UNIVERSITAT DE BARCELONA

Molecular recognition and chemical space navigation in drug discovery

Serena Piticchio

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L’ALIMENTACIÓ

DEPARTAMENT DE FARMÀCIA I TECNOLOGIA FARMACÈUTICA I FISICOQUÍMICA

MOLECULAR RECOGNITION AND CHEMICAL SPACE NAVIGATION IN DRUG DISCOVERY

SERENA PITICCHIO
2020

Programa de Doctorat en Biomedicina, Universitat de Barcelona
Director de Tesi: Dr. Xavier Barril Alonso

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

DEPARTAMENT DE FARMÀCIA I TECNOLOGIA FARMACÈUTICA I FISICOQUÍMICA

PROGRAMA DE DOCTORAT EN BIOMEDICINA

MOLECULAR RECOGNITION AND CHEMICAL SPACE NAVIGATION IN DRUG DISCOVERY

Aquesta tesi ha estat realitzada per Serena Gaetana Piticchio sota la direcció del Dr. Xavier Barril Alonso, Professor d'Investigació ICREA en el Departament de Farmàcia i Tecnologia Farmacèutica i Fisicoquímica de la Facultat de Farmàcia i Ciències de l'Alimentació de la Universitat de Barcelona. Es presenta aquesta memòria per optar al títol de doctor per la Universitat de Barcelona en el Programa de Doctorat en Biomedicina.

Piticchio Serena Gaetana

Xavier Barril Alonso

Director de tesi

Serena Gaetana Piticchio

Doctorand

Serena Piticchio

Desembre 2020

To all the people I love.

Table of Contents

Summary

Efficient discovery of bioactive molecules is an essential goal of Computer-Aided Drug Design (CADD). The molecules can be used as chemical probes, to validate novel targets, or as starting points for drug discovery. This endeavour is particularly challenging in the case of proteins that are considered undruggable or for which no ligands are known. These are precisely the type of proteins that must be targeted in order to expand the "druggable genome" and extend the range of therapeutic opportunities.

CADD tools available nowadays are numerous but have limitations that must be overcome in order to improve the efficacy and efficiency of drug discovery. Particularly because they should also be able to exploit non-standard sites, such as protein-protein interfaces, allosteric sites or cryptic pockets. They should also be adapted to address specific needs in the drug discovery process. Finally, they can be used to gain a fundamental understanding of the behaviour of molecular systems and the rules of molecular recognition that govern the recognition of a drug by its target. In this thesis, I have explored each one of these aspects.

Initially, I developed an automatic pipeline that can be used in Fragment-Based Drug Discovery (FBDD) to navigate the "fragment chemical space". Starting from a fragment hit with a known binding mode to its target, the platform automatically seeks non-obvious analogues (scaffold hops) within large chemical collections, delivering fragment hits, with novel structures that would, otherwise, be missed. I validated the platform using a fragment hit of the first bromodomain of the Bromodomain-containing protein 4 (BRD4) taken from the literature as a starting point. The platform identified multiple fragments with novel scaffolds and excellent ligand efficiencies. For some, their binding modes could be corroborated experimentally.

The optimized fragment identified in the first study allowed us to investigate the unusual behaviour of structural water molecules in BRD4(1) and their role in molecular recognition. Paradoxically, a hydrophobic binding hot spot of BRD4(1) is lined with water molecules. A series of compounds were derived to probe the preference of this site for chemical groups with various degrees of polarity. Molecular dynamics (MD) and free energy calculations allowed us to rationalize the experimental results.

I have then used *de novo* design (DND) methods to further grow the most active fragment into a very potent and efficient drug-like BRD4 ligand.

Finally, I have discovered the first ever described inhibitors of the Three Prime Repair Exonuclease 2 (TREX2) protein. [redacted]

List of Abbreviations

ADMET	Absorption Distribution Metabolism Excretion Toxicity
BB	Building Block
BRD4	Bromodomain-containing protein 4
CADD	Computer-Aided Drug Design
DND	De Novo Design
F2L	Fragment-to-Lead
FBDD	Fragment-Based Drug Design/Discovery
FEP	Free Energy Perturbation
GPU	Graphical Processing Units
HAC	Heavy Atom Counts
HB	Hydrogen Bond
HTS	High-Throughput Screening
LE	Ligand Efficiency
MD	Molecular Dynamics
MDmix	Mixed-solvent Molecular Dynamics
ML	Machine Learning
MMGBSA	Molecular Mechanics Generalized Born Surface Area
MMGPSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MOO	Multi-Objective Optimization
NME	New Molecular Entities
QSAR	Quantitative Structure-Activity Relationship
SA	Synthetic Accessibility
TI	Thermodynamic Integration
TREX2	Three Prime Repair Exonuclease 2
VS	Virtual Screening

Chapter 1

Introduction

“Everything should be made as simple as possible, but not simpler”

Albert Einstein

1.1. The Quest For A New Drug

When one considers the transformation of drug discovery from the ancient world – when it was only possible to randomly test raw materials and observe if they have an effect on the specific illness – to modern medicine, the progress is astounding. Yet, we are still far away from the point where, for any disease, it is possible to find the proper treatment in a short period of time. This has been made painfully clear during the on-going COVID-19 crisis. Even though science has made great strides in recent years, scientists around the world are still struggling to understand and address the molecular basis of unsolved biomedical problems. The consequence is that there are still major unmet medical needs to be addressed. To name but a few, orphan diseases, deadly tropical diseases, new upcoming conditions, and the replacement of life-saving drugs with severe side-effects, are a case in point. A strong desire to ameliorate and prolong life are at the base of the search for new, more effective and safer drugs.

In this arduous quest, the tools available to the “drug hunters” nowadays are numerous, but still have limitations due to insufficient understanding about the “small world” where the molecules inhabit. Research and innovation are paramount to bridge this gap. The discovery of computers at the beginning of last century, and their application to drug discovery in the 1980’s, brought a great advancement in the field. In spite of rapid and continuous progress, these methods still suffer from important limitations and better tools are needed to improve the efficacy and efficiency of drug discovery.

In this introduction I am going to discuss the importance of Computer-Aided Drug Design (CADD) in the drug discovery process and describe some important computational tools available to drug discovery practitioners. I will also present the importance of expanding the “druggable genome” and the application of CADD in challenging therapeutic targets.

1.1.1 The Druggable Genome

After the sequencing of the human genome in the early 2000, it was estimated that about 30,000 genes can express proteins [1,2] and scientists began to ask how many of them could be considered potential therapeutic targets.

With the seminal paper of 2002 by Hopkins and Groom [3] a first attempt to calculate this number led to the estimate that 10-14% of the predicted proteome could be considered druggable and that the proteins targeted until that point represented only about the 1% of the total (399 over 30,000), of which only 0,5% led to a marketed drug (120 over 30,000). While approximate, this assessment made it clear that there was room for improvement and a race towards increasing the number of druggable proteins has since taken place. Three great initiatives have appeared recently (Illuminating the Druggable Genome, Open Targets and

Target 2035 [4,5,6]) with the specific objective to expand the druggable genome. In a recent paper [4] the number of targeted proteins is increased to 9%, but still 40% of the proteome is underexplored and may include new opportunities for therapeutic targets.

But what makes a protein druggable? If we know the rules that define a drug (Lipinski's rule of 5 [7], for example), we can expect to have the exact complementary properties on the protein [8], like a negative film. Some proteins are considered difficult to drug or even undruggable due to the lack of these properties (i.e. a deep cavity, for shape complementarity, or specific features for functional interactions). For example, transcription factors and protein-protein interactions are usually considered undruggable because their surfaces are rather flat. In other cases, the 'undruggable' label is earned after many drug discovery attempts fail (for example MYC [9]).

Other authors place the focus on the "ability to bind anything" from very small (fragments) to lead-like size compounds, that can be used as a biological tool to understand cellular mechanisms [10] (chemical probes). Under this perspective, we should talk about "bindability" instead of "druggability". In any case, the challenge nowadays is not only to find new druggable targets but also to succeed in finding binders for these intractable targets (making the 'undruggable' 'druggable') because many such targets have a role in important pathways and cellular mechanisms involved in disease.

This is a first and fundamental step to validate the target and open the way for drug discovery. Yet, finding a chemical compound that binds the target, is only the beginning of a long and expensive process to develop a new drug.

1.1.2 The Drug Discovery Process

The path to a drug is usually depicted as a linear track like in **Figure 1.1**, where a consecutive application of different approaches inevitably leads to success.

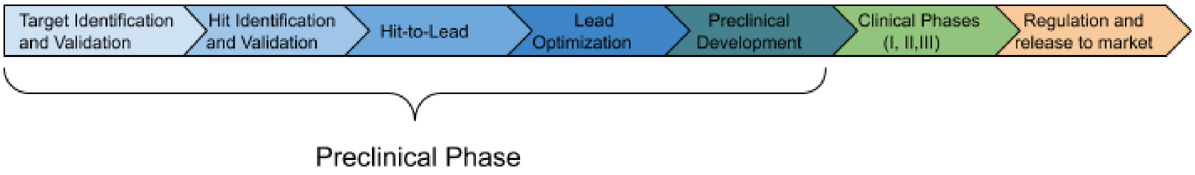

Target Identification and Validation
Hit Identification and Validation
Hit-to-Lead
Lead Optimization
Preclinical Development
Clinical Phases (I, II,III)
Regulation and release to market
Preclinical Phase

Figure 1.1: *Phases of Drug Discovery.*

In reality it is more likely a tortuous road with many obstacles to overcome and barriers that obligate you to retrace your steps (**Figure 1.2**), particularly in the preclinical stage.

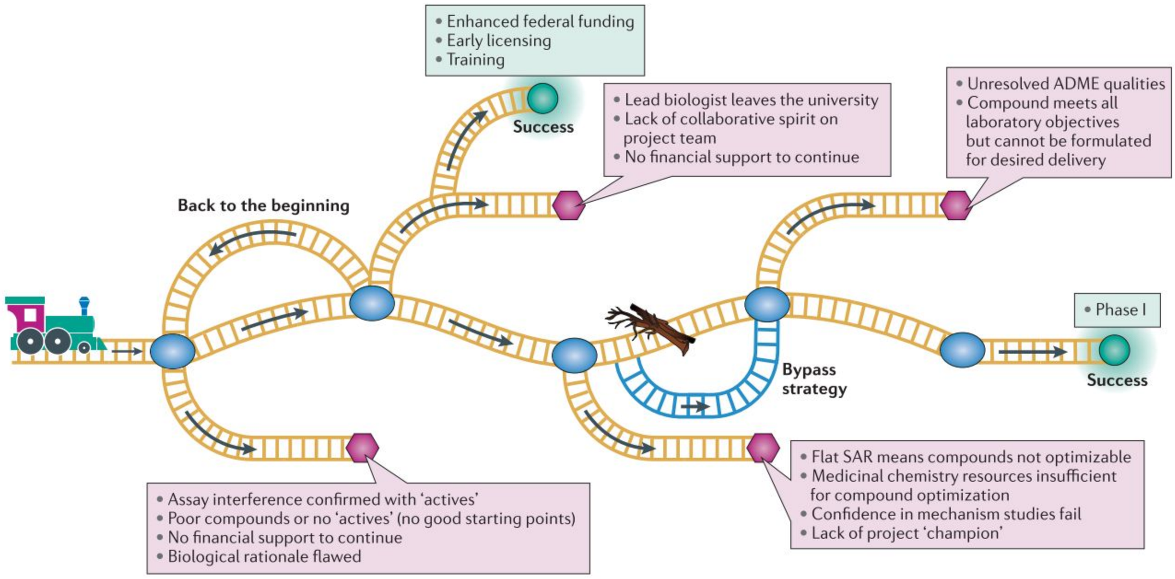

• Enhanced federal funding
• Early licensing
• Training
Success
• Lead biologist leaves the university
• Lack of collaborative spirit on project team
• No financial support to continue
• Unresolved ADME qualities
• Compound meets all laboratory objectives but cannot be formulated for desired delivery
Back to the beginning
• Phase I
Bypass strategy
Success
• Flat SAR means compounds not optimizable
• Medicinal chemistry resources insufficient for compound optimization
• Confidence in mechanism studies fail
• Lack of project 'champion'
• Assay interference confirmed with 'actives'
• Poor compounds or no 'actives' (no good starting points)
• No financial support to continue
• Biological rationale flawed

Figure 1.2: *The tortuous path of preclinical drug discovery. Reproduced from [11]*

The most expensive failures happen in the clinical stage. They could be a consequence of an inadequate design from the beginning or to the unpredictable nature of some biological responses. “Fail early, fail fast” is the mantra of the pharmaceutical industry to avoid loss of money and time on a project with a dead end.

In fact, it is estimated that the path from the first identification of a disease-related target to the release of a drug in the market lasts an average of 12 years [12] and costs more than $1 billion [13,14]. For this reason, the largest effort should be concentrated in the preclinical stage, as it has the responsibility to deliver an optimal candidate that facilitates development and minimises the risk of future failures. In consequence, advances and improvements in any step of the preclinic are extremely important.

It is possible to define 5 different phases in the preclinical stage, following the linear representation of **Figure 1.1**:

1) **Target Identification and Validation**: once a disease is recognized, the most important targets involved in the development of the disease are identified. By means of different techniques, the mechanisms of action in the cell and the biological pathway/s involved are described and it is proved that a perturbation on a selected target (e.g. by deletion, overexpression or modulation) can be beneficial for the condition.

2) **Hit Identification and Validation**: During this phase a series of compounds (hits) are identified through experimental (e.g. high-throughput screening) or virtual screening of libraries of molecules. The efficacy of the hits on the target are validated with different assays. Among them, biophysical assays like Differential Scanning Fluorimetry (DSF), Surface Plasmon Resonance (SPR), Isothermal Calorimetry (ITC), Nuclear Magnetic Resonance (NMR) or those based on fluorescence (e.g. FP, TR-FRET) are preferred to assess the binding. Confirmation of the binding pose by X-ray crystallography is ideal. Biochemical and biological assays (i.e. phenotypic assays) can also be performed.

3) **Hit-to-Lead**: The potency, selectivity or other particular properties of the validated hits are improved through iterative steps of chemical modifications and biological tests until a successful outcome is obtained.

4) **Lead Optimization**: The ADMET properties (Absorption, Distribution, Metabolism, Excretion and Toxicity) of the selected lead are improved to obtain a candidate drug.
5) **Preclinical Development**: formulation studies and more in-depth effectivity tests are carried out in vitro and in vivo. The resulting drug candidate will be filed for clinical trials in humans (clinical phases I, II and III).

The work described in this thesis can be related to the first 3 steps of the drug discovery process, namely target validation, hit identification and hit-to-lead.

In many of these steps, it is possible to use computer programs to facilitate some tasks, significantly reducing the preclinical time length and increasing the chances of obtaining a successful candidate drug. From its first appearance more than 30 years ago, this field of research is called Computer-Aided Drug Design (CADD).

1.1.3. Computer-aided Drug Design (Cadd)

In the early 1980s there was an increased interest in “designing drugs by computers”[15]. At the time it was believed that drugs could be designed atom by atom, but the initial hype led to disappointment. In the early 1990s the paradigm changed to “design as many molecules as possible and screen them” thanks to technologies like combinatorial chemistry and high-throughput screening.
From then, a series of different computational methodologies have emerged to help in different parts of the drug discovery process. Several of them have been used here:

- The first step to any drug design project is to be sure that a target can properly bind a small molecule and determine where (target “bindability” assessment). Mixed-solvent molecular dynamics and programs for cavity detection can be used for this purpose (*Section 3.3.1*).
- In Hit Identification and Validation, chemoinformatics methods can be applied for chemical space navigation (*Section 3.1*), docking programs can be used for screening a virtual library of compounds (*Section 3.2*), dynamic undocking can help remove false positives (*Section 3.3.2*) and molecular dynamics can validate the binding of the hits found (*Section 3.3*).
- In Hit-to-Lead, chemical space is explored with the aim to increase the activity of the hits. The most straightforward method is simply search molecules with similar structures in vendor databases (“SAR-by-catalogue”). With *de novo* drug design, novel molecules that are not present in any databases can be created. Free energy calculations can be used to predict the affinity of the hits (*Section 3.3.3*).

All the previous methods are applicable when a structure of the target is known (Structure-based drug design). Structures can be obtained experimentally by X-ray Crystallography, NMR or cryo-EM.
If ligands of the target are known, they can be used to guide the drug design (Ligand-based drug design). Similarity methods, Quantitative Structure-Activity Relationships (QSAR), including the modern machine-learning approaches, and pharmacophore-based approaches, belong to this class of CADD.

1.2. Principles Of Molecular Recognition In Protein-ligand Binding

When we administer a drug to treat a medical condition, we can see a visible effect of amelioration. This effect is the global result of the intricate voyage of the drug through the body. In this microscopic world, the active principle of the drug is crossing compartments and making countless interactions with many proteins, until it reaches the final destination, namely the therapeutic target for that medical condition. At that point, drug and target bind with each other, leading to an alteration of the target function and a biological response.
In the development of a drug, it is important to consider all these events, but the first and foremost is the binding with the target protein, without which the effect would not be present.

At the basis of any binding event, there are mechanisms of molecular recognition between a protein and a ligand. Molecular recognition refers to the complementarity of protein and ligand by matching electronic properties and shape [16]. This is a dynamic process where different events may occur and have different weights. We can, though, recognize three main structural factors that are important: the specific interactions, the shape and the flexibility of the system. These three structural factors affect enthalpy and entropy in different ways and the outcome of the combination of them will determine the free energy of binding and if a favourable interaction can happen.

1.2.1 Protein-ligand Interactions

If the chemical world offers infinite possibilities, the interactions in proteins correspond to only 20 basic units, the aminoacids. All the possible interactions that a protein can do with its counterpart, being a natural substrate, a natural product or a synthetic drug, are a mere recombination of those units. While the number of possible 3D arrangements and particular variations is still infinite, we can organize the molecular interactions in a few classes (**Table 1.1**). This knowledge helps understanding the properties that a drug must have to engage with the appointed protein.

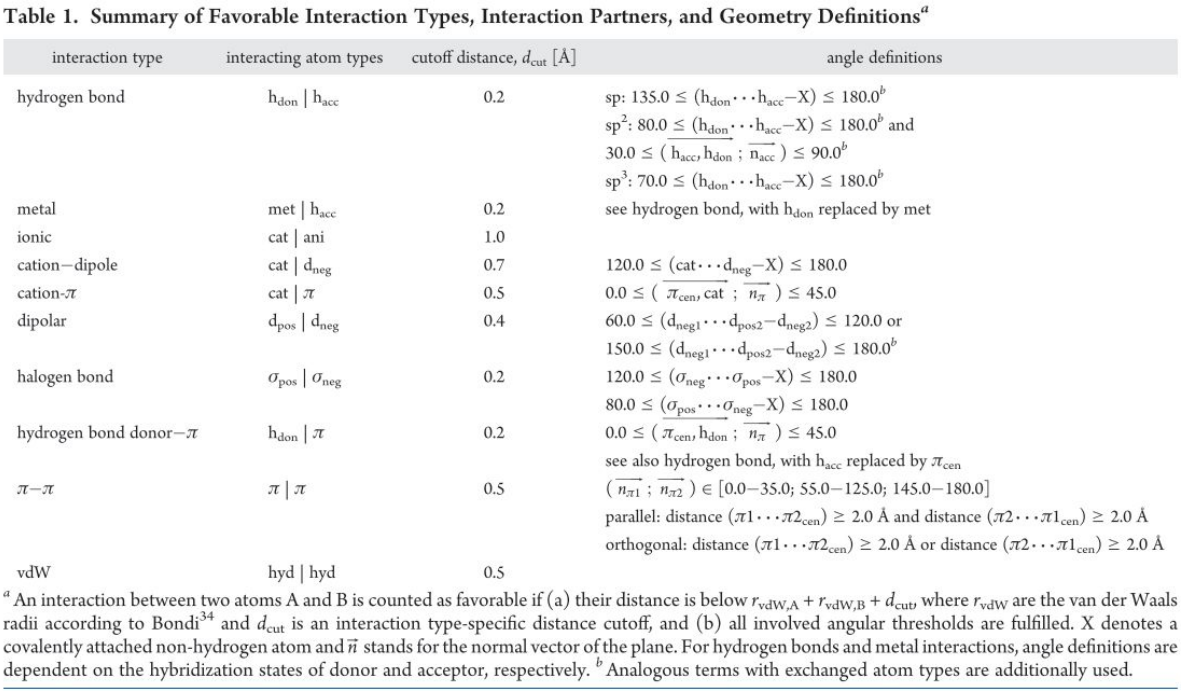


Table 1. Summary of Favorable Interaction Types, Interaction Partners, and Geometry Definitions[a]

interaction type	interacting atom types	cutoff distance, d_{cut} [Å]	angle definitions
hydrogen bond	$h_{don} \mid h_{acc}$	0.2	sp: $135.0 \leq (h_{don} \cdots h_{acc}-X) \leq 180.0$[b] sp^2: $80.0 \leq (h_{don} \cdots h_{acc}-X) \leq 180.0$[b] and $30.0 \leq (\overrightarrow{h_{acc}, h_{don}} ; \overrightarrow{n_{acc}}) \leq 90.0$[b] sp^3: $70.0 \leq (h_{don} \cdots h_{acc}-X) \leq 180.0$[b]
metal	$met \mid h_{acc}$	0.2	see hydrogen bond, with h_{don} replaced by met
ionic	cat | ani	1.0	
cation−dipole	$cat \mid d_{neg}$	0.7	$120.0 \leq (cat \cdots d_{neg}-X) \leq 180.0$
cation-π	$cat \mid \pi$	0.5	$0.0 \leq (\overrightarrow{\pi_{cen}, cat} ; \overrightarrow{n_{\pi}}) \leq 45.0$
dipolar	$d_{pos} \mid d_{neg}$	0.4	$60.0 \leq (d_{neg1} \cdots d_{pos2}-d_{neg2}) \leq 120.0$ or $150.0 \leq (d_{neg1} \cdots d_{pos2}-d_{neg2}) \leq 180.0$[b]
halogen bond	$\sigma_{pos} \mid \sigma_{neg}$	0.2	$120.0 \leq (\sigma_{neg} \cdots \sigma_{pos}-X) \leq 180.0$ $80.0 \leq (\sigma_{pos} \cdots \sigma_{neg}-X) \leq 180.0$
hydrogen bond donor−π	$h_{don} \mid \pi$	0.2	$0.0 \leq (\overrightarrow{\pi_{cen}, h_{don}} ; \overrightarrow{n_{\pi}}) \leq 45.0$ see also hydrogen bond, with h_{acc} replaced by π_{cen}
π−π	$\pi \mid \pi$	0.5	$(\overrightarrow{n_{\pi1}} ; \overrightarrow{n_{\pi2}}) \in [0.0-35.0; 55.0-125.0; 145.0-180.0]$ parallel: distance $(\pi1 \cdots \pi2_{cen}) \geq 2.0$ Å and distance $(\pi2 \cdots \pi1_{cen}) \geq 2.0$ Å orthogonal: distance $(\pi1 \cdots \pi2_{cen}) \geq 2.0$ Å or distance $(\pi2 \cdots \pi1_{cen}) \geq 2.0$ Å
vdW	hyd | hyd	0.5	

[a] An interaction between two atoms A and B is counted as favorable if (a) their distance is below $r_{vdW,A} + r_{vdW,B} + d_{cut}$, where r_{vdW} are the van der Waals radii according to Bondi[34] and d_{cut} is an interaction type-specific distance cutoff, and (b) all involved angular thresholds are fulfilled. X denotes a covalently attached non-hydrogen atom and $\vec{n}$ stands for the normal vector of the plane. For hydrogen bonds and metal interactions, angle definitions are dependent on the hybridization states of donor and acceptor, respectively. [b] Analogous terms with exchanged atom types are additionally used.

***Table 1.1:** Reproduced from "Rationalizing tight ligand binding through cooperative interaction networks. Bernd Kuhn, Julian E. Fuchs, Michael Reutlinger, Martin Stahl, and Neil R. Taylor. J. Chem. Inf. Model. 2011, 51, 3180–3198. dx.doi.org/10.1021/ci200319e" [17]*

Of the favourable interactions shown in **Table 1.1**, the most frequent non-covalent interactions occurring between a ligand and a protein are hydrophobic interactions, hydrogen bonds (standard or weak), π-stacking and salt-bridge interactions (**Figure 1.3**).

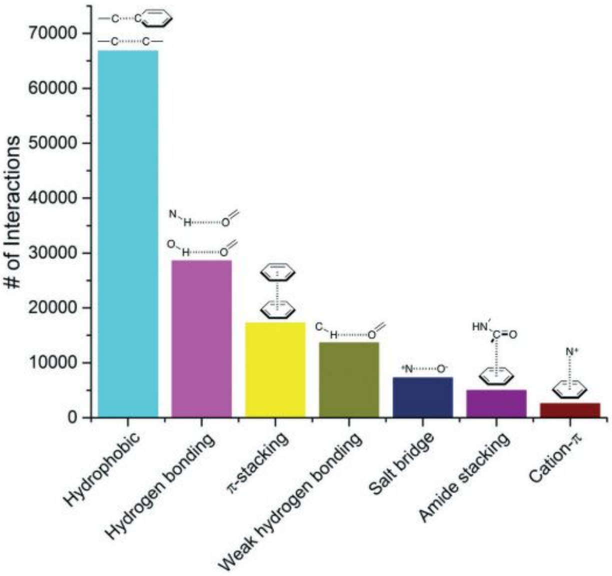


***Figure 1.3:** Frequency distribution of interactions observed in protein–ligands extracted from the PDB. Reproduced from Ferreira de Freitas, R. & Schapira, M. A systematic analysis of*

atomic protein–ligand interactions in the PDB. Med. Chem. Commun. 8, 1970–1981 (2017) [18]

Hydrogen bonds will be discussed more extensively in paragraph 1.2.1.1. Apart from oxygen and nitrogen of the backbone, aminoacids with hydrogen donors or acceptors in the side chain (Asparagine, Glutamine, Serine, Threonine, Tyrosine and weakly Cysteine) can make hydrogen bonds. Salt-bridge are charge-reinforced hydrogen bonds that occur between a positively charged nitrogen and a negatively charged oxygen [18] with a median distance of 2.79 Å [19]. Negatively (Aspartate and Glutamate) and positively (Arginine, Lysine, Histidine) charged aminoacids can make salt-bridges with the respective opposite charges present in the ligand.

Hydrophobic interactions can occur between aliphatic carbons, between aliphatic and aromatic carbons and between carbon and sulphur atoms with distances between 3.7 and 4.4 Å [19]. Aminoacids with hydrophobic side chains (Alanine, Leucine, Isoleucine, Methionine, Proline, Valine, Phenylalanine and Tryptophan) and carbon atoms in the backbone are involved in this type of interactions. π-stacking occurs between aromatic rings and can be considered a special case of hydrophobic interaction. Three types of geometries are possible (face-to-face, edge-to-face and parallel displaced) and distances can range from 3.4 to 3.8 Å [19]. Phenylalanine, Tyrosine, Tryptophan and Histidine can make this interaction with aromatic rings in the ligand.

The energetic yield that ligands can extract from the protein surface is unevenly distributed. Some residues (or interaction points), also called “binding hotspots”, are particularly important in drug design because they contribute the most to the free energy of binding [20]. In fact, in some cases, a small change in one substituent interacting in these areas can provide major differences in activity called “activity cliffs” [21].

The main interactions can be simplified to 6 pharmacophoric features: Hydrogen-Bond Donor (HBD), Hydrogen-Bond Acceptor (HBA), Anionic, Cationic, Hydrophobic and Aromatic. The binding site can be represented by a 3D ensemble of these features, or a “pharmacophore model” that should emphasize the binding hot spots. A pharmacophore can also be obtained from ligand-based methods, and the matching of the corresponding features in the two models implies an excellent understanding of the interaction properties of the target [22].

1.2.1.1 A Focus On Hydrogen Bonds

Hydrogen Bonds are the most important polar interactions involved in protein-ligand molecular recognition. They follow strict geometric rules with the distance between the hydrogen bond donor and acceptor in the range of 2.5 and 3.5 Å and an angle very close to linearity.

They are particularly important hotspots and can contribute to the binding energy up to 1.8 kcal/mol for one hydrogen bond [23], but there is a huge variability that depends on the environment around the hydrogen bond. In particular, it has been calculated that in hydrophobic environments, which is usual in a druggable cavity [24], the contribution to the free energy of binding is 1.2 kcal/mol more than in other environments [25].

It has been found that hydrogen bonds are especially important for structural stability of the complex, with water-shielded hydrogen bonds acting as kinetic traps [26], slowing down the release of the ligand (k_{off}) from the complex. Strong hydrogen bonds and clusters of hydrogen bonds are usually robust anchors for the ligands [27]. For example, the conserved Asparagine in most bromodomains is making an important anchoring hydrogen bond inside the cavity. For these reasons, it is common to use hydrogen bonds as reference points (or pharmacophoric restraints) in VS/SBDD studies.

Hydrogen bonds can also occur between the ligand and water molecules present in the cavity and this is particularly important when structural waters play a major role.

1.2.1.2 The Role Of Structural Waters

Structural waters are, by definition, water molecules that have an important role in the stabilization of the tridimensional structure of proteins. Those waters are frequently found in crystal structures of proteins and can be difficult to displace by ligands. They can also play a role in the recognition of substrates and ligands [28].

A single water molecule can make up to 4 hydrogen bonds with other molecules, 2 as a hydrogen donor (through the hydrogens) and 2 as a hydrogen acceptor (through the lone pairs of the oxygen). This number of interactions is found in the structure of ice. In bulk water, the average number of hydrogen bonds per water molecule is lower than the theoretical maximum (between 2 and 4 depending on the experimental technique [29,30]) because water molecules can interact dynamically in different ways.

When interacting with a solute, they have to rearrange and this can cause an increase or loss of free energy. The role of water in solution is of major importance for the thermodynamics of protein-ligand binding. This is usually called the “solvation effect”. The hydrophobic effect is the most fundamental solvation effect and can be captured with a statistical approach (i.e. 25 cal/Å^2 for apolar surfaces)[31], but aqueous solvation causes many more effects that are difficult to predict. For example, Klebe *et al.* discovered that waters present in the first layer of solvation of ligand-protein complexes can have a role in the energy of binding [32]. Displacing water molecules that occupy unfavourable positions (colloquially referred to as “unhappy waters”) can yield significant binding free energy, but predicting this effect is not easy [33].

In a study it has been discovered that structural waters can create networks of different polygon shapes [34], in the same way they do in solution. The interactions that they make in these polygons can affect the free energy of the water molecules and can determine the structural stability of networks of conserved waters in proteins. Restrained water molecules can behave like ice molecules and make highly favourable interactions [35]. The breaking of this interaction by a ligand can thus create a loss of enthalpy that is not compensated by the increase of entropy for the liberation of water molecules in the bulk. For this reason, the interaction with the ligand is not favoured and only ligands that do not disrupt this perfect network are accepted, like hydrophobic ligands.

In *section 1.4.1* I will discuss the specific case of structural water molecules in BRD4 and how they affect ligand binding.

1.2.2 Ligand Architecture

As explained above, not all the interactions that a protein can make with a given ligand have the same importance. Some interactions drive the binding because they contribute disproportionately to the binding free energy. Without these key interactions, it is not possible to form tight protein-ligand complexes. For this reason, when searching for new ligands, it is imperative to satisfy these key interactions. Then, the ligand must branch out to form secondary interactions that improve overall complementarity and binding free energy. Optimising these interactions is important for potency and selectivity of the ligand. The "ligand architecture", understood as a molecular framework that enables the positioning of interacting atoms in the correct position to form optimal interactions with the protein, plays a major role in ligand optimization.

On one hand, the size and shape of the ligand should be adequate to fit the cavity, but on the other hand, the spatial distribution of the different interactions is as important as the interactions themselves. The most active compound should both have adequate shape and match precisely the geometric arrangement of features in the cavity (**Figure 1.4 A**). A ligand with perfect shape but that does not match the features of the protein (**Figure 1.4 B**) is as inactive as a ligand with wrong shape (**Figure 1.4 D**). The shape a ligand must have does not have to be unique for a specific cavity but different options are usually possible if the feature arrangement is maintained (**Figure 1.4 C**).

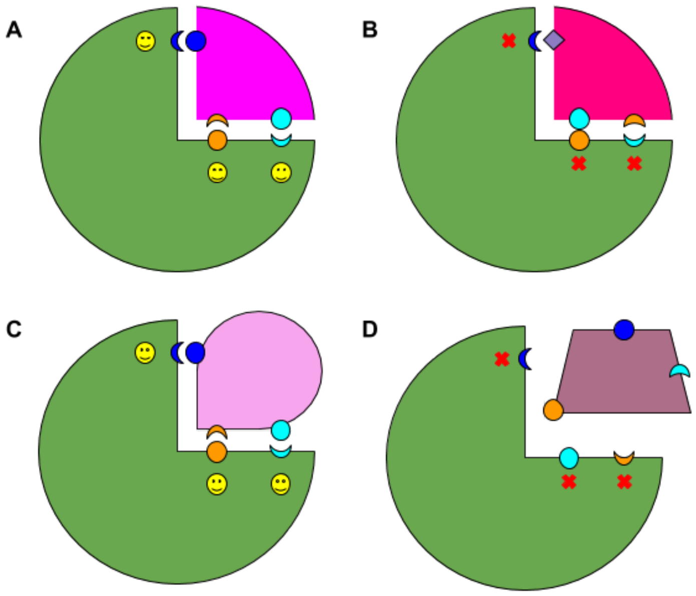


Figure 1.4: *Importance of scaffolds in molecular recognition. A) Perfect match B) Ligand shape fits perfectly in the cavity but the spatial distribution of the features doesn't match the features in the cavity. C) Scaffold hopping: even though the scaffold is not a perfect match, it still fits the cavity and provides an adequate spatial distribution of the features that can match the ones of the cavity. D) complete mismatch of features and shape doesn't fit in the cavity.*

In drug discovery the shape of the ligand and the position of the features is largely dominated by its scaffold. This is the “core structure” of the ligand that serves as a skeleton from where the substituents that are making the interactions are “hanging”.

Scaffolds are extremely important in drug discovery. They largely dictate the synthetic procedures and the type of structural modifications that can be introduced with relative ease. This has a major impact on the investigation of Structure-Activity Relationships (SAR), i.e. the effect of structural changes on the activity. In turn, it also affects the possibility of obtaining a New Chemical Entity (NCE) that belongs to a novel “chemotype”. Chemical novelty is the basis for intellectual property in the actual pharmaceutical regulation. Without a solid intellectual property position, usually in the form of a patent, it is not possible to fund the clinical studies that lead to the market. Specifically, pharmaceutical patents describe a Markush formula that defines the essential scaffold and R-groups needed for activity. Furthermore, the scaffold also dominates *in vivo* off-target effects, and a change of scaffold (or re-using a known scaffold for a different target) may be necessary to attain adequate ADMET properties.

The formal definition of scaffold in computational drug design was given by Bemis and Murcko (BM scaffolds) 20 years ago [36]. It consists in the removal of all substituents until only ring systems and frameworks (rings connected by a linker) are obtained (**Figure 1.5**). The concept of BM scaffold has been further extended to generate hierarchies of scaffolds [37].

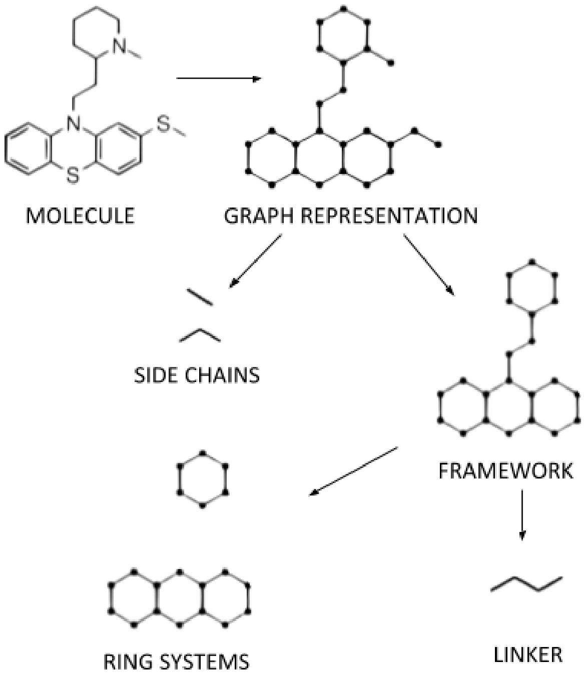


Figure 1.5: *BM scaffolds. Adaptation from Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. 1996, 39, 2887−2893. [36]*

If a ligand for a target is already known, a change in the scaffold while maintaining the correct spatial distribution of features (**Figure 1.4 C**) or “Scaffold Hopping” can be beneficial for many reasons like facilitate the synthesis, simplify a complex natural product, improve some pharmacokinetics properties (e.g. metabolism) of active molecules, increase selectivity or circumvent pre-existing intellectual property [38].

From an analysis in 2014 [39], it was detected that, from all the drugs in the market until 2013, the number of unique framework and ring systems were only 1197 and 351, respectively. Considering that the number of carbon-based ring systems up to 14 atoms has been estimated to be 916,130 [40], it is clear that there are still plenty of possibilities for scaffold exploration and scaffold hopping.

1.2.3. Protein Flexibility

Protein flexibility also plays an important role in the mechanisms of molecular recognition. Proteins are not static bodies but rather dynamic, constantly rearranging their structure in response to their local environment and to exert their intrinsic functions [41,42].

The old model of Lock and Key (where a rigid cavity matches a rigid ligand) has been already overcome by two other models (**Figure 1.6**): conformational selection and induced fit [43,44].

In conformational selection, it is hypothesized that the ligand preferably binds one conformer from an ensemble, stabilizing said conformation and shifting the equilibrium of the ensemble. In induced fit, the cavity undergoes a reorganization and opening determined by the arrival of the ligand so that binding can occur. There is still controversy on which is the correct model but it is mostly accepted that it is in fact a combination of both mechanisms.

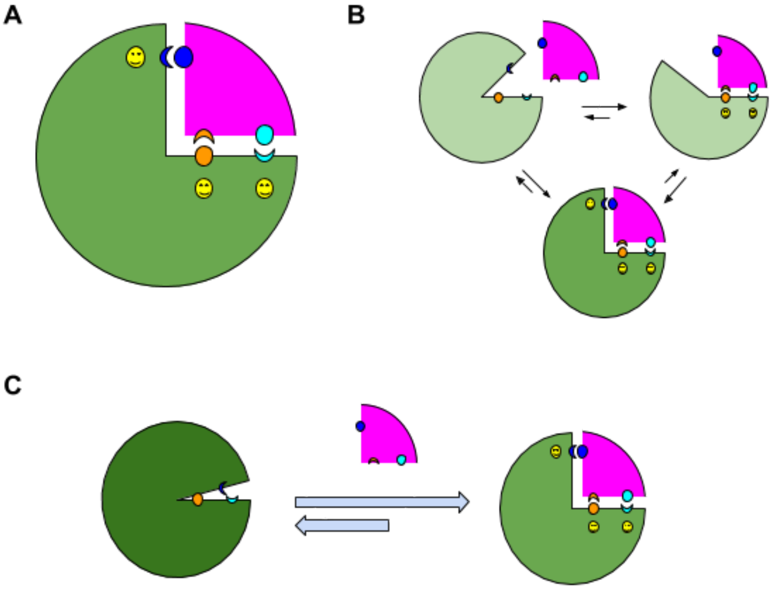


Figure 1.6: *Types of models of molecular recognition. A) Lock & Key B) Conformational selection: there is an equilibrium between different conformations of the protein, the ligand selects the one that matches, shifting the equilibrium towards that conformations. C) Induced fit: the ligand adapts the cavity conformation upon arrival/union.*

Programs for cavity detection or for binding assessment (like docking programs) that consider the protein rigid can often incur in errors due to this simplification. Methods that take into account protein flexibility (e.g. molecular dynamics) are more accurate but usually more computationally expensive. Nonetheless, a proper exploration of the protein flexibility should be considered mandatory before starting any drug discovery project. This can help assess the risk of considering the protein as a rigid body, and open other options, for example using an ensemble of structures taken from different crystal structures or from snapshots of a molecular dynamics simulation [45].

But if protein flexibility poses difficulties for structure-based drug design, it also brings about substantial opportunities. When dealing with proteins that lack a suitably druggable binding site, conformational changes may open cavities that remain hidden in the basal (experimentally observed) structure. These hidden cavities, called “cryptic pockets", can bind ligands and modulate the activity of the protein. Several examples are known [46,47].

Proteins may contain, apart from the main functional binding site, additional or allosteric binding sites. These sites are involved in other important functions of the protein such as signaling in cellular pathways, auto-regulation, protein-protein interactions and complex formation [48,49]. A binding event in the allosteric site affects the activity on the main (or another) site through a conformational change. Targeting allosteric binding site can expand the “druggable” genome in case of proteins previously considered undruggable [50]. Particularly interesting is the case of hidden (cryptic) allosteric binding sites, as they may enable modulation of difficult targets whose main catalytic site is not druggable and they lack any visible allosteric site [51].

This type of cavities cannot be easily detected both experimentally and computationally. Fragment screening (*section 1.3.3.1*) can help identify these pockets. Computational methods such as Molecular dynamics might fail to explore the conformational landscape and more specialized techniques are needed [47]. One technique that has become useful in this problem is Mixed-solvent molecular dynamics (*section 3.3.1*).

1.3 Finding Binders

1.3.1 High-throughput Screening

With the genomic revolution in the 1990s, there was an acceleration on the discovery of novel targets and new diseases. The rational drug design approach that was used until then, where drugs were created starting from natural products, could not work for targets without a known structure or whose natural ligands are unknown [52]. At that time high-throughput screening (HTS) started to increase its application, helped also by the advent of combinatorial chemistry and faster assay technologies. The focus was on screening larger and larger libraries as fast as possible, with the promise of obtaining NCEs in a really short time. But the expectations were frustrated when it was realized that it was very difficult, if not impossible, to discover a ready-to-use drug directly from HTS. Instead, HTS hits also needed long cycles of drug discovery and development [52,53]. It was also noticed that the success rates of the screening were different according to the target class with no hits at all for some targets, considering this an evidence for “undruggability” of the whole class. Drug discovery practitioners started to consider the idea that these results were instead a consequence of the library design and that maybe the target was not undruggable but the library lacked the right compounds[52]. Then, the focus shifted on more properly designed libraries with higher diversity of compounds to increase the probability to find the correct one and increase the success rate. However, diversity should be balanced with other factors, like novelty, intellectual property potential, synthetic vectors, price and target focus [53]. Combinatorial chemistry relies on a few reactions and, even though they can produce large numbers of compounds, their diversity is limited [53]. Diversity-oriented synthesis approach can be used, but an upfront cost for the synthesis should be considered [54].

In spite of these limitations, nowadays high-throughput screening (HTS) is still the gold standard in pharmaceutical industry to find new binders.[55] The automatization of HTS by means of robots has led to the possibility to screen more than 100.000 compounds per day (ultra-high-throughput screening, uHTS) [56]. Big pharma companies can screen in-house libraries of millions compounds in a few days. Nonetheless, increasing the number of compounds present in the libraries means increasing costs for purchase, storage and screening. For this reason, the size of a library normally wouldn’t exceed 3 million compounds [53], and only few organizations, apart from “Big pharma” companies, can afford it. [57]

To solve these issues, two options are available:

- “Going larger” by expanding the chemical space. DNA-encoded libraries are one possibility.[58] But navigation of virtual screening collections by computational means offers a more systematic approach (*section 1.3.2*)
- “Going smaller” by using Fragment Screening and Fragment-based Drug Design, where computation also plays a prominent role (*section 1.3.3*)

1.3.2 Chemical Space Navigation And Virtual Screening

Libraries from pharma companies are not usually accessible by outside scientists. For this reason, many chemical vendors offer selected libraries of purchasable compounds for screening [57]. However, the number of molecules included in these libraries rarely is more than 2 million compounds. Even combining them in databases of commercially-available compounds like ZINC [59] the maximum number achieved for unique readily available purchasable compounds is about 17 million (in March 2017[57]).

A seminal paper from Bohacek estimated that the size of the Chemical Space (the set of all possible enumerated compounds according to some property rules) of compounds with drug-like features is 10^{60} [60]. A more recent paper estimates that it can be much less, around 10^{33} [61]. In any case, it is an unmanageable number of compounds, more than the stars in the Universe.

Even though the screening libraries contain in the region of 10^6 compounds, thus covering an incredibly small part of the drug-like chemical space, the intention is that they should maximise coverage, with the inclusion of diverse compounds to represent less-explored areas of chemical space. The hope is that these molecules can make enough suboptimal interactions that, as a whole, can still display good affinity, and this initial hit, even if imperfect, can be optimized in subsequent phases of the drug development process. But a good outcome is not always achievable if the starting molecule is not the most appropriate. Moreover, for targets deemed undruggable the compounds available in the libraries could not be good enough to obtain hits. For these reasons, it is important to expand the accessible chemical space to discover unexplored areas.

The bottom-up systematic virtual enumeration of all the compounds was initiated by Reymond's group [62] using a topology-based method (unlike the reaction-based method used by combinatorial libraries until then), but the exponential combinatorial explosion of new compounds for each heavy atom added forced them to stop enumerating at 17 heavy atoms [62d]. In order to enumerate compounds with increased numbers of heavy atoms (more "drug-like"), the recent trends have shifted again to a reaction-based focus, virtually assembling compounds from commercial building blocks using a set of trusted reactions. Therefore, these "virtual libraries" contain compounds with "high probability of being synthesized", but that don't physically exist and they will be only synthesized when they are purchased [63,64]. The most comprehensive case, which is Merck MASSIV library, contains 10^{20} molecules, that is still far from the estimated "drug-like chemical space". This can be due to prediction limitations or to a lack of new reaction rules. However, this is still a proprietary database and it is not available to everyone. The most comprehensive available library is ENAMINE REAL with 1.36×10^{9} enumerated compounds and ENAMINE REAL SPACE with 15.5×10^{9} virtually synthesizable compounds.[65] Even though this library can theoretically be entirely purchased, physically screening this enormous number of molecules is a Herculean task that is not profitable. Besides, it is supposed that only a small percentage of the molecules contained in these libraries would be active on the target of study. For this reason, computational methods can be applied to filter out inactive compounds and reduce the number of molecules to eventually test experimentally. This option is especially important for "possible synthesizable compounds", to avoid a costly and time-consuming synthesis of inactive molecules [66].

In this context, a very important computational technique is Virtual Screening (VS). It consists in predicting the binding and affinity of molecules by means of computational tools, like a machine learning classifier or a docking program. Standard VS methods (docking-based) can be applied to the scale of 10^{9} [67,68], but with larger collections, these methods face other limitations like the calculation time, disk space or bandwidth, that depends on the capability of the research lab and is impractical even for supercomputing centres. Adapted VS protocols could be introduced in these cases. This will be further discussed in the "Docking" section (*3.2*).

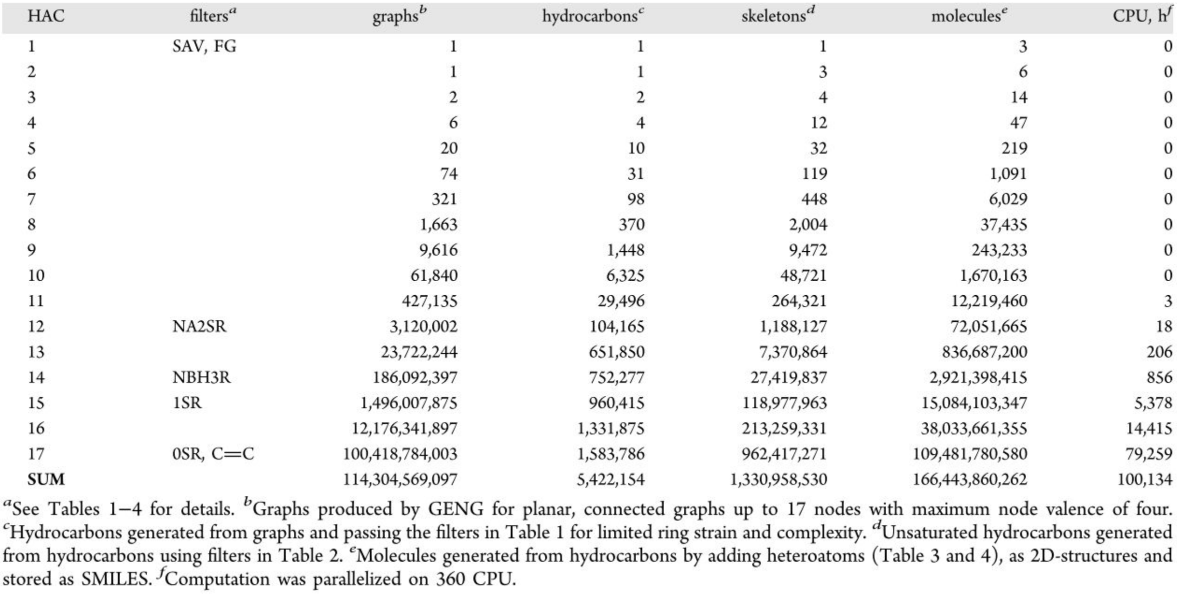
1.3.3 Fragment-based Drug Discovery

The other solution to the problem of increasing chemical diversity without having to deal with combinatorial explosion is to decrease the molecular size of the compounds in the libraries. The idea that it is possible to partition the whole energy of binding into some discrete contribution given by structural building blocks was stated in 1981 by Jencks [69]. In a seminal paper of Hann et al [16] it was demonstrated that as molecular complexity increases, the chances to see a binding event of a specific type decrease. Smaller compounds, called "fragments", contain a limited number of interaction points and less rotatable bonds which make them suitable for more efficient interactions leading to perfect match in subpockets or smaller parts of the cavity. When combined through a linking or merging approach, this would lead to improved affinity and selectivity [70]. On the contrary, bigger molecules such as standard HTS hits contain interactions that are suboptimal due to the increased molecular complexity that increases the probability of mismatch [16].

According to the Rule-of-Three (Ro3) [71], inspired by Lipinski's Rule of Five (Ro5) for drug-like molecules [7], it is considered that a fragment is a molecule with molecular weight between 150 and 300 Da, a number of hydrogen bond donors ≤3, a number of hydrogen bond acceptors ≤3 and ClogP ≤3.

1.3.3.1 Fragment Screening

The use of fragments in a screening campaign can offer several advantages over a standard HTS. Of note, fragments can cover a bigger part of the corresponding chemical space. If we consider GDB-17 [62d] as the closer equivalent of the fragment chemical space, we can conclude that the typical fragment screening collection (ca. 1000 molecules) represents only a tiny fraction of the total number of theoretical molecules or even graphs (**Table 1.2**). Yet, we can also observe that the chemical space grows exponentially with the number of atoms. Thus, screening 1000-fold more compounds (10^6), as done in HTS, does not even start to compensate the gigantic increase in size of the drug-like chemical space (≤35 heavy atoms) relative to the fragment space (≤22 heavy atoms).



HAC	filters[a]	graphs[b]	hydrocarbons[c]	skeletons[d]	molecules[e]	CPU, h[f]
1	SAV, FG	1	1	1	3	0
2		1	1	3	6	0
3		2	2	4	14	0
4		6	4	12	47	0
5		20	10	32	219	0
6		74	31	119	1,091	0
7		321	98	448	6,029	0
8		1,663	370	2,004	37,435	0
9		9,616	1,448	9,472	243,233	0
10		61,840	6,325	48,721	1,670,163	0
11		427,135	29,496	264,321	12,219,460	3
12	NA2SR	3,120,002	104,165	1,188,127	72,051,665	18
13		23,722,244	651,850	7,370,864	836,687,200	206
14	NBH3R	186,092,397	752,277	27,419,837	2,921,398,415	856
15	1SR	1,496,007,875	960,415	118,977,963	15,084,103,347	5,378
16		12,176,341,897	1,331,875	213,259,331	38,033,661,355	14,415
17	0SR, C=C	100,418,784,003	1,583,786	962,417,271	109,481,780,580	79,259
SUM		114,304,569,097	5,422,154	1,330,958,530	166,443,860,262	100,134

[a] See Tables 1−4 for details. [b] Graphs produced by GENG for planar, connected graphs up to 17 nodes with maximum node valence of four. [c] Hydrocarbons generated from graphs and passing the filters in Table 1 for limited ring strain and complexity. [d] Unsaturated hydrocarbons generated from hydrocarbons using filters in Table 2. [e] Molecules generated from hydrocarbons by adding heteroatoms (Table 3 and 4), as 2D-structures and stored as SMILES. [f] Computation was parallelized on 360 CPU.

Table 1.2: *Statistics of GDB17. Reproduced from [62d].*

A second factor in favour of fragments is their promiscuity, resulting in hit rates of 1% to 10% [72] meaning that the method can identify meaningful starting points for almost any target. In fact, fragment screening can help identify allosteric or cryptic pockets in difficult targets, previously considered undruggable or help develop small inhibitors for PPI. Indeed, the average protein presents 2.2 fragment binding sites. [73]

But, obviously, fragment hits should be seen as probes highlighting privileged areas of chemical space, from where a deeper exploration can be applied. Indeed, since a fragment can only satisfy a limited number of interactions with the protein, the binding affinities are in the high micromolar or millimolar range [74]. More sensitive techniques are thus needed to detect the binding of fragments. In 1996 Abbott Laboratories were the first to use biophysical techniques for this purpose with their SAR-by-NMR [75], where it was possible to identify the weak binding of fragments to a target with nuclear magnetic resonance. Nowadays the methods of choice for Fragment Screening are biophysical techniques like DSF, SPR, FRET, ITC, NMR and XRAY. They display different sensitivity and throughput levels. For this reason, often a cascade of screening assays is considered. In a primary assay for example DSF, FRET and SPR have higher throughput, while ITC and NMR can be used as secondary assays to confirm the binding. Generally, X-ray is the last but essential step of the cascade. It gives very valuable information on the binding mode of the fragment.

Another important consideration is the efficiency of binding. The average atomic contribution to the binding free energy, called Ligand Efficiency (LE) is defined as follows:

$\Delta G_{bind} = -RT\ln K_d$

$LE = \Delta G_{bind}/N_{non\text{-}hydrogen\ atoms}$

It has been observed that LE has a maximum of -1.5kcal/mol,[76] and tends to decrease as the size of the ligand increases.[77] For this reason, fragment hits are more efficient than HTS hits. This metric is used to compare the fragment hits, but also to guide the fragment evolution process, with the aim that fragment-derived drugs are smaller than the HTS counterparts.

There are also some limitations of FBDD. First, it is not possible to apply the method when the target is not known (for example in cellular assays)[54]. Second, structural data are mandatory for the fragment hits because the following “fragment-to-lead” process cannot occur without knowing how the fragment binds. Third, combining or optimizing these fragment hits into leads can be cumbersome in some cases and computational tools are needed.

1.3.3.2 Computational Approaches For Fragment-to-lead

If various fragments are bound to the same cavity but in different subpockets, or if a known inhibitor binds differently, linking or merging approaches, where parts of various molecules are joined together can be an efficient way of jumping in potency [78]. The main limitation is how to combine these ligands preserving the optimal orientation of each component. A different strategy, known as fragment growing, consists of sequential additions of moieties to the initial fragment to capture additional interactions. This “Fragment-to-Lead (F2L)” approach must be strictly supervised by measuring LE, which should not decrease too much during growing.

F2L can be quite time consuming for medicinal chemists if no other tool can come to help. For this reason, in the last 15 years, several groups have developed computational tools especially designed for fragment growing and optimization and it has increased the number of fragment-to-lead applications [79].

Considering that, we have developed a computational platform that aims to exploit the available chemical space around a particular fragment hit, exploring non-obvious scaffold hops that may confer advantages in terms of potency, novelty, synthetic feasibility or accessible growth vectors. In order to facilitate experimental follow up, we only navigate commercial chemical space either in stock (from ZINC database [59]) or on demand (e.g. Enamine Real or Real Space[65]). In *section 4.1.1* I will present the development and application of this computational platform.

Another option for F2L is to create new chemical space. *De novo* drug design methods are a useful computational tool in this endeavour. Born some 30 years ago, these programs generate novel molecules from an initial ligand whose binding mode is known. Generally, the binding mode of the fragment is determined experimentally, but when that is not the case, docking programs or molecular dynamics simulations [80] can help place the fragment in the cavity. The target structure is used as a template to build the molecules directly inside the cavity, calculating the affinity on the fly. Scoring functions can be force field [81], empirical [82], or knowledge-based [83]. Sometimes, ligand-based restraints can also be applied. In particular, if active ligands are known, they can be used to generate 2D [84] or 3D pharmacophores [85] that are compared to the novel molecules to maintain some important features.

Two main classes of *de novo* programs can be considered, depending on how they grow the initial molecule: Atom-based or Fragment-based. The first programs of DND were using mainly the atom-by-atom approach, but even though it has advantages such as the theoretically infinite possible molecules that can be generated and the fine-tuning due to the small step size [86], the big disadvantages are that the chemical space increases enormously and that many of the solutions are impossible to synthesize. The last atom-based program was published in 1996. It was clear that different building blocks had to be used. For this reason, all recent programs are Fragment-based. Fragments can include anything, from atoms to functional groups or rings. The fragments can be generated by breaking down known drugs [87] or by using commercial building blocks [85]. The cleavage of the molecules can be done at single bond [88] or with retrosynthetic rules [89]. Sometimes generated building blocks can contain chemical handles useful for the following recombinations. Likewise, different ways of adding portions are possible. The most straightforward is to join fragments by single bond, but some additional rules on how to join them are needed. For example, in FOG [90] and its fragment-focused successor OpenGrowth[91] the frequency of connections between fragments extracted by a database of known drugs are used to combine the building blocks. If the building blocks are generated by retrosynthetic cleavage, the same rules can be used to combine them [92]. Another option is "in silico synthesis" where a set of rules mimicking real reactions obtained from literature are used to join functionalized building blocks, usually from commercial collections [84,93]. This is particularly interesting, because in this way not only it is ensured that the resulting molecules are synthetically accessible, but also a synthetic route can be described.

Due to the combinatorial explosion problem, it is impossible to do an exhaustive search. Heuristic algorithms are usually a good compromise between optimal solutions and fast runtime [94]. Depth-first search (DFS) strategy retains only one of the possible solutions at each iteration, reducing the search space but it is not expected to be the most optimal solution. A breadth-first search (BFS) strategy, on the other hand, explores systematically all the possible solutions. A possible application is to first use DFS to reduce the search space and then apply BFS to the best solutions [83]. Monte Carlo search (random sampling) can be used [81], also in addition to a Metropolis criterion [95]. But, probably the most used and useful algorithms are the evolutionary ones [87, 92, 93, 96, 97]. Candidate compounds are generated in a first run. Those compounds are then evaluated and the best ones are considered parents of a next generation of compounds. Small changes to a parent molecule ("mutations") or recombination between parents ("crossover") are creating children molecules, which in turns will be selected to be parents in another iteration. The process is repeated for a number of iterations or until a condition is reached [98]. A problem of stochastic algorithms is that two runs can give different solutions, with each solution not being the most optimal to the problem of interest.

To help further reduce the search space, secondary constraints can be considered. Properties like the Lipinski rules, aqueous solubility, ADMET properties, or synthetic accessibility influence the clinical development and should be calculated for the candidate molecules either after the runs, or even better, during the search. Scoring considering both the primary and secondary constraints can be included with multi-objective optimization (MOO) tools [97,99]

In particular, one of the most important properties that has to be considered is synthetic accessibility. Novel molecules that cannot be synthesized are of no use. It is possible to assess synthetic accessibility of the candidate molecules after the runs according to different scores (for example: SYLVIA [100]). However, the most useful way is to consider synthetic accessibility during the building runs, to help guide the search. Programs can consider synthetic accessibility in an implicit or explicit way. The implicit way uses knowledge from drugs, like the frequency of fragment connections [90], to ensure that the resulting molecules are synthesizable. Another way to implicitly consider synthesizability that is gaining ground in the last years is the use of Deep Neural Network. If the model is trained on known drugs and fine-tuned to compounds active to the target of interest, the generated molecules not only are active to the target but also synthesizable. The model implicitly learns the synthetic rules from the training set and there is no need to explicitly provide them. [101]

The explicit way makes use of rules to add the building blocks. TOPAS uses the RECAP rules to retrosynthetically break molecules and uses the same rules to join them [87]. Other methods use "in silico synthesis" to replicate the synthesis process by adding building blocks with real reaction rules [84,93]. Programs that perform virtual synthetic reactions [102] as a growing strategy include Autocouple [103], DOTS [104], PINGUI [105] and NAOMInext [106]. Those programs not only suggest new molecules to synthesize based on binding prediction, but also the possible synthetic reactions that can be performed to obtain those molecules.

An explanation of NAOMInext program is in *section 3.6*. An application of this program to grow a fragment hit into a potent ligand is presented in *section 4.1.3*.

1.4. Application To Therapeutic Targets

1.4.1. BRD4

In 1992 *brahma* protein was identified in Drosophila [107]. From then, more and more analogue proteins began to appear in yeast, mice and human [108, 109], and from the name of this first protein, the family of bromodomains was identified. They were first discovered as an independent domain in HAT (Histone Acetyl Transferase) proteins like P/CAF [110] that are important proteins in transferring acetyl-coA to lysines in Histone Tails. Nonetheless bromodomains don't catalyze this reaction. In 1999 it was recognized that bromodomains are acetyl lysine "readers"[111]: they bind selectively to acetyl-lysine and can distinguish from not acetylated lysine [110]. They also bind acetylated small molecules like acetyl-histamine [110].

There are 61 known bromodomains present in 46 proteins in the human genome that can be grouped in 8 families (**Figure 1.8**) based on their structure similarity [112, 113] .

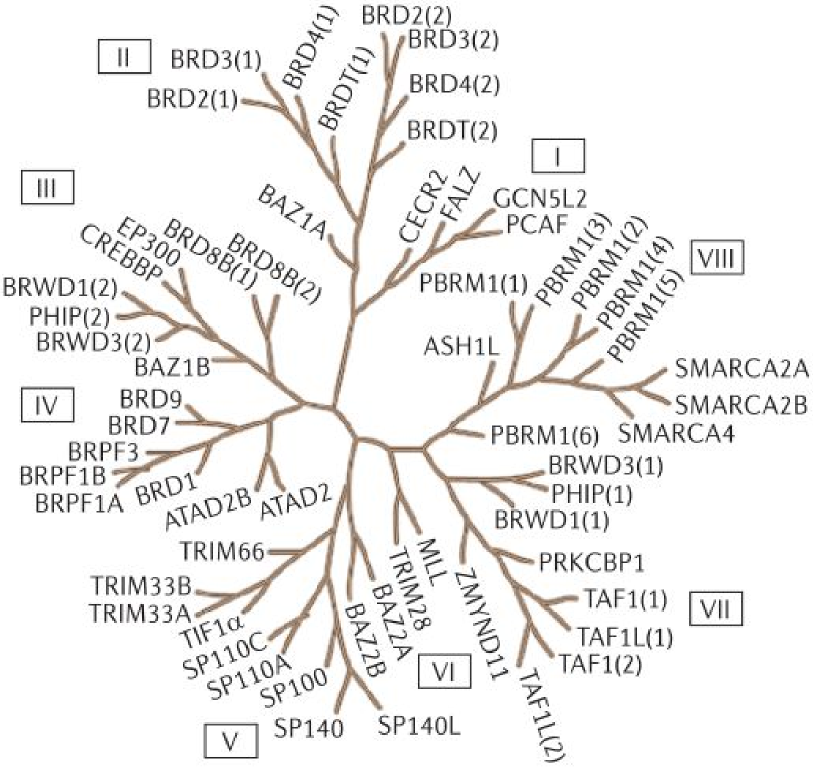


***Figure 1.8**: Phylogenetic tree of human bromodomains. Reproduced from [113]*

Bromodomains are important in epigenetics because they are selective readers of the histone code. In fact, in a study [112] it was demonstrated that each bromodomain recognizes (more or less selectively depending on the bromodomain) specific modification in Histone tails. These tails are lysine-rich and can show different patterns of acetylation (and other modifications like methylation, phosphorylations, etc…). Bromodomains anchor the acetyllysine of the histones tails but the flanking amino acids present in the sequence of the tails are what is recognized selectively by the different bromodomains.

Despite the high difference in sequence, the 3D structure is really conserved. It consists of 4 alpha-helices A, B, C, Z, connected by 2 loops ZA and BC (**Figure 1.9**).

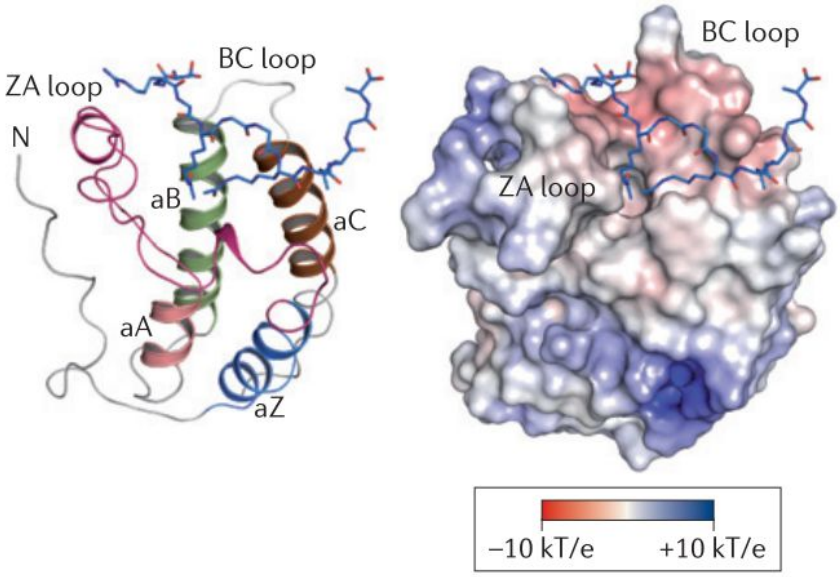


Figure 1.9*: Three-dimensional structure of bromodomains. Adapted from [113]*

The alpha-helices are more conserved across bromodomains than the loops. The histone peptides are laying in a cavity formed by the 2 loops and the lesser conservation in these points can explain the selectivity for different histone peptides in different bromodomains.

Almost all bromodomains contain a conserved asparagine in the BC loop that is in charge of anchoring the carbonyl of the acetyl lysine. A network of waters at the bottom of the cavity is well-conserved in all bromodomains. [114]

In 2010 the first potent ligands for bromodomains were identified (JQ1 [115], IBET [116]), in particular they were selective binders of the BET family. From that moment an increased number of bromodomain binding molecules appeared every year (reviews [113, 117, 118]), motivated by the growing discovery of bromodomain-disease associations and the interesting challenge of targeting epigenetics mechanisms. In the most recent years the efforts were committed to finding ligands selective for BD1 over BD2 and viceversa, and for other families of bromodomains over BET family.

Bromodomain-containing protein 4 (BRD4) is part of the BET family of bromodomains. BRD4 has a role in the regulation of genes and it is dysregulated in cancer [119]. It contains 2 bromodomains (BD1 and BD2) that are less similar. In the first bromodomain, ASN140 is the anchoring aminoacid for acetyl lysine and 7 structural waters are present at the bottom of the cavity.

After all the attempts to target BRD4 over the years, finding new chemotypes for BRD4(1) can be challenging, but the high number of possibilities in chemical space suggest that there can be many opportunities for scaffolds neglected in previous studies. As such, this system represents an ideal test bed for novel computational methods. In *section 4.1.1* I will present the application of a computational pipeline to mine chemical space and discover new BRD4 ligands.

All bromodomains contain a conserved network of structural water molecules at the bottom of the acetyllysine-binding cavity. Superimposing all the crystal structures of BRD4(1) present in the PDB, it is possible to recognize at least 7 water molecules in this site that are conserved. Ligands of BRD4 contain a methyl or similar substituent in contact with these water molecules, mimicking the methyl part of the acetyl lysine, with few exceptions [120]. Even though some efforts in SAR analysis have been made to study other substituents [121], none improves the activity. It is not clear why hydrophobic substituents are preferred in a polar environment. In order to understand this behaviour, we have carried out multiple computational studies and designed a series of substituents to the novel scaffold discovered for BRD4(1). These will be presented in *section 4.1.2*.

1.4.2 TREX2

During the life of a cell, DNA replication, repair and recombination are essential tasks. Several proteins participate in the process and, among those, 3’-5’-exonuclease have an important role in 3’-excision of impaired or damaged nucleotides. The 3’ terminus is then ready for next steps of DNA metabolism [122, 123]. 3’-5’ exonuclease activity has been recognized in several proteins. Human polymerases γ, δ and ε have an intrinsic proofreading 3’-5’ exonuclease activity that remove wrong nucleotides during DNA synthesis [124]. p53 also contains 3’-5’-exonuclease activity. Independent 3’-5’ exonucleases have also been found, in eukaryotic cells some examples are Mre11, WRN, RAD1, RAD9, APE1 and TREX1 and TREX2. They have a role of support for polymerases that do not have intrinsic proofreading activity and it was proved that they enhance accuracy of polymerases in error-prone conditions [125]. However, the diversity of proteins suggest that they may have more specific functions in the cells and they can help in different types of external genotoxic stress [124].

The major non-processive, autonomous 3’-5’ exonucleases in humans are TREX1 and TREX2 proteins. They are homologous proteins pertaining to the DnaQ-like family or DEDDh family (from the type of aminoacids in the catalytic site), that account also proteins in bacteria and yeast (**Figure 1.10**) [126].

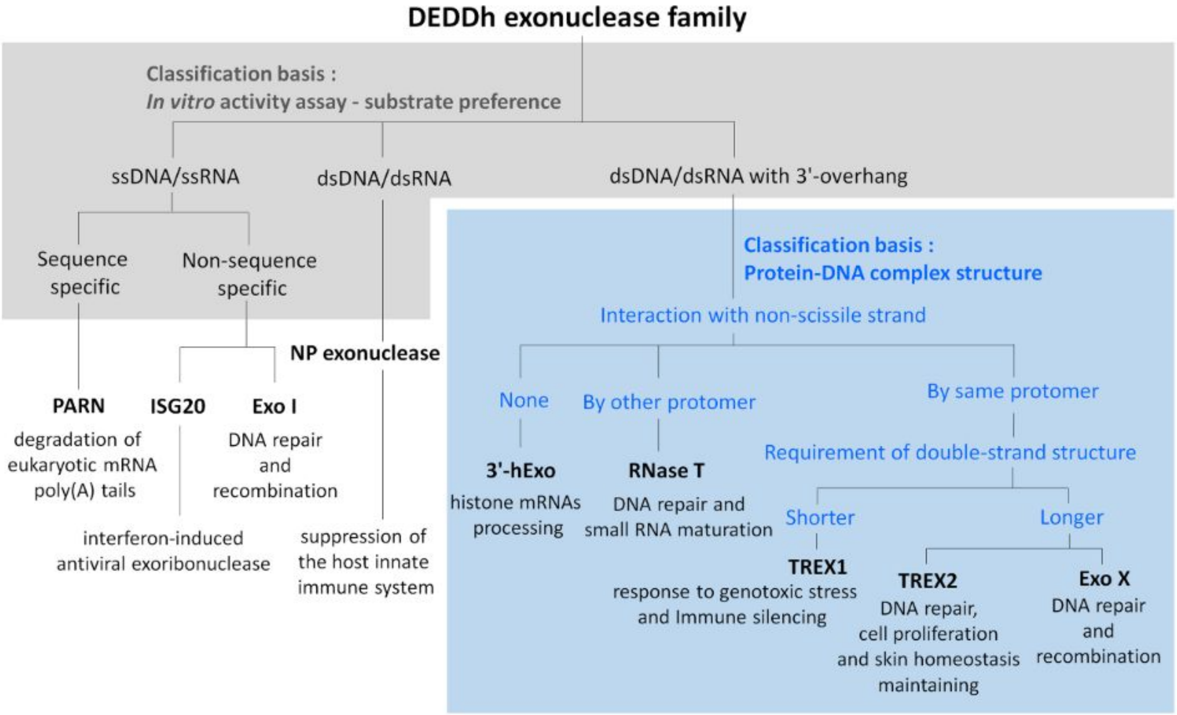

DEDDh exonuclease family
Classification basis :
In vitro activity assay - substrate preference
ssDNA/ssRNA
dsDNA/dsRNA
dsDNA/dsRNA with 3'-overhang
Sequence specific
Non-sequence specific
NP exonuclease
PARN
ISG20
Exo I
degradation of eukaryotic mRNA poly(A) tails
DNA repair and recombination
interferon-induced antiviral exoribonuclease
suppression of the host innate immune system
Classification basis :
Protein-DNA complex structure
Interaction with non-scissile strand
None
By other protomer
By same protomer
3'-hExo
histone mRNAs processing
RNase T
DNA repair and small RNA maturation
Requirement of double-strand structure
Shorter
Longer
TREX1
response to genotoxic stress and Immune silencing
TREX2
DNA repair, cell proliferation and skin homeostasis maintaining
Exo X
DNA repair and recombination

Figure 1.10: *DEDDh exonuclease family. TREX2 belongs to this family of proteins with a conserved catalytic site. Reproduced from [126]*

In this family the catalytic site contains 4 conserved anionic residues (aspartate and glutamate) that coordinate 2 metal ions (Mg2+) (**Figure 1.11**). In the TREX proteins a histidine at the catalytic site deprotonates a water molecule, creating the nucleophile that will attack the phosphodiester bond, provoking the hydrolysis [127, 128].

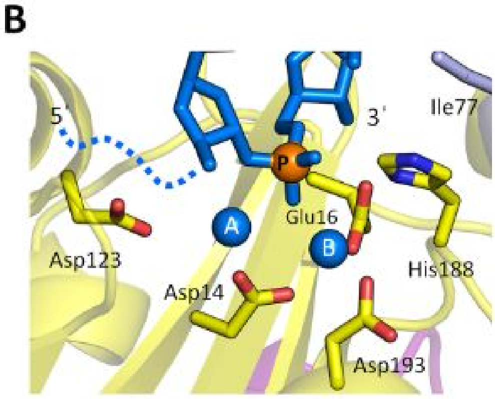


Figure 1.11: *The catalytic site of TREX2. Adapted from [126]*

TREX1 and TREX2 only share 40% of the sequence even though they have a similar structure that they share also with bacterial exonucleases. This suggests a similar evolution in the 3'-5' exonuclease function but a different specific cellular function for the 2 proteins [128]. The preferred substrate of both is DNA over RNA, and it can be single-stranded DNA (ssDNA) or double stranded DNA (dsDNA). Both of the proteins are homodimers, and the dimerization positions the 2 DNA binding sites on the same face of the dimer but in opposite edges [127].

In TREX2 three arginines in a mobile loop next to the catalytic site are responsible for DNA binding. The dimer shows cooperation in DNA binding and catalysis between the two protomers [129]. Expression studies of TREX2 in tissues and cell lines show a higher expression in squamous epithelial tissues like the skin [130]. In particular TREX2 is specifically expressed in keratinocytes [130]. This suggests a specific function in skin homeostasis.

Trex2 knockout mice are healthy and do not develop spontaneous tumors, but they are more susceptible to skin carcinogenesis with respect to wild type when they are exposed to carcinogenic chemicals or UVB light [130,131]. This suggests that TREX2 is involved in DNA repair mechanisms, degrading damaged DNA, that will affect keratinocytes apoptosis and activation of an inflammatory immune response to remove damaged cells, to avoid skin carcinogenesis [130,131,132]. This is confirmed by the discovery of polymorphisms and abnormal expression of TREX2 in samples of human squamous cell carcinomas [131].

It has also been found that TREX2 is highly expressed in psoriatic skin lesions with respect to normal skin [133]. In murine models of psoriasis, with imiquimod or IL-23 induced inflammation, it has been detected that trex2 knockout mice show a significant reduction of psoriatic inflammation signs such as erythema or epidermal thickness [131, 132]. Also, decreased transcription of inflammation-related genes, like IL-23 or tumor necrosis factor-alpha, and inhibition of keratinocytes apoptosis were observed. All these observations suggest that TREX2 has a pro-inflammatory role and can be a new therapeutic target for psoriasis.

Psoriasis is a chronic, noncommunicable, painful, disfiguring and disabling disease for which there is no cure [134]. It consists of skin lesions with a higher turnover of keratinocytes. It is also associated with comorbidities and psychological disorders. Current treatments only aim to control symptoms and usually are lifelong. There is no complete remission of the disease and relapses are common. Side effects, low adherence, non-responders and loss of efficacy of treatments are common issues. As a result, new, better and safer treatments are needed.

Finding a drug-like molecule that binds in the main catalytic site can be challenging, not only because it would have to compete with a much larger molecule (the DNA substrate), but also due to the magnesium ions present and the four negatively charged aminoacids, which would necessarily involve the development of a constitutively charged ligand [8]. Furthermore, the active site is conserved across similar proteins, and the development of selective inhibitors would be very difficult.

With our methods, [redacted] giving rise to the first TREX2 inhibitors ever described. This will be presented in *section 4.2*.

Chapter 2

Objectives

General Objectives

The main objective of this work is to extend the capabilities of state-of-the-art computer-aided drug discovery, and to apply these tools to biological systems of pharmacological interest. In particular, we aim to apply and improve our understanding of the principles of molecular recognition, create new tools to explore chemical space and discover novel bioactive molecules that can create therapeutic opportunities and expand the druggable proteome.

Specific Objectives

1. Explore the fragment-size chemical space to unravel new chemical scaffolds that can be used as starting points for drug discovery. This can be broken down into the following secondary objectives:

1.1. Develop a computational pipeline for mining the fragment chemical space around a selected fragment hit, delivering novel chemotypes with a very high probability of being active.

1.2. Test the computational pipeline on a prospective study, discovering active compounds for BRD4(1) starting from a pre-existing fragment.

1.3. Reach objective success criteria: high hit rates and scaffold novelty through scaffold hopping.

2. Use the novel series of ligands to investigate the unusual behaviour of structural water molecules in BRD4(1). Secondary objectives are:

2.1. Probe, experimentally and computationally, a hydrophobic hotspot in BRD4(1) lined by structural water molecules.

2.2. Demonstrate the correlation between experimental and computational free energy of binding of the different compounds to the protein

2.3. Provide a molecular explanation for the counterintuitive molecular recognition properties of the structural water molecules.

3. Develop the most active fragment of the novel series with *de novo* virtual synthesis. Secondary objectives are:

3.1. Apply the NAOMInext program to the BRD4(1) ligand, to increase potency through formation of additional protein-ligand interactions.

3.2. Validate experimentally the predictions and learn strength and limitations of the method.

4. Rational discovery of potent inhibitors of the TREX2 protein, to investigate their application in psoriatic disease. This can be broken down into the following secondary objectives:

- **4.1.** Investigate TREX2 druggability [redacted]
- **4.2.** Virtual screening to obtain the first binders of this novel target.
- **4.3.** Experimental determination of compound activities.

Chapter 3

Methods

In this thesis I have used a variety of computational tools. In this section I will first provide some background about these methods, then I will proceed to provide a detailed account of how the methods have been used to attain the specific objectives.

3.1 Background – Chemoinformatics

Chemoinformatics are a range of methods for efficient manipulation of chemical data in electronic format.[135]. They can handle large data volumes and are ideally suited to navigate the chemical space. Given a molecule of reference (e.g. a binder to the target of interest), a substructural search can rapidly identify molecules containing the query substructure but with different substituents or additional parts. This is particularly useful when one wants to navigate the chemical space around a specific scaffold (like in SAR analysis). When scaffold hopping is sought, other computational approaches can help, especially methods based on molecular similarity (**Figure 3.1**). The fastest and most common for conventional applications are fingerprint-based methods (presence or absence of particular substructure encoded in a linear vector).[136]In a similarity search, features of the molecules are annotated as bits in a vector (fingerprint). If a feature is present is coded as 1, if not it's a 0. One of the most known fingerprints is MACCS (Molecular ACCess System) fingerprints [137]. Fingerprints of 2 molecules can be compared bitwise and a score assessing the number of matching features can be calculated (for example the Tanimoto similarity index). Molecules within a library with a score higher than a determined cut-off can be selected for further analysis. The selection of the cut-off is very important because it will determine if molecules are too similar to the original query (and maybe they are not adding any useful information to the analysis) or too different (and they do not contain the features of the original molecules that are important for the binding).

Simple changes like atomic substitutions can be considered scaffold hopping but, in this sense, the detection of more distantly related compounds while retaining the activity is more meaningful, especially if the change is impossible to predict by a medicinal chemistry expert [38]. This is better achieved with methods that sit at the edge between chemoinformatics and computational chemistry, such as pharmacophore searches (molecules with similar pharmacophoric features are obtained),[138] shape similarity,[139] and machine learning methods trained on 3D descriptors.[140] Ultimately, molecular docking can also be considered as a reverse similarity search, since it assesses structural complementarity to a common template (the receptor).

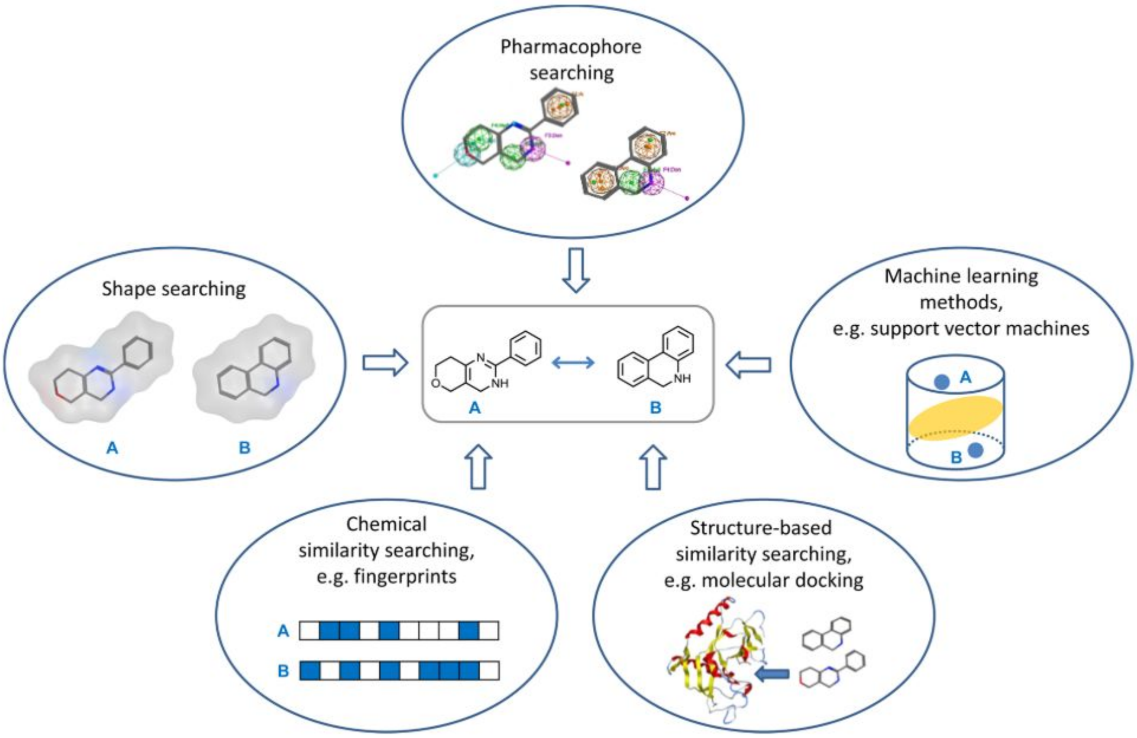

Pharmacophore searching
Shape searching
A
B
Machine learning methods, e.g. support vector machines
A
B
N
O
NH
NH
A
B
Chemical similarity searching, e.g. fingerprints
A
B
Structure-based similarity searching, e.g. molecular docking

Figure 3.1: *Computational approaches for scaffold hopping. Reproduced from Recent Advances in Scaffold Hopping. Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath. Journal of Medicinal Chemistry 2017 60 (4), 1238-1246. DOI: 10.1021/acs.jmedchem.6b01437 [38]*

3.2 Background – Docking

Docking programs try to fit a ligand in a binding site of a target, giving it a score to quantify this fitness. The majority of docking applications consider the target rigid, so the dynamics of binding is not considered. The structure has to be cautiously selected from experimental structures or from molecular dynamics snapshots according to the best solution for the considered problem (for example if the target has open or closed conformations). It has to be prepared by eliminating/maintaining water molecules and protonating at biological pH. If ligands in a library are given with a 2D structure, they also have to be prepared considering all the possible stereoisomers, tautomers, ring conformations and protonation states of the 3D structure. Docking programs cannot do that on-the-fly.[141]

The specific binding site where to dock should be chosen beforehand. If a ligand is known, it can be used to identify the binding site. If no information is known (for example in the case of a new target or allosteric binding site), the location of a cavity can be extrapolated from cavity detection methods.

There are multiple docking programs, each using specific scoring functions and search algorithms.[142] In this thesis I have used **rDock**, [143] a docking program supported in our group. It uses a genetic algorithm to generate poses in the binding site and an empirical scoring function to score protein-ligand goodness of fit. Ligands can be docked freely or can be guided by tethered docking (maintain some atoms fixed and explore conformations of the rest) or with pharmacophore restraints (an atom with a specific feature should be placed in a definite position with a distance-dependent penalty if the restraint is not fulfilled). Pharmacophoric restraints can be derived from known ligands or from hotspots of MDmix (see below). The docking program generates a certain number of poses per ligands (50 is recommended for convergence when docking drug-like molecules), but the process can be accelerated using the HTVS (High-throughput virtual screening) mode, which discards molecules on-the-fly if they do not pass some predetermined score filters.

Like all docking methods, rDock shows some limitations of accuracy that stem from the compromise between speed and accuracy [141]. The main limitations are due to the rigid body approximation (i.e. protein flexibility is not properly accounted for) and the scoring functions, which take a statistical approach that does not consider the effect of the local environment and is particular error prone for polar, ionic and metal interactions which have a large entropic and solvation component.

In a virtual screening, post-filtering of wrong poses and reranking of ligands are needed to overcome these limitations.This is possible applying other orthogonal methods to obtain consensus score (Dynamic Undocking), to confirm the poses (Molecular Dynamics) or to calculate energy of binding more accurately (Free Energy Calculations)

3.3. Background – Molecular Dynamics

Classical molecular dynamics (MD) simulate motion and interactions of atoms according to Newton’s laws. Universal parameters for proteins and molecules, such as mass, bond, angles, dihedral angles, improper angles and non-bonded terms, are annotated in a “forcefield”. A very used forcefield is **AMBER** and the homologous software can be used to conduct a series of different simulations and analysis [144].

Simulations are very useful in structure-based drug design as they can reproduce real events in ligand-targets interactions. Since the binding of a ligand to a protein is not a static event, protein and ligand flexibility is very important [145]. Even though the whole binding event of a ligand has been described [146,147], its routine implementation as a screening tool remains unfeasible due to the long simulation time. For this reason, unbiased MD to compute binding free energy from the observation of binding and unbinding events is rarely used. More common uses are observation of movements (e.g. conformational rearrangement upon ligand binding) and the extraction of conformational ensembles that can be used to compute properties, including binding free energies with end-state methods such as MM-PBSA.[148] It can also be used to confirm the binding pose of a ligand predicted by docking. After a few nanoseconds, it is already possible to see changes in the binding pose (to another more stable pose with more affinity) or even the unbound event (ligand leaving the cavity). In those cases, the initial binding pose suggested by docking cannot be considered reliable.

MD has given rise to a plethora of SBDD methods that rely on its ability to explore the energetic landscape of the system and generate meaningful ensembles. Some such methods have been used in this thesis with different purposes and they will be described in the next sections.

3.3.1. Mixed-Solvent Molecular Dynamics

Given the 3D structure of a target, it is important to determine which chemical moieties can form optimal interactions at particular sites. In particular, one should detect the “binding hotspots”, that is, regions on the protein surface where residues make the strongest interactions with the ligand and contribute considerably to the energy of binding [149]. In 1985 the program GRID first considered the use of probe molecules to map the surface of protein [20]. In Multiple Copy Simultaneous Search (MCSS), many copies of solvent molecules are minimized on the protein surface, in a similar way as GRID, and density maps can be obtained [150]. Subsequently, with Multiple Solvent Crystal Structure (MSCS) it was proved experimentally that, when co-soaked in crystal structure, multiple organic solvents overlap in regions of biological importance [151].

Inspired by these methods, in Mixed-Solvent Molecular Dynamics the protein is solvated in water and a co-solvent. Unbiased MD simulations allow the co-solvent molecules to compete with waters and density maps can be obtained [152]. The method, unlike the others, takes into account protein flexibility and water effects[153,154]. Mixed-solvent molecular dynamics can also give an estimate of the free energy associated with fulfilling each hotspot. This is possible thanks to the fast diffusion rates of very small molecules like solvents, which allow to observe a high number of binding-unbinding events in a relatively short MD simulation time range (tens to hundreds of nanoseconds).

In 2014 our group developed **MDMix**, a software for Mixed-solvent MD and it’s been used successfully in a series of proteins [152,153,154]. One of the most straightforward applications is the creation of protein-based pharmacophore, particularly useful when no additional knowledge on the protein is available except its structure. Pharmacophores are generated from the given hotspots (which represent desired features in a ligand). The created pharmacophores can then be used in virtual screening and they have been applied with success [155]. Since the first description of the method in 2009 by Seco *et al.*, new methods and approaches have appeared [156,157] and they have been applied to a variety of problems. A way of detecting druggable binding sites of a protein is by identifying sites where hotspots cluster together [8]. Given the physics-based nature of mixed-solvent MD, druggability predictions based on this method are non-parametric and can be used to detect non-standard binding sites, such as allosteric binding sites [158,159] and protein-protein surfaces [160]. In proteins that were never studied before, the identification of clusters of hotspots provides the most robust druggability assessment.

Apart from druggability assessment and protein-based pharmacophore creation [161], Mixed-solvent Molecular Dynamics can be used for binding energy prediction [156] and water displacement prediction [154,162]. In the last years, it has grown importance the application of Mixed-Solvent Molecular Dynamics for the detection of cryptic pockets [163]. These are pockets that open in the presence of a ligand and it is thought that the underlying mechanism of opening can be a mixture of induced fit and conformational selection [163e]. These pockets may be too lipophilic to open spontaneously in aqueous solvation, but can open in the presence of apolar solvent molecules, mimicking the ligand effect.[163a]

[redacted]

[redacted]

3.3.2. Dynamic Undocking

Dynamic Undocking (DUck) is a special form of steered MD developed in our group [165] in which a key protein-ligand hydrogen bond is pulled with a certain force from an initial distance of 2.5Å to a final distance of 5Å where the hydrogen bond is broken. From there, one determines the quasi-bound state as the point where the work has the highest value and defines W_{QB} as the work necessary to take the hydrogen bond from the optimal position (usually around 3.0Å) to the quasi-bound state. This parameter provides an assessment of the structural stability of the investigated hydrogen bond. Surprisingly, it was found that W_{QB} can be used as a measure to differentiate between binders and non-binders. Thus, DUck is used as a post-docking method that filters out a large proportion of docking false positives. To do so, a clear hydrogen bond donating or accepting moiety in the binding site of the protein must be chosen beforehand as the anchoring point. This is usually derived from MDmix simulations and will be used during docking as a pharmacophoric restraint, to ensure that the predicted binding poses of the selected ligands fulfil the key interaction. W_{QB} values above 6 kcal mol–1 have been associated with robust complexes[27] and true actives [165], so ligands with values around this cut-off can be progressed and ligands with very low W_{QB} (below 2 kcal/mol) are indicative of unstable binding mode and can be discarded as false positives.

Docking scores do not correlate with W_{QB} so a consensus score between the 2 scores (docking score and DUck W_{QB}) can help to identify highly probable true ligands.

3.3.3 Free Energy Calculations

The Holy Grail of computer simulations is the calculation of free energy of binding, which has a direct relationship with the experimentally observable binding constant:

$\Delta G_{bind} =- RTln\ K_D$

To be useful, the prediction should be very accurate, as even a small error in the binding free enegy value causes a very high error in K_D owing to the exponential relationship between them. Even though the idea behind free energy methods was proposed almost a century ago [166], it was not until the 1980's that it was applied to real molecular systems [167,168]. Nowadays there is a new increased interest in the topic thanks to recent advances with GPUs that can speed up the calculation and have results in much shorter times [169, 170]. At the same time, researchers continue to develop new and more accurate methods [171].

Free energy is a state function, so it depends only on the final state, not the path between them. For this reason it is possible to calculate the free energy of binding from a closed thermodynamic cycle (**Figure 3.2**).

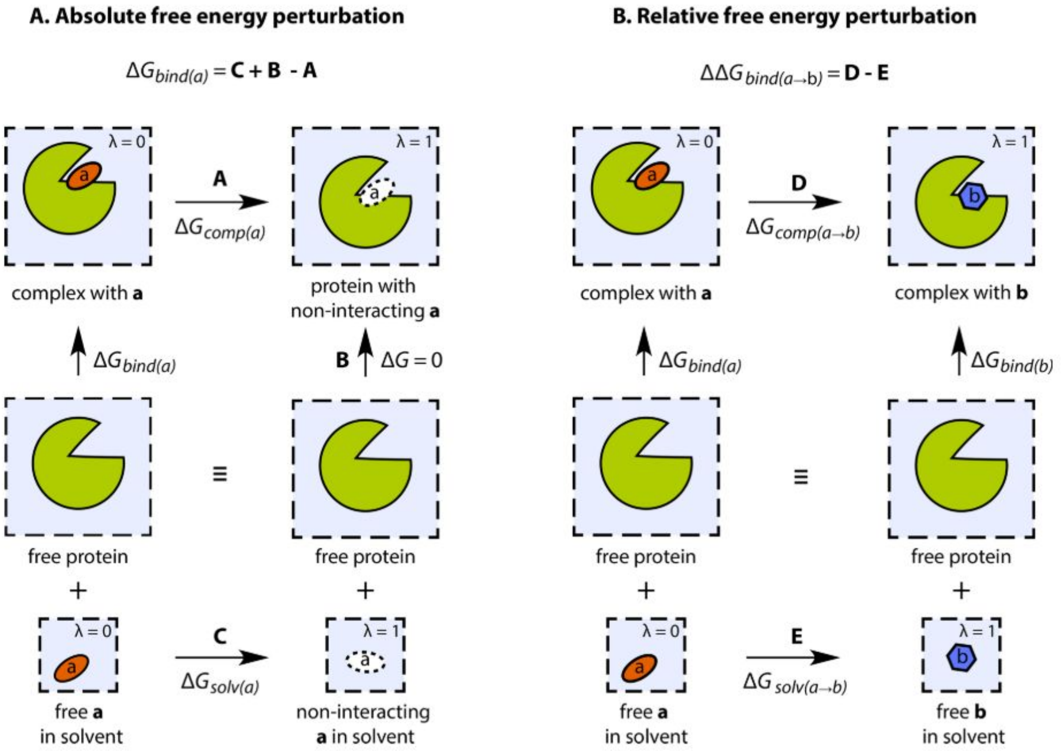


Figure 3.2: *Thermodynamic cycles for the calculation of free energies of ligand binding. Adapted from [172]*

Even though unbiased MD cannot be used routinely to calculate free energy of binding, enhanced sampling methods like Steered Molecular Dynamics, Umbrella Sampling, Alchemical Transformations, Replica-Exchange Molecular Dynamics and Metadynamics can be used. [172] In Alchemical Transformation, an initial ligand is mutated through non-physical states into another one (or making it "disappear") (**Figure 3.3**).

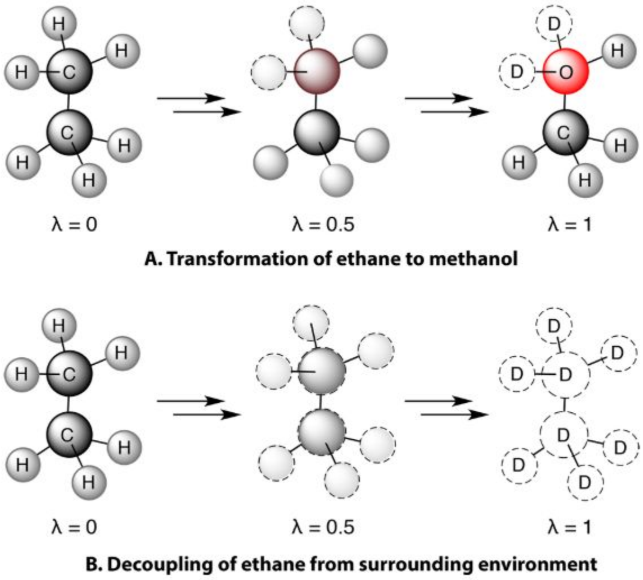

λ = 0
λ = 0.5
λ = 1
A. Transformation of ethane to methanol
λ = 0
λ = 0.5
λ = 1
B. Decoupling of ethane from surrounding environment
H
C
O
D

Figure 3.3*: Alchemical Transformations. Adapted from [172]*

The energy of binding (relative or absolute) is then calculated considering the thermodynamic cycle with different formulas (FEP, MBAR, TI, etc) [173].
Other methods to calculate binding free energies, so called end state methods, only take into account the bound and unbound states. The most common methods in this category are the molecular mechanics generalized Born surface area (MMGBSA) and molecular mechanics Poisson-Boltzmann surface area (MMPBSA). They both rely on implicit continuum solvation models and tend to provide mixed results. [148]
Free Energy Calculation can be especially important when you have ligands that differ only in a very small part, to understand the importance that that part has in the binding. This has application in Structure-Activity Relationship analysis, particularly for activity cliffs [21]. The method can accurately predict the difference in free energy of binding and used for the selection of new ligands. Extensive evaluation on diverse datasets shows the potential but also the limitation of the method, which routinely delivers root-mean square errors in the 1.0 to 2.0 kcal/mol range, depending on the target or specific chemical series[174]) It is thought that thanks to recent advances it may become possible to use Free Energy Calculations as a Virtual Screening tool [175] but this is far from practical yet.

3.4 A Computational Pipeline to Explore Fragment Chemical Space

Ligand selection. After collecting all the fragments co-crystalized with BRD4(1) in PDB, we choose fragment 1XA (PDB code 4LR6)[176]. The scaffold of this fragment is an isoxazole and derivatives are contained in some known binders of BRD4 [113]

Protein structure selection and preparation. The PDB structure of BRD4 (PDB code 4LR6) was prepared with MOE program [177] and set the protonation states at pH 7.0. From an internal study of conserved waters carried out overlapping the PDB structure of BRD4(1) we decided to maintain 7 water molecules in the cavity (HOH 302, 305, 311, 322, 327, 331, 332). The final structure was saved in the standard Tripos MOL2 format.

Database preparation. We downloaded ZINC15 (version of October 2015) database [59] filtering by reactivity (“clean” was selected) and purchasability (“in-stock” was selected, meaning that only the molecules ready to be delivered are included). The database was downloaded as SMILES. It contained 15,662,223 molecules. Then it was divided into subsets based on the number of heavy atoms (HA) in order to allow a quicker search (from 4 to 36 HA) Molecules containing more than 36 atoms were discarded because their size is not optimal for a lead-like compound (number after filtering: 15,247,008). After categorization by HA, the MACCS fingerprints [137] were generated using Openbabel [178] (see “Similarity search, MCS and superposition”).

The 3D conformations were calculated with LigPrep [179] (see “Compound preparation”) and stored as SDF files. For 0.17% of molecules it was not possible to obtain the 3D structure. The final number of molecules included in the database is 21,192,046 states (tautomers, protonation states, stereoisomers for racemic mixtures and ring conformations) for

15,220,882 unique molecules. All the states of the same unique molecule are contained in a single SDF file identified by the molecule ZINC ID.

Compound preparation. Each compound was prepared with Ligprep [179]. The protocol we applied generates a maximum of 8 stereoisomers, 6 tautomers, ring conformations up to 8 kJ/mol more than the lowest energy conformer, and creates all the structures present at pH 7 (± 1) for each molecule.

Similarity search, MCS and superposition. For the similarity search we used MACCS fingerprints [65]. We used Openbabel fastsearch format [178, 180] to store precalculated MACCS fingerprint to speed up the search. To grow a few atoms at a time, the protocol searches only for molecules with a number of atoms ± 2 relative to the query. The similarity index used was Tanimoto. The similar molecules found must be superposed to the initial fragment to ensure that the binding mode is preserved. In order to have a correct alignment, we first calculate the Maximum Common Substructure (MCS) between the similar molecule and a predefined rigid block of the fragment. As this block does not comprise rotatable bonds, there is no need to perform a flexible alignment. We have chosen the isoxazole structure as a fixed part in the MCS in the first iteration. MCS is calculated with RDkit [181]. The definition of the rigid block is updated at each iteration, as it may incorporate significant chemical modifications. From this step we obtain the SMARTS of the common and continuous atoms that form the rigid block, which will be fed to *sdtether* script (part of rDock)[182] to obtain a superposition that preserves the original binding mode. If there is more than one possible SMARTS match, *sdtether* generates all the possible superpositions.

Docking. For docking we used rDock [143] an open source program. The cavity was defined with *rbcavity* using the "reference ligand method" with the co-crystalized ligand as reference with 6Å radius. We applied the "tethered docking" function of rDock, which consists in restricting a selected part of the ligand and to optimize the rest. This fixed part will be the one superposed to the initial fragment. We allowed translation and rotation of the constrained part of just 0.01 Å per iteration in the genetic algorithm. Also we added a pharmacophore restraint for H-bond acceptor at 2 Å from the Nδ of Asn140, which is known to interact with BRD4 ligands[113], with a tolerance of 1 Å around it.

We applied the high-throughput mode of rDock: any ligand is discarded after the first run if *SCORE.INTER* is greater than -5 or the pharmacophore penalty score is greater than 2. If the first filter is passed, then after 3 additional runs, the program will check again the two scores with more restricted filters, discarding any ligand with *SCORE.INTER* greater than -10 and pharmacophore score greater than 1. All molecules passing this filter complete a total of 20 runs, which is sufficient to ensure convergence in the tethered docking mode.

After docking, the poses of all the states of the same molecules are ranked together and the best pose (lowest *SCORE* value) is selected and minimized without constraints in order to improve the match with the receptor while still preserving the binding mode. At this point we have one minimized pose per molecule. All molecules are then ranked by *SCORE.INTER*.

From the first iteration we noticed that the majority of the molecules showing good scores in the docking step contained a quaternary nitrogen that binds in the pocket where the structural water molecules are present. The cost of burying a charge often exceeds the gain of forming intermolecular interactions. In this case, as the interaction is formed with interfacial water molecules, the net result of placing a charge is difficult to predict and well beyond the capabilities of a docking score. For this reason, and noting that all ligands in the

literature present a hydrophobic group in this position rather than a charged group, we decided to filter out molecules containing the quaternary nitrogen at the bottom of the cavity.
The first molecule (ranked by SCORE.INTER) is chosen, then the rest of the molecules are chosen following the same ranking if they have a Tanimoto similarity less than 0.9 with the all the previously chosen molecules. This is done to discard very similar molecules, keeping only the best-scoring molecule as representative. When 500 molecules are selected, the selection program stops and the molecules are sent to the next step.

DUck. DUck (Dynamic Undocking) is a method developed in our group[165]. It consists in calculating the work necessary to break an H-bond important for the binding. The parameters used are the same as in the original paper: energy minimization for 1,000 cycles, 4 steps of gradual warming to 300 K in NVT for 400 ps and 1 step of equilibration at 300K for 1 ns in NPT (1atm), 5 runs of MD, each followed by 2 runs of SMD at 2 temperatures (300 and 325K). The cutoff for the Wqb was set to 4 kcal/mol. After minimization, heating and equilibration the compliance of this cutoff is checked at every run of Molecular Dynamics (MD) and Steered Molecular Dynamics (SMD) and if the W_{qb} of that simulation is below this cutoff no more runs are performed. Only molecules with W_{qb} greater or equal to 4 Kcal/mol will complete all the runs. The simulations were performed at the Barcelona Supercomputing Center using NVIDIA Tesla M2090GPUs.
Only a subset of the protein (hereinafter referred to as "chunk") is used as receptor in DUck calculation. Specifically, for BRD4 the chunk was prepared manually selecting residues within 6 Å from ASN140 (TRP81-PRO82-PHE83-GLN84-GLN85-PRO86-VAL87-ASP88-ALA89, LYS91-LEU92-ASN93-LEU94, TYR97, ILE101, PRO104-MET105, THR131,ASN135-CYS136-TYR137,TYR139-ASN140, ASP144-ASP145-ILE146, MET149, HOH 302,305,311,322,327,331,332).
ASN 140 was chosen as the H-bond donor on the protein side, and the DUck set-up scripts automatically identify the partner H-bond acceptor on the ligand side, based on distance. The final W_{qb} is the lowest value of all the SMD, as explained [165]. Since W_{QB} and *SCORE.INTER* scoring terms are orthogonal and complementary [165], a consensus scoring was implemented. Each score was ranked in ascending order and an index was assigned to each molecule. The consensus score was the sum of both indexes.

Spawning. From the ranking of the consensus score, we select the top 50 molecules. Each of the 50 molecules is used subsequently as query molecule for the similarity search calculation. Many of the query molecules ("parents") can have some similar molecules ("children") in common. The similar molecules undergo a refinement process: duplicates are removed and only unique molecules are kept, the molecules that were already docked in previous iterations are filtered out and the remaining ones are ranked by number of parents. Among these, the 100.000 children molecules that have the highest number of parent molecules in common are selected. The rigid core of the parent molecules that interacts with the target (close to the original fragment no more than 0.7 Å) is calculated with a MOE script and its SMILES code is used to identify the MCS with their respective children molecules. The parent with the largest MCS is selected and used as reference to superpose with *sdtether*. Then, the top 100,000 molecules are sent to the docking and all subsequent steps using exactly the same protocol as for the first iteration.

The process was repeated until reaching the fourth generation of molecules.

3.5 Analysis Of The Molecular Recognition Properties Of Brd4 Structural Water Molecules

Unbiased molecular dynamics*.*

Ligand ETH was co-crystallized in the previous study with PDB code **6ZF9**. The ligand pose was extracted from the PDB and missing hydrogens were added. Ligands MEH, OHH, NHH, NH3, OCH and SHH were prepared from this pose using the “Build” tool in MOE [183]. The coordinates were maintained, changing only the position of the particular substituent. This part was minimized with MOE and each ligand was saved as a MOL2 file.

The protein was extracted from the PDB and it was prepared with MOE (“Protein Preparation” tool). Seven water molecules were maintained (301, 304, 316, 320, 322, 324, 331). It was saved as MOL2 with AMBER nomenclature.

Gaussian optimization (Gaussian 09 [184]) was applied to each of the ligands and Antechamber [185] RESP method was used to obtain the partial charges. Parameterization with the GAFF force field was applied. Some parameters needed manual intervention.

Each ligand, the APO protein) and the seven protein-ligand complexes were neutralized and solvated in an octahedral box of TIP3P water with *tleap* [185]. Topology (prmtop) and coordinates (prmcrd) Amber files were obtained for each of them.

The same molecular dynamics steps were applied in all cases with *pmemd.cuda* of Amber18 [185]:

1) A first step of energy minimization for 1,000 cycles (maxcyc). The default algorithm is used with 10 cycles of steepest descent method and then conjugate gradient is switched on. Energy information is printed every 100 steps (ntpr). Atoms of the ligands and the protein are restrained using an harmonic potential with a force constant of 25 kcal/mol·Å^2 (ntr). Constant volume (ntb 1).

2) A second step of minimization for 1,000 cycles. Energy information is printed every 100 steps. Atoms of the ligands and the protein are restrained using an harmonic potential with a force constant of 5 kcal/mol·Å^2. Constant volume (ntb 1).

3) Four steps of progressive heating in NVT. Energy information is printed every 2000 steps. The coordinates are written every 2000 steps (ntwx). Atoms of the ligands and the protein are restrained using an harmonic potential with a force constant of 5 kcal/mol·Å^2. 100,000 MD-steps are performed (nstlim) with a time step of 0.002 psec (dt), using SHAKE (ntc). In the NVT ensemble the Langevin thermostat (ntt 3) was used with a collision frequency γ of 4.0 (gamma_ln) with a random seed generator (ig=-1). The initial temperature was set to 100K and was increased by 50K in 4 steps (150K, 200K, 250K, 300K). Constant volume (ntb 1).

4) Two steps of equilibration in NPT. Energy information is printed every 2000 steps. The coordinates are written every 2000 steps. Atoms of the ligands and the protein are restrained using an harmonic potential with a force constant of 5 kcal/mol·Å^2. In the NPT ensemble the Langevin thermostat (ntt 3) was used with a collision frequency γ of 4.0 (gamma_ln) with a random seed generator (ig=-1). The temperature was set to 300K. Constant pressure (ntb 2) periodic boundary conditions are used with isotropic position scaling (ntp 1) with a Berendsen barostat for a pressure of 1.0 atm and a Pressure relaxation time of 2.0 ps. In the first step 10,000 MD-steps are performed (nstlim) with a time step of 0.002 psec (dt), using

SHAKE (ntc). In the second step 490,000 MD-steps are performed (nstlim) with a time step of 0.002 psec (dt), using SHAKE (ntc).

5) Five steps of progressive release of the restraints in NVT. Energy information is printed every 2000 steps. The coordinates are written every 2000 steps. Atoms of the ligands and the protein are restrained using an harmonic potential with a force constant of 5, 4, 3, 2 & 1 kcal/mol·Å^2 for the first, second, third, fourth & fifth simulation, respectively. 100,000 MD-steps are performed (nstlim) with a time step of 0.002 psec (dt), using SHAKE (ntc). In the NVT ensemble the Langevin thermostat (ntt 3) was used with a collision frequency γ of 4.0 (gamma_ln) with a random seed generator (ig=-1). The temperature was set to 300K. Constant volume (ntb 1)

6) A final step of equilibration in NVT without restraints. Energy information is printed every 2000 steps. The coordinates are written every 2000 steps. 100,000 Number of MD-steps are performed (nstlim) with a time step of 0.002 psec (dt), using SHAKE (ntc). In the NVT ensemble the Langevin thermostat (ntt 3) was used with a collision frequency γ of 4.0 (gamma_ln) with a random seed generator (ig=-1). The temperature was set to 300K. Constant volume (ntb 1)

The restraints were applied to both the protein and the ligand in the case of complexes.

7) 200 production steps of 1ns in NVT (total 200ns). Energy information is printed every 5000 steps. The coordinates are written every 5000 steps. 500,000 MD-steps are performed (nstlim) with a time step of 0.002 psec (dt), using SHAKE (ntc). In the NVT ensemble the Langevin thermostat (ntt 3) was used with a collision frequency γ of 4.0 (gamma_ln) with a random seed generator (ig=-1). The temperature was set to 300K. Constant volume (ntb 1).

For the ligands in water only 20ns were performed. Nonbonded cutoff was 8.0 Å for free ligand simulation and 9.0 Å for APO and complex simulations. For each element (apo, ligand, complex), 3 replicas were performed.

Free Energy Calculations.

Minimization, heating and equilibration were performed following the protocol described above. The restart file after the last equilibration step was used to prepare the prmtop and prmcrd files for the alchemical transformations.

For each transformation the "START" ligand coordinates were obtained from the restart file. The coordinates for the common atoms were used for the "END" ligand. The atoms of the different substituent attached to the heterocycle were considered unique. tleap added them in the end ligand from the OFF file. Prmtop and prmcrd with both ligands in the same file were obtained (needed for pmemd). The same procedure was applied for ligands and complexes.

21 lambdas were used: 0.0, 0.02, 0.04, 0.06, 0.08, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.92, 0.94, 0.96, 0.98, 1.00. Each lambda is run in parallel. For each lambda, it is performed 1 step of equilibration of 1ns and 20 steps of production of 1ns each (total 20ns).

Energy information is printed every 4000 steps (ntpr). The coordinates are written every 4000 steps (ntwx). 1,000,000 MD-steps are performed (nstlim) with a time step of 0.001 psec (dt), without using SHAKE (ntc). The Langevin thermostat (ntt 3) was used with a collision frequency γ of 4.0 (gamma_ln) with a random seed generator (ig=-1). The temperature was set to 300K. NVT was used in the equilibration step (ntb 1) while the production step was in the NPT ensemble (ntb 2). Constant Pressure periodic boundary conditions are used in production with isotropic position scaling (ntp 1) with a Monte Carlo

barostat (barostat 2) for a pressure of 1.0 atm and a Pressure relaxation time of 2.0 ps. Nonbonded cutoff was 10.0 Å.
timask1 was the starting ligand, timask2 the ending ligand. softcore potential was applied to the unique atoms. ∂V/∂ λ values and the Bennett acceptance ratio scheme were calculated at every step. For some transformation to avoid using unnecessary space, the coordinates are written every 10,000 steps. Amber18 was used with the patch update.16 [186].
A scheme of the protocol is shown in **Figure 3.4**. For the analysis , a program from the Mobley lab was used (https://github.com/MobleyLab/alchemical-analysis)[173].

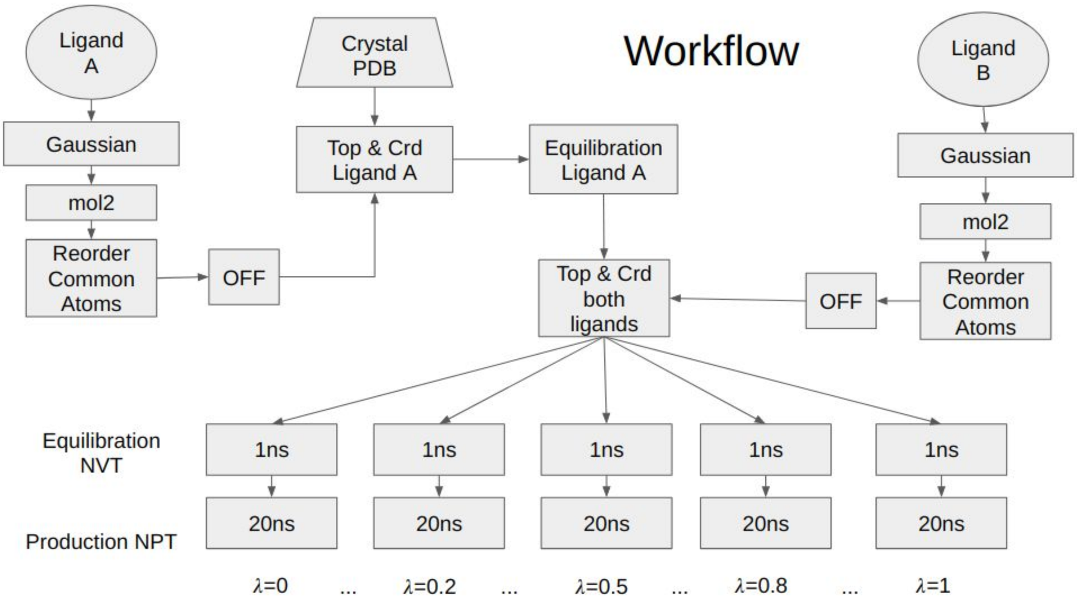


Figure 3.4: *Workflow used for the Alchemical Transformation to calculate the Relative Free Energies of Binding.*

3.6 Application of NAOminext to the BRD4 Ligand.

De novo techniques have been extensively used in fragment-based drug design, particularly in the Fragment-to-Lead (F2L) step [79]. Synthetic accessibility is still a common issue of these programs and various solutions have appeared (*Section 1.3.3.1*). **NAOMInext** [106] is a program developed in the group of Matthias Rarey at the University of Hamburg in Germany, where I did a research stay of 3 months in 2019. NAOMInext performs “constrained synthetically feasible fragment growing”. Building Blocks (BB) are virtually attached to a fragment *in situ* in his crystallized structure within the protein. The new bonds are formed according to a set of 58 virtual reactions [102] that are annotated in the program as “Reaction SMARTS”. By means of the “Reaction SMARTS” the program is able not only to recognize compatible reactions with the query fragment but also compatible BBs from the given library. Once the “virtual reaction” is made, a series of conformations of the new added part are generated and evaluated with a score. The conformation with the best score is taken as the final pose. The new generated compounds can be ranked by this score and the top ranking can be considered for lab synthesis. In this way 3 objectives are achieved at the same time: (1) Growth of the fragment, (2) Synthetic accessibility is considered by using a

set of rules (3) The best conformation per molecule is obtained by the cavity-constrained conformational search.
NAOMInext has been used in section 4.1.3 to grow a fragment hit into a potent ligand.The following specific NAOMInext parameters were used:

- Number of start pose to use(default 50): 30
- Number of poses to write: 10
- number of automatic growing steps 1
- Number of parallel threads 10
- user info update interval 10 secs
- ligand result chunk size 10
- thread timeout 300secs
- level of info output 4
- minimum allowed building block size (e.g. size of 1 allows attaching atoms) 1
- angle used to rotate start pose (default 15.0): 15.0
- active site radius (default 6.5): 6.5
- show additional ligands
- SASA element coloring
- Sampling method
- skip input molecules on write
- number of molecules to write 2147483647
- number of poses to write (for each result molecule): 10

3.7 Discovery of TREX2 Inhibitors

[redacted]

[redacted]

Chapter 4

Results

4.1 BRD4(1)

4.1.1 Fragment-size Chemical Space Exploration For New Scaffold Opportunities

Herein, we devise an automated platform that navigates the chemical space around any specific fragment, rapidly evolving the initial hit into novel fragments that share a binding motif but are structurally diverse, including non-obvious scaffold changes.This will help discover neglected fragments that can display improved efficiency, good starting points for growing or represent novel patentable chemical matter.

In a prospective application on BRD4, we recycle a known fragment to obtain active compounds with scaffolds that are different from the initial fragment.Due to their therapeutic interest, they have been thoroughly investigated and a large number of inhibitors exist, covering a multitude of chemical scaffolds [113,117,118]. As such, it is a challenging test to investigate if deeper exploration of the available fragment space can unearth novel chemical matter.

4.1.1.1 Development Of The Computational Pipeline

In this work a new pipeline was developed for mining the fragment chemical space to find more efficient fragments with novel scaffolds while at the same time increase the size and evaluate the binding by docking and dynamic undocking. The algorithm can repeat the search in subsequent iteration using the previously found molecules as “parents” for the new generation of molecules.

We devised a fragment mining tool with the following key characteristics:

1) efficient and scalable exploration of chemical collections; 2) performs scaffold hopping, in order to explore the diversity around a fragment hit; 3) maximises the probability of finding active compounds by exploiting the structural information; 4) performs iterative and step-wise fragment growing in order to increase potency while maintaining or improving ligand efficiency; 5) automatic and unsupervised process.

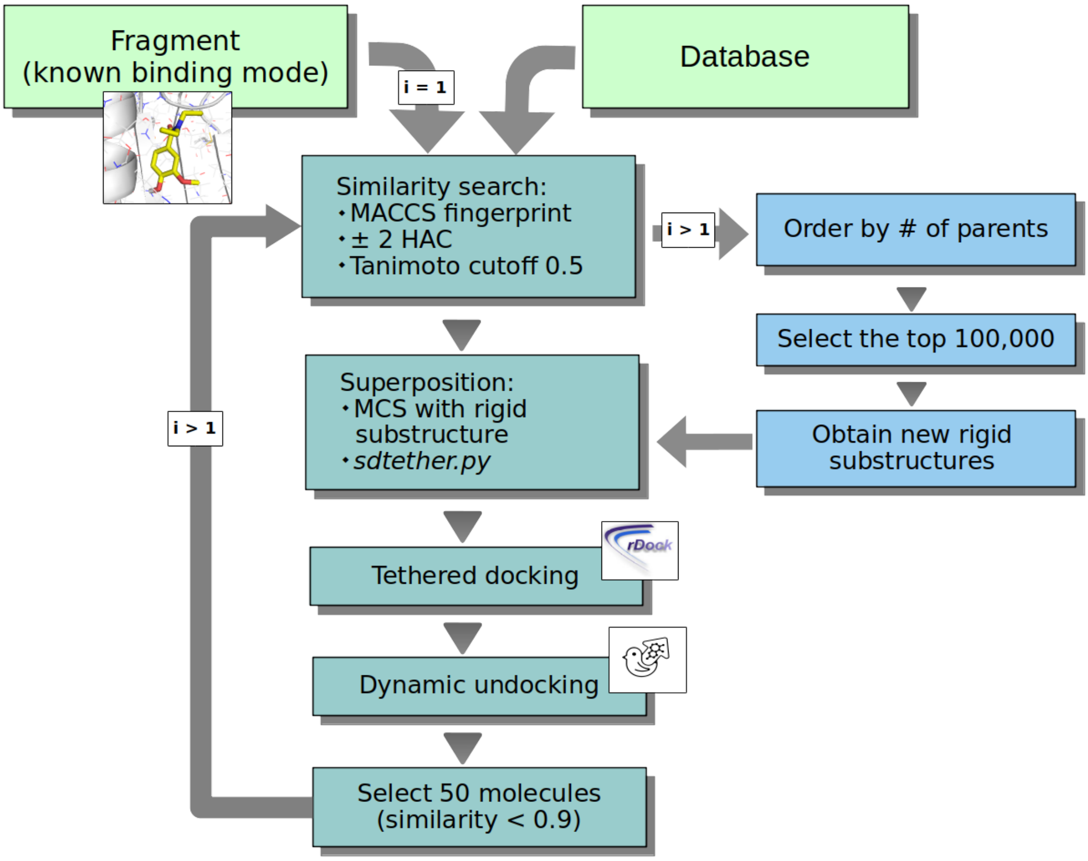


***FIGURE 3.1**. Steps of the fragment evolution platform. For the first iteration 5 steps are performed. For the next iterations 3 steps (on the right) are added between the first and the second step. See text for explanation. i=iteration.*

Figure 3.1 provides a visual summary of the platform.The process starts with an initial fragment, for which the binding mode is known. An essential interaction for binding, generally a hydrogen bond, will be preserved throughout the optimisation process to increase the chance of selecting active compounds. Then, a permissive similarity search (50% similarity cutoff) is performed on a chemical collection of choice. For convenience, here we use the ZINC15 “in-stock” subset.[59]. The search is limited to molecules of a similar size (+/- 2 heavy atoms). The size limit rewards ligand efficiency and avoids biasing the search towards larger molecules, which are more abundant and attain better absolute scores.

The retrieved molecules are then superposed to the query molecule in a two-step process to ensure that the key interaction and binding mode is maintained: i) identification of the Maximum Common Substructure (MCS) with the original fragment, which must include the key interaction point; ii) superimposition of the atoms in the MCS followed by optimisation of intermolecular interactions formed by the rest of the molecule by means of tethered docking.[143] The 500 top-scoring molecules are selected, always ensuring that no two molecules are more than 90% similar.

Dynamic undocking (DUck) is then applied to filter out docking false positives that cannot form structurally robust interactions.[165]. The 50 top-scoring molecules are selected and can be used for another iteration.This number is considered sufficient for a hit-optimisation exercise and, as the molecules are ranked by score, generating a shorter list is trivial.

If bigger molecules are sought, the process can be repeated iteratively, always using the diverse set of 50 hits from the previous iteration as starting point. In this case, a similarity search is performed for each of the parent compounds. An important point is that the number of molecules to consider can grow rapidly at each step. This is caused by multiple factors. First and foremost, the expansion of the chemical space as bigger molecules are considered.

The number of theoretical molecules increases exponentially with the number of heavy atoms.[62a] In commercial catalogues the increase is far less pronounced but still important (**Figure 3.2**). A second factor adding to this trend is the natural tendency of the Tanimoto index to increase as more complex (bigger) molecules are considered (**Figure 3.3**). Finally, the use of 50 query (parent) molecules instead of 1 also multiplies the number of candidate molecules. To ensure that the protocol remains computationally efficient, all similarity hits are rank ordered by the number of common parents (out of 50) and only the 100,000 top-ranking compounds are further considered (the value can be adapted to match the available computational resources). This approach is important to focus the search towards the most promising areas of the chemical space and prevent excessive scattering.

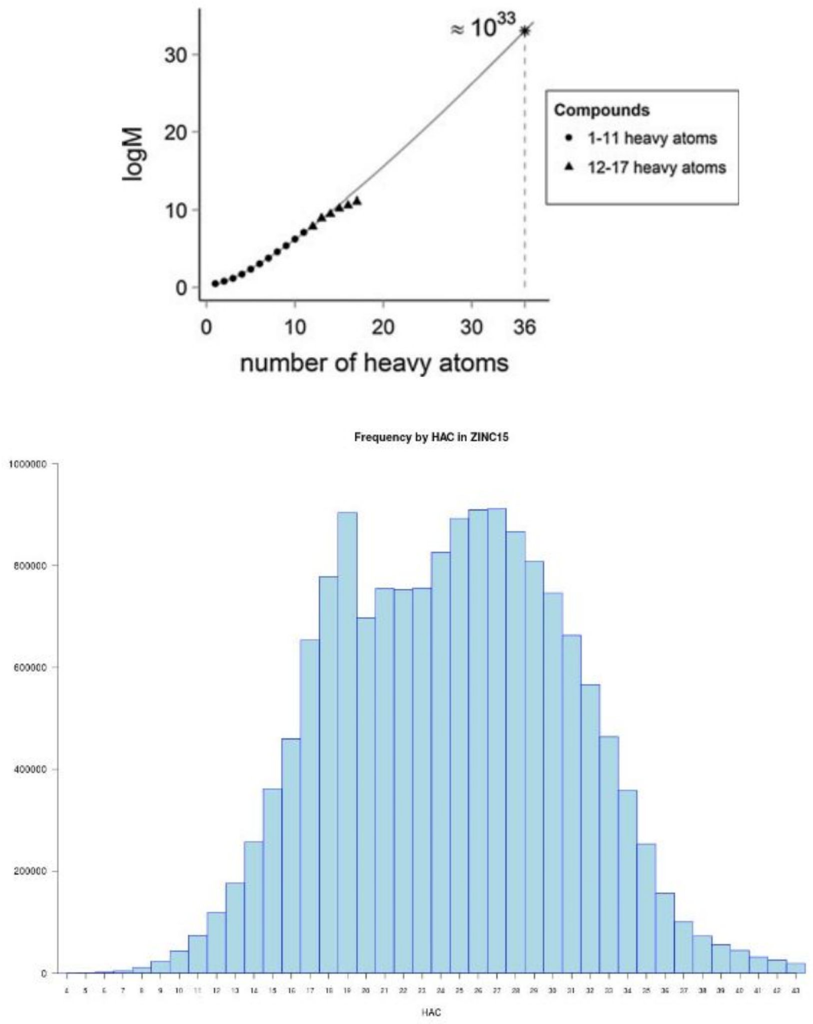


***Figure 3.2**. (a) Extrapolation of the compounds number (M) as a function of the number of heavy atoms (N) based on data taken from GDB-17. Adapted from [56] (b) Number of compounds present in ZINC15 database (version October 2015) by number of heavy atoms (HAC=Heavy Atoms Counts)*

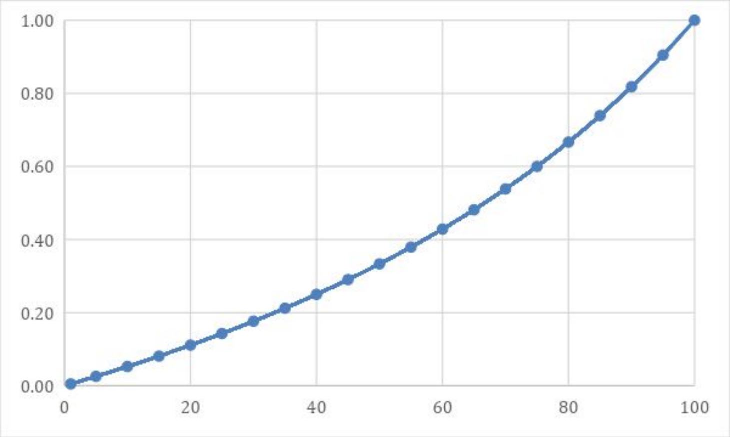


***Figure 3.3**. Expected Tanimoto similarity value as a function of the % bits set in the fingerprint of the molecules that are being compared (assuming an equal number of bits set in both molecules). As the number of bits set increases with the size of the molecule, the probability of finding molecules above a given threshold also increases for bigger (and more complex) molecules.*

4.1.1.2 Scaffold Exploration for BRD4(BD1)

The initial fragment (**1XA**, an amino-isoxazole) was taken from the literature [176] and was selected for optimization. The binding mode was available (PDB code **4LR6**), and its structure had already been used in a fragment evolution exercises, in this case, merging 1XA with a pre-existing ligand, (+)-JQ1, as shown in **Figure 3.4**.

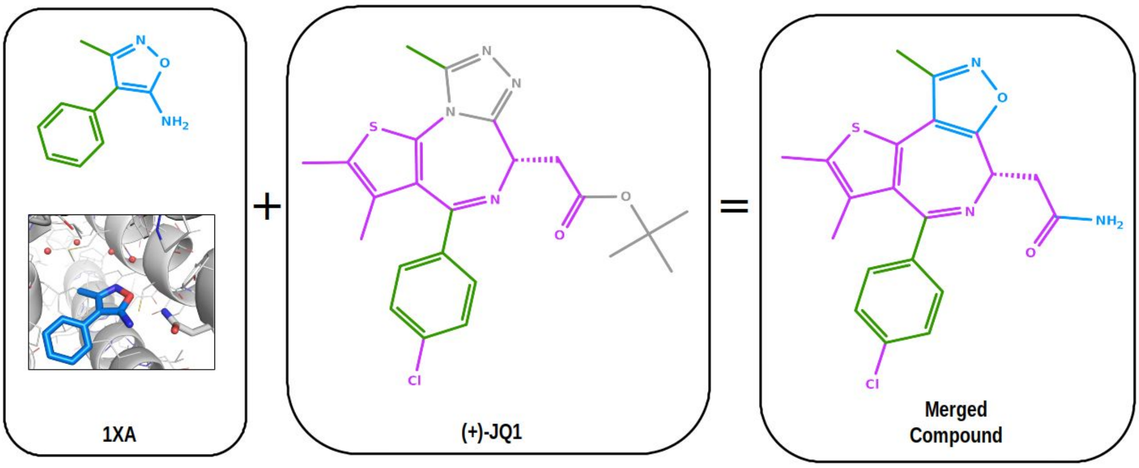


***FIGURE 3.4.** Merging approach described in Gehling et al. (2013)[176]. Two features of compound 1XA (crystal structure shown) were substituted in the structure of the known BRD4 inhibitor (+)-JQ1 giving the merged compound. Blue: substructure coming from 1XA; Magenta: Substructure coming from (+)-JQ1; Green: substructure common to 1XA and (+)-JQ1; Grey: substructure of (+)-JQ1 not present in the merged compound.*

The computational protocol was applied for 4 iterations. At the end of each iteration, 50 molecules are selected (**Figure 3.5**).

***Figure 3.5**. Chemical structures of compounds selected by the computational platform, sorted by consensus score. Those purchased are indicated by red squares.*

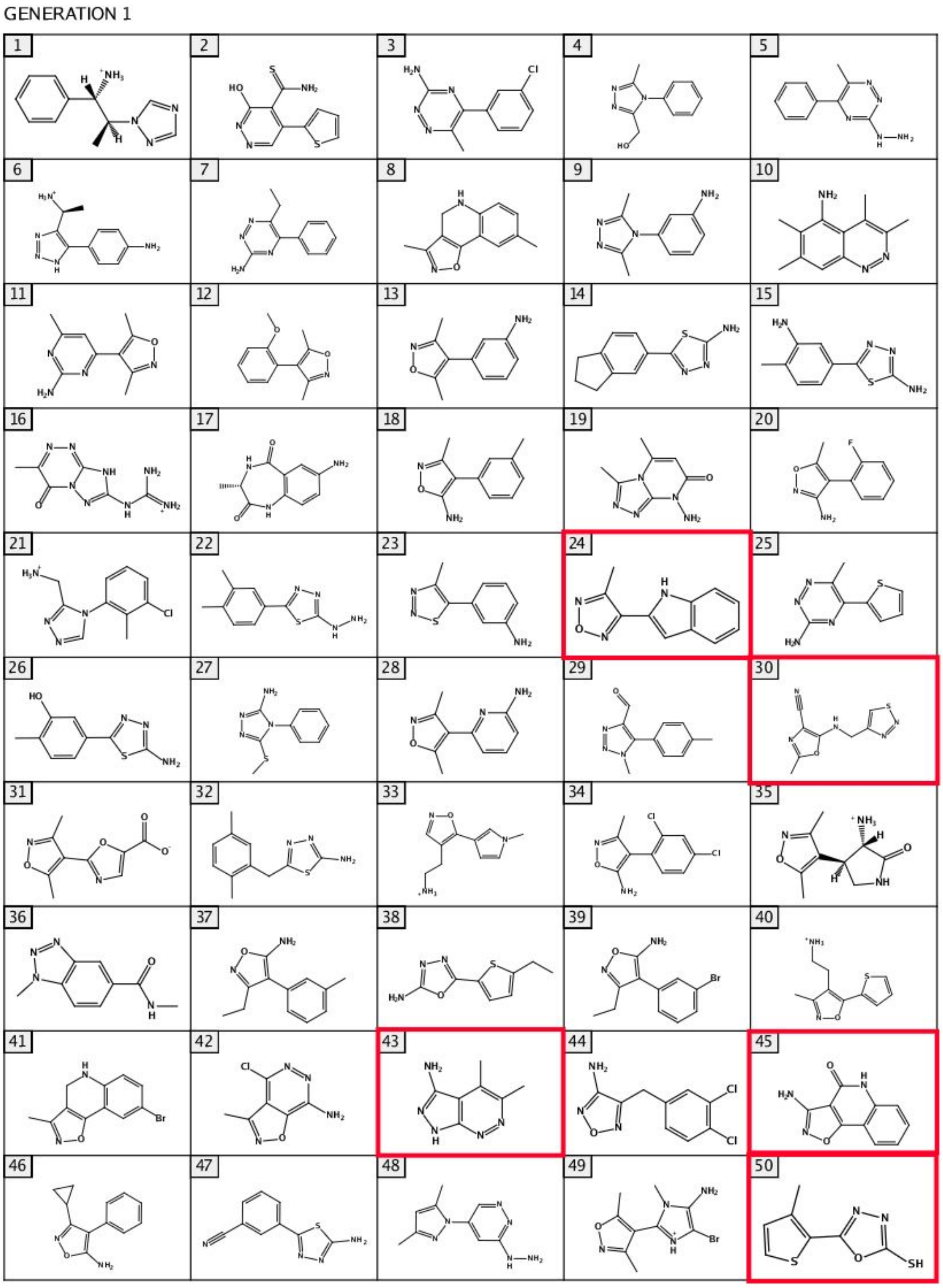


Figure 3.5. (continued)

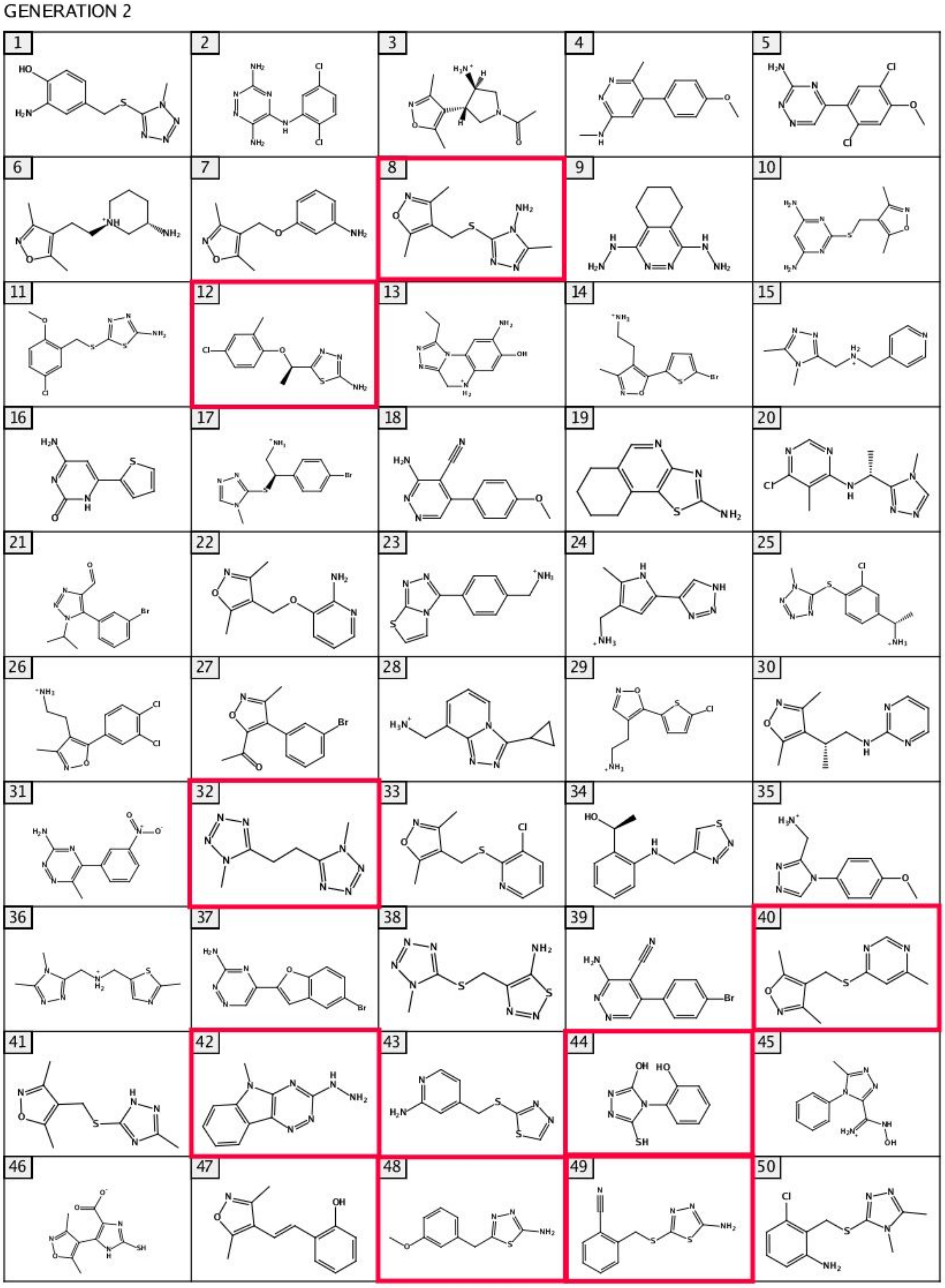

GENERATION 2

Figure 3.5. (continued)

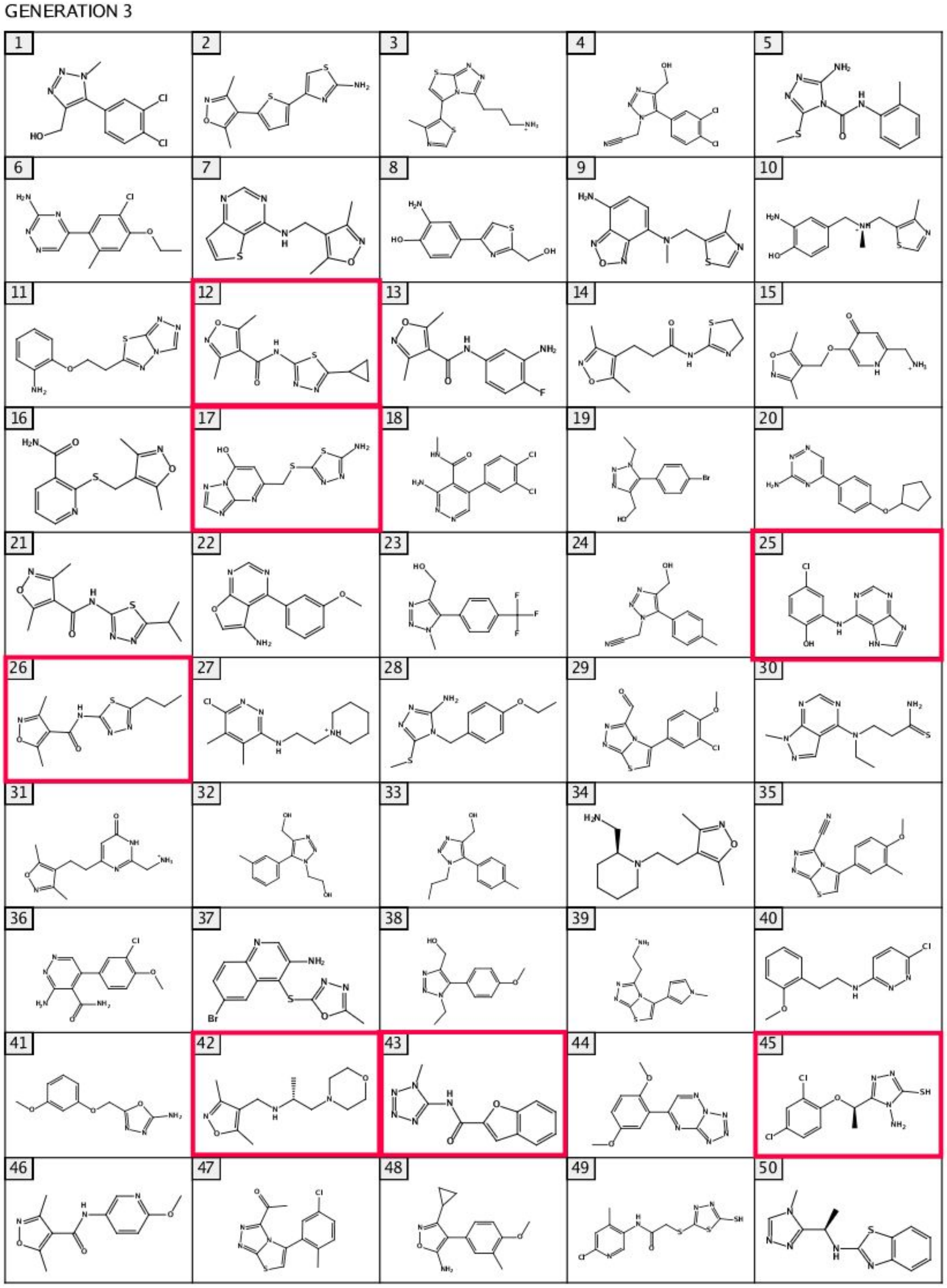


***Figure 3.5**. (continued). Yellow squares indicate members of a family with a common thiazolo[2,3-c]-1,2,4-triazole scaffold, all of which share a common predicted binding mode.*

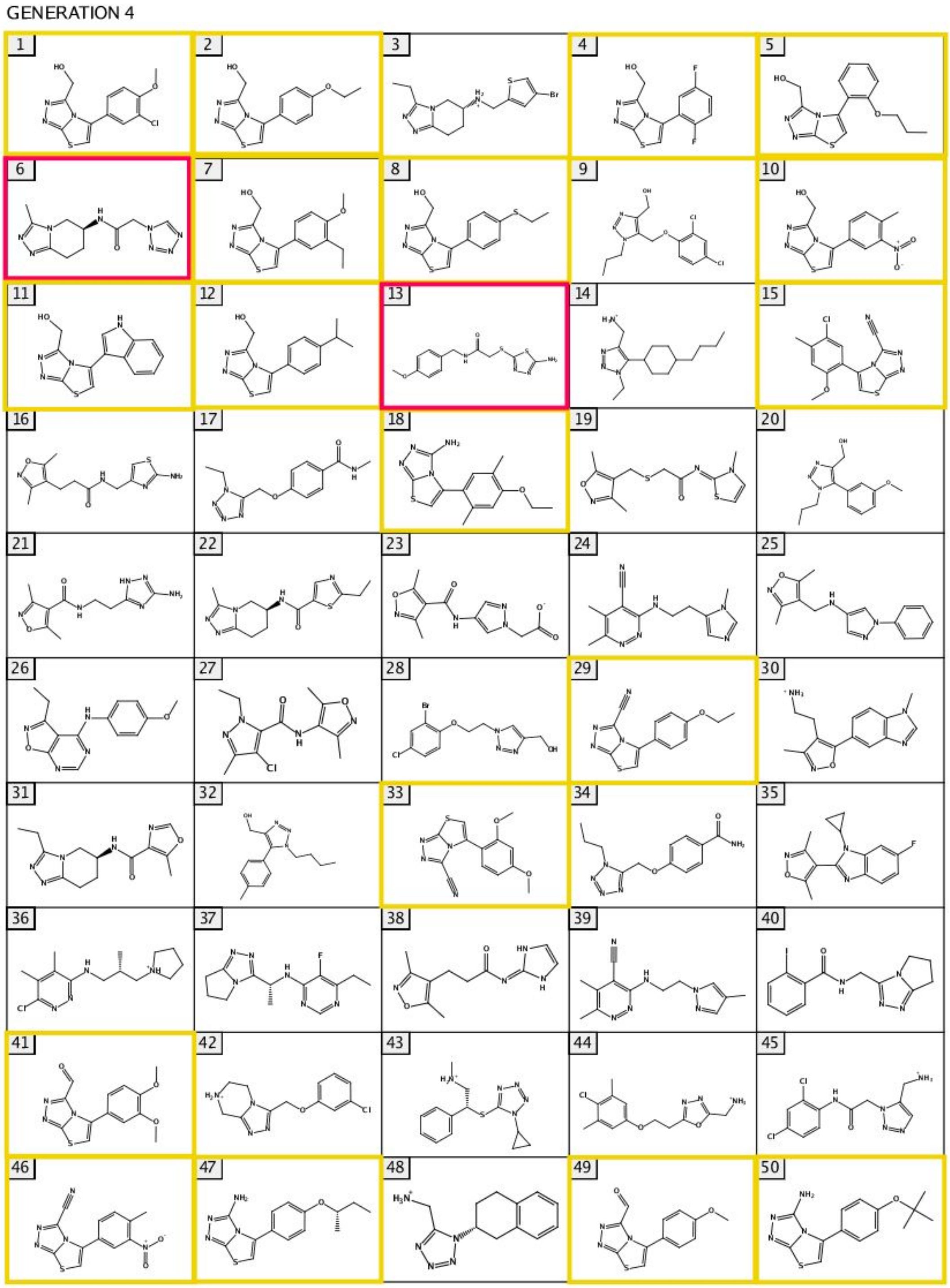


Table 3.1 summarises the number of molecules considered at each step. In order to validate the computational predictions, we proceeded to buy and test a sample of compounds (**Table 3.2**). Though ZINC15 contains compounds that are, in principle, available for purchase, in fact only a subset could be acquired and tested. We applied two orthogonal experimental

methods to measure binding: Differential Scanning Fluorimetry (DSF) at a single concentration (10uM) and a substrate displacement assay by TR-FRET at a range of concentrations, to attain an IC50 value. DSF measures protein stabilisation caused by ligand binding. It can be very sensitive and give clear signals even for weak-binding compounds, but is relatively prone to false positives and false negatives.[187] TR-FRET is a competitive displacement assay. It is more quantitative than DSF, but also very sensitive to environmental conditions.[188] Both methods have been abundantly used in bromodomain research.[189] We consider compounds as active if they give a clear signal by either method, but will focus on those where the methods coincide.

TABLE 3.1. *Statistics of the molecules obtained in each iteration and each step of the computational platform. HAC= Heavy Atom Count.*

	Iteration 1 **≤15 HAC**	**Iteration 2** **≤17 HAC**	**Iteration 3** **≤19 HAC**	**Iteration 4** **≤21 HAC**
Molecules in DB	986,524	2,141,996	3,779,575	5,037,693
Similar molecules (%)	40,231 (4.1%)	587,699 (27.4%)	1,306,135 (34.6%)	2,334,865 (46.3%)
Docked molecules (no. of parents)[a]	40,231 (1)	100,000 (7,11,37)	100,000 (8,10,28)	100,000 (13,15,34)
Molecules passing docking (%)	9,622 (24 %)	12,656 (13%)	16,066 (16%)	11,023 (11%)
DUcked molecules	500	500	500	500
Molecules passing DUck (%)	58 (12%)	86 (17%)	114 (23%)	90 (18%)
Selected	50	50	50	50
BM Scaffolds[b]	33	43	39	32
AG Scaffolds[c]	18	26	35	29
Tested	5	8	7	3
Active[d] (%)	4 (80%)	5 (63%)	3 (43%)	3 (100%)

[a] *values in brackets represent the minimum, median and maximum number of common parents*

[b] *Bemis-Murcko (BM) Scaffolds: unique scaffolds using Schuffenhauer fragmentation [36,37], taking into account the "least-pruned" fragment. Carbonyl groups are not removed in the fragmentation and elements are kept.*

[c] *Atom-Generic (AG) Scaffolds: Considering the unique scaffolds as formed only by carbons but maintaining the hybridization [37]. Tautomers are grouped together.*
[d] *In any experimental assay.*

Results of first iteration

As 1XA has 13 non-hydrogen ("heavy") atoms (HA), the protocol considered molecules in ZINC15 ranging from 11 to 15 HA. This represents almost 1 million molecules, of which only 4% were ≥ 50% similar to 1XA (MACCS fingerprints and Tanimoto similarity)[137] and were considered in the next steps. After superimposition of the MCS and tethered docking, 9,622 molecules (24%) presented a docking score as good or better than 1XA (-16.2 Kj/mol; **Table 3.2**). The top 500 were then subjected to DUck,[165] to calculate the work needed to break the key hydrogen bond with Asn140 (WQB). A WQB threshold of 4 Kcal/mol was chosen according to the value obtained by 1XA (4.5 Kcal/mol) and because it is a value most ligands fulfil.[27]

Only 58 of the 500 molecules passed this threshold (**Table 3.1**). Interestingly, the final list of 50 molecules (**Figure 3.5)** represents 33 different Bemis-Murcko (BM) scaffolds and 18 Atom-Generic (AG) scaffolds. The original scaffold is well represented in this list, with 7 members (14%), increasing to 18 members (36%) if we consider AG scaffolds [37]. But, as intended, a wide range of alternative scaffolds is also present (**Table 3.2**).

Only five compounds in the final list were available from the vendors and could be tested. Luckily, each represented a different scaffold. Pleasingly, 4 of these compounds (80%) were active (**Table 3.2**). The predicted binding mode of these compounds is shown in Figure S4. Compound 3, presented an IC50 value of 72 μM, which makes it more potent and ligand efficient than the parent compound. The binding mode of this compound was also confirmed by X-ray crystallography (**Figure 3.6**). Of note, compound 3 preserves two adjacent hydrogen bond acceptors that interact with Asn140, but the 5-membered ring (isoxazole) is replaced by a 6-membered ring (pyridazine). At the same time, the exocyclic amine of 1XA (which donates a hydrogen bond to Asn140) is now cyclised into a pyrazole ring. This radical change of scaffold is possible thanks to the high tolerance in the similarity search (50% similarity cutoff). This result encapsulates all the features that we wanted from the computational platform: a non-obvious transformation using chemical matter that is immediately available, excellent ligand efficiency, and a novel scaffold (never reported for BRD4 ligands) with completely different development potential compared to its parent compound.

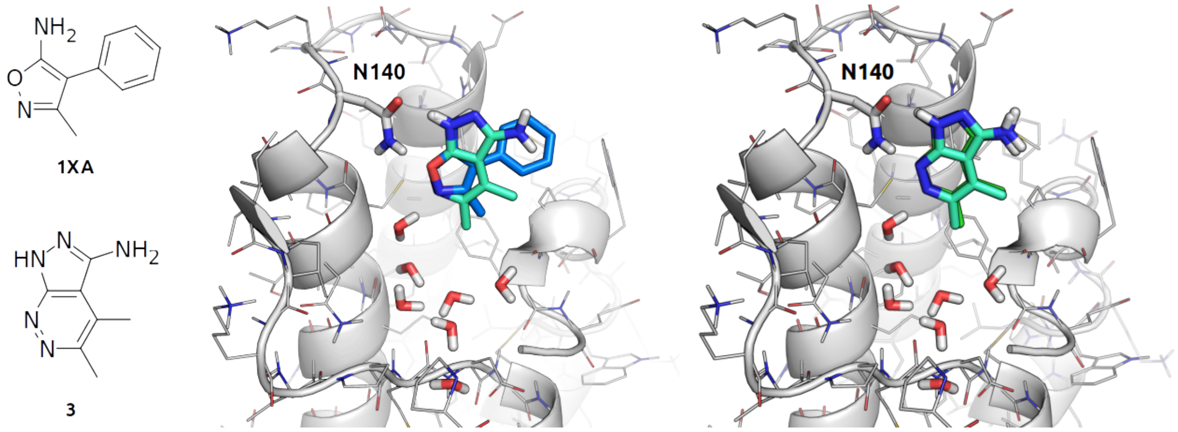


FIGURE 3.6. *Crystal binding mode of* ***3*** *compared with* ***1XA*** *(left) and with the binding mode predicted by the computational platform (right).*

TABLE 3.2: *Summary of the experimental results for the molecules tested shown by Iteration and ordered by consensus score. Number of heavy atoms (HAC= Heavy Atoms Count), similarity with reference, Docking and DUck scores and values of the different assays are shown.*

ID	**HAC**	**sim with 1XA**	**rDock (Score .inter)**	**DUck (W_{qb})**	**DSF (ΔT at 10μM)**	**TR-FRET (IC_{50})**	**XRAY**	**LE**
1XA	13	1	-16.2	4.5	4.55 ± 0.62	91μM	4LR6	0.52
				Iteration 1				
1 **(SPF17)**	15	0.7	-17.7	5.1	2.64 ± 0.47	n.s.		
2 **(SPF18)**	15	0.57	-18.1	4.9	3.41 ± 0.38	n.s.		
3 **(SPF1)**	12	0.53	-19.7	4.5	0.89 ± 0.22	72 μM	6ZED	0.58
4 **(SPF19)**	15	0.57	-19.6	4.4	3.09 ± 1.10	n.s.		
5 **(SPF2)**	12	0.55	-19.4	4.3	n.s.	n.s.		

Iteration 2								
6 (SPF3)	16	0.5	-23.4	5.0	n.s.	191μM		0.45
7 (SPF4)	17	0.5	-23.4	4.4	0.83 ± 0.27	79μM		0.4
8 (SPF20)	14	0.39	-22.1	4.0	n.s.	n.s.		
9 (SPF5)	16	0.5	-21.6	4.1	1.83 ± 0.27	30 μM	6ZEL	0.38
10 (SPF6)	16	0.43	-21.5	4.2	n.s.	n.s.		
11 (SPF7)	14	0.46	-21.4	4.2	n.s.	644μM		0.57
12 (SPF8)	15	0.46	-21.3	4.1	1.59 ± 0.24	n.s.		
13 (SPF15)	16	0.51	-21.1	4.3	n.s.	n.s.		
Iteration 3								
14 (SPF9)	18	0.65	-24.6	4.9	n.s.	192 μM		0.4
15 (SPF21)	18	0.43	-23.4	6.3	n.s.	n.s.		
16 (SPF22)	18	0.4	-23.1	6.6	2.36 ± 0.52	491μM		0.43
17 (SPF10)	18	0.56	-23.9	4.5	0.88 ± 0.19	n.s.		
18 (SPF11)	18	0.42	-22.6	6.1	n.s.	n.s.		
19 (SPF12)	18	0.58	-22.6	5.8	n.s.	n.s.		
20 (SPF13)	18	0.42	-22.6	5.4	n.s.	n.s..		
Iteration 4								
21 (SPF23)	20	0.43	-26.6	7.3	2.88 ± 0.16	n.s.		

22 **(SPF14)**	19	0.43	-26.1	5.1	0.85 ± 0.12	n.s.		
23 **(SPF16; SSR4; ETH)**	16	0.37	-20.1	6.8	3.32 ± 0.4	26µM	6ZF9	0.38

n.s. No signal

Results of successive iterations.

The platform offers the possibility to continuously grow the virtual hits in a stepwise manner. Thus, we also assessed this functionality prospectively. This option faces multiple challenges. Ideally, one would like to perform experimental testing after each cycle, to identify true binders and remove the false positives (i.e. molecules that receive a good score but turned out to be inactive). But this would not be practical, as it would take a long time. Thus, the platform should be able to execute multiple iterations in one go. This means that the parent compounds after iteration 1 are virtual hits that might be false positives. This implies a risk of error propagation that we will be investigating in this prospective study.

Finally, we obtained 50 molecules per iteration (**Figure 3.5**). For prospective validation, again we face the handicap of a very low procurement success rate, with only 8, 7 and 2 molecules available for purchase in iteration 2, 3 and 4, respectively (**Table 3.2 and Figure 3.7**). We noticed that the last iteration is converging to a family of compounds with a common 5-phenylthiazolo[2,3-c]-1,2,4-triazole scaffold (17 out of 50 molecules), but none of the compounds were available for purchase. Thus, we decided to synthesize 23 as a representative member of the family (**Figure 3.8**), making a total of 18 tested compounds. 11 of those (61%) are active, breaking down into 5 out of 8 (63%), 3 out of 7 (43%) and 3 out of 3 (100%) active at iterations 2, 3 and 4 respectively. The excellent hit rate confirms that the evolution process has not drifted into an area of spurious hits.

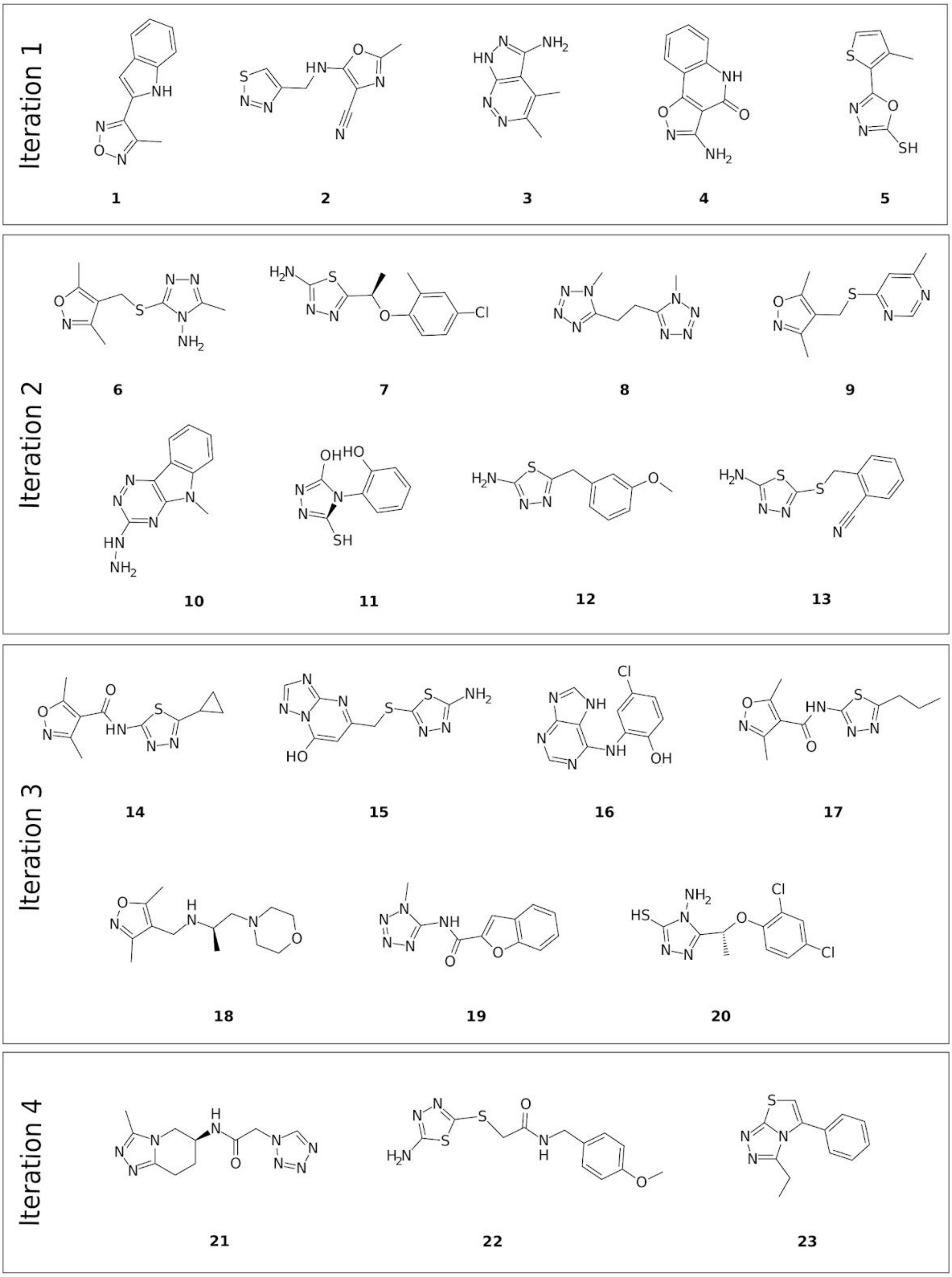


Figure 3.7: *Compounds found with the platform and tested with biophysical techniques.*

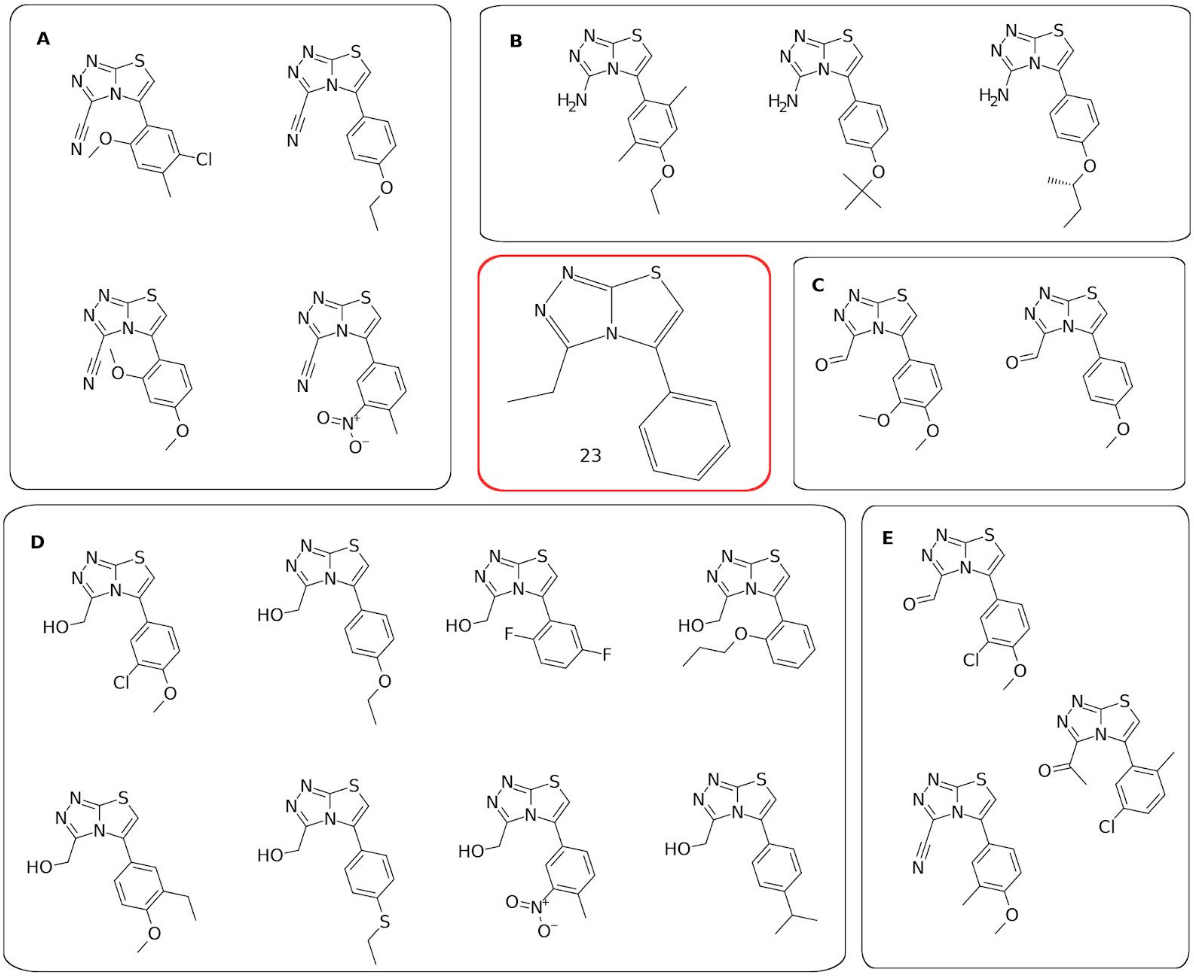


Figure 3.8: *Compound **23** in the center was synthesized as a representative of the family of compounds found in iteration 3 (E) and iteration 4 (A-D) with scaffold n.10 (Table 3->5). In iteration 4 the compounds can be grouped by the substituent (A: CN, B: NH2, C: C=O, D: OH)*

The predicted binding mode of the active compounds (**Figure 3.9**) preserves the expected interaction pattern while changing the chemical scaffold sometimes quite significantly.

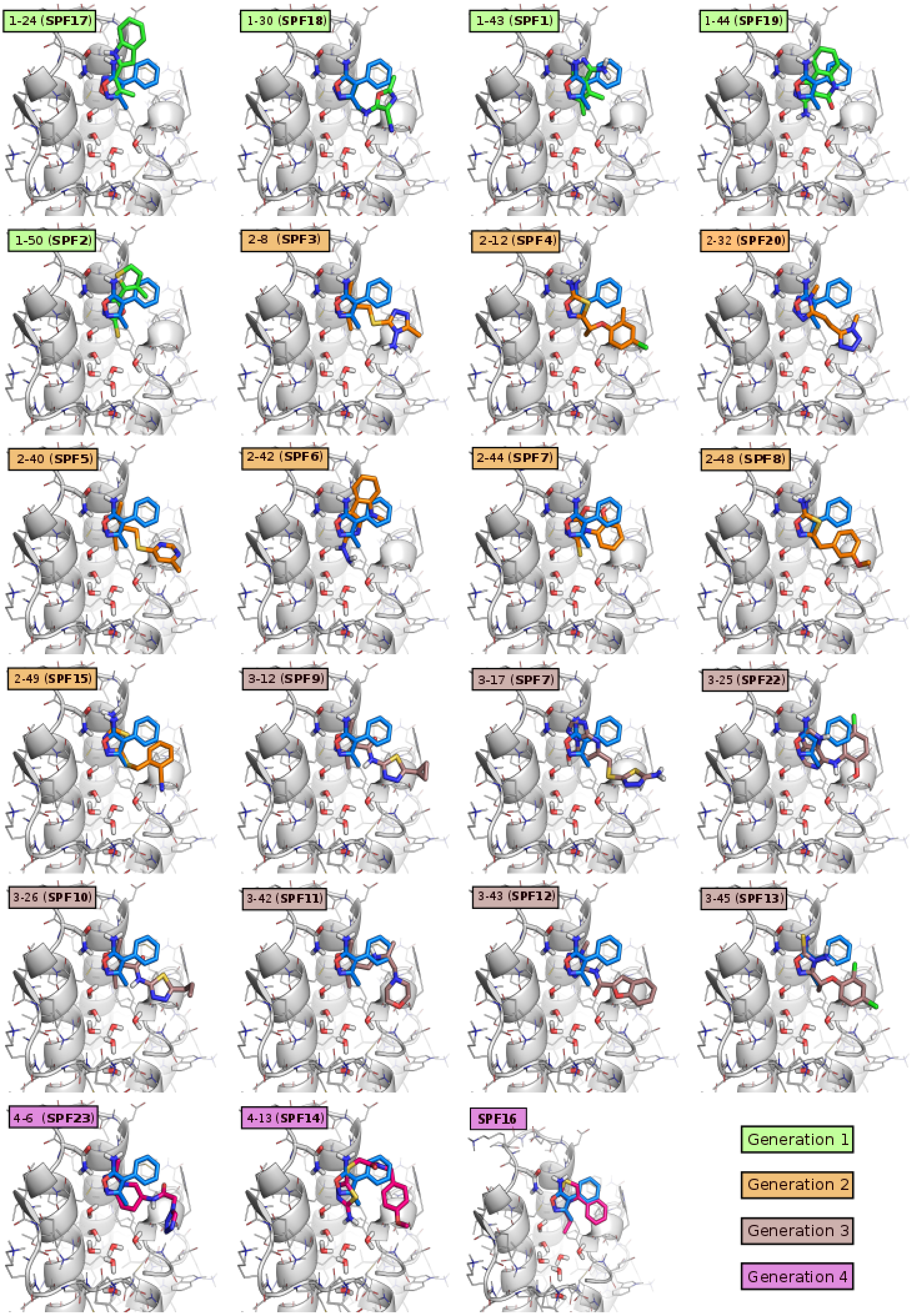


Figure 3.9. *Predicted pose with docking for the purchased compounds of the 4 iterations.*

Of note, the binding mode of the two most potent compounds **9** (IC_{50} = 30 μM) and **23** (IC_{50} = 26 μM) could be determined by X-ray crystallography (**Figure 3.10;** PDB codes **6ZEL** and **6ZF9**) and is in agreement with the predicted binding mode (**Figure 3.11**). Compound **23** is of particular interest because, like compound **3** in the first iteration, it represents a drastic scaffold transformation that is far from obvious, presents an excellent ligand efficiency (0.38) and has never been described as bromodomain ligand. As such, it opens the opportunity to develop a whole new family of compounds against this target.

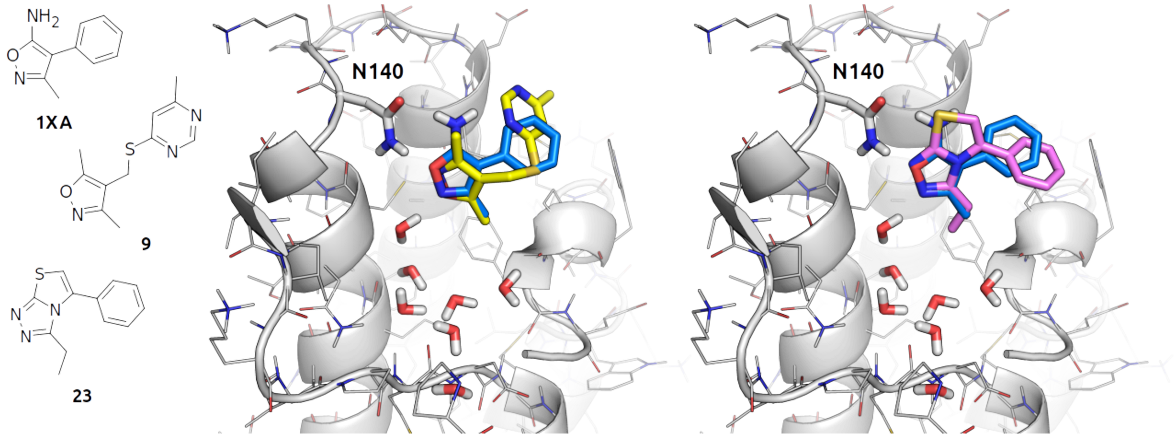


FIGURE 3.10. *X-ray crystallography binding mode of* ***9*** *(left) and* ***23*** *(right) compared with* ***1XA****.*

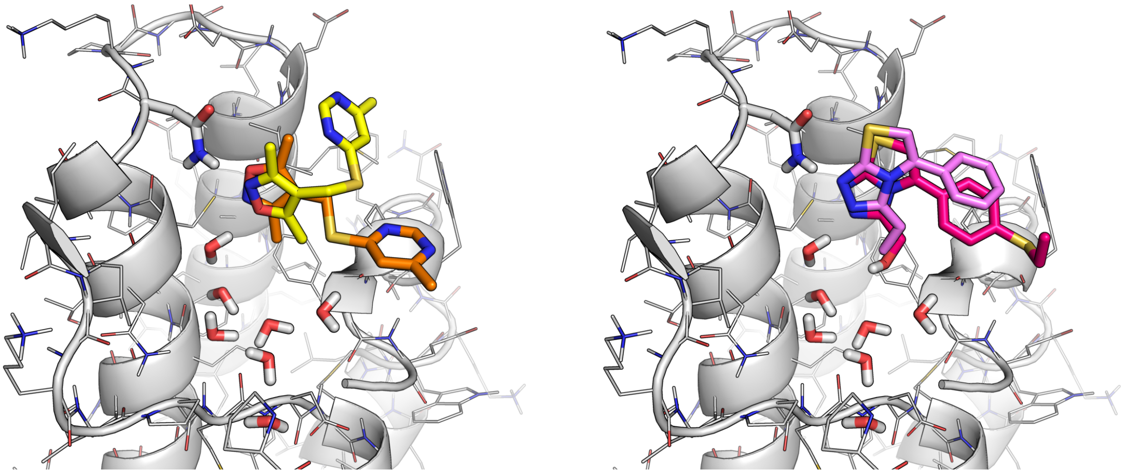


FIGURE 3.11*. X-ray crystallography binding mode of 9 (left; yellow) and 23 (right; pink), compared to their respective computational predictions (in orange and fuchsia, respectively). As compound 23 was synthesized as a representative example of an abundant family of virtual hits (but was not itself present in the virtual collection), the binding mode is compared to a representative of this family (I4_8).*

Structural drift and chemical navigation.

As discussed, one of the salient features of our pipeline is the ability to identify non-obvious analogues of the initial fragment hit. As expected, the similarity of the selected molecules with the reference decreases from iteration to iteration. Notably, compound **23** is only 37% similar to 1XA, and could not be found in a standard search for analogues. Another common strategy to evolve fragments consists in performing substructural searches, which may enable larger modifications.

To study the ability of our protocol to explore non-obvious changes also from a substructural perspective, we considered the scaffolds of the selected molecules, primarily as Bemis and Murcko (BM) scaffolds (where terminal side chains are removed)[36] and as Atom-Generic (AG) scaffolds (atoms replaced by carbons)[37]. (**Table 3.3 and Figure 3.12**).

***TABLE 3.5**: AG scaffolds found 4 or more times in the 200 selected molecules. Occurrences of the scaffolds are shown by iterations, total and on the molecules that have been tested experimentally.*

Cluster	i1	i2	i3	i4	Total	Tested (cpd ID)
1	24	6	10	2	42	2 (**1XA**, **11**)
2	6	4			10	1 (**5**)
3	3	5	4		12	0
4	2	2			4	1(**12**)
5	1	4			5	2 (**2, 6**)
6		14	3	3	20	4 (**7, 9, 13, 20**)
7		2	1	2	5	0
8			3	17	20	1 (**23**)
9			2	2	4	2 (**14, 17**)

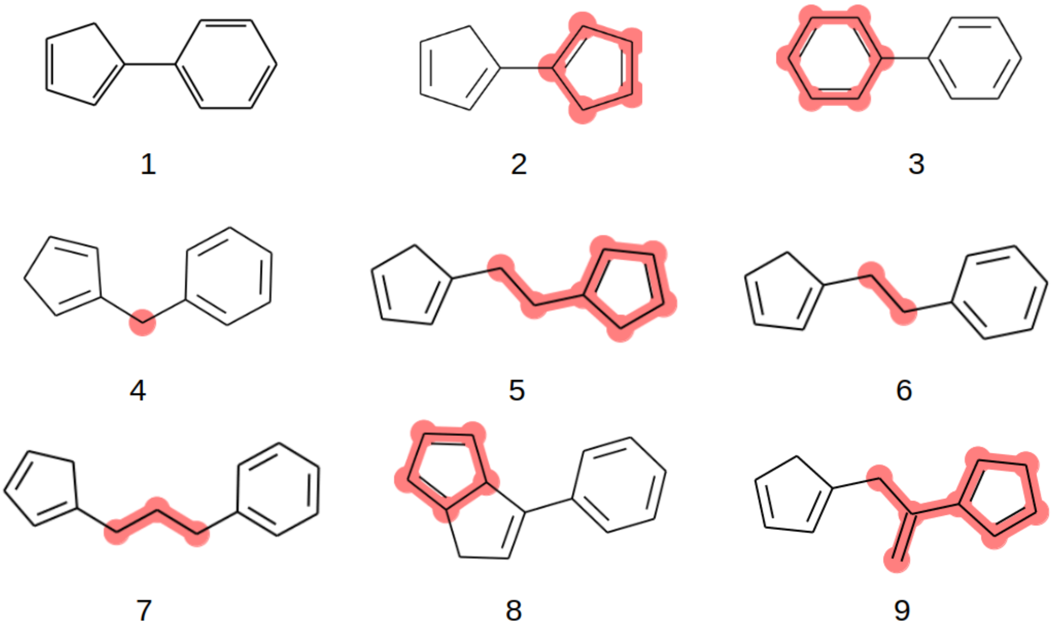


Figure 3.12*: Clusters of AG scaffolds with the highest number of members. In red is shown the difference of the scaffold from the original one.*

The topology of the original fragment (5-membered ring directly bound to a 6-membered ring) is the most prevalent, with 33 cases (17%) (**Figure 3.13**), of which 18 cases in the first iteration. But only 8 molecules retain the specific molecular framework of 1XA (isoxazole directly bound to benzene). The remaining distribute into 10distinct heterocyclic arrangements. Such isosteric replacements already represent significant scaffold hops, but the majority of the selected molecules (83%) represent larger transformations. In the first iteration, 16 cases (32%) correspond to moderate topological transformations, such as change of ring size (from 5-6 to 5-5 and 6-6) or extension of the ring-ring connection by 1 atom. The 16 remaining molecules (including compound 3) represent individual topological scaffolds, with a notable presence of fused ring systems. This analysis confirms that a substantial proportion of the selected molecules can be considered remote analogues that would not be retrieved in a typical search based on the original fragment. Furthermore, our protocol selects a very diverse set of molecules (52 AG scaffolds), thus maximizing the probability of finding suitable novel scaffolds.

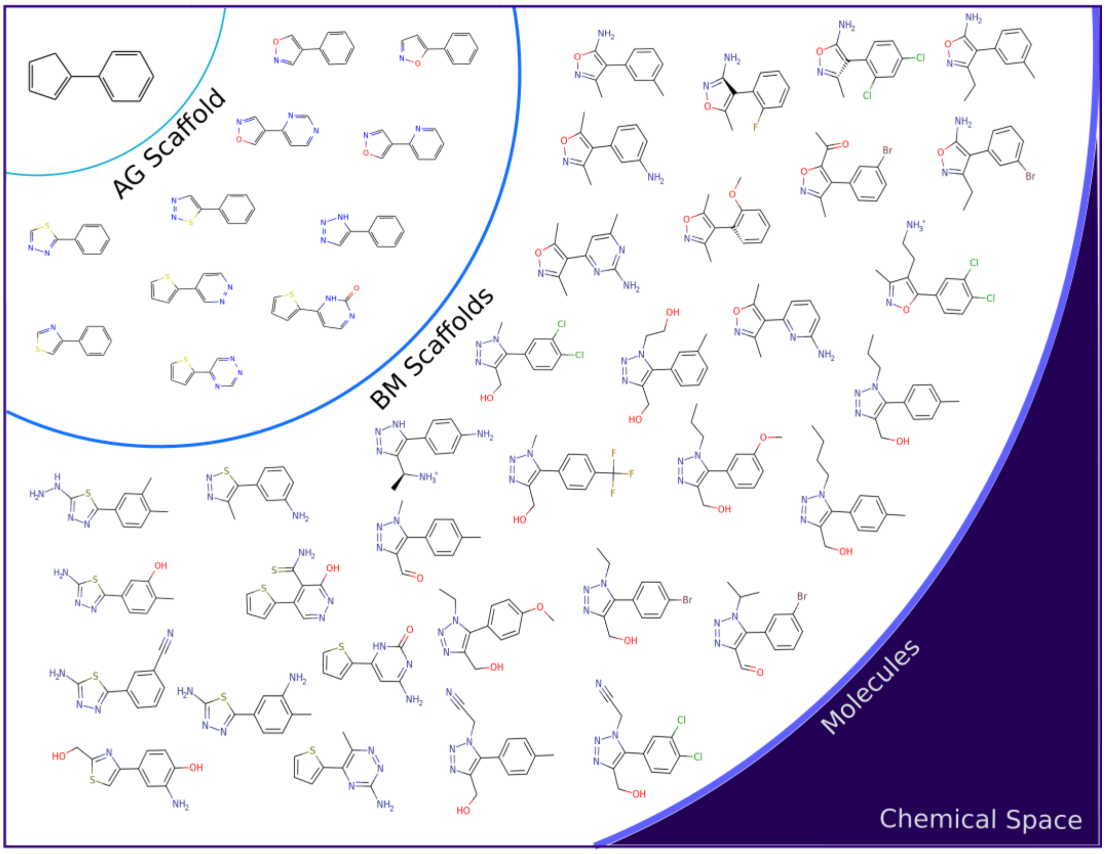


***Figure 3.13**: Ensemble of different molecules obtained from the AG scaffold of 1XA. 11 BM scaffolds are found and a total of 33 molecules derived from them.*

Successive iterations take the search even further away. The 5 more frequent topological scaffolds in iteration 1 remain in iteration 2, but with decreased frequency (70% vs 34%). Interestingly, in iteration 2 the most populated topology, with 14 cases (28%), contains a two-atom linker between the aromatic rings (compound 9 belongs to this category), confirming a gradual drift into more substantial topological modifications. The trend continues in iteration 3, which represents an exploratory step, with 28 compounds (56%) corresponding to topological singletons. This highlights that in the absence of any constraints, independent walkers exploring a large chemical space from the same starting point would finally lead to a scatter of unrelated molecules. We curb this tendency by considering only those molecules with a larger number of parents in common. In this way, the number of topological singletons in iteration 4 retains a similar proportion (48%). But in this case the algorithm has identified a privileged topology, represented by compound 23, with 17 members (34%). In terms of molecular frameworks, our protocol explores distinct structures in iterations 2, 3 and 4, respectively. Again, confirming the scaffold diversity of the selected set. Albeit the number of compounds that could be purchased is relatively low, they are representative of the overall set of selected compounds, with 11 molecules covering 6 out of the 9 most populated topological scaffolds (**Table 3.5**) and 11 additional singleton topological scaffolds.

4.1.2 Analysis of Water Network of BRD4(1) with a Novel Series of Ligands

In the 4th iteration of the previous study, we noticed that rDock program and the DUck method score well those molecules containing a polar or positively charged substituent close to the waters at the bottom of the cavity of BRD4(1). While, in principle, this makes sense (water molecules should favour polar interactions), the majority of known BRD4(1) ligands, starting with the natural substrate (acetyl-lysine) contain a hydrophobic substituent in that position (**Figure 3.14**).

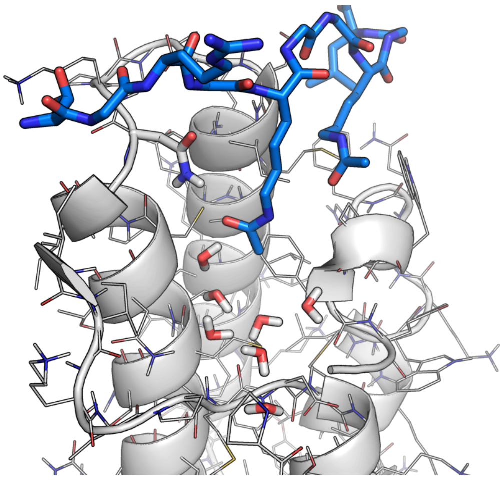


Figure 3.14: *Brd4(1) binding to a H4 peptide with two acetyl lysine.*

This indicates that, while counter-intuitive, these water molecules create a hydrophobic environment. As the synthetic protocol used to synthesize compound 23 allows for the introduction of various functional groups in that position, we decided to synthesize derivatives of 23 with different substituents near the waters (**Figure 3.15**). Salvo Scaffidi and Carmen Escolano's group carried out the synthesis.

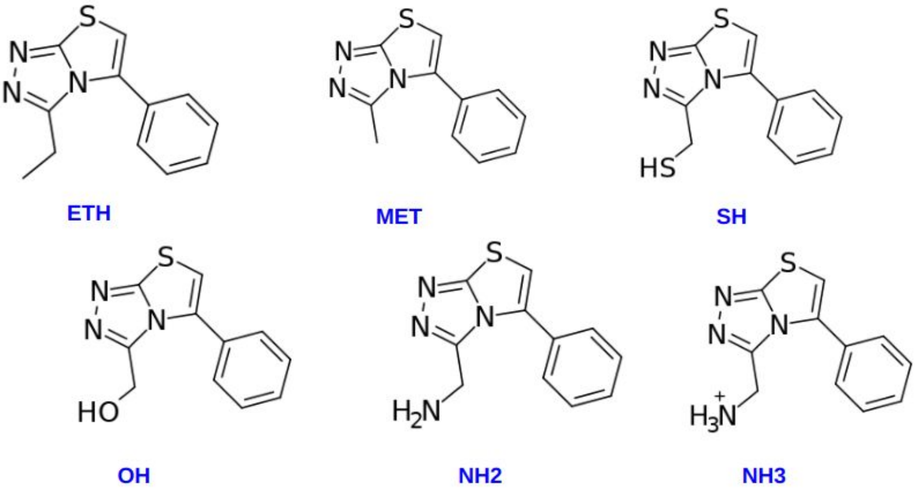


Figure 3.15: *Ligand series with the different substituents interacting with the network of waters. Ethyl: ETH; Methyl: MEH; Ammine: NHH; Ammonium: NH3; Thiol: SHH.*

If we hypothesize that the molecules will have the same binding mode as 23/ETH, the different substituents would be close to the waters of the cavity and we can analyze if a polar substituents can in fact stay there, and if it will provide more stability to the specific compound.

The amine derivatives can exist in mainly two states at physiological pH (**Figure 3.16**). The protonation gives to the nitrogen different electronical properties that influence the environment around it, in the specific case they could influence the waters of the cavity. For this reason the two states are considered as different molecules in this study.

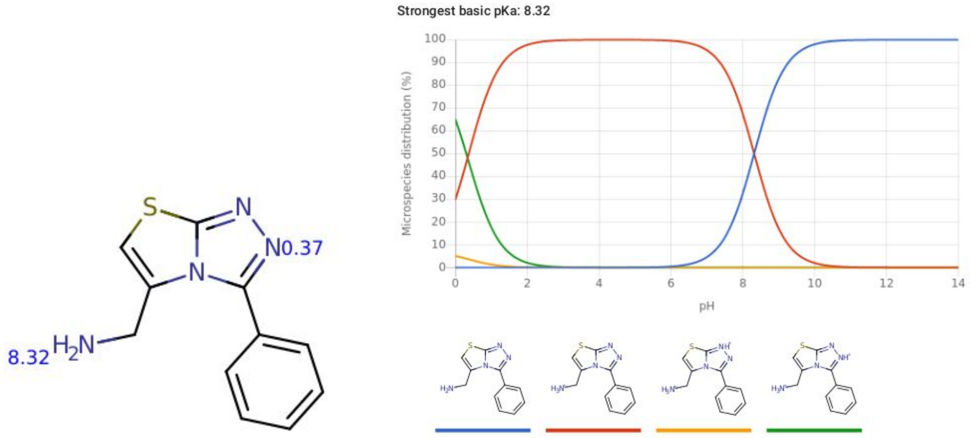


Figure 3.16: *Protonation of compound NH2-NH3 as a function of pH (source: https://chemicalize.com)*

We run docking calculation with both rDock and Glide to see if the scoring problem was only related to rDock. The results was that both programs preferred the polar (hydroxyl and amine) derivatives and charged (ammonium) derivatives close to the water molecules (**Figure 3.17**).

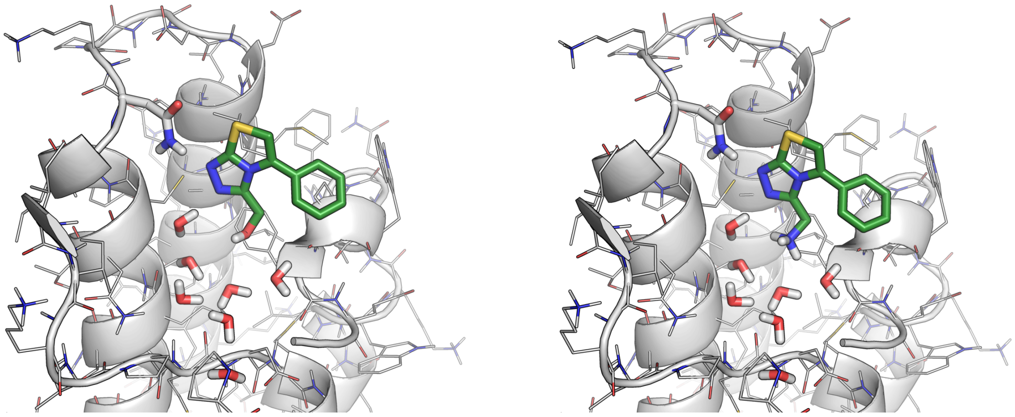


Figure 3.17: *docking pose of OHH and NH3*

We could obtain the crystals of the Ethyl (ETH), Methyl (MET) and Hydroxyl (OH) substituents. The Methyl- and Ethyl-substituted molecules place the hydrophobic substituent as it was expected but the Hydroxyl-substituted molecules prefer instead a reversed pose **(Figure 3.18**).

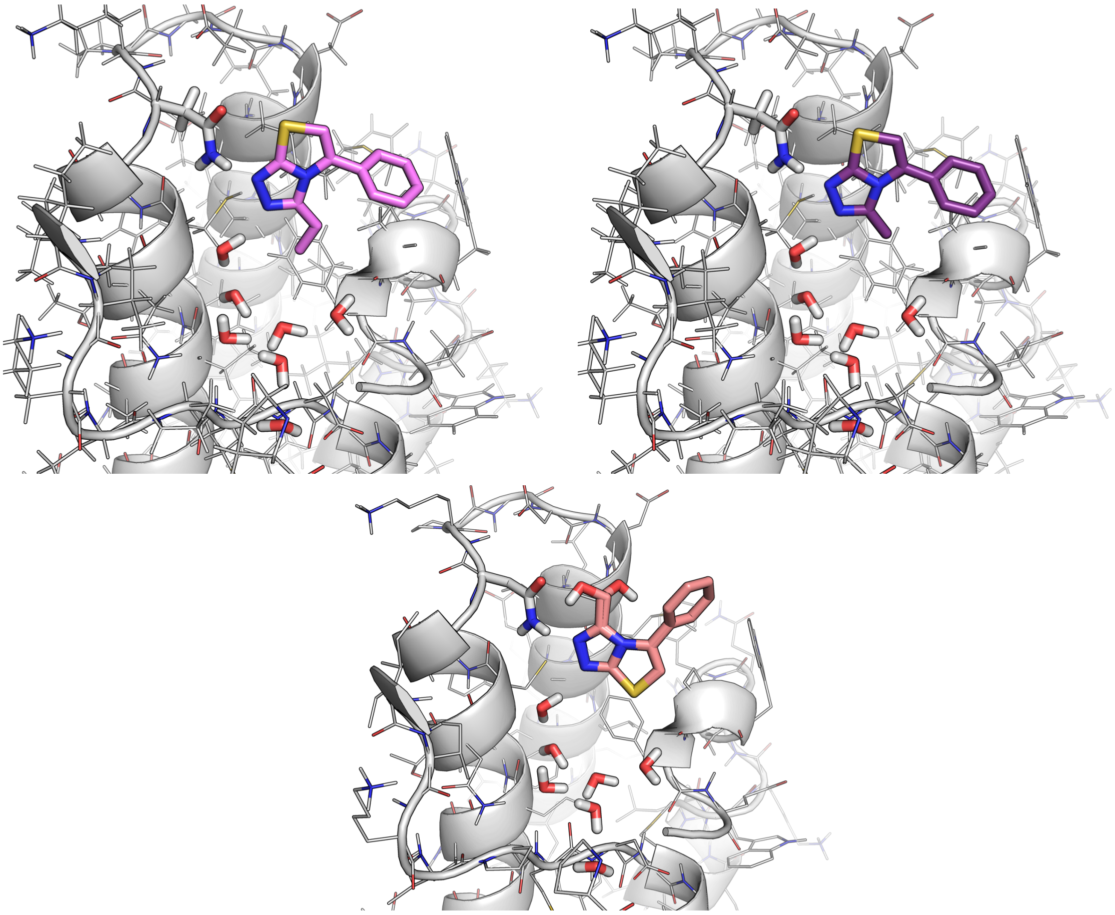


Figure 3.18: *Crystal structures of ETH (left) and MEH (right) and OHH (below)*

This demonstrates not only that a polar substituent in that position is not favourable, but that it's even repulsed, changing completely the position of the molecule. From this result we can predict that the ammonium-substituted molecule will have the same preferred pose as the Hydroxyl molecules. As a proof, a simulation of Methylammonium starting from the position near the waters showed that the molecule rapidly leaves the cavity (data not shown).

4.1.2.1 Prediction Of Relative Affinity

We could assess the experimental IC50 of ETH as 26uM. To predict the relative affinity of the derivatives we decided to perform free energy calculations. The comparison of these results with experimental analysis will also be a confirmation that it is not a problem of forcefield but only of docking score approximations.

After the free energy calculations studies of the molecules were performed, it was confirmed that the preferred substituents in vicinity to the waters are hydrophobic ones (**Figure 3.19**).

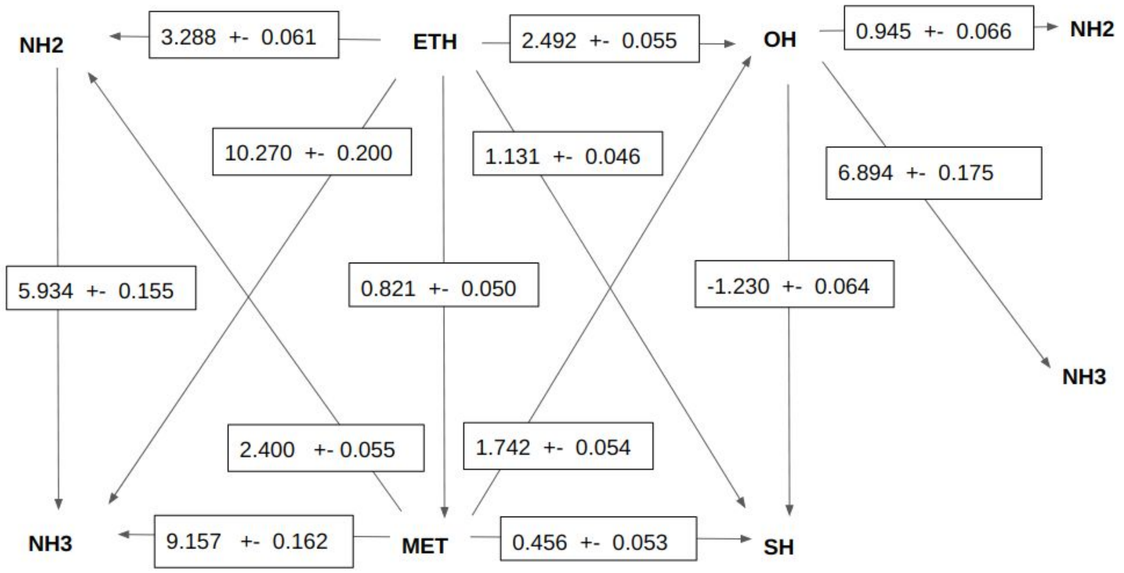


Figure 3.19: *Relative Free Energies of the series.*

According to these studies we can also prepare a ranking of affinity where hydrophobic substituents (ETH, MET, SH) have higher affinity than polar (OH and NH2) and that NH3+ can be considered inactive (**Figure 3.20**). From these results we can also predict the relative Kd outcome of the experimental results (**Figure 3.21**).

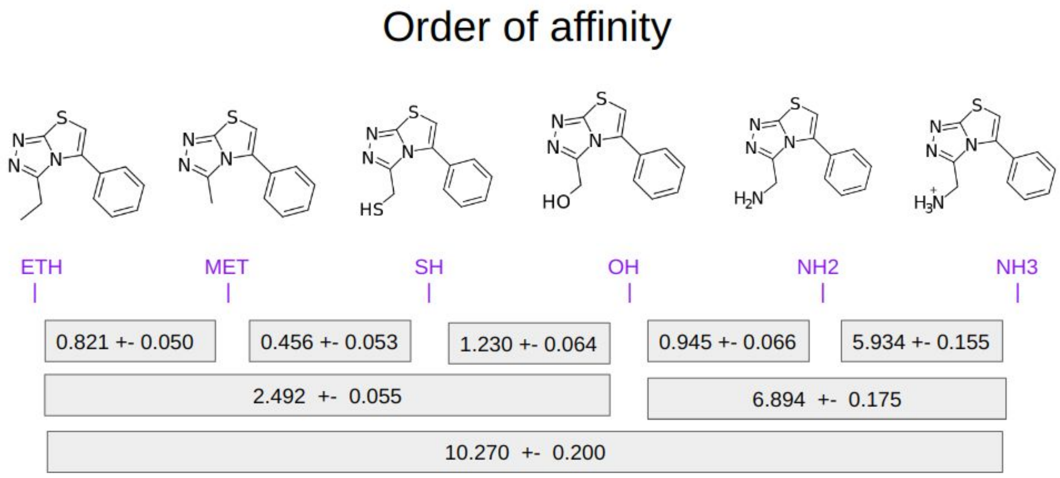


Figure 3.20: *Order of affinity of the series according to the calculated energies.*

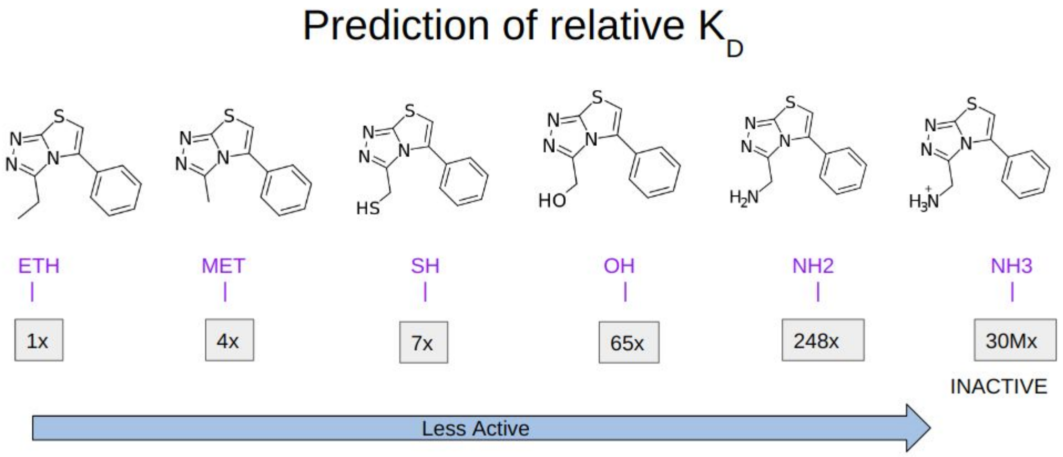


Figure 3.21: *Prediction of relative Kd*

Salvo Scaffidi from our group performed TR-FRET of this series and the results are in **Table 3.6**.

Table 3.6: *TR-FRET results*

Compound	IC50
SSR4/ETH	26 uM
SSR3/MET	54 uM
SSR2/OHH	65 uM
SSR6/NH2-NH3+	157 uM

From these results we can see that the predicted affinity order of the series is confirmed, but that polar substituents present more affinity than predicted. This could be due to the different binding mode (as stated by the crystal structure of OH). To study this free energy calculations starting from the other binding mode can be performed in the future.

But the question we want to answer is not why polar derivatives have good affinity, but instead why hydrophobic derivatives have so good affinity.

Another analysis that can be performed is how these water molecules behave in different environments.

4.1.2.2 Structural Changes Influence Water Occupancy

In free energy calculation the focus is centered on the behaviour of the ligands when interacting with the protein in a dynamic situation. But in these types of calculations the physical properties don't reflect a real situation (“alchemical calculations”) so in order to understand the real behaviour of water molecules we have to run “real world” simulation, that is Molecular dynamics simulations.

In this case we will focus on the behaviour of the water molecules in the cavity of BRD4(1) and the ligands are just considered as a disruption in the environment around them.

From the analysis of the trajectories we can see that the binding of hydrophobic ligands stabilizes the BC and ZA loops of BRD4, while polar and charged ligands destabilize these loops as it happens in the apo structure (**Figure 3.17**), with the charged ligand with a more similar behaviour to the apo (trajectories of the charged group are analysed until the ligand remains in the cavity)

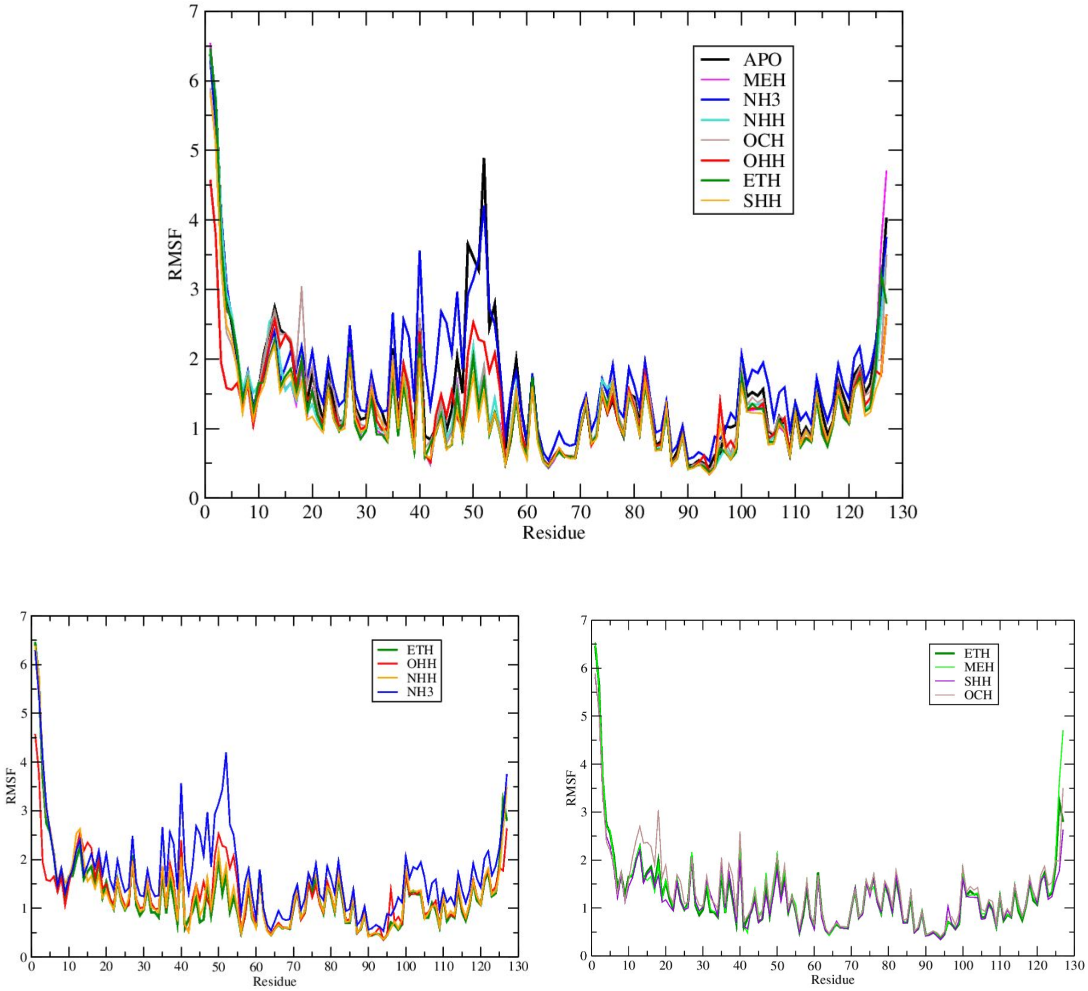


Figure 3.22: *RMSF per residue for all the ligands and the apo protein (a); only ETH and polar ligands (b); only hydrophobic ligands (c).*

To understand how this destabilization influences the water, we can analyse the interchange of waters in the cavity. Only water closed to the aminoacids of the cavity (**Figure 3.23**) were considered.

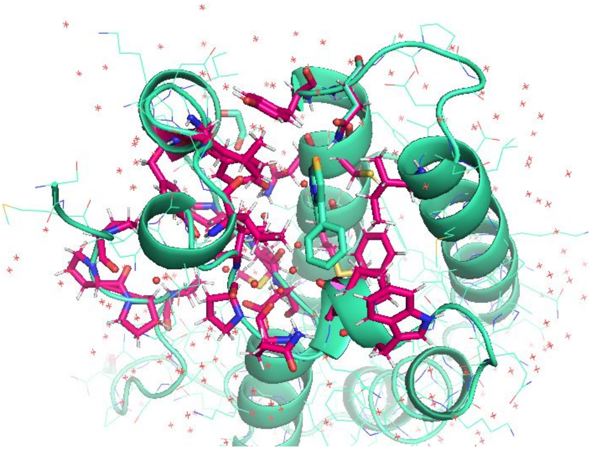


Figure 3.23: *Aminoacids composing the cavity of BRD4(1)*

We can see that when a ligand like ETH is bound, some of the water molecules are interchanging less with the bulk (they are present during more frames) with respect to the apo protein. When the ligand is charged (NH3+), the behaviour is similar to the apo protein (**Figure 3.24 and Table 3.7**).

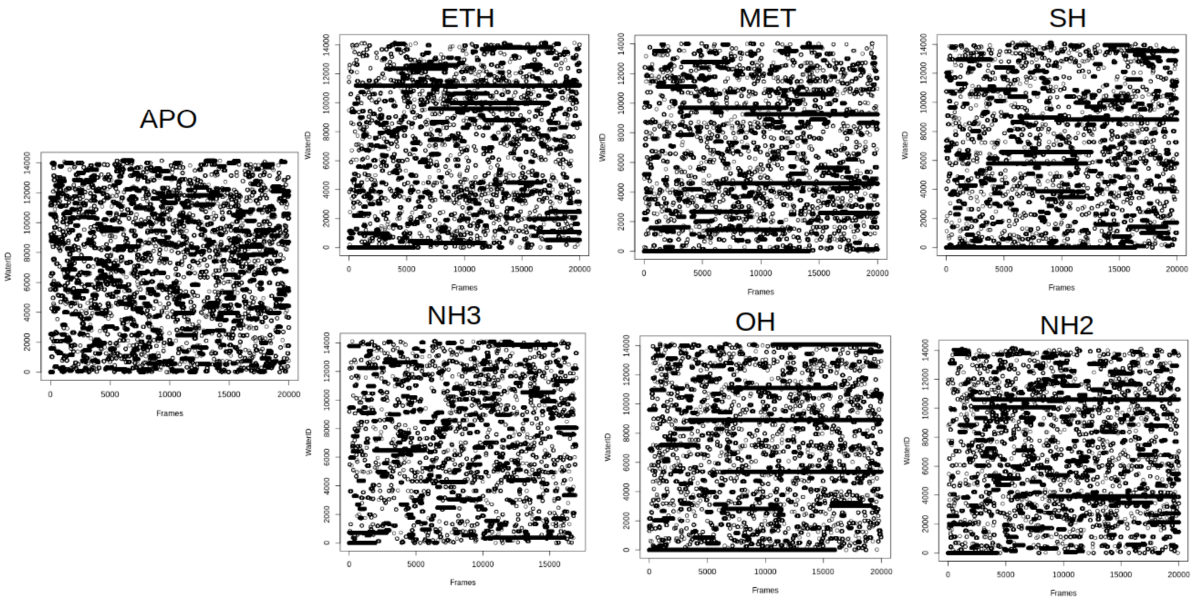


Figure 3.24: *Interchange of water molecules in the cavity in the presence of the ligands and in the apo protein during 200ns of MD simulation. With the charged ligand (NH3). only the first 170ns are considered because the ligand is leaving the cavity afterwards.*

Table 3.7. *Measure of "water trapping": how many frames in the MD trajectory , the same water molecule is found in the cavity. For the APO protein simulation, in presence of the ligands (200ns), and in presence of NH3+ (before the ligand leaves the cavity, 170ns).*

Water Ranking	Number of Frames						
	APO	ETH	MET	SH	OH	NH2	NH3
1	2296	19414	14045	16367	16191	17810	5359
2	1813	8408	13761	10356	15994	10802	4483
3	1731	7601	11247	8516	9224	6937	3792
4	1587	6176	9020	7502	8849	4244	2800
5	1568	5622	6131	6932	8045	4103	2373
6	1550	5422	4940	4771	6404	2856	2350
7	1535	5196	4917	4142	5908	2789	2251
8	1521	4272	3910	3448	5065	2574	1945
9	1471	4046	3527	3346	5051	2566	1864
10	1392	3309	3205	3283	3370	2372	1525

If we analyse the trajectories of these interchanging water we can see that there are mainly two directions of interchange: from the top of the cavity (mainly from the ZA channel direction when a ligand is bound) and from a small tunnel directly connecting the bottom of the cavity to the side of the protein. When a ligand is tightly bound the interchange from the top is limited and the only direction is from the other tunnel. This tunnel is so small that only 1 molecule of water can fit in the diameter. This means that in order to have interchange between the water of the cavity and the exterior, the waters have to go away orderly one by one. The waters closer to the asparagine have in fact the more occupancy. When the ligand is hydrophobic there is no disruption of this mechanism because the tunnel is maintained and the interchange mechanism from this side is kept. When the ligand is polar the water molecules are obligated to interact with the polar substituent, disrupting the network of interactions between them and obligating the protein to move also its loop to fit. This mechanism is less energetic favourable and the ligand prefers to change position instead (hydroxyl case). This "water trap" works like a magnet for hydrophobic substituents, explaining also the high specificity of bromodomains for acetyl lysine and not for lysine or other polar aminoacids.

4.1.3 Fragment Growing Of A Novel Scaffold For Brd4(1) With *De Novo* Virtual Synthesis

After confirming that 23 is the most active substituent of the series, we want to grow this fragment to increase the affinity by adding more functional substituents.
During a stay in Matthias Rarey's Lab in the University of Hamburg I used their program NaomiNext to do fragment growing with a de novo approach. NaomiNext [106] uses a list of 58 known reactions stored as "Reaction Smarts". When given a ligand bound to a structure and a database of building blocks, the program automatically defines the possible reactions to apply to the ligand and the respective building blocks that are possible to use. Once the "virtual reaction" is made, a series of conformations of the new added part are generated and evaluated with a score. The conformation with the best score is taken as the final pose and afterwards the new generated compounds can be ranked by this score.

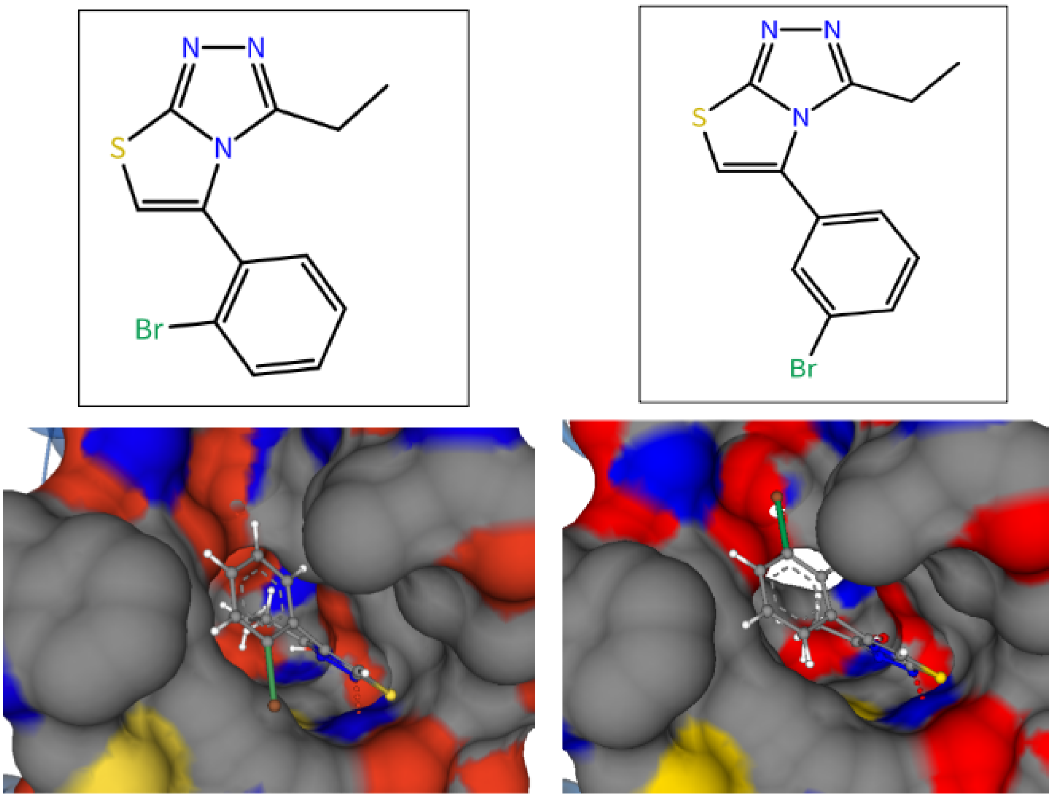
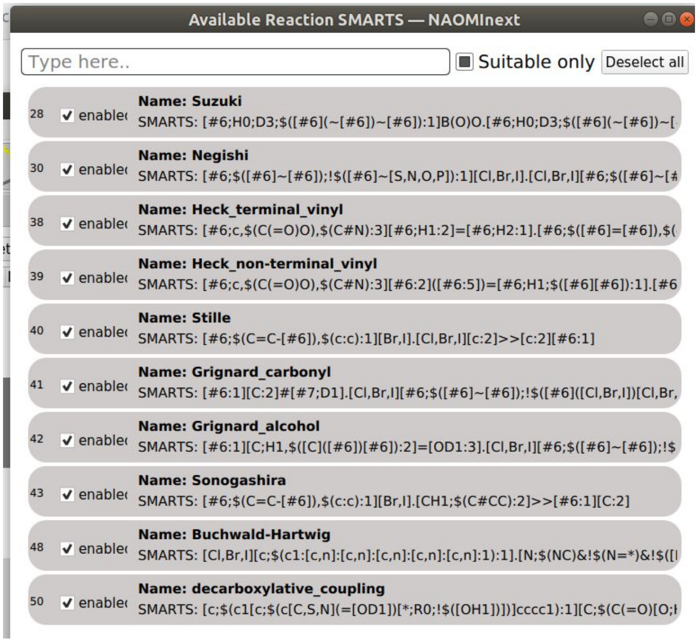


Figure 3.25: *Bromide "decoration" added in position ortho (left) and meta (right) of the phenyl ring.*

The crystal obtained in the previous study (6ZF9) was used for the correct pose of the ligand. 23/ETH was not recognized directly as "reactive" by the program, so a Bromide was added as a "decoration" to the phenyl group to obtain the following compatible reactions: Suzuki, Negishi, Heck, Stille, Grignard, Sonogashira, Buchwald-Hartwig, decarboxylative coupling. All of them have a halide as one of the reactants. Consulting with the medicinal chemists that are performing the synthesis of the final compounds (Carmen Escolano's group), they confirmed that this intermediate could be synthesized with the same methodology used to obtain 23, because the bromine was not expected to interfere in the previous reaction steps. Two vectors of growing, ortho and meta, were found considering the crystal pose, whereas the other vectors are blocked by a tryptophan or by the wall of the cavity. The Br-derivatives of ETH in ortho and meta (**Figure 3.20**) were prepared with MOE[183].

NAOMInext [106] installation includes a set of building blocks from ChembridgeBB [190] in SMILES. This set includes 17998 molecules and was used for the calculation. The database was divided in 9 files of 2000 molecules to ease the calculation.

All the 10 reactions compatible with Bromide were enabled (**Figure 3.26**)



```
28  ☑ enable  Name: Suzuki
              SMARTS: [#6;H0;D3;$([#6](~[#6])~[#6]):1]B(O)O.[#6;H0;D3;$([#6](~[#6])~[
30  ☑ enable  Name: Negishi
              SMARTS: [#6;$([#6]~[#6]);!$([#6]~[S,N,O,P]):1][Cl,Br,I].[Cl,Br,I][#6;$([#6]~[#
38  ☑ enable  Name: Heck_terminal_vinyl
              SMARTS: [#6;c,$(C(=O)O),$(C#N):3][#6;H1:2]=[#6;H2:1].[#6;$([#6]=[#6]),$(
39  ☑ enable  Name: Heck_non-terminal_vinyl
              SMARTS: [#6;c,$(C(=O)O),$(C#N):3][#6:2]([#6:5])=[#6;H1;$([#6][#6]):1].[#6
40  ☑ enable  Name: Stille
              SMARTS: [#6;$(C=C-[#6]),$(c:c):1][Br,I].[Cl,Br,I][c:2]>>[c:2][#6:1]
41  ☑ enable  Name: Grignard_carbonyl
              SMARTS: [#6:1][C:2]#[#7;D1].[Cl,Br,I][#6;$([#6]~[#6]);!$([#6]([Cl,Br,I])[Cl,Br,
42  ☑ enable  Name: Grignard_alcohol
              SMARTS: [#6:1][C;H1,$([C]([#6])[#6]):2]=[OD1:3].[Cl,Br,I][#6;$([#6]~[#6]);!$
43  ☑ enable  Name: Sonogashira
              SMARTS: [#6;$(C=C-[#6]),$(c:c):1][Br,I].[CH1;$(C#CC):2]>>[#6:1][C:2]
48  ☑ enable  Name: Buchwald-Hartwig
              SMARTS: [Cl,Br,I][c;$(c1:[c,n]:[c,n]:[c,n]:[c,n]:[c,n]:1):1].[N;$(NC)&!$(N=*)&!$([
50  ☑ enable  Name: decarboxylative_coupling
              SMARTS: [c;$(c1[c;$(c[C,S,N](=[OD1])[*;R0;!$([OH1])])]cccc1):1][C;$(C(=O)[O;
```

Figure 3.26: *Reactions compatible with Bromide.*

4.1.3.1 Rescoring and Selection

The program generated 10 poses per BB. I used KNIME workflow [191] to sort the results from NAOMInext by score in ascending order (nodes: SDF reader, Sorter, GroupBy) and select the top scoring pose for each BB. Some BB were quite big so only resulting molecules below 30 heavy atoms were considered and only molecules with a score below -28 (the score of SSR4/ETH/23).

The resulting number of molecules are shown in the **Table 3.8**. At the end I had 7346 resulting molecules for the ortho-Br and 7508 for the meta-Br.

The program Seesar was used to rescore the results. SeeSar [192] is a program developed by BioSolveIT for visual compound evaluation in ligand optimization. It contains different scores for affinities, physicochemical properties and torsional analysis of the bonds. It uses some tools developed in Matthias Rarey's Lab, among them the Hyde scoring function [193]. It is a function based on physical properties like hydrogen bond and dehydration energies and it gives scores of affinities either by atom or for the whole molecule.

After selecting only the molecules below a Hyde score of 1000nM and visual inspection, a final number of 53 compounds for the ortho-Br and 38 for the meta-Br were selected. Of these, 85 compounds were obtained with Buchwald-Hartwig reaction, 4 with Negishi, 1 with Grignard and 1 with Sonogashira reaction.

Table 3.8: Results from NaomiNext run.

	original smi	**ORTHO**	**META2**
TOT	17998	133054*	135198*
Top score conf. x mol		12126	12353
<30 HAC		7481	7562
< -28 Score**		7346	7508

**10 poses x BB. Theoretical number= 200.000*

*** better than the original mol.*

Considering the number of compounds and the convenience of using a single synthetic procedure, only compounds from the Buchwald-Hartwig reaction were selected to be synthesized. After specific searches in Scifinder [194] to confirm the feasibility of the reaction with the specific building blocks, 3 molecules for the ortho-Br and 4 for the meta-Br were selected for synthesis (**Figure 3.27**).

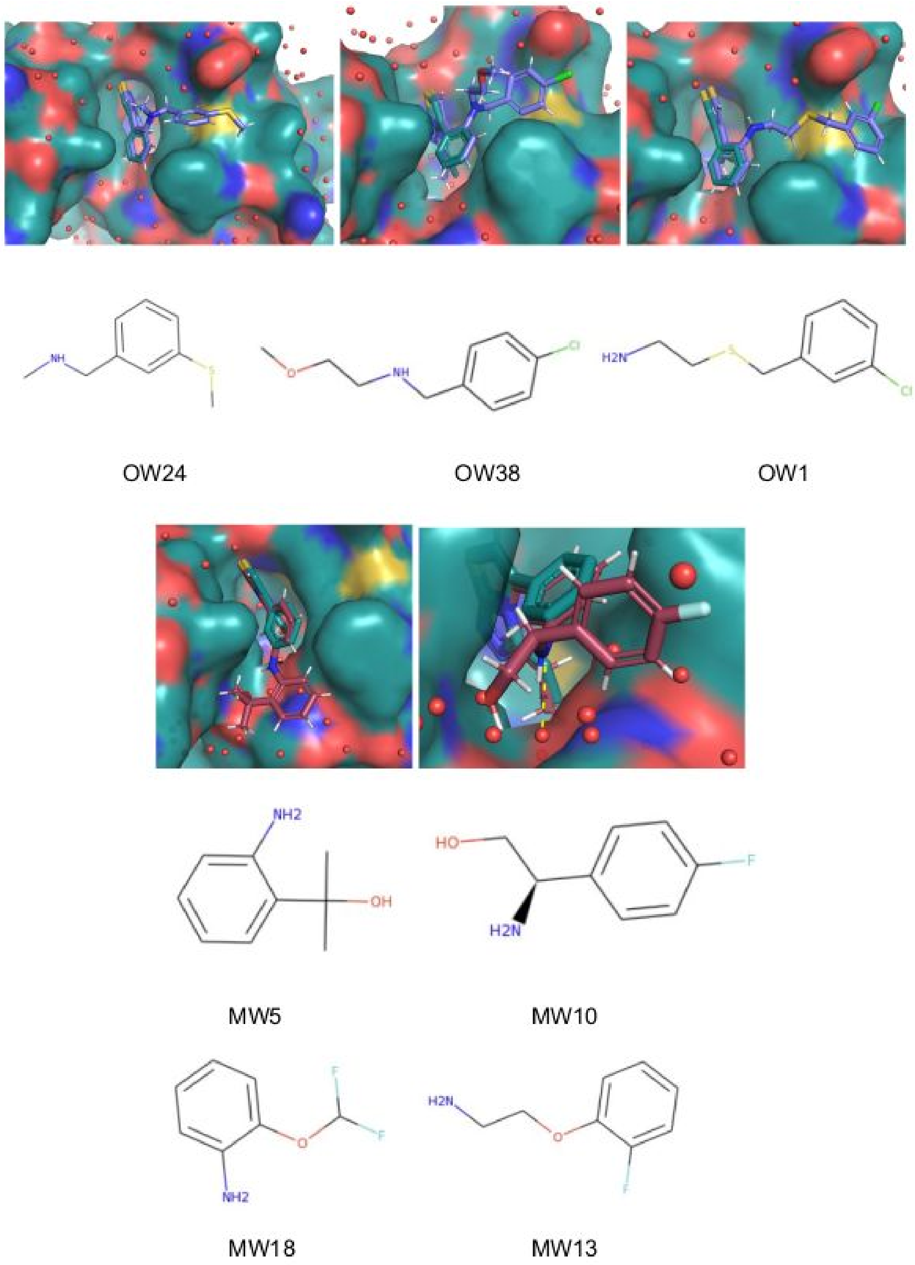


Figure 3.27: *Final Building Blocks for the synthesis de novo of derivatives of 23*

4.1.3.2 Experimental Results

Two compounds were synthetized (SSR11 from BB MW13 and SSR12 from BB MW18) (**Figure 3.28 and 3.29**)

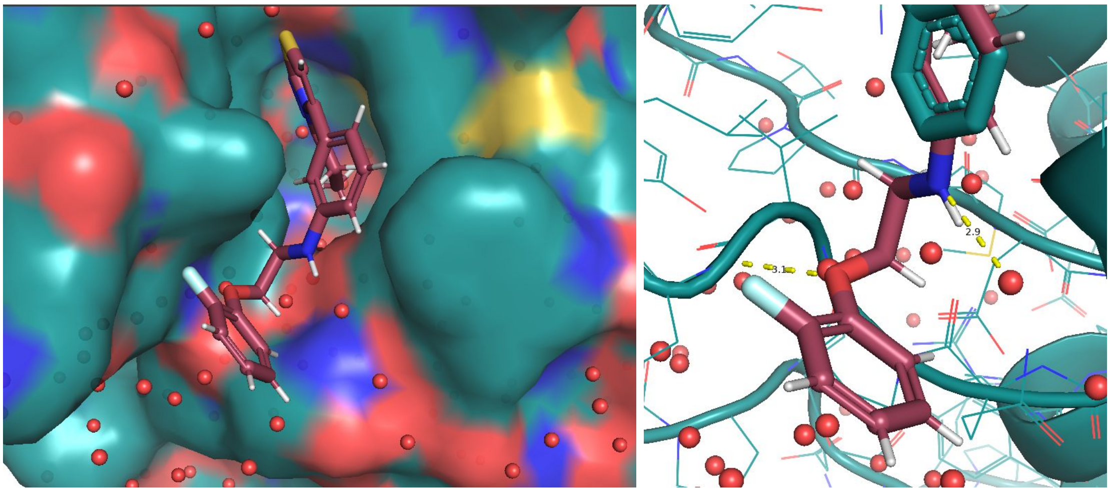


Figure 3.28: Predicted pose of compound SSR11 from BB MW13.

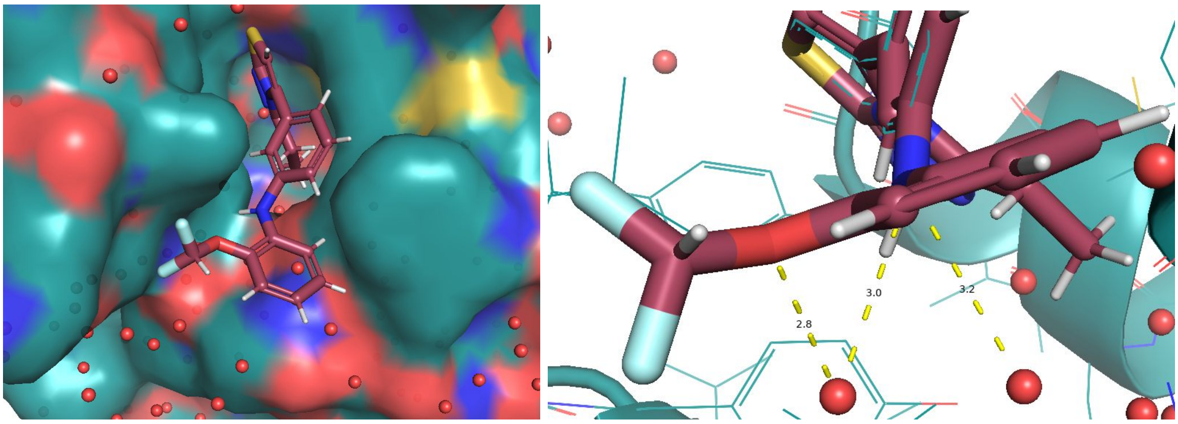


Figure 3.29: Predicted pose of compound SSR12 from BB MW18.

From the TR-FRET experiment, it was measured an IC50 of 9.7uM for SSR11 (comparable to 23) and 40nM for SSR12 (650-fold better than 23).

4.2 TREX2

[redacted]

[redacted]

4.2.1 [redacted]

[redacted]

[redacted]

Figure 3.30: [redacted]

[redacted]

[redacted]

[redacted]

4.2.2 [redacted]

[redacted]

[redacted]

Figure 3.31: [redacted]

[redacted]

[redacted]

4.2.3 [redacted]

[redacted]

[redacted]

Figure 3.32: [redacted]

[redacted]

[redacted]

Figure 3.33: [redacted]

[redacted]

Chapter 5

Discussion

The expansion of the "druggable genome" is a difficult task but is absolutely important to extend the knowledge of biological functions and create new therapeutic opportunities. The study of molecular recognition mechanisms can help to understand how unusual proteins, like BRD4 (with its functional waters) and TREX2 [redacted], work. Understanding the molecular recognition mechanism is important in designing new drugs. In particular, Understanding the role of structural waters in BRD4, either in the bottom of the cavity for the main "anchoring" of the ligand, or the ones present in the ZA channel that is an obvious vector of growing, is crucial to design more active and selective compounds.[redacted]

Navigation of chemical space is paramount in the quest of New Molecular Entities (NME) for known targets like BRD4 as well as to determine the "druggability" of new targets (TREX2). Different methodology can be used: 1) A Fragment-based Drug Design approach, where fragments discovered by experimental methods are used as starting points for a computational drug design strategy 2) [redacted]
In both cases, computational tools (docking, dynamic undocking) help discriminate active molecules from inactive ones, facilitating the navigation of chemical space to areas of activity.

The two projects are very different and need different approaches. In TREX2 the protein has never been targeted for small binders and a study of bindability was needed previously. [redacted] BRD4 inhibitors, instead, are known and can be used to define the cavity and the pharmacophore.
In TREX2 a virtual screening was done [redacted]In the case of BRD4 many inhibitors are already found so a more spread search is needed to find novel compounds. Novelty is what is searched. This can be achieved with bigger databases of compounds but a compromise for only commercial compounds has to be done due to the resources of the lab (no synthetic resources available). For this reason ZINC15 can be used. It contains the double of compounds than our in-house library (15M compounds). Fragment mining and growing are useful in this case. Afterward, hit compounds are chosen and tested with different assays for the two proteins [redacted]For BRD4 the crystal structure of the hits was available and to understand the SAR of the substituent close to the water in the cavity, Free energy calculations were carried out. Since the hit compound chosen for BRD4 was found not available to buy (sometimes it happens that a vendor discharges compounds that don't have high demand, or any other reasons. If it was in ZINC, at some moment it was in-stock) and any derivatives of that were not available, de novo design was carried out.

As we can see there is not a defined pathway for a drug design project but it depends on many reasons. CADD programs can help and guide the different steps. Many limitations that influence the projects are Docking Score, Library chosen, Computational resources availability, Laboratory resources. Some of these are not major problems in pharma companies but can be big issues in academia. For this reason an accurate selection of complementary methods is necessary.

5.1 Navigation Of The Fragment Chemical Space

Fragment screening has emerged in the last decades as a very effective and cost efficient hit identification method for drug discovery. The infrastructure and technical requisites are relatively low, enabling its implementation in big pharma, biotech and academic institutions alike. However, the evolution of fragment hits into suitable leads remains challenging and largely artisanal.

A general FS collection of typical size is 10^3 and, frequently, great importance is given to a diversity-based design of FS collections[72], which aims to provide uniform coverage of the fragment chemical space. But what these collections have in breadth, they lack in depth, meaning that hits should be seen as beacons indicating privileged areas of chemical space, to be further explored. Instead, the diversity around a fragment hit is rarely exploited, and the decision to progress a fragment is generally based on the chemical structure of the particular hit, rather than on the potential of the chemical space that it is meant to represent.

The pressure to attain rapid improvements in potency, combined with the natural bias towards convergent synthesis and scaffold-centric optimisation favours conservative growing strategies over deeper exploration of the privileged chemical space represented by the fragment hit. Indeed, most of the fragment-evolution processes described in the literature preserve or introduce only minor changes to the scaffold of the initial fragment[79] , meaning that the outcome depends crucially on the chemical structure of the starting fragment. Such practices represent a dire loss of opportunities because each fragment in the screening collection is a mere sample of the 10^5 fragments available for immediate purchase,[192] 10^7 fragments that can be synthetized on demand [193] and up to 10^9 theoretically possible fragments.[194] Furthermore, the fragments in a screening collection are often selected based on pragmatic reasons such as commercial availability or cost rather than their potential to deliver leads.[72] Also, fragments in general screening collections must comply with certain physical and chemical properties, such as the “rule-of-three” [718] or high solubility[72], as well as be amenable to synthesis, provide adequate vectors for growth, and lead to a patentable series of compounds, meaning that certain chemical structures that could be perfectly good starting points are not even considered.

We argue that hits should be seen as probes highlighting privileged areas of the chemical space rather than actual starting points. A systematic exploration of the chemical space around each fragment hit could afford novel and non-obvious analogues (including scaffold change) with increased probabilities of being active. Such fragment mining process could be invaluable, because the progression of a fragment into a lead is the slowest and most expensive part of FBDD, and its success depends largely on the quality of the initial hit. Exploring the fragment space in this way could potentially exploit many fragment hits that are considered sub-optimal in terms of ligand efficiency, growth vectors or synthetic accessibility but can ultimately generate better lead compounds. Even for an already grown fragment, such approach could reveal unexplored venues that are neglected in more traditional (synthesis-oriented) fragment evolution processes, which can help finding new chemical matter even for well-studied targets.

We have presented our platform for navigating the fragment chemical space around any specific fragment, rapidly evolving the initial hit into novel fragments that share a binding motif but are structurally diverse. The platform is able to mine a library of compounds for growing at the same time of doing scaffold hopping in an automatic and scalable way. Ideally, the platform starts from a fragment with a known binding mode and uses the structural information given to ensure the main interaction is not lost during the process, giving more probability that the selected molecules will be active.Application of this fast and inexpensive procedure has the potential to uncover many hidden opportunities and improve the overall performance of FBDD.

On a prospective application on a famous fragment target, BRD4(1), we have evolved a known fragment to obtain an array of compounds that are novel and more potent, displaying non-obvious scaffold changes. We carried out 4 iterations and obtained 50 molecules from each of them with a total of 200 molecules. Of this, only 22 were purchasable, even though we used ZINC15 as a library, which contains commercial compounds. In fact, vendors could have removed from their catalogues molecules that were once available, probably for end of stock or less demand or maybe they were not real compounds, but “virtual”. This is an important issue in CADD since a library is usually downloaded and prepared only once, at the beginning of a project, and can be used for many years and different projects. The downloaded version can be a simple photograph of that time and updates should be considered regularly, which is a significant effort. This was specifically dramatic in the case of the family of compounds obtained in iteration 4, of which none of them was available to buy. The high number of members of the family (34% of iteration 4) made clear that we had to test some of them. We decided to synthesize a compound containing the “minimal” scaffold of phenylthiazolo[2,3-c]-1,2,4-triazole.

We tested the 23 compounds with 2 experimental methods. 12 molecules were active in DSF assay, 8 in TR-FRET, and 3 could be confirmed by X-ray crystallization. The total active molecules were 15 which gives a total hit rate of 65%, which is very promising considering that the 22 molecules were chosen by commercial availability, which is close to random. The analysis of the scaffolds of the 200 molecules shows a movement into “islands of chemical space” as can be seen by the family of compounds of scaffold 1 in iteration 1 and 3 with 24 and 10 members, respectively, of scaffold 6 in iteration 2 with 14 members, and scaffold 8 in iteration 4 with 17 members. This may correspond to a “lake of activity” as can be seen for compound 9 in representation of scaffold 6 and compound 23 in representation of scaffold 8. It may be worth exploring these “activity lakes” further by testing more compounds with active scaffolds. We didn’t buy any compounds for scaffold 1 but 1XA is a representative of the family. We were also able to obtain crystal structures for three diverse ligands, confirming the binding mode predicted by the platform which led to the selection of the compounds. After the analysis of known brd4 inhibitors it was also possible to show that many of the active compounds correspond to novel scaffolds for this target, confirming the objective of this study to obtain novel chemical series.

One important limitation of the method is the library selection, that is up to the user. Ideally the library should have a high procurement rate, and databases of commercial collections (Zinc15, Molport, Enamine) are the most adequate. The drawback is that purchasable compounds are dynamic and maybe not purchasable after a while, so the library to use for the platform should always be the most up to date possible. The platform was also designed to theoretically scale up to any library of compounds, including multi-billion libraries e.g.Enamine REAL. In addition, the database should be diverse, or the search will be biased towards the scaffolds that are more numerous in the database.

The platform is also limited by the inherent issues of the docking program and for the iterative feature of the platform, this can lead to error propagation. Consensus scoring using additional programs e.g. DUck or MMGBSA, can overcome this issue. Also, some very interesting scaffolds are not further explored in subsequent iteration (compound 3 as example). The automatic nature of the platform prevents the user to choose which scaffold weighs more in the search. This issue will be addressed in future implementations.

The step-wise approach (1 iter at a time) is useful in the exploration maintaining ligand efficiency, but we would be missing the opportunity of more drastic jumps. In further implementation this can be decided by the user. In future implementation we will also apply to other targets.

5.2 Hydrophobic Behaviour Of Structural Waters In Brd4

It is known that bromodomains contain a network of 4-7 structural waters in contact with the methyl of Acetyl-lysine and with the hydrophobic moiety of ligands. Drug design efforts in the last 10 years have been either to acknowledge the preference for hydrophobic substituents from SAR analysis or try to displace them with little success [121]. Some researchers have tried to understand the behaviour of these waters in the Apo protein with computational methods [195], focusing on the displaceability of the single waters or the network as a whole but not taking into account the effect upon ligand binding.

We have analyzed with a series of compounds the bizarre hydrophobic behaviour of the structural water molecules of BRD4(1) by computational methods. The compounds contained different substituents that were selected according to the electronic nature, namely hydrophobic (Ethyl, Methyl, Thiol), polar (Hydroxyl, Amine) and positively charged (Ammonium). Free energy calculations confirmed the preference for hydrophobic ligands instead of polar or charged. Within the hydrophobic group Ethyl was preferred to Methyl and to Thiol, meaning that a complete hydrophobic substituent (Ethyl) which fill the hollow created between the waters and the wall of the cavity is more suitable in this case, while Methyl fill only partially and the thiol has a mix behaviour. Longer chains were proved to be also suitable (Propionylated, Butyrylated and Crotonylated Lysine) [196] but are less active than the Acetyl lysine for BRD4. In another SAR exercise also unsaturated chains in a different scaffold [33] were tested and none showed better affinity than the methyl. So we can consider that the Ethyl with our scaffold is the perfect match for BRD4 cavity and longer chains won’t improve the affinity.

Hydroxyl and Amine substituents have similar affinity, with hydroxyl slightly better, maybe because the electronic similarity with the water molecules favour more stable interactions with the network, and also the single hydrogen bond can favour a geometrical stability, while for the amine more rearrangement is needed by the water network to fit both hydrogens. The

charged amine has a very low affinity that can even be considered inactive. The burial of the charge is not favoured and the water network doesn't compensate for the loss of the solvent layer present in the bulk. In fact, in molecular dynamics simulations the ligand leaves quickly the cavity.
With Molecular dynamics simulation it was possible to understand that the presence of ligands with hydrophobic substituents were stabilizing the ZA loop and “trapping” the waters in the cavity, while polar and moreover charged ligands were disrupting the network of water and allowing the loop to move, shifting the equilibrium to an apo-like structure of the protein.
It seems that the increased entropy from the loop movement is not enough to justify the loss of enthalpy from the break of the interaction of the water network. In fact it is a strange behaviour that in order to have a more favourable binding for the ligand the water must lose some degrees of freedom to remain more or less still in the cavity.

The combination of all these events is favoring hydrophobic substituents instead of polar ones in the cavity of BRD4.

5.3 Virtual Synthesis Of New Compounds For Brd4

Once we have proved that the Ethyl-substituted fragment **23** is the most active ligand of the series, in order to increase the affinity we needed to grow this optimized fragment to a lead-size ligand. Since there are no commercial derivatives available, we needed to synthetize them. For this task we decided to use a program for virtual synthesis to create molecules from de novo.
The program NaomiNext allowed to grow fragment **23** from two different vectors in the phenyl ring (ortho and meta) adding commercial building blocks by means of a set of known reactions, thus considering synthetic accessibility, while, at the same time, it was predicting the binding pose of the newly created molecules and assessing the score of binding. Seven Building blocks were chosen at the end of the exercise and 2 molecules were synthesised. SSR11 showed similar affinity as the original fragment but SSR12 was 650-fold more active. The binding modes of the two molecules were quite different, with SSR11 displacing 2 water molecules (W1,W2) present in the ZA channel, substituting one interaction of W1 to the protein with the oxygen of the methoxy group, and interacting with another water molecule (W3). SSR12, instead, is interacting with W1 and W2 and displacing W3. The 2 molecules contain very similar chemical parts but distributed in a different way, and this distribution is extremely significant for the affinity, confirming the importance of the ligand architecture explained in the section *1.2.2.*
The importance of the water molecules in the ZA channel was already assessed [197] and our study confirms that it is preferable to use these water molecules as a bridge between the ligand and the protein instead of displacing them. This confirms the importance of considering not only “proper” structural water molecules in the design of ligands but also the first layer of solvation of the protein even though they are solvent exposed, especially if they are interacting with protein hotspots, contributing most to the free energy.
Additional discussion should be made on the affinity of SSR12. From our experiments we could assess an IC50 of 91uM for 1XA, and compound 23 was 3.5-fold more active (IC50 26uM) and with only 1-step structure change with NAOMInext, we have a compound 650-fold more active (~40nM) than compound 23 and 2275-fold more active than 1XA,

placing the compound in a similar range than the standard BET inhibitor (+)-JQ1 (29nM from our experiment) and the isoxazole azepine final compound of Gehling et al (26nM according to their paper) resulting from the merging approach of 1XA (Figure 3.2).

In a future effort we are trying to join substituents either in ortho and in meta to further increase the affinity.

5.4 First Binders of TREX2

[redacted]

Chapter 6

Conclusions

General Conclusions

This thesis described the study and application of principles of molecular recognition and exploration of chemical space to improve the drug discovery process and help expand the druggable proteome.

Specific Conclusions

1. The exploration of the fragment-size chemical space, starting from a known fragment, with a computational pipeline was able to deliver chemical entities from completely novel chemical series, which have been confirmed experimentally.

2. The series of ligands obtained with a novel scaffold for BRD4 could be used to investigate the unusual behaviour of structural water molecules in BRD4(1) and their role in molecular recognition of ligands.

3. The further growth of the ethyl-derivative of this series with *de novo* virtual synthesis allowed it to explore new points of interactions in BRD4(1) and the role of waters in the ZA channel, delivering a large jump in potency and a 40nM inhibitor.

4. Compounds capable of inhibiting the exonuclease activity of TREX2 have been found. [redacted]

Bibliography

(1) Lander, E. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).

(2) Venter, J. C. et al. The sequence of the human genome. Science 291, 1304–1351 (2001)

(3) The druggable genome Andrew L Hopkins 1, Colin R Groom. Nat Rev Drug Discov. 2002 Sep;1(9):727-30. doi: 10.1038/nrd892.

(4) Oprea, T.I. et al. (2018) Unexplored therapeutic opportunities in the human genome. Nat. Rev. Drug Discov. 17, 317–332 ;

(5) Koscielny, G. et al. (2017) Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 45, D985–D994;

(6) Carter, A.J. et al. (2019) Target 2035: probing the human proteome. Drug Discov. Today 24, 2111–2115

(7) Lipinski, C., Lombardo, F., Dominy, B. & Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 23, 3–25 (1997).

(8) Druggability predictions: Methods, limitations, and applications. Barril, Xavier. WIREs Comput Mol Sci 2013, 3: 327–338 doi: 10.1002/wcms.1134

(9) Drugging the ‘undruggable’ cancer targets.Dang, Chi V. Reddy, E. Premkumar. Shokat, Kevan M. Soucek, Laura. Nature Reviews Cancer, 2017, 7 (8), 502-508.

(10) Lipinski, C., Hopkins, A. Navigating chemical space for biology and medicine. Nature 432, 855–861 (2004). https://doi.org/10.1038/nature03193

(**11)** Dahlin, J., Inglese, J. & Walters, M. Mitigating risk in academic preclinical drug discovery.Nat Rev Drug Discov 14, 279–294 (2015). https://doi.org/10.1038/nrd4578

(12) Gail A. Van Norman, "Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs", JACC: Basic to Translational Science, 2016, 1(3), 170-179. https://doi.org/10.1016/j.jacbts.2016.03.002

(13) Joseph A. DiMasi, Henry G. Grabowski, Ronald W. Hansen (2016). "Innovation in the pharmaceutical industry: New estimates of R&D costs". Journal of Health Economics. 47: 20–33. doi: https://doi.org/10.1016/j.jhealeco.2016.01.012;

(14) Wouters, Olivier J.; McKee, Martin; Luyten, Jeroen (2020-03-03). "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018". JAMA. 323 (9): 844–853. doi: https://doi.org/10.1001/jama.2020.1166

(15) Van Drie, J.H. Computer-aided drug design: the next 20 years. J Comput Aided Mol Des 21, 591–601 (2007). https://doi.org/10.1007/s10822-007-9142-y

(16) Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. Michael M. Hann,* Andrew R. Leach, and Gavin Harper. Journal of Chemical Information and Computer Sciences 2001 41 (3), 856-864. DOI: 10.1021/ci000403i

(17) Rationalizing tight ligand binding through cooperative interaction networks. Bernd Kuhn, Julian E. Fuchs, Michael Reutlinger, Martin Stahl, and Neil R. Taylor. J. Chem. Inf. Model. 2011, 51, 3180–3198. dx.doi.org/10.1021/ci200319e

(18) Ferreira de Freitas, R. & Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med. Chem. Commun. 8, 1970–1981 (2017)

(19) A Medicinal Chemist’s Guide to Molecular Interactions. Caterina Bissantz, Bernd Kuhn, and Martin Stahl*. J. Med. Chem. 2010, 53, 5061–5084 5061 DOI: 10.1021/jm100112j

(20) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. J. Med. Chem. 1985, 28, 849−857.

(21) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. J. Med. Chem. 2014, 57,18−28

(22) Leach AR, Gillet VJ, Lewis RA, Taylor R. Three-dimensional pharmacophore methods in drug discovery. J Med Chem. 2010;53(2):539-558. doi:10.1021/jm900817u

(23) Fersht, A. R. The hydrogen bond in molecular recognition. Trends Biochem. Sci. 1987, 12, 301–304.

(24) Schmidtke, P.; Barril, X. Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. J. Med. Chem. 2010, 53 (15), 5858–5867. https://doi.org/10.1021/jm100574m.

(25) Gao, J.; Bosco, D. A.; Powers, E. T.; Kelly, J. W. Localized thermodynamic coupling between hydrogen bonding and micro-environment polarity substantially stabilizes proteins. Nature Struct. Mol. Biol. 2009, 16, 684–690

(26) Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. Schmidtke, Peter, Javier Luque, F., Murray, James B., Barril, Xavier. J. Am. Chem. Soc. 2011, 133, 18903–18910. dx.doi.org/10.1021/ja207494u

(27) 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, 110 (2019). https://doi.org/10.1038/s42004-019-0205-5.

(28) The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. Francesca Spyrakis, Mostafa H. Ahmed, Alexander S. Bayden, Pietro Cozzini, Andrea Mozzarelli, and Glen E. Kellogg. Journal of Medicinal Chemistry 2017 60 (16), 6781-6827. DOI: 10.1021/acs.jmedchem.7b00057

(29) Rastogi A, Ghosh AK, Suresh SJ. Hydrogen bond interactions between water molecules in bulk liquid, near electrode surfaces and around ions. In: Moreno-Piraj ́an JC, editor. Thermodynamics - physical chemistry of aqueous systems. Chapter 13. London,UK: In-TechOpen; 2011. p. 351–364.

(30) Liu, J., He, X., Zhang, J., & Qi, L. W. (2017). Hydrogen-bond structure dynamics in bulk water: insights from ab initio simulations with coupled cluster theory. Chemical science, 9(8), 2065–2073. https://doi.org/10.1039/c7sc04205a

(31) Karplus P a. Hydrophobicity regained. Protein Sci 1997;6:1302–7. https://doi.org/10.1002/pro.5560060618.

(32) (a) Krimmer, S. G.; Betz, M.; Heine, A.; Klebe, G. Methyl, Ethyl, Propyl, Butyl: Futile but Not for Water, as the Correlation of Structure and Thermodynamic Signature Shows in a Congeneric Series of Thermolysin Inhibitors. ChemMedChem 2014, 9 (4), 833–846. https://doi.org/10.1002/cmdc.201400013. **(b)** Klebe, G. Applying Thermodynamic Profiling in Lead Finding and Optimization. Nat. Rev. Drug Discov. 2015, 14 (2), 95–110. https://doi.org/10.1038/nrd4486. **(c)** Krimmer, S. G.; Cramer, J.; Betz, M.; Fridh, V.; Karlsson, R.; Heine, A.; Klebe, G. Rational Design of Thermodynamic and Kinetic Binding Profiles by Optimizing Surface Water Networks Coating Protein-Bound Ligands. J. Med. Chem. 2016, 59 (23), 10530–10548. https://doi.org/10.1021/acs.jmedchem.6b00998. **(d)** Cramer, J.; Krimmer, S. G.; Heine, A.; Klebe, G. Paying the Price of Desolvation in Solvent-Exposed Protein Pockets: Impact of Distal Solubilizing Groups on Affinity and Binding Thermodynamics in a Series of Thermolysin Inhibitors. J. Med. Chem. 2017, 60 (13), 5791–5799. https://doi.org/10.1021/acs.jmedchem.7b00490. **(e)** Schiebel, J.; Gaspari, R.; Wulsdorf, T.; Ngo, K.; Sohn, C.; Schrader, T. E.; Cavalli, A.; Ostermann, A.; Heine, A.; Klebe, G. Intriguing Role of Water in Protein-Ligand Binding Studied by Neutron

Crystallography on Trypsin Complexes. Nat. Commun. 2018, 9 (1). https://doi.org/10.1038/s41467-018-05769-2.

(f) Hüfner-Wulsdorf, T.; Klebe, G. Role of Water Molecules in Protein-Ligand Dissociation and Selectivity Discrimination: Analysis of the Mechanisms and Kinetics of Biomolecular Solvation Using Molecular Dynamics. J. Chem. Inf. Model. 2020, 60 (3), 1818–1832. https://doi.org/10.1021/acs.jcim.0c00156.

(33) Crawford, T. D.; Tsui, V.; Flynn, E. M.; Wang, S.; Taylor, A. M.; Côté, A.; Audia, J. E.; Beresini, M. H.; Burdick, D. J.; Cummings, R.; Dakin, L. A.; Duplessis, M.; Good, A. C.; Hewitt, M. C.; Huang, H. R.; Jayaram, H.; Kiefer, J. R.; Jiang, Y.; Murray, J.; Nasveschuk, C. G.; Pardo, E.; Poy, F.; Romero, F. A.; Tang, Y.; Wang, J.; Xu, Z.; Zawadzke, L. E.; Zhu, X.; Albrecht, B. K.; Magnuson, S. R.; Bellon, S.; Cochran, A. G. Diving into the Water: Inducible Binding Conformations for BRD4, TAF1(2), BRD9, and CECR2 Bromodomains. J. Med. Chem. 2016, 59 (11), 5391–5402. https://doi.org/10.1021/acs.jmedchem.6b00264.

(34) Lee J, Kim SH. Water polygons in high-resolution protein crystal structures. Protein Sci. 2009;18(7):1370-1376. doi:10.1002/pro.162

(35) Agrawal, K., Shimizu, S., Drahushuk, L. et al. Observation of extreme phase transition temperatures of water confined inside isolated carbon nanotubes. Nature Nanotech 12, 267–273 (2017). https://doi.org/10.1038/nnano.2016.254

(36) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. 1996, 39, 2887−2893.

(37) (a) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.;Waldmann, H. The Scaffold Tree–Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. J. Chem. Inf. Model. 2007, 47,47−58. **(b)** rdScaffoldNetwork: The Scaffold Network Implementation in RDKit. Franziska Kruger, Nikolaus Stiefl, and Gregory A. Landrum. Journal of Chemical Information and Modeling 2020 60 (7), 3331-3335. DOI: 10.1021/acs.jcim.0c00296

(38) Recent Advances in Scaffold Hopping. Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath. Journal of Medicinal Chemistry 2017 60 (4), 1238-1246. 10.1021/acs.jmedchem.6b01437

(39) Rings in drugs. Richard D. Taylor, Malcolm MacCoss and Alastair D. G. Lawson. J. Med. Chem. 2014, 57, 5845−5859. dx.doi.org/10.1021/jm4017625

(40) Virtual Exploration of the Ring Systems Chemical Universe. Ricardo Visini, Josep Arús-Pous, Mahendra Awale, and Jean-Louis Reymond. Journal of Chemical Information and Modeling 2017 57 (11), 2707-2718. DOI: 10.1021/acs.jcim.7b00457

(41) Teilum K, Olsen JG, Kragelund BB. Functional aspects of protein flexibility. Cell Mol Life Sci. 2009;66(14):2231-2247. doi:10.1007/s00018-009-0014-6 ;

(42) Craveur Pierrick, Joseph Agnel Praveen, Esque Jeremy, Narwani Tarun JaiRaj, Noel Floriane, Shinada Nicolas, Goguet Matthieu, Sylvain Léonard, Poulain Pierre, Bertrand Olivier, Faure Guilhem, Rebehmed Joseph, Ghozlane Amine, Swapna Lakshmipuram, Bhaskara Ramachandra, Barnoud Jonathan, Téletchéa Stéphane, Jallu Vincent, Cerny Jiri, Schneider Bohdan, Etchebest Catherine, Srinivasan Narayanaswamy, Gelly Jean-Christophe, de Brevern Alexandre.Protein flexibility in the light of structural alphabets. Front. Mol. Biosci., 27 May 2015 | https://doi.org/10.3389/fmolb.2015.00020

(43) Structural biology and drug discovery of difficult targets: The limits of ligandability
Surade, Sachin, Blundell, Tom L. Chemistry & Biology 19, January 27, 2012

(44) Csermely, P.; Palotai, R.; Nussinov, R. Induced fit, conforma- tional selection and independent dynamic segments: an extended view of binding events. Trends Biochem. Sci. 2010, 35, 539−46

(45) Ensemble Docking in Drug Discovery. Amaro, Rommie E. Baudry, Jerome, Chodera, John, Demir, Özlem, McCammon, J. Andrew, Miao, Yinglong, Smith, Jeremy C. 2018.Biophysical Journal 114(10),2271-2278

(46) Vajda, S.; Beglov, D.; Wakefield, A. E.; Egbert, M.; Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. Curr. Opin. Chem. Biol. 2018, 44,1−8.

(47) Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. Antonija Kuzmanic, Gregory R. Bowman, Jordi Juarez-Jimenez, Julien Michel, and Francesco L. Gervasio. Accounts of Chemical Research 2020 53 (3), 654-661. DOI: 10.1021/acs.accounts.9b00613

(48) Tsai, C.J.; del Sol, A.; Nussinov, R. Protein allostery, signal transmission and dynamics: A classification scheme of allosteric mechanisms. Mol. BioSyst. 2009, 5, 207;

(49) Nussinov & Tsai 2013, Allostery in Disease and in Drug Discovery. Cell, 153(2): 293-305

(50) Vigil, D.; Cherfils, J.; Rossman, K. L.; Der, C. J. Ras superfamily GEFs and GAPs: validated and tractable targets for cancer therapy? Nat. Rev. Cancer 2010, 10, 842−857.

(51) Lu S, Ji M, Ni D, Zhang J. Discovery of hidden allosteric sites as novel targets for allosteric drug design. Drug Discov Today. 2018 Feb;23(2):359-365. doi: 10.1016/j.drudis.2017.10.001. Epub 2017 Oct 10. PMID: 29030241.

(52) Critical review of the role of HTS in drug discovery. Ricardo Macarron. Drug Discovery Today. 2006 Apr;11(7-8):277-279. doi:10.1016/j.drudis.2006.02.001

(53) Macarron R, Banks MN, Bojanic D, et al. Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov. 2011;10(3):188-195. doi:10.1038/nrd3368

(54) A question of library design.Hajduk, Philip J., Galloway, Warren R. J. D., Spring, David R. Nature, 2011, 470, 7332, 42-43

(55) Schneider, G. Automating drug discovery. Nat Rev Drug Discov 17, 97–113 (2018). https://doi.org/10.1038/nrd.2017.232

(56) A Robotic Platform for Quantitative High-Throughput Screening. Sam Michael, Douglas Auld, Carleen Klumpp, Ajit Jadhav, Wei Zheng, Natasha Thorne, Christopher P. Austin, James Inglese, and Anton Simeonov. ASSAY and Drug Development Technologies 2008 6:5, 637-657

(57) Dmitriy M. Volochnyuk, Sergey V. Ryabukhin, Yurii S. Moroz, Olena Savych, Alexander Chuprina, Dragos Horvath, Yuliana Zabolotna, Alexandre Varnek, Duncan B. Judd, Evolution of commercially available compounds for HTS, Drug Discovery Today, Volume 24, Issue 2, 2019, Pages 390-402, https://doi.org/10.1016/j.drudis.2018.10.016

(58) Franzini, R. M.; Randolph, C. Chemical Space of DNA-Encoded Libraries: Miniperspective. J. Med. Chem. 2016, 59 (14), 6629–6644. https://doi.org/10.1021/acs.jmedchem.5b01874.

(59) Sterling and Irwin, J. Chem. Inf. Model, 2015 http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559.

(60) The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. Medicinal Research Reviews, Vol. 16, No. 1, 3-50 (1996)

(61) Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des. 2013;27(8):675-679. doi:10.1007/s10822-013-9672-4

(62) **(a)** Virtual Exploration of the Small Molecule Chemical Universe below 160 Daltons. T. Fink, H. Bruggesser, J.-L. Reymond, Angew. Chem. Int. Ed. 2005, 44, 1504-1508.

(b) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physico-chemical properties, compound classes and drug discovery. T. Fink, J.-L. Reymond, J. Chem. Inf. Model. 2007, 47, 342-353.

(c) 970 Million Drug-like Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. L. C. Blum, J.-L. Reymond, J. Am. Chem. Soc. 2009, 131, 8732-8733.

(d) 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. 2012, 52, 2864−2875

(63) The next level in chemical space navigation: going far beyond enumerable compound libraries. Torsten Hoffmann and Marcus Gastreich. Drug Discovery Today, 2019, 24(5), Pages 1148-1156,https://doi.org/10.1016/j.drudis.2019.02.013.

(64) Virtual Chemical Libraries. W. Patrick Walters. Journal of Medicinal Chemistry 2019 62 (3), 1116-1124. DOI: 10.1021/acs.jmedchem.8b01048

(65) (a) https://enamine.net/library-synthesis/real-compounds/real-database

(b) https://enamine.net/library-synthesis/real-compounds/real-space-navigator

(66) Xavier Barril (2017) Computer-aided drug design: time to play with novel chemical matter, Expert Opinion on Drug Discovery, 12:10, 977-980, DOI: 10.1080/17460441.2017.1362386

(67) Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. Nature 566, 224–229 (2019).

(68) Gorgulla, C., Boeszoermenyi, A., Wang, Z. et al. An open-source drug discovery platform enables ultra-large virtual screens. Nature 580, 663–668 (2020). https://doi.org/10.1038/s41586-020-2117-z

(69) Jencks, W. P. On the attribution and additivity of binding energies. Proc. Natl. Acad. Sci. U.S.A. 1981, 78, 4046-4050.

(70) Fragment-Based Drug Discovery. Daniel A. Erlanson,* Robert S. McDowell,* and Tom O'Brien*. Journal of Medicinal Chemistry, 2004, Vol. 47, No. 14. 10.1021/jm040031v

(71) A 'Rule of Three' for fragment-based lead discovery? Miles Congreve, Robin Carr, Chris Murray and Harren Jhoti. Drug Discovery Today, 2003, 8 (19), 876-877

(72) (a) Keseru, G. M.; Erlanson, D. A.; Ferenczy, G. G.; Hann, M. M.; Murray, C. W.; Pickett, S. D. Design Principles for Fragment Libraries – Maximizing the Value of Learnings from Pharma Fragment Based Drug Discovery (FBDD) Programs for Use in Academia. J. Med. Chem. 2016, acs.jmedchem.6b00197. https://doi.org/10.1021/acs.jmedchem.6b00197

(b) Hall, R. J.; Mortenson, P. N.; Murray, C. W. Efficient Exploration of Chemical Space by Fragment-Based Screening. Prog. Biophys. Mol. Biol. 2014, 116 (2–3), 82–91. https://doi.org/10.1016/j.pbiomolbio.2014.09.007.

(73) Ludlow RF, Verdonk ML, Saini HK, Tickle IJ, Jhoti H. Detection of secondary binding sites in proteins using fragment screening. Proc Natl Acad Sci U S A 2015;112:15910–5. https://doi.org/10.1073/pnas.1518946112.

(74) Molecular complexity and fragment-based drug discovery: Ten years on. Leach, Andrew R. Hann, Michael M. Current Opinion in Chemical Biology 2011, 15:489–496

(75) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. Science 1996, 274, 1531-1534

(76) The maximal affinity of ligands I. D. KUNTZ*†,K. CHEN*, K. A. SHARP‡§, AND P. A. KOLLMAN*. Proc. Natl. Acad. Sci. USA Vol. 96, pp. 9997–10002, August 1999

(77) Ligand efficiency: A useful metric for lead selection. Hopkins, Andrew L. Groom, Colin R. Alex, Alexander. DDT Vol. 9, No. 10 May 2004

(78) Fragment-Based Approaches in Drug Discovery and Chemical Biology. Duncan E. Scott, Anthony G. Coyne, Sean A. Hudson, and Chris Abell. Biochemistry 2012, 51, 4990−5003. dx.doi.org/10.1021/bi3005126

(79) Fragment-to-Lead Tailored In Silico Design. Moira Rachman, Serena Piticchio, Maciej Majewski, Xavier Barril. Drug Discovery Today: Technologies. *Submitted*

(80) Pearlman, D. A. & Murcko, M. A. CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. J. Med. Chem. 39, 1651–1663 (1996).

(81) LEGEND:Nishibata Y, Itai A. Confirmation of Usefulness of a Structure Construction Program Based on Three-Dimensional. Receptor Structure for Rational Lead Generation. J Med Chem 1993;36:2921–8. https://doi.org/10.1021/jm00072a011

(82) LUDI:Bohm H. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Mol Des Vol 1992;6:61–78. https://doi.org/10.1007/BF00124387

(83) DeWitte RS, Shakhnovich E. SMoG: de novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. J Am Chem Soc 1996;118:11733-44

(84) Hartenfeller M, Zettl H, Walter M, Rupp M, Reisen F, Proschak E, et al. Dogs: Reaction-driven de novo design of bioactive compounds. PLoS Comput Biol 2012;8:1–25. https://doi.org/10.1371/journal.pcbi.1002380

(85) Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 2005;45:160–9. https://doi.org/10.1021/ci049885e

(86) Peter S Kutchukian & Eugene I Shakhnovich (Professor) (2010) De novo design: balancing novelty and confined chemical space, Expert Opinion on Drug Discovery, 5:8, 789-812, DOI: 10.1517/17460441.2010.497534

(87) Schneider G, Lee ML, Stahl M, Schneider P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. J Comput Aided Mol Des 2000;14:487–94. https://doi.org/10.1023/A:1008184403558

(88) Douguet D, Munier-Lehmann H, Labesse G, Pochet S. LEA3D: A computer-aided ligand design for structure- based drug design. J Med Chem 2005;48:2457–68. https://doi.org/10.1021/jm0492296

(89) Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci 1998;38:511–22. https://doi.org/10.1021/ci970429i.

(90) Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment Optimized Growth algorithm for the de novo generation of molecules occupying druglike chemical space. J. Chem. Inf. Model. 2009, 49 (7), 1630−42

(91) OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. Nicolas Chéron, Naveen Jasty, and Eugene I. Shakhnovich. Journal of Medicinal Chemistry 2016 59 (9), 4171-4188. DOI: 10.1021/acs.jmedchem.5b00886

(92) (a) Fechner U, Schneider G. Flux (1): a virtual synthesis scheme for fragment-based de novo design. J Chem Inf Model 2006;46:699-707 (b) Fechner U, Schneider G. Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. J Chem Inf Model 2007;47:656-67

(93) Vinkers HM, de Jonge MR, Daeyaert FF, et al. SYNOPSIS: SYNthesize and OPtimize System in Silico. J Med Chem 2003;46:2765-73

(94) Computer-based de novo design of drug-like molecules. Gisbert Schneider and Uli Fechner. Nature Reviews Drug Discovery, 2005, 4(8), 649-663

(95) Pearlman, D. A. & Murcko, M. A. CONCEPTS: new dynamic algorithm for de novo design suggestion. J. Comput. Chem. 14, 1184–1193 (1993).

(96) Wang R, Gao Y, Lai L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. J Mol Model 2000;6:498–516. https://doi.org/10.1007/s0089400060498.

(97) Dey F, Caflisch A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. J Chem Inf Model 2008;48:679-90

(98) Schneider G, Hartenfeller M, Reutlinger M, et al. Voyages to the (un)known: adaptive design of bioactive compounds. Trends Biotechnol 2009;27:18-26

(99) **(a)** Degen J, Rarey M. FlexNovo: structure-based searching in large fragment spaces. ChemMedChem 2006;1:854-68 **(b)** Fischer JR, Lessel U, Rarey M. LoFT: similarity-driven multiobjective focused library design. J Chem Inf Model 2010;50:1-21 **(c)** Nicolaou CA, Apostolakis J, Pattichis CS. De novo drug design using multiobjective evolutionary graphs. J Chem Inf Model 2009;49:295-307

(100) Boda K, Seidel T, Gasteiger J. Structure and reaction based evaluation of synthetic accessibility. J Comput Aided Mol Des 2007;21:311-25

(101) **(a)** D. Merk, L. Friedrich, F. Grisoni, G. Schneider. De Novo Design of Bioactive Small Molecules by Artificial Intelligence Mol. Inf. 2018, 37, 1700153. **(b)** Schneider, P., Walters, W.P., Plowright, A.T. et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 19, 353–364 (2020). https://doi.org/10.1038/s41573-019-0050-3

(102) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. A collection of robust organic synthesis reactions for in silico molecule design. J. Chem. Inf. Model. 2011, 51, 3093−3098

(103) Chemical Space Expansion of Bromodomain Ligands Guided by in Silico Virtual Couplings (AutoCouple). Laurent Batiste, Andrea Unzue, Aymeric Dolbois, Fabrice Hassler, Xuan Wang, Nicholas Deerain, Jian Zhu, Dimitrios Spiliotopoulos, Cristina Nevado, and Amedeo Caflisch. ACS Central Science 2018 4 (2), 180-188. DOI: 10.1021/acscentsci.7b00401

(104) Integrated Strategy for Lead Optimization Based on Fragment Growing: The Diversity-Oriented-Target-Focused-Synthesis Approach. Laurent Hoffer, Yuliia V. Voitovich, Brigitt Raux, Kendall Carrasco, Christophe Muller, Aleksey Y. Fedorov, Carine Derviaux, Agnès Amouric, Stéphane Betzi, Dragos Horvath, Alexandre Varnek, Yves Collette, Sébastien Combes, Philippe Roche, and Xavier Morelli. Journal of Medicinal Chemistry 2018 61 (13), 5719-5732. DOI: 10.1021/acs.jmedchem.8b00653

(105) Binding-Site Compatible Fragment Growing Applied to the Design of β2-Adrenergic Receptor Ligands. Florent Chevillard, Helena Rimmer, Cecilia Betti, Els Pardon, Steven Ballet, Niek van Hilten, Jan Steyaert, Wibke E. Diederich, and Peter Kolb. Journal of Medicinal Chemistry 2018 61 (3), 1118-1129. DOI: 10.1021/acs.jmedchem.7b01558

(106) NAOMInext e Synthetically feasible fragment growing in a structure- based design context. Kai Sommer, Florian Flachsenberg, Matthias Rarey.European Journal of Medicinal Chemistry 163 (2019) 747-762

(107) Tamkun JW, Deuring R, Scott MP, et al. brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. Cell. 1992;68(3):561-572. doi:10.1016/0092-8674(92)90191-e

(108) The bromodomain: A conserved sequence found in human, Drosophila and yeast proteins. Susan R.Haynes, Catherine Dollard', Fred Winston', Stephan Beck2, John Trowsdale2 and Igor B.Dawid. Nucleic Acids Research,1992, Vol. 20, No. 10, 2603.

(109) Jeanmougin F, Wurtz JM, Le Douarin B, Chambon P, Losson R. The bromodomain revisited. Trends Biochem Sci. 1997; 22(5):151-153. doi:10.1016/s0968-0004(97)01042-6

(110) Dhalluin, C., Carlson, J., Zeng, L. et al. Structure and ligand of a histone acetyltransferase bromodomain. Nature 399, 491–496 (1999). https://doi.org/10.1038/20974

(111) Winston F, Allis CD. The bromodomain: a chromatin-targeting module?. Nat Struct Biol. 1999;6(7):601-604. doi:10.1038/10640

(112) Filippakopoulos P, Picaud S, Mangos M, et al. Histone recognition and large-scale structural analysis of the human bromodomain family. Cell. 2012;149(1):214-231. doi:10.1016/j.cell.2012.02.013

(113) Filippakopoulos, P., Knapp, S. Targeting bromodomains: epigenetic readers of lysine acetylation. Nat Rev Drug Discov 13, 337–356 (2014). https://doi.org/10.1038/nrd4286

(114) Zhang X, Chen K, Wu YD, Wiest O (2017) Protein dynamics and structural waters in bromodomains. PLOS ONE 12(10): e0186570. https://doi.org/10.1371/journal.pone.0186570

(115) Filippakopoulos P, Qi J, Picaud S, et al. Selective inhibition of BET bromodomains. Nature. 2010;468(7327):1067-1073. doi:10.1038/nature09504

(116) Nicodeme, E., Jeffrey, K., Schaefer, U. et al. Suppression of inflammation by a synthetic histone mimic. Nature 468, 1119–1123 (2010). https://doi.org/10.1038/nature09589

(117) Galdeano C, Ciulli A. Selectivity on-target of bromodomain chemical probes by structure-guided medicinal chemistry and chemical biology. Future Med Chem. 2016;8(13):1655-1680. doi:10.4155/fmc-2016-0059

(118) Cochran, A.G., Conery, A.R. & Sims, R.J. Bromodomains: a new target class for drug development. Nat Rev Drug Discov 18, 609–628 (2019). https://doi.org/10.1038/s41573-019-0030-7

(119) Fujisawa, T., Filippakopoulos, P. Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. Nat Rev Mol Cell Biol 18, 246–262 (2017). https://doi.org/10.1038/nrm.2016.143

(120) Ember, S. W. J.; Zhu, J. Y.; Olesen, S. H.; Martin, M. P.; Becker, A.; Berndt, N.; Georg, G. I.; Schonbrunn, E. Acetyl-Lysine Binding Site of Bromodomain-Containing Protein 4 (BRD4) Interacts with Diverse Kinase Inhibitors. *ACS Chem. Biol.* **2014**, *9* (5), 1160–1171. https://doi.org/10.1021/cb500072z.

(121) **(a)**Filippakopoulos, P.; Picaud, S.; Fedorov, O.; Keller, M.; Wrobel, M.; Morgenstern, O.; Bracher, F.; Knapp, S. Benzodiazepines and Benzotriazepines as Protein Interaction Inhibitors Targeting Bromodomains of the BET Family. Bioorganic Med. Chem. 2011, 20 (6), 1878–1886. https://doi.org/10.1016/j.bmc.2011.10.080. **(b)** Zhang, G.; Plotnikov, A. N.; Rusinova, E.; Shen, T.; Morohashi, K.; Joshua, J.; Zeng, L.; Mujtaba, S.; Ohlmeyer, M.; Zhou, M. M. Structure-Guided Design of Potent Diazobenzene Inhibitors for the BET Bromodomains. J. Med. Chem. 2013, 56 (22), 9251–9264. https://doi.org/10.1021/jm401334s. **(c)** Fedorov, O.; Lingard, H.; Wells, C.; Monteiro, O. P.; Picaud, S.; Keates, T.; Yapp, C.; Philpott, M.; Martin, S. J.; Felletar, I.; Marsden, B. D.; Filippakopoulos, P.; Müller, S.; Knapp, S.; Brennan, P. E. [1,2,4]Triazolo[4,3-a]Phthalazines: Inhibitors of Diverse Bromodomains. J. Med. Chem. 2014, 57 (2), 462–476. https://doi.org/10.1021/jm401568s. **(d)** Hügle, M.; Lucas, X.; Weitzel, G.; Ostrovskyi, D.; Breit, B.; Gerhardt, S.; Einsle, O.; Günther, S.; Wohlwend, D. 4-Acyl Pyrrole Derivatives Yield Novel Vectors for Designing Inhibitors of the Acetyl-Lysine Recognition Site of BRD4(1). J. Med. Chem. 2016, 59 (4), 1518–1530. https://doi.org/10.1021/acs.jmedchem.5b01267

(122) Dan J. Mazur and Fred W. Perrino.Identification and Expression of the TREX1 and TREX2 cDNA Sequences Encoding Mammalian 3'-5' Exonucleases. THE JOURNAL OF BIOLOGICAL CHEMISTRY. Vol. 274, No. 28, Issue of July 9, pp. 19655–19660, 1999

(123) Biochemical and cellular characteristics of the 3'-5' exonuclease TREX2 Ming-Jiu Chen*, Sheng-Mei Ma, Lavinia C. Dumitrache and Paul Hasty.Nucleic Acids Research, 2007, Vol. 35, No. 8 doi:10.1093/nar/gkm151.

(124) Shevelev, Igor V. Hübscher, Ulrich. The 3′-5′ exonucleases. Nature Reviews Molecular Cell Biology. 2002. 3(5), 364-375

(125) Shevelev, I. V., Ramadan, K. & Hubscher, U. The TREX2 3′–5′ exonuclease physically interacts with DNA polymerase δ and increases its accuracy. The Scientific World Journal 2, 275–281 (2002)

(126) Structural insights into the duplex DNA processing of TREX2. Hiu-Lo Cheng1, Chun-Ting Lin2, Kuan-Wei Huang2, Shuying Wang3,4, Yeh-Tung Lin2, Shu-Ing Toh2,5 and Yu-Yuan Hsiao. Nucleic Acids Research, 2018, Vol. 46, No. 22. 12166–12176

(127) Cooperative DNA Binding and Communication across the Dimer Interface in the TREX2 3? 35?-Exonuclease. Fred W. Perrino1, Udesh de Silva, Scott Harvey, Edward E. Pryor, Jr., Daniel W. Cole, and Thomas Hollis. THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 283, NO. 31, pp. 21441–21452, August 1, 2008

(128) DNA binding induces active site conformational change in the human TREX2 3'-exonuclease Udesh de Silva, Fred W. Perrino and Thomas Hollis. Nucleic Acids Research, 2009, Vol. 37, No. 7 2411–2417 doi:10.1093/nar/gkp025

(129) Perrino FW, de Silva U, Harvey S, Pryor EE Jr, Cole DW, Hollis T. Cooperative DNA binding and communication across the dimer interface in the TREX2 3' --> 5'-exonuclease. J Biol Chem. 2008;283(31):21441-21452. doi:10.1074/jbc.M803629200

(130) Increased Susceptibility to Skin Carcinogenesis in TREX2 Knockout Mice
David Parra,1 Joan Manils,1Ba`rbara Castellana,1 Arnau Vin˜a-Vilaseca,1 Eva Mora´n-Salvador,1 Nuria Va´zquez-Villoldo,1 Gemma Taranco´n,1 Miquel Borra`s,2 Sara Sancho,4 Carmen Benito,3 Sagrario Ortega,5 and **Concepcio´** Soler. Cancer Res 2009;69:6676-6684. Published OnlineFirst August 4, 2009

(131) Multifaceted role of TREX2 in the skin defense against UV-induced skin carcinogenesis Joan Manils1 , Diana Gómez1 , Mercè Salla-Martret1 , Heinz Fischer2 , Jason M. Fye3 ,Elena Marzo1, Laura Marruecos1, Inma Serrano1, Rocío Salgado4, Juan P. Rodrigo5,Juana M. Garcia-Pedrero5, Anna M. Serafin6, Xavier Cañas6, Carmen Benito7 , Agustí Toll4, Sònia-Vanina Forcales8, Fred W. Perrino3, Leopold Eckhart2, **Concepció** Soler. Oncotarget. 2015, 6(26). 22375-22396.

(132) Mathers: TREX through Cutaneous Health and Disease Alicia R. Mathers. Journal ofInvestigative Dermatology (2016) 136, 2337e2339. doi:10.1016/j.jid.2016.06.628

(133) The Exonuclease Trex2 Shapes Psoriatic Phenotype. Joan Manils1, Eduard Casas2,3, Arnau Vin˜a-Vilaseca1, Marc Lo´pez-Cano1,4, Anna Dı´ez-Villanueva2, Diana Go´mez1, Laura Marruecos1, Marta Ferran5, Carmen Benito6, Fred W. Perrino7, Tanya Vavouri2,3, Josep Maria de Anta1, Francisco Ciruela1,4 and **Concepcio´** Soler1. Journal of Investigative Dermatology (2016) 136, 2345e2355; doi:10.1016/j.jid.2016.05.122

(134) global report: Global Report on Psoriasis, World Health Organization. 2016.

(135) Gasteiger, J., & Engel, T. (2003). Chemoinformatics: A textbook. Weinheim: Wiley-VCH. Chicago

(136) Willett P. Similarity searching using 2D structural fingerprints. Methods Mol Biol. 2011;672:133-58. doi: 10.1007/978-1-60761-839-3_5. PMID: 20838967.

(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. 2002, 42, 1273−1280

(138) Hessler, G.; Baringhaus, K.-H. The Scaffold Hopping Potential of Pharmacophores. Drug Discovery Today: Technol. 2010, 7, e263−e269.

(139) Kumar, Ashutosh, and Kam Y J Zhang. “Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery.” Frontiers in chemistry vol. 6 315. 25 Jul. 2018, doi:10.3389/fchem.2018.00315

(140) Schneider, P.; Tanrikulu, J.; Schneider, G. Self-Organizing Maps in Drug Discovery: Compound Library Design, Scaffold-Hopping, Repurposing. Curr. Med. Chem. 2009, 16, 258−266.

(141) Charting a Path to Success in Virtual Screening. Stefano Forli. Molecules 2015, 20, 18732-18758; doi:10.3390/molecules201018732

(142) (a) Chen, Y. C. Beware of Docking! Trends Pharmacol. Sci. 2015, 36 (2), 78–95. https://doi.org/10.1016/j.tips.2014.12.001 **(b)** Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev. 2017 Apr;9(2):91-102. doi: 10.1007/s12551-016-0247-1. Epub 2017 Jan 16. PMID: 28510083; PMCID: PMC5425816.

(143) Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, et al. (2014) rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. PLoS Comput Biol 10(4): e1003571. doi:10.1371/journal.pcbi.1003571

(144) D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, San Francisco

(145) Role of Molecular Dynamics and Related Methods in Drug Discovery. Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Journal of Medicinal Chemistry 2016 59 (9), 4035-4061. DOI: 10.1021/acs.jmedchem.5b01684

(146) Buch, I.; Giorgino, T. Complete reconstruction of an enzyme- inhibitor binding process by molecular dynamics simulations. Proc. Natl. Acad. Sci. U. S. A. 2011, 108 (25), 10184−10189.

(147) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How does a drug molecule find its target binding site? J. Am. Chem. Soc. 2011, 133, 9181−9183.

(148) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. Expert Opin. Drug Discovery 2015, 10, 449−461

(149) Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. Phani Ghanakota and Heather A. Carlson. Journal of Medicinal Chemistry 2016 59 (23), 10383-10399. DOI: 10.1021/acs.jmedchem.6b00399

(150) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. Proteins: Struct., Funct., Genet. 1991, 11,29−34

(151) (a) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. J. Phys. Chem. 1996, 100, 2605−2611**; (b)** Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. J. Mol. Biol. 2006, 357, 1471−1482; **(c)** English, A. C.; Done, S. H.; Caves, L. S. D.; Groom, C. R.; Hubbard, R. E. Locating Interaction Sites on Proteins: The Crystal Structure of Thermolysin Soaked in 2% to 100% Isopropanol. Proteins: Struct., Funct., Genet. 1999, 37, 628−640.

(152) Binding Site Detection and Druggability Index from First Principles. Jesus Seco, F. Javier Luque, and Xavier Barril. J. Med. Chem. 2009, 52, 2363–2371. DOI: 10.1021/jm801385d

(153) Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design Daniel Alvarez-Garcia†,‡ and Xavier Barril. J. Chem. Theory Comput. 2014, 10, 2608−2614. DOI:dx.doi.org/10.1021/ct500182z ;

(154) Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. Daniel Alvarez-Garcia and Xavier Barril. J. Med. Chem. 2014, 57, 8530−8539. dx.doi.org/10.1021/jm5010418

(155) (a) Molecular Dynamics in Mixed Solvents Reveals Protein–Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. Juan Pablo Arcon, Lucas A. Defelipe, Carlos P. Modenutti, Elias D. López, Daniel Alvarez-Garcia, Xavier Barril, Adrián G. Turjanski, and Marcelo A. Martí. Journal of Chemical Information and Modeling 2017 57 (4), 846-863. DOI: 10.1021/acs.jcim.6b00678; **(b)** Cosolvent-Based Protein Pharmacophore for Ligand Enrichment in Virtual Screening. Juan Pablo Arcon, Lucas A. Defelipe, Elias D. Lopez, Osvaldo Burastero, Carlos P. Modenutti, Xavier Barril, Marcelo A. Marti, and Adrian G. Turjanski. Journal of Chemical Information and Modeling 2019 59 (8), 3572-3583. DOI: 10.1021/acs.jcim.9b00371

(156) Guvench, O.; MacKerell, A. D., Jr. Computational Fragment- Based Binding Site Identification by Ligand Competitive Saturation. PLoS Comput. Biol. 2009, 5, e1000435.,

(157) Lexa, K. W.; Carlson, H. A. Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping. J. Am. Chem. Soc. 2011, 133, 200−202.

(158) Bakan, A.; Nevins, N.; Lakdawala, A.S.; Bahar, I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. J. Chem. Theory Comput. 2012, 8, 2435–2447.

(159) Ghanakota, P.; Carlson, H.A. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. J. Phys. Chem. B 2016, 120, 8685–8695.

(160) Ghanakota, P.; van Vlijmen, H.; Sherman, W.; Beuming, T. Large-Scale Validation of Mixed-Solvent Simulations to Assess Hotspots at Protein-Protein Interaction Interfaces. J. Chem. Inf. Model. 2018, 58, 784–793.

(161) Yu, W.; Lakkaraju, S.K.; Raman, E.P.; Fang, L.; MacKerell, A.D., Jr. Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules. J. Chem. Inf. Model. 2015, 55, 407–420

(162) Graham, S.E.; Smith, R.D.; Carlson, H.A. Predicting Displaceable Water Sites Using Mixed-Solvent Molecular Dynamics. J. Chem. Inf. Model. 2018, 58, 305–314

(153) (a) 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. 2016, 138, 14257–14263. **(b)** 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. 2017, 57, 1388–1401. **(c)** Comitani, F.; Gervasio, F.L. Exploring Cryptic Pockets Formation in Targets of Pharmaceutical Interest with SWISH. J. Chem. Theory Comput. 2018, 14, 3321–3331. **(d)** Cosolvent-Enhanced Sampling and Unbiased Identification of Cryptic Pockets Suitable for Structure-Based Drug Design. Denis Schmidt, Markus Boehm, Christopher L. McClendon, Rubben Torella, and Holger Gohlke. Journal of Chemical Theory and Computation 2019 15 (5), 3331-3343. DOI: 10.1021/acs.jctc.8b01295

(e) Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. Antonija Kuzmanic, Gregory R. Bowman, Jordi Juarez-Jimenez, Julien Michel, and Francesco L. Gervasio. Accounts of Chemical Research 2020 53 (3), 654-661. DOI: 10.1021/acs.accounts.9b00613

(164) Schmidtke P, Bidon-Chanal A, Luque FJ, Barril X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics. 2011; 27(23):3276-3285. doi:10.1093/bioinformatics/btr550

(165) Ruiz-Carmona, S., Schmidtke, P., Luque, F. et al. Dynamic undocking and the quasi-bound state as tools for drug discovery. Nature Chem 9, 201–206 (2017). https://doi.org/10.1038/nchem.2660

(166) (a) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. J. Chem. Phys. 1935, 3, 300−313. **(b)** Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. J. Chem. Phys. 1954, 22, 1420−1426

(167) Wong, C. F.; McCammon, J. A. Dynamics and Design of Enzymes and Inhibitors. J. Am. Chem. Soc. 1986, 108, 3830−3832.

(168) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. J. Chem. Phys. 1985, 83, 3050−3054.

(169) Mermelstein DJ, Lin C, Nelson G, Kretsch R, McCammon JA, Walker RC. Fast and flexible gpu accelerated binding free energy calculations within the amber molecular dynamics package. J Comput Chem. 2018;39(19):1354-1358. doi:10.1002/jcc.25187

(170) He, X.; Liu, S.; Lee, T. S.; Ji, B.; Man, V. H.; York, D. M.; Wang, J. Fast, Accurate, and Reliable Protocols for Routine Calculations of Protein-Ligand Binding Affinities in Drug Design Projects Using AMBER GPU-TI with Ff14SB/GAFF. *ACS Omega* **2020**, *5* (9), 4611–4619. https://doi.org/10.1021/acsomega.9b04233.

(171) Fast, Accurate, and Reliable Protocols for Routine Calculations of Protein–Ligand Binding Affinities in Drug Design Projects Using AMBER GPU-TI with ff14SB/GAFF. Xibing He, Shuhan Liu, Tai-Sung Lee, Beihong Ji, Viet H. Man, Darrin M. York, and Junmei Wang. ACS Omega 2020 5 (9), 4611-4619. DOI: 10.1021/acsomega.9b04233

(172) Free Energy Methods in Drug Design: Prospects of “Alchemical Perturbation” in Medicinal Chemistry. Billy J. Williams-Noonan, Elizabeth Yuriev, and David K. Chalmers
Journal of Medicinal Chemistry 2018 61 (3), 638-649. DOI: 10.1021/acs.jmedchem.7b00681

(173) Klimovich PV, Shirts MR, Mobley DL. Guidelines for the analysis of free energy calculations. J Comput Aided Mol Des. 2015;29(5):397-411. doi:10.1007/s10822-015-9840-9

(174) Schindler, C. E. M., Baumann, H., Blum, A., Böse, D., Buchstaller, H. P., Burgdorf, L., Cappel, D., Chekler, E., Czodrowski, P., Dorsch, D., Eguida, M. K. I., Follows, B., Fuchß, T., Grädler, U., Gunera, J., Johnson, T., Jorand Lebrun, C., Karra, S., Klein, M., … Kuhn, D. (2020). Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. Journal of Chemical Information and Modeling, 60(11), 5457–5474. https://doi.org/10.1021/acs.jcim.0c00900

(175) Rigorous Free Energy Simulations in Virtual Screening. Zoe Cournia, Bryce K. Allen, Thijs Beuming, David A. Pearlman, Brian K. Radak, and Woody Sherman. Journal of Chemical Information and Modeling Article ASAP. DOI: 10.1021/acs.jcim.0c00116

(176) Discovery, Design, and Optimization of Isoxazole Azepine BET Inhibitors
Victor S. Gehling, Michael C. Hewitt, Rishi G. Vaswani, Yves Leblanc, Alexandre Côté, Christopher G. Nasveschuk, Alexander M. Taylor, Jean-Christophe Harmange, James E. Audia, Eneida Pardo, Shivangi Joshi, Peter Sandy, Jennifer A. Mertz, Robert J. Sims, Louise Bergeron, Barbara M. Bryant, Steve Bellon, Florence Poy, Hariharan Jayaram,

Ravichandran Sankaranarayanan, Sreegouri Yellapantula, Nandana Bangalore Srinivasamurthy, Swarnakumari Birudukota, and Brian K. Albrecht
ACS Medicinal Chemistry Letters 2013 4 (9), 835-840
DOI: 10.1021/ml4001485

(177) Molecular Operating Environment (MOE), **2014.01**; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

(178) O'Boyle, N.M., Banck, M., James, C.A. et al. Open Babel: An open chemical toolbox. J Cheminform 3, 33 (2011). https://doi.org/10.1186/1758-2946-3-33

(179) LigPrep, Schrödinger, LLC, New York, NY

(180) https://openbabel.org/docs/dev/FileFormats/Fastsearch_format.html

(181) RDKit: Open-source cheminformatics; http://www.rdkit.org

(182) rDock User Guide. http://rdock.sourceforge.net/documentation/

(183) Molecular Operating Environment (MOE), **2019.01**; Chemical Computing Group ULC

(184) M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, ***Gaussian 09*** (Gaussian, Inc., Wallingford CT, 2009).

(185) D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2018), **AMBER 2018**, University of California, San Francisco

(186) https://ambermd.org/bugfixes/18.0/update.16

(187) Ciulli, Alessio. "Biophysical screening for the discovery of small-molecule ligands." Methods in molecular biology (Clifton, N.J.) vol. 1008 (2013): 357-88. doi:10.1007/978-1-62703-398-5_13

(188) Degorce, François et al. "HTRF: A technology tailored for drug discovery - a review of theoretical aspects and recent applications." Current chemical genomics vol. 3 22-32. 28 May. 2009, doi:10.2174/1875397300903010022

(189) **(a)** Fedorov, O.; Lingard, H.; Wells, C.; Monteiro, O. P.; Picaud, S.; Keates, T.; Yapp, C.; Philpott, M.; Martin, S. J.; Felletar, I.; Marsden, B. D.; Filippakopoulos, P.; Müller, S.; Knapp, S.; Brennan, P. E. [1,2,4]Triazolo[4,3-a]Phthalazines: Inhibitors of

Diverse Bromodomains. J. Med. Chem. 2014, 57 (2), 462–476. https://doi.org/10.1021/jm401568s. **(b)** Huang, L.; Li, H.; Li, L.; Niu, L.; Seupel, R.; Wu, C.; Cheng, W.; Chen, C.; Ding, B.; Brennan, P. E.; Yang, S. Discovery of Pyrrolo[3,2- d]Pyrimidin-4-One Derivatives as a New Class of Potent and Cell-Active Inhibitors of P300/CBP-Associated Factor Bromodomain. J. Med. Chem. 2019, 62 (9), 4526–4542. https://doi.org/10.1021/acs.jmedchem.9b00096.

(190) https://www.hit2lead.com/

(191) Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, Bernd Wiswedel. (2008) KNIME: The Konstanz Information Miner. In: Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_38

(192) SeeSAR version 10.0; BioSolveIT GmbH, Sankt Augustin, Germany, 2020, https://www.biosolveit.de/SeeSAR/

(193) Schneider N, Lange G, Hindle S, Klein R, Rarey M. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. J Comput Aided Mol Des. 2013 Jan;27(1):15-29. doi:10.1007/s10822-012-9626-2.

(194) https://scifinder.cas.org

(192) ZINC15 in-stock consulted in february 2020

(193) Enamine REAL consulted in february 2020

(194) Fragment Database FDB-17. Ricardo Visini, Mahendra Awale, and Jean-Louis Reymond. Journal of Chemical Information and Modeling 2017 57 (4), 700-709. DOI: 10.1021/acs.jcim.7b00020

(195) Carlo, M.; Aldeghi, M.; Ross, G. A.; Bodkin, M. J.; Essex, J. W.; Knapp, S.; Biggin, P. C. Large-Scale Analysis of Water Stability in Bromodomain Binding Pockets with Grand Canonical Monte Carlo. Commun. Chem. 2018. https://doi.org/10.1038/s42004-018-0019-x

(196) **(a)** Vollmuth, F.; Geyer, M. Interaction of Propionylated and Butyrylated Histone H3 Lysine Marks with Brd4 Bromodomains. Angew. Chemie - Int. Ed. 2010, 49 (38), 6768–6772. https://doi.org/10.1002/anie.201002724. and **(b)** Flynn, E. M.; Huang, O. W.; Poy, F.; Oppikofer, M.; Bellon, S. F.; Tang, Y.; Cochran, A. G. A Subset of Human Bromodomains Recognizes Butyryllysine and Crotonyllysine Histone Peptide Modifications. Structure 2015, 23 (10), 1801–1814. https://doi.org/10.1016/j.str.2015.08.004.

(197) Bharatham, N.; Slavish, P. J.; Young, B. M.; Shelat, A. A. The Role of ZA Channel Water-Mediated Interactions in the Design of Bromodomain-Selective BET Inhibitors. J. Mol. Graph. Model. 2018, 81, 197–210. https://doi.org/10.1016/j.jmgm.2018.03.005

(198) Huang, K. W.; Hsu, K. C.; Chu, L. Y.; Yang, J. M.; Yuan, H. S.; Hsiao, Y. Y. Identification of Inhibitors for the DEDDh Family of Exonucleases and a Unique Inhibition Mechanism by Crystal Structure Analysis of CRN-4 Bound with

2-Morpholin-4-Ylethanesulfonate (MES). J. Med. Chem. 2016, 59 (17), 8019–8029. https://doi.org/10.1021/acs.jmedchem.6b00794.