpenMM. Represented in light yellow there is the equilibrated initial position of the protein-ligand complex, and in light blue the SMD endpoint. The green circles show the steps where the vector is updated to keep orthogonality to the arginine nitrogen (marked with a blue sphere).

It is important to note that the native implementation of SMD on AMBER and OpenMM is different. In OpenDuck, an ad-hoc implementation of SMD in OpenMM was developed to mimic the AMBER behavior. In AMBER's implementation of SMD, a distance between two atom masks (receptor and ligand atoms that perform the H-bond) is defined with no end-point nor specific vector force. This allows for, if the steering speed is correctly set, the dissociating trajectory to relax orthogonal degrees of freedom *'on the fly'* when the ligand moves not in the optimal direction due thermal fluctuations. On the contrary, in OpenMM we define a harmonic potential with a stiff spring and use a vector to a specific endpoint calculated based on the initial H-bond direction. This implementation is very sensitive to deviations of the ligand orthogonally to the

steering vector, which usually result in clashing with the receptor during dissociation. To correct this,we update the vector direction every 400 fs (Figure 6.10). The force needed to maintain the steering velocity is integrated accounting for the real distance transversed (including 'horizontal' movement) and reported as work. Finally the normalized work is used as the work threshold at the end of each SMD step to decide if continuing the simulation or not.



System : 2ACK_SER200_OG	
W_{QB}	7.73
Eucl. AVG	9.22
Jarzynski	8.82
Jarzynski σ	0.32

Figure 6.11. Example of an OpenDUck report of the SER200_OG H-bond from the 2ACK pdb in Iridium. A) A histogram of the normalized work values obtained for each SMD trajectory together with the resulting p value of the Shapiro-Wilk normality test. B) The work profiles per replica (in blue) and the jarzynski free energy per distance step during the dissociation. C) The jarzynski free energy with the bootstrap σ as the error area. D) A sample output table.

OpenDUck incorporates an analysis module to calculate and report W_{QB} per ligand or per replica, ΔG_{QB} from the jarzynski equality (equation 3.12.b) and perform bootstrapping to assess the error in the ΔG_{QB}. Finally, plotting is also handled by the module, creating a

final report with the normality assessment, work trajectories and free energies with their confidence interval (Figure 6.11).

6.3.2 Benchmarking and validation

6.3.2.1 The SERAPhiC and Iridium datasets

To validate the OpenDuck implementation I compared the ΔG_{QB} predictions for all H-bonds formed by the protein-ligand complexes of two datasets: Iridium and SERAPhiC. These two datasets were employed previously in the research group by Maciej Majewski, to investigate the structural stability of H-bonds in crystalized protein-ligand complexes and to use that information to predict their correct binding mode.[177,181] Maciej filtered the original dataset from 120 protein-ligand complexes to 78, excluding complexes with coordination with metal ions, multiple ligands and ligands without polar interactions. The resulting protein-ligand complexes covered a wide variety of protein families (mainly kinases, proteases and nuclear receptors), with the majority of the ligands bound to the orthosteric binding site. The SERAPhiC data set consists of fragment-protein complexes, which after a similar filtering was reduced to 53 systems. The distribution of the calculated W_{QB} on both datasets shows that whilst the Iridium complexes have higher structural stability than the SERAPhiC, their proportion of robust to labile interactions is very similar (Figure 6.12). The widespread W_{QB} distributions and the differences in molecular weight between the fragments and ligands from SERAPhiC and Iridium respectively, make these datasets very adequate to benchmark OpenDUck.

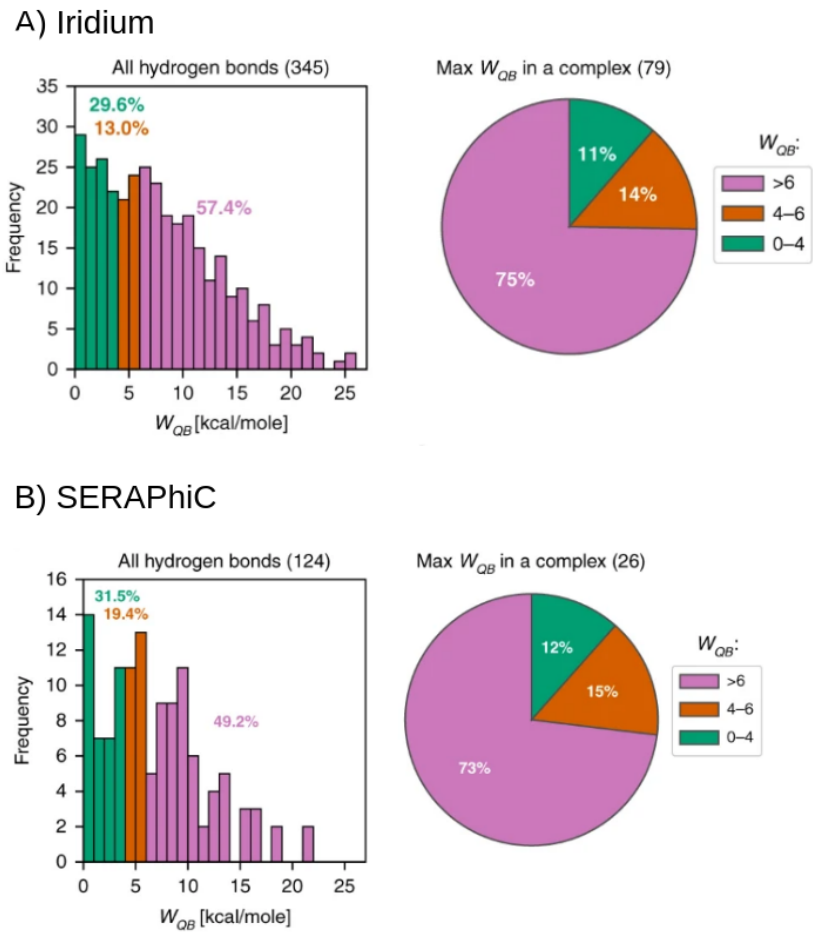


Figure 6.12. Distribution of W_{QB} in the Iridium and SERAPhiC datasets. The labile, intermediate and robust interactions are represented in green, red and pink respectively. A) shows the distribution of W_{QB} for the Iridium dataset, with the proportion of labile and robust classification of the most robust H-bond in each complex. The mirrored analysis is shown for SERAPhiC in B). Adapted from [181].

I simulated the same prepared chunks used by Maciej using different conditions to assess the reproducibility of the AMBER results in OpenMM, to test new conditions that can potentially enhance the performance of the method and to exemplify the flexibility of the OpenDUck implementation.

6.3.2.2 Reproducibility of AMBER results with OpenMM

Firstly I addressed the reproducibility of the results obtained in the original AMBER execution with the new steering schema in OpenMM. For that, I compared the results obtained from the MOE preparation and AMBER execution to the simulations using the same force field, but simulated in OpenMM (see section 3.2.3.2 for the simulation details). Figure 6.13 shows that there is a significantly good correlation (r = 0.85) between the energy values obtained through both engines, however the reported mean absolute error (MAE) is higher to the standard shown by other free energy calculating methods (i.e. being the accepted discrepancies between converged predictions of different MD engines around 1 kcal/mol,[288] versus the 2.6 kcal/mol I obtained). However, the uncertainty of the work calculations (σ) for most H-bonds is lower than 1 kcal/mol, with a higher proportion in AMBER (89,7%) and in OpenMM (82,4%). This trend seems to be system dependent and mostly restricted to the Iridium dataset, where calculations with low σ moved from 92% to 80% from AMBER to OpenMM. This could be optimized by adjusting the frequency of updating the steering vector, at the expense of a higher simulation cost.

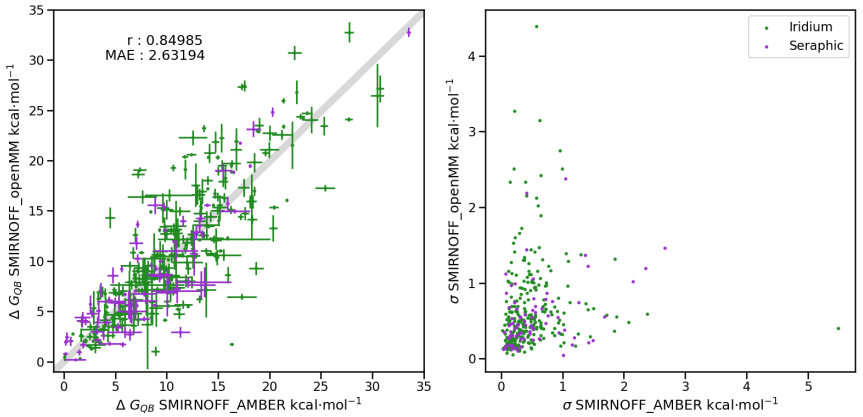


Figure 6.13. Comparison of predicted ΔG_{QB} and their standard deviation between OpenMM and AMBER. The Iridium and SERAPhiC datasets are shown in green and violet respectively. The error bars correspond to the standard deviation. The Pearson coefficient (r) and the mean absolute error are shown for the joint comparison.

6.3.2.3 Access to different force fields and system conditions

On top of emancipating DUck from licensed programs, the aim of OpenDUck is to allow for a higher flexibility in the simulations setup and analysis. In that regard, OpenDUck facilitates to easily change many parameters such as forcefields, water models, the ionic strength, the restraining mask or the collective variable that might be needed for specific projects. To verify the validity of the preparation I simulated the systems modifying these parameters (Table 3.1). The subsequent comparisons were not aimed at evaluating the accuracy or adequacy of the specific parameters, but more so on the ability of the library to handle them.

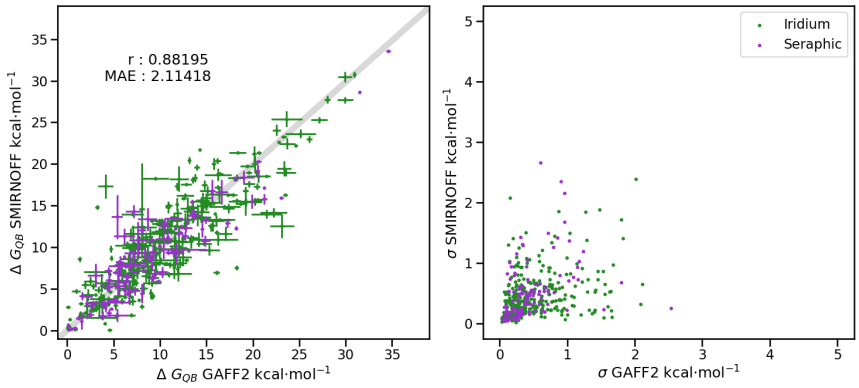


Figure 6.14. Comparison of predicted ΔG_{QB} and their standard deviation using GAFF2 and SMIRNOFF small molecule force fields. The Iridium and SERAPhiC datasets are shown in green and violet respectively. The error bars correspond to the standard deviation. The Pearson coefficient (r) and the mean absolute error are shown for the joint comparison of both datasets.

In Figure 6.14 there is the comparison between the GAFF2 and the SMIRNOFF99Frosst small molecule force fields. With a Pearson correlation of 0.88 and a MAE around 2 kcal/mol, it shows a closer correspondence than when changing the simulation engine. Moreover, the dispersion of the predictions is also very consistent between the two force fields and datasets. This, however, can be very much dependent on the nature of the ligands to parametrize and outliers could appear in future projects. Through providing the flexibility to select the desired force field I can ensure OpenDUck users can adopt the best forcefield for their needs.

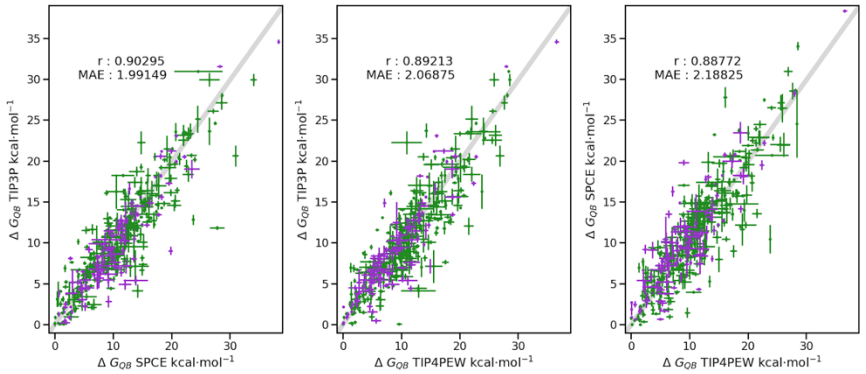


Figure 6.15. Comparison of predicted ΔG_{QB} using SPCE, TIP3P and TIP3PEW water models. The Iridium and SERAPhiC datasets are shown in green and violet respectively. The error bars correspond to the standard deviation. The Pearson coefficient (r) and the mean absolute error are shown for the joint comparison of both datasets.

A similar situation can be observed when changing the water model. In this case I simulated using extended simple point charge (SPCE), TIP3P and TIP4PEW water models. The different levels of complexity in the description of waters and the different physicochemical properties these models have, might be beneficial for different protocols (e.g. simulations on intrinsic disordered proteins). The preparation of these solvation boxes is very similar, however in TIP4PEW an extra charged particle is added to each water to better represent the dipole moment. At the moment, 5-point and polarizable water models are not compatible with openMM. Thus, I excluded them to maintain interoperability between the two MD engines. The correlation between the calculated ΔG_{QB} with the different water models and the error is very similar to what was observed when changing the small molecule force field (Figure 6.15). This indicates that this correlation might be the baseline between independent executions of OpenDUck. In Figure 6.16 we can see that the hysteresis in all the calculations is very similar independently on the water model. The TIP3P model shows a slight

improvement in respect to the others with a higher peak at σ ≅0.5 kcal/mol. This matches the recommendation to use TIP3P when parameterizing the proteins with AMBER f14SB, as is the case in these simulations. However, other water models have shown higher accuracy reproducing experimental phenomena during simulations (e.g. TIP4PEW for simulating IDP conformational behavior[289]).

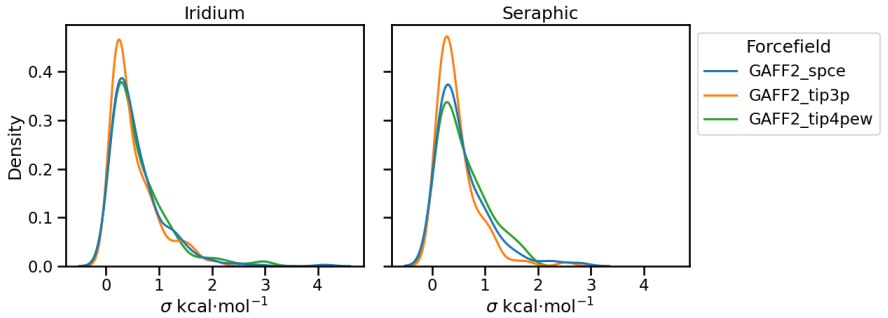


Figure 6.16. Distribution of the standard deviation from the bootstrapped ΔG_{QB} calculations using SPCE, TIP3P and TIP3PEW water models.

While the high MAE depending on the conditions used is concerning, OpenDuck is devised primarily as a VS tool, discarding molecules with labile or unstable H-bonds after docking. Following that criteria, I compared the classification of stable (ΔG_{QB} >= 6 kcal/mol) and unstable (ΔG_{QB} < 6 kcal/mol) using the different water models. There is a very low mismatch (<10%) in the classification between the three comparisons (Figure 6.17). Furthermore, the alignment in the qualitative predictions could be optimized by changing the ΔG_{QB} threshold, especially for the fragments in the SERAPhiC dataset. We usually consider H-bonds with a ΔG_{QB} < 4 kcal/mol robust (or stable) if they come from a fragment. This would push the predictions mismatch down to a 6%.

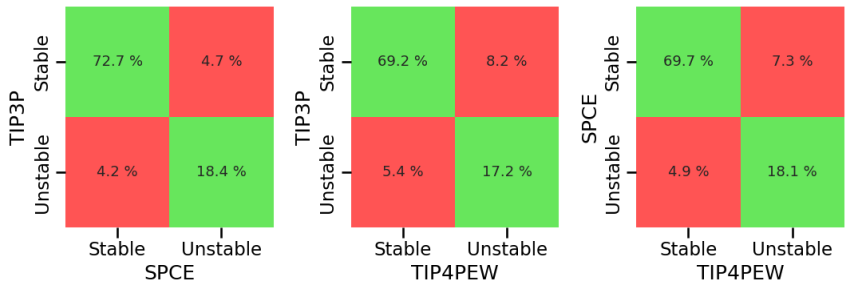


Figure 6.17. Comparison of H-bond classification using different SPCE, TIP3P and TIP3PEW water models. The mismatching classifications are colored in red and the matching classifications are colored in green.

Adding counter ions to neutralize the system's charge is a necessary step for simulations using PBC. This is particularly important when using particle mesh Ewald (PME) to calculate long range electrostatics. In addition, the presence of ions in the solvent has an important effect in many properties of the bulk solvation that might impact protein folding, ligand binding, phase-separation phenomena, etcetera.[290] On top of that, most, if not all, experiments are performed under non-zero salt conditions. Thus, I considered it very important being able to easily control the ionic strength for OpenDUck simulations. As seen in Figure 6.18 the correlation in ΔG_{QB} when adding a flat ionic strength on top of the neutralizing ions is similar to the previous comparisons changing other parameters ($r \cong 0.9$ and $MAE \cong 2$). Yet, when pushing the ionic strength to higher concentrations (e.g. 0.2-0.5M) I observed a significant increase in the instability of the integrators, regardless of the MD engine. Increasing the buffer space of the solvation box reduced the integrator instability events, although it resulted in a significantly lower simulation speed. For the sake of comparing single changes in parameters I restricted the correlations to 0.1M ionic strengths.

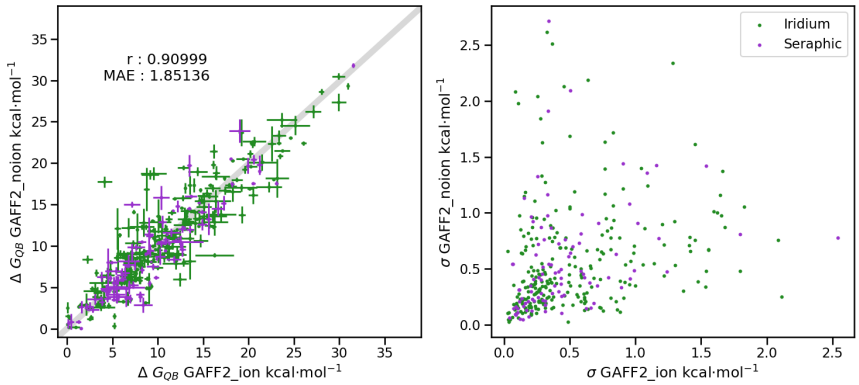


Figure 6.18. Comparison of predicted ΔG_{QB} and their standard deviation changing the solvation ionic strength. Simulations were performed only using counter ions (*noion*) and using a 0.1M ionic strength (ion). The Iridium and SERAPhiC datasets are shown in green and violet respectively. The error bars correspond to the standard deviation. The Pearson coefficient (r) and the mean absolute error are shown for the joint comparison of both datasets.

6.3.2.4 Evaluating speed and convergence

Dynamic Undocking, therefore OpenDuck too, is a program meant to yield a description of the protein-ligand complexes in a fast, high throughput manner. As seen in section 6.2, this comes with a high compromise on accuracy, especially with ligands with complex dissociation pathways. ΔG_{QB} corresponds to the steepness of the first dissociation barrier and does not correlate directly with ΔG_{bind}. Hence, DUck was not designed to compete in accuracy with rigorous free energy calculation methods like FEP or TI, but to quickly discard ligands with labile H-bonds. To this extent, speed is a very important parameter to optimize in OpenDUck.

The first considerations in speeding up the OpenDUck pipeline focus on the preparation and parametrization steps. In that regard, the current implementation supports a highly parallelizable set up. When the CLI is fed multiple ligands in a single *sdf*, multiple child processes

are spawned for the ligands' parametrization. On the other hand, shared steps in the pipeline as the receptor preparation are shared between the child processes. The re-sampling strategies are also parallelized during the analysis to evaluate the ΔG_{QB} and convergence of large libraries in a matter of minutes.

On the other hand, I optimized the production phase by adjusting parameters such as the time-step and the solvation box size. HMR is a modification of the topology where the mass of heavy atoms is partially shared with the hydrogens bound to them.[194] The optimal ratio of mass sharing has been shown to be engine dependent, but the general consensus has been increasing the mass of hydrogen atoms to 3 atom mass units. Combined with SHAKE, HMR allows to increase the time step to 4 fs, effectively doubling the simulation speed.

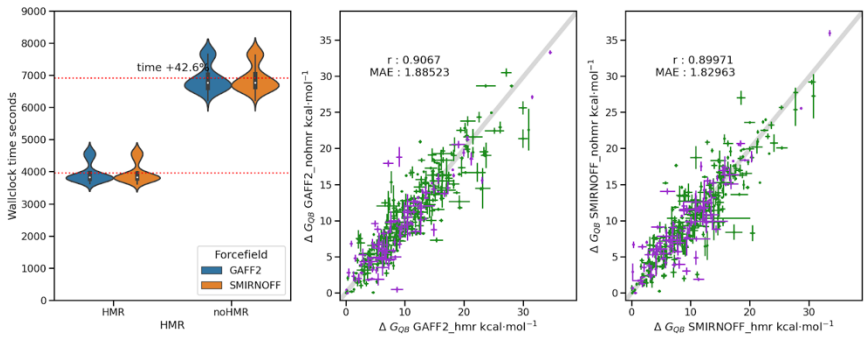


Figure 6.19. Comparison of predicted ΔG_{QB} and the accumulated simulated time with and without HMR. Simulations were performed using the GAFF2 and SMIRNOFF99Frosst forcefields. On the left-most plot, the distributions of wallclock time spent on the simulation of all the H-bonds applying HMR and not, using the GAFF2 (blue) and SMIRNOFF (orange) in AMBER. On the right plots, the correlation of the predicted ΔG_{QB}. The Iridium and SERAPhiC datasets are shown in green and violet respectively. The Pearson coefficient (r) and the mean absolute error are shown for the joint comparison.

In Figure 6.19 we can observe that applying HMR halves the simulation time almost in half. The mismatch with the theoretical gain in speed comes from the minimization, where the simulation steps do not apply and the data transferring time spent between the disks and the CPU, which is constant. The calculated ΔG_{QB} show a comparable correlation as the other parameters ($r \cong 0.9$ and $MAE \cong 2$), even when combining it with a change in small molecule force field.

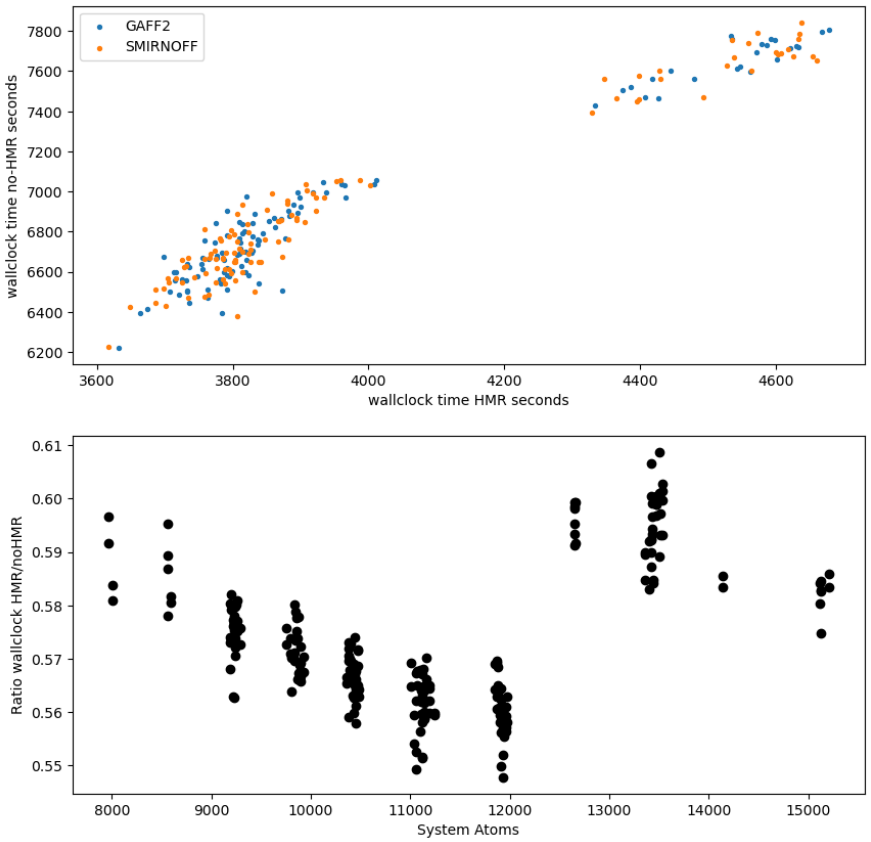


Figure 6.20. Speed increase in HMR and non-HMR simulations and its dependency on the number of atoms. At the top, the relation of wallclock time for simulations with the GAFF2 (blue) and SMIRNOFF (orange) force fields. At the bottom, the ratio of HMR/nonHMR wallclock times in respect to the atoms in the solvated system.

Surprisingly, the relationship between simulation times when using HMR is not constant, and decays for bigger systems (Figure 6.20). Moreover, there is a jump in the ratio when surpassing the 12K atoms in the system, going from a 0.55-0.58 reduction in wallclock time, to 0.59-0.60 at 13K atoms. This behavior seems to be hardware dependent (data not shown), and I hypothesize that it is related to the cache capacity to communicate between the GPU, CPU and then the disk. Nonetheless, the computational time saving is significant even in the least optimized systems, and very close to the theoretical 0.5 ratio in all cases.

An alternative way of quickening the simulations is reducing even more the amount of atoms in the system. It is possible to reduce the atoms by either being more stringent with the chunking procedure (i.e. selecting a smaller distance cutoff) or reducing the buffer space between the protein and the PBC at the end of the solvation box. Reducing the residues in the chunk is very system dependent and can easily introduce artifacts on waters quickly permeating to regions supposedly shielded. On the other hand, due to the very stiff restraints applied to the chunked receptor, reducing the size of the solvation box appears to be a more generalizable and safe manner to reduce the atoms to simulate. I simulated two different buffer distances for the solvation box: 18Å as in the original DUck implementation (big box) and 10Å (small box). The box size had no impact in the calculation of ΔG_{QB} nor the apparent dispersion of the results (Figure 6.21), which is consistent with similar studies previously reported.[291] The small box size can create artifacts, however, if the chunk or ligand is highly charged and their coulombic potential might reach to themselves through the PBC. With the small box schema, I obtained a ~35% reduction in the simulation time,

which combined with the ~40% of using HMR significantly speeds up the simulations in OpenDUck compared to DUck.

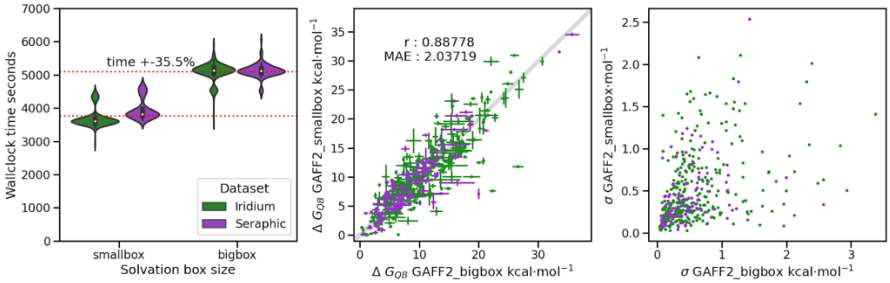


Figure 6.21. Comparison of predicted ΔG_{QB}, dispersion and the accumulated simulated time with the different solvation box sizes. On the left-most plot, the distribution of wallclock time spent on the simulation of all the H-bonds using the two solvation box sizes is represented. On the right plots, the correlation of the predicted ΔG_{QB} and the standard deviation are shown. The Iridium and SERAPhiC datasets are shown in green and violet respectively. The Pearson coefficient (r) and the mean absolute error are shown for the joint comparison.

Since all comparisons yielded a very similar correlation coefficient, error and dispersion I simulated a set of conditions using a very high amount of replicas (i.e. 500 SMD cycles) to assess convergence. When bootstrapping different amounts of sampling steps from the 500 replicas there is a general trend of lower ΔG_{QB} values and in narrower distributions for larger subsamples (i.e. lowering the ΔG_{QB} approximately 1-3 kcal/mol and the σ 0.5-1.5 kcal/mol). Figure 6.22 shows an example of a H-bond following this trend. For other, more noisy systems, bigger shifts can occur when a new ligand conformation is sampled after several nanoseconds. These changes in binding mode are not necessarily relevant, especially as all the benchmark dataset come from crystallized structures and are most likely an artifact created by the chunk. On the other hand, HTVS poses might change when equilibrated in MD in a much faster manner.

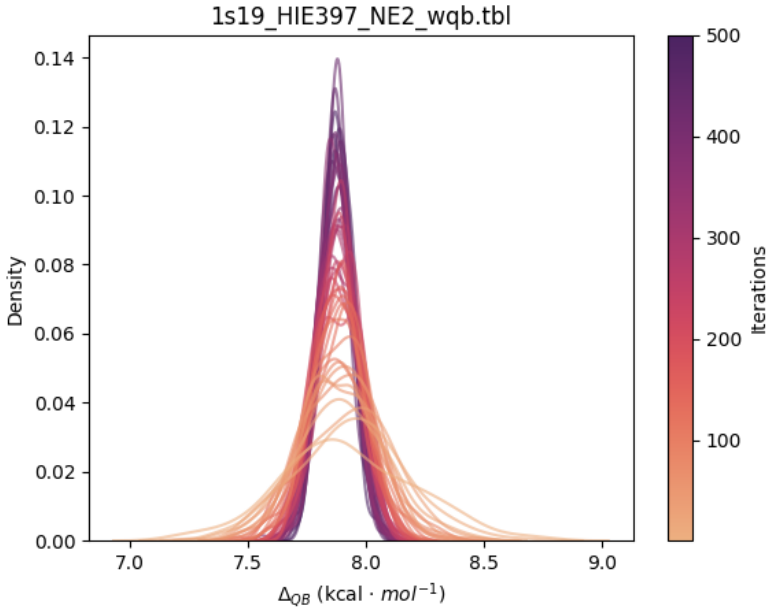


Figure 6.22. Distribution of bootstrapped ΔG_{QB} predictions across different amounts of iterations for the HIE397 H-bond in the 1S19 PDB structure.

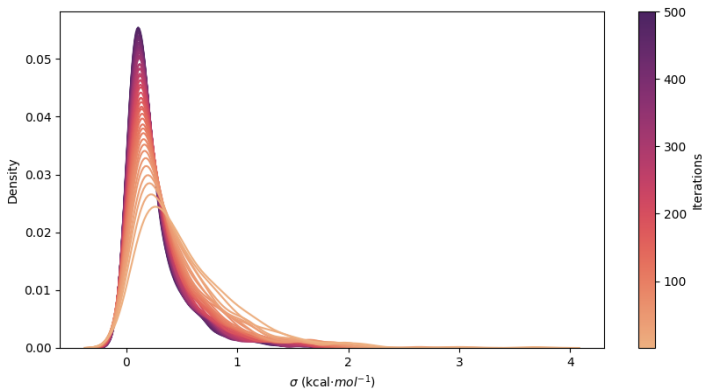


Figure 6.23. Distribution the standard deviations for the predicted ΔG_{QB} across different amounts of iterations for the Iridium and SERAPhiC datasets.

Figure 6.24 shows that the default 20 cycles compares very tightly with a very oversampled simulation, with a correlation of 0.99 and a MAE of only 0.35 kcal/mol. This result confirms that 20 SMD cycles

has reached enough convergence and, especially in the context of a HTVS campaign, the gain with further sampling is negligible. With higher replicas, however, the work distributions are narrower, especially in noisy systems (Figure 6.23). This high correlation contrasts with the baseline noise that all the other comparisons share. To evaluate the source of the hysteresis I repeated the same analysis, but comparing two completely independent executions, with different equilibrations instead of extending a single execution to a higher number of SMD cycles. With independent equilibrations and with a similar amount of replicas, the correlation worsens back to the baseline noise observed in the other comparisons (Figure 6.25). This highlights the impact of the initial conditions (i.e. velocities) in the final ΔG_{QB} calculations.

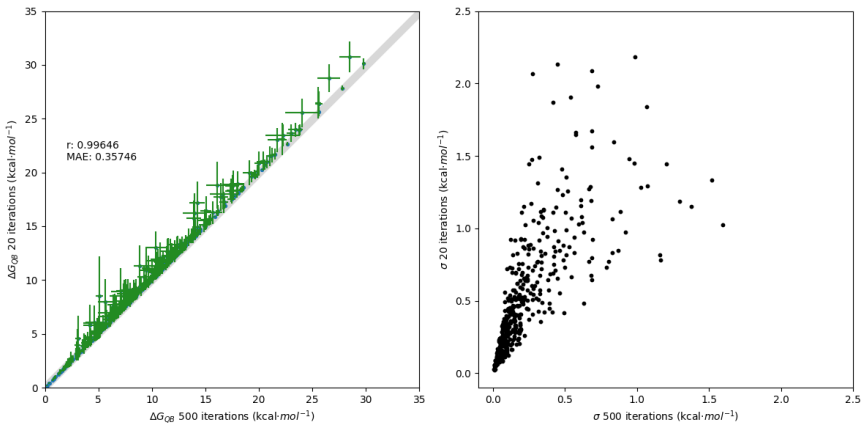


Figure 6.24. Correlation of calculated ΔG_{QB} and their dispersion from 500 and 20 SMD cycle simulations from the same equilibration. Values shown correspond to both Iridium and SERAPhiC simulations. The error bars correspond to the standard deviation.

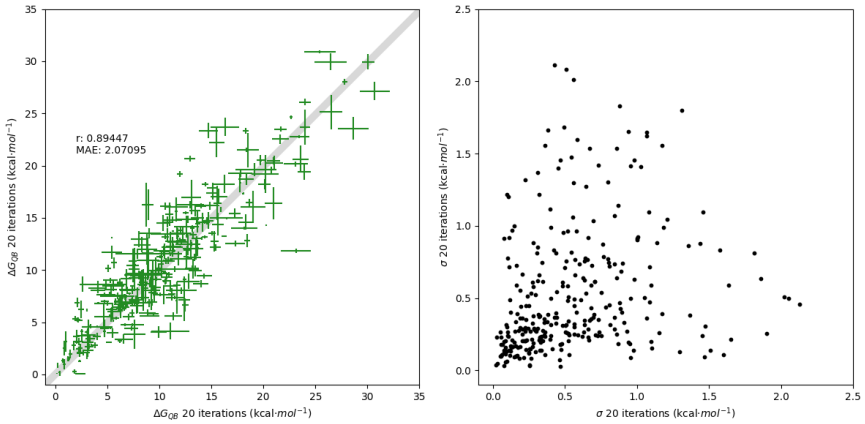


Figure 6.25. Correlation of calculated ΔG_{QB} and their dispersion from 20 SMD cycle simulations from different equilibrations. Values shown correspond to both Iridium and SERAPhiC simulations. The error bars correspond to the standard deviation.

6.3.2.5 OpenDUck library as the base for new protocols

Whilst OpenDUck's CLI allows for many changes and variations, as shown in the previous sections, it is still focused around the DUck pipeline. The current developing efforts exploiting the structural stability concept need broader flexibility. In order to facilitate such efforts I have developed the *duck* python library to accompany the CLI. The library includes several functions to parametrize and generate topologies, to set up SMD in AMBER and OpenMM, analyze the resulting work profiles and to extract free energies from them. Ideally, following the community driven philosophy of the project, after testing and validating new applications, they could be added as independent modules in the main *openduck.py* executable.

6.3.2.6.1 Steering water-mediated hydrogen bonds

To prove the ease of generating such pipelines, I adapted the DUck procedure to assess the structural stability of water bridges. From

RESULTS: STEERED MOLECULAR DYNAMICS FOR DRUG DISCOVERY

Maciej initial evaluation of the Iridium dataset predictions he encountered a high mismatch of some ligands' W_{QB} with their potency. These three ligands come from the 1YVF, 1F0S and 1G9V PDB entries and all interact with stable water networks in addition to the protein receptor (Figure 6.26). Even then, he hypothesized then that the most stable interaction of those complexes was the water-mediated H-bond with water networks. Still, due to the available DUck implementation it was impractical to assess them, and the investigation was abandoned. Now with OpenDUck I recovered his idea and created a water-steering protocol using the *duck* python library. In this case the pipeline is very similar to the standard OpenDuck execution but the CV is generated using the water as reference. Moreover, the restraining schema needed to be adapted to maintain the water network in place, equilibrating the orientation of the hydrogens and assessing the new work minimum. The resulting work profiles in Figure 6.26 show that the water-mediated in 1F0S ligand is very robust (i.e. W_{QB} = 9.2 kcal/mol), in contrast with the direct H-bonds with the protein. This corresponds with the high reported Ki. On the other hand, the ligand in 1YVF has a very labile water-mediated H-bond. In this case, however, the H-bond is performed with two waters, which can create a *pivoting* effect when steering where the acid swaps to the other water. Finally, despite the high dispersion in the 1G9V complex, it has a W_{QB} around 4 kcal/mol which was higher than the other reported H-bond.

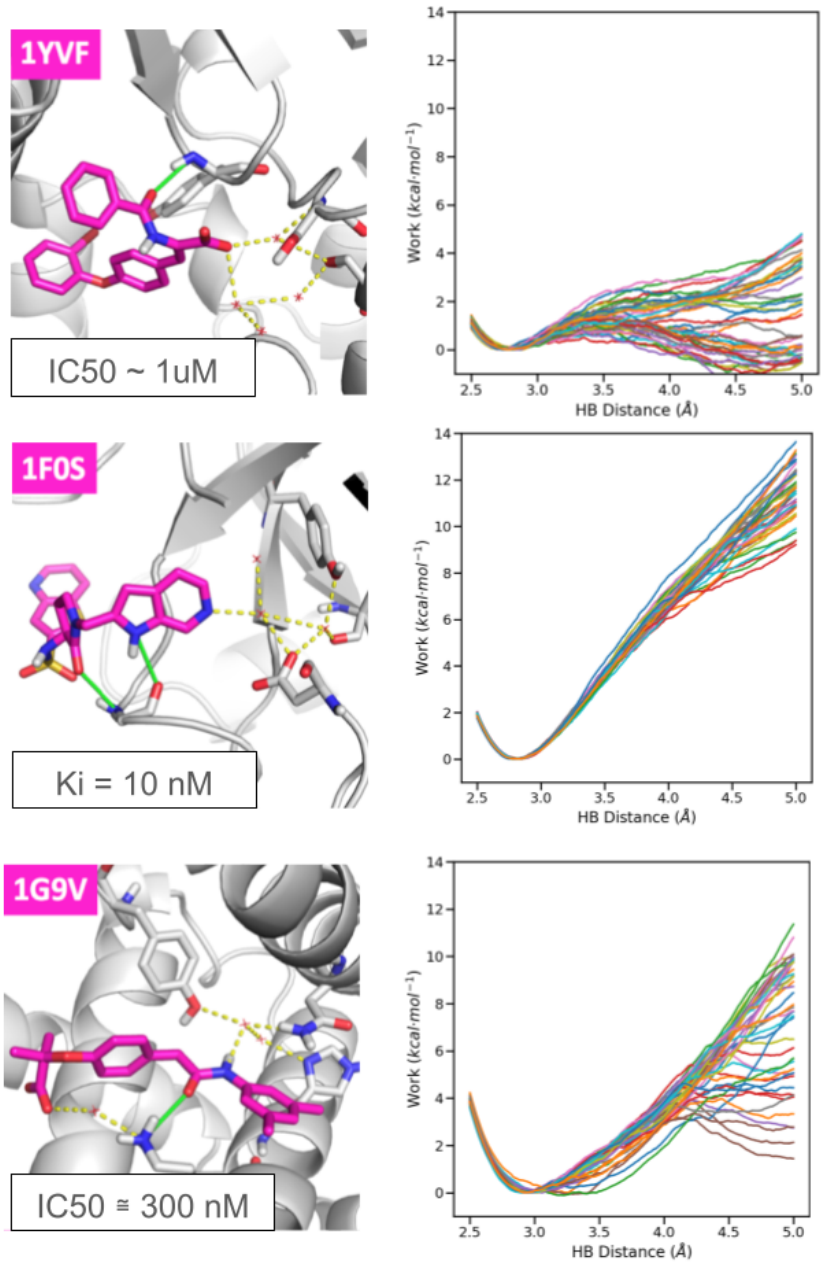


Figure 6.26. Representation and work profiles of the water-mediated H-bonds in 1VYF, 1F0S and 1G9V structures. H-bonds between the ligands (magenta) and protein receptors (gray) are highlighted in green. The H-bonds with waters are highlighted in the yellow dotted lines. The shown work profiles correspond to steering the water-mediated H-bonds.

Several other similar procedures have been tested by me and other members of the group, such as steering ligands in a lipidic environment (i.e. POPC membrane), using metal coordinations or evaluating protein-protein interactions. Altogether, it demonstrates the potential of OpenDUck to facilitate further development regarding SMD and structural stability for drug discovery and modeling of relevant interactions.

6.4 Non-equilibrium free energy calculations using the Crooks fluctuation theorem

All the estimates of free energy values from the SMD work profiles I have shown until now utilize the Jarzynski equality (JE). An alternative, more general relationship between out-of-equilibrium work (W) values with free energy was proposed in 1999 by Gavin Crooks in the framework of the fluctuation theorem, the Crooks fluctuation theorem (CFT).[174] The CFT states that if the dynamics of the system fulfill the microscopic reversibility principles, then the forward trajectory is exponentially more probable than the reverse, due to the entropy generated (equation 6.1). Interestingly, the JE can be recovered by integrating Eq. 6.1 over W.

$$\frac{P_{\mathrm{Forward}}(W)}{P_{\mathrm{Reverse}}(-W)} = e^{\beta(W-\Delta G)} \qquad (6.1)$$

In the Crooks Equation, $P_{\mathrm{Forward}}(W)$ and $P_{\mathrm{Reverse}}(W)$ are the normalized probability distributions of the work values obtained from simulating the forward and reverse transformation. Given enough overlap between the work distributions the free energy difference (ΔG) can be directly estimated through the Eq. 6.2.

$$\widehat{\Delta G} = W + k_\beta T ln \frac{P_{Forward}(W)}{P_{reverse}(-W)} \qquad (6.2)$$

This equation simplifies to G=W at the intersection of the work distributions, where the entropy generated in the forward process (in the form of dissipative work) is minimized. This phenomenon has been validated experimentally in the context of RNA folding and unfolding.[292]

Theoretically, the CFT has the benefit over the JE that the rare events of low work dissipation which occur in the forward process, can be encountered more frequently through exploring the reversed process. By virtue of making those rare events more frequent, the calculation of ΔG should converge faster. However, this approach has known limitations: firstly it may be difficult obtaining the needed overlap between the forward and reverse transitions.[293] Secondly, it might be difficult to converge the reverse processes to the initial state of the forward simulations. Finally, the tails of the work distributions, defined by the rare events, still contribute highly to the free energy differences. To partially address these problems, the Crooks Gaussian Intersection (CGI) was proposed,[294] which approximates the work distribution to Gaussian functions to analytically derive the theoretical intersection. The CGI relies on the Gaussian approximation, which does not stand for many of the cases I have studied previous. Regardless of previous findings, our final goal is to obtain good free energy estimations in low sampling scenarios, which is incompatible with the requirements for Gaussian distributions. For this reason I set to apply the CFT without assuming a Gaussian distribution of works, using non-parametric methods such as the kernel density estimation (KDE). The end goal was to

develop a protocol that speeds up obtaining estimates of ΔG_{QB} ,while maintaining comparable levels of robustness and accuracy.

6.4.1 Free energy calculations of CRBN-CK1α-molecular glues ternary complex

First, I aimed at exploring the application of CFT to assess the strength of H-bonds in ligand-stabilized protein-protein interactions. To this end I reproduced the potential mean force (PMF) calculations previously performed in the research group by Miñarro et. al. on the lenalidomide dependent CRBN-CK1α complex.[217] The PMF they calculated are equivalent to what I have considered as ΔG_{QB} in the previous sections of the chapter (i.e. the free energy obtained from bringing a H-bond from 2.5Å to 5Å). Thus, for the sake of consistency I will keep the ΔG_{QB} nomenclature. They characterized the stability of the complex by the ΔG_{QB} extracted by breaking the three H-bonds between CRBN and CK1α. They observed that the presence of lenalidomide modulates the work needed to break the H-bonds by shielding them from water molecules (Figure 6.27). Moreover this descriptor was sensitive to mutations in CK1α and translated very well to ternary complex stabilization measured experimentally.

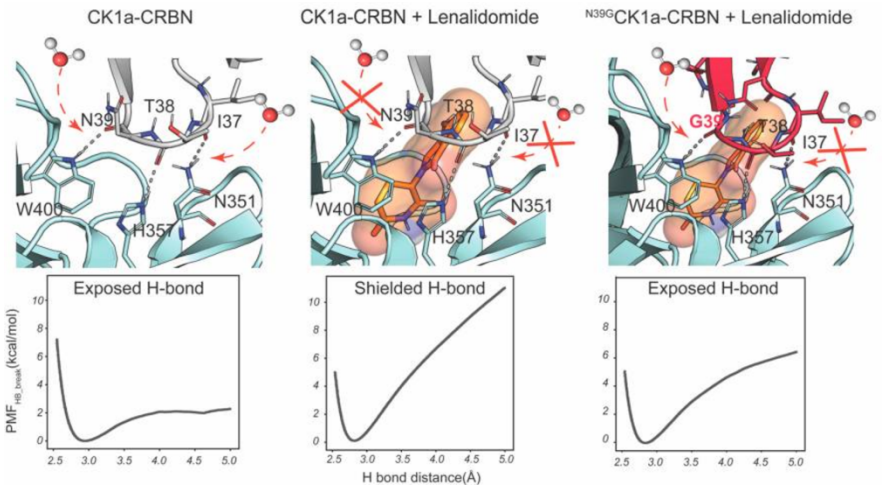


Figure 6.27. Representation of the proposed mechanism of stabilization of the CRBN-CK1α H-bonds by lenalidomide and the effect of CK1α mutations. On the top row, the representation of the CRBN (light blue), CK1α-wt (white), and CK1α-N39G (red) complexes with lenalidomide (orange). The drawn waters indicate the hypotheses of water shielding by lenalidomide, which reflects on the PMF profiles shown at the bottom. Extracted from [217].

In order to apply the CFT, I added a final step to the H-bond breaking SMD simulations, reversing it by reforming the interaction broken in the previous step (Figure 6.28). To assess the intersection between the work distributions I performed a KDE using Gaussian hills. In this manner the KDE will be equivalent to a Gaussian if the given process fulfills normality, allowing for direct comparison to CGI. Moreover, the magnitude of the smoothing can be recovered to assess the confidence on the probabilities derived from it.

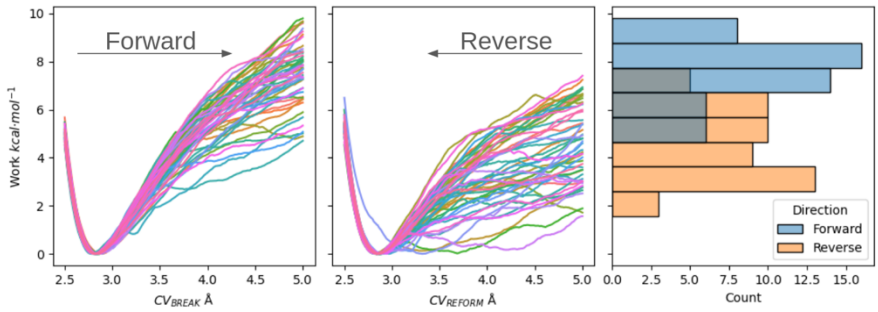


Figure 6.28. Example of forward and reverse work profiles from the SMD simulations. The represented simulations correspond to the breaking and reformation of the I29-N542 H-bond of the CRBN-CK1α-lenalidomide complex.

Due to the nature of the systems, the authors opted to steer at a constant speed of 0.5 Å/ns in contrast to the 5 Å/ns I use for OpenDUck. Moreover their protocol included extensive sampling (100 replicas) with reassignment of velocities according to a Boltzmann distribution at the simulation temperature, which might contribute to address the convergence mismatch I detected in section 6.3.2.4. In addition to reproducing the results reported in the article using the CFT, I assessed the convergence of the calculated free energy values when using higher steering speeds and lowering the sampling. The smoothed forward (H-bond break) and reverse (H-bond reforming) work distributions are shown in Figures 6.29 and 6.30 for the complete set of complexes reported in the original study. Most distributions overlap at a single point. In some cases, seen in both figures, some distributions intersect in more than one work value. This is due two main reasons; (i) conformational changes that create shifts in the work profiles, changing the location of the minima and (ii) very wide distributions (i.e. with a high hysteresis). By selecting the work intersection with the highest probability, given by its density, the estimated ΔG_{QB} values deviated from previously reported values 0.9 kcal/mol on average. At higher steering speeds

(from 0.5 to 5 Å/ns), the work distributions shift to higher values for both directions. The works in the forward direction change from 8.5 ± 2.7 kcal/mol at 0.5 Å/ns to an average of 9.9 ± 3.0 kcal/mol at 5 Å/ns, and from 5.7 ± 3.0 to 7.4 ± 3.6 in the reverse simulations. Theoretically, despite the change in simulation speed, the calculated free energies should be constant. However the estimation of ΔG_{QB} values range from 7.1 ± 2.1 kcal/mol at 0.5 Å/ns to 8.4 ± 2.3 kcal/mol at 0.5 Å/ns. This effect seems interaction dependent, which might be due to differences in the entropy generated during the process; either due to being more solvent accessible or to the flexibility of the flanking residues (Figure 6.31). The apo wildtype and I35G and N39G mutations share a trend where the ΔG_{QB} calculations deviate for I29-N542, N31-W591 and T30-H548 in a decreasing manner. On the other hand, the N31-W591 have a higher discrepancy than the other H-bonds for the wildtype and G40N and I37E mutations.

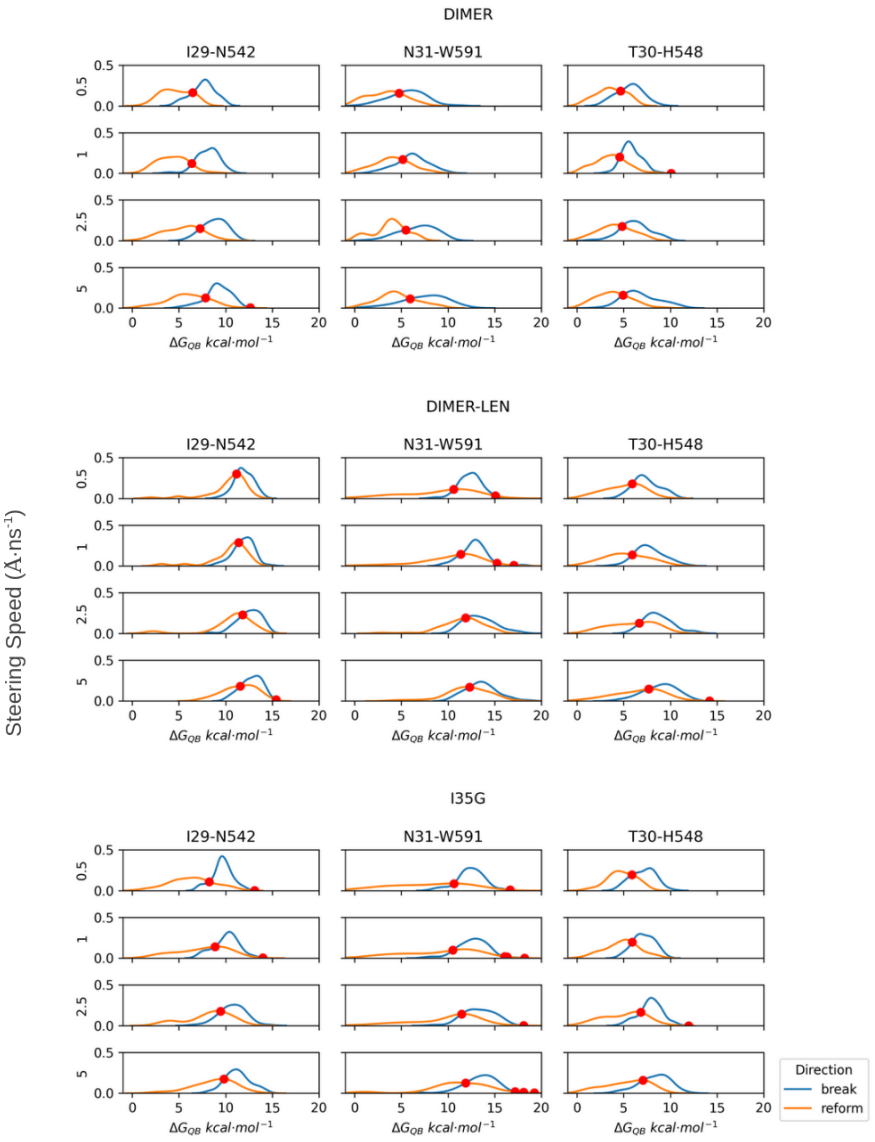


Figure 6.29. Gaussian smoothing of the work distributions for the forward and reverse H-bond breaking simulations of CRBN-CK1α complex. The labels dimer, dimer-len and I35G correspond to the wildtype complex without lenalidomide, with lenalidomide and the I35G CK1α mutation respectively. The x-axis labels correspond to the interactions steered. The work values with intersecting probabilities are highlighted in red.

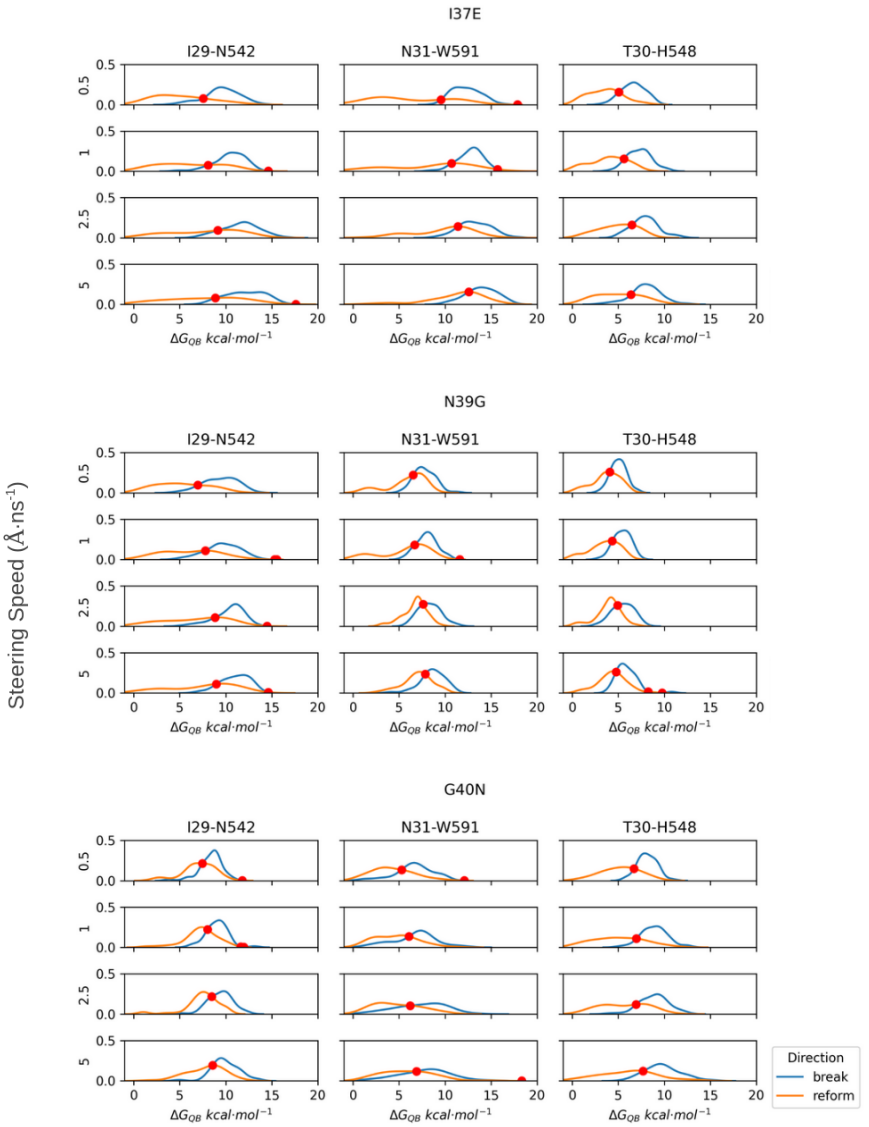


Figure 6.30. Gaussian smoothing of the work distributions for the forward and reverse H-bond breaking simulations of CRBN-CK1α mutated complexes. The labels I37E, M39G and G40N correspond to the CK1α mutations. The x-axis labels correspond to the interactions steered. The work values with intersecting probabilities are highlighted in red.

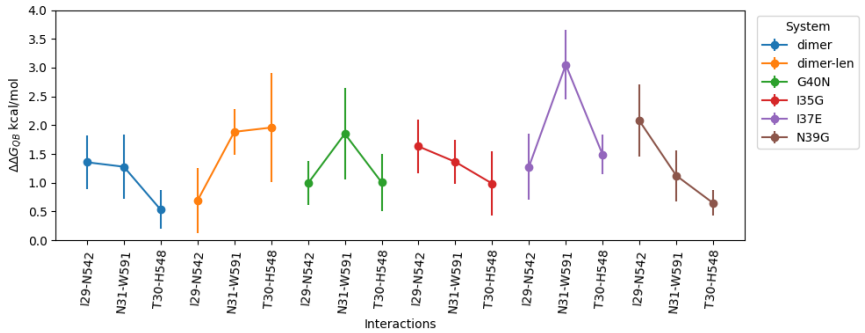


Figure 6.31. Differences in ΔG_{QB} between the slowest (0.5 Å/ns) and fastest (5 Å/ns) simulations per interaction per system.

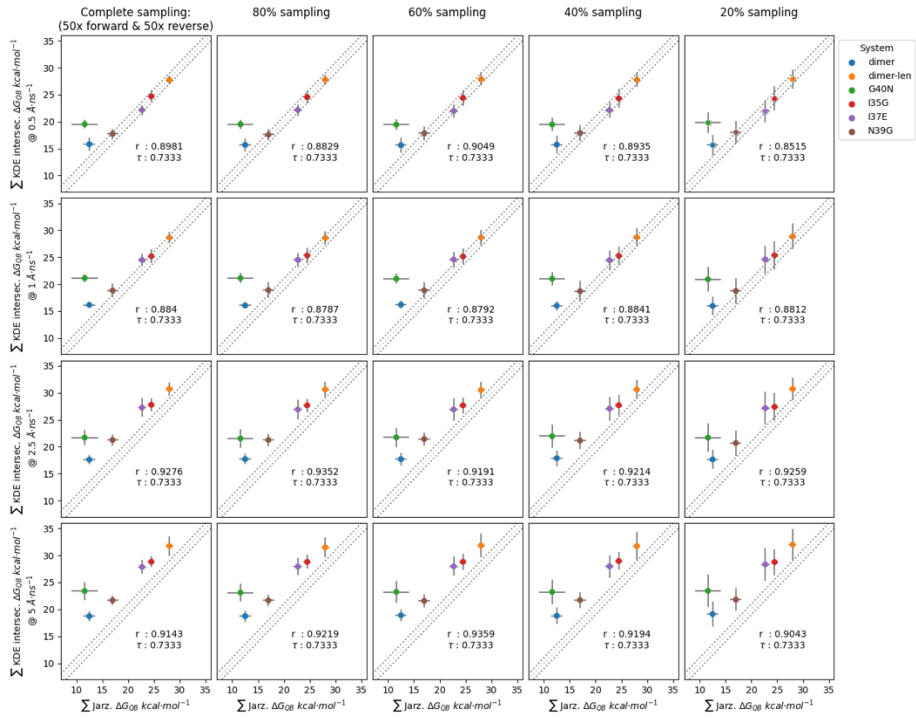


Figure 6.32. Comparison of H-bond ΔG_{QB} predictions using the JE and the CFT intersection in the CRBN-CK1α complex. The reported r are the Fisher corrected averages of the r for each comparison from the bootstrapping subsamples. The τ correspond to the Kendall test of the averaged predictions. The dotted lines mark the 1:1 relationship and the 1 kcal/mol error width band.

Following the approximation used in the original publication, I estimated the strength of the PPI by summing the ΔG_{QB} for the individual H-bonds. I then compared these values with the ones originally reported using the JE. Given the same sampling and speed conditions (Figure 6.32, top-left square) the correlation coefficient between both systems is 0.89, with a Kendall-τ of 0.73. The ternary complex involving the G40N mutant is an outlier on this correlation. Importantly, this system also showed the highest hysteresis in the original simulations and is the main outlier for all the comparisons presented in Figure 6.32. When reducing the sampling from the 50 replicas per direction to 10, there is an increase in the average standard deviation of the bootstrapped predictions from 1.26 to 2.61 kcal/mol. Nonetheless, the correlation coefficient is comparable and the ranking between the systems, reflected in the Kendall-τ, remains the same. Similarly, when increasing the pulling speed during the simulations, the ranking and correlations are also maintained. However, in agreement with the fluctuation theorem, pulling at higher speeds involves generating a higher dissipative work, as seen in the $\Delta\Delta G_{QB}$ comparisons from Figure 6.31.

The original publication by Miñarro et. al. showed that the summation of the three individual ΔG_{QB} values was a good predictor of the experimentally determined ternary complex stability. The high correlation of the ΔG_{QB} estimated with CFT with the ones estimated with JE is transferred to the comparison with experimental values, reaching an squared Pearson correlation of 0.79 when using 20% of the original sampling at ten times the steering speed (Figure 6.33). This procedure is 50 times faster than the one reported in the paper, which opens the door to future prospective applications in a HTVS manner.

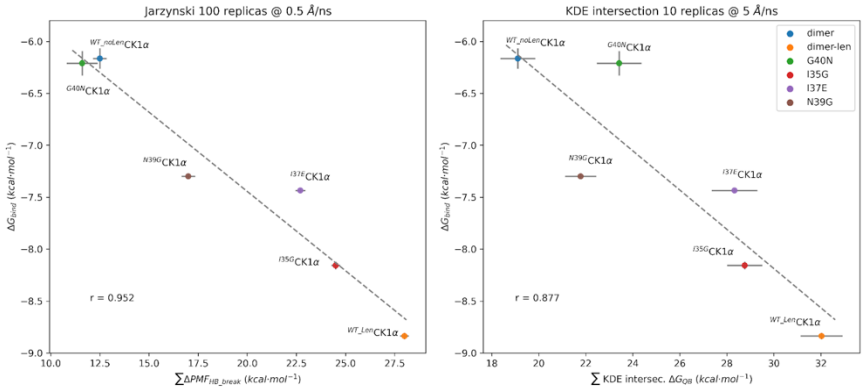


Figure 6.33. Correlation between the predicted ΔG_{QB} and the experimental ΔG_{bind} calculated from K_d estimations. The ΔG_{QB} are calculated as the sum of the interaction's ΔG_{QB}, using Jarzynski (left plot) and Crooks (right plot) equations. Left plot was adapted from [217].

6.4.2 ΔG_{QB} calculations of HSP90 protein-ligand complexes

Similarly to the ΔG_{QB} calculation of the CRBN-CK1α protein-protein interactions, CFT can be applied to protein-ligand complexes. Given the possibility of hastening convergence of accurate free energy calculations, it is specially relevant in the context of the development of quick SMD methods for drug discovery, such as OpenDuck. Hence, I applied a similar protocol to a set of 9 HSP90 ligands using the same interaction with D93 described in the section 6.2.1. Initial results showed that, during simulations of the dissociations at high speed, the HSP90 pocket collapsed due to the inability of waters to diffuse inside when the ligand exited the cavity, creating a vacuum effect. This created artifacts in the reverse simulations, as ligands would clash with the collapsed pocket yielding artificially high work values (i.e. in the range of 30-50 kcal/mol). The narrower geometry of the HSP90 cavity compared to the protein-protein interactions is most probably the reason this phenomena does not appear in the

CRBN-CK1α simulations. To circumvent this effect, I added a restraint to the heavy atoms during the dissociation and association trajectories to prevent the collapse of the pocket.

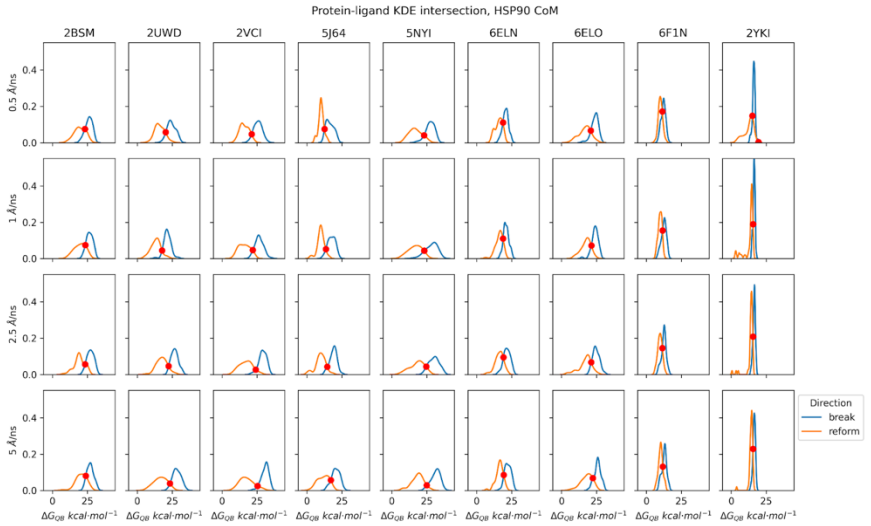


Figure 6.34. Gaussian smoothing of the work distributions for the forward and reverse H-bond breaking simulations of HSP90 protein-ligand complexes. The x-axis labels correspond to the PDB entry where the ligand structure comes from. The y-axis corresponds to the steering speed. The work values with intersecting probabilities are highlighted in red.

The work distributions for the forward and reverse trajectories are narrower than for the protein-protein interactions (Figure 6.34). This is also reflected in the proportion of single intersections seen for most complexes. The distributions overlap is also smaller, which reflects on the density associated with the intersection (i.e. the HSP90 intersections occur at a density of 0.09±0.04 compared to the 0.16±0.06 seen in CRBN-CK1a). There are two possible causes for the small overlap between the work distributions. The first possibility is a higher dissipated work during the forward trajectory, which might be a direct consequence of the solvation artifacts explained above. These effects will most likely be the limiting factor

when further scaling down the sampling for these simulations. The second reasoning is that the initial and final binding modes are different, indicating that the forward and reverse trajectories are exploring two different (albeit similar) processes. To discard this, we investigated the differences between the initial and final binding poses. Figure 6.35 shows that the RMSD between the starting and final binding poses are all are within what it is usually considered equivalent binding poses (i.e. < 2Å). Moreover, 7 out of the 9 show RMSD lower than 1Å. Only in the case of 2VCI we can see a second peak around 1.5Å, which corresponds to a rotation in a benzene group, and 6F1N having a wider distribution reaching further than 1Å.

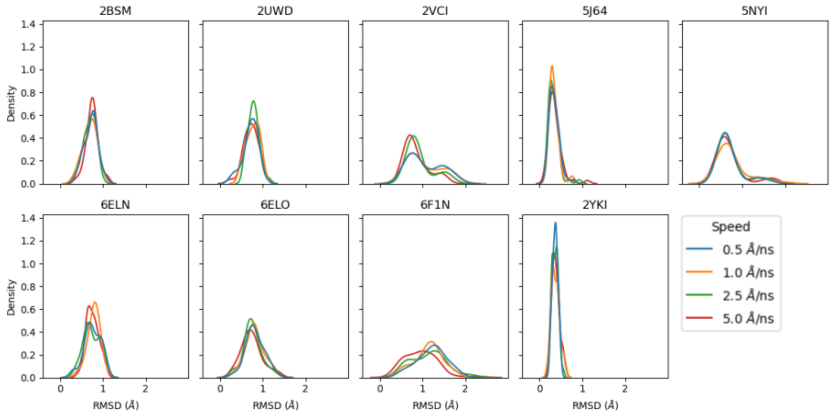


Figure 6.35. Distribution of HSP90 ligands' RMSD between the forward starting point and reverse endpoints per steering speed.

The comparison between the ΔG_{QB} predicted by the JE and CFT for the HSP90 protein-ligand complexes yielded slightly better results. Given the same sampling and speed conditions (Figure 6.36, top-left square) the correlation coefficient between both systems is 0.97, with a Kendall-τ of 0.89. Nonetheless, it follows the same trend seen

in CRBN-CK1α. Firsly, the dispersion of the bootstrapped intersection calculations increases with less sampling (from an average σ of 0.4 kcal/mol with 50 replicas to 0.95 kcal/mol at 10 replicas). Secondly, the overall magnitude of the calculated free energies increases when steering at higher speeds, from an average of 18.4 kcal/mol at 0.5 Å/ns to an average of 20.45 kcal/mol at 5.0 Å/ns, a similar increase as seen in Figure 6.31 and 6.32).

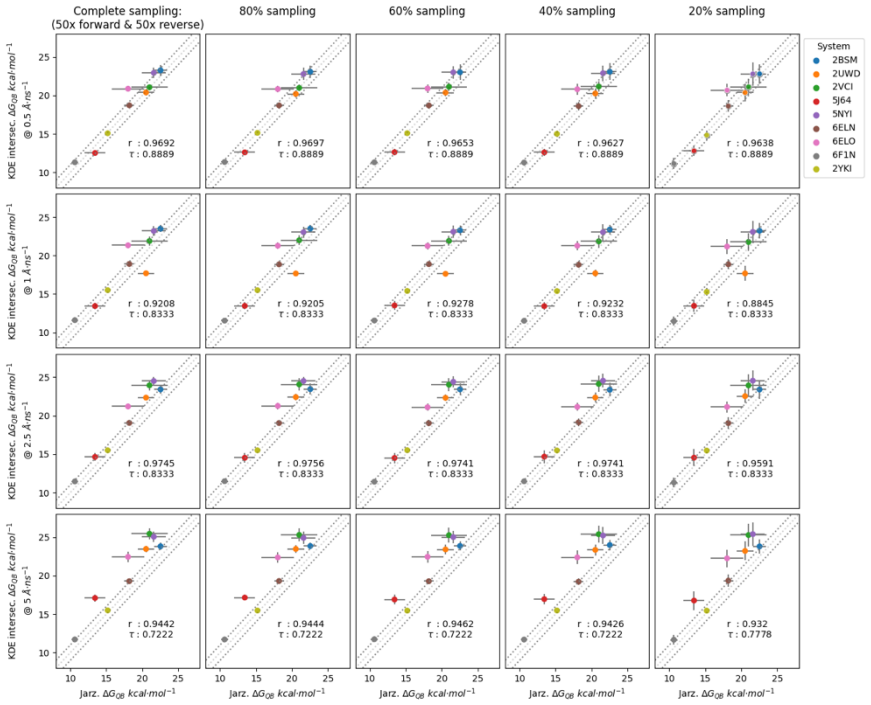


Figure 6.36. Comparison of H-bond ΔG_{QB} predictions using the JE and the CFT intersection in HSP90 protein-ligand complexes. The reported r are the Fisher corrected averages of the r for each comparison from the bootstrapping subsamples. The τ correspond to the Kendall estimator of the averaged predictions. The dotted lines mark the 1:1 relationship and the 1 kcal/mol error width band.

Interestingly, the standard deviations obtained with the CFT calculations are much smaller than with the JE with the same conditions, with the JE average σ being 1.27 kcal/mol with full

sampling and at 0.5Å/ns steering speed. The dispersion only equalizes when the CFT is 10 times faster with 20% of the sampling (Figure 6.36, bottom right square).

6.4.3 Evaluating the null hypothesis by speeding up Jarzynski calculations

Overall the free energy calculations using the CFT implementation hold their accuracy when reducing the sampling 5 folds and increasing the speed 10 folds. Moreover, in the case of HSP90 the CFT implementation shows a smaller dispersion than the JE through all conditions. However, while all the comparisons have been done with the JE using an over-sampled and slow-steering simulations, the JE calculations might also converge with the faster and smaller simulations sets. Thus I simulated and compared the ΔG_{QB} obtained with the JE under these conditions for the CRBN-CK1α (Figure 6.37) and the HSP90 (Figure 6.38) complexes. Interestingly, the ΔG_{QB} calculations using the JE maintain a very high correlation when speeding up and lowering the sampling, comparable to the CFT intersection. In the case of CRBN-CK1α, the G40N outlier appears again which indicates a shift in the dissociation pathways for this particular mutation, which was not seen in the original simulations. That outlier, together with the dispersion of I35G and I37E causes the ranking to vary more than with the CFT predictions (section 6.4.1). This becomes more evident in the case of the HSP90, where most ligands are concentrated in the 18-25 kcal/mol range. Surprisingly, despite the high dispersion of the HSP90's ΔG_{QB} calculation with the most 'conservative' conditions, the standard deviation does not increase significantly when altering the speed and

sampling, reaching an average sigma of 1.4 kcal/mol from the original 1.27 kcal/mol (Figure 6.38).

In summary, the alternative free energy calculations using the Crooks fluctuation theorem intersection show a high accuracy and performance even at steering high speeds and low sample size. However, the more established Jarzynski calculation shows comparable results. Despite the differences seen in the dispersion of the HSP90 simulations, I do not have sufficient data to assess which formulation actually converges faster to accurate results. In this direction, pushing the simulations further with higher speeds and lower sampling sizes will be evaluated in the future.

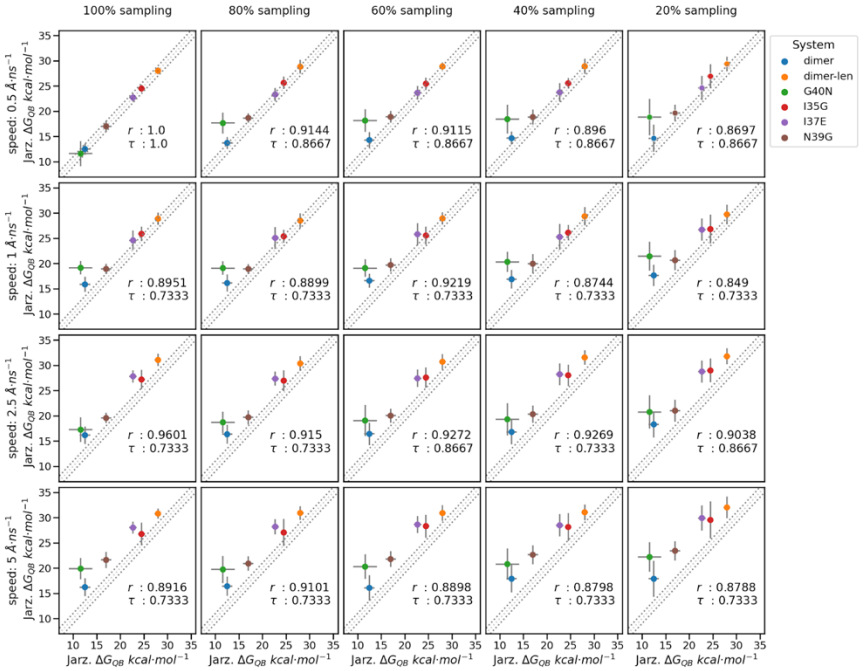


Figure 6.37. Comparison of H-bond ΔG_{QB} calculations using the JE at different speeds and sample sizes on the CRBN-CK1α complexes. The reported r are the Fisher corrected averages of the r for each comparison from the bootstrapping subsamples. The τ correspond to the Kendall estimator of the averaged predictions. The dotted lines mark the 1:1 relationship and the 1 kcal/mol error margin.

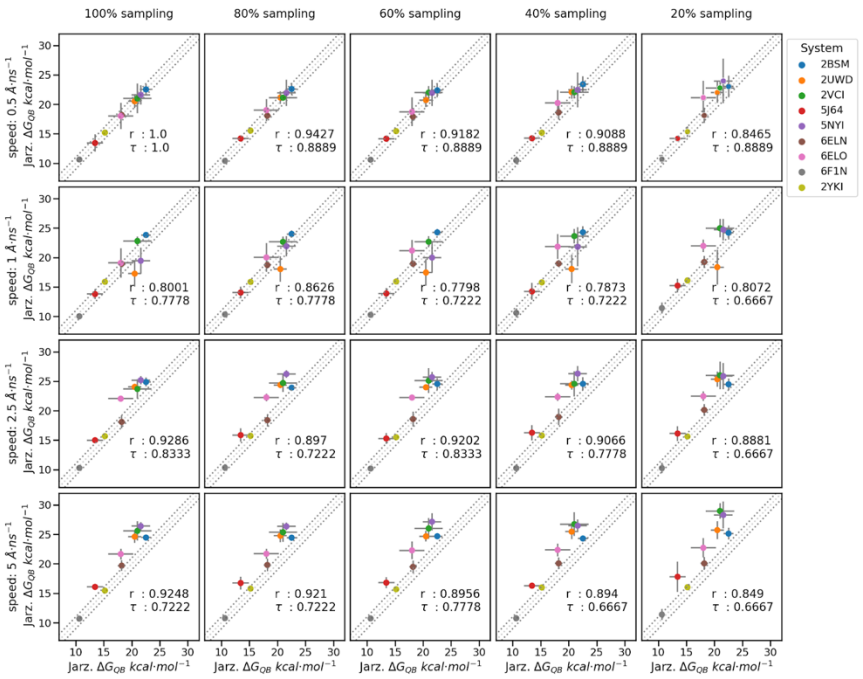


Figure 6.38. Comparison of H-bond ΔG_{QB} calculations using the JE at different speeds and sample sizes on the HSP90 protein-ligand complexes. The reported r are the Fisher corrected averages of the r for each comparison from the bootstrapping subsamples. The τ correspond to the Kendall estimator of the averaged predictions. The dotted lines mark the 1:1 relationship and the 1 kcal/mol error margin.

CHAPTER 7: DISCUSSION

7. DISCUSSION

In the last decades, research in the early stages of drug discovery has seen huge advances, opening the field to new challenges and opportunities. Specifically, the emergence of new and higher-throughput experimental methodologies to explore biological systems demands tools capable of exploiting the new and large volume of data generated, to create useful knowledge that can help solve unmet medical needs. In that regard, computational methods have become an essential component of drug discovery, from target identification and validation, to hit discovery and lead development. Indeed, the main motivation of this thesis has been the development and application of structure-based computational tools and processes to improve these stages. Firstly, in Chapter 4 I identified promising ligandable pockets in E3 ligases previously labeled as undruggable, and then aided disentangling the mechanism of action (MOA) of a first-in-class FBW7 modulator. In Chapter 5 we proposed a computational pipeline to efficiently navigate ultra-large chemical libraries to find more and better hits during virtual screening. Finally, Chapter 6 describes the efforts in improving and expanding the applicability of steering molecular dynamics methods to evaluate the structural stability of molecular complexes and its application to guide ligand optimisation. In the following pages I'll discuss in detail the results, their implications and limitations of each chapter.

The main objective of Chapter 4 was to evaluate the ligandability of a subset of E3 ligases, and find new pockets that could bind small molecules, thus expanding the available toolbox for targeted protein degradation (TPD). E3 ligases are a key component of the UPS

machinery that recognize the substrates for degradation. Despite the surge in popularity and efforts in the TPD field, few E3 ligases have been targeted and are currently used for PROTACs or molecular glue applications. To help expand the available E3 ligases, we implemented a pipeline using molecular dynamics with mixtures with organic solvents and water (MDMix) to identify regions favorable to interact with organic molecules. Then clustering and evaluating these regions to ensure drug-likeness of the predicted pockets. There are many tools that aim to predict ligandable pockets reported in the literature.[295–298] However, most represent the receptors as static entities, or cover very little range of conformational variability. Notably, empirical druggability methods are trained on datasets that contain relatively few allosteric pockets.[299,300] Thus it is not clear if they are capable of detecting such pockets. Physics-based methods, on the other hand, have the advantage of a general applicability at the expense of higher cost.[301] In particular, molecular simulations with organic solvents have been reported to facilitate the opening of cryptic pockets and ligand-induced conformational changes.[80,302,303]

The limitations of static pocket detection tools show up in a previously reported E3 pocketome database (ELIOT).[304] The ELIOT pocket predictions are based on FLAP,[305] which uses the comparison of GRID-based molecular interaction fingerprints to known E3 pockets and adds information mined from the literature to predict its *PROTACability*. Whilst covering a large spectrum of the E3 space, they were limited by the crystallographic structures and the similarity with known pockets. Still, static approximations offer good opportunities as a wide exploration to then perform a deeper analysis on interesting targets.[306] On the other hand, we chose to build a smaller selection of E3 and propose a higher curation of the

results with a high conformational depth in the analysis. However, we also employed X-ray structures, which limited the sequence coverage of our input structures and the availability of relevant functional domains and adaptor proteins. Nonetheless, adapting the pipeline to using alphafold predicted structures and complexes should be trivial and not have a negative impact on the predictions.[307] Naturally, the method remains limited to well-structured proteins, and is not applicable to protein regions with abundance of disordered sequences.

From the 23 E3 ligases evaluated, 22 presented at least one potential ligandable pocket, with an average of 2.3 pockets per E3 ligase. This highlights the untapped potential of the yet-unexplored E3 ligases to design novel small molecule binders. Firstly, we validated retrospectively the accuracy of the predictions by inspecting pockets with known orthosteric binders. We recovered as ligandable 5 out of the 8 orthosteric ligand binding sites. The pipeline discarded the SOCS2 covalent ligand binding site due to the flat surface and GID4 pocket due to its conformational instability during the simulations, exceeding the volume limit specified. Flat binding surfaces expose the ligands to solvation, which is generally undesirable as unbinding becomes easier, but covalent binders solve this by not allowing unbinding. Thus, pockets for covalent binders fall outside the scope of the method. Additionally, not all other untargeted degron recognition sites were identified as ligandable. This emphasizes the importance of the physics-based approach adopted here, as it does not rely solely on the protein's surface geometry, but also on its composition, flexibility and desolvation properties.This is particularly relevant for protein-protein interfaces and allosteric binding sites, which are scarcely

represented in the datasets of druggable pockets. One of the main limitations of the pipeline is the nature of the organic solvents we have used. The pipeline did not recover the recently reported DCAF1 binder[237] interaction patterns in the degron recognition site, as there is a big change induced by the ligand. Despite the ability of organic solvents to induce conformational changes similar to those modulated by a ligand,[308] their efficacy depends on the solvent used, larger pockets requiring bulkier and more hydrophobic solvents.[165] In our case, we limited the analysis to the ethanol solvent, but the DCAF1 ligand seems to rely on pi-pi interactions to modulate a loop in the pocket. Another relevant example of pocket induced by a ligand appeared in VHL, whose cryptic pocket was opened by the organic solvents and recovered as ligandable in the pipeline. In this case, the pocket was smaller and the ethanol simulations were sufficient to open and detect it.

In addition to known binding sites, 66% of the predicted pockets are novel and correspond to non-orthosteric sites. In the context of TPD, non-orthosteric binders, even those without allosteric activity, are still relevant to modulate the function of E3 ligases or even substrates to degrade. Moreover, these non-orthosteric binding sites offer unique opportunities since (a) they do not compete with the natural substrate and are less likely to disrupt the natural function of the E3 ligase, (b) can be very selective by targeting non-conserved regions and (c) can avoid resistant-inducing mutations. These binding sites are key to increase the number of available E3 for TPD applications, especially those with unknown or undruggable degron recognition sites.

Finally, we detected non-orthosteric ligandable pockets with clear indications of being able to produce allosteric modulation of the E3 ligase (i.e. due to previously reported motions or their close location to an adaptor or substrate binding domain). In particular, one of the studied E3 ligases, FBW7, yielded a pocket situated in a hinge region between the two functional domains (i.e. a WD40, where the degron recognition site is located, and a F-BOX domain, where FBW7 interacts with the adaptor to the rest of the UPS domain). Moreover, the modulation of this region had been proposed previously to enhance substrate's lysine presentation to the E2 and thus ubiquitination and degradation.[250] Nevertheless, FBW7 was deemed undruggable and it had no previously known binders.

After a virtual screening (VS) campaign, first-in-class FBW7 binders targeting the predicted pocket were discovered. We detected that a specific chemical family of those binders induced an increased degradation of c-Myc, a natural FBW7 substrate. Out of the family, the compound A5 showed the best binding affinity and degradation potency. Andrea Bertran-Mostazo and I set out to investigate the MOA of the A5 family through experimental and computational means respectively. I performed long MD simulations of the SCF^{FBW7} (FBW7's UPS complex) to assess if A5 was modulating the complex in a similar manner as proposed by Liu and Nussinov. Normal mode analysis of these simulations showed that compound A5 caused the opposite effect than expected. During the simulations with the holo structure the WD40 domain (corresponding to the substrate presentation unit) was farther to the E2 than in the apo simulations. Hypothesizing that, despite having a more open conformation, the ligand hindered the movement of the hinge, I performed a markov state analysis of the simulations which was not conclusive.

Furthermore, a recent analysis has shown that the rigidification of the ternary complex does not necessarily correlate with an increase in protein degradation.[309] Overall, these analyses showed that A5 alters the conformational space of the SCF^{FBW7} complex, however it is still unknown how that translates into the specific increased degradation of c-Myc.

In parallel, Andrea performed ubiquitination assays which showed that the A5 compound family does not increase the levels of ubiquitinated c-Myc (neither when blocking the proteasome). Yet, the modulation of c-Myc degradation showed FBW7 and proteasome dependence (i.e. recovering normal c-Myc levels when using a FBW7 knockout cell line and a proteasome inhibitor respectively). These findings suggest that A5's MOA might rely on pathways alternative to the direct ubiquitination of c-Myc. We currently envision several hypotheses. First, it has been described that FBW7 dimerization determines its substrate specificity and the degradation.[310] Especifically, c-Myc depends on the dual recognition of its degrons. Despite the dimerization domain being sequentially distant from A5's binding pocket, the conformational changes might shift the dimer/monomer equilibrium in FBW7 populations, thus increasing c-Myc degradation. Alternatively, it has been shown that lysosomal degradation plays an important role in protein degradation, independently from ubiquitination, and that it gets reinforced when the UPS pathways are blocked.[311–313] Finally, A5 could be modifying the substrate recognition patterns in a non-degron dependent manner, indirectly reducing c-Myc levels. This phenomena has been extensively described in CRBN where different IMIDs induce the degradation of a specific repertoire of neo-substrates through a non-degron recognition domain (i.e. the

CULT).[314] Similarly A5 could be inducing the degradation of a neo-substrate involved in the c-Myc stability (e.g. AURKA,[315] NEMO[316] or BTRC[317]). Further experiments will be needed to disentangle the MOA of the A5 family of compounds.

In spite of A5's unclear MOA, I performed a *SAR-by-catalog* to determine the key groups in the family and optimize their potency. While maintaining the scaffold that showed the most stability during the MD simulations, I evaluated changes in the linker and terminal region of the ligand. From the eleven compounds bought to Enamine, only AMB8, with a sulfonamide linker, yielded a K_d comparable to A5 (i.e. 2-5 μM). Finding more potent FBW7 binders within the family would ease the experiments to understand the family's MOA, by increasing the signal, reducing toxicity and possible off-targets, and even allowing success in crystallization.

The difficulty in finding very potent ligands is a common limitation of VS. In Chapter 5 we propose a novel approach to VS that can potentially yield more potent hits by providing efficient access to ultra large chemical libraries, even using higher-quality computational methods. Ultra-large compound libraries are virtual collections of compounds, usually stored as graphs of building blocks and reactions, that can be synthesized *on-demand*. These libraries have grown exponentially in recent years, reaching hundreds of billions of compounds, while maintaining a high success ratio in their synthesis.

The promising potential of the ultra-large libraries for drug discovery is the possibility of finding more and better hits, allowing for better starting points for hit-to-lead.[125] This claim has been heavily debated, as while bigger libraries have higher chances of containing a very potent ligand, they also contain significantly more non binders and

ligands with low potency.[318] Moreover, a balance between the depth and diversity of the chemotypes present in the databases is key to ensure their success on new targets.[319] At the moment, there are several *on-demand* ultra-large libraries (REAL, CHEMriya, GalaXi, eXplore, AMBrosia) with small overlap between them,[124] all having good representation of certain chemotypes, but Enamine's REAL space has a significantly higher diversity than the others,[320] and Enamine has a proven track record in delivering a high proportion of the requested compounds. Another significant concern around the ultra-large libraries is how many drug-like compounds populate them.[321]

Navigating through libraries of such magnitudes also presents specific challenges. Firstly the cost of enumerating, storing and screening the libraries is outside the reach of most academic groups, and it is only increasing. As an example, the enumeration of GDB-17 (circa. 10^{11} molecules with less than 17 heavy atoms) took more than 20000 cpu-hours and occupies ~400GB disk space in zipped format.[322] Thus, the community has adopted alternative library architectures that do not require enumeration, such as the fragment spaces. Traditionally, search and evaluation algorithms would rely on enumerated libraries, with singular exceptions like FTrees.[257] Hence several strategies have appeared to either alleviate or eliminate the need for enumeration. The group of Prof. Matthias Rarey has spearheaded the research regarding similarity,[144] substructure,[145] shape[323] and physico-chemical property[324] searches in compound libraries using the fragment space, graph-based, architecture. Furthermore, many strategies have been proposed regarding the evaluation of compounds from such libraries for virtual screening. The main three approaches in the literature are (i) brute-

force docking, enabled by the use of HPC and cloud computing,[252–254,258] (ii) active learning of a library subset to predict iteratively the scores for the rest of the library,[259–262] and (iii) limiting the exploration to specific building-blocks or fragments within the libraries.[266–268]

As a variation of the third strategy, we developed the bottom-up exploration of the chemical space and validated it by finding hits for the BRD4-BD1 system. We divided the pipeline into two main independent modules, the fragment screening and the scaffold growing. The first step involves an exhaustive search of promising scaffolds in the low molecular weight space, whose extension is limited even when considering the theoretical chemical space. Specifically, there are only approximately 10^9 possible fragments under fourteen heavy atoms,[269] which would be already possible to exhaustively evaluate. Noteworthy, the fragment space is exhaustive and independent of the composition of the on-demand chemical collections. Thus it only needs to be explored once and does not require previous knowledge about the building blocks or chemical reactions used to generate each on-demand chemical collection. On the other hand, the evaluation and binding pose generation of such fragments is crucial to the pipeline, as the scaffold growing stage heavily depends on it. However, in-silico evaluation of fragment hits is significantly harder than with larger molecules due to their high promiscuity and overall lower potency than drug-like molecules.[325–327] We tackled this issue by layering computational methods and employing a reference set of known BRD4 binders for comparison. Nonetheless, to evaluate the scaffold growing independently from the fragment screening, we combined the selection of fragment hits with scaffolds from experimentally validated fragment hits and drug candidates.

The second stage of the bottom-up pipeline involves the expansion of the selected scaffolds by creating scaffold-focused libraries extracted from the ultra-large libraries. By leveraging the capabilities of SpaceMACS[145] to create complex substructure queries using SMARTS patterns, we were able to tailor the specific subspace of the library to evaluate; guiding the growing vectors, bioisosteres, specific key functional groups or even isomery, through the scaffold definition. The high control on the chemical subspace selection differentiates the bottom-up from other synthon or fragment-based methods like V-SYNTHES.[266] Typically these methods are limited by the predefined building blocks within the library to stepwise construct the larger drug-like molecules. By comparison, the scaffold patterns we designed match multiple building blocks at the same time; either through two partial MCS matches or due to multiple building blocks matching the SMARTS generalization. After the enumeration of the scaffold-focused libraries, we selected the compounds with drug-like properties, using Lipinski's rule of five, rotatable bonds, and reactive and PAINS groups filters. The potential usefulness of stricter filters at this step became noticeable when the experimental validation started. In particular, solubility has been a significant limitation during the biophysical experiments. Subsequently, we have currently incorporated LogS predictions (using the ESOL algorithm[328]) and restrictions to the amount of chirality centers and tautomers allowed.

I performed tethered docking with rDock as the pose generation tool for the 240 million prepared conformers from the 11 scaffold-focused libraries (i.e. 5 from virtual fragment hits, 3 from experimental fragment hits and 3 from scaffolds derived from drug candidates). Tethering the common scaffold of the library helps us significantly reduce the degrees of freedom, and thus the computational cost,

during the calculation. However, this relies heavily on the premise of the scaffold's binding mode not changing. This working hypothesis is very common in SBDD, for instance, when performing RBFE calculations,[329,330] but given the magnitude of the changes between the fragment scaffold and the ligands, there is a chance that it might not be fulfilled. Thankfully, the crystallography experiments confirmed that at least two of the three hits crystallized had exactly the predicted scaffold's binding mode (with an RMSD < 0.2 Å). The case where we observe a change in binding mode corresponds to a drug scaffold where the reference binding mode was experimentally determined. Interestingly, the binding mode it adopted matches another experimentally determined fragment we did not employ in the validation (with PDB id 5LVQ). Finally, a significant challenge when evaluating large chemical spaces for drug discovery purposes is the strain on scoring algorithms that need to quickly select the top compounds. All the scoring schemes rely on enriching the selection of binders with respect to non-binders. However, the acceptable enrichment factor needs to be stricter as the amount of compounds to evaluate increases. Moreover, docking scoring has been described to be more proficient in ranking diverse chemotypes rather than differentiating closely related molecules,[270] as we would need for the scaffold-focused libraries we created. In the last decades, an abundance of strategies have been developed involving MD[176,331,332] and ML[333–335] to apply as post-docking filters. In the bottom-up pipeline we employed two orthogonal MD methodologies, MM/GBSA and DUck, which have previously shown good results when combined.[201]

Despite the high throughput of the post-docking MD methodologies we used, it was outside our capabilities to evaluate the *circa.* 13 million compounds yielded by docking. Thus, we performed a diversity selection and clustering of the molecules, reducing the libraries to 1000 cluster representatives each (i.e. a total of 11000 molecules). Discarding molecules exclusively due to molecular similarity entails missing on potential good binders in the library, reducing the chemotype depth provided by using the ultra-large libraries. Nonetheless, this creates the opportunity to recover the clusters from the best hits for later optimization, using more expensive but accurate RBFE such as FEP. In that regard, there are published examples of selection and optimization of perturbation networks from large pools of molecules,[90] and the *a priori* estimation of the confidence for the related calculations.[89,336,337]

Overall, we created 11 scaffold focused libraries from Enamine's REAL space ultra-large library, starting from three different scenarios: (i) a fully computational pipeline, (ii) starting from a fragment of known binding mode, and (iii) starting from optimal scaffolds that are capable of yielding drug candidates. Then docked the 230 million compound conformers from those libraries, clustered the results and evaluated a diverse set of 1000 representatives with MM/GBSA and DUck. The accumulated computational cost of the full procedure was approximately $3 \cdot 10^4$ cpu hours, using the BSC and IQTC HPC facilities. By comparison, using an average of 10 seconds per molecule in docking, evaluating the equivalent library through brute-force docking would take the order of $3 \cdot 10^8$ cpu hours without considering the library enumeration. As such, by constraining the space to evaluate to promising scaffolds, we have achieved a theoretical ten thousand fold reduction in the

computational time, even when evaluating the ligands through more sophisticated means than docking.

Ultimately, Andrea Bertran-Mostazo experimentally tested the compounds proposed through the bottom-up VS pipeline; first through a single dose screening with DSF and SPR and then characterized their potency (IC_{50}) through TR-FRET. 29.4% of the compounds were determined as hits using both screening techniques, being the hit rate significantly higher for the experimental and virtual fragment scenarios than for the drug-derived scaffolds. Furthermore, 16 of those compounds yielded IC_{50} between 1 and 100 nM, comparable to the drug candidate (+)-JQ1. Despite the orthogonal nature of the techniques, there was a good accordance in the best hits and most potent molecules. The high hit-rates and potency of the compounds obtained support the claims associated with ultra-large virtual screening. Moreover, by using diverse initial scaffolds to study we ensured a high chemical diversity in the resulting hits, reporting new chemotypes dissimilar from the known BRD4 binders.

Through the exploration of ultra-large libraries from the bottom up, we have been able to successfully exploit the opportunities they present, obtaining more diverse and more potent hits in a highly efficient manner. Still, the reliance on clustering due to the limited throughput of more accurate methodologies emphasize the need to improve our post-docking toolset.

Thus, in Chapter 6 I explored the use of SMD simulations for evaluating the structural stability of protein-ligand and protein-protein complexes. In particular I used DUck to obtain RBFE estimations and to predict activity cliffs in congeneric series of compounds.

Additionally I developed OpenDUck, an open source implementation of the method to facilitate further development of SMD applications. In this context, I also evaluated the use of the Crooks fluctuation theorem, an alternative formalism to calculate free energies from non-equilibrium work, with the objective of obtaining converged free energy estimations faster than with the Jarzynski equality.

The use of structural stability, measured as the work needed to bring a H-bond to the quasi-bound state (W_{QB}) with DUck, to perform qualitative classifications between *robust* and *labile* binders has been shown to yield significant enrichment as a post-docking filter.[176,178,201,284] However, a quantitative use of this magnitude remained in question. Previous work showed that W_{QB}, as a local property, does not directly correlate with ΔG_{bind}, which is a global property of the systems.[176] Still, in a context of congeneric compounds it could spotlight and rank favorable interactions due to error canceling effects. In addition, prediction of AC is one of the most important tasks in hit-to-lead, and does not require an exact solution, but correct classification. ACs underline non-additive effects whose impact might be better captured in local measurements such as W_{QB}. Hence, I investigated the use of quasi-bound free energy (ΔG_{QB}) to evaluate congeneric series of protein-ligand complexes and to predict the AC in them.

Firstly, I evaluated a dataset of HSP90 inhibitors, which had reported kinetic rates and activities. Following our hypothesis, we obtained a better correlation between $\Delta\Delta G_{QB}$ and $\Delta\Delta G_{Bind}$, than what was previously reported between W_{QB} and ΔG_{Bind}. Whereas I could also recall most AC, the predictions improved when comparing with kinetic off-rates rather than with binding affinities. Considering that

kinetic on-rates are relatively constant within chemical series,[338–340] the improvement in MCC when employing off-rates is a reflection of the imperfect correlation between k_{off} and K_d. In turn, these results suggest that DUck could be well suited to guide kinetic optimization of ligands, which has been postulated as a promising route towards improving in vivo efficacy.[341,342] Furthermore, they also indicate that DUck could be a useful tool to better understand AC.

Interestingly, τRAMD simulations of the ligands showed that most ligands followed a one-step dissociation pathway, but the minority that followed a complex dissociation pathway were present in all the false positive predictions. To evaluate this phenomena, I extended the analysis to 4 distinct congeneric series of CDK2 inhibitors. The simulations on selected serie's representatives showed the same behaviour; ΔG_{QB} calculations of ligands with complex dissociation states yielded worse predictions. Moreover, a one-step dissociation was not a sufficient condition to ensure predictability, but a high overall ΔG_{QB} was also needed as seen in series CDK2-D. Only ligands in series CDK2-B and D showed a one-step dissociation but the only series where DUck yielded good predictability was the serie CDK2-B due to serie CDK2-D low work profiles.

Considering the results seen for HSP90 and CDK2, we proposed two conditions to assess *a priori* the applicability of DUck to calculate $\Delta\Delta G_{Bind}$ and the prediction of AC. First, identifying a robust interaction within the protein-ligand complexes, with high resulting ΔG_{QB}. While this property is very system dependent, we postulate that only those complexes with low dispersion of the work profiles and a minimum W_{QB} of 8 kcal/mol fall within the scope of the method. Secondly, a dissociation analysis of a representative compound

should be performed to evaluate the unbinding mechanism and confirm a one-step dissociation mechanism. In this work we employed τRAMD, but alternative protocols such as metadynamics are equally valid. If both conditions are fulfilled, we expect high confidence in the AC predictions using DUck. Finally we confirmed the protocol using BACE1 protein–ligand complexes. After ensuring a one-step dissociation with a representative ligand and high ΔG_{QB} profile with low dispersion, we proceeded to evaluate all with DUck. Followingly, we obtained good results, comparable to the HSP90 and CDK2-B datasets.

The aforedescribed results show the promising applicability DUck for quick RBFE calculations with a throughput adequate for post-docking steps. This adds to the increasing pool of applications that range from binding mode predictions[177] to ternary complex formation in the context of molecular glues.[217] Thus I developed OpenDUck, a python-based open source implementation of DUck. OpenDUck aims to facilitate further developments and applications of the structural stability concept without the dependence on licensed programs. OpenDUck was built as a python library with functions to assist the implementation of new applications, leveraging the effort from other open source initiatives in the computational chemistry community such as OpenForcefields and OpenMM. In addition, it includes a command-line executable to run the DUck protocol with further optimization parameters such as the small molecule force field, using HMR or the solvation model.

Using Iridium and Seraphic, two datasets previously analyzed with DUck, I benchmarked the performance of OpenDUck to DUck, and also a variety of parameters available to change in OpenDUck. While

OpenDUck allows for the execution of simulations in Amber, just like DUck, the implementation of SMD in OpenMM needed a specific adaptation to mimic DUck's results. Finally, OpenDUck calculations always yielded correlation coefficients around 0.9 and MAE in the 2 kcal/mol range. Moreover these correlations were also maintained when changing the small molecule force field, box size, water models, solvation ionic strength, and applying HMR. Significantly, a smaller box size and the use of HMR enabled a reduction of approximately 70% of the computational time, almost tripling the simulation speed without a change in the ΔG_{QB} distributions.

The convergence analysis spotlights the high impact of independent equilibrations to the reproducibility of the results. The correlation coefficients observed when using different software or simulation conditions are on par with those obtained when running independent replicas. However, we observe a surprising importance of the equilibration step in the reproducibility of the results, with Pearson correlation of 0.996 for trajectories derived from the same equilibrated states compared to the circa. 0.9 obtained for all the other independent equilibrations. Two main possibilities arise from this observation, changes in binding mode during the heating and equilibration or a high dependence of the work from the initial velocities. Conformational changes during the equilibrium sampling steps in the procedure usually translate in big shifts in the dissociation work profiles, changing the location of the minima. While I did not observe such shifts when evaluating the benchmark datasets, it is possible that they have a larger impact when evaluating docking poses rather than structures obtained from crystallization. Alternatively, the high impact on the initial velocities raises a concern on the reliability of the predictions with the DUck

pipeline as it is. Nonetheless, it could be easily solved by reassigning velocities from a Bolzmann distribution at the simulation temperature before each SMD and reevaluating the convergence under those conditions. This would, however, significantly change the original DUck pipeline and was outside the scope of this investigation. The effect due to independent equilibrations is reflected on the correlation of all the comparisons made previously, which puts in question the real underlying impact of the change in parameters.

To highlight the ease of creating new applications within the DUck framework using the OpenDUck library I created a protocol to measure the structural stability of water-mediated H-bonds. I evaluated three protein-ligand complexes from the benchmark datasets where, despite their high potency, no robust protein-ligand H-bond could be found, but water-mediated interactions existed. The change from H-bonds to water-mediated H-bonds may seem trivial, but involves changing the definition of the collective variable and the restraining schema. With that, we recovered three very different work profiles that correctly ranked them to their reported activities. Moreover, other similar modifications are in development in the group and have been facilitated by OpenDUck, such as steering ligands in a lipidic environment (i.e. inside a membrane), steering ligands coordinated with metal ions, or evaluating the stability of protein-protein interactions.

Finally, I evaluated the possibility to further improve the methodologies following the DUck framework, by employing the Crooks fluctuation theorem (CFT) as an alternative to the Jarzynski equality (JE). Theoretically, by also simulating the reversed process (i.e. reforming the H-bonds) where the entropy gain is reduced,

pathways with low dissipated work should be more probable, thus converging faster to accurate free energy values. To evaluate the benefits of using CFT, I compared it to PMF calculations on the three H-bonds from CRBN-CK1α performed by Miñarro et al. [217] and the ΔG_{QB} of 9 HSP90 protein-ligand complexes calculated with the JE.

I evaluated the high-throughput capabilities of both approximations by increasing the steering speed (from 5 to 0.5 Å/ns) and employing less sampling (from 100 replicas to 20). Through these simulations the estimations of ΔG_{QB} were very stable for both the CFT and the JE. By increasing the steering speed, the resulting work increased. This should be corrected in the estimation to free energy, but due to insufficient sampling was not completely eliminated when using the CFT nor the JE. Interestingly, while the upward shift in free energy was comparable between CTF and JE for the CRBN-CK1α simulations, the CFT performed better in the HSP90. This might be a direct result of the overlap between the forward and reverse simulations in that system, or by the low hysteresis of the reverse simulations. The requirement of a significant overlap between the two work distributions is a critical limitation of the CFT.[343] This limitation accentuates when the sampling is restricted and the distributions are wide, as the probability of encountering events at the work intersection is low. The multiple intersections due to widened reverse work distributions in the CRBN-CK1α simulations reflected in the high variance of the predictions larger than in the HSP90 protein-ligand complexes, specially when lowering the sampling pool.

I was not able to observe any significant improvement in the convergence of the free energy estimations when using the CFT

compared to the JE. Nonetheless, the correlation to experimental values was maintained under the fastest conditions, resulting in a 50 fold increase in speed compared to the protocol employed by Miñarro et al. Further analysis of the performance of the different formalisms under more strain conditions (i.e. faster steering and lower sampling) should reveal the limits of the methodology. Moreover, a protocol to estimate *a priori* the fastest convergeable conditions to specific systems will be highly relevant to expand the applicability of these protocols in a high-throughput manner.

Overall, I have shown the applicability of the robustness of protein-ligand complexes to RBFE predictions. Moreover, I have developed OpenDUck as a tool to facilitate the research of new applications of the framework started with DUck. Finally, I explored in detail an example of such new applications by employing an alternative formalism to estimate free energies. With these, I have aimed to solidify the position of structural stability as a key system descriptor for drug design.

CHAPTER 8: CONCLUSIONS

8. CONCLUSIONS

Computational methodologies have proven crucial to address current challenges in early drug discovery such as expanding the druggable space, supporting the development and understanding of new drug modalities, and providing tools to improve hit identification and hit-to-lead stages.

In particular I have performed a computational assessment of the ligandability of 23 E3 ligases using mixed organic/aqueous solvent MD. I found an average of 2.3 ligandable pockets per E3 ligase, 66% of which were novel and corresponded to non-orthosteric regions. A retrospective analysis showed that the proposed pockets recover most of the binding pockets with ligands previously described. A prospective validation resulted in the discovery of A5, a first-in-class FBW7 allosteric modulator that decreased the levels of c-Myc in cells through an unknown mechanism of action. A conformational analysis of the effect of A5 to the SCF^{FBW7} complex discarded previously proposed mechanisms of action. Finally, I generated a set of A5 analogs through a structure-based SAR, but did not find any that increased the affinity to FBW7 nor the degradation of c-Myc.

I have developed a pipeline to efficiently explore the ultra-large on-demand chemical libraries from the bottom-up for hit identification. The pipeline consists on an exhaustive search of promising scaffolds in the low molecular weight space and to grow those scaffolds into drug-like molecules. We validated the pipeline in the BRD4-BD1 system using eleven scaffolds from three different scenarios. The scaffold growing was performed through tethered docking and evaluated through a hierarchy of computational techniques. The

computational resources needed to do so accounted for less than a ten thousand part of what would be needed through brute-force docking. 85 compounds where successfully synthesized and evaluated experimentally through a single dose screening with DSF and SPR, resulting in a combined hit rate of 29.4%. The compounds' binding affinities were assessed through TR-FRET, finding 16 comparable to (+)-JQ1. The binding mode employed for the tethered docking was corroborated through X-ray crystallography in two out of three resolved structures.

I have defined the applicability domain of DUck to calculate RBFE and to find AC in congeneric series of compounds. Differences in structural stability ($\Delta\Delta G_{QB}$) have shown a better correspondence to kinetic experimental values such as Δpk_{off} than to $\Delta\Delta G_{bind}$ in an HSP90 inhibitors dataset. Moreover, I found that DUck is capable of yielding high confidence AC predictions given that the ligands follow a one-step dissociation mechanism and have at least one robust interaction. I validated this *a priori* conditions through four CDK2 and a BACE1 datasets.

I have developed OpenDUck, an open-source implementation of DUck. OpenDUck is capable of reproducing the results obtained from DUck, and to easily change simulation parameters such as the engine, the force fields, water models, ionic strength, box size, the use of HMR, and the steering CV. Through the use of smaller solvation box and HMR I obtained a reduction of 70% of computational cost in respect to DUck, without reduction in accuracy. Finally, I showed how the OpenDUck library can be employed to generate new protocols within the DUck framework, by studying water-mediated interactions.

Finally, I have evaluated the use of CFT as a new formalism to calculate free energies from SMD simulations and compared it to the JE. I evaluated the structural stability of a set of CRBN-CK1α-lenalidomide ternary complex and HSP90 protein-ligand key H-bonds, challenging the formalism through quicker simulations and reduced sampling. Given the same conditions, CFT yielded similar predictions and dispersion than JE in the CRBN-CK1α simulations, but a significantly smaller hysteresis in the HSP90 simulations. At high steering speed and small sampling size, however, the performance of CFT and JE was not significantly different.

CHAPTER 9: BIBLIOGRAPHY

9. BIBLIOGRAPHY

1. Villoutreix, B. O. Post-Pandemic Drug Discovery and Development: Facing Present and Future Challenges. *Frontiers in Drug Discovery* **1**, 728469 (2021).
2. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision Public Health for the Era of Precision Medicine. *Am J Prev Med* **50**, 398–401 (2016).
3. Davis, R. L. Mechanism of Action and Target Identification: A Matter of Timing in Drug Discovery. *iScience* **23**, 101487 (2020).
4. Mak, I. W. Y., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res* **6**, 114 (2014).
5. Blay, V., Tolani, B., Ho, S. P. & Arkin, M. R. High-Throughput Screening: today's biochemical and cell-based approaches. *Drug Discov Today* **25**, 1807–1821 (2020).
6. Temml, V. & Kutil, Z. Structure-based molecular modeling in SAR analysis and lead optimization. *Comput Struct Biotechnol J* **19**, 1431–1444 (2021).
7. Dou, T., Lian, T., Shu, S., He, Y. & Jiang, J. The substrate and inhibitor binding mechanism of polyspecific transporter OAT1 revealed by high-resolution cryo-EM. *Nat Struct Mol Biol* **30**, 1794–1805 (2023).
8. O' Donovan, D. H., De Fusco, C., Kuhnke, L. & Reichel, A. Trends in Molecular Properties, Bioavailability, and Permeability across the Bayer Compound Collection. *J Med Chem* **66**, 2347–2360 (2023).
9. Pradeep, P., Paul Friedman, K. & Judson, R. Structure-based QSAR models to predict repeat dose toxicity points of departure. *Computational Toxicology* **16**, 100139 (2020).
10. Khojasteh, S. C., Wong, H., Zhang, D. & Hop, C. E. C. A. DMPK Lead Optimization. *Discovery DMPK Quick Guide* 63–139 (2022).
11. Singh, N. *et al.* Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery* **3**, 1201419 (2023).
12. Miller, J. A., Fredrickson, M. E., Greene, J. M., Jay, M. & Oyewumi, M. O. Reimagining drug manufacturing paradigm in today's pharmacy landscape. *Journal of the American Pharmacists Association* **62**, 1761–1764 (2022).
13. Dimasi, J. A. Research and Development Costs of New Drugs. *Journal of America Medical Association* **324**, 517–517 (2020).
14. Sun, D., Gao, W., Hu, H. & Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B* **12**, 3049–3062 (2022).
15. Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *Journal of American Medical Association* **323**, 844–853 (2020).
16. de la Torre, B. G. & Albericio, F. The Pharmaceutical Industry in 2023: An Analysis of FDA Drug Approvals from the Perspective of Molecules. *Molecules* **29**, 585 (2024).
17. May, M. How to fight antibiotic resistance. *Nat Med* **29**, 1583–1586 (2023).
18. McGowan, J. E. Antimicrobial Resistance in Hospital Organisms and Its Relation to Antibiotic Use. *Rev Infect Dis* **5**, 1033–1048 (1983).
19. Larsson, D. G. J. & Flach, C. F. Antibiotic resistance in the environment. *Nat Rev Microbiol* **20**, 257–269 (2021).
20. Abushaheen, M. A. *et al.* Antimicrobial resistance, mechanisms and its clinical significance. *Disease-a-Month* **66**, 100971 (2020).
21. Reygaert, W. C. & Reygaert, W. C. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol* **4**, 482–501 (2018).

22. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. *Adv Pharm Bull* **7**, 339–348 (2017).
23. Wang, X., Zhang, H. & Chen, X. Drug resistance and combating drug resistance in cancer. *Cancer Drug Resist* **2**, 141–160 (2019).
24. Salzberg, S. L. Open questions: How many genes do we have? *BMC Biol* **16**, 1–3 (2018).
25. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* **1**, 727–730 (2002).
26. Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* **17**, 317–332 (2018).
27. Miljković, F. & Bajorath, J. Computational Analysis of Kinase Inhibitors Identifies Promiscuity Cliffs across the Human Kinome. *ACS Omega* **3**, 17295–17308 (2018).
28. Levit, A., Beuming, T., Krilov, G., Sherman, W. & Niv, M. Y. Predicting GPCR promiscuity using binding site features. *J Chem Inf Model* **54**, 184–194 (2014).
29. Hardy, J. A. & Wells, J. A. Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol* **14**, 706–715 (2004).
30. Kufareva, I., Ilatovskiy, A. V. & Abagyan, R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res* **40**, D535–D540 (2012).
31. Fuller, J. C., Burgoyne, N. J. & Jackson, R. M. Predicting druggable binding sites at the protein–protein interface. *Drug Discov Today* **14**, 155–161 (2009).
32. Francois, M., Donovan, P. & Fontaine, F. Modulating transcription factor activity: Interfering with protein-protein interaction networks. *Semin Cell Dev Biol* **99**, 12–19 (2020).
33. Panjkovich, A. & Daura, X. Assessing the structural conservation of protein pockets to study functional and allosteric sites: Implications for drug discovery. *BMC Struct Biol* **10**, 1–14 (2010).
34. Dang, C. V., Reddy, E. P., Shokat, K. M. & Soucek, L. Drugging the 'undruggable' cancer targets. *Nat Rev Cancer* **17**, 502–508 (2017).
35. Blair, H. A. Sotorasib: First Approval. *Drugs* **81**, 1573–1579 (2021).
36. Schmidtke, P., Alvarez-garcia, D., Seco, J. & Barril, X. Expanding the Target Space: Druggability Assessments. in *Physico-Chemical and Computational Approaches to Drug Discovery* 302–318 (The Royal Society of Chemistry, 2012).
37. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc Natl Acad Sci U S A* **112**, 15898–15903 (2015).
38. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
39. Vellosillo, P. & Minguez, P. A global map of associations between types of protein posttranslational modifications and human genetic diseases. *iScience* **24**, 102917 (2021).
40. Overall, C. M. The Human Proteome: 90% in the Light, 10% on the Dark Side. *J Proteome Res* **19**, 4731–4734 (2020).
41. Azria, D. & Rosenstein, B. S. Use of genomics to balance cure and complications. *Nat Rev Clin Oncol* **17**, 9–10 (2019).
42. Dowden, H. & Munro, J. Trends in clinical success rates and therapeutic focus. *Nat Rev Drug Discov* **18**, 495–496 (2019).
43. Ingber, D. E. Human organs-on-chips for disease modelling, drug development and personalized medicine. *Nat Rev Genet* **23**, 467–491 (2022).
44. Uzozie, A. C. & Aebersold, R. Advancing translational research and precision medicine with targeted proteomics. *J Proteomics* **189**, 1–10 (2018).

45. Kiriiri, G. K., Njogu, P. M. & Mwangi, A. N. Exploring different approaches to improve the success of drug discovery and development projects: a review. *Futur J Pharm Sci* **6**, 1–12 (2020).
46. Sahin, S. & Benet, L. Z. The operational multiple dosing half-life: A key to defining drug accumulation in patients and to designing extended release dosage forms. *Pharm Res* **25**, 2869–2877 (2008).
47. Pérez Santín, E. *et al.* Toxicity prediction based on artificial intelligence: A multidisciplinary overview. *Wiley Interdiscip Rev Comput Mol Sci* **11**, e1516 (2021).
48. Blanco, M. J., Gardinier, K. M. & Namchuk, M. N. Advancing New Chemical Modalities into Clinical Studies. *ACS Med Chem Lett* **13**, 1691–1698 (2022).
49. Disney, M. D. Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine. *J Am Chem Soc* **141**, 6776–6790 (2019).
50. Hoy, S. M. Patisiran: First Global Approval. *Drugs* **78**, 1625–1631 (2018).
51. Scott, L. J. Givosiran: First Approval. *Drugs* **80**, 335–339 (2020).
52. Fosgerau, K. & Hoffmann, T. Peptide therapeutics: current status and future directions. *Drug Discov Today* **20**, 122–128 (2015).
53. Valeur, E. *et al.* New Modalities for Challenging Targets in Drug Discovery. *Angewandte Chemie International Edition* **56**, 10294–10323 (2017).
54. Liu, X. & Ciulli, A. Proximity-Based Modalities for Biology and Medicine. *ACS Cent Sci* **9**, 1269–1284 (2023).
55. Hu, Z. & Crews, C. M. Recent Developments in PROTAC-Mediated Protein Degradation: From Bench to Clinic. *ChemBioChem* **23**, e202100270 (2022).
56. Price, E. *et al.* Beyond Rule of Five and PROTACs in Modern Drug Discovery: Polarity Reducers, Chameleonicity, and the Evolving Physicochemical Landscape. *J Med Chem* **67**, 5683–5698 (2024).
57. Caron, G., Digiesi, V., Solaro, S. & Ermondi, G. Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discov Today* **25**, 621–627 (2020).
58. Boike, L., Henning, N. J. & Nomura, D. K. Advances in covalent drug discovery. *Nat Rev Drug Discov* **21**, 881–898 (2022).
59. Sakamoto, K. M. *et al.* Protacs: Chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proc Natl Acad Sci U S A* **98**, 8554–8559 (2001).
60. Li, K. & Crews, C. M. PROTACs: past, present and future. *Chem Soc Rev* **51**, 5214–5236 (2022).
61. Takahashi, D. *et al.* Second-Generation AUTACs for Targeted Autophagic Degradation. *J Med Chem* **66**, 12342–12372 (2023).
62. Chen, P. H. *et al.* Modulation of Phosphoprotein Activity by Phosphorylation Targeting Chimeras (PhosTACs). *ACS Chem Biol* **16**, 2808–2815 (2021).
63. Henning, N. J. *et al.* Deubiquitinase-targeting chimeras for targeted protein stabilization. *Nat Chem Biol* **18**, 412–421 (2022).
64. Bertran-Mostazo, A. *et al.* Proximity-Induced Pharmacology for Amyloid-Related Diseases. *Cells* **13**, 449 (2024).
65. Bajorath, J. Computer-aided drug discovery. *F1000Res* **4**, 630 (2015).
66. Frye, L., Bhat, S., Akinsanya, K. & Abel, R. From computer-aided drug discovery to computer-driven drug discovery. *Drug Discov Today Technol* **39**, 111–117 (2021).
67. Muratov, E. N. *et al.* QSAR without borders. *Chem Soc Rev* **49**, 3525–3564 (2020).
68. Johnson, S. R. The trouble with QSAR (or how i learned to stop worrying and embrace fallacy). *J Chem Inf Model* **48**, 25–26 (2008).

69. Bajorath, J. Duality of activity cliffs in drug discovery. *Expert Opin Drug Discov* **14**, 517–520 (2019).
70. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin Drug Discov* **16**, 949–959 (2021).
71. Lv, Q., Zhou, F., Liu, X. & Zhi, L. Artificial intelligence in small molecule drug discovery from 2018 to 2023: Does it really work? *Bioorg Chem* **141**, 106894 (2023).
72. Renaud, J. P. *et al.* Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat Rev Drug Discov* **17**, 471–492 (2018).
73. Okumura, H. Perspective for Molecular Dynamics Simulation Studies of Amyloid-β Aggregates. *Journal of Physical Chemistry B* **127**, 10931–10940 (2023).
74. Zhao, J. *et al.* Understanding the Mechanism of Increased IgG/IgE Reactivity but Decreased Immunodetection Recovery in Thermally Induced Shrimp (Litopenaeus vannamei) Tropomyosin via Multispectroscopic and Molecular Dynamics Simulation Techniques. *J Agric Food Chem* (2022).
75. Das, P. *et al.* Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng* **5**, 613–623 (2021).
76. Dixon, T. *et al.* Predicting the structural basis of targeted protein degradation by integrating molecular dynamics simulations with structural mass spectrometry. *Nat Commun* **13**, 1–24 (2022).
77. Agoni, C., Olotu, F. A., Ramharack, P. & Soliman, M. E. Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say? *J Mol Model* **26**, 1–11 (2020).
78. Seco, J., Luque, F. J. & Barril, X. Binding site detection and druggability index from first principles. *J Med Chem* **52**, 2363–2371 (2009).
79. Brenke, R. *et al.* Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **25**, 621–627 (2009).
80. Guvench, O. & MacKerell, A. D. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput Biol* **5**, e1000435 (2009).
81. Basse, N. *et al.* Toward the Rational Design of p53-Stabilizing Drugs: Probing the Surface of the Oncogenic Y220C Mutant. *Chem Biol* **17**, 46–56 (2010).
82. Ghanakota, P. & Carlson, H. A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J Med Chem* **59**, 10383–10399 (2016).
83. Stumpfe, D. & Bajorath, J. Current trends, overlooked issues, and unmet challenges in virtual screening. *J Chem Inf Model* **60**, 4112–4115 (2020).
84. Zhao, H. The Science and Art of Structure-Based Virtual Screening. *ACS Med Chem Lett* **15**, 436–440 (2024).
85. Walters, W. P. & Wang, R. New Trends in Virtual Screening. *J Chem Inf Model* **60**, 4109–4111 (2020).
86. Sahakyan, H. Improving virtual screening results with MM/GBSA and MM/PBSA rescoring. *J Comput Aided Mol Des* **35**, 731–736 (2021).
87. Cournia, Z., Chipot, C., Roux, B., York, D. M. & Sherman, W. Free Energy Methods in Drug Discovery - Introduction. *ACS Symposium Series* **1397**, 1–38 (2021).
88. Schindler, C. E. M. *et al.* Large-scale assessment of binding free energy calculations in active drug discovery projects. *J Chem Inf Model* **60**, 5457–5474 (2020).
89. Scheen, J., Mackey, M. & Michel, J. Data-driven generation of perturbation networks for relative binding free energy calculations. *Digital Discovery* **1**, 870–885 (2022).

90. Pitman, M., Hahn, D. F., Tresadern, G. & Mobley, D. L. To Design Scalable Free Energy Perturbation Networks, Optimal Is Not Enough. *J Chem Inf Model* **63**, 1776–1793 (2023).
91. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug–target residence time and its implications for lead optimization. *Nat Rev Drug Discov* **5**, 730–739 (2006).
92. Yin, N., Pei, J. & Lai, L. A comprehensive analysis of the influence of drug binding kinetics on drug action at molecular and systems levels. *Mol Biosyst* **9**, 1381–1389 (2013).
93. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nat Rev Clin Oncol* **8**, 189–190 (2011).
94. Vijaykumar, A., Bolhuis, P. G. & Ten Wolde, P. R. The intrinsic rate constants in diffusion-influenced reactions. *Faraday Discuss* **195**, 421–441 (2017).
95. Lee, S., Wang, D., Seeliger, M. A. & Tiwary, P. Calculating Protein-Ligand Residence Times Through State Predictive Information Bottleneck based Enhanced Sampling. *bioRxiv* 2024.04.16.589710 (2024).
96. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat Commun* **6**, 1–10 (2015).
97. Kokh, D. B. *et al.* Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J Chem Theory Comput* **14**, 3859–3869 (2018).
98. Potterton, A. *et al.* Ensemble-Based Steered Molecular Dynamics Predicts Relative Residence Time of A 2A Receptor Binders. *J Chem Theory Comput* **15**, 3316–3330 (2019).
99. Alquraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
100. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
101. Dessimoz, C. & Thomas, P. D. AI and the democratization of knowledge. *Sci Data* **11**, 1–5 (2024).
102. Bajorath, J. State-of-the-art of Artificial Intelligence in Medicinal Chemistry. *Future Sci OA* **7**, 702 (2021).
103. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1**, 206–215 (2019).
104. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat Mach Intell* **2**, 573–584 (2020).
105. Sadeghi, M. & Noé, F. Large-scale simulation of biomembranes incorporating realistic kinetics into coarse-grained models. *Nat Commun* **11**, 1–13 (2020).
106. Jung, J. *et al.* Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *J Comput Chem* **40**, 1919–1930 (2019).
107. Yu, A. *et al.* A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophys J* **120**, 1097–1104 (2021).
108. Markidis, S. & Laure, E. Solving software challenges for exascale: International Conference on Exascale Applications and Software, EASC 2014 Stockholm, Sweden, April 2–3, 2014 revised selected papers. *Lecture Notes in Computer Science* **8759**, (2015).
109. Pérez, A., Luque, F. J. & Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J Am Chem Soc* **129**, 14739–14745 (2007).
110. Duan, Y. & Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science (1979)* **282**, 740–744 (1998).

111. Izrailev, S., Crofts, A. R., Berry, E. A. & Schulten, K. Steered molecular dynamics simulation of the Rieske subunit motion in the cytochrome bc1 complex. *Biophys J* **77**, 1753–1768 (1999).
112. Gamini, R., Han, W., Stone, J. E. & Schulten, K. Assembly of Nsp1 Nucleoporins Provides Insight into Nuclear Pore Complex Gating. *PLoS Comput Biol* **10**, e1003488 (2014).
113. Reddy, T. *et al.* Nothing to Sneeze At: A Dynamic and Integrative Computational Model of an Influenza A Virion. *Structure* **23**, 584–597 (2015).
114. Liu, C. *et al.* Cyclophilin A stabilizes the HIV-1 capsid through a novel non-canonical binding site. *Nat Commun* **7**, 1–10 (2016).
115. Perilla, J. R. & Schulten, K. Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. *Nat Commun* **8**, 1–10 (2017).
116. Durrant, J. D. *et al.* Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent Sci* **6**, 189–196 (2020).
117. Schlick, T. & Portillo-Ledesma, S. Biomolecular modeling thrives in the age of technology. *Nat Comput Sci* **1**, 321–331 (2021).
118. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med Res Rev* **16**, 3–50 (1996).
119. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).
120. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J Chem Inf Model* **52**, 2864–2875 (2012).
121. Cieplinski, T., Danel, T., Podlewska, S. & Jastrzebski, S. We Should at Least Be Able to Design Molecules That Dock Well. *J Chem Inf Model* **63**, 3238–3247 (2020).
122. Gironda-Martínez, A., Donckele, E. J., Samain, F. & Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges. *ACS Pharmacol Transl Sci* **4**, 1265–1279 (2021).
123. Kodadek, T. The rise, fall and reinvention of combinatorial chemistry. *Chemical Communications* **47**, 9757–9763 (2011).
124. Bellmann, L., Penner, P., Gastreich, M. & Rarey, M. Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. *J Chem Inf Model* **62**, 553–566 (2022).
125. Warr, W. A., Nicklaus, M. C., Nicolaou, C. A. & Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J Chem Inf Model* (2022).
126. Kuan, J., Radaeva, M., Avenido, A., Cherkasov, A. & Gentile, F. Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* **13**, e1678 (2023).
127. Eichler, G. S., Imbert, G., Branson, J., Balibey, R. & Laramie, J. M. Democratizing data at Novartis through clinical trial data access. *Drug Discov Today* **27**, 1533–1537 (2022).
128. Hopkins, A. M., Rowland, A. & Sorich, M. J. Data sharing from pharmaceutical industry sponsored clinical studies: Audit of data availability. *BMC Med* **16**, 1–6 (2018).
129. Modi, N. D. *et al.* A 10-year update to the principles for clinical trial data sharing by pharmaceutical companies: perspectives based on a decade of literature and policies. *BMC Med* **21**, 1–9 (2023).
130. Kaufman, L., Gore, K. & Zandee, J. C. Data Standardization, Pharmaceutical Drug Development, and the 3Rs. *ILAR J* **57**, 109–119 (2016).
131. Lippert, J. *et al.* Open Systems Pharmacology Community—An Open Access, Open Source, Open Science Approach to Modeling and Simulation in

Pharmaceutical Sciences. *CPT Pharmacometrics Syst Pharmacol* **8**, 878–882 (2019).
132. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).
133. Mobley, D. L. *et al.* Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J Chem Theory Comput* **14**, 6076–6092 (2018).
134. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 1–9 (2016).
135. Boby, M. L. *et al.* Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science (1979)* **382**, eabo7201 (2023).
136. Leach, A. R. & Gillet, V. J. An introduction to chemoinformatics. *An Introduction To Chemoinformatics* 1–255 (2007).
137. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **42**, 1273–1280 (2002).
138. Smith, D. H., Carhart, R. E. & Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J Chem Inf Comput Sci* **25**, 64–73 (1985).
139. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50**, 742–754 (2010).
140. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J Chem Doc* **5**, 107–113 (1965).
141. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* **28**, 31–36 (1988).
142. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
143. Korn, M., Ehrt, C., Ruggiu, F., Gastreich, M. & Rarey, M. Navigating large chemical spaces in early-phase drug discovery. *Curr Opin Struct Biol* **80**, 102578 (2023).
144. Bellmann, L., Penner, P. & Rarey, M. Topological Similarity Search in Large Combinatorial Fragment Spaces. *J Chem Inf Model* **61**, 238–251 (2021).
145. Schmidt, R., Klein, R. & Rarey, M. Maximum Common Substructure Searching in Combinatorial Make-on-Demand Compound Spaces. *J Chem Inf Model* (2021).
146. Schmidt, R., Krull, F., Heinzke, A. L. & Rarey, M. Disconnected Maximum Common Substructures under Constraints. *J Chem Inf Model* **61**, 167–178 (2021).
147. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys Rev* **9**, 91–102 (2017).
148. Jhoti, H. & Leach, A. R. Structure-based drug discovery. *Structure-Based Drug Discovery* 1–249 (2007).
149. Shen, C. *et al.* From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdiscip Rev Comput Mol Sci* **10**, e1429 (2020).
150. Morley, S. D. & Afshar, M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®. *J Comput Aided Mol Des* **18**, 189–208 (2004).
151. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput Biol* **10**, e1003571 (2014).
152. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput J* **7**, 308–313 (1965).

153. Moreno, D. *et al.* DFFR: A New Method for High-Throughput Recalibration of Automatic Force-Fields for Drugs. *J Chem Theory Comput* **16**, 6598–6608 (2020).
154. Bayly, C. I. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* **117**, 5179–5197 (1995).
155. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* **4**, 187–217 (1983).
156. Van Gunsteren, W. F. & Berendsen, H. J. C. The GROMOS Software for (Bio)Molecular Simulation GROMOS87 Groningen Molecular Simulation (GROMOS) Library Manual.
157. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118**, 11225–11236 (1996).
158. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J Comput Chem* **25**, 1157–1174 (2004).
159. Verlet, L. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **159**, 98 (1967).
160. Frenkel, D., Smit, B., Frenkel, D. & Smit, B. *Chapter 1 – Introduction*. *Understanding Molecular Simulation* (2002).
161. Liepinsh, E. & Otting, G. Organic solvents identify specific ligand binding sites on protein surfaces. *Nat Biotechnol* **15**, 264–268 (1997).
162. Mattos, C. *et al.* Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J Mol Biol* **357**, 1471–1482 (2006).
163. Alvarez-Garcia, D. & Barril, X. Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. *J Med Chem* **57**, 8530–8539 (2014).
164. Lexa, K. W. & Carlson, H. A. Improving protocols for protein mapping through proper comparison to crystallography data. *J Chem Inf Model* **53**, 391–402 (2013).
165. Yanagisawa, K., Moriwaki, Y., Terada, T. & Shimizu, K. EXPRORER: Rational Cosolvent Set Construction Method for Cosolvent Molecular Dynamics Using Large-Scale Computation. *J Chem Inf Model* **61**, 2744–2753 (2021).
166. Lee, T. S. *et al.* GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *J Chem Inf Model* **58**, 2043–2050 (2018).
167. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip Rev Comput Mol Sci* **1**, 826–843 (2011).
168. Kästner, J. Umbrella sampling. *Wiley Interdiscip Rev Comput Mol Sci* **1**, 932–942 (2011).
169. Park, S. & Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J Chem Phys* **120**, 5946–5961 (2004).
170. Prinz, J. H. *et al.* Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *Journal of Chemical Physics* **134**, (2011).
171. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **314**, 141–151 (1999).
172. Deb, I. & Frank, A. T. Accelerating Rare Dissociative Processes in Biomolecules Using Selectively Scaled MD Simulations. *J Chem Theory Comput* **15**, 5817–5828 (2019).
173. Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys Rev Lett* **78**, 2690 (1997).

174. Crooks, G. E. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys Rev E* (1999).
175. Schmidtke, P., Javier Luque, F., Murray, J. B. & Barril, X. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *J Am Chem Soc* **133**, 18903–18910 (2011).
176. Ruiz-Carmona, S. *et al.* Dynamic undocking and the quasi-bound state as tools for drug discovery. *Nat Chem* **9**, 201–206 (2016).
177. Majewski, M. & Barril, X. Structural Stability Predicts the Binding Mode of Protein-Ligand Complexes. *J Chem Inf Model* **60**, 1644–1651 (2020).
178. Rachman, M. *et al.* Discovery of a novel kinase hinge binder fragment by dynamic undocking. *RSC Med Chem* **11**, 552–558 (2020).
179. Chemical Computing Group ULC. Molecular Operating Environment (MOE). Preprint at (2022).
180. An Informal AMBER Small Molecule Force Field: parm@Frosst. http://www.ccl.net/cca/data/parm_at_Frosst/.
181. Majewski, M., Ruiz-Carmona, S. & Barril, X. An investigation of structural stability in protein-ligand complexes reveals the balance between order and disorder. *Commun Chem* **2**, 1–8 (2019).
182. Kokh, D. B., Kaufmann, T., Kister, B. & Wade, R. C. Machine Learning Analysis of τRAMD Trajectories to Decipher Molecular Determinants of Drug-Target Residence Times. *Front Mol Biosci* **6**, (2019).
183. Liu, L. *et al.* UbiHub: a data hub for the explorers of ubiquitination pathways. *Bioinformatics* **35**, 2882–2884 (2019).
184. Bateman, A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
185. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res* **41**, (2013).
186. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J Chem Inf Model* **56**, 599–604 (2016).
187. Case, D. *et al.* AMBER 2017. Preprint at (2017).
188. Hart, P. E., Nilsson, N. J. & Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* **4**, 100–107 (1968).
189. Horn-Ghetko, D. *et al.* Ubiquitin ligation to F-box protein targets by SCF-RBR E3-E3 super-assembly. *Nature* **590**, 671–676 (2021).
190. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).
191. Stein, A. & Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS One* **8**, e63090 (2013).
192. Frisch, M. J. *et al.* Gaussian 16. Preprint at (2016).
193. He, X., Man, V. H., Yang, W., Lee, T. S. & Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *Journal of Chemical Physics* **153**, (2020).
194. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J Chem Theory Comput* **11**, 1864–1874 (2015).
195. Shkurti, A. *et al.* pyPcazip: A PCA-based toolkit for compression and analysis of molecular simulation data. *SoftwareX* **5**, 44–50 (2015).
196. Baell, J. B. & Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 - Utility and Limitations. *ACS Chem Biol* **13**, 36–44 (2018).
197. ChemAxon (https://www.chemaxon.com). cxcalc suite.
198. Molecular Networks Altamira (www.mn-am.com). 3D Structure Generator CORINA Classic.

199. Martínez Cartró, M. Biophysical and structural approaches to discover small-molecule modulators of challenging targets in cancer. *Tesis Doctorals - Departament - Farmàcia, Tecnologia farmacèutica i Físicoquímica* (2021).
200. Miñarro Lleonar, M. Technological developments in Virtual Screening for the discovery of small molecules with novel mechanisms of action. *Tesis Doctorals - Facultat - Farmàcia i Ciències de l'Alimentació* (2023).
201. Piticchio, S. G. *et al.* Discovery of Novel BRD4 Ligand Scaffolds by Automated Navigation of the Fragment Chemical Space. *J Med Chem* **64**, 17887–17900 (2021).
202. Duran-Frigola, M. *et al.* Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat Biotechnol* **38**, 1087–1096 (2020).
203. O'Boyle, N. M. *et al.* Open Babel: An Open chemical toolbox. *J Cheminform* **3**, 1–14 (2011).
204. Kokh, D. B. *et al.* A workflow for exploring ligand dissociation from a macromolecule: Efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *Journal of Chemical Physics* **153**, (2020).
205. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* **52**, D1180–D1192 (2024).
206. Schrödinger LLC. Schrödinger Release: LigPrep. Preprint at (2020).
207. Liu, Z. *et al.* Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc Chem Res* **50**, 302–309 (2017).
208. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* **35**, D198–D201 (2007).
209. Benson, M. L. *et al.* Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Res* **36**, D674–D678 (2008).
210. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7**, 1–13 (2015).
211. Barca, G. M. J. *et al.* Recent developments in the general atomic and molecular electronic structure system. *J Chem Phys* **152**, 154102 (2020).
212. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J Comput Chem* **22**, 501–508 (2001).
213. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *Journal of Chemical Physics* **153**, 44130 (2020).
214. Bell, J. A. *et al.* Prime and the Schrödinger computational chemistry suite of programs. in 534–538 (2012).
215. Chicco, D., Tötsch, N. & Jurman, G. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min* **14**, 1–22 (2021).
216. JE, M. & PM, D. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J Comput Aided Mol Des* **10**, 607–622 (1996).
217. Miñarro-Lleonar, M., Bertran-Mostazo, A., Duro, J., Barril, X. & Juárez-Jiménez, J. Lenalidomide Stabilizes Protein-Protein Complexes by Turning Labile Intermolecular H-Bonds into Robust Interactions. *J Med Chem* **66**, 6037–6046 (2023).
218. Petzold, G., Fischer, E. S. & Thomä, N. H. Structural basis of lenalidomide-induced CK1α degradation by the CRL4CRBN ubiquitin ligase. *Nature* **532**, 127–130 (2016).

219. Schrodinger LLC. The PyMOL Molecular Graphics System. Preprint at (2010).
220. Liao, Q., Pabis, A., Strodel, B. & Kamerlin, S. C. L. Extending the Nonbonded Cationic Dummy Model to Account for Ion-Induced Dipole Interactions. *Journal of Physical Chemistry Letters* **8**, 5408–5414 (2017).
221. Neria, E., Fischer, S. & Karplus, M. Simulation of activation free energies in molecular systems. *J Chem Phys* **105**, 1902–1921 (1996).
222. Case, D. A. *et al.* AmberTools. *J Chem Inf Model* **63**, 6183–6191 (2023).
223. Kim, H., Park, J. & Kim, J. M. Targeted Protein Degradation to Overcome Resistance in Cancer Therapies: PROTAC and N-Degron Pathway. *Biomedicines* **10**, 2100 (2022).
224. Zhao, L., Zhao, J., Zhong, K., Tong, A. & Jia, D. Targeted protein degradation: mechanisms, strategies and application. *Signal Transduct Target Ther* **7**, 1–13 (2022).
225. Dikic, I. Proteasomal and autophagic degradation systems. *Annu Rev Biochem* **86**, 193–224 (2017).
226. Luh, L. M. *et al.* Prey for the Proteasome: Targeted Protein Degradation—A Medicinal Chemist's Perspective. *Angewandte Chemie International Edition* **59**, 15448–15466 (2020).
227. Ester Morreale, F. & Walden, H. Types of Ubiquitin Ligases. (2016).
228. Fang, Y. *et al.* Targeted protein degrader development for cancer: advances, challenges, and opportunities. *Trends Pharmacol Sci* **44**, 303–317 (2023).
229. Galdeano, C. *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. *J Med Chem* **57**, 8657–8663 (2014).
230. Lucas, X., Van Molle, I. & Ciulli, A. Surface Probing by Fragment-Based Screening and Computational Methods Identifies Ligandable Pockets on the von Hippel-Lindau (VHL) E3 Ubiquitin Ligase. *J Med Chem* **61**, 7387–7393 (2018).
231. Lupas, A. N., Zhu, H. & Korycinski, M. The Thalidomide-Binding Domain of Cereblon Defines the CULT Domain Family and Is a New Member of the β-Tent Fold. *PLoS Comput Biol* **11**, e1004023 (2015).
232. Han, T. *et al.* Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. *Science (1979)* **356**, (2017).
233. Baird, L. & Yamamoto, M. The Molecular Mechanisms Regulating the KEAP1-NRF2 Pathway. *Mol Cell Biol* **40**, (2020).
234. Gowans, F. A. *et al.* Covalent Degrader of the Oncogenic Transcription Factor β-Catenin. *bioRxiv* 2023.10.31.565018 (2023).
235. Watanabe, N. *et al.* M-phase kinases induce phospho-dependent ubiquitination of somatic Wee1 by SCFβ-TrCP. *Proc Natl Acad Sci U S A* **101**, 4419–4424 (2004).
236. Bhuniya, R. *et al.* Design, Synthesis, and Biological Evaluation of Apcin-Based CDC20 Inhibitors. *ACS Med Chem Lett* **13**, 188–195 (2022).
237. Kimani, S. W. *et al.* Discovery of a Novel DCAF1 Ligand Using a Drug–Target Interaction Prediction Model: Generalizing Machine Learning to New Drug Targets. *J Chem Inf Model* (2023).
238. Hickey, C. M. *et al.* Co-opting the E3 ligase KLHDC2 for targeted protein degradation by small molecules. *Nat Struct Mol Biol* **31**, 311–322 (2024).
239. Ramachandran, S. *et al.* Structure-based design of a phosphotyrosine-masked covalent ligand targeting the E3 ligase SOCS2. *Nat Commun* **14**, 1–17 (2023).
240. Asatsuma-Okumura, T., Ito, T. & Handa, H. Molecular mechanisms of cereblon-based drugs. *Pharmacol Ther* **202**, 132–139 (2019).

241. James, L. C., Keeble, A. H., Khan, Z., Rhodes, D. A. & Trowsdale, J. Structural basis for PRYSPRY-mediated tripartite motif (TRIM) protein function. *Proc Natl Acad Sci U S A* **104**, 6200–6205 (2007).
242. Uljon, S., Xu, X., Marto, J. A., Pear, W. S. & Blacklow Correspondence, S. C. Structural Basis for Substrate Selectivity of the E3 Ligase COP1. *Structure/Folding and Design* **24**, 687–696 (2016).
243. Rueda, A. J. V., Monzon, A. M., Ardanaz, S. M., Iglesias, L. E. & Parisi, G. Large scale analysis of protein conformational transitions from aqueous to non-aqueous media. *BMC Bioinformatics* **19**, 1–10 (2018).
244. Balasubramaniam, D. *et al.* How the ankyrin and SOCS box protein, ASB9, binds to creatine kinase. *Biochemistry* **54**, 1673–1680 (2015).
245. Hatakeyama, S. TRIM Family Proteins: Roles in Autophagy, Immunity, and Carcinogenesis. *Trends Biochem Sci* **42**, 297–311 (2017).
246. Yumimoto, K. & Nakayama, K. I. Recent insight into the role of FBXW7 as a tumor suppressor. *Semin Cancer Biol* **67**, 1–15 (2020).
247. Yeh, C. H., Bellon, M. & Nicot, C. FBXW7: a critical tumor suppressor of human cancers. *Mol Cancer* **17**, 1–19 (2018).
248. Shimizu, K., Nihira, N. T., Inuzuka, H. & Wei, W. Physiological functions of FBW7 in cancer and metabolism. *Cell Signal* **46**, 15–22 (2018).
249. Singh, N., Zeke, A. & Reményi, A. Systematic Discovery of FBXW7-Binding Phosphodegrons Highlights Mitogen-Activated Protein Kinases as Important Regulators of Intracellular Protein Levels. *Int J Mol Sci* **23**, 3320 (2022).
250. Liu, J. & Nussinov, R. Molecular Dynamics Reveal the Essential Role of Linker Motions in the Function of Cullin–RING E3 Ligases. *J Mol Biol* **396**, 1508–1523 (2010).
251. Irwin, J. J. & Shoichet, B. K. Docking Screens for Novel Ligands Conferring New Biology. *J Med Chem* **59**, 4103–4120 (2016).
252. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
253. Stein, R. M. *et al.* Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614 (2020).
254. Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
255. Gloriam, D. E. Bigger is better in virtual drug screens. *Nature* **566**, 193–194 (2019).
256. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J Chem Inf Model* **55**, 2324–2337 (2015).
257. Rarey, M. & Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J Comput Aided Mol Des* **15**, 497–520 (2001).
258. Alon, A. *et al.* Structures of the σ2 receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759–764 (2021).
259. Gentile, F. *et al.* Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat Protoc* **17**, 672–697 (2022).
260. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem Sci* **12**, 7866–7881 (2021).
261. Sivula, T. *et al.* Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries. *J Chem Inf Model* **63**, 5773–5783 (2023).
262. Cao, Z., Sciabola, S. & Wang, Y. Large-Scale Pretraining Improves Sample Efficiency of Active Learning-Based Virtual Screening. *J Chem Inf Model* **64**, 1882–1891 (2024).

263. Ghorbani, M., Genentech, L. G., Genentech, P. B. & Keiser, M. J. Autoregressive fragment-based diffusion for pocket-aware ligand design. (2023).
264. Guan, J. *et al.* 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. (2023).
265. Gao, W. & Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J Chem Inf Model* **60**, 5714–5723 (2020).
266. Sadybekov, A. A. *et al.* Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2021).
267. Beroza, P. *et al.* Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nat Commun* **13**, 1–10 (2022).
268. Penner, P. *et al.* FastGrow: on-the-fly growing and its application to DYRK1A. *J Comput Aided Mol Des* **36**, 639–651 (2022).
269. Arús-Pous, J. *et al.* Exploring the GDB-13 chemical space using deep generative models. *J Cheminform* **11**, 1–14 (2019).
270. Bender, B. J. *et al.* A practical guide to large-scale docking. *Nat Protoc* **16**, 4799–4832 (2021).
271. Soares, T. A. *et al.* The (Re)-Evolution of Quantitative Structure-Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. *J Chem Inf Model* **62**, 5317–5320 (2022).
272. Van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J Chem Inf Model* **62**, 5938–5951 (2022).
273. Dablander, M., Hanser, T., Lambiotte, R. & Morris, G. M. Exploring QSAR models for activity-cliff prediction. *J Cheminform* **15**, 1–16 (2023).
274. Stumpfe, D., Hu, H. & Bajorath, J. Advances in exploring activity cliffs. *J Comput Aided Mol Des* **34**, 929 (2020).
275. Wang, E. *et al.* End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem Rev* **119**, 9478–9508 (2019).
276. Raniolo, S. & Limongelli, V. Ligand binding free-energy calculations with funnel metadynamics. *Nat Protoc* **15**, 2837–2866 (2020).
277. Muegge, I. & Hu, Y. Recent Advances in Alchemical Binding Free Energy Calculations for Drug Discovery. *ACS Med Chem Lett* **14**, 244–250 (2023).
278. Pérez-Benito, L., Casajuana-Martin, N., Jiménez-Rosés, M., Van Vlijmen, H. & Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *J Chem Theory Comput* **15**, 1884–1895 (2019).
279. Cournia, Z., Allen, B. & Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J Chem Inf Model* **57**, 2911–2937 (2017).
280. Mey, A. S. J. S. *et al.* Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living J Comput Mol Sci* **2**, 18378–18378 (2020).
281. Wright, L. *et al.* Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem Biol* **11**, 775–785 (2004).
282. Kumar, B. K. *et al.* Pharmacophore based virtual screening, molecular docking, molecular dynamics and MM-GBSA approach for identification of prospective SARS-CoV-2 inhibitor from natural product databases. *J Biomol Struct Dyn* **40**, 1363–1386 (2022).
283. Bray, S. *et al.* Galaxy workflows for fragment-based virtual screening: a case study on the SARS-CoV-2 main protease. *J Cheminform* **14**, 1–13 (2022).
284. Rachman, M. *et al.* DUckCov: a Dynamic Undocking-Based Virtual Screening Protocol for Covalent Binders. *ChemMedChem* **14**, 1011–1021 (2019).

285. Hedges, L. O. *et al.* BioSimSpace: An interoperable Python framework for biomolecular simulation. *J Open Source Softw* **4**, 1831 (2019).
286. Wang, Y. *et al.* End-to-end differentiable construction of molecular mechanics force fields. *Chem Sci* **13**, 12016–12033 (2022).
287. Shirts, M. R. *et al.* Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *J Comput Aided Mol Des* **31**, 147–161 (2017).
288. Rizzi, A. *et al.* The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations. *J Comput Aided Mol Des* **34**, 601–633 (2020).
289. Emperador, A., Crehuet, R. & Guàrdia, E. Effect of the Water Model in Simulations of Protein–Protein Recognition and Association. *Polymersç* **13**, 176 (2021).
290. El Hage, K., Hédin, F., Gupta, P. K., Meuwly, M. & Karplus, M. Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size. *Elife* **7**, (2018).
291. Parameswaran, S. & Mobley, D. L. Box size effects are negligible for solvation free energies of neutral solutes. *J Comput Aided Mol Des* **28**, 825–829 (2014).
292. Collin, D. *et al.* Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature* **437**, 231–234 (2005).
293. Aldeghi, M., De Groot, B. L. & Gapsys, V. Chapter 2 Accurate Calculation of Free Energy Changes upon Amino Acid Mutation. in *Computational Methods in Protein Evolution* 19–47 (2019).
294. Goette, M. & Grubmüller, H. Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes. *J Comput Chem* **30**, 447–456 (2009).
295. Schmidtke, P., Bidon-chanal, A., Luque, F. J. & Barril, X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **27**, 3276–3285 (2011).
296. Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).
297. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform* **10**, 1–12 (2018).
298. Meller, A. *et al.* Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat Commun* **14**, 1–15 (2023).
299. Li, S. *et al.* PocketAnchor: Learning structure-based pocket representations for protein-ligand interaction prediction. *Cell Syst* **14**, 692-705.e6 (2023).
300. Gagliardi, L. & Rocchia, W. SiteFerret: Beyond Simple Pocket Identification in Proteins. *J Chem Theory Comput* **19**, 5242–5259 (2023).
301. Barril, X. Druggability predictions: methods, limitations, and applications. *Wiley Interdiscip Rev Comput Mol Sci* **3**, 327–338 (2013).
302. Smith, R. D. & Carlson, H. A. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. *J Chem Inf Model* **61**, 1287–1299 (2021).
303. Kimura, S. R., Hu, H. P., Ruvinsky, A. M., Sherman, W. & Favia, A. D. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J Chem Inf Model* **57**, 1388–1401 (2017).
304. Palomba, T., Baroni, M., Cross, S., Cruciani, G. & Siragusa, L. ELIOT: A platform to navigate the E3 pocketome and aid the design of new PROTACs. *Chem Biol Drug Des* **101**, 69–86 (2023).
305. Siragusa, L., Cross, S., Baroni, M., Goracci, L. & Cruciani, G. BioGPS: Navigating biological space to predict polypharmacology, off-targeting, and

selectivity. *Proteins: Structure, Function, and Bioinformatics* **83**, 517–532 (2015).
306. Palomba, T. *et al.* Exploiting ELIOT for Scaffold-Repurposing Opportunities: TRIM33 a Possible Novel E3 Ligase to Expand the Toolbox for PROTAC Design. *Int J Mol Sci* **23**, (2022).
307. Lyu, J. *et al.* AlphaFold2 structures template ligand discovery. *bioRxiv* 2023.12.20.572662 (2024).
308. Oleinikovas, V., Saladino, G., Cossins, B. P. & Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J Am Chem Soc* **138**, 14257–14263 (2016).
309. Zhao, H. Structural Basis of Conformational Dynamics in the PROTAC-Induced Protein Degradation. *ChemMedChem* e202400171 (2024).
310. Welcker, M. *et al.* Fbw7 dimerization determines the specificity and robustness of substrate degradation. *Genes Dev* **27**, 2531–2536 (2013).
311. Zaarur, N. *et al.* Proteasome Failure Promotes Positioning of Lysosomes around the Aggresome via Local Block of Microtubule-Dependent Transport. *Mol Cell Biol* **34**, 1336–1348 (2014).
312. Li, C. *et al.* Proteasome Inhibition Activates Autophagy-Lysosome Pathway Associated With TFEB Dephosphorylation and Nuclear Translocation. *Front Cell Dev Biol* **7**, 472905 (2019).
313. Gain, C. *et al.* Proteasomal inhibition triggers viral oncoprotein degradation via autophagy-lysosomal pathway. *PLoS Pathog* **16**, e1008105 (2020).
314. Sievers, Q. L. *et al.* Defining the human C2H2 zinc finger degrome targeted by thalidomide analogs through CRBN. *Science (1979)* **362**, (2018).
315. Dauch, D. *et al.* A MYC–aurora kinase A protein complex represents an actionable drug target in p53-altered liver cancer. *Nat Med* **22**, 744–753 (2016).
316. Kim, B. Y., Yang, J. S., Kwak, S. Y., Zhang, X. kun & Han, Y. H. NEMO stabilizes c-Myc through direct interaction in the nucleus. *FEBS Lett* **584**, 4524–4530 (2010).
317. Popov, N., Schülein, C., Jaenicke, L. A. & Eilers, M. Ubiquitylation of the amino terminus of Myc by SCFβ-TrCP antagonizes SCFFbw7-mediated turnover. *Nat Cell Biol* **12**, 973–981 (2010).
318. Clark, D. E. Virtual screening: Is bigger always better? Or can small be beautiful? *J Chem Inf Model* **60**, 4120–4123 (2020).
319. Lyu, J., Irwin, J. J. & Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nat Chem Biol* **19**, 712–718 (2023).
320. Perebyinis, M. & Rognan, D. Overlap of On-demand Ultra-large Combinatorial Spaces with On-the-shelf Drug-like Libraries. *Mol Inform* **42**, 2200163 (2023).
321. Neumann, A., Marrison, L. & Klein, R. Relevance of the Trillion-Sized Chemical Space “eXplore” as a Source for Drug Discovery. *ACS Med Chem Lett* **14**, 466–472 (2023).
322. Van Hilten, N., Chevillard, F. & Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J Chem Inf Model* **59**, 644–651 (2019).
323. Hönig, S. M. N. *et al.* SpaceGrow: efficient shape-based virtual screening of billion-sized combinatorial fragment spaces. *J Comput Aided Mol Des* **38**, 1–17 (2024).
324. Lübbers, J., Lessel, U. & Rarey, M. Enhanced Calculation of Property Distributions in Chemical Fragment Spaces. *J Chem Inf Model* **64**, 2008–2020 (2024).
325. Schulz, M. N. & Hubbard, R. E. Recent progress in fragment-based lead discovery. *Curr Opin Pharmacol* **9**, 615–621 (2009).
326. de Souza Neto, L. R. *et al.* In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front Chem* **8**, 507499 (2020).

327. Rachman, M., Piticchio, S., Majewski, M. & Barril, X. Fragment-to-lead tailored in silico design. *Drug Discov Today Technol* **40**, 44–57 (2021).
328. Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* **44**, 1000–1005 (2004).
329. Mobley, D. L. & Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **17**, 489–498 (2009).
330. Kaus, J. W. *et al.* How to deal with multiple binding poses in alchemical relative protein-ligand binding free energy calculations. *J Chem Theory Comput* **11**, 2670–2679 (2015).
331. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* **10**, 449–461 (2015).
332. Rastelli, G. & Pinzi, L. Refinement and rescoring of virtual screening results. *Front Chem* **7**, 467450 (2019).
333. Hassan-Harrirou, H., Zhang, C. & Lemmin, T. RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *J Chem Inf Model* **60**, 2791–2802 (2020).
334. Sánchez-Cruz, N., Medina-Franco, J. L., Mestres, J. & Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **37**, 1376–1382 (2021).
335. McGibbon, M., Money-Kyrle, S., Blay, V. & Houston, D. R. SCORCH: Improving structure-based virtual screening with machine learning classifiers, data augmentation, and uncertainty estimation. *J Adv Res* **46**, 135–147 (2023).
336. Liu, S. *et al.* Lead optimization mapper: Automating free energy calculations for lead optimization. *J Comput Aided Mol Des* **27**, 755–770 (2013).
337. Furui, K. & Ohue, M. Fastlomap: faster lead optimization mapper algorithm for large-scale relative free energy perturbation. *Journal of Supercomputing* **80**, 14417–14432 (2024).
338. Kumar, P. *et al.* Update of KDBI: Kinetic Data of Bio-molecular Interaction database. *Nucleic Acids Res* **37**, D636–D641 (2009).
339. Norval, L. W. *et al.* KOFFI and Anabel 2.0—a new binding kinetics database and its integration in an open-source binding analysis software. *Database* **2019**, 101 (2019).
340. Schuetz, D. A., Richter, L., Martini, R. & Ecker, G. F. A structure–kinetic relationship study using matched molecular pair analysis. *RSC Med Chem* **11**, 1285–1294 (2020).
341. Vauquelin, G. Effects of target binding kinetics on in vivo drug efficacy: koff, kon and rebinding. *Br J Pharmacol* **173**, 2319–2334 (2016).
342. Bernetti, M., Cavalli, A. & Mollica, L. Protein–ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling. *Medchemcomm* **8**, 534–550 (2017).
343. Dellago, C. & Hummer, G. Computing Equilibrium Free Energies Using Non-Equilibrium Molecular Dynamics. *Entropy* **16**, 41–61 (2013).