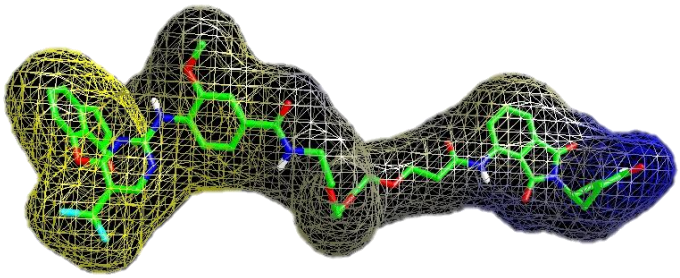
Università degli Studi di Torino

PhD in Pharmaceutical and Biomolecular Sciences

# A rational control of molecular properties to discover new oral drugs in the beyond-Rule-of-5 (bRo5) chemical space



Candidate: Diego García Jiménez

Supervisor: Professor Giulia Caron, PhD

Cycle XXXVI, 2020-2023

CYCLE XXXVI

2020-2023

DOCTORAL THESIS

# A rational control of molecular properties to discover new oral drugs in the beyond-Rule-of-5 (bRo5) chemical space


| | |
|---|---|
| *Doctoral candidate* | *Diego García Jiménez* |
| *Supervisor* | *Prof. Giulia Caron* |
| *PhD Program coordinator* | *Prof. Roberta Cavalli* |
| *Head of the Doctoral School* | *Prof. Alberto Rizzuti* |
| *Jury Members* | *Jan Kihlberg (external jury)*<br>*Carles Galdeano (external jury)*<br>*Giulia Caron (internal jury)* |


SCIENTIFIC FIELD: CHIM/08 - PHARMACEUTICAL CHEMISTRY

## Acknowledgments

*First and foremost, I am deeply grateful to my PhD supervisor, Giulia Caron, for the invaluable opportunity and support provided during this academic journey. From the day I entered your lab, you gave me the tools to develop myself as an independent researcher. Throughout these years, I have developed my own thinking style and mentality thanks to you. I am uncertain about what the future holds, but I doubt that I will encounter a better mentor.*

*I extend my appreciation to other beloved members of my research group, especially Giuseppe Ermondi and Maura Vallaro. You have carefully guided me in the computational and experimental part, respectively. Moreover, I want to express my gratitude to Matteo Rossi Sebastiano, my greatest colleague. Working with you, I realized that science is commitment, but also passion and fun. In addition, I must thank Cosmin Butnarasu, a former member of the lab. From the moment I arrived at the lab, you warmly welcomed and supported me, making my transition to Italy much easier. In addition, I would like to thank Jan Kihlberg for being my host in Sweden. My time there was very productive, and I am more than grateful for the scientific and personal support.*

*Additionally, I want to thank my family, including my sister Claudia and my parents, Mercedes and Flaviano. Especially the last two, for supporting me day by day in every aspect of my life. Since I was little, you have tried to give me the best opportunities that were available. I am fully aware and thankful for the efforts you have made. Moreover, my father Flaviano is the best example of what it means to be a good scientist.*

*A especial mention goes to Cristina for the personal support during this adventure, especially during my first year. Even more importantly, thank you for bringing out the best in me since we started pharmacy together. Thank you for pulling me into the pharmaceutical chemistry labs when I did not even know they existed. Without your help, I would not have even had the possibility of doing a PhD.*

*Lastly, I would also like to thank Italy and the city of Turin for the invaluable experience. Adjusting to life thousands of kilometers away is not easy, yet I have been fortunate to find my place in this country. Throughout my stay, I have encountered incredible people and forged numerous friendships that I will always keep by my side. I am pretty much certain that a piece of Italy will always stay with me, reminding me of the wonderful memories I made during my time here.*

*Grazie di cuore a tutti*

# Index

# 1. Abstract

bRo5 (beyond-Rule-of-5) compounds such as PROTACs, macrocycles and non-macrocyclic compounds exhibit revolutionary mechanisms of action but possess challenging physicochemical properties (i.e., extreme polarity and lipophilicity). As a result, they face *in vitro* ADME limitations (i.e., low water solubility and cell permeability) that impact their oral bioavailability. However, nature has provided some examples of bRo5 molecules that can behave as "molecular chameleons". This behavior can make the compound adapt to the environment by simultaneously displaying decent water solubility and cell permeability. Nevertheless, despite the interest in bRo5 compounds, the lack of property-based strategies slows down their development as oral drugs. Therefore, this thesis aims to design and implement experimental and *in silico* molecular property strategies, tailored for bRo5 compounds.

First, the experimental property profile (i.e., ionization, lipophilicity, polarity, intramolecular hydrogen bond formation) of a large set of bRo5 compounds was evaluated. Then, an innovative strategy combining polarity and lipophilicity was proposed to monitor the thermodynamic solubility of PROTACs. Furthermore, a novel chromatographic method to measure the chameleonicity of bRo5 compounds was disclosed. In practice, this method is able to capture the property change of a compound, when modifying the polarity of the environment. Notably, it is a high-throughput (HT) method that deserves to be included in any drug discovery program. Moreover, these molecular property strategies were integrated to rationalize oral bioavailability. Finally, the obtained experimental proofs provided a reasonable opportunity to rationalize or model bRo5 properties from *in silico* predictions.

Consequently, the thesis shifted its focus towards *in silico* approaches. 2D descriptor-based strategies (i.e., chemical space analysis) were tentatively employed to detect subregions of reference compounds with privileged permeability or bioavailability. Moreover, several 2D descriptor-based property models also succeeded in predicting the oral absorption of macrocyclic drugs. In addition, *in silico* strategies that account for the 3D conformation of bRo5 molecules were designed. In particular, these methods generate a set of possible conformations in polar and nonpolar environments, mimicking the outside and inside of the cell. Conformational sampling (CS) and molecular dynamics (MD) were optimized for this purpose. Subsequently, a series of 3D molecular descriptors were calculated for the obtained conformations and their results were interpreted. Overall, these strategies succeeded in reproducing specific experimental properties (polarity and chameleonicity) and shed light on the importance of dynamic intramolecular hydrogen bonds (dIMHBs) in chameleonicity. Moreover, work is in due course to integrate 3D descriptor-based descriptors and dIMHBs into property-based design strategies.

Overall, in my opinion this thesis has decisively contributed to fill the molecular property void between the Ro5 and bRo5 space.

## 2. Publications and academic merits

A. Publication list, ordered by appearance in the text

**I.** Garcia Jimenez, D.; Poongavanam, V.; Kihlberg, J. Macrocycles in Drug Discovery–Learning from the Past for the Future. *J. Med. Chem.* **2023**. *66* (8), 5377-5396.

**II.** Ermondi, G.; Garcia Jimenez, D.; Caron, G. Protacs and Building Blocks: The 2D Chemical Space in Very Early Drug Discovery. *Molecules* **2021**, *26* (3), 672.

**III.** Garcia Jimenez, D.; Rossi Sebastiano, M.; Caron, G.; Ermondi, G. Are We Ready to Design Oral PROTACs®? *ADMET DMPK.* **2021**. *9* (4), 243-254.

**IV.** Garcia Jimenez, D.; Rossi Sebastiano, M.; Vallaro, M.; Mileo, V.; Pizzirani, D.; Moretti, E.; Ermondi, G.; Caron, G. Designing Soluble PROTACs: Strategies and Preliminary Guidelines. *J. Med. Chem.* **2022,** *65* (19), 12639-12649.

**V.** Garcia Jimenez, D.; Vallaro, M.; Rossi Sebastiano, M.; Apprato, G.; D'Agostini, G..; Ermondi, G.; Caron, G. Chamelogk: A Chromatographic Chameleonicity Quantifier to Design Orally Bioavailable Beyond-Rule-of-5 Drugs. *J. Med. Chem.* **2023,** *66* (15), 10681–10693.

**VI.** Ermondi, G.; Garcia Jimenez, D.; Rossi Sebastiano, M.; Kihlberg, J.; Caron, G. Conformational Sampling Deciphers the Chameleonic Properties of a VHL-Based Degrader. *Pharmaceutics* **2023**, *15* (1), 272.

**VII.** Rossi Sebastiano, M.; Garcia Jimenez, D.; Vallaro, M.; Caron, G.; Ermondi, G. Refinement of Computational Access to Molecular Physicochemical Properties: From Ro5 to BRo5. *J. Med. Chem.* **2022**, *65* (18), 12068-12083.

**VIII.** Ermondi, G.; Garcia Jimenez, D.; Rossi Sebastiano, M.; Caron, G. Rational Control of Molecular Properties Is Mandatory to Exploit the Potential of PROTACs as Oral Drugs. ACS Med. Chem. Lett. **2021**, *12* (7), 1056-1060 (not discussed as part of the text).

B. Congress contributions and awards

- Invited speaker at the 2023 Genentech DMPK Symposium, San Francisco, USA.
- Poster presentation at the EFMC (European Federation for Medicinal Chemistry) 2021, EFMC-YMCS (European Federation for Medicinal Chemistry-Young Medicinal Chemists' Symposium) 2021, EFMC-YMCS 2022 and SEQT (Sociedad Española de Química Terapéutica) 2022.
- Featured cover for the *Journal of Medicinal chemistry*: "New therapeutic modalities" (October 2022).
- *Molecules* 2020 Best Paper Award.
- Flash talks at DMDG (Drug Metabolism Discussion Group) 2021, AMYC (Autumn Meeting for Young Chemists) 2021 and AMYC 2023.

## 3. Abbreviations and keywords

ADME: Absorption, Distribution, Metabolism, Excretion; ANNs: artificial neural networks; AR: Androgen Receptor; BCS: Biopharmaceutics Classification System; bRo5: Beyond Rule of 5; CDK: Cyclin-Dependent Kinase; CRBN: Cereblon; CS: Conformational Sampling; DL: deep learning; DMPK: Drug Metabolism and Pharmacokinetics; DMSO: Dimethyl Sulfoxide; DNNs: deep neural networks; DTs: decision trees; E3: Ubiquitin ligase enzyme; EPSA: Experimental Polar Surface Area; F%: Oral bioavailability; FDA: Food and Drug Administration; FN: False Negative; FP: False Positive; GB/SA: Generalized Born/Surface Area; GI: Gastrointestinal; GMean: Geometric Mean; GSE: General solubility equation; HBA: Hydrogen Bond Acceptor; HBD: Hydrogen Bond Donor; HCV: Hepatitis C Virus; HPLC: High-Performance Liquid Chromatography; HT: High Throughput; IAPs: Inhibitor of Apoptosis Proteins; IMHB: Intramolecular Hydrogen Bond; IUPAC: International Union of Pure and Applied Chemistry; LR: linear regression; MAPK: Mitogen-Activated Protein Kinase; MC: Monte Carlo; MCC: Matthews Correlation Coefficient; MD: Molecular Dynamics; ML: Machine Learning; MLP: Molecular Lipophilicity potential; LR: multiple linear regression: MLP; MMP: Matched molecular pair; MW: Molecular Weight; NAR: Number of Aromatic Rings; nC: Number of Carbons; ND: Non determined; NHR: Nuclear Hormone Receptor; NMR: Nuclear Magnetic Resonance; NOE: Nuclear Overhauser Effect; NPs: Natural products; NRotB: Number of Rotatable Bonds; PAMPA: Parallel Artificial Membrane Permeability Assay; PBPK: Physiologically-Based Pharmacokinetic; PCA: Principal Component Analysis; PD: Pharmacodynamics; PHI: Kier´s flexibility index; PK: Pharmacokinetics; PLRP: Polymeric Reversed Phase; PLSR: partial least squares regression; POI: Protein of Interest; PROTAC: Proteolysis Targeting Chimera; PSA: Polar Surface Area; QSAR: Quantitative Structure-Activity Relationship; QSPR: Quantitative Structure-Property Relationship; RF: Random Forest; $R_{gyr}$: Radius of Gyration; RMSD: Root Mean Square Distance; RSD: Relative Standard Deviation; SD: Standard Deviation; SFC: Supercritical Fluid Chromatography; SMD: Steered Molecular Dynamics; SOM: Self-Organizing Map; STK: Serine/Threonine Kinase; SVMs: support-vector machines; TC: Ternary complex; $T_m$: Melting Temperature; TN: True Negative; TP: True Positive; TPSA: Topological Polar Surface Area; UPS: Ubiquitin-Proteasome System; VHL: Von Hippel-Lindau protein.

## 4. Thesis organization

The content of this thesis is organized into thirteen chapters (1-13). Chapter 1, the abstract, summarizes the main objectives and findings. Chapter 2 includes the list of publications and academic merits derived from the thesis, while Chapter 3 presents the abbreviations used throughout the text. Chapter 4 provides instructions on how to read the thesis. Chapter 5 contains a general introduction, followed by the thesis objectives in Chapter 6. Chapters 7, 8, and 9 describe the original contribution of the candidate to specific fields, leading to a final conclusion in Chapter 10. Chapter 11 outlines the methods used, and Chapter 12 provides a list of bibliographic references. Finally, Chapter 13 contains the supplementary material.

The scientific content begins in Chapter 5, in which a general introduction to the characteristics of compounds belonging to the bRo5 space is provided. This chapter mainly focuses on their pharmacokinetic properties and the resulting challenges of their design for the oral route. Additionally, this section includes a critical overview of the available experimental and computational strategies for assessing molecular properties. The objective of this part is to assess the applicability of these strategies to bRo5 compounds, and to identify their strengths, weaknesses and opportunities for improvement. Chapter 6 will focus on the specific aims to be pursued in this thesis.

The following chapters (7, 8, and 9) are dedicated to the use of (7) *in silico* 2D descriptor-based strategies, (8) experimental strategies and (9) *in silico* 3D descriptor-based strategies to monitor molecular properties in the bRo5 space.

The order of the chapters has been chosen to follow the logical workflow used in bRo5 drug discovery projects. First, 2D descriptors are employed to refine the leads based on reference compounds or drugs (retrospective analysis). Next, the compounds are synthesized or obtained, and the experimental molecular property proofs are obtained. Lastly, 3D descriptor-based strategies are developed based on the aforementioned experimental evidence. In practice, once optimized, these strategies are intended to be employed immediately after the application of 2D descriptor-based strategies and just before the synthesis of a future lead.

This thesis is oriented versus the identification of molecular property strategies. Consequently, although constituting part of the published articles, content related to pharmacodynamics will not be discussed in the text.

Throughout the thesis, the scientific content is organized in an article-based style, each of which is identified in the text by Roman numbers (I-VII). Despite being divided into three chapters with different topics, the articles (I-VII) are intentionally interconnected and include contextualization that allows them to be read consecutively.

At the beginning of each chapter, there is a brief introduction to connect the contents. Finally, after each chapter, a summary of the main findings is provided. The text will exclusively narrate the most important findings of each article. Some results with less

relevance for the reader will be located either in the "Supplementary Material" (Chapter 13) or in the corresponding publication (Chapter 2). Throughout the text, the reader is referred to the Methods section exclusively when the employed techniques are innovative or crucial for the understanding of the article. The remaining details can be found in the methods section of each published article (Chapter 11).

The goal of each chapter is not only to discuss the most significant events but also to interpret them from a broader perspective and to evaluate them from the present point of view. Thus, the writing in this thesis is original and may include critiques or judgments that may differ from or complement the original publications. Therefore, data previously published with a given explanation now offer new insights and conclusions. Consequently, references now include literature that was not available at the time of publication (Chapter 12).

Specifically, in the experimental part (Chapter 8), all the mentioned techniques were applied, although only the key results were discussed. On the other hand, the *in silico* strategies discussed in Chapter 9 are supported by data that were unpublished at the time of publication, with particular emphasis on the information provided by the chromatographic descriptor "Chamelogk." This provided a stronger interpretation of the computational results.

Lastly, Chapter 10 summarizes the main findings discussed throughout the thesis, discusses the challenges requiring further comprehension, and proposes the next steps to be taken.

# 5. General introduction to the bRo5 technology

## 5.1 Drug discovery: a historical overview

Since the 19th century, the pharmaceutical industry has developed at a rapid pace, providing effective drugs for many diseases that were previously considered untreatable. This growth has been driven by advances in many scientific fields.[1] For example, the development of analytical chemistry triggered the discovery of drugs derived directly from nature (natural products, NPs) such as morphine or penicillin.[2,3] In addition, advances in organic chemistry made it possible to reproduce or modify (partially or completely) the biosynthetic pathways found in nature to obtain NPs or NP derivatives (i.e., semi-synthetic antibacterials).[4] Moreover, the progress in pharmacology led to the rational design of the first candidates for specific targets (i.e., agonists or antagonists) and the consequent development of the first synthetic *de novo* drugs.[5]

As a result, the 20th century was characterized by the rise of organic compounds with low molecular weight (MW), vaguely defined as small molecules. However, at the end of the century their approval suffered a slowdown caused by the exhaustion of synthetic pathways and traditional approaches, and the need for new therapeutic strategies often referred to as new chemical modalities. Moreover, at the same time, the development of biotechnology began to provide the first biologics (i.e., antibodies, cell therapies), which now represent a promising but still expensive alternative (beyond the scope of this thesis).[6,7] Thus, today we see how small molecules and biologics share the market.[7,8]

## 5.2 New chemical modalities

At the turn of the century, advances in genomics and human biology provided a new plethora of possible drug targets (genes, proteins, etc.) that were previously unknown or considered "undruggable" by traditional approaches.[9,10] This opportunity represented a boost for small molecule drug discovery which had remained stuck in traditional targets, but also allowed to develop new chemical modalities which occupied an interface between classical medicinal chemistry and biotechnology. Thus, this new modality included biotech-based products (i.e., RNA-based strategies, gene therapy), antibody-drug conjugates and specific categories of small molecules (covalent inhibitors, small peptides, protein degraders and macrocycles).[11] In particular, these categories of small molecules have innovative mechanisms of actions that require specific and complex structural moieties. As a result, these molecules do not normally meet the structural requirements, grouped under the Lipinski's Rule of 5 (Ro5), that a compound should have to be orally available (see Chapter 5.4).[12] Consequently, these molecules fall into a chemical space (ensemble of all possible synthesizable molecules) named beyond-Rule-of-5 (bRo5) space, with challenging pharmacokinetic properties (see Chapter 5.5).

This thesis will focus on chemical modalities belonging to the bRo5 space, mainly degraders (often referred to as PROTACs) and macrocyclic and non-macrocyclic compounds. In this section, their unique structural and pharmacological features will be reviewed.

### A. PROTACs

Degraders or PROteolysis TArgeting Chimeras (PROTACs) were first described by Crews and Deshaies[13] in 2001 and quickly gained popularity due to their capacity to attack "undruggable proteins", of great use majorly for cancer. Chemically speaking, they are heterobifunctional compounds composed of three building blocks (Figure 1): a warhead that binds to a protein of interest (POI), a ligand (E3 ligand) that recruits an E3 ligase enzyme, and a linker that connects both regions (structural details of the building blocks are provided as part of Papers II and III).

Their mode of action involves three consecutive steps: (1) the formation of a ternary complex (TC)(POI-PROTAC-E3 ligase complex), (2) the polyubiquitination of the lysine residues of the POI by the activated E3 ligase complex and (3) the degradation of the POI by the UPS (ubiquitin-26S proteasome system).[14] PROTACs function through an event-driven mode of action, in contrast to typical small molecules that exert their pharmacological effects by occupancy.[15] Thus, unlike classical inhibitors that rely on potency, PROTACs require the recruitment of the POI. Indeed, this mechanism provides a new toolbox of "druggable" proteins. In addition, PROTACs act through a catalytic cycle, reducing the required dose and therefore the likelihood of side effects.[16] In addition, PROTACs are expected to solve many treatment resistances (mainly in cancer), although cell resistance is not yet fully understood.[16,17]

However, despite their pharmacological advantages, degraders are bRo5 compounds that suffer from pharmacokinetic limitations (pharmacokinetic considerations will be discussed separately in Chapter 5.5). In fact, it was not until 2020 that the first degraders entered in clinical trials (ARV-110 and ARV-471; www.arvinas.com), both as oral treatments for cancer. Currently, up to 18 PROTACs are in clinical trials in 2023, mostly for cancer. In addition, PROTAC technology is expanding to autoimmune diseases, neurodegenerative diseases, infections etc.[18,19]

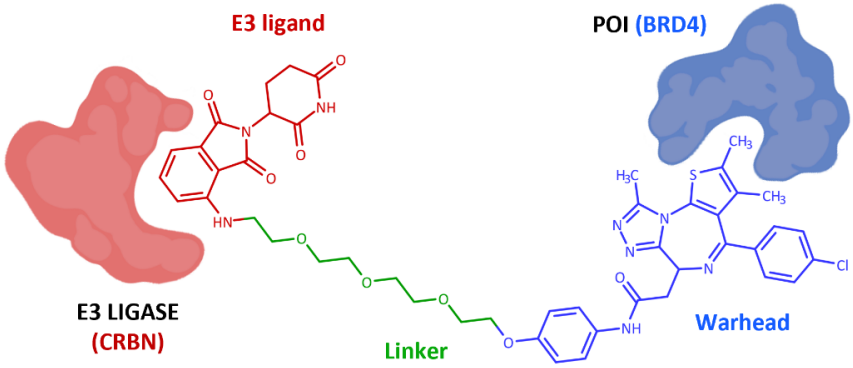



**Figure 1.** ARV-825 as a PROTAC example (E3 ligand-linker-warhead). It targets BRD4 as the POI and CRBN as the E3 ligase, colored in blue and red, respectively. Figure adapted from Garcia Jimenez *et al.,* 2021.[20]

B. <u>Macrocycles (peptidic or non-peptidic)</u>

The concept of a *"macro-cycle"* refers to an organic compound containing at least 12 heavy atoms in a ring. Macrocycles, despite being considered as a new therapeutic modality, are far from being recent. In fact, nature was the first to provide antibacterial and immunosuppressive macrocycles, such as erythromycin and cyclosporin, respectively (Figure 2).[21] Notably, these first macrocycles were poly-peptidic in origin and were baptized as "cyclic peptides". However, non-peptidic *de novo* designs are also becoming popular in the past years.[22]

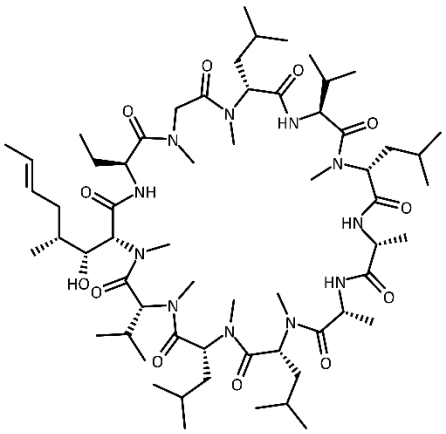


**Figure 2.** The structure of cyclosporin, a macrocyclic peptide.

Structurally, macrocycles have a semi-flexible and pre-organized disposition determined by the three-dimensional conformation of the ring and its side chains. These unique traits often give them the ability to bind with greater affinity and selectivity to targets that are inaccessible to conventional ring-opened equivalents.[23] Moreover, ring cyclization also provides pharmacokinetic advantages for oral dosing.[24] Thus, the strategy of rationally inducing ring closure, called "macrocyclization", is now a common optimization route for many acyclic compounds.[23–27] As a result, macrocycles (natural and *de novo)* are gaining interest from a pharmacodynamic and pharmacokinetic point of view, despite their often complicated synthesis (macrocycles will be treated separately in Chapter 7.1).

C. Non-macrocycles

The last category of the bRo5 space corresponds to non-macrocyclic compounds. These are large compounds that have in common a linear structure type. Overall, they have a varied origin that covers from natural terpenoids like paclitaxel[28] to synthetic antivirals as ritonavir.[29]

## 5.3 The oral route: bioavailability and molecular properties

### A. Factors governing oral bioavailability

The pharmaceutical research of a drug mainly involves the study of pharmacodynamics (PD) and pharmacokinetics (PK). PD focuses on the activity of a drug in the body (i.e., ligand-receptor affinity, activity, etc.), whereas PK focuses on the way in which the body deals with the administered substance. Traditionally, PK has been decomposed in 4 steps: the absorption, distribution, metabolism and elimination (ADME), which together with PD determine the final administration route of a drug.[30] Generally, the oral route is the most desirable route for any treatment with a desired systemic effect for several reasons (a comprehensive evaluation of the administration routes is beyond the scope of this thesis). From an industrial perspective, a drug intended for oral absorption is easier to manufacture and stock and is more cost-effective. From the patient's perspective, it implies better patient compliance and safety (non-invasiveness).[31,32] For example, in oncology, the oral route is preferred by 54 to 89% of the patients.[33]

The oral route, however, involves a series of steps (Figure 3) that limit the type of compounds that can be administered. First, the compound must be sufficiently soluble and stable in the variable pH environments of the gastrointestinal tract (GI); acidic pH in the stomach (variable depending on the fasting state), neutral in the small intestine and slightly basic in the large intestine.[34,35] Then, the compounds need to permeate passively (following an energy-free concentration gradient) or actively (with energy expenditure) through the intestinal membranes. During the whole thesis, the term "permeability" will refer to passive diffusion unless specifically mentioned. Notably, 3 main mechanisms are relevant: the paracellular (the compound permeates through the intercellular space between the enterocytes), transcellular (diffuses through the enterocyte) and active transport (uses protein transporters).[36] Generally, except for some exceptions that use active transport, drugs use passive transcellular transport. However, inside the enterocyte, drugs are prone to several efflux transporters ((i.e, P-glycoprotein (P-gp), multi-drug resistance-associated proteins (MRPs)) that can prevent the drug from passing through the enterocyte to the portal vein (Figure 3).[37,38] Eventually, drugs reach the liver where they are metabolized before entering the circulatory system.[34] Once in the systemic circulation the drug can be distributed and can carry out its action on extracellular and intracellular targets (after a new cell permeation).

Thus, the alternance of polar and nonpolar environments in specific steps (i.e., solubility, membrane permeability) calls for the monitoring of the chemical, physical and structural characteristics of the molecule of interest, named as molecular properties.[39] Moreover the drug formulation also determines the success in obtaining an oral drug but will not be reviewed in this thesis, since part of the development stage.

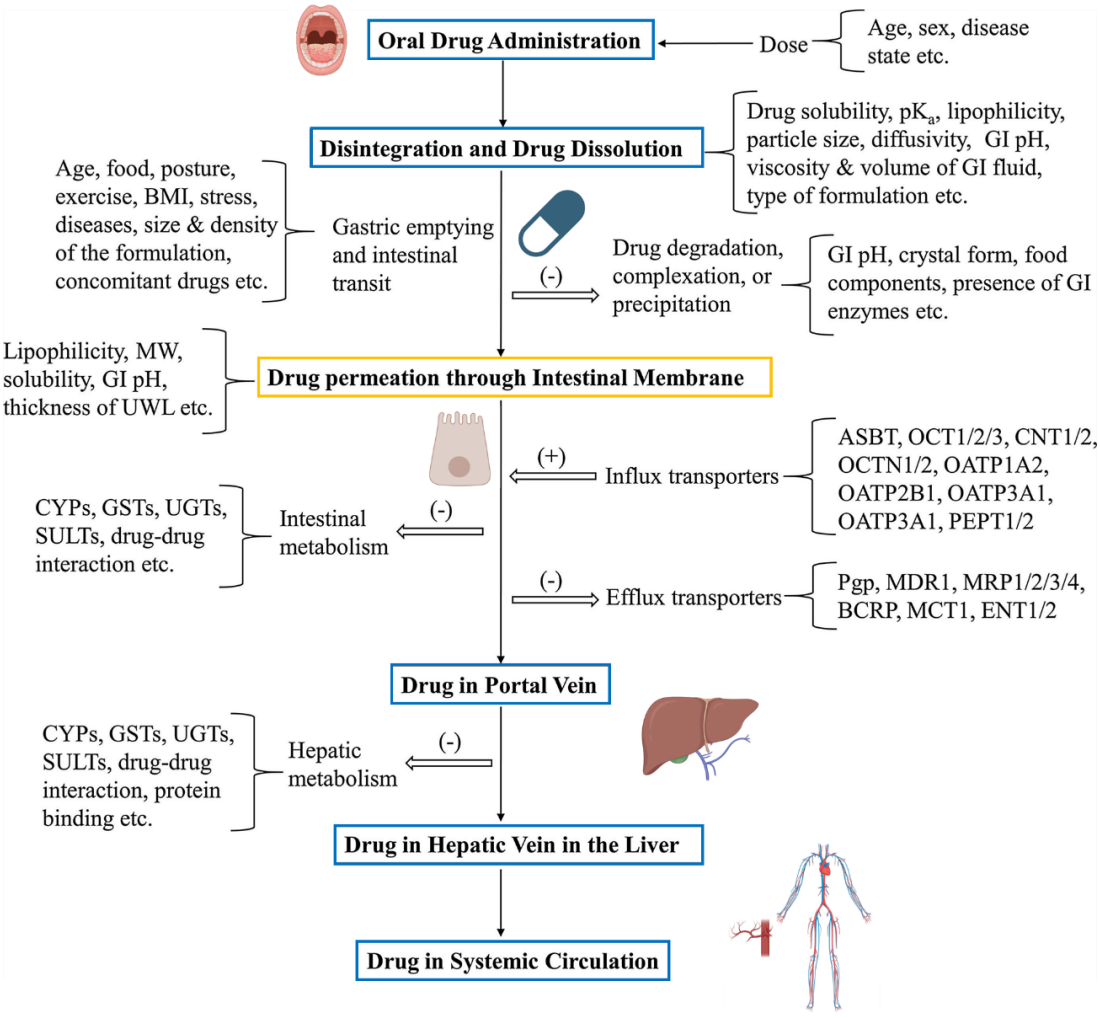



**Figure 3.** Schematic representation of the steps in the oral absorption process. Blue and yellow boxes refer to the presence of limiting polar and nonpolar environments, respectively. Figure adapted from Abuhelwa *et al.,* and Huang *et al.,* 2017 and 2009, respectively.[34,40]

Overall, the efficiency of a molecule's journey into the blood is measured by a parameter named oral bioavailability (F%), defined as the percentage of the molecule that reaches the systemic circulation after an oral administration. Therefore, to have a decent F%, the molecule of interest must have favorable molecular properties. Formally, these can be divided into *in vitro* ADME (metabolic stability, water solubility, cell permeability, etc.)[41] and physicochemical properties (polarity, lipophilicity, ionization, size, conformational features, etc.).[39] Normally, obtaining a favorable molecular property profile is feasible for small and Ro5 compliant molecules. However, it becomes notably more complicated in the bRo5 space (i.e., PROTACs and macrocycles, etc.).

B. Solubility and permeability: the BCS classification

In the design of an oral drug, a key pharmacokinetic requirement is for the molecule to be soluble enough to guarantee a desired concentration in an aqueous media. Solubility becomes highly relevant at various stages of the discovery process.[42] For instance, sufficient solubility must be guaranteed in the GI fluids (for absorption), in the blood plasma (for distribution) and in the body fluids (for elimination). Moreover, permeability represents another crucial molecular property for a drug to be absorbed in the GI tract, reach the systemic circulation, and penetrate into tissues and cells. Thus, a candidate must simultaneously have enough water solubility and cell permeability to be orally absorbed (other properties involved in the absorption, distribution, metabolism and elimination such as water and hepatic stability, protein binding and clearance will not be treated in this thesis, although they are also critical for the success of an oral candidate).

Oral bioavailability can be monitored by the balance between solubility and permeability. This strategy was addressed some time ago by the Biopharmaceutics Classification System (BCS), which groups drugs into four classes based on their water solubility and membrane permeability. Class I includes high solubility and permeability, Class II high permeability and low solubility, Class III high solubility and low permeability and Class IV, low solubility and low permeability.[43] Consequently, Class I represents the ideal category for the development of oral drugs. It is noteworthy that advances in pharmaceutical formulations have made it possible to partially solve some of these issues by introducing new dosage forms. For example, the use of salts represents a common strategy to improve the dissolution rate of Class II compounds.[44] However, this strategy can only be applied in the latest stages of preclinical development and is only performed for specific candidates (formulation will not be discussed). Moreover, in many other cases, a chemical modification of the lead is required, especially for Class III and IV compounds.[45] This is particularly evident when bRo5 molecules are considered. Consequently, from a drug design point of view, efforts must be made to refine the chemical structures in order to resolve this situation at later stages.

C. Modeling of oral route-related events: QSPR and PBPK models

An important aim of any drug discovery project is to obtain efficient models to predict the PK behavior of a drug candidate (bioavailability, tissue exposure, clearance, etc.). Two different but complementary approaches can be used (Figure 4). First, quantitative structure-property relationships (QSPR) relate the structural properties of a molecule, for instance polarity or lipophilicity, to a monitored parameter of interest, for instance, bioavailability.[46] Specifically, QSPR models can be built based on simple statistical regression or more complex machine learning (ML) models (see Chapter 5.8). Second, physiologically based pharmacokinetic (PBPK) models are also built based on the molecular properties of the candidate, but also include information about the pharmaceutical formulation and the characteristics of the living organism (a detailed analysis of PBPK models is beyond the scope of this work).[47]

Thus, both approaches effectively consider molecular properties (*in vitro* ADME and physicochemical molecular properties) to model critical steps of the oral absorption process (i.e., bioavailability, tissue exposure, clearance). Moreover, in both cases the quality of the experimental data is essential, often at risk when dealing with complex molecules.

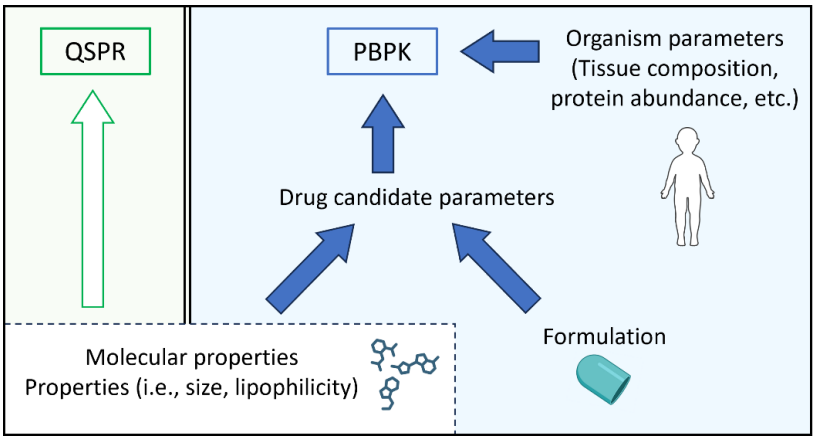



**Figure 4.** Schematic representation of the factors governing QSPR and PBPK models.

## 5.4 The definition of the bRo5 space

The growing speed of drug research programs at the end of the past century called for a rule of thumb to assess the capacity of a molecule to become an oral drug before it entered development. Thus, In 1997 Lipinski and coworkers came up with the "Rule of 5", a set of 5 criteria that a compound should meet to be orally available or to have "drug-like" properties.[12] These *in silico* rules indicate that an oral candidate must have no more than 5 hydrogen bond donors (HBD) and 10 acceptors (HBA), a molecular weight (MW) lower than 500 Da and a calculated LogP (logarithm of the octanol/water partition coefficient) not greater than 5.

These guidelines defined the limit between the Ro5 and the beyond the Ro5 chemical space and have been updated and improved over the years. For instance, a polarity quantifier named the topological polar surface area (TPSA or 2D PSA) was included by Veber. 140 $Å^2$ was therefore set as a polarity limit for oral drugs. Moreover, he also included a flexibility constraint, by limiting the number of rotatable bonds (NRotB) to 10.[48] In addition, other relevant studies (i.e., Kihlberg,[39] Whitty,[49] and DeGoey[50]) have also contributed to mapping the limits of the oral bRo5 space, mainly including new chemical modalities such as degraders, macrocycles and non-macrocyclic bRo5 drugs. In fact, Maple and coworkers recently defined, for the first time, the molecular property space of PROTACs.[51] These guidelines only include two-dimensional descriptors that are fast to calculate and easy to interpret. Their calculation requires a 2D representation of their structure, making the process easily automatable for large data sets.

However, many drugs are orally bioavailable despite having many Ro5 violations.[39] This has been evident over the past decades due to the increasing size (MW) of FDA-approved oral drugs.[52] Consequently, scientists are now questioning the rigidity with which Ro5 guidelines have been interpreted by medicinal chemists and are seeking new solutions adapted to bRo5 drug discovery programs.[53]

## 5.5 The bottleneck in the discovery of oral bRo5 drugs

The pharmacokinetics of Ro5 molecules can be optimized with traditional molecular property guidelines (i.e., Ro5, Veber, etc.).[54] However, bRo5 molecules cannot be optimized with the same rules of thumb due to structural and strategy limitations. These issues will be treated in the following sections.[55]

A. Inherent molecular properties

bRo5 molecules are known for their big size, high flexibility and extreme polarity or lipophilicity, which can confer several disadvantages. Their large size (MW) decreases solubility and permeability across the membrane.[56,57] Likewise, the efflux-pump (i.e., P-gp) which expels the drug outside of the cells, is also triggered by molecules with high MW. In addition, high polarity (HBDs, HBAs and TPSA) can provide higher solubility but limit notably their cell permeability. bRo5 molecules can also be known for their high lipophilicity (log P), which works in the opposite direction of polarity. High lipophilicity favors cell permeability but limits oral solubility. Furthermore, an excessively highly lipophilic drug can be retained in the membranes and compromise oral bioavailability. Moreover, high lipophilicity also promotes metabolic clearence,[58] protein-binding,[59] general toxicity (i.e., hERG cardiac channel),[60] etc.[61] Consequently, it is challenging to obtain a correct balance between polarity and lipophilicity that guarantees enough solubility and permeability. Thus, the search for oral bRo5 drugs becomes even more complicated when the improvement of molecular properties goes against the interests of potency or activity (PD) (Figure 5).

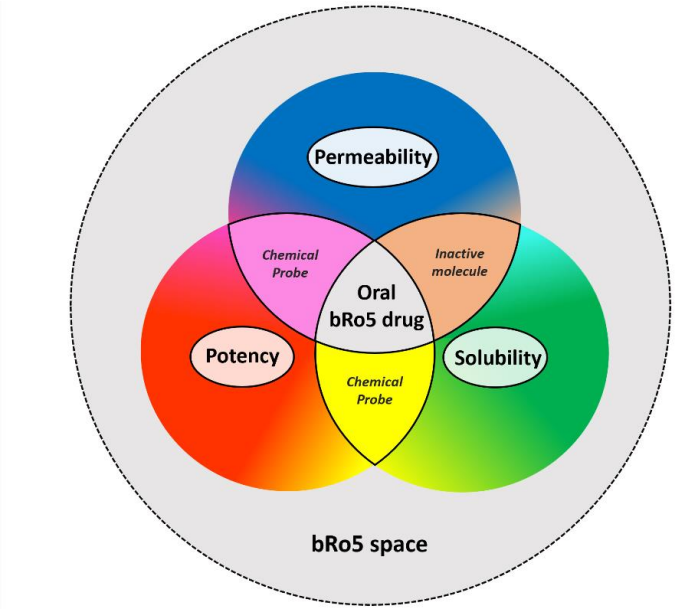



**Figure 5.** The quest for oral bRo5 drugs in the earliest stage of drug discovery: balance of permeability, solubility and potency. Figure adapted from Garcia Jimenez *et al.,* 2021.[20]

B. The conformational challenge

Notably, their high flexibility[62] complemented with their structural complexity confers to some bRo5 compounds the ability to adopt different 3D conformations depending on the specific environment considered (intestine, membrane, blood, etc.). Furthermore, each distinct conformer not only has a unique molecular property profile but also experiences different conformation populations based on the environment. This makes the biorelevant conformation in each environment difficult to know. However, to date, water solubility is known to be less affected by conformational variability than permeability.[63,64]

C. Inadequate tools to assess molecular properties

bRo5 compounds have unique molecular properties that make traditional strategies optimized for Ro5 compounds inefficient. Moreover, the relationship between the conformation in each environment and molecular properties, specific for the bRo5 space, also needs to be assessed with *ad hoc* experimental and computational strategies for the bRo5 space. However, most of these tools, if available, have theoretical and technical issues that will be separately discussed as part of Chapter 5.7 and 5.8 due to their importance for the thesis.

D. Misunderstanding of property-based strategies

Finally, the progress in the bRo5 space is slowed down by the almost exclusive use of traditional 2D descriptors by medicinal chemists.[65,66]

## 5.6 Opportunities for the discovery of oral bRo5 drugs: the role played by chameleonicity

As introduced earlier, some bRo5 compounds can modify their conformations and molecular properties based on the environment. This behavior was first detected by Carrupt and coworkers in 1991 and was termed "chameleonicity". More formally, it was defined as the capacity of a compound to display open and polar conformations in water and folded and less polar in nonpolar environments such as cellular membranes.[67] They observed this behavior for small glucuronides that varied their activity due to their dynamic molecular properties. Interestingly, chameleonicity had previously been observed in larger molecules, such as cyclosporin (CsA), depicted as a macrocycle in Figure 2.[68] In the case of CsA, this phenomenon had direct implications in solubility and permeability. Therefore, CsA represented the first rationalized example of a bRo5 drug having sufficient solubility and cell permeability to become an oral drug through molecular chameleonicity.

In practice, chameleonicity arises from conformational changes that modify the molecular properties of a compound in an environment-dependent manner. However, these structural changes can be very different in nature. The first and most well-known chameleonicity mechanism is the formation of dynamic intramolecular hydrogen bonds. In 2011, Alex *et al.* observed looking at experimental conformations that CsA in water was forming hydrogen bonds with the solvent whereas in nonpolar solvents the HBs were being formed within the CsA amides, therefore known as dIMHBs. In addition, these dIMHBs produced a conformational change from an extended to a more folded conformation inside the membrane interior. Consequently, they hypothesized that the dynamic formation of IMHBs was the key for the surprisingly high permeability of CsA.[69] Since then, the use of IMHB-mediated chameleonicity to enhance permeability has been studied and applied for several compounds, including PROTACs (Figure 6) (see Chapter 9).[70,71]

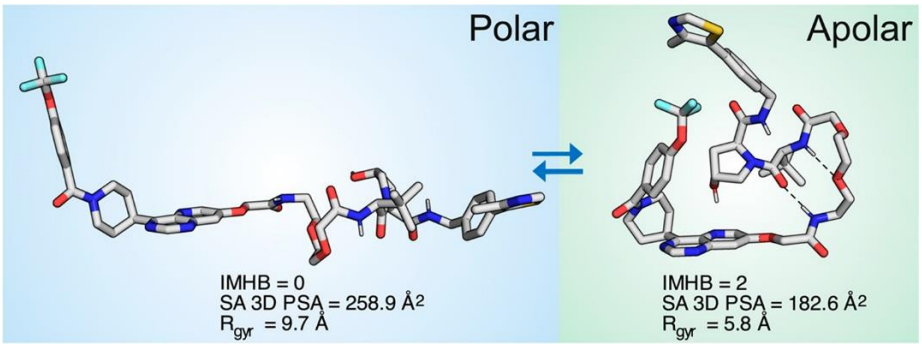



**Figure 6.** PROTAC-1, the first NMR-proven chameleonic degrader. Adapted from Atilaw *et al.,* 2021.[70]

Another common strategy involves polarity shielding by bulky and lipophilic substituents.[55,72] For example, the change of a cyclic penta-leucine with a cyclic hexa-

leucine peptide would theoretically provide higher polarity (2D PSA) and lower permeability (Figure 7).[72] However, the hexa-leucine undergoes a conformational rearrangement that orients polar residues toward the core of the macrocyclic ring, decreasing the 3D PSA and increasing permeability and bioavailability (F%). Finally, other less known strategies include increasing flexibility and modifying stereochemistry. Furthermore, the rearrangement of lateral chains and the establishment of dynamic π-π and van der Waals interactions can also be rationally planned.

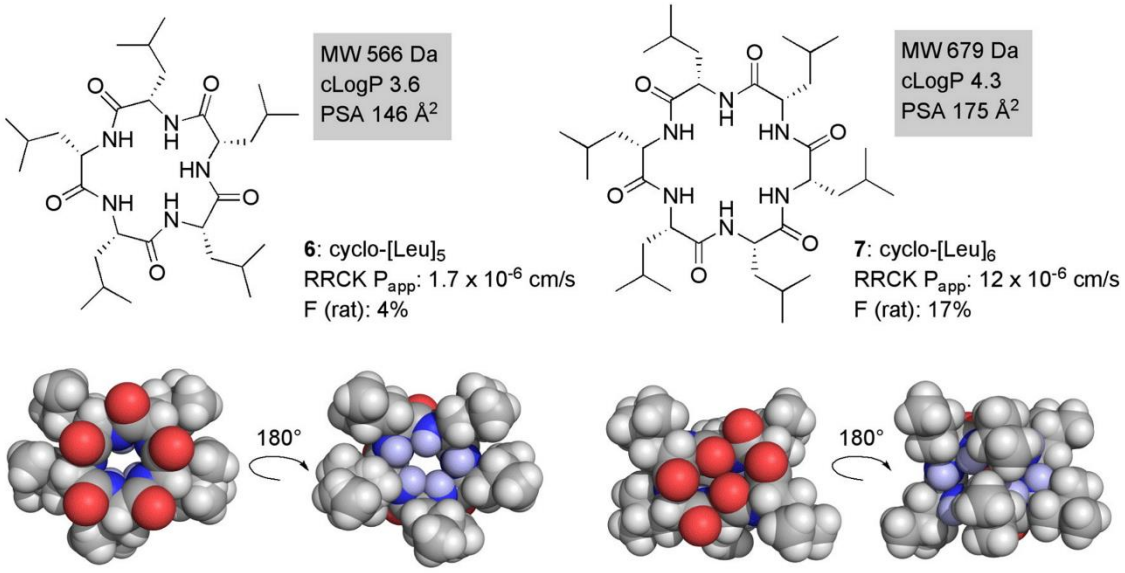



**Figure 7**. Structures, calculated lipophilicity, cell permeability and oral bioavailability in rat for cyclic penta- and hexaleucines 6 and 7 and their structures, determined by X-ray crystallography and NMR spectroscopy, respectively.[72] Figure adapted from Matsson *et al.,* 2016.[55]

Overall, chameleonicity embodies an opportunity to design *de novo* oral compounds beyond the Ro5.[49,64,73,74] However, the application of chameleonicity is for the moment a case-by-case strategy whose measurement and/or prediction is still under investigation (see Chapters 8 and 9).

## 5.7 Critical view of the experimental strategies to assess molecular properties in the bRo5 space

In this section, the available experimental strategies to assess physicochemical (ionization, lipophilicity, IMHBs, polarity and chameleonicity) and *in vitro* ADME properties (solubility and cell permeability) are reviewed and discussed. Special emphasis is given to the feasibility of their application to the bRo5 space (see Chapter 8).

### A. Physicochemical properties

#### 1) Ionization

The acidity constant ($pK_a$) is an important parameter to be evaluated for ionizable molecules. In the context of oral pharmacokinetics, the proportions of the major microspecies vary at the different pHs of the human body (gastric, intestinal, plasma), having a decisive impact on the *in vitro* ADME properties. For instance, an ionized molecule is more water soluble but has a lower cell permeability than the neutral form.[75,76] Thus, the ionization becomes relevant also in the bRo5 space (i.e., PROTACs, macrocycles). For instance, erythromycin derivatives have ionizable lateral chains that are essential for their chameleonicity.[77] Moreover, many orally available PROTACs use piperidine or piperazine linkers, which are charged at neutral pH.[78,79] In addition, some bRo5 compounds have 2 or more ionization sometimes even existing as zwitterions (i.e., rifampicin), further complicating the situation.[80]

Traditionally, the ionizable site has been determined by simple potentiometric methods.[81] However, this technique is highly sensitive, manual and requires high solubilities, making it unsuitable for many bRo5 molecules. A second technique is the UV-Vis spectrophotometric titration, which quantifies the absorbance (concentration) variation as a function of the pH. Finally, Caron's group observed that the ionization state of a molecule can be monitored by the capacity factor in the PLRP-S system.[82] In practice, the variation of the molecule's log k'80 (80% ACN in the eluent) at different pHs reveals its acidic, neutral or basic nature. However, this strategy does not reveal the exact $pK_a$ of the molecule but only the acidic/basic behavior of the analyte.

#### 2) Lipophilicity

Lipophilicity is formally defined by the IUPAC as the affinity of a molecule for a lipophilic environment. Testa and coworkers deconvoluted lipophilicity into a positive contribution of hydrophobicity (water repulsion) and a negative contribution of polarity (separation of electronic charges).[83] Thus, hydrophobic effects (hydrophobic and steric interactions, hydrophobic collapse, polar group shielding, etc.) increase the lipophilicity of a molecule, whereas the opposite is true for polar effects (HBD acidity, HBA basicity, polarizability of the solute, etc.).[83]

It is traditionally measured by the "shake-flask" method,[84] in which an organic solvent (often n-octanol that has been saturated with water, hereafter named octanol) is added to an aqueous solution of a chemical in a flask (liquid-liquid partition). The flask is then shaken to reach equilibrium and their phases are separated. Lastly, the concentration in both environments is measured and the ratio is calculated. The partition coefficient (P or log P) is expressed as:

$$(1)\ \text{Partition coefficient (P)} = [\text{neutral}]_{\text{octanol}}/[\text{neutral}]_{\text{water}}$$

The lipophilicity of a compound is expressed as log P for neutral molecules. However, in the case of ionizable centers in the molecule, the lipophilicity is expressed as log D at a specific pH. In this case, log D at a given pH is recalculated by replacing the neutral concentrations in Equation 1 with the ionized and neutral concentrations at that pH in octanol and water.[85] In both cases, the protocol is often slow and its outcome depends on many other factors such as the purity and solubility of the compound of interest.[86] Besides, due to its quantitative principle, it is very sensitive to the detection limits of the instrument used.

Moreover, other strategies for measuring lipophilicity, such as potentiometry or high-performance liquid chromatographic-based methods (HPLC), have been reported. The former is based on the quantification of the $pK_a$ change after the addition of octanol as a cosolvent to an aqueous solution.[87] However, as recalled earlier, this technique has several limitations and can only be used for molecules with ionizable centers.

HPLC-based methods, on the other hand, have gained popularity due to their high degree of automation and their ability to deal with impurities and low concentrations.[84,88] Moreover, they can adapt their mobile phase composition to the solubility of the analyte and avoid precipitation issues.

In this context, the lipophilicity of a certain molecule is measured as the capacity or retention factor (k' or log k') on a certain reverse-phase column (RP-HPLC) at a certain mobile phase composition (liquid-solid partition). Formally, k' is defined as the difference between the retention time of the analyte ($t_R$) and the dead time of the column ($t_0$), divided by the dead time itself ($t_0$). Mathematically, it is described by Equation 2:

$$(2)\quad k' = (t_R - t_0)/t_0$$

In addition, the measurement of the log k' values is insensitive to the concentration of the analyte (the retention on the column is maintained), which is ideal for low solubility compounds. This makes HPLC methods more suitable than shake-flask and potentiometric experiments for insoluble compounds with extreme lipophilicity values, such as bRo5 compounds.[89]

Under reverse RP-HPLC conditions, the compounds are retained proportionally to their partition coefficient on the generated system (generally reproducing the octanol/water, toluene/water partition, etc.). The system is created by the effect of the eluent (a mixture of aqueous and organic solvents) on the analytical column used. Columns are generally assembled by long hydrocarbon chains (i.e., C8 and C18, nonpolar) chemically attached onto a silica stationary phase (polar). Consequently, the high variety of columns and eluents available allow lipophilicity to be measured in different environments.

The use of RP-HPLC for lipophilicity has been exploited in the last two decades by various academic groups and companies, mainly to provide octanol/water surrogates.[85] Valko and coworkers developed the chromatographic hydrophobicity index (CHI) at GlaxoSmithKline (GSK)[89] which was correlated with cLogP values. In addition, Pfizer developed its own lipophilicity descriptors (ElogP or ElogD)[90,91] and validated (correlated) them with shake-flask experiments (octanol/water). Caron's group developed BRlogD, a surrogate of the octanol/water coefficient, focused on bRo5 compounds.[92]

BRlogD was conceived to provide newer alternatives to the well-known descriptor ElogP.[92] ElogP was designed as the capacity factor of the compound in the LC-ABZ column extrapolated to a 0% concentration of the organic solvent ($k'_W$ or $k'_{WATER}$).[93] This column contains a silica-based stationary phase attached to a an alkylamide siloxane chain with an internal amide group inserted near the silica surface. This amide hides the polarity of the silanol group from polar analytes, improving the efficiency of the assay over alkyl siloxane-bonded phases. However, this system is unstable at pH above 8, making it unsuitable for many experiments. Therefore, Caron and coworkers used the XBridge Shield RP18, a column with the same properties as LC-ABZ but stable over the entire pH range. Moreover, they verified that the k' at 60% of ACN (Acetonitrile), named BRlogD, is a quicker parameter than $k'_W$ to measure the lipophilicity of the octanol/water system for neutral compounds.[94] In addition, they also observed that for a set of bRo5 drugs there was a very good, although not perfect correlation with ElogP.[95]

Furthermore, RP-HPLC methods also have the capacity to generate environments that mimic other binary systems than the octanol/water. For example, immobilized artificial membranes (IAMs) are surfaces composed of a phospholipid attached to a silica surface, intended to mimic the environment of the cell membrane.[96] This type of biomimetic system has been used to model permeability and the blood-brain barrier (BBB) passage, and is suitable for neutral, charged or partially charged molecules.[97–99] The capacity factor for this system at 0% organic solvent (extrapolated value) has been defined as log $k_W^{IAM}$. In addition, the retention of charged molecules depends, not only on the hydrophobicity and polarity contribution in the system but also on the electrostatic interactions with the phospholipid heads.[100,101] Thus, while BRlogD is a pure surrogate of the octanol/water (average membrane lipophilicity), log $k_W^{IAM}$ measures lipophilicity in an environment that mimics the exterior of the membrane.

Just as the shake-flask method for octanol and water (log $P_{oct}$) became popular to model the average lipophilicity of the membrane, the log $P_{alk}$ (alkane/water) and log $P_{tol}$

(toluene/water) provided lipophilicity values for the interior of the membrane where polarity is low (dielectric constant ~ 2).[102,103] However, compounds especially bRo5, are barely soluble in alkane solvents. Toluene, for its part, absorbs in the UV-range and complicates HPLC-separation. Notably, HPLC-based columns that provide similar nonpolar environments are available and represent a better alternative. For instance, the PLRP-S (polystyrene/divinylbenzene) is a polymeric column that mimics the tails of the phospholipid.[82] In fact, Caron's group revealed that the capacity factor at 80% ACN (log k'80 PLRP-S) can be a surrogate of the log $P_{tol}$, by correlating the retention capacity to the calculated log $P_{tol}$ for a set of simple Ro5 compounds.

3) IMHB formation

The ability to form IMHBs is of great interest for bRo5 molecules. In fact, the dynamic formation of IMHBs in an environment-dependent manner can confer chameleonic properties to some molecules (see Chapter 9). To experimentally assess the formation of IMHBs, the difference between the octanol/water and toluene/water systems can be used.[104,105] Toluene provides a nonpolar environment (dielectric constant ~2.4) and promotes the folding (if possible) of the molecule through its IMHBs backbone. This is possible because there are no HBDs or HBAs available in the solvent to interfere with the molecule. On the contrary, octanol is more polar (dielectric constant ~9.8) and should favor more open conformations. In addition, the formation of IMHBs is unfavorable because the HBDs and HBAs of the octanol could interact with the HBA /HBD pairs of the compound, avoiding the formation of IMHB. Thus, the $\Delta$ log $P_{oct\text{-}tol}$ has been validated as an optimal strategy to reveal IMHB formation.[104,105] However, its validation involved shake-flask/potentiometric methodologies that hinder its application to the bRo5 space. Therefore, the application of this concept to HPLC methods seems vital for bRo5 compounds.[106]

4) Polarity

Polarity refers to the separation of electric charges in a molecule, which is determined by its HBD acidity, HBA basicity and polarizability.[83] As mentioned in the introduction, polarity is inversely correlated with cell membrane permeability. Experimentally, it can be measured by several chromatographic methods.[95,107]The experimental polar surface area (EPSA) is a superfluid ($CO_2$ as part of the eluent) chromatographic method (SFC) that uses a normal phase column, the Chirex 3014. It displays a balance of lipophilic and polar attributes that efficiently separates compounds according to their polarity. In this system, the polarity of the eluent is modified as a gradient by varying the percentage of methanol in an ammonium formate solution. This generates a nonpolar system that favors folded conformations, allowing the monitoring of cell permeability and to some extent also IMHBs.[108,109]

Polarity can also be quantified with a chromatographic descriptor named $\Delta$ log $k_W^{IAM}$. In 1997, Barbato and colleagues observed that the correlation between log $k_W^{IAM}$ and log P for a series of structurally unrelated neutral compounds was much better than for log $D_{7.4}$. This revealed that when charged states appear, the log $k_W^{IAM}$ is determined not only

by lipophilicity but also by the interactions with the phospholipid heads.[110] Moreover, to assess the lipophilicity of future compounds, an equation was built on a set of non-ionizable isolipophilic compounds (i.e., benzene, toluene) (log $k_W^{IAM}$ = 0.816*log P - 1.055). Consequently, the difference between the log $k_W^{IAM}$ value of an analyte and the extrapolated log $k_W^{IAM}$ value obtained by interpolating log P in the previous equation was defined as $\Delta$ log $k_W^{IAM}$.[111] Therefore, this value represents the excess of polar contribution (hydrogen bonding and electrostatic interactions) of the analyte. More recently, they also observed that the experimental log $k_W^{IAM}$ of compounds with PSA=0 (calculated polarity) correlated with their log $P_{oct}$ lipophilicity.[98,99] Furthermore, given the previously mentioned correlation between log $P_{oct}$ and BRlogD,[92] Caron and coworkers used the BRlogD descriptor as a surrogate of log P in the calculation of the extrapolated log $k_W^{IAM}$, from now on, named clog $k_W^{IAM}$ (clog $k_W^{IAM}$ = 0.92*BRlogD - 1.03).[97] Consequently, the final obtention of $\Delta$ log $k_W^{IAM}$ is feasible from pure chromatographic descriptors ($\Delta$ log $k_W^{IAM}$ = log $k_W^{IAM}$ – clog $k_W^{IAM}$). Moreover, this descriptor has recently been inversely correlated with permeability.[97,112]

5) Chameleonicity

The experimental evaluation of chameleonicity represents one of the biggest opportunities in drug discovery, but it is complex. At present, few tools are available to determine chameleonicity, all of which with important limitations. The most classical approach is the use of X-ray crystallography to explore the crystallized conformations of a molecule and analyze their structural and molecular property differences.[64,77,113–116] A low superposition of their poses and a different molecular property space are indicators of high chameleonicity. However, molecular structures are often crystallized in the same solvent due to solubility limitations, which is not ideal for studying dynamic behaviors. Moreover, the X-ray-based chameleonicity analysis strictly depends on the number of available crystals. X-ray also suffers from the “crystal packing effect”, which implies that crystallized conformations may not be seen in solution.

Nuclear Magnetic Resonance (NMR) serves as a more sophisticated approach, relying on the generation of molecular constraints extracted from the analysis of Nuclear Overhauser Effects (NOE) and proton coupling constants (*J* coupling). NOE and *J* coupling data form the basis for the NMR Analysis of Molecular Flexibility In Solution (NAMFIS). Through NAMFIS, the NMR-derived constraints are fitted with a series of computationally generated conformations (referenced in Chapter 9), allowing the extraction of relative population distributions in solution.[77,117–120] This strategy has the advantage of focusing on actual solution conformers, but it still mostly uses a semiquantitative case-by-case study method. Additionally, it is time-consuming, suffers from solubility issues and requires skilled training, making it inappropriate for early drug discovery. In addition, this strategy relies on computationally generated conformations to fit NMR constraints, which makes it partly dependent on predicted values.[121]

Furthermore, HPLC methods can also be used for assessing chameleonicity. As described in the lipophilicity section, ElogD and BRlogD constitute two correlated lipophilicity descriptors obtained in different environments.[95] Caron and coworkers theorized that

the difference between the two descriptors would reveal the chameleonic behavior, if any, of the compounds. This descriptor was therefore named ChameLogD.[95] Thus, ChamelogD represented the first HPLC-based descriptor for chameleonicity suitable for HT. However, since it independently measures two different environments, it lacks information about the behavior of the molecule in intermediate environments. Furthermore, it does not provide information about the structure of the retained conformations.

B. *In vitro* ADME properties

1) Solubility

Solubility, formally defined by the IUPAC as the "analytical composition of a saturated solution expressed in terms of the proportion of a solute in a given solvent," can be measured using two approaches in the pharmaceutical industry: kinetic and thermodynamic. Normally, kinetic solubility is the common choice, which consists of precipitating the compound by adding an aqueous phase to a DMSO solution.[12,122] It is a rapid method that requires only a small sample. However, it denotes solubility as "apparent," because it determines solubility before the establishment of the equilibrium between the stable phase or polymorph and the solvent.[122] Moreover, kinetic solubility often tends to overestimate solubility.[123] In contrast, thermodynamic solubility evaluates solubility after a "shake-flask" experiment conducted in aqueous media. This approach considers solubility at the equilibrium, presenting a more accurate measurement. Nonetheless, it requires a longer experimental procedure and larger amounts of sample and solid.

2) Permeability

The ease with which a compound passes through a membrane is defined as the apparent permeability ($P_{app}$) and is typically assessed in drug development by Parallel Artificial Membrane Permeability Assay (PAMPA) and cell-based models.[124,125] PAMPA quantifies the diffusion of a compound through an artificial phospholipid bilayer, which is a rough measure of passive permeability across a cell membrane. This method is popular due to its simplicity, cost-effectiveness, and relatively quick execution through automated platforms. However, it has limitations: it is unsuitable for high molecular weight (MW) compounds, which often encounter solubility issues and can exhibit non-specific binding.[66] Additionally, PAMPA lacks the presence of transporters, restricting its applicability primarily to compounds falling within Ro5 space.[50,126] On the contrary, cell-based models like Caco-2, MDCKII and or LLC-PK1 also evaluate active transport mediated by efflux transporters. These models are more complex and expensive but are better suited for bRo5 compounds. Hence, many bRo5 compounds that are absorbed by active transport and are susceptible to efflux transporters (i.e., saquinavir),[127] can only be studied by cellular models. Nevertheless, a direct relationship permeability and oral bioavailability has only been seen in specific cases.[55,126]

## 5.8 Classical and innovative *in silico* strategies to assess molecular properties in the bRo5 space

At the moment, there is a lack of default strategies for predicting the molecular properties of bRo5 compounds. In this section, the application of general and innovative *in silico ad hoc* strategies to bRo5 compounds is discussed.

### A. Molecular descriptors

They are defined as the result of logical and mathematical processes that transform symbolic representations of chemical information in a molecule into meaningful numerical values.[128] Moreover, molecular descriptors can generally be classified as 2D or 3D.

#### 1) 2D descriptors

2D descriptors capture the structural information of the molecule in a maximum of two dimensions and provide insights into their atomic composition, molecular connectivity and fragmental arrangement of the atoms. They can be divided into size-related (i.e., MW), hydrophobicity-related (i.e., clogP[129]), polarity-related (i.e., TPSA[130,131]), descriptors related to electronic effects (i.e., polarizability[132]), hydrogen-bonding related (i.e., HBD,[133] HBA,[134] Abraham´s acidity and basicity[135]) and topological descriptors (i.e., Wiener index[136]).[137,138] Generally, they are fast to calculate and provide straightforward information.

However, the use of 2D descriptors has several limitations. First the charge prediction represents a major issue. Most $pK_a$ calculators are trained on Ro5 compounds with just one ionizable group. Thus, the predicted $pK_a$ of bRo5 compounds is the first source of error that can considerably impact the calculation of molecular descriptors. For instance, lipophilicity (cLogD) depends on the predicted ionization state of the molecule ($pK_a$), often inaccurate. Moreover, the HBD count also depends on the considered ionization state. The second and most limiting issue is that 2D descriptors, normally atom- or fragment-based, do not consider the 3D conformation and are not environment-dependent.[66] This makes them only partially useful for the bRo5 space. Notably, the use of 2D descriptors will be explored in detail in Chapter 7.

#### 2) 3D descriptors

*In silico* strategies in the bRo5 space require the generation of 3D conformations in both polar and nonpolar environments (mimicking biological environments outside and inside the cell membrane, respectively) and the subsequent calculation of 3D molecular descriptors for the generated conformations.[64,66,131,139]

Their main goal is to identify biorelevant conformations with specific molecular properties that can be used to model experimental physicochemical properties and pharmacokinetic-related descriptors. Important applications of *in silico* 3D descriptor-based strategy include, for instance, the identification of a pool of low polarity conformers in nonpolar environments (i.e., chloroform or toluene) that are relevant for permeability modeling. Another application is the screening of high polarity conformers in water that can be used to model solubility.[64] Moreover, additional applications integrate the conformational energetic penalties when moving between environments to model permeability.[140,141]

In practice, these strategies suffer from limitations in the generation of conformers and in the nature of the 3D descriptors. On one hand, the generation of conformations depends on the algorithm used and the environment selected, which requires a deep and complex study. On the other hand, the 3D descriptors were not initially built or validated on training sets of bRo5 drugs, which complicates their interpretation and use. Furthermore, the choice of the best 3D descriptors depends on the specific aim. Additionally, the ionization state of the molecule is a crucial factor for conformation generation and molecular property prediction. However, introducing ionization consideration into the exploration of conformational space is not a trivial task. In fact, in these simulations different polarity environments are considered and several microspecies should coexist. Therefore, conformational search algorithms, designed to simulate one species at a time, tend to oversimplify the situation. In the following section, the current strategies and their limitations will be discussed.

### 2.1 Conformational search algorithm and solvent treatment method

The first step is to choose the conformer generation engine that provides the best ensemble of conformations in a given environment. Several methods are available to perform this task. Overall, they can be classified based on the algorithm and the solvent model used (Table 1).

| Conformation generation algorithm (C) | Solvent treatment (S)* | Approach name (C+S) |
|---|---|---|
| Distance geometry (DG) | Implicit | Conformational sampling (CS) |
| Dihedral angle-based sampling | | |
| Inverse kinematics (IK) formation | | |
| Monte Carlo (MC) | | |
| Molecular dynamics-based (MD) | Explicit | MD |

**Table 1.** Available conformation generation algorithms (C) and solvent treatment methods (S). S* reflects the most popular combinations of algorithm and solvent treatment method, although other combinations are possible. (C+S)* refers to the simplified names for the used combinations in this thesis.

According to the theory supporting the generation algorithms, the methods can be classified in distance geometry (DG),[142] dihedral angle-based sampling,[143] inverse kinematics (IK), Monte Carlo (MC) torsional methods and molecular dynamics (MD).[77,116,144] Moreover, mixed models are also available.[145] Of all, MC and MD-based methods are the most popular. MC methods sample the states of the system by randomly adjusting structural features (i.e., dihedral angles of rotatable bonds). In particular, the random sampling is based on the Metropolis-Hastings algorithm.[146] Then, the structures of the obtained conformers are checked, their energy windows are calculated and the redundant conformations are eliminated. In contrast, MD is based on the interactions between atoms according to Newton's physics. First, the atom-atom interactions and the energy of the system are calculated based on the application of an energy equation or force field.[147,148] This provides the position and velocity (trajectory) for each atom at a given time. Next, the application of this force as a function of time allows to simulate the evolution of a system over a desired time.

The solvation model can generally be either implicit, which treats the solvent as a continuous medium surrounding the compound, and explicit, which simulates the interaction between the compound and the solvent molecule itself.[149] The main difference between the two methods is the computation time and the information provided. Implicit methods simulate the effect of the solvent based on the solvation model used. The most common ones are the Poisson-Boltzmann (PB) and the Generalized Born (GB), which calculate the forces based on the temperature, the dielectric constant, the electrostatic potential of the solvent, etc.[150] In addition, this last model can be complemented with the hydrophobic solvent accessible surface area (ASA) to obtain the GB/SA model. On the other hand, explicit models mainly use the particle mesh Ewald (PME) method, which considers the simulation as a periodic cell. These methods allow the evaluation of the individual interactions between each solvent molecule and the compound of interest.[150] The selection of the strategy can vary upon the desired scope, but it normally considers the simulation of a compound in a nonpolar and polar environment. Overall, in this thesis we will exploit the use of conformational sampling (hereafter referred to as any conformational search algorithm except molecular dynamics) methods with implicit models and molecular dynamics with explicit models (Table 1).

## 2.2 Selection of the 3D molecular descriptors

Once the conformational search has been performed in a given environment, the next step is to calculate their 3D molecular properties (polarity, size, shape, lipophilicity, IMHBs, etc.).[139] Moreover, the RMSD (root mean square deviation), which measures the average distance between the backbone atoms, is normally assessed to evaluate the superposition between the conformers.

Polarity has been traditionally assessed in 2D by the TPSA, which uses a fragment-based approach that sums up the polar contribution (generally N, O, and the bound hydrogens) of each fragment in a rapid and effective way.[130,131] However, the bRo5 space needs the application of a 3D PSA to explore polarity differences between conformers. In practice,

its calculation first requires the calculation of the van der Waals surface, defined by the sum of the unburied atomic surfaces in the molecule (Figure 8).[151] Then, the outer boundary of the area defined by the motion of a probe sphere as it rolls over the molecular van der Waals surface, defined as the molecular surface area, is calculated. Accordingly, the molecular PSA (M 3D PSA), reported by Rossi Sebastiano and coworkers is calculated as the molecular surface area corresponding to only the polar and other partially charged atoms.[64] Moreover, they also reported the solvent-accessible PSA (SA 3D PSA) that measures the polarity accessible to a water-sized probe of 1.4 Å radius around the molecular surface (Figure 8).

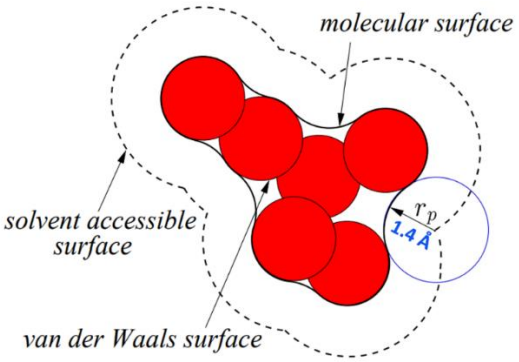



**Figure 8.** Graphical representation of the various surfaces of a molecule: van der Waals, molecular and solvent accessible. Abbreviation: $r_p$: radius probe. Figure adapted from Wei *et al.,* 2005.[151]

Molecular shape is quantified by the radius of gyration ($R_{gyr}$) as the root mean square distance of a set of atoms from their center of gravity.[152] For small molecules (MW ≤ 550), the relationship between MW and $R_{gyr}$ is fair. This suggests that $R_{gyr}$ is informative about the size for small molecules. This fact was also highlighted by our group very recently.[106] Moreover we observed that $R_{gyr}$ is rather a shape descriptor measuring the sphericity of the conformer. Furthermore, another common shape descriptor is the normalized principal moment of inertia (nPMI) plot. It is based on the assessment of the similarity of the conformation to a rod, disk, and sphere shape (Figure 9).[153]

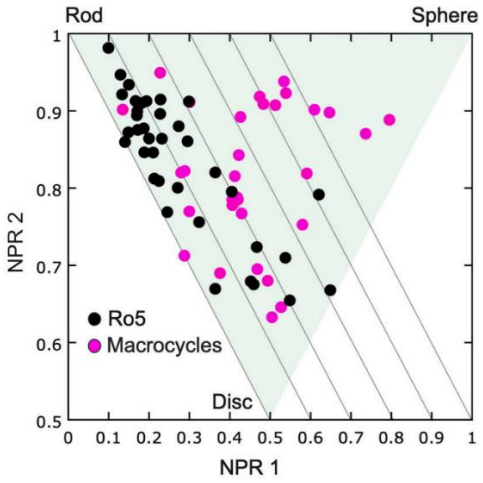


**Figure 9.** PMI plot (PMI coordinates NPR1 and NPR2) illustrating the shapes of the target-bound conformations of a set of FDA-approved macrocyclic drugs compared to the shapes of a set of Ro5 compliant drugs. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

Lastly, intramolecular interactions are also specific features of 3D conformations. In particular IMHBs (Figure 10) and polar and nonpolar interactions are of great importance for molecular properties such as chameleonicity. Above all, the predicted number of IMHBs (nIMHBs) is the most informative parameter.

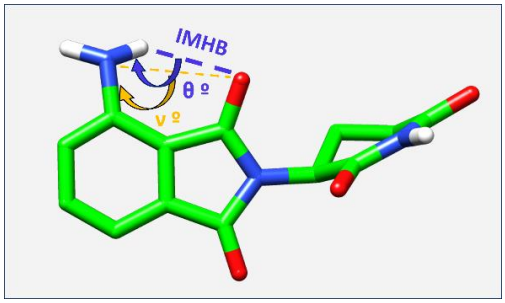



**Figure 10.** Representation of a predicted IMHB for pomalidomide, a Ro5 compound. The IMHB is present as a blue dashed line. Relevant angles and distances for the prediction of the IMHB formation are specified as suggested by Mills and coworkers.[154]

B. <u>The chemical space</u>

The goal of any drug discovery project is to predict a desired property directly from the use of simple descriptors. However, when more variables become relevant, the use of a chemical space becomes reasonable.[155,156] The chemical space is defined as the physicochemical or chemoinformatics limits that define an ensemble of compounds. Moreover, its creation relies on the selection of several descriptors used to monitor a desired property. More specifically, this concept becomes useful to explore structure-property (SPR) and structure-activity relationships (SAR) in pharmacokinetic and pharmacodynamic studies, respectively.[155,156] In fact, it can reveal regions with common characteristics, for instance the detection of zones with improved *in vitro* ADME properties. Generally, it is assumed that the chemical space needs to be visually explicable, typically using 2D or 3D plots.[155,156] However, the chemical space is an abstract concept that can use n variables in n dimensions, depending on the complexity of the property being monitored. Consequently, it is now common to use unsupervised ML techniques to deal with a great number of variables.[157] These strategies, such as principal component analysis (PCA) or self-organizing maps (SOM), allow dimensionality reduction while minimizing the information loss. Moreover, other subfields of ML like deep learning are beginning to be applied, not only to the visualization of chemical spaces, but also to the generation of efficient QSPR models (see below).[157]

Furthermore, the advances in the development of molecular fingerprints[158] have allowed the development of chemical spaces based exclusively on the "similarity" or

"dissimilarity" of the molecules under study. These fingerprints, which are essentially coding vectors containing the information of the substructure fragments of a molecule, have made it possible to create chemical spaces explained upon similarity networks.[159] This concept, often named as "similarity map", is useful to identify not only general molecular property characteristics but also specific structural moieties with relevance for the monitored property. A common application is the study of "activity landscapes", described as graphical representations that integrate similarity and activity relationships. These maps are useful for identifying pairs of compounds (matched molecular pairs, MMPs) that exhibit remarkable differences in activity despite their strong similarity, often referred to as "activity cliffs".[160,161]

1) Conventional use of the chemical space in drug discovery

The most common strategy is to create the chemical space upon 2D descriptors (i.e., MW, TPSA) in which to search for compound regions with privileged *in vitro* ADME property (see Chapter 7). This represents the fastest strategy and normally succeeds when atomic or fragmental differences between the studied compounds account for the property variation. This concept can also be related to property cliffs when 2D fingerprints are used or combined with 2D property descriptors. In practice, this tool represents the most classical approach in the optimization of molecular properties. For instance, it is obvious that most oral CRBN-based degraders (Cereblon as E3 ligase) in clinical trials have taken the property profile of ARV-110 and ARV-471 (the first oral CRBN-based PROTACs to enter in clinical trials) as a reference.[78] In fact, many of them use the same semi-rigid linkers (piperidine or piperazine derivatives) as ARV-110 and ARV-471, and are therefore located in the same chemical space regions.

2) Non-conventional use of the chemical space in drug discovery

The conventional approach often fails when the studied candidates are in the far bRo5 space. In these regions, the conformation can sometimes have a higher impact on the final molecular properties than the atomic or fragmental composition. In addition, the conformation can play a crucial role in the obtention of sound molecular properties for oral absorption (i.e., by chameleonicity). Therefore, the 3D descriptor-based chemical space needs to be implemented to explore certain regions of the 2D descriptor-based. In practice, Poongavanam and coworkers recently proposed the use of $R_{gyr}$ and 3D PSA as size/shape and polarity descriptors, respectively.[77] Nevertheless, the challenge is still how to generate 3D conformations, in which environment, and which are the biorelevant conformations. In Chapter 9, we focused on the refinement of this tool.

C. QSPR

QSAR/QSPR can be defined as computational methods that mathematically correlate the molecular structure of a compound, using molecular descriptors, to its molecular properties (chemical, physical or biological). These terms are denoted as QSAR when related to the pharmacological activity of the compound and as QSPR when related to

*in vitro* ADME properties.[162,163] Generally, these methods are founded on the principle of poly-linearity, which means that a linear relationship between the property and one or more structural features is required. However, non-linear (logarithmic and exponential relationships) are also possible.[163] Therefore, the identification of these mathematical correlations provides the opportunity to screen numerous compounds, even those not yet synthesized, for a given molecular property directly from their structure.

1) QSPR model building

To create these models, one should first search for the simplest correlations based on easy to calculate 2D descriptors (i.e., TPSA, MW, NRotB). From there, the complexity of the included descriptors can increase, including also 2D quantum chemistry-derived descriptors of the electronic structure (i.e., dipole moment, net atomic charges) obtained by semi-empirical calculations. Moreover, the final level of complexity involves the use of 3D descriptors (i.e., PMI, molecular volume, surface accessible surface area) that depend on the calculated conformation.

Afterwards, the QSPR generation requires the mathematical modelling of the found correlation.[163] Simple ML algorithms like linear regression (LR), multiple linear regression (MLR), and partial least squares regression (PLSR) are examples of statistical techniques to model linear relationships. Moreover, non-linear relationships can also be modelled by the use of other ML algorithms such as support-vector machines (SVMs), decision trees (DTs), Random Forest (RF), etc.[164] Lastly, deep learning (DL) algorithms like artificial neural networks (ANNs) or deep neural networks (DNNs) represent the latest advances for QSAR/QSPR modeling.[165,166]

2) QSPR derivatives

Since the definition of QSPR only includes direct mathematical relationships, the use of Rules of thumb (Ro5) cannot be formally defined as a QSPR strategy. Though, these rules were hypothesized from known correlations between molecular descriptors and *in vitro* ADME properties. For instance, Lipinski included log P in the Rule of 5 given the effect of hydrophobicity in solubility, introduced by Yalkowsky.[167] Other examples of QSPR derivatives are descriptor-based models whose output is not a numerical value but a discrete variable (i.e., soluble or insoluble). In these cases, the application of logistic regression and non-linear classification models are common.[168]

# 6. Aims

In line with the limitations and opportunities related to molecular properties, the main aim of this thesis is to expand the knowledge of the bRo5 space at the experimental and computational level and to develop new *ad hoc* strategies to be used in drug design projects. Thus, the thesis specifically aims to:

- Exploit the use of ***in silico* 2D descriptor-based** strategies to set up the 2D chemical space of bRo5 compounds (mainly macrocycles and degraders) and propose applicative strategies for the design of oral bRo5 compounds (Papers I, II and III).

- Design, implement and fine-tune **experimental** methods to measure *in vitro* ADME (permeability and solubility) and physicochemical (ionization, IMHB formation, lipophilicity and polarity) molecular properties of bRo5 compounds. Furthermore, an important goal is to develop new chromatographic strategies to measure molecular chameleonicity (Papers IV and V).

- Design new ***in silico* 3D descriptor-based** strategies to predict molecular properties through the modeling of their conformational variability by CS and MD-based methods. Again, a special focus will be put on monitoring the chameleonic behavior. Moreover, the developed *in silico* strategies will be integrated with the obtained experimental evidence to generate predictive models (Papers VI and VII).

## 7. *In silico* 2D descriptor-based strategies to monitor molecular properties

*In this chapter, the chemical space of several groups of bRo5 compounds is created in order to identify regions of drug discovery interest (Figure 11). To this end, the first step requires the evaluation of the simplest approaches, the use of 2D descriptors. As previously explained, they have several limitations but can be of great utility in the refinement of ADME properties. In particular, the articles in this section discuss the creation of 2D descriptor-based (i.e., MW, TPSA, etc.) chemical spaces that provide information on relevant molecular properties for the oral absorption route (i.e., permeability).*[24]

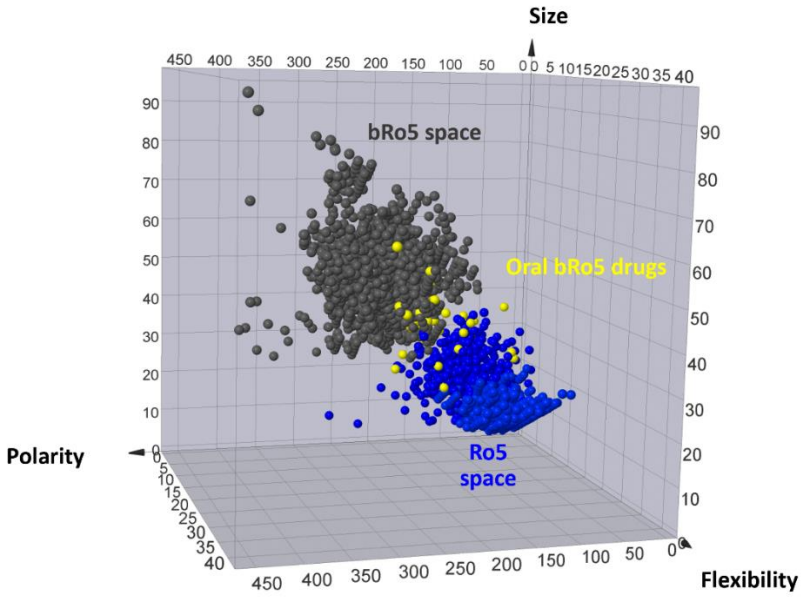



***Figure 11.*** *Schematical representation of a simple 2D descriptor-based chemical space using polarity (TPSA), size (MW) and flexibility (PHI) to cluster Ro5 and bRo5 molecules.*

## 7.1 Macrocycles in Drug Discovery–Learning from the Past for the Future (Paper I)[22]

Macrocycles have been extensively popular in the clinic since the XX$^{th}$ century thanks to their diverse clinical indications (i.e., antibacterials, immunosuppressants). However, it is only since the *de novo* design of complex bRo5 structures became feasible that medicinal chemists have begun to study their privileged PD (mechanism of action) and PK (*in vitro* ADME properties) in order to mimic them.[21,24]

Macrocycles therefore require an individual state-of-the-art analysis focusing on their pharmacodynamic and pharmacokinetic potential. With this aim, different analyses were performed mainly on two generated data sets: macrocycles approved by the FDA[169] (n=67) and macrocyclic candidates in clinical trials in the US (n=34) by 2022.[170] Moreover, two additional data sets included: literature entries containing information of the past 15 years (509 articles)[22] and a data set with all the macrocycles reported in ChEMBL (n = 28052).[171]

This manuscript discusses important advances in the pharmacological uses (indications, targets, etc.) and modes of action (binding form, site, etc.) of macrocycles. However, since the main scope of this work is the study of molecular properties, this information will not be discussed. For additional information on this topic, the reader is referred to the original publication (Paper I).

### A. <u>The molecular property space</u>

The first goal was to examine the potential of macrocycles to serve as oral drugs through a retroprospesctive study based on their molecular properties. Among the FDA-approved macrocycles, 39% of the macrocyclic drugs are orally bioavailable and 41% are parenteral (administered elsewhere in the body other than the digestive system). In addition, the ratio of approved oral to parenteral macrocycles has remained stable over time (Figure S1). Moreover, only 32% of the clinical candidates are intended for oral absorption. Thus, these similar trends reveal that despite the efforts to achieve oral dosing, oral macrocycles are still challenging to obtain.

To investigate the potential factors governing this fact, the molecular property space of the approved macrocycles was calculated. Although 2D molecular property descriptors may be less useful than 3D descriptors for characterizing the chemical space of bRo5,[64,95] they are simple to compute and can provide useful information, especially for comparisons with Ro5 candidates.[24,39] For this purpose, their 2D structures were retrieved from Drugbank,[172] adapted to pH 7 (average for physiological absorption) and submitted to the calculation of 10 molecular descriptors (see Methods). Subsequently, the approved oral and parenteral macrocycles were compared based on their molecular descriptors (Figure 12 and Figure S2). The most informative descriptors were MW and the number of carbon atoms (nC) for size, TPSA, HBA and HBD for polarity, NRotB for flexibility and calculated logP (cLogP) for lipophilicity. Moreover, the logarithm of

aqueous solubility (cLogS), included as a solubility predictor also showed notable differences between the two groups. Results suggest that oral macrocycles are smaller (MW), less flexible (NRotB), less polar (TPSA, HBD and HBA) and more lipophilic (cLogP) (Figure 12). These limits reveal that oral absorption is favored by low polarity and high lipophilicity. In fact, these values do not respect Lipinski's and Veber's guidelines[12,48] for MW, TPSA, and HBA (see Introduction), which reveals that these guidelines need an update, at least for macrocycles.

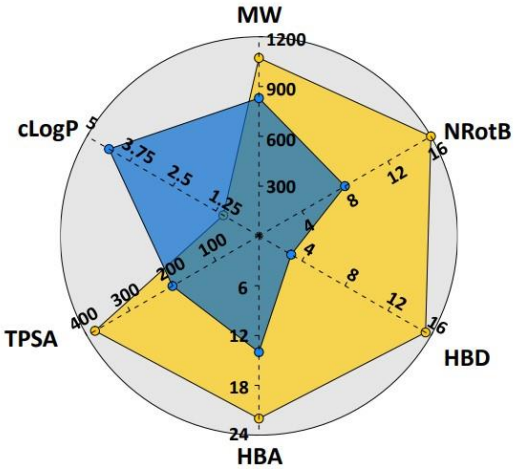



**Figure 12.** Radar plot comparing the median values for the descriptors employed in Lipinski's Ro5 and Veber's rule for the oral (blue, n = 24) and parenteral (gold, n = 38) subsets of FDA-approved macrocyclic drugs. Only Lipinski's and Veber's descriptors are represented. The remaining descriptors (i.e., cLogS) can be found in Figure S2. Another 5 additional compounds were discarded due to molecular property calculation errors. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

Moreover, we performed a PCA as an alternative way to plot and compare the molecular property space of macrocycles (Figure 13). Principal components 1 and 2 (PC1 and PC2) contain 91% of the data set variability, being lipophilicity (cLogP) the most contributing variable. As denoted by simple 2D descriptors, oral macrocycles occupy a narrower space than the parenteral space. More specifically, the oral space is formed by 4 major subregions. The main subregion is represented by most oral macrocycles, as represented by the blue centroid in Figure 13. Lorlatinib, pacritinib and moxidectin form a second cluster with molecular properties close to the Ro5 space. Cyclosporin and voclosporin (cyclosporin derivative) represent a third area, outside of the Ro5 space. In theory, these two are too lipophilic and polar, which would compromise the solubility-permeability balance. However, cyclosporin has been proved to be a molecular chameleon that can hide its polarity, therefore obtaining enough cell permeability and solubility to be an oral drug.[114,173] Given its similarity to cyclosporin, voclosporin also behaves as a chameleon. This was recently proved by our new chameleonicity descriptor, the Chamelogk.[121] Finally, octreotide[174] and desmopressin,[175] two cyclic peptides, constitute the final cluster. Both are orally absorbed but have low F% (4 and < 0.16%, respectively). Oral octreotide is available as an alternative to subcutaneous

administration and desmopressin is potent enough to achieve a decent effect even with low bioavailability.

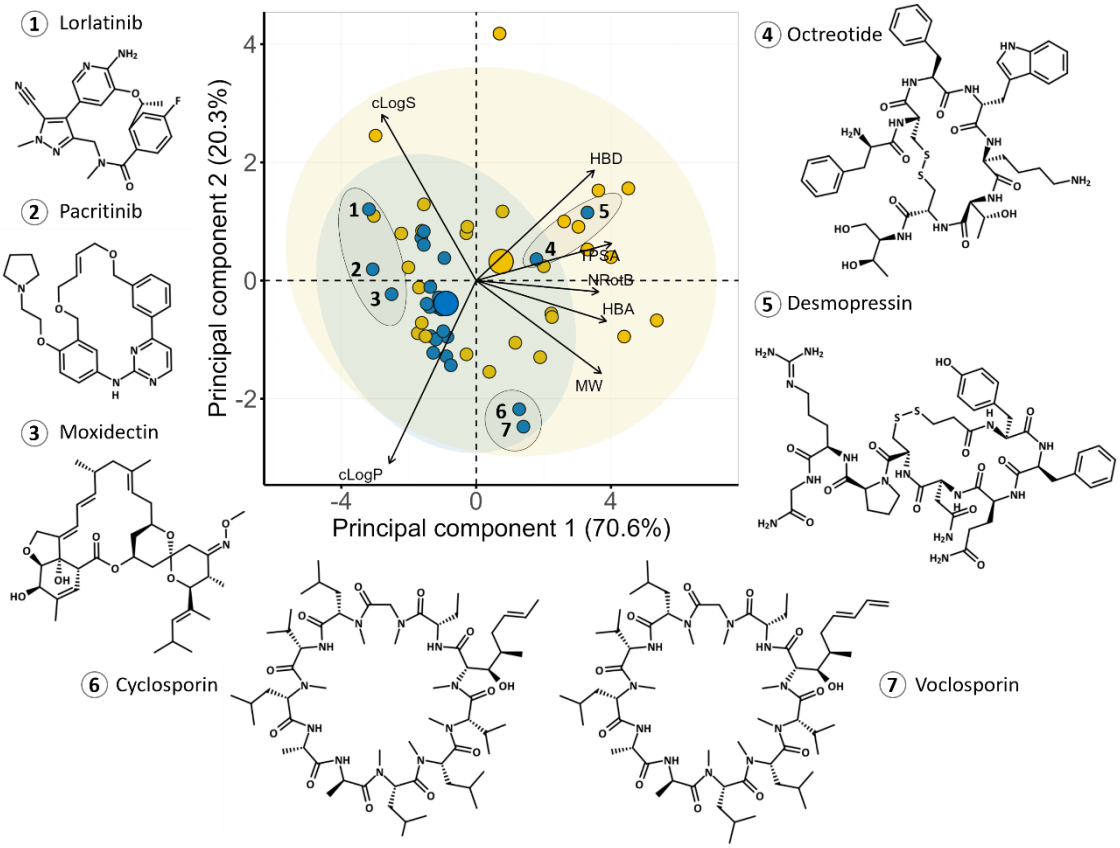



**Figure 13.** PCA of the chemical space of the macrocyclic drugs data set (n = 53). The PCA was based on the descriptors of Lipinski's and Veber's rules, as well as cLogS, calculated at pH 7.0. Ellipses in blue and yellow shading show the 95% confidence intervals for orally and parenterally administered macrocycles, respectively. The centroid of each class is indicated with a large circle in the color of the respective class. The contributions of individual descriptors to the PCAs are indicated by the length of the arrows. The structures of three Ro5 compliant macrocycles (1−3), two analogues of cyclic peptide hormones (4 and 5), as well as cyclosporin (6) and voclosporin (7) are provided. Another 5 additional compounds were discarded due to molecular property calculation errors. Nine parenterals with MW > 1500 Da were excluded in the PCA to provide a better dissection of the chemical space. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

Clinical candidates followed a similar same trend. Interestingly, their molecular property distribution remained closer to the Ro5 space than the approved oral macrocycles (Figure S3). Moreover, PCA suggested that odalasvir, in clinical trials for the oral treatment of HCV infection, behaved as an outlier with an extremely high lipophilicity. It should be noted that its structure is atypical for a macrocyclic drug. However, clinical trials are prone to changes in their formulation and administration route. Therefore, although the general picture remains similar, property distributions may undergo slight changes.

Lastly, to gain some insights into the future, all the macrocycles contained in ChEMBL (a database of bioactive molecules) were downloaded and their properties calculated.[171] ChEMBL macrocycles explore a wider chemical space than approved macrocycles and macrocycles in clinical trials, but their median values did not show major differences (Figure S4). However, the macrocycles in the ChEMBL set are slightly smaller, more lipophilic, less polar (TPSA and HBA), less flexible (NRotB) and have rings of lower size than the drugs and clinical candidates. Moreover, its biggest difference compared to the drugs and clinical candidates is the low number of stereocenters. This could reveal that ChEMBL is mostly based on *de novo* designs. In addition, when compared to all the macrocycles with oral and parenteral absorption, CHEMBL was revealed to be surprisingly similar to the oral space.

B. Predictive models for oral macrocycles

The comparison of molecular descriptors and the PCA analysis revealed remarkable differences between orals and parenterals for both the approved and clinical trial candidates (Figure 13 and Figure S3, respectively). Consequently, we attempted to find simple cut-offs that could be used in the screening of novel macrocyclic drugs intended for oral absorption. For all the descriptors, a density plot including orals and parenterals was performed and the major separation between the clouds was taken as a cut-off (Figure 14). This intersection allowed the evaluation of single descriptor models to separate both populations, using the FDA-approved set as the training set (to build the model) and an external set of macrocycles as the test set.[24,39]

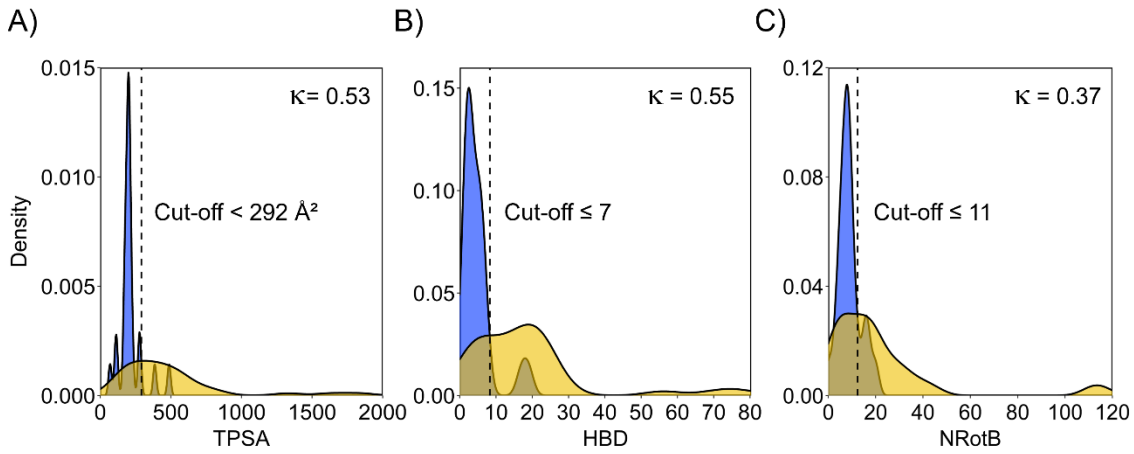



**Figure 14.** Single-property distributions for HBD A), TPSA B), and NRotB C) for the oral (blue) and parenteral (gold) subsets of the macrocyclic drugs training set (n = 62). The black dashed line indicates the intersection point of the density plot, and the derived cutoff value is given adjacent to the dashed line. The reliability of single-property models based on each of the three descriptors for the differentiation of oral and parenteral drugs in the training set is given by the Cohen's kappa (κ) value: a value higher than 0.6 suggests that the level of agreement between the predictions is moderate.[176] The evaluation of the 10 single-descriptor models is reported as Table S1. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

Thereafter, a confusion matrix (see Methods) was created, and several statistical and performance parameters were calculated to evaluate the models. The ML evaluators suggested that HBD (≤ 7) and TPSA (< 292 Å$^2$) were the best performing descriptors, while NRotB (≤ 11), for example, was one of the worst ones (Cohen's kappa = 0.53, 0.55 and 0.37, respectively) (Figure 14 and Table S1). The single HBD and TPSA models were found to be very efficient in the detection of orals (TP), given their high sensitivity (88% and 92%, respectively). Nevertheless, their low specificity values (71% and 66%, respectively) revealed that parenterals (TN) were not as efficiently predicted. In addition, similar results were observed for the test set of clinical candidates (Table S1).

Single descriptor models were not accurate enough and therefore, bi-descriptor models were explored. Combinations of the best 4 descriptors involved in single-descriptor models (MW, HBD, TPSA and cLogP) were performed. Overall, HBD (≤ 7) plus any of the other three descriptors (MW, cLogP and TPSA) gave the best models in the training set (Table 2). In fact, the best improvements over single descriptor models were in specificity or capacity to identify parenterals. Thus, sensitivity remained equally high (83-92%), while parenterals were discriminated with 74-79% specificity, which is an important improvement over any single-descriptor model. Furthermore, desmopressin and octreotide (cyclic peptides) are the only oral drugs that are misclassified by all the three models (Figure S5). In the case of the test set (83-94% sensitivity, 67-71% specificity), the prediction of parenterals also improved notably.

**Table 2.** Most accurate bi-descriptor models for prediction of oral bioavailability for macrocycles. Cutoffs were selected based on the major intersection between orals and parenterals in the density plots for each descriptor as calculated for the approved macrocycles (Figure 14). Positive values stand for "oral". Abbreviations: Sens. (Sensitivity), Spec. (Specificity), Acc. (Accuracy), Kappa (Cohen's kappa) and Geometric Mean (GMean). Other combinations are found in Table S2. Table adapted from Garcia Jimenez *et al.,* 2023.[22]

| Bi-property models | 1st cut-off | 2nd cut-off | Confusion matrix | | | | Sens. | Spec. | Acc. | k |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | TP | TN | FP | FN | | | | |
| Training set (n=62) | HBD (≤ 7) | MW (< 982 Da) | 20 | 30 | 8 | 4 | 0.83 | 0.79 | 0.81 | 0.6 |
| | | cLogP (> 2.22) | 21 | 30 | 8 | 3 | 0.88 | 0.79 | 0.82 | 0.64 |
| | | TPSA (< 292 Å$^2$) | 22 | 28 | 10 | 2 | 0.92 | 0.74 | 0.81 | 0.62 |
| Test set (n=60) | HBD (≤ 7) | MW (< 982 Da) | 15 | 30 | 12 | 3 | 0.83 | 0.71 | 0.75 | 0.48 |
| | | cLogP (> 2.22) | 17 | 28 | 14 | 1 | 0.94 | 0.67 | 0.75 | 0.51 |
| | | TPSA (< 292 Å$^2$) | 17 | 28 | 14 | 1 | 0.94 | 0.67 | 0.75 | 0.51 |

Moreover, the 3 bi-descriptor models were also examined with the data set of macrocycles in clinical trials (Table S3). In this case, the model based on HBD ≤ 7 and cLogP > 2.22 improved the other 2 (91% sensitivity and 67% specificity). However, this data set only had 11 orals and results might not be fully reliable.

Finally, the novelty and utility of these models was compared to already existing ones. The AbbVie's multiparameter scoring function (AB-MPS) [AB-MPS = Abs(cLogD-3) + NAR + NRotB] predicts the oral F% for bRo5 candidates (NAR, number of aromatic rings).[50] A value ≤ 15 differentiates between oral and parenteral bRo5 drugs. Consequently, our data sets were submitted to the AB-MPS score. The approved set and the test set were classified with moderate sensitivity (79 and 61%, respectively) and good specificity (71 and 79%, respectively) (Table 3). However, for macrocycles the AB-MPS was outperformed by our three HBD bi-property models which predicted orals with 83-94% sensitivity and discriminated parenterals with 67-79% specificity (Table 2).

**Table 3.** AB-MPS evaluation of the approved macrocycle data set (test set 1) and test set. Table adapted from Garcia Jimenez *et al.,* 2023.[22]

| AB-MPS score | Cut-off | Confusion matrix | | | | Sensitivity | Specificity | Accuracy | Kappa |
|---|---|---|---|---|---|---|---|---|---|
| | | TP | TN | FP | FN | | | | |
| **Test set 1** | **AB-MPS ≤ 15** | 19 | 27 | 11 | 5 | 0.79 | 0.71 | 0.74 | 0.48 |
| **Test set 2** | | 11 | 33 | 9 | 7 | 0.61 | 0.79 | 0.73 | 0.38 |

Thus, our results conclude that HBD ≤ 7 in combination with one of the following: MW < 1000 Da, cLogP > 2.5, or TPSA < 300 Å$^2$ are rules that are easy to remember and can be applied to screen the oral potential of new macrocyclic candidates. Beyond these limits, the likelihood of finding an oral candidate is low.

C. Macrocycle origin (*de novo* or nature-based)

As recalled in the introduction, most macrocycles have been obtained from nature, many of them as cyclic peptides. This trend is reflected in an analysis of the FDA-approved macrocycles: 88% are nature-based (directly obtained from nature or partially modified), 67% of which are oral (Figure S1). Overall, the main reasons for optimizing natural compounds were pharmacokinetics (i.e., oral bioavailability, half-life, chemical stability, resistance to proteases and solubility) and pharmacodynamics (improved potency, broader spectrum of activity and reduced side effects).

In addition, the number of *de novo* drugs has increased in recent years, although only 12% are *de novo*. In fact, since 2008 (plerixafor) several *de novo* drugs have been approved (Figure S1). As expected from their complex synthesis, *de novo* compounds stay close to the Ro5 space while natural-based compounds can also be far outside. Consequently, all but one is orally absorbed.

The fact that almost all *de novo* candidates are dosed orally prompted us to analyze the effect of HBDs on this trend.[177] As expected, *de novo* have significantly fewer HBDs than nature-based macrocycles (Figure 15), and they are almost exclusively amidic. Nature-

based macrocycles, on the other hand, have a striking quantity of alcoholic HBDs (phenols and aliphatic alcohols).[178,179] Therefore, it seems that oral *de novo* macrocycles need to have a maximum of 2 HBDs of an amidic nature. Macrocycles in clinical trials showed similar trends; 82% are nature-based and *de novo* candidates (18%), mainly contain amide HBDs over alcoholic HBDs (data not shown).

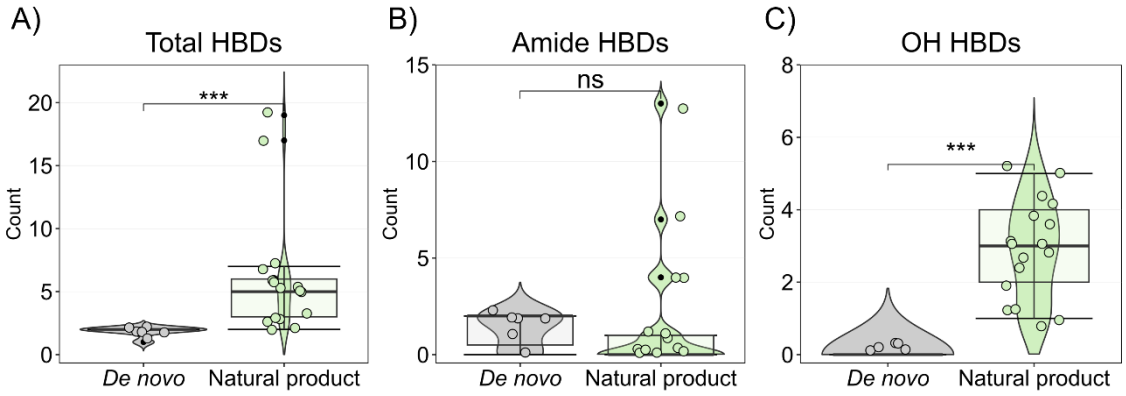



**Figure 15.** A) Comparison of the number of HBDs in orally bioavailable macrocyclic drugs discovered by *de novo* design (n = 7) or from natural products (n = 17) for the charge state calculated at pH 7.0. Frequencies of HBDs originating from B) amide moieties and C) phenols and aliphatic alcohols (OH) in the two classes of drugs. Box plots show the 50th percentiles as horizontal bars, the 25th and 75th percentiles as boxes, and the 25th percentile minus 1.5× the interquartile range and the 75th percentile plus 1.5× the interquartile range as whiskers. Black dots represent values higher than 1.5× the interquartile range and less than 3× the interquartile range at either end of the box. Violin shapes represent the data density at each count value. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

### D. Paper I conclusions

In this work, the past, present and future of macrocycles in drug discovery has been investigated. For this purpose, several data sets were built, including a) FDA-approved macrocycles, b) macrocycles in clinical trials in the US, c) research articles on macrocycles since 2005 and d) macrocycles available in the ChEMBL database. The analyses of these datasets enabled us to acquire a deeper understanding of their indications, targets and binding modes (PD, data not shown), molecular property space, ADME models for oral absorption and macrocyclic origin.

The study of their administration routes suggested that only 40 and 30% of the FDA-approved macrocycles and clinical candidates are absorbed orally, respectively. In general, oral macrocycles occupy a restricted region in the chemical space, which is only expanded by verified molecular chameleons such as cyclosporin. Parenterals partially share the chemical space of orals but are mostly spread towards regions with higher size, flexibility, polarity and lower lipophilicity. With this knowledge, simple bi-descriptor models to distinguish orals from parenteral macrocycles at pH7 were built and tested with decent performance (approximately 90% sensitivity and 70% specificity). Thus, HBD

≤ 7 in combination with either MW < 1000 Da, cLogP > 2.5 or TPSA < 300 $Å^2$ were set as simple guidelines to be used in the design of oral macrocycles. For macrocycles outside, the shielding of HBDs through molecular chameleonicity should be considered.

Lastly, an in-depth study of the origin of approved and clinical candidates suggested that most of them are derived or obtained from nature (> 80%). However, since 2008 *de novo* macrocycles have started to appear in the clinic. However, their progress is slow due to their complex synthesis. As a result, nature-based macrocycles occupy a wider chemical space whereas *de novo* ones stay closer to the Ro5 space. This is reflected by the lower number of HBDs and their HBD types: d*e novo* macrocycles have almost exclusively amide HBDs and lack alcoholic HBDs. Nature-based compounds, on the other hand, are mainly composed of alcoholic HBDs and only a few amide HBDs in cyclic peptides. Thus, *de novo* designs of macrocycles, at least from our experience, should not have more than 2 amide HBDs. Overall, this study has shed light on the potential of macrocycles as drugs.

## 7.2 PROTACs and Building Blocks: The 2D Chemical Space in Very Early Drug Discovery (Paper II)[180]

Despite their growing popularity, PROTACs represent quite a novel technology. From the discovery of the first degrader until 2017, only ten articles were published annually. Since then, the interest in PROTACs has grown exponentially, reaching a yearly frequency of nearly 100 by 2020.[181] At that time, little was known about their pharmacodynamics and pharmacokinetics. In fact, only a few studies had focused on the PK behavior of PROTACs and their bRo5 nature.[182] In 2019, Edmonson and coworkers[183] calculated several molecular descriptors of about 40 PROTACS, including Lipinski's[12] and Veber's[48] descriptors and the AB-MPS score.[50] Later that year Maple[51] generated a manually curated data set of 422 public degraders and explored their molecular properties, comparing them to other bRo5 compounds.

Due to the exponential rise of public PROTACs in the following years, we reasoned that these studies needed an update. Moreover, the creation of two databases for PROTACs, the PROTAC-DB[184] (http://cadd.zju.edu.cn/protacdb/) and the PROTACpedia, (http://protacdb.weizmann.ac.il/ptcb/stats) provided more than 2500 PROTACs and represented an opportunity to expand the knowledge introduced by previous studies. Moreover, PROTACs are a novel type of bRo5 compounds (i.e., macrocycles, non-macrocycles) that require *ad hoc* molecular descriptors for their design. For example, it was recently published that the NRotB and the calculated logP are inaccurate for bRo5 compounds.[62,112]

Thus, in this paper we aimed to a) reveal useful resources providing PROTAC-related structures, b) select and calculate a set of *ad hoc* 2D descriptors for PROTACs and discuss their potential advantages and disadvantages and c) create a PROTAC chemical space, focusing also on their building blocks (i.e., linker, warhead and E3 ligand) and their contribution to the overall PROTAC structure.

### A. Database availability

When this work was performed (2020), the PROTAC-DB[184] contained the chemical structures, biological activities, a few physicochemical characteristics, and the relevant references of 1662 PROTACs and their building blocks. PROTACpedia, on the contrary, contained only 779 and highly overlapped with the chemical space PROTAC-DB (Figure S6). Therefore, the PROTAC-DB was used as the source of PROTACs. In both cases, the experimental information on molecular properties is missing and the information on degradation activity is in many cases incomplete.

### B. 2D descriptors

Calculated descriptors, such as those reported by Lipinski[12] and Veber[48] can be divided into structural descriptors (MW, HBA, HBD, NRotB, etc.) and experimental predictors

(clogP, cLogS, etc.). Overall, they have been used efficiently in many drug projects, mostly involving Ro5 candidates. However, the flexibility descriptor NRotB, has been proved to be inaccurate especially when comparing cyclic (macrocycles) and non-macrocyclic (PROTACs) structures.[62] The underlying reason is the neglect of flexibility in cyclic structures, which obviously have a certain flexibility. Other issues arise from the lack of correlation between experimental predictors (cLogP) and experimental values for new modalities such as PROTACs. Obviously, the success in the use of these descriptors for the analysis of large data sets is undoubtable, but their use in early stages of bRo5 projects where few compounds are available is not clear.[112] These problems are exaggerated when the $pK_a$, whose prediction for bRo5 compounds is already imprecise, is considered.[112] Therefore, in order to provide a solid starting point, we decided not to use descriptors that may cause misunderstanding such as cLogP and NRotB.

As an alternative to NRotB, the Kier's flexibility index (PHI) was selected as flexibility descriptor because of its better capacity to handle the flexibility differences between cyclic and acylic structures. Moreover, MW and TPSA were selected as molecular size and polarity descriptors, respectively. Finally, a selection of count descriptors related to the polar (HBA and HBD) and nonpolar part (the number of carbon atoms, nC and NAR) were selected (Figure 16). In total, three correspond to polarity (TPSA, HBD and HBA), two to hydrophobicity (NAR and nC) and one to size (MW) and flexibility (PHI). nC, however, can also be considered as a size descriptor (normally larger molecules include higher carbon counts).

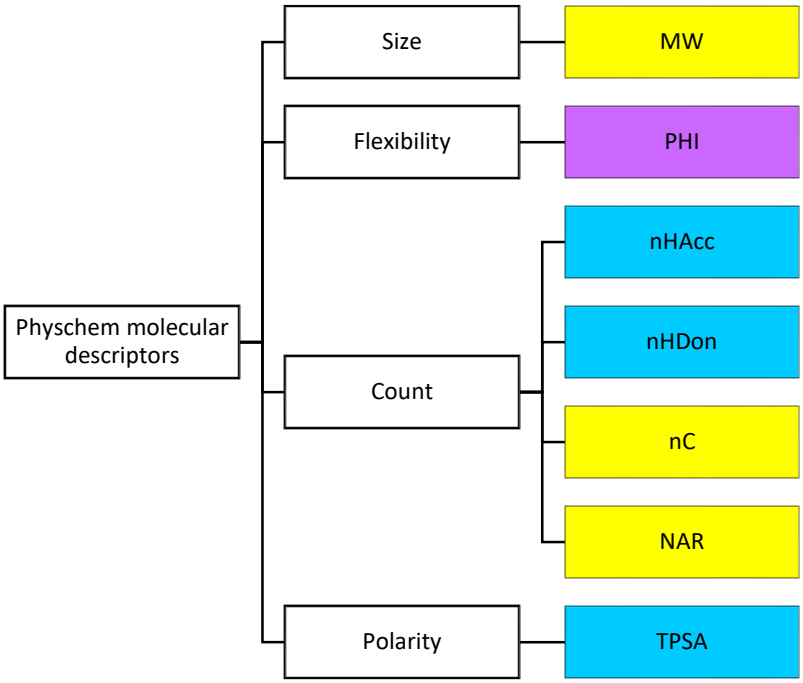


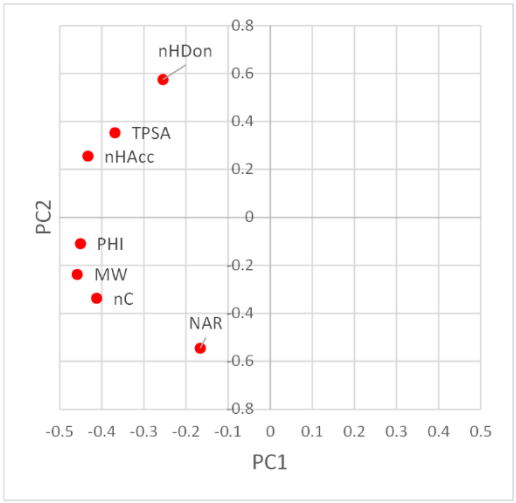

**Figure 16.** Molecular descriptors used in this study (yellow: descriptors related to hydrophobicity, light blue: descriptors related to polarity, violet: flexibility descriptors). HBA and HBD were named as nHAcc and nHDon in the original publication, respectively. Figure adapted from Ermondi *et al.,* 2021.[180]

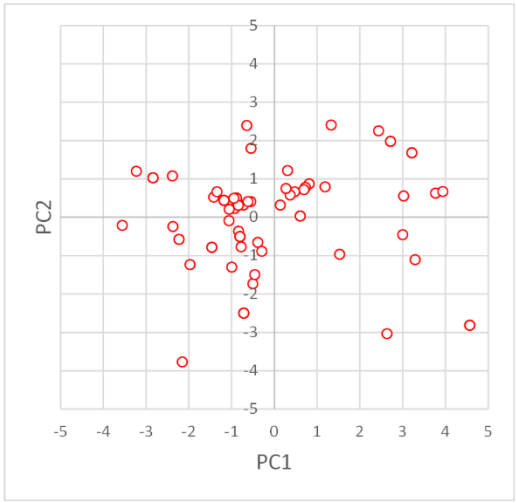
The structural complexity of PROTAC resides not only on the complete structure but on the individual contribution of its 3 parts (warhead, E3 ligand and linker). Consequently, the structural codes for 1662 PROTACs and their building blocks (806 linkers, 65 E3 ligands and 202 warheads) were downloaded from the PROTAC-DB and submitted to the calculation of several 2D descriptors (see Methods). Additionally, the same 7 descriptors were calculated for 52 oral bRo5 drugs published in the literature for comparative purposes.[39,185]

Overall, PROTACs are larger (MW), more polar (TPSA, HBA, HBD), more flexible (PHI) and more hydrophobic (NAR) than their building blocks (Table S4). Moreover, to unveil the molecular property trends originated from their structure, the relationship between the 2D descriptors was examined for each group using a correlation matrix (Figure S7). For each class, polarity (TPSA) was not related to nC (maximum Pearson's correlation coefficient, R observed < 0.6), which supports the choice of nC as a good descriptor of hydrophobicity. TPSA increases with the molecular size (MW) of the molecule (R range= 0.5-0.8). The underlying reason for this is not clear; the inclusion of nitrogens (amides) or oxygens (esters) may be a shortcut for synthesis or may have a pharmacodynamic purpose (increase binding affinity). This fact remains inconclusive but is key to understanding the issues of PROTAC with permeability. Furthermore, PHI was mainly correlated with MW (R = 0.9). Thus, flexibility inevitably increases with size, as expected. Linkers, specifically, showed a weak inverse correlation between PHI and NAR (R = -0.24), suggesting that aromatic rings in the linker reduce the flexibility of the linker and thus of the PROTAC.

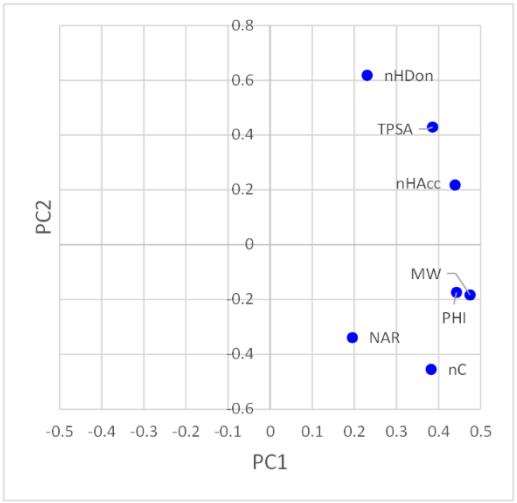
These relationships were evaluated using PCA (Figure 17). In all cases, PC1 and PC2 explained 70%-85% of the variance within the groups (PROTACs and building blocks). Furthermore, the loading plots (first panel) showed that the 2D descriptors for warheads, E3 ligands, and PROTACs behaved in a similar manner (Figure 17A, B, and D). The polarity descriptors were located in the same quadrant (TPSA, HBD, and HBA). However, size (MW and nC) and flexibility (PHI) were in opposite quadrants. In addition, NAR is placed in a different region of the second quadrant and is the least contributing variable to PC1 but the most to PC2. Interestingly, nC was placed between MW and NAR, revealing a mixed information profile between hydrophobicity and size. However, these last two descriptors were more variable within the subgroups. Finally, the loading plot for the linkers (Figure 17C) showed that they behaved differently from other compounds. PC1 explained most of the variance (59.9%) by polarity, size, and flexibility descriptors, while PC2 (17.4%) was only governed by flexibility (PHI) and aromaticity (NAR). These last descriptors were placed in opposite regions, justifying the negative correlation discussed earlier. Moreover, PC2 (score plot, second panel) divided the linkers into those at the top lacking aromatic rings and those in the middle and bottom with 1 and 2, respectively. Thus, the linker, which is the most variable part in the design of a PROTAC, must be studied separately.

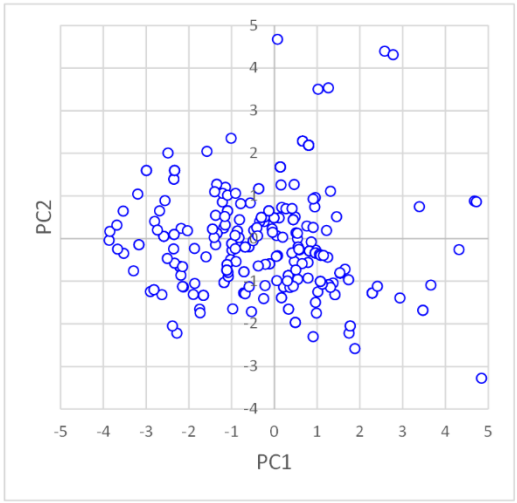
A)


nHDon
TPSA
nHAcc
PHI
MW
nC
NAR
PC1
PC2

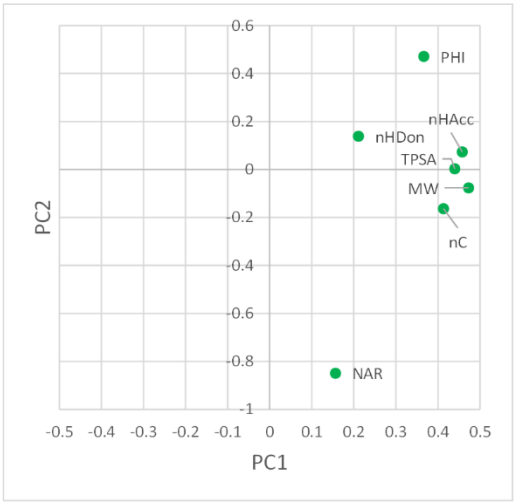

PC1
PC2

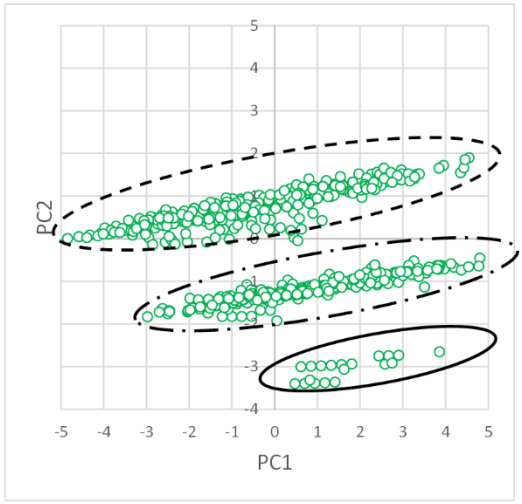
B)


nHDon
TPSA
nHAcc
MW
PHI
NAR
nC
PC1
PC2


PC1
PC2

C)


PHI
nHAcc
nHDon
TPSA
MW
nC
NAR
PC1
PC2


PC1
PC2

D)

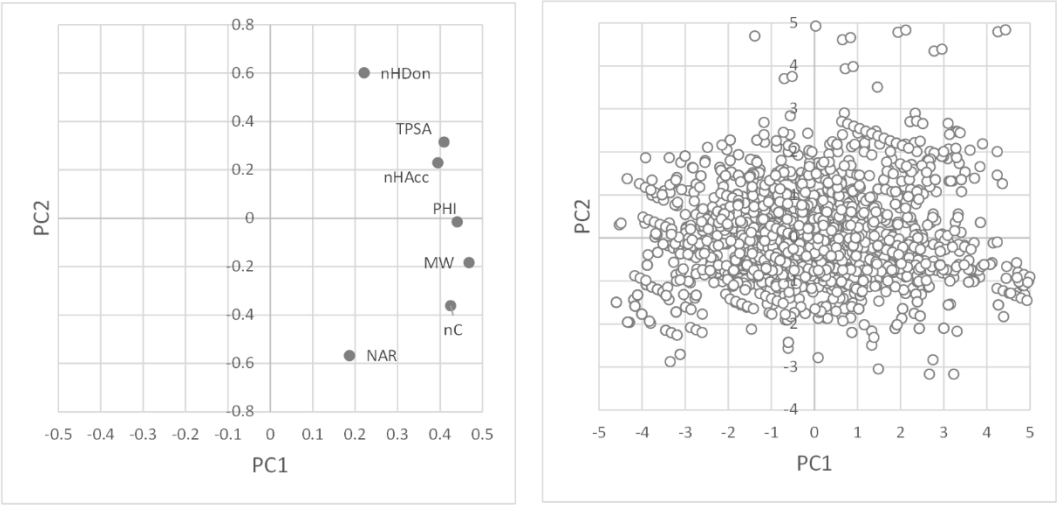



**Figure 17.** PCA. A) E3 ligands (PC1 = 59.1%, PC2 = 25.0%), B) warhead (PC1 = 56.9%, PC2 = 24.2%), C) linkers (PC1 = 59.9%, PC2 = 17.4%) and D) PROTACs (PC1 = 59.7%, PC2 = 17.4%). HBA and HBD were named as nHAcc and nHDon in the original publication, respectively. Figure adapted from Ermondi *et al.,* 2021.[180]

Another important feature to assess is the property change when considering individual building blocks compared to the full PROTAC (Figure S8). As expected, the sum of the MW medians of the building blocks is higher than the PROTAC median. Very often linkers and E3 ligands display terminal carboxylic and aminic groups in their structures which are converted into amides in the final PROTAC. This reflects that building blocks lose several groups when forming the corresponding PROTAC. The same occurs for nC, while NAR is the exact sum of the three parts. Interestingly, since warheads require aromatic moieties to bind the target, NAR is higher for warheads than for E3 ligands. Additivity is not respected for TPSA because of the elimination of polar groups in the reaction step that joins the two or three parts. In general, polarity (TPSA, HBD and HBA) is equally provided by the warhead and the E3 ligand. Finally, PHI is mostly provided by the linker, which determines the flexibility of the entire PROTAC.

C. <u>The 2D chemical space of PROTACs</u>

To obtain orally available drug candidates in the early stages of drug discovery, the comparison of simple 2D descriptors with reference compounds is the first step. However, outside the Ro5 space, in this case for PROTACs, the way to go is the comparison (evaluation of similarities and differences) with approved oral bRo5 drugs. However, this approach is generally limited for two reasons: the lack of available information on reference bRo5 drugs, and the neglect of the 3D effect at this early stage. Though, the availability of several oral bRo5 drugs allowed a reasonable comparison with the data set of PROTACs. For this purpose, we used a dataset of 52 orally approved bRo5 compounds.[39,185]

By simply analyzing the set of 2D descriptors, it can be seen that PROTACs are larger (MW and nC), more flexible (PHI), more polar (TPSA and HBA) and have a higher NAR (Figure 18A). HBDs, on the other hand, do not vary as much and are equally within the Ro5 guidelines (≤ 5).

Moreover, to perform comparisons within the data sets, the three descriptors that represented the biggest differences were selected to build a 3D plot or chemical space: nC (size), PHI (flexibility) and TPSA (polarity) (Figure 18B). Importantly, PROTACs share a small region with oral bRo5 drugs, which indicates that a small proportion of the entire PROTAC ensemble might have the possibility to be orally absorbed (figure 18B). We also observed that linkers (yellow), warheads (blue) and E3 ligands (red) belong to the Ro5 space but do not share the same region. Warheads and E3 ligands were mostly superimposable and partially overlapped with the zone of oral bRo5 candidates (orange). This means that E3 ligands and warheads already have decent molecular property profiles that become complicated when included as part of PROTACs (grey).

A)

| Name | MW | nC | NAR | PHI | nHDon | nHAcc | TPSA |
|---|---|---|---|---|---|---|---|
| Mean bRo5 | 765 | 39 | 2 | 14 | 3 | 13 | 177 |
| Mean PROTAC | 984 | 50 | 5 | 18 | 4 | 18 | 242 |

B)

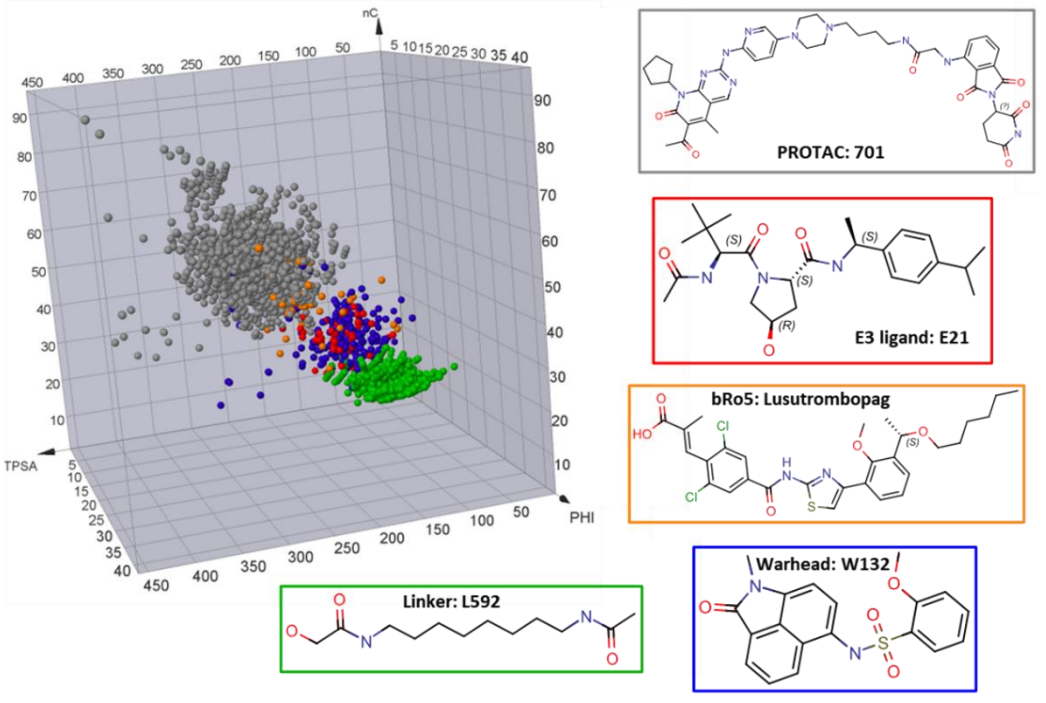


**Figure 18.** PROTACs vs oral bRo5. A) Mean value of the calculated descriptors; B) 3D plot: nC, TPSA, and PHI. Color codes are: PROTACs = grey, bRo5 = orange, linkers = green, E3 ligands = red, and warheads = blue. A representative structure for each class is shown for comparison. MW is expressed as dalton (Da) and TPSA as square angstrom (Å²). HBA and HBD were named as nHAcc and nHDon in the original publication, respectively. Figure adapted from Ermondi *et al.,* 2021.[180]

Thus, this method requires experimental data (permeability and/or bioavailability) to be useful.[126] Currently, only a few commercial PROTACs with reliable and comparable permeability[112] or bioavailability[186] are available.[187]

To provide the first case study, a preliminary example was performed using three PROTACs (Figure 19A, chemical structures in Figure S9) with consistent *in vitro* (Caco-2) permeability values:[112] BI-3663 (log $P_{app}$ = 4.8, high), ACBI1 (log $P_{app}$ = 5.7, medium) and MZ1 (log $P_{app}$ = 7.5, low). A simple view of their molecular properties reveals how MZ1 is larger (MW), more polar (TPSA), and more flexible (PHI), which explains its lower permeability.[48] Moreover, when plotted as part of the chemical space (Figure 19B and C), MZ1 is located further away from the oral bRo5 region and from ACBI1 and BI-3663.

A)

| Name | MW | nC | NAR | PHI | nHDon | nHAcc | TPSA |
|---|---|---|---|---|---|---|---|
| MZ1 | 1002.77 | 49 | 5 | 18.074 | 4 | 16 | 267.97 |
| ACBI1 | 936.23 | 49 | 5 | 15.039 | 6 | 17 | 236.84 |
| BI-3663 | 917.93 | 44 | 4 | 15.563 | 4 | 22 | 244.47 |

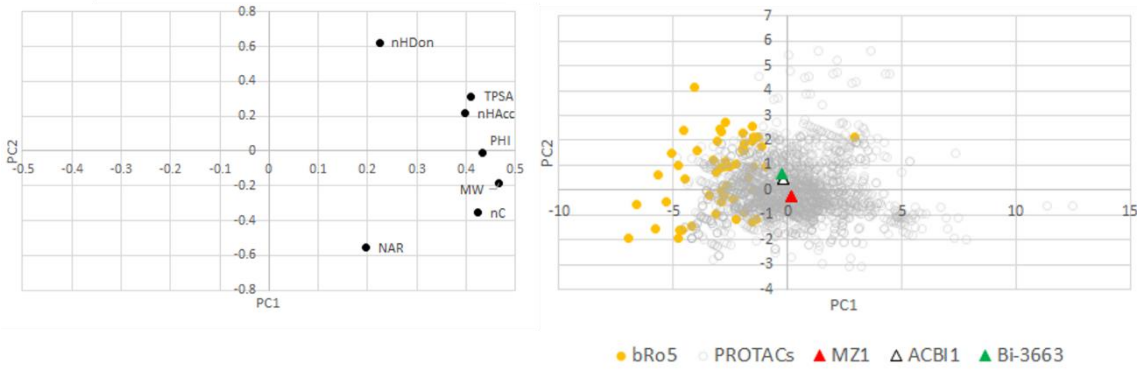



**Figure 19.** A) Mean descriptor values for three PROTAC with low (MZ1), average (ACBI1) and high (BI-3663) Caco-2 log $P_{app}$.[112] B) loading plot and C) score plot; color code: PROTACs=grey; bRo5=orange; MZ1=red; ACBI1=white and BI-3663 in green. MW is expressed as dalton (Da) and TPSA as square angstrom (Å²). HBA and HBD were named as nHAcc and nHDon in the original publication, respectively. Figure adapted from Ermondi *et al.,* 2021.[180]

As recalled in the introduction, the oral bioavailability of bRo5 compounds depends on their structural conformations in specific environments (i.e., 3D PSA). However, the application of 3D strategies is slow and can only be applied to a few compounds. Therefore, another use of the PROTAC chemical space is to individualize regions with similar 2D properties. At this point, the application of 3D strategies can be used to explain property differences between more similar compounds, named as “property cliff”.

D. Paper II conclusions

Maple and Edmonson were the first to describe the molecular properties of PROTACs. However, the high number of new publications and availability of databases required an update and extension of their work. The PROTACpedia and more importantly, the PROTAC-DB contain more than 1500 PROTACs and building blocks with structural information, providing an optimal opportunity to study the molecular properties of PROTACs and build the first PROTAC chemical space.

7 descriptors were selected to breakdown PROTAC complexity: three nonpolar (MW, nC, and NAR), three polar (TPSA, HBA, and HBD), and one flexibility (PHI) descriptor. Notably, cLogP and NRotB should be avoided in this type of study. Moreover, several analyses using PCA and statistical tests have revealed remarkable differences between building blocks. The linker is the most flexible building block, and their variability is dominated by the NAR in the structure. This reflects the importance of linker design for optimizing the DMPK properties of PROTACs. On the other hand, the warhead is the PROTAC substructure that includes more aromatic rings, presumably to improve binding.

Furthermore, a PROTAC chemical space was built by plotting size (nC), polarity (TPSA) and flexibility (PHI) descriptors. This allowed us to visually identify a region of oral bRo5 drugs and compare it to the space occupied by PROTACs. PROTACs are generally bigger, more polar and more flexible, which can compromise permeability and oral bioavailability. However, they share a subregion of the chemical space, suggesting that suitable property design strategies can make PROTACs oral. Moreover, a preliminary example was used to demonstrate how the use of the PROTAC chemical space can help in the earliest stage of drug discovery. However, we acknowledge that this application of the chemical space is retrospective and needs reference compounds to work.

## 7.3 Are we Ready to Design Oral PROTACs? (Paper III)[20]

In the previous article, we discussed the structural differences between the PROTACs and their building blocks included in the PROTAC-DB.[180] 7 2D descriptors were selected to reveal their property differences and a PROTAC chemical space was built upon size (nC), polarity (TPSA) and flexibility (PHI) descriptors, which allowed PROTACs and their building blocks to be compare with oral bRo5 drugs. Finally, a preliminary example showed that the identification of oral bioavailability zones may be feasible once more experimental data are available.

However, other aspects of the PROTAC database were missing, such as comprehensive analyses of the E3 ligases and POIs used. In addition, the database was updated to include 2258 degraders. Among these new additions, several have been experimentally tested for permeability, including ARV-110 and ARV-471, the first oral PROTACs to enter clinical trials. For these reasons, we conducted a complementary study to complete, update, and refine the PROTAC 2D chemical space.[20]

In our recent PROTAC-DB analysis,[180] we focused only on the chemical properties of PROTACs and building blocks but did not provide information on the pharmacodynamics of PROTACs (out of the scope of this thesis).[20] In this article, an analysis of the E3 ligases suggested that despite the progress made in the past years, the selection of E3 ligases is still limited, which makes the design of E3 ligands quite restrictive (data not shown).[188–190] In fact, most PROTAC entries use the von Hippel-Lindau (VHL) or CRBN ligase. Moreover, the POIs targeted by the PROTACs in the database are very diverse, though mostly involved in cancer processes (data not shown).

### A. The 2D PROTAC chemical space 2.0.

In our previous publication the chemical space of 1662 PROTACs was created and compared to their building blocks and a set of oral bRo5 drugs. The update of the PROTAC-DB to 2258 degraders allowed us to expand the covered area and include new PROTACs with available experimental properties. By applying the same descriptors (MW, nC, NAR, TPSA, HBA, HBD and PHI) a newer molecular property space was generated. Overall, these PROTACs had an average property distribution of: MW= 972.9 Da, nC= 49.5, NAR= 4.5, HBD= 4.5, HBA= 17.7, TPSA= 240 $Å^2$ and PHI= 17.3) (Figure 20A). Compared to the initial version (Figure 18, Paper II), the new additions seem to have similar properties. Moreover, a direct comparison with the set of 50 oral bRo5 drugs, suggested that PROTACs are, as expected, bigger, more polar, more flexible and more lipophilic (Figure 20). Consequently, from a simple 2D descriptor consideration, most PROTACs will hardly become oral candidates.

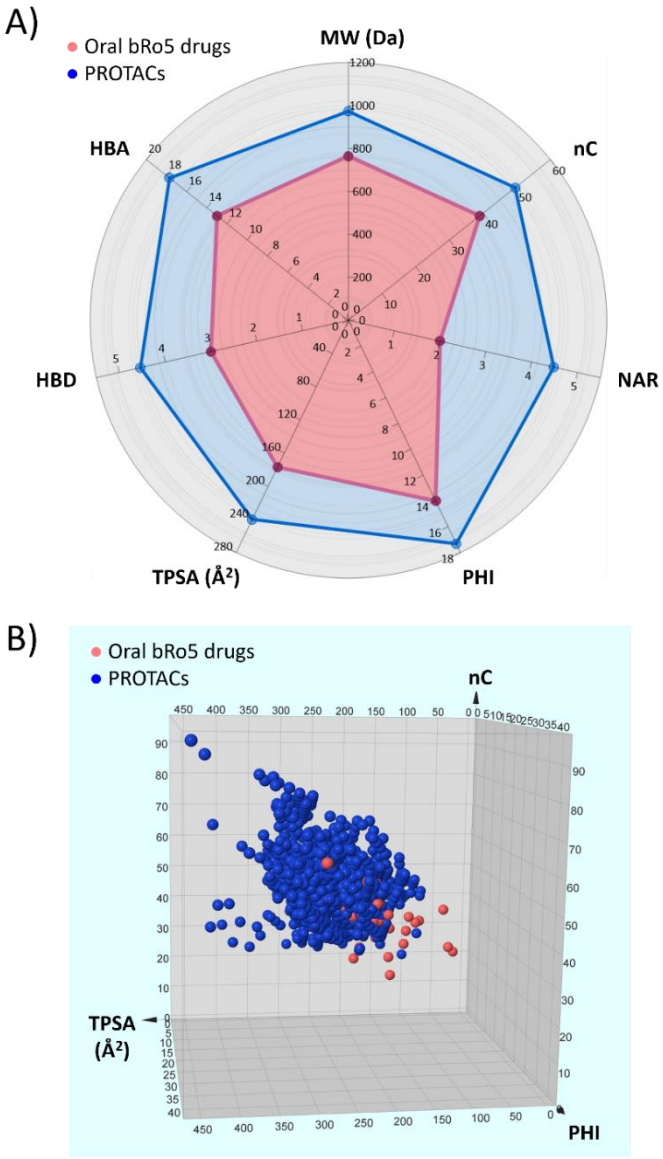



**Figure 20.** A) Means of molecular descriptor radar plot and B) 3D chemical space of the 2258 PROTACs in the PROTAC-DB (pink) and the 50 oral bRo5 drugs (blue). Figure adapted from Garcia Jimenez *et al.,* 2021.[20]

As described earlier, they mostly target VHL and CRBN ligases. Furthermore, the targeting of an E3 ligase is performed by similar E3 ligands with common properties. This has an impact on the molecular properties of the entire PROTAC. Thus Poongavanam *et al.,* recently showed how CRBN-based PROTACs are closer to the Ro5 space (smaller, less polar and less flexible) than VHL and other E3 ligase-based PROTACs, therefore having more possibilities to be absorbed orally.[191]

B. Permeability and bioavailability subregions in the PROTAC chemical space

As recalled earlier, the identification of regions with privileged DMPK properties is crucial to set simple descriptor guidelines. Therefore, several examples with bioavailability and/or permeability were studied. By 2021, ARV-110 and ARV-471 were the only oral PROTACs with disclosed structures in clinical trials (Figure S10, Table S5). They both use a CRBN-based E3 ligand, which places them in the low edge of the PROTAC chemical space (Figure 21). In fact, they are both smaller (nC= 41 and 45, respectively), less flexible (PHI= 12 and 10, respectively) and less polar (182.86 $Å^2$ and 96.43 $Å^2$, respectively) than most PROTACs. In addition, they are superimposable on the oral bRo5 space. Furthermore, the rigid nature of their linker enormously reduces their flexibility. This, at least in theory, prevents them from displaying a high chameleonicity. In their case, however, chameleonic properties are not needed since they already sit in a privileged area of the chemical space.

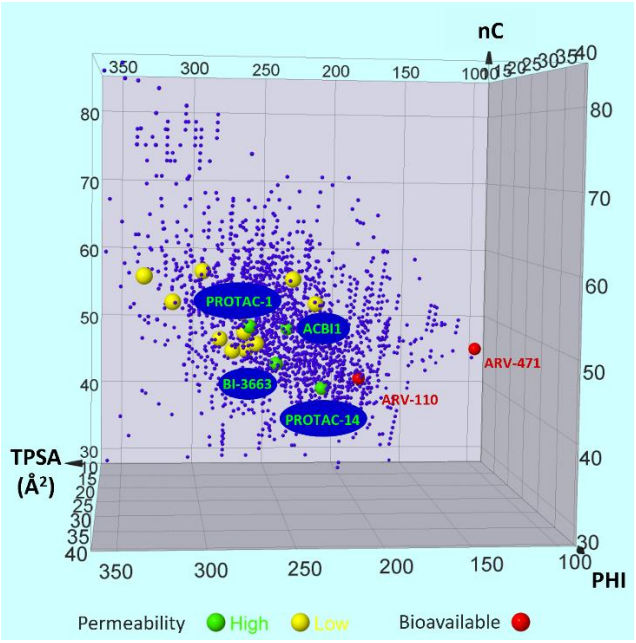



**Figure 21.** Chemical space of a subset of permeable and bioavailable degraders. Figure adapted from Garcia Jimenez *et al.,* 2021.[20]

Moreover, several PROTACs with available permeability data (PAMPA and Caco-2) were retrieved and placed in the chemical space (Table S5). PAMPA and Caco-2 are the two most common methods to assess passive permeability, despite their methodological issues (see Chapter 5.7). Consequently, only PROTACs with consistent permeability values were considered. Despite the variability of the methods, the overall classification should be comparable. Thus, PROTACs were classified as permeable ($P_{app} > 1 \times 10^{-6}$ cm/s; green dots) or low permeable ($P_{app} < 1 \times 10^{-6}$ cm/s; yellow dots) (Figure 21) (structures are found in Figure S10).

Permeable degraders (PROTAC-1, PROTAC-14, ACBI1 and BI-3663) occupy a centered area in the chemical space. In particular, PROTAC-14[192] is placed very close to ARV-110 as it has the lowest nC, TPSA and PHI. In addition, BI-3663[112] is the next closest PROTAC. ACBI1[112] and PROTAC-1[70], on the other hand, are placed in a more distant region. Interestingly, the last two PROTACs are VHL-based whereas the first two are CRBN-based. This confirms the fact that CRBN-based PROTACs are generally easier to be absorbed orally. In addition, VHL-based PROTACs do not sit too far out, which makes CRBN and VHL partially superimposable from a property perspective. Finally, PROTAC-1 occupies the farthest permeability region, displaying high polarity (TPSA) which normally challenges permeability.[55] Nevertheless, this compound has been proved as a molecular chameleon able to partially hide its polarity in nonpolar environments. Thus, PROTAC-1 is the proof of concept of how chameleonicity could help to solve permeability issues for PROTACs.[70]

### C. Paper III conclusions

In this study, we updated, refined and completed the chemical space of PROTACs with two examples of bioavailable PROTACs (ARV-110 and ARV-471) and PROTACs with known permeability values. Overall, the bioavailable PROTACs of our set lie in a region that is partially overlaid with an external data set of 50 oral bRo5 drugs, demonstrating that PROTACs can be orally absorbed when size, polarity and flexibility are controlled. In addition, the permeable PROTACs were located in a central region of the PROTAC chemical space. Among them, the 2 CRBN PROTACs were closer to the orally available PROTACs than the other 2 (VHL-based), proving that CRBN are generally more “drug-like”. Interestingly, PROTAC-1, the first NMR-proven chameleon, was the most distant compound to the oral region.

Thus, to obtain oral PROTACs, their molecular structure should be designed to have a size, polarity and flexibility as similar as possible to those areas in which the most similar orally available compounds are located. Indeed, this represents a classical property-based strategy that works for the Ro5 space and some neighboring regions belonging to the bRo5 space. However, more complex bRo5 compounds (i.e., PROTAC-1) can show chameleonic properties that improve permeability and/or solubility. In fact, this property is a key feature yet to be exploited in the bRo5 space. Finally, while we acknowledge that the relationship of molecular property descriptors with bioavailability is not trivial to assess, we believe that the first step of any drug discovery needs to exploit the use of 2D descriptors to obtain clues on this matter.

## 7.4 Chapter 7 conclusions

*In this chapter, we have created the 2D descriptor-based chemical space of PROTACs and macrocycles in comparison to FDA-approved drugs and identified regions of interest from an ADME perspective. Moreover, we have created 2D descriptor-based models distinguishing oral macrocycles in early drug discovery. As discussed earlier, these approaches are feasible when reference compounds are available (retrospective tools). Nonetheless, they deserve to be included as molecular property optimization strategies.*[66]

## 8. Experimental strategies to monitor molecular properties

*The assessment of experimental molecular properties (lipophilicity, polarity, etc.) is a key step to understand the potential of bRo5 compounds to become oral drugs. To date, most predictive ADME models are based on experimental data obtained for Ro5 compounds*[193] *and thus, the application of these models to bRo5 compounds is partially inaccurate. An additional problem is the difficulty for medicinal chemists to obtain experimental molecular property data in the bRo5 space. Consequently, the experimental monitoring of the molecular properties of bRo5 compounds is the first compulsory step to obtain good quality information upon which ADME models can potentially be built.*

*In this chapter, the main contributions of this thesis to the discovery of ad hoc experimental strategies to monitor the molecular properties of bRo5 compounds are discussed. Special emphasis is placed on solubility and chameleonicity and their relationship with other molecular properties (the most used strategies during this thesis are shown in Figure 22).*

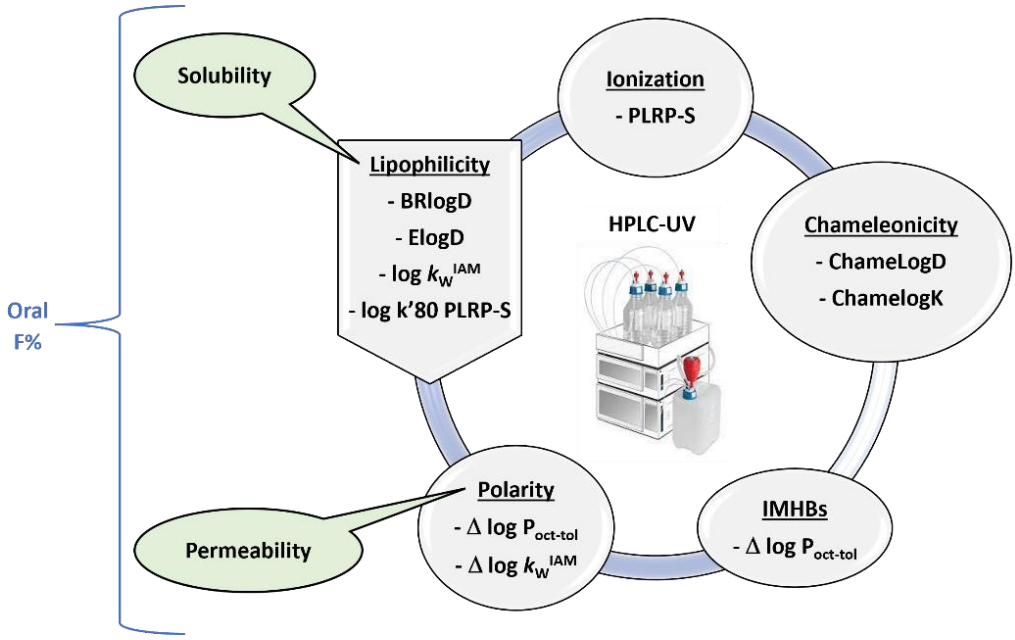



***Figure 22.*** *Oral bioavailability (F%) monitoring from the solubility/permeability balance. Scheme of the used and/or available chromatographic strategies to assess the physicochemical properties (colored in grey) determining crucial in vitro ADME properties (colored in green) of bRo5 compounds.*

## 8.1 Designing Soluble PROTACs: Strategies and Preliminary Guidelines (Paper IV)[63]

### A. Solubility as a DMPK property

There is no clear agreement on the minimal experimental solubility that guarantees oral absorption. Some researchers set > 60 µg/mL as the desired goal for solubility, whereas others at GSK proposed < 30 µM, 30-200 µM and > 200 µM as cut-offs for low, intermediate and high solubility, respectively. Moreover, solubility measurements often suffer from interlaboratory variability which also depends on the method used.[194]

From a mechanistic perspective, the solubility process is complex and can be studied in several steps. Once the active compound is released from the pharmaceutical form (i.e, capsule, tablet), it then dissociates from the crystal structure. This step depends on the crystal packing of each compound (thermodynamically stable or amorphous),[195] which is conditioned by the intermolecular interactions in the solid state or the "brick-dust" effect.[196] In this step, the main determinant is the melting point ($T_m$). For example, in compounds with $T_m$ higher than 200, the crystal effect will have a high final impact on solubility. Moreover, this represents the key step improved by formulation efforts, in which the apparent solubility and dissolution rate of the compound is increased.[197] Next, a cavity must be formed in the solvent to receive the compound. In this step, the MW, which is a rough indicator of the volume, is the key determinant. Finally, the third step involves the insertion of the molecule in the cavity. In this case the hydrophobicity of the molecule determines the degree of interaction with the solvent, monitored by the log P. Thus, from a property-based (QSPR) perspective, the modeling of the intrinsic solubility is expressed as a function of $T_m$ and log P in the General Solubility Equation (GSE), established by Jain and Yalkowsky in 2001.[198] Moreover, other predictive strategies rely on physics-based methods, which model the thermodynamic cycle. However, none of these strategies have been trained on bRo5 compounds and their application to the bRo5 space remains challenging.[196,199]

bRo5 compounds, and PROTACs in particular, lack a property-based strategy for solubility. In fact, the solubility of bRo5 molecules, has only been reported and studied in few cases.[64,200,201] Moreover, no evidence has ever been published for PROTACs. Consequently, we aimed at designing an experimental strategy to determine the thermodynamic solubility of PROTACs and its underlying factors. Then, experimental and computed molecular properties for bRo5 compounds, such as lipophilicity and polarity, were also evaluated with the aim of providing solubility classification models to be used in early drug discovery. In addition, the experimental solubility values were compared to solubility predicted values to examine the potential of available solubility predictors and the relationship between the solubility of the PROTAC structure and that of its building blocks.

B. Experimental thermodynamic solubility (log S)

In the last three articles (Papers I, II and III) the concept of the bRo5 chemical space was developed and several regions of interest were discussed. With this in mind, a subset of commercially available degraders that efficiently represented the entire chemical space of PROTACs was selected. Thus, for the 21 selected degraders, their calculated descriptors showed that they cover a big area of the chemical space: TPSA (166−335 Å$^2$), nC (34−58), and PHI (9−27) (Figure 23). Moreover, the data set was selected to include VHL and CRBN ligands and diverse types of linker families (PEGylated, alkylic, rigid, etc.) and warheads (Table S6). In addition, given the importance of the ionization state for solubility, we included both PROTACs calculated to be charged and neutrals at pH 7 (Table S7). Experimental measurements by potentiometry were not feasible (Sirius T3) due to solubility issues in aqueous environments (water, water-methanol).

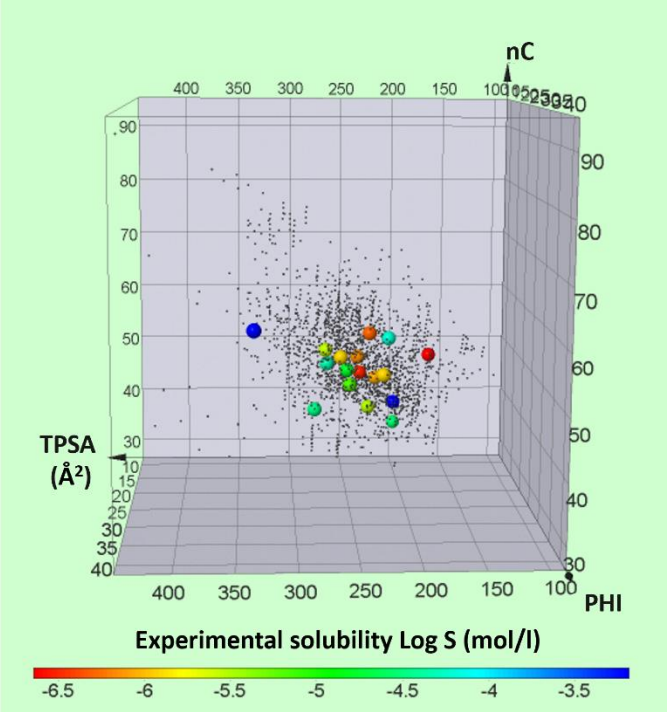



**Figure 23.** Graphical representation of the PROTAC solubility data set based on 2D descriptors (nC, PHI and TPSA). Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

The experiment attempted to measure solubility under reproducible conditions for early stages of drug discovery (see Methods).[63,202] Thus, the well-known shake-flask approach was employed at a pH representative of an average aqueous media in the body (stomach, intestine, blood). Since the chromatographic descriptors were validated at pH 7, the experimental solubility was performed under the same conditions to obtain comparable data (although pH 7.4 is the average in the body). Moreover, the same rationale was followed to select the temperature, 25ºC. In addition, solubility predictors are optimized at this temperature. Finally, PBS-based solvents were selected over biorelevant dissolution media (i.e., FeSSIF) to study the correlation with other

descriptors (computational and experimental) or predictors and drive conclusions to strictly structural reasons. Solubility values were then classified according to the GSK classification into low (< 30 µM), intermediate (30-200 µM) and high solubility (> 200 µM) (Table 4). Notably, the three categories are well populated, although some degraders are poorly soluble. ACBI1, *cis*ACBI1, ARV-825, Mcl1 degrader-1 and MD-224 have solubilities below the quantification limit and thus they were excluded from any quantitative analysis. Moreover, the solubility trends were not explained by their position in the chemical space (nC, PHI, and TPSA) (Figure 23).

**Table 4.** PROTAC Solubility classification based on GSK Guidelines. SD, standard deviations of the solubility measurements; RSD, relative standard deviations with respect to the mean; ND, not detectable. Values expressed as Log S (mol/l) are present in Table S8. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| PROTACs | Solubility average (µM) | SD (µM) | RSD (%) | Classification |
|---|---|---|---|---|
| **BRD9 degrader-1** | 660.69 | 4.64 | 0.70 | **High (> 200 µM)** |
| **CM 11** | 630.96 | 17.71 | 2.86 | |
| **CMP 98** | 331.13 | 1.90 | 0.57 | |
| **THAL-SNS-032** | 52.48 | 1.10 | 2.09 | **Intermediate (30-200 µM)** |
| **VZ185** | 50.12 | 1.43 | 2.84 | |
| ***Cis*MZ1** | 50.12 | 0.34 | 0.66 | |
| **MZ1** | 38.02 | 5.20 | 13.50 | |
| **dBET57** | 30.20 | 0.03 | 0.11 | |
| **BSJ-03-123** | 17.78 | 1.06 | 6.00 | **Low (< 30 µM)** |
| **BI-3663** | 6.92 | 0.09 | 1.26 | |
| **ZXH-3-26** | 2.95 | 0.21 | 7.28 | |
| **BI-0319** | 2.63 | 0.01 | 0.32 | |
| **CRBN-6-5-5-VHL** | 1.23 | 0.05 | 3.78 | |
| **Gefitinib-based PROTAC 3** | 1.23 | 0.01 | 0.46 | |
| ***Cis*ACBI1** | < 0.87 | ND | ND | |
| **Mcl1 degrader-1** | < 0.87 | ND | ND | |
| **ACBI1** | < 0.72 | ND | ND | |
| **BI-4206** | 0.58 | 0.02 | 4.07 | |
| **MZP-54** | 0.51 | 0.01 | 1.87 | |
| **ARV-825** | < 0.32 | ND | ND | |
| **MD-224** | < 0.23 | ND | ND | |

C. Physicochemical determinants of PROTAC solubility

1) Experimental descriptors

The prediction of solubility by QSPR strategies (GSE) requires the assessment of lipophilicity (log P). As introduced earlier, solubility is expected to have an inverse relationship with lipophilicity. Consequently, a set of chromatographic *ad hoc* lipophilicity chromatographic descriptors were measured for the set of PROTACs (see Methods). BRlogD, provides an average lipophilicity equivalent to the average of the membrane (octanol/water). Notably, it showed a promising inverse correlation with thermodynamic solubility ($R^2$ = 0.67) (Figure 24A). Moreover, log $k_W^{IAM}$, which provides a lipophilicity value at the interface with the phospholipid tails, also showed a strong correlation with solubility ($R^2$ = 0.61) (Figure 24B). Moreover, *ad hoc* experimental descriptors of polarity were also assessed, given the relationship between polarity and solubility in other bRo5 studies.[64] EPSA showed a poor positive linear correlation ($R^2$ = 0.23) (Figure 24C) while $\Delta$ log $k_W^{IAM}$ performed worse ($R^2$ = 0.09) (Figure 24D). However, the 3 most soluble PROTACs (CM 11, CMP 98, and BRD9) had high $\Delta$ log $k_W^{IAM}$ values. No correlation was observed between the two polarity descriptors (data not shown). Thus, the solubility of the PROTAC structure does not seem to correlate with its experimental polarity.

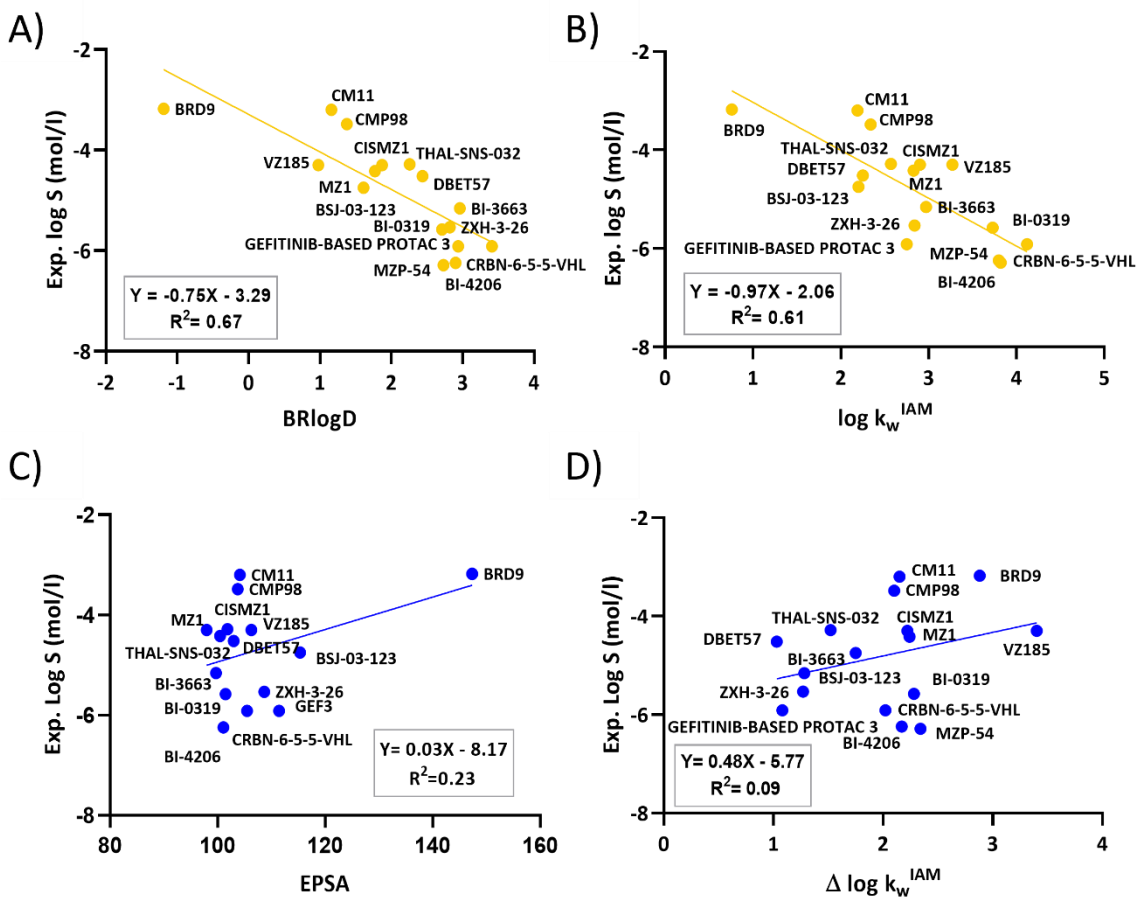
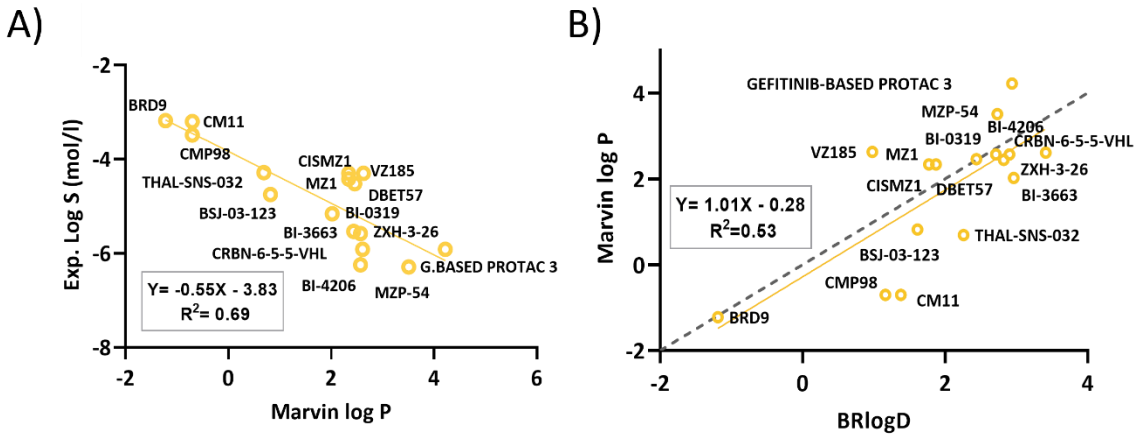


**Figure 24.** Experimental solubility versus A) BRlogD, B) log $k_W^{IAM}$, C) EPSA and D) $\Delta$ log $k_W^{IAM}$ for the PROTAC data set. The yellow color is used for lipophilicity and blue for polarity. The linear trend is present as a colored line. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

2) *In silico* descriptors

Experimental lipophilicity descriptors were highly correlated with experimental solubility. Moreover, it is of high interest to determine the relationship between experimental solubility and computed descriptors, such as the calculated log P. Despite the big limitations in the use of calculated log P descriptors for bRo5 molecules, they are widely used by the medicinal chemistry community and thus they were also studied. Several calculated log Ps were assessed, representing every algorithm available; atom-based, fragment-based, chemical descriptor-based, and 3D-based (Tables S9 and S10, see Methods). Lipinski's MLOGP was revealed as a less appropriate calculator than atom- and fragment-based models. Additionally, adding a third-dimensional component (VolSurf+) had no better effects on the models. Notably, the consideration of the charge (log D) did not outperform (when available) intrinsic lipophilicity (log P). The log P calculator with the best correlation to experimental solubility was Marvin's log P (atom-based) ($R^2$=0.69) (Figure 25A), very similar to BRlogD. BRlogD and Marvin's log P did not fully correlate with each other, indicating that the lipophilicity of bRo5 compounds may depend on more than just the 2D structure ($R^2$ = 0.53) (Figure 25B).

The most popular 2D polarity descriptor, TPSA, was plotted against solubility (Figure 25C) ($R^2$=0.34). TPSA seems weakly related to solubility, but the three most soluble PROTACs have the highest solubility. Compared to the experimental polarity descriptors, TPSA seems to be slightly more correlated. Moreover, TPSA showed an extremely poor correlation with $\Delta$ log $k_W^{IAM}$ ($R^2$ = 0.08) (data not shown), which reveals that the theoretical maximal polarity that a molecule can have (TPSA) is not normally achieved in aqueous solution. Overall, experimental and 2D calculated descriptors suggest that lipophilicity is more important for solubility, although polarity can still have an impact. Moreover, other 2D descriptors like MW were assessed but no correlations were found with solubility (data not shown).



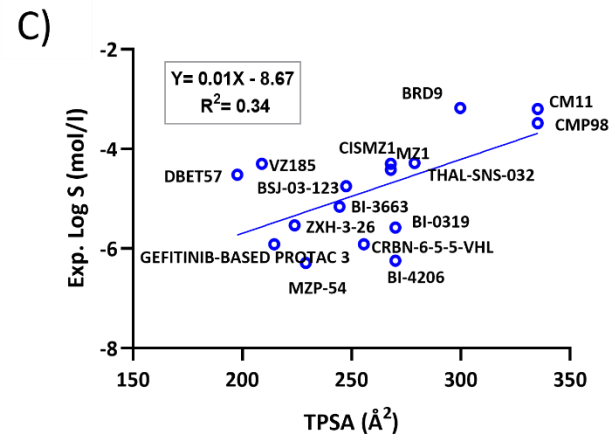



**Figure 25.** Experimental solubility versus A) Marvin log P. B) Correlation between BRlogD and Marvin log P. C) Experimental solubility versus TPSA. The yellow color is used for lipophilicity and blue for polarity. The linear trend is present as a colored line. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

In 2018, Rossi Sebastiano and coworkers observed that 3D descriptors (3D PSA) correlated better with solubility than 2D descriptors (TPSA) for a set of bRo5 compounds.[64] Thus, we examined the effect of the conformation on polarity for our PROTACs. The relationship between 3D PSA and solubility was studied for computationally generated conformations in water of every PROTAC (see Methods). Conformers were generated using conformational sampling and steered molecular dynamics (SMD) (Figure 26), only for the neutral PROTACs (to simplify the use of CS and SMD) (a detailed explanation of these methods is found in Chapter 9). Notably, the removal of the charged PROTACs improved the correlation of TPSA ($R^2$=0.42) compared to the full set (Figure 26).

SMD (blue violins) generated conformers have closer 3D polarities to TPSA value (green dots) in comparison to CS (red violins) generated ones. Statistical analyses showed how the correlation of the lowest, median and highest 3D PSA models with solubility was performing similarly ($R^2$= 0.35-0.40) to TPSA ($R^2$=0.42), both for SMD and CS. This suggests that conformation does not have an impact on solubility. Notably, Atilaw and colleagues observed that in water PROTACs can have different size and polarity conformations.[70] Thus, solubility might be guided by the equilibrium of all the conformers rather than by a single conformer population.

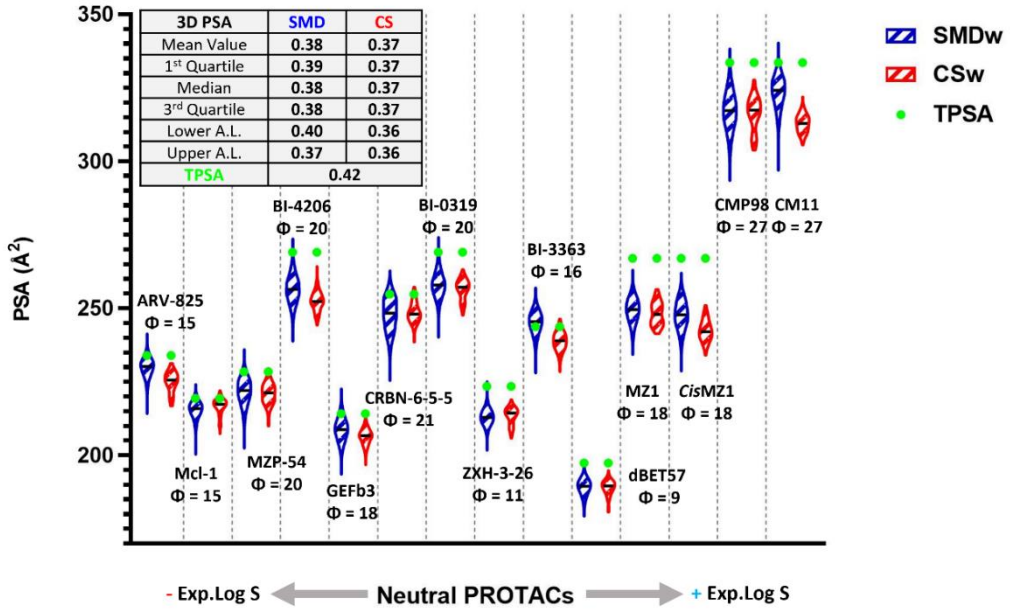



**Figure 26.** Violin plot representation of SMD (blue) versus CS (red) ordered by experimental log S. PHI values (flexibility) are expressed as Φ. Medians are presented as black horizontal lines. $R^2$ are present for every statistical group (although neutral, ARV-825 and Mcl-1 are not considered in the statistical analysis, since accurate solubility values are not available); the lower adjacent value (Lower A.L.) is the smallest value that is equal to or higher than the lower inner fence (first quartile - 1.5 × interquartile range). The upper adjacent value (Upper A.L.) is the highest value that is equal to or smaller than the upper inner fence value (third quartile + 1.5 × interquartile range). Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

D. Solubility predictors

In addition, the performance of online solubility predictors (cLogS) was evaluated (see Methods). As recalled earlier, solubility can be predicted by QSPR strategies (GSE) and by physics-based methods that model the thermodynamic cycle. In this study, we calculated the solubility using several algorithms: atom-based, 3D-based, etc. (Table S11). Notably, moderate correlations were found with intrinsic MarvinSketch ($R^2$ = 0.56), MarvinSketch at pH 7 ($R^2$ = 0.57), and VolSurf ($R^2$ = 0.57) (Figure S11). However, the predictivity is not good enough to be used in bRo5 projects.

E. A solubility decision-making tool

The previous sections have highlighted the importance of polarity and lipophilicity in thermodynamic solubility. Thus, we hypothesized that the simultaneous use of both molecular properties could be used to monitor solubility. Therefore, a 3D plot using two experimental lipophilicity descriptors (BRlogD, and log $k_W^{IAM}$) and one computed polarity descriptor (TPSA) was built. Notably, the distribution of the 16 measurable PROTACs in the 3D plot agreed with the experimental solubility (Figure 27). Moreover, when

classified upon their GSK classification group, results were equally informative except for BSJ-03-123, which was apparently misclassified (Figure S12). Notably, Avdeef and coworkers recently discussed that supersaturation issues can have an impact in solubility differences.[203]

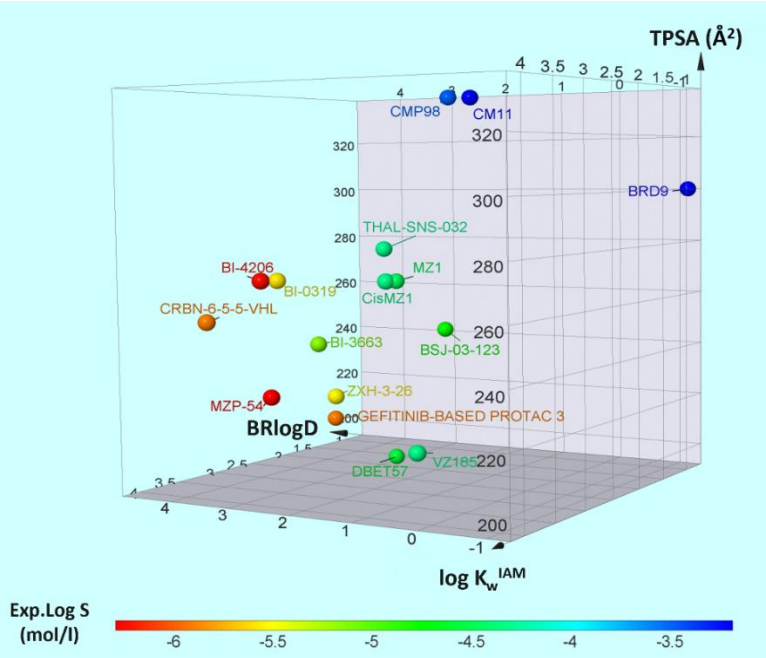



**Figure 27.** PROTAC solubility distribution based on log $k_W^{IAM}$, BRlogD, and TPSA (3D plot) for the 16 PROTACs with quantifiable solubility. Figure adapted from Garcia Jimenez *et al.,* 2022.[63.]

Based on the excellent performance of the 3D plot to model solubility, the next step was to define cut-offs that could distinguish between low and high solubility PROTACs. ML algorithms, such as Random Tree and Random Forest, are supervised methods appropriate for this purpose. Random Tree makes a single prediction guided by the desired outcome, suffering from overfitting. Random Forest, conversely, makes multiple random predictions and then provides the best result based on the majority of the answers. Obviously, we know that the generation of reliable models requires large data sets, but the opportunity to provide a first-ever solubility classification model for PROTACs prompted us to adopt this strategy. Therefore, the 3 descriptors (log $k_W^{IAM}$, BRlogD, and TPSA) and the expected results (GSK classification limits; low, intermediate and high) for the set of PROTACs were used to construct the models (see Methods). Figure S12 suggested that BSJ-03-123 was misclassified in the low solubility group, so we removed it from the models.

The results revealed that 86.7% of the PROTACs were correctly classified for both models (Figure S13). Notably, the low classification group was always correctly predicted (100%), whereas the intermediate and high solubility groups exchanged one PROTAC (80 and 66.7%, respectively). Consequently, the Matthew's Correlation Coefficient (MCC), a quality parameter of ML models, suggested that low solubility classifications

were more performant than intermediate and high solubility predictions (MCC= 1, 0.7 and 0.58, respectively).

In practice, the results of the Random Tree model suggest that PROTACs are directly classified as highly soluble if their TPSA value is equal to or greater than 289.31 Å$^2$. Less polar PROTACs with BRlogD values lower than 2.58 are classified as intermediate, and those with values equal to or higher than 2.58 are classified as poorly soluble (Figure 28). Log $k_W^{IAM}$ was not considered by the models probably due to a partial redundancy with BRlogD, at least for neutral compounds. These results suggest that chromatographic and computed descriptors can be used to indirectly classify the solubility of PROTACs, of particular interest for large pharma companies. In addition, they can also be used to discard low solubility PROTACs, given their perfect classification performance.

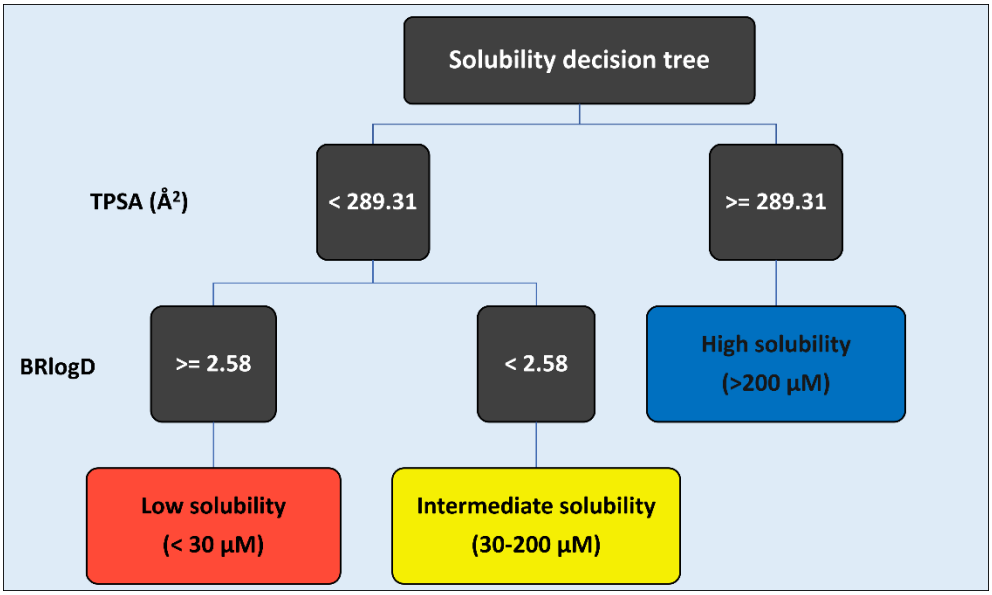



**Figure 28.** PROTAC Random Tree model (BRlogD and TPSA) colored by the experimental classification: low (red), intermediate (yellow), and high (blue) Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

F. Impact of building blocks on PROTAC solubility

PROTACs are synthesized by combining three (warhead, linker, and E3 ligand) or two (E3 ligand bound to the linker and warhead) building blocks. From a medicinal chemistry perspective, it is of high interest to understand how the behavior of a PROTAC depends on its building blocks. Notably, the complexity of the PROTAC structure (bRo5 molecule) provides new properties (i.e., chameleonicity) that single building blocks (normally Ro5 molecules) do not have. Thus, it is expected that building blocks do not entirely explain the behavior of the PROTAC. To examine this hypothesis, 3 PROTAC pairs mainly differing in their warhead, linker and E3 ligand were selected from the data set.

Moreover, their individual building blocks were purchased and their molecular properties measured (Table S12).

We first investigated the role of the warhead. To do so, MZ1[204] and MZP-54[205] were selected (Figure 29). Both are neutral VHL-based degraders with a PEGylated linker and differ only in the warhead (JQ1 and I-BET726, respectively). MZ1 (log S = -4.42) is significantly more soluble than MZP-54 (log S = -6.29). Notably, JQ1 (terminal carboxylic acid) is much more soluble than I-BET726 (terminal carboxylic acid), justifying the difference in the PROTAC solubility. In addition, JQ1 has lower BRlogD and log $k_W^{IAM}$ values and higher TPSA values compared to I-BET726, as expected from the relationship of solubility with lipophilicity and polarity.

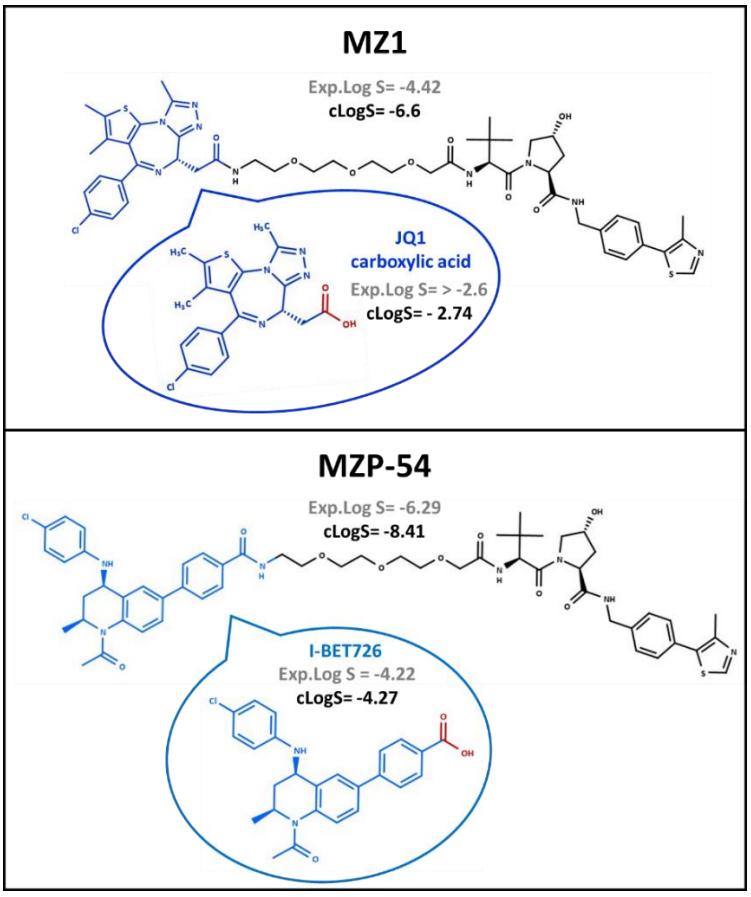



**Figure 29.** Comparison between MZ1 and MZP-54. Experimental (Exp. Log S, S in mol/L) and/or calculated (cLogS, Marvin pH 7) values are presented for PROTACs and their warheads. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

The linker impact was explored by the pair dBET57 and ZXH-3-26.[206] In this case the main difference is in the length of the linker (green dashed circle in Figure 30). dBET57 and ZXH-3-26 contain an alkyl linker, ethylamine and pentyl-1-amine, respectively. Moreover, they have a similar, though not identical, warhead (JQ1 in dBET57 and PROTAC BET-binding moiety 2 in MZP-54). Overall, we observe that dBET57 (log S = -4.52), with the shorter linker, has a higher solubility than ZXH-3-26 (log S = -5.53) by

about 1 log unit. As for the warheads, both have high solubilities although they differ in the attaching point and the presence of the ester, making the comparison challenging. Nevertheless, despite not having identical warheads, the solubility is expected to be due to the difference in the linker. However, other pairs should be evaluated to conclude this fact.

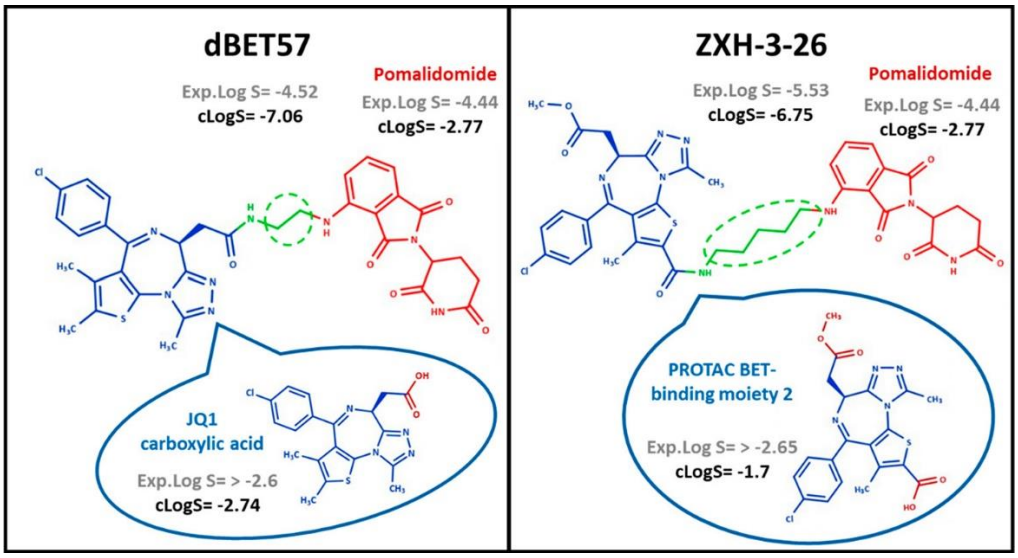



**Figure 30.** Comparison between dBET57 and ZXH-3-26. Experimental and/or calculated log S (S in mol/L) values (Marvin pH 7) are presented for PROTACs and their building blocks. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

Finally, the role of the E3 ligand was evaluated (Figure 31). The PROTAC pair BI-3663-BI-0319/BI-4206 was selected.[207] These are PROTACs based on the BI-4464 warhead, with a pegylated linker, differing mainly in the E3 ligand and its attaching point. BI-3663 is CRBN-based (pomalidomide), whereas BI-3663 and its negative control, BI-4206 are VHL-based (S,R,S-AHPC HCl and S,S,S-AHPC 2HCl). BI-3663 (log S = -5.16) is more soluble than BI-0319/BI-4206 (log S = -5.58 and -6.24, respectively). Surprisingly, pomalidomide itself is less soluble (log S = -4.44) than S,R,S-AHPC HCl and S,S,S-AHPC 2HCl (log S = -2.68 and -2.69, respectively). This contradictory information can be explained by the charged group of the VHL-based ligands. This charge, present in the individual building block at pH 7, is lost when being part of the full PROTAC (verified by potentiometry, see Methods) (data not shown). Pomalidomide, on the other hand, is maintained unionized. Thus, the prediction from the building blocks is not feasible for this case. Overall, in this section we have analyzed the problems in predicting solubility from building blocks.

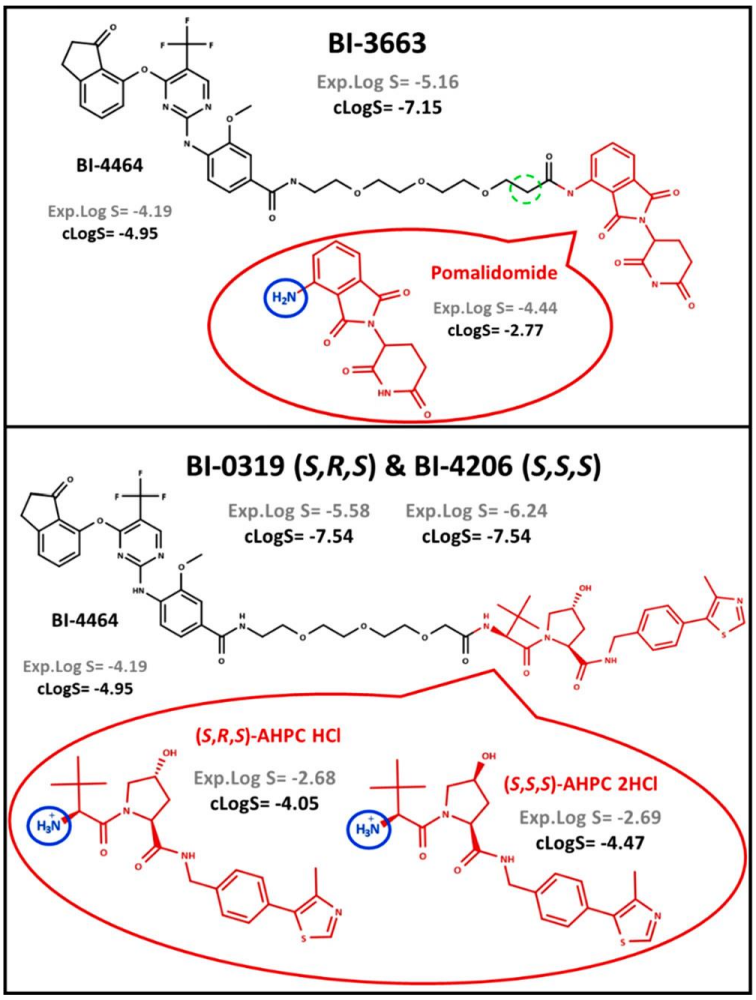



**Figure 31.** Comparison between BI-3663 and BI-0319/BI-4206. Experimental and/or calculated log S (S in mol/L) values (Marvin pH 7) are presented for PROTACs and their building blocks. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

## G. Paper IV conclusions

Solubility is a molecular property that needs to be measured at an early stage in drug discovery projects. However, PROTACs lack structure-property relationships, which makes their design challenging. Therefore, we aimed to fill this gap in order to obtain rational design strategies.

In this manuscript we measured the thermodynamic solubility of 21 commercial PROTACs and assessed their key molecular determinants (lipophilicity and polarity), both experimentally and computationally. Moreover, their relationship with solubility was established. As expected, solubility is strongly inversely correlated with experimental (BRlogD and log $k_W^{IAM}$) and calculated (log P) lipophilicity. Experimental polarity descriptors showed no correlation. However, TPSA showed a small but noticeable effect on polarity. Furthermore, we demonstrated that conventional solubility predictors, which are often used in Ro5 drug development pipelines, are not

accurate enough for the use in PROTAC design. In addition, we verified how the 3D conformation of the structure is not relevant for solubility, probably driven by an equilibrium between conformations. We also explored the limitations of predicting PROTAC solubility directly from its building blocks.

Thus, the use of the 3 most relevant descriptors (log $k_W^{IAM}$, BRlogD, and TPSA) allowed us to create a 3D plot that explained solubility reasonably well. Furthermore, ML strategies (Random Forest and Random Tree) were applied to obtain a classification model based on these parameters. Notably, both models performed very well, especially in distinguishing low solubility PROTACs from intermediate and high solubility PROTACs. Consequently, BRlogD and TPSA values of 2.58 and 289 $Å^2$ were set as the lipophilicity and polarity thresholds, respectively. Thus, in this article we have revealed how the use of calculated and *ad hoc* experimental descriptors for the bRo5 space can help to classify and eventually even predict PROTAC solubility.

## 8.2 Chamelogk: A Chromatographic Chameleonicity Quantifier to Design Orally Bioavailable Beyond-Rule-of-5 Drugs (Paper V)[121]

In the previous article, we presented new insights and QSPR strategies for the design of soluble PROTACs. Solubility, however, is only one of the multiple factors affecting the oral absorption process. In addition, the balance between solubility and permeability is another factor that needs to be addressed.[39] For traditional Ro5 molecules this balance is feasible, unlike the bRo5 space, which requires ad hoc strategies.[66] This issue, however, can be partially compensated by chameleonicity (see Introduction). The concept of chameleonicity, however, includes any environment-driven conformational and/or property change. Thus, many types of intramolecular interactions can generate chameleonic behavior, complicating its study and application (IMHBs, hydrophobic collapse, amide shielding, etc.).[64,71,208]

Since the first examples of chameleonic bRo5 drugs were discovered (i.e., CsA), chameleonicity has been of greater interest in the search of an equilibrium between cell permeability and water solubility. It had its greatest applicability in macrocycles and non-macrocycles (i.e., saquinavir), but it is now being extensively applied to PROTACs. In fact, PROTACs are more flexible and should be, at least in theory, more benefited from chameleonicity.[70,71,106,120,209]

Overall, the available experimental strategies to assess chameleonicity are X-ray, NMR and ChameLogD (see Chapter 5.7). However, these strategies are mostly limited to case-by-case studies, cannot be applied to larger data sets, and cannot explore different environments with a single system. Thus, there is a need not only to improve our understanding of the mechanisms of chameleonicity, but also to develop novel experimental tools to efficiently measure chameleonicity. Consequently, we put our efforts in designing a fast and HT chameleonicity descriptor that can explore different environments using the same chromatographic system.

### A. The chamelogk: method and design

ChameLogD also uses chromatography strategies but using two different systems. Therefore, we aimed at creating a chromatographic descriptor that measures chameleonicity in a unique and dynamic system. Ideally, the system should mimic the passage of the molecule from the outside to the inside of the membrane.

With this goal in mind, we selected the PLRP-S column which provides the most hydrophobic environment after analyzing a pool of commercially available chromatographic columns. In fact, the PLRP-S column, used in combination with a mobile phase composed of 80% MeCN (dielectric constant $\varepsilon \sim 37.5$) and 20% aqueous phase (dielectric constant $\varepsilon \sim 80$), is able to provide an environment similar to the interior of the membrane, often measured as the log P (toluene/water). Therefore, we hypothesized that by varying the proportion of the organic phase in the eluent, the polarity of the environment could change, similar to what would happen in the passage

through the membrane. In previous studies we verified how the behavior of a molecule can change in the different environments provided by this chromatographic system. For instance, saquinavir showed a complete and sudden deviation from the linear trend in the most lipophilic environments (maximized at 100% MeCN). This behavior change, also observed for two PROTACs (MZ1[106,112] and PROTAC-1[210]), suggested the possibility of quantifying it as a chameleonicity descriptor.

To obtain the chameleonicity descriptor, the first step involved the experimental determination of the capacity factor (log k' PLRP-S) at 50, 60, and 70% of MeCN. Then the linear correlation between the log k' PLRP-S and the percentage of MeCN was determined (Figure 32) (Figure S14). The equation derived from the linear correlation (if validated; $R^2$ equal or higher than 0.9) was then applied to extrapolate the value at 100% (named Ext. Log *k'* 100 PLRP-S). Moreover, the experimental value at 100% MeCN was recorded (named Exp. Log k' 100 PLRP-S) and the difference with the extrapolated value was calculated. Thus, Chamelogk was defined as the capacity factor difference (Δ log k') between the experimental log k' measured at 100% MeCN (Exp. Log k' 100) and the extrapolated value (Ext. Log *k'* 100), as shown by equation 3 and Figure 32A.

(3) Chamelogk = Exp. Log k' 100 − Ext. Log *k'*100

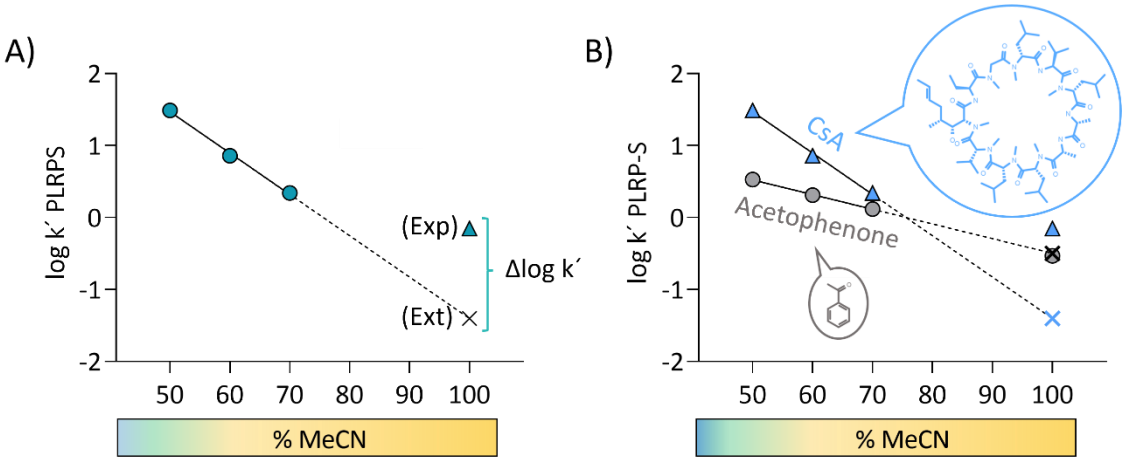



**Figure 32.** A) Graphical scheme of Chamelogk. B) Chamelogk plot of cyclosporin (CsA, neutral bRo5, blue triangles) and acetophenone (neutral Ro5 compound, gray circles). Exp. log k' 100 PLRP-S and Ext. log *k'* 100 PLRP-S values are presented as colored symbols and colored crosses, respectively. Figure adapted from Garcia Jimenez et al., 2023.[121]

In theory, for a non-chameleonic molecule we would not expect a deviation from the linear trend (Figure 32B). This was verified for acetophenone, a rigid Ro5 compound with a Chamelogk value of -0.04. On the contrary, chameleonic compounds should maximize their behavior differences at 100 % MeCN, and thus have a higher retention to the stationary phase compared to the extrapolated value. For example, CsA the most well-known chameleon (see Introduction), has a Chamelogk value of 1.25. Thus, Chamelogk can be used as a strategy to quantify and compare molecular chameleonicity.

B. The dataset

The design of Chamelogk opened up the possibility to measure chameleonicity for large data sets of compounds and gain knowledge on its underlying factors. Thus, 55 commercially available neutral molecules were selected; 25 Ro5 molecules, divided into classic Ro5 molecules and PROTAC building blocks (E3 ligands, warheads and linkers) and 35 bRo5 molecules, including PROTACs, macrocycles, non-macrocycles and some complex PROTAC building blocks. Overall, the data set included a wide representation of the chemical space of small molecules. Cyclic peptides, though, were not included in this study due to their particular behavior in chromatographic systems. Their Chamelogk and $R^2$ values are reported as part of Table 5.

**Table 5.** Chamelogk and $R^2$ values for the neutral data set (n = 55). Entries were ordered sequentially by class, subclass, and Chamelogk. The classification into Ro5 and bRo5 was based on Lipinski's guidelines.[12] The Ro5 class was defined to have just one violation of the following: MW < 500 Da and no more than 5 and 10 HBD and HBA, respectively. Nelfinavir was manually classified as a bRo5 drug despite being formally Ro5 compliant. This was due to the violation of Veber's guidelines and similarity to the bRo5 antiviral series. MW, TPSA and PHI are reported as descriptors of size, polarity, and flexibility, respectively. Table adapted from Garcia Jimenez et al., 2023.[121]

| Compound | Class | Subclass | Chamelogk | $R^2$ | MW | TPSA | PHI |
|---|---|---|---|---|---|---|---|
| 3-Bromoquinoline | Ro5 | Classic Ro5 | 0.13 | 1.00 | 208 | 13 | 2 |
| Acetone | Ro5 | Classic Ro5 | -0.13 | 0.96 | 58 | 17 | 1 |
| Acetophenone | Ro5 | Classic Ro5 | -0.04 | 1.00 | 120 | 17 | 2 |
| Bifonazole | Ro5 | Classic Ro5 | 0.45 | 1.00 | 310 | 18 | 4 |
| Clotrimazole | Ro5 | Classic Ro5 | 0.71 | 0.99 | 345 | 18 | 4 |
| Diazepam | Ro5 | Classic Ro5 | 0.30 | 1.00 | 285 | 33 | 3 |
| Diethylstilbestrol | Ro5 | Classic Ro5 | 0.44 | 1.00 | 268 | 40 | 5 |
| Hydrochlorothiazide | Ro5 | Classic Ro5 | -0.22 | 1.00 | 298 | 135 | 3 |
| Hydrocortisone | Ro5 | Classic Ro5 | 0.10 | 0.91 | 363 | 95 | 4 |
| Naphthalene | Ro5 | Classic Ro5 | 0.02 | 1.00 | 128 | 0 | 1 |
| Phenol | Ro5 | Classic Ro5 | 0.16 | 1.00 | 94 | 20 | 1 |
| Toluene | Ro5 | Classic Ro5 | 0.06 | 0.99 | 92 | 0 | 1 |
| 4-F-thalidomide | Ro5 | E3 Ligand | 0.23 | 1.00 | 276 | 85 | 3 |
| 4-Hydroxy thalidomide | Ro5 | E3 Ligand | -0.16 | 0.84 | 274 | 105 | 3 |
| Cis-OH-VH298 (S,S,S) | Ro5 | E3 Ligand | 0.45 | 1.00 | 540 | 184 | 8 |
| Cis-phenol-VH032 (S,S,S) | Ro5 | E3 Ligand | 0.65 | 0.99 | 489 | 160 | 8 |
| OH-VH298 (S,R,S) | Ro5 | E3 Ligand | 0.48 | 1.00 | 540 | 184 | 8 |
| Phenol-VH032 (S,R,S) | Ro5 | E3 Ligand | 0.37 | 1.00 | 489 | 160 | 8 |
| Pomalidomide | Ro5 | E3 Ligand | 0.00 | 1.00 | 273 | 111 | 3 |
| BI-0115 | Ro5 | Warhead | 0.19 | 0.99 | 288 | 51 | 4 |
| BI-1580 | Ro5 | Warhead | 0.10 | 0.99 | 253 | 51 | 3 |
| CPI203 | Ro5 | Warhead | 0.54 | 0.98 | 400 | 114 | 5 |
| HJB97 | Ro5 | Warhead | 0.15 | 0.99 | 501 | 136 | 6 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| MS-417 | Ro5 | Warhead | 0.29 | 0.97 | 415 | 98 | 5 |
| OTX-015 | Ro5 | Warhead | 0.64 | 0.97 | 492 | 121 | 6 |
| Cyclosporin | bRo5 | Macrocycle | 1.25 | 1.00 | 1203 | 279 | 34 |
| Everolimus | bRo5 | Macrocycle | 0.45 | 1.00 | 958 | 205 | 23 |
| Pimecrolimus | bRo5 | Macrocycle | 0.43 | 0.99 | 811 | 158 | 17 |
| Sirolimus | bRo5 | Macrocycle | 0.25 | 1.00 | 914 | 195 | 21 |
| Temsirolimus | bRo5 | Macrocycle | 0.23 | 1.00 | 1030 | 242 | 24 |
| Atazanavir | bRo5 | Non-macrocycle | 0.33 | 1.00 | 705 | 171 | 15 |
| Nelfinavir | bRo5 | Non-macrocycle | 0.74 | 1.00 | 568 | 127 | 11 |
| Paclitaxel | bRo5 | Non-macrocycle | 0.15 | 1.00 | 854 | 221 | 12 |
| Ritonavir | bRo5 | Non-macrocycle | 0.67 | 1.00 | 721 | 202 | 16 |
| Saquinavir | bRo5 | Non-macrocycle | 1.23 | 1.00 | 671 | 167 | 12 |
| Telaprevir | bRo5 | Non-macrocycle | 0.31 | 1.00 | 680 | 180 | 12 |
| PEG4-PH-NH2-Pomalidomide | bRo5 | E3 Ligand | 0.12 | 1.00 | 541 | 160 | 11 |
| ARV-825 | bRo5 | PROTAC | 0.72 | 0.98 | 924 | 235 | 15 |
| BI-0319 | bRo5 | PROTAC | 0.99 | 1.00 | 1061 | 270 | 20 |
| BI-3663 | bRo5 | PROTAC | 0.50 | 1.00 | 918 | 244 | 16 |
| BI-4206 | bRo5 | PROTAC | 0.70 | 0.99 | 1061 | 270 | 20 |
| BRD4 degrader AT1 | bRo5 | PROTAC | 1.26 | 0.99 | 973 | 266 | 17 |
| CisMZ1 | bRo5 | PROTAC | 1.27 | 0.98 | 1003 | 268 | 18 |
| CRBN-6-5-5-VHL | bRo5 | PROTAC | 1.05 | 1.00 | 972 | 256 | 20 |
| dBET1 | bRo5 | PROTAC | 0.80 | 0.99 | 785 | 224 | 11 |
| dBET57 | bRo5 | PROTAC | 0.68 | 1.00 | 699 | 198 | 9 |
| dBET6 | bRo5 | PROTAC | 0.86 | 0.99 | 841 | 224 | 14 |
| Gefitinib-based PROTAC 3 | bRo5 | PROTAC | 1.36 | 0.98 | 935 | 215 | 18 |
| MZ1 | bRo5 | PROTAC | 1.15 | 0.99 | 1003 | 268 | 18 |
| MZP-54 | bRo5 | PROTAC | 1.18 | 1.00 | 1037 | 229 | 20 |
| PROTAC BET Degrader-10 | bRo5 | PROTAC | 0.83 | 0.99 | 783 | 215 | 11 |
| PROTAC FAK degrader 1 | bRo5 | PROTAC | 0.79 | 0.99 | 996 | 254 | 18 |
| PROTAC Mcl degrader-1 | bRo5 | PROTAC | 0.93 | 0.99 | 910 | 220 | 15 |
| PROTAC-1 | bRo5 | PROTAC | 1.07 | 0.99 | 1034 | 265 | 19 |
| ZXH-3-26 | bRo5 | PROTAC | 0.65 | 0.99 | 785 | 224 | 11 |

To learn about the chameleonicity distribution in our data set, several statistical tests were performed. A simple comparison between Ro5 and bRo5 investigated compounds suggested that bRo5 derivatives are much more chameleonic than Ro5 compliant molecules (median values 0.19 and 0.77, respectively) (Figure 33A). Furthermore, the subdivision of the bRo5 set into its structural subgroups suggested that PROTACs are the most chameleonic group (median value 0.9), followed by non-macrocyclic (median value 0.5) and macrocyclic bRo5 compounds (median value 0.43), when compared to classic Ro5 compounds (median value 0.12) (Figure 33B).

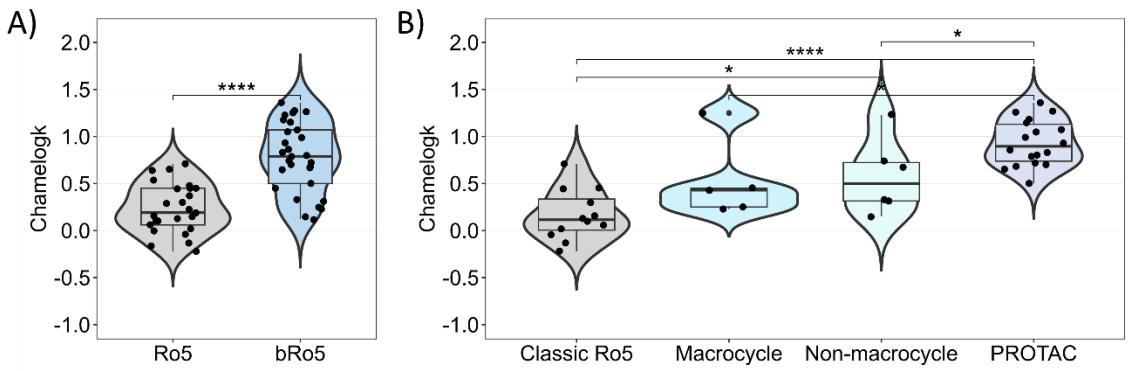



**Figure 33.** A) Chamelogk distribution of Ro5 (n=25) and bRo5 compounds (n=30). B) Chamelogk distribution of bRo5 subclasses (n=29); macrocycle (n=5), non-macrocycle (n=6) and PROTAC (n=18). For comparative purposes, just classical Ro5 were displayed (n=12) (E3 ligands and warheads were removed). Statistical significance is presented as p-value from Wilcoxon's test: 0-0.0001 (****), 0.0001-0.001 (***), 0.001-0.01 (**), 0.01-0.05 (*), 0.05-1 (ns). Figure adapted from Garcia Jimenez et al., 2023.[121]

In addition, we performed a detailed analysis of the most populated class, PROTACs (n= 18). As expected, the higher chameleonicity of PROTACs agrees with their complexity (polarity, flexibility, etc.). Structurally, they are divided into 2 or 3 building blocks, which suggests, as performed in the solubility analysis,[63] to analyze the role of each building block. Thus, the individual building blocks were plotted as a function of their chameleonicity, revealing that warheads and E3 ligands are, as expected, significantly less chameleonic (median values 0.24 and 0.30, respectively) than PROTACs (median value 0.90) due to their simpler structure (Figure 34A). In particular, the structure of the warhead can be very variable, depending mainly on the project. E3 ligands, on the contrary, mostly belong to two classes: Von Hippel-Lindau (VHL) ligands and CRBN ligands.[20] Thus, the E3 ligands and PROTACs were classified upon their E3 ligand and their chameleonicity differences were compared (Figure 34B). Apparently, the bigger structure of VHL ligands not only makes them more chameleonic than CRBN ligands (median values 0.47 and 0.06, respectively, first panel), but confers higher chameleonicity to their corresponding PROTAC structures (median values 1.16 and 0.8, respectively, second panel).

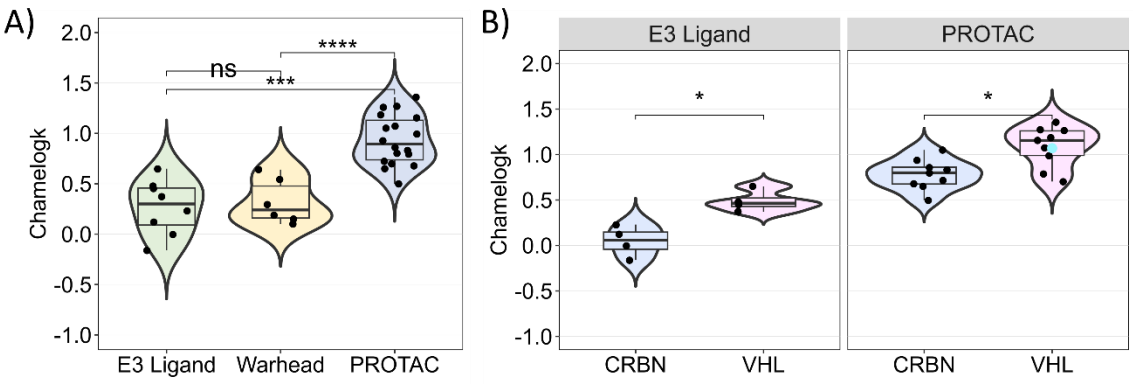


**Figure 34.** A) Chamelogk distribution of E3 ligands (n = 8), warheads (n = 6), and PROTACs (n = 18). (B) Panel 1: Chamelogk distribution of E3 ligands (n = 8) (CRBN, n = 4; VHL, n = 4). Panel 2: PROTACs (n = 18) (CRBN-based, n = 9; VHL-based, n = 9). PROTAC-1 is presented as a light-blue dot. Statistical significance is presented as p values: 0-0.0001 (****), 0.0001-0.001 (***), 0.001-0.01 (**), 0.01-0.05 (*), 0.05-1 (ns) (Wilcoxon's test). Figure adapted from Garcia Jimenez et al., 2023.[121]

## C. Chamelogk interpretation and validation with literature data

The Chamelogk descriptor efficiently classifies chameleonicity for Ro5 and bRo5 molecules. However, to further confirm these trends, Chamelogk needs to be studied from a molecular property perspective that allows to deconvolute the factors that favor chameleonicity. Thus, the relationship between the molecular properties and chameleonicity was studied using the set of 7 computed descriptors reported earlier (see Paper II)[180] by means of PCA (Figure 35). PC1 is mainly driven by MW, TPSA, nC, HBA and PHI while PC2 is guided by NAR, both explaining 90% of the total PCA variance. This fact confirms that molecular complexity (size, polarity, flexibility and hydrophobicity) favors chameleonicity, making it a unique characteristic of the bRo5 chemical space. However, as expected, no trends are revealed within the bRo5 space. Thus, *in silico* studies are mandatory to determine the individual factors between bRo5 subclasses (see Chapter 9).

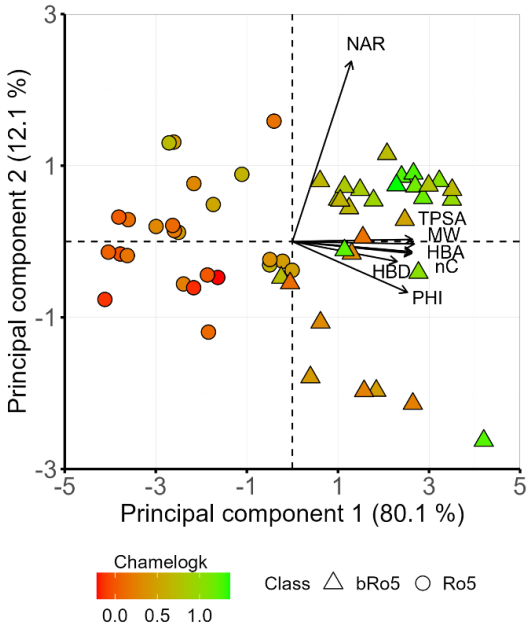



**Figure 35.** Chameleonicity distribution (red-green color scale) according to a 2D descriptor-based PCA. The contributions of individual descriptors to the PCAs are indicated by the length of the arrows. Vertical and horizontal jitter (0.13 PC1 and PC2 units) was introduced to avoid point overlap. Figure adapted from Garcia Jimenez et al., 2023.[121]

Moreover, we aimed to validate our Chamelogk descriptor with well-known chameleons reported in the literature. Overall, chameleonicity had been experimentally assessed by X-ray, NMR and ChameLogD. These techniques have their advantages and limitations (see Chapter 5.7) but fail to provide comparable data other than the presence or absence of chameleonicity. Thus, Chamelogk was compared for its capacity to detect weak and strong chameleons in a list of suspected and/or confirmed chameleons by these strategies (Table 6).

**Table 6.** Chameleonicity assessment of neutral bRo5 compounds classified by bRo5 subtypes and ordered in increasing value of Chamelogk. Table adapted from Garcia Jimenez et al., 2023.[121]

| Compound | Subclass | ChameLogD [95] | Chamelogk [121] | X-ray | | NMR[70,211] |
|---|---|---|---|---|---|---|
| | | | | Δ 3D-PSA[64] | Crystal analysis[114–116,211] | |
| Temsirolimus | Macrocycle | 2 | 0.23 | ND | | |
| Sirolimus | Macrocycle | 1.4 | 0.25 | ND | | |
| Everolimus | Macrocycle | 1.7 | 0.45 | ND | | |
| Cyclosporin | Macrocycle | 2.3 | 1.25 | 79 | Chameleon | Chameleon |
| Paclitaxel | Non-macrocycle | 0.3 | 0.15 | 23 | | |
| Telaprevir | Non-macrocycle | 0.9 | 0.31 | 32 | | |
| Atazanavir | Non-macrocycle | 1.6 | 0.33 | 34 | | |
| Ritonavir | Non-macrocycle | 1.6 | 0.67 | 53 | | |
| Nelfinavir | Non-macrocycle | 1.4 | 0.74 | ND | | |
| Saquinavir | Non-macrocycle | 2.3 | 1.23 | 21 | | |
| PROTAC-1 | PROTAC | ND | 1.07 | ND | | Chameleon |

1) ChameLogD

First, Chamelogk was compared with ChameLogD[95] (Table 6). Both methods show a decent positive correlation (Figure 36) ($R^2$ = 0.48) when all bRo5 compounds are grouped together. Moreover, when macrocycles and non-macrocycles are clustered separately, they show a positive trend ($R^2$= 0.56 and 0.74, respectively). Thus, there are several differences between the methods, although they both agree on the chameleonicity of cyclosporin[114–116,211] and saquinavir.[106] Methodologically, despite being both chromatographic techniques, they are different in nature. Chamelogk, in our opinion, is more suitable as a chameleonicity descriptor because it explores different environments with just one stationary phase. ChameLogD, on the contrary, compares different chromatographic systems (BRlogD and ElogD).

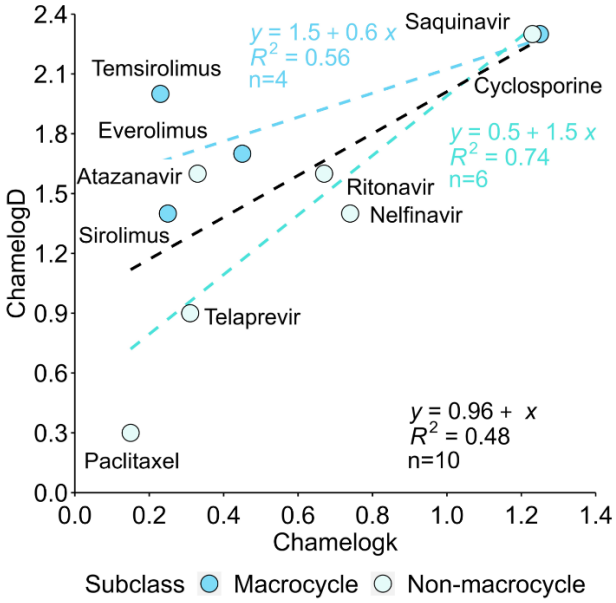



**Figure 36.** ChameLogD vs Chamelogk (n=10). Dashed lines represent the linear regression for the neutral bRo5 compounds (black), macrocycles (blue) and non-macrocyclic bRo5 compounds (light blue). Figure adapted from Garcia Jimenez et al., 2023.[121]

2) X-ray and NMR

Chameleonicity studies through X-ray analysis typically involve comparing crystallized conformations. Generally, a low structural superposition (RMSD, root mean square distance) and a large molecular property window (Δ 3D-PSA) are indicators of chameleonicity.[64,77] Thus, from the work of Rossi Sebastiano,[64] we collected the available Δ 3D-PSA data and compared them with Chamelogk for our compounds. However, a poor correlation was found ($R^2$ = 0.23, Figure S15). Both strategies agreed for cyclosporin but disagreed for saquinavir. Overall, X-ray revealed chameleonicity, such as cyclosporin is always an imperious proof, but the absence of chameleonicity is not equally informative. Since the X-ray analysis is based on the available crystals, it is not possible to confirm that other conformers revealing chameleonicity cannot potentially be seen. Thus, in practice, X-ray is not able to discard non-chameleons. Moreover, NMR combined to NAMFIS provides the most likely conformations that a molecule can adopt in polar or nonpolar solution. In our data set, cyclosporin[211] agrees with NMR, as does PROTAC-1, the first permeable and NMR-proven PROTAC chameleon. However, the limited availability of NMR data makes data set comparison difficult.[70]

D. Chamelogk threshold for chameleonicity

An essential question is what specific value of Chamelogk can effectively differentiate molecular chameleons. With this purpose, a density plot of the available data was performed and the intersection of the Ro5 and bRo5 areas (Chamelogk = 0.6) was found (Figure 37A). Moreover, the application of this threshold to the bRo5 subclass division discarded most macrocycles and non-macrocycles (Figure 37B). It is obvious that this

threshold selection was based solely on a population distribution and needs to be re-defined based on its utility for DMPK purposes. However, it can be used for the first screening tests.

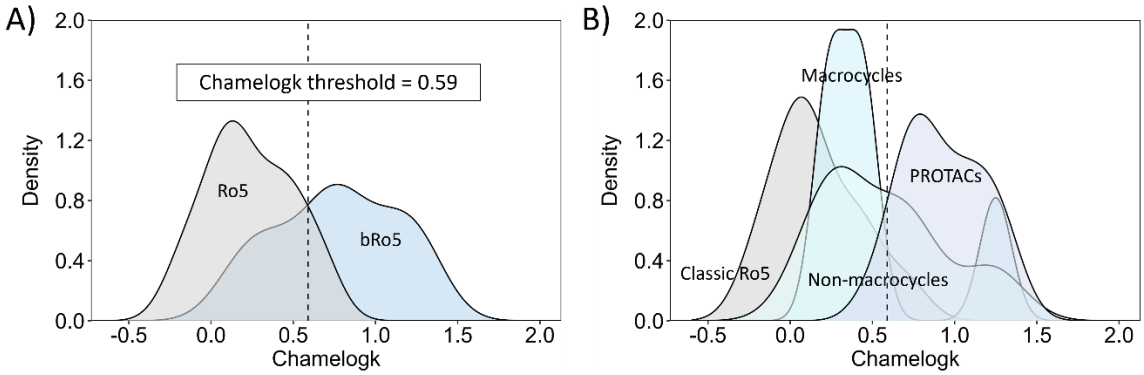



**Figure 37.** A) Chamelogk density distribution for neutral Ro5 (n=25) and bRo5 compounds (n=30). B) Chamelogk density distribution for bRo5 subclasses (n=29); macrocycle (n=5), non-macrocycle (n=6) and PROTAC (n=18). For comparative purposes, just classical Ro5 were displayed (n=12) (E3 ligands and warheads were removed). Color coding is maintained with respect to Figure 33. Figure adapted from Garcia Jimenez et al., 2023.[121]

### E. Chameleonicity in practice

The potential of a candidate to become an oral drug depends on several factors (permeability, solubility, metabolic stability, etc.). In fact, as recalled earlier, candidates can be divided into 4 classes (BCS classification)[43,212] according to their permeability and solubility balance. The assessment of these properties, however, often comes along with experimental issues that slow down the drug discovery process.[55,63] For example, kinetic solubility assessment is not representative enough and PAMPA is not suitable for bRo5 compounds. Consequently, the use of chromatographic descriptors is a fast and intelligent alternative to the measurement of these properties. In fact, $\Delta$ log $k_W^{IAM}$ and BRlogD are related to, permeability and solubility, respectively.[63,112] Notably, chameleonicity is not related to any of these descriptors. In our experience, a compound with BRlogD > 5 is too lipophilic, while one with $\Delta$ log $k_W^{IAM}$ > 1.5 is too polar. Consequently, we theorized that chameleonicity may be the key to balancing these two molecular properties. Therefore, the use of 0.6 as a the Chamelogk threshold could help distinguish compounds that can benefit from chameleonicity. To verify this fact, the experimental molecular properties of several FDA-approved chameleons were measured (see Methods) and related to chameleonicity (Figure 38).

The focus was set first on macrocycles. Among the data set, CsA showed low polarity and high lipophilicity, resulting in low solubility. Nevertheless, CsA is orally absorbed and has a Chamelogk of 1.25. On the other hand, pimecrolimus,[213] which is administered as a cream, has similar polarity and lipophilicity properties but is barely chameleonic (Chamelogk = 0.43). Thus, the chameleonic properties of CsA are crucial to compensate for its low solubility (Class II).[214] However, chameleonicity is not always necessary. In

fact, two oral macrocycles, sirolimus and everolimus[39,215] (a pharmacokinetically improved version of sirolimus with better bioavailability), already have balanced polarity and lipophilicity.

| Pair | Compound | Abs. Route | Oral% F Humans | Chamelogk | BRlogD | Δ log $k_W^{IAM}$ | log $k_W^{IAM}$ |
|---|---|---|---|---|---|---|---|
| A | Cyclosporin | Oral | 30 | 1.25 | 5.9 | 0.31 | 4.71 |
| | Pimecrolimus | Cream | ND | 0.43 | 6.52 | -1.62 | 3.35 |
| B | Everolimus | Oral | 20 | 0.45 | 4.86 | 0.71 | 4.15 |
| | Sirolimus | Oral | 14 | 0.25 | 4.97 | 0.42 | 3.96 |

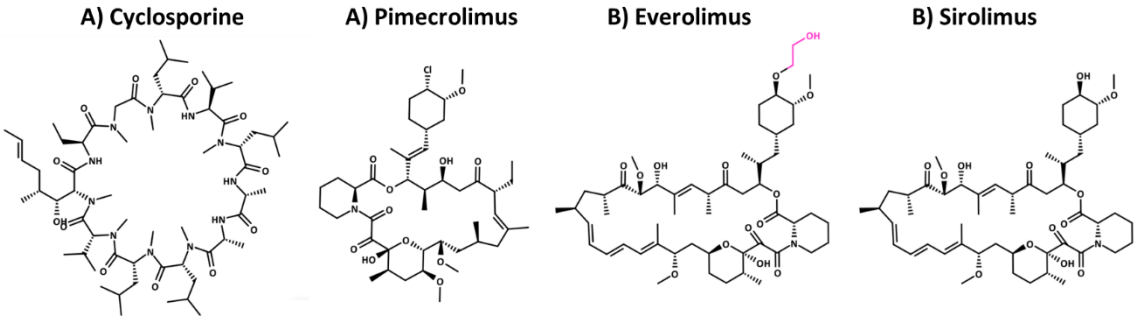



**Figure 38.** Macrocyclic drugs and their molecular properties. Abbreviations: absorption route (Abs. Route). Figure adapted from Garcia Jimenez et al., 2023.[121]

Moreover, several FDA-approved non-macrocycles were confronted (Figure 39). All of them are orally absorbed and show balanced polarity and lipophilicity properties, which makes them orally available despite their low-intermediate chameleonicity. However, saquinavir behaves in a different way; it is very polar (Δ log $k_W^{IAM}$ = 1.85)[106] and not very lipophilic), suffering from cell permeability issues. However, its high chameleonicity (Chamelogk = 1.25) may compensate and guarantee cell permeability (even if partially absorbed by active transport).[106]

| Compound | Absorption route | Chamelogk | BRlogD | Δ log $k_W^{IAM}$ | log $k_W^{IAM}$ |
|---|---|---|---|---|---|
| Telaprevir | Oral | 0.31 | 3.60 | 0.75 | 3.02 |
| Atazanavir | Oral | 0.33 | 3.15 | 0.87 | 2.74 |
| Ritonavir | Oral | 0.67 | 3.29 | 1.02 | 3.02 |
| Nelfinavir | Oral | 0.74 | 4.73 | -0.63 | 2.69 |
| Saquinavir | Oral | 1.23 | 2.85 | 1.85 | 3.44 |

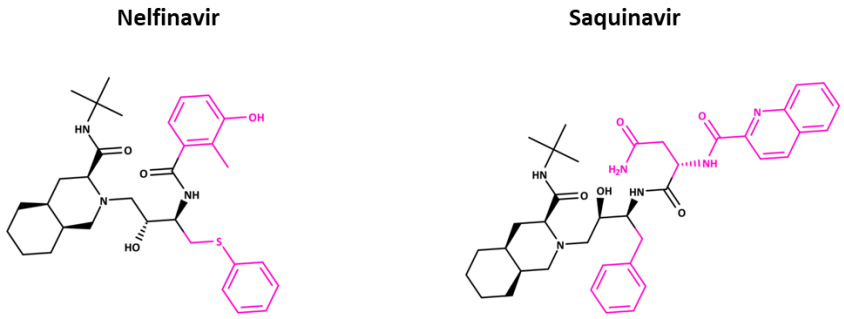



**Figure 39.** Non-macrocyclic oral drugs and their molecular properties. The common atomic scaffold for saquinavir and nelfinavir is colored in black (different chiral centers). Figure adapted from Garcia Jimenez et al., 2023.[121]

PROTACs occupy a different space compared to oral bRo5 drugs.[180] Thus, it is expected that PROTACs behave in a rather different way, given their high flexibility and polarity. For the moment, not much is known about the potential to obtain oral PROTACs by chameleonicity strategies. Only a few studies by Kihlberg's group have reported passive permeability by chameleonicity.[70,71] Our data set showed that PROTACs are extremely polar (median $\Delta$ log $k_W^{IAM}$ = 2) and not so lipophilic (median BRlogD = 2.6). Moreover, all of them are chameleons (median Chamelogk = 0.9). Consequently, most PROTACs will have their major limitation in permeability, although solubility is also an obstacle.[126] Consequently, not many oral PROTAC candidates are in clinical trials.

In our data set, ARV-825 is the only PROTAC proven to be orally available.[216] It is a CRBN-based PROTAC with intermediate lipophilicity (BRlogD = 3.49), low polarity ($\Delta$ log $k_W^{IAM}$ = 1.31) and intermediate chameleonicity (Chamelogk = 0.72) (Figure 40). Their molecular properties are already balanced and therefore ARV-825 will not benefit from chameleonic effects. However, the non-oral PROTACs in our data set, MZ1[204] and its negative control *cis*MZ1, behave differently (Figure 40).[217] These two PROTACs are VHL-based degraders with low lipophilicity (BRlogD= 1.77-1.87) and very high polarity ($\Delta$ log $k_W^{IAM}$ = 2.24-2.20), respectively. Moreover, both PROTACs are strong chameleons (1.15 and 1.27, respectively). Thus, these two compounds have a worse permeability profile than ARV-825 but a better solubility.[63] In this case, despite their high chameleonicity, they cannot compensate for their excessively high polarity, leading to a non-oral PROTAC.

| Compound | Building blocks | Abs. Route | Chamelogk | BRlogD | $\Delta$ log $k_W^{IAM}$ | log $k_W^{IAM}$ | Log S |
|---|---|---|---|---|---|---|---|
| ARV-825 | OTX-015 + PEG + Pomalidomide | Orally active | 0.72 | 3.49 | 1.31 | 3.49 | Low range |
| MZ1 | JQ1 + PEG + VH-032 (S,R,S) | Non-oral | 1.15 | 1.77 | 2.24 | 2.83 | Int. range |
| CisMZ1 | JQ1 + PEG + VH-032 (S,S,S) | Non-oral | 1.27 | 1.87 | 2.20 | 2.90 | Int. range |

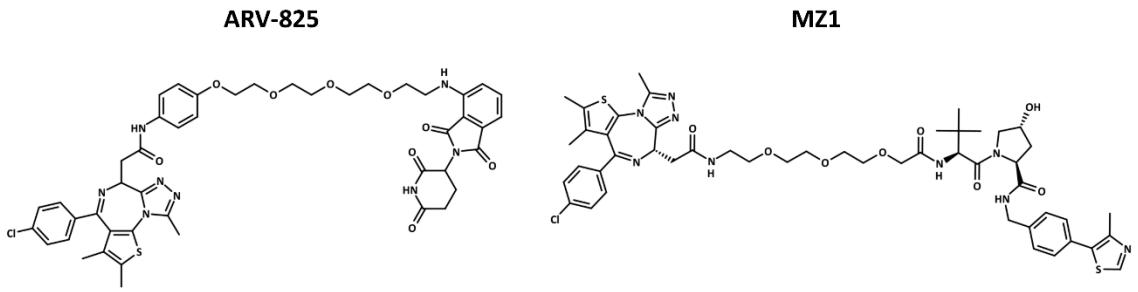



**Figure 40.** PROTACs and their molecular properties. Abbreviations: absorption route (Abs. Route), intermediate (Int.). Figure adapted from Garcia Jimenez et al., 2023.[121]

These results show that chameleonicity can be helpful when lipophilicity and/or polarity are in an affordable range to be corrected. Thus, from a medicinal chemistry point of view, Chamelogk needs to be monitored along BRlogD and $\Delta$ log $k_W^{IAM}$, and studied by means of an experimental property plot (Figure 41). In the interest of prospectively applying this predictive tool, we measured voclosporin (an oral cyclosporin derivative recently approved by the FDA). Notably, voclosporin behaves very similarly to CsA (BRlogD = 5.96, $\Delta$ log $k_W^{IAM}$ = -0.96, Chamelogk = 1.15). Consequently, voclosporin represents a new example of how chameleonicity can correct extreme lipophilicity values and sheds light on the use of chromatographic descriptors in drug design (Figure 41).

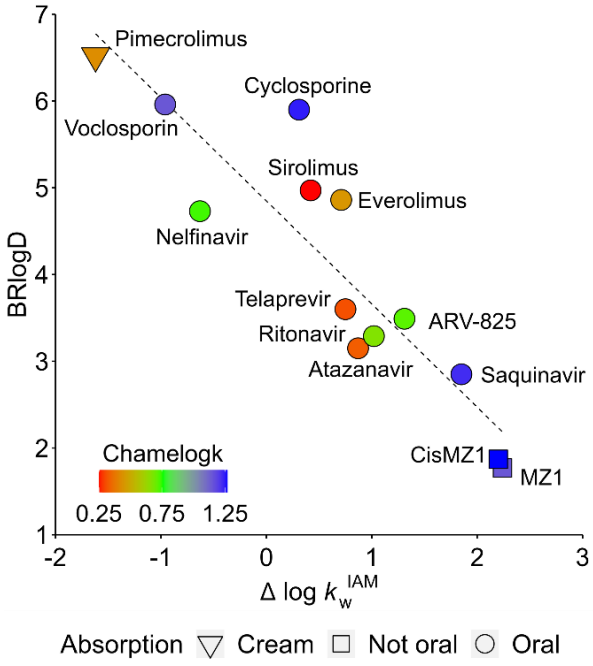



**Figure 41.** Polarity, lipophilicity, and chameleonicity representation for a set of bRo5 compounds. The absorption route is represented by different shapes (if approved). The dashed line represents the ideal linear slope for both variables. Figure adapted from Garcia Jimenez et al., 2023.[121]

F. Paper V conclusions

Chameleonicity is a crucial molecular property in the bRo5 chemical space to achieve simultaneous water solubility and cell permeability. However, the available tools to experimentally measure it have several drawbacks that limit its implementation in drug design. Therefore, in this article we present Chamelogk, a fast chromatographic method capable of quantifying chameleonicity for large sets of compounds with a single stationary phase, the PLRP-S. In essence, it is based on the creation of different polarity environments by modifying the eluent composition. Thus, Chamelogk is able to capture a property change compared to to its original behavior at more polar environments (defined by the linear trend in the PLRP-S system). Finally, we measured 55 Ro5 and bRo5 neutral compounds, revealed several chameleonicity trends, and set a preliminary Chamelogk threshold at 0.6.

The integration of chameleonicity (Chamelogk) with polarity ($\Delta$ log $k_W^{IAM}$) and lipophilicity (BRlogD) allowed to justify the absorption profile (solubility/permeability balance) of several FDA-approved drug candidates. A few selected examples of macrocycles, non-macrocyclic compounds and PROTACs highlighted the importance of balancing polarity and lipophilicity with chameleonicity. Our evidence suggests that cases with extremely high lipophilicity (BRlogD = 6) and polarity ($\Delta$ log $k_W^{IAM}$ = 2) can be balanced. However, our examples also suggest that chameleonicity cannot compensate for higher polarities. We ignore, though, if greater chameleonicity values might be able to solve this situation. Additionally, the measurement and integration of chameleonicity with other molecular properties represents a key tool to prioritize compounds with future as oral candidates. Besides, Chamelogk can also be used to refine and validate computational strategies to predict chameleonicity (Papers VI and VII).

## 8.3 Chapter 8 conclusions

*Overall, in this chapter the major contributions of this thesis to the development of experimental molecular property determination strategies have been exposed.*

*First, a set of chromatographic tools was used to successfully determine the physicochemical properties (mainly polarity and lipophilicity in different systems) of many bRo5 compounds (PROTACs, macrocycles and non-macrocycles). Notably, the experimental measurement of the capacity to form IMHBs was limited by the solubility of the compounds in biphasic systems (octanol/water and toluene/water) and was only applied to a few compounds used to explore in silico strategies (Paper VII). Moreover, a new method to measure chameleonicity was developed and its integration with lipophilicity and polarity allowed to hypothesize a property-based strategy intended to screen compounds with higher potential of being orally available.*

*Finally, in terms of in vitro ADME properties, the thermodynamic solubility of PROTACs was assessed and rationalized upon experimental lipophilicity and calculated polarity descriptors. This allowed us to build a new classification model to be used in the screening of soluble PROTACs, for large datasets of compounds in the academic or industrial world. Permeability on the other hand was measured for a series of PROTACs using the PAMPA method, but no consistent results were obtained (data not shown). In fact, many discrepancies are found in the literature for this method, especially in the bRo5 space.*

## 9. *In silico* 3D descriptor-based strategies to monitor molecular properties

*As discussed in Chapter 7, the modeling of in vitro ADME properties (i.e., solubility, permeability) with simple 2D descriptors (i.e., TPSA, MW, NRotB) provides some pieces of information but has several limitations for bRo5 compounds. Thus, the application of 3D descriptor-based in silico strategies that consider and efficiently model the conformational variability of bRo5 compounds in both polar and nonpolar environments is necessary. Moreover, chameleonicity, which is a property entirely based on a conformational change, can only be computationally studied with 3D descriptor-based strategies.*

*In this chapter, the goal is to explore CS and MD-based tools to generate conformations in polar and nonpolar environments. Moreover, a set of 3D descriptors (IMHB count, $R_{gyr}$ and 3D-PSA) that are expected to be relevant in the modeling of molecular properties are calculated and interpreted for the resulting conformations (general information on the theorical concepts is available in Chapter 5.8). A general scheme of the procedure is shown in Figure 42. Notably, the availability of experimental data obtained from this thesis and/or the literature (Chapter 8) allowed us to obtain a few proofs of concept.*

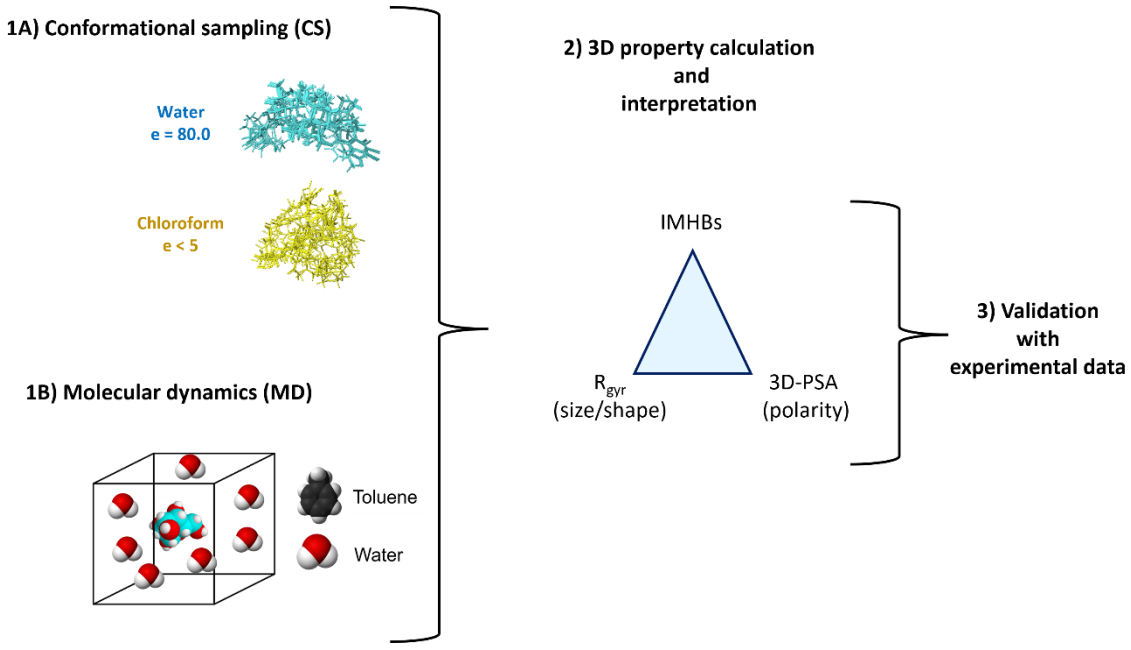



***Figure 42.*** *Schematic workflow of the 3D in silico strategies used in this thesis.*

## 9.1 Conformational Sampling Deciphers the Chameleonic Properties of a VHL-Based Degrader (Paper VI)[210]

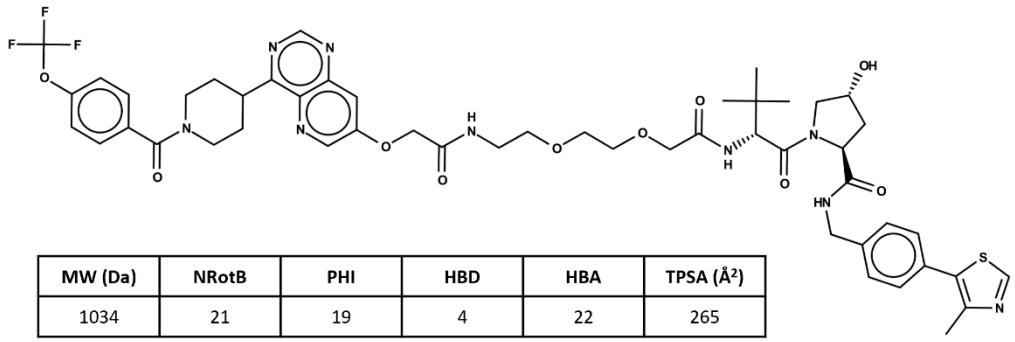
Not much is known yet about the chameleonicity of PROTACs. Since the discovery of the first NMR-proven chameleon (PROTAC-1)[70] by Kihlberg and coworkers, only a few studies on the chameleonicity of PROTACs have been published.[71,121] However, PROTACs are the subclass of bRo5 compounds with the highest flexibility and polarity, key factors that can trigger chameleonicity.[180] Nonetheless, while new experimental methods to measure chameleonicity are emerging, its computational prediction remains limited.

Notably, there is no record of computational tools to predict PROTAC chameleonicity. Thus, the arrival of the first NMR-proven chameleonic PROTAC provided an opportunity to evaluate the use of CS to predict its chameleonicity. To this end, in the following paper a series of CS protocols were evaluated, and the 3D properties of the generated conformers were calculated and compared to the NMR conformations.

### A. PROTAC-1, the first experimentally-verified chameleonic PROTAC

PROTAC-1 is a permeable (-log $P_{app}$ = 5.85 cm/s in PAMPA) and anticancer VHL-based degrader, predicted to be neutral at pH 7.4 (Figure 43). Its conformations in solution were determined by NMR in a mixture of polar solvents (water and DMSO, hereafter referred to as water), mimicking the extracellular regions and in a nonpolar solvent (chloroform), mimicking the interior of the membrane.[70] Eight conformations were obtained in chloroform and ten in water. 90% of the conformations in chloroform were highly folded and 10% semi-folded. On the other hand, the water ensemble consisted of a greater variety, mainly containing semi-folded (56%) but also folded (29%) and linear (15%) conformations.[70] Thus, NMR evidenced, for the first time, a chameleonic PROTAC. Notably, the obtention of solution conformers from NMR requires the application of NAMFIS. This algorithm uses previously generated conformations to match the conformational constraints provided by NMR, providing a set of final conformations in each solvent. For PROTAC-1, the conformational search was based on Monte Carlo algorithms followed by an energy minimization with the generalized-born/surface-area (GB/SA) implicit solvent model.[70] Consequently, NMR-based methods have a computational bias which makes them partially dependent on the method.

Moreover, its chameleonic behavior was also recently verified by Chamelogk (1.07) (Table 5, Figure S16). As shown in paper V, VHL-based PROTACs have a higher tendency than CRBN-based PROTACs to be chameleonic due to the higher complexity of the VHL ligand structure. In this case, PROTAC-1 is more chameleonic than any CRBN PROTAC and most macrocyclic and non-macrocyclic bRo5 compounds (figure 33).[121]



| MW (Da) | NRotB | PHI | HBD | HBA | TPSA (Å²) |
|---|---|---|---|---|---|
| 1034 | 21 | 19 | 4 | 22 | 265 |

**Figure 43.** Chemical structure and 2D descriptors of PROTAC-1. Figure adapted from Ermondi *et al.,* 2023.[210]

B. <u>3D properties of NMR-based conformations</u>

As discussed in Chapter 5.6, a molecular chameleon tends to display folded and less polar conformations in non-polar solvents and more open and polar conformations in polar solvents. Consequently, to evaluate the presence of this behavior in the NMR-based conformations, their 3D molecular properties (3D PSA, $R_{gyr}$ and IMHBs) were calculated and compared (Figure 44A) (see Methods).[106]

The sphericity ($R_{gyr}$) vs polarity (3D PSA) plot suggests that PROTAC-1 behaves as a chameleon, exhibiting more spherical (lower $R_{gyr}$) and less polar (lower 3D PSA) conformations in chloroform compared to water (Figure 44A). Moreover, IMHBs are equally present in both solvents, highlighting that the presence of IMHBs does not necessarily imply chameleonicity (Figure 44B). In particular, the conformations with more spherical shape and lower polarity have 2 IMHBs instead of 1. In fact, even the linear and more open conformations have 1 IMHB. Moreover, none of the conformations reach the TPSA (265 Å$^2$, the maximum polarity the molecule could reach attending to its atoms), suggesting that the molecule is never completely exposed versus the polar solvent.

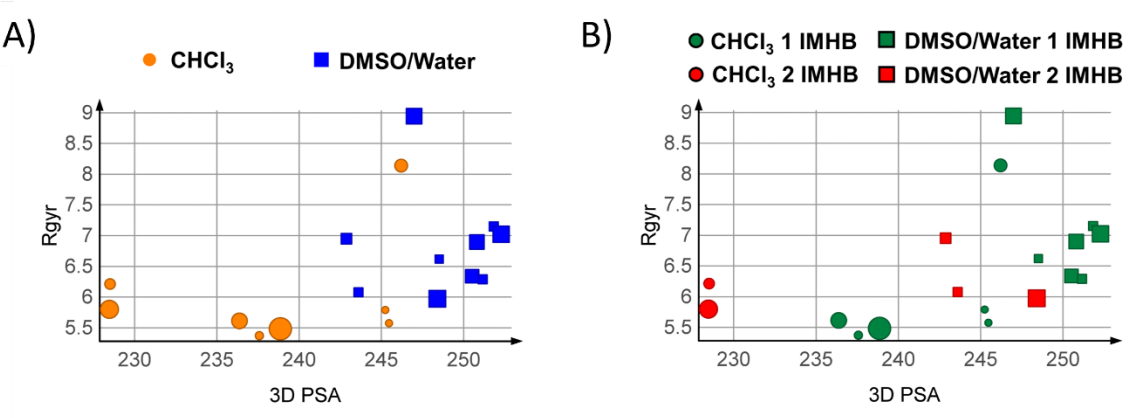


**Figure 44**. $R_{gyr}$ plotted vs 3D PSA for the NMR-derived solution conformations of PROTAC-1 coloured by A) shape by solvent, B) nIMHBs. The marker shape indicates the solvent and the size of the relative population in%. Figure adapted from Ermondi *et al.,* 2023.[210]

C. 3D properties of CS-derived conformations

Since NAMFIS is based on a conformational sampling protocol, CS was preferred over molecular dynamics in this work. CS methods mainly depend on two main factors (see Chapter 5.8); the environment (polar or nonpolar) and the algorithm used. Other relevant factors are the number of maximum steps and the allowed energy range, which limits the final number of generated conformations. Notably, energy-based interpretations, such as the use of the minimum energy conformer (MEC) to model permeability, are known to be risky.[77] Therefore, several CS protocols (Sc01-Sc06) were designed (details are provided in Table S13) (see Methods). Mostly, these CS methods differ in the algorithm, using either Monte Carlo methods (MCMM) (Sc04-06) or combinations of MCMM with low-frequency vibrational modes (LMOD) (Sc01-03).[145] Other modified parameters were the number of generated steps.

The ability of the CS methods to predict the chameleonic behavior of PROTAC-1 and to reproduce the molecular property space of NMR-derived conformations was evaluated. The polarity and sphericity of the obtained conformations were evaluated for the different protocols (Figure 45) using different selection criteria; all conformations (1$^{st}$ column), only those conformers falling within an energy window of less than 3 kcal/mol from the minimum energy conformer (MEC) (2$^{nd}$ column) and this last representation weighted according to their energy (3$^{rd}$ column). Sc03 and Sc06 are not reported since they provided identical ensembles to Sc02 and Sc05, respectively.

Overall, all methods suggest that PROTAC-1 behaves as a chameleon, displaying lower polarity conformations (3D PSA) in the nonpolar solvent (chloroform). Indeed, in all cases there is a shared polarity region between the conformations generated in chloroform and water. Moreover, it seems that MCMM (Sc04-05, Figure 45C and D) finds a smaller polarity difference compared to the mixed models (Sc01-02, Figure 45A and B). In fact, the best polarity overlap with NMR conformations is found with the mixed models.

Notably, in terms of sphericity ($R_{gyr}$), the mixed protocols (Figure 45A and B) display similar results regardless of the number of steps, with water conformers displaying more spherical shapes (lower $R_{gyr}$). Surprisingly, Sc04 (Figure 45C), which uses the MCMM method, shows a similar pattern. Sc05, on the other hand, seems to find water conformations with more extended shapes (higher $R_{gyr}$).

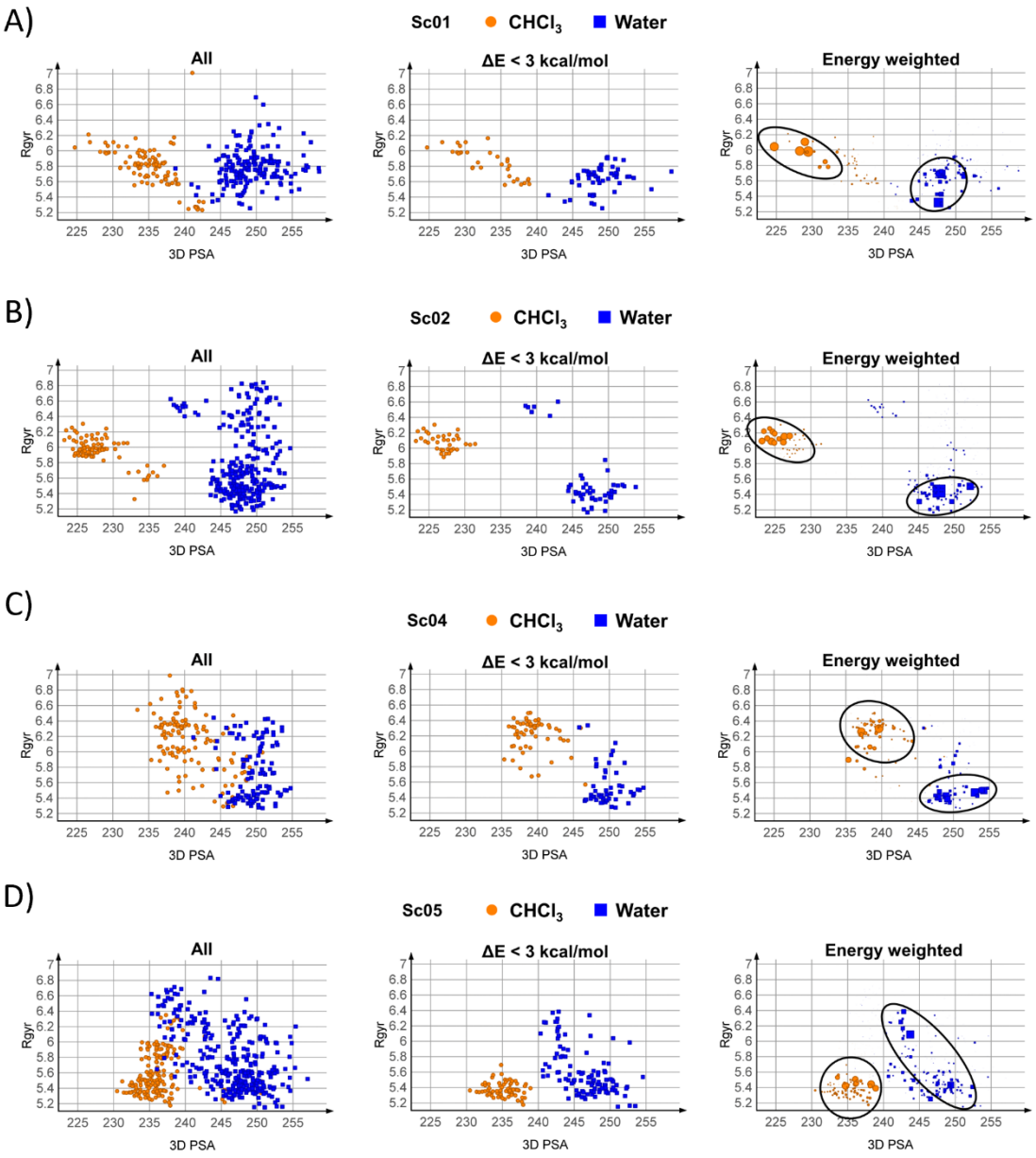



**Figure 45.** $R_{gyr}$ vs. 3D PSA plots for the ensembles obtained with the different protocols. A) Sc01, B) Sc02, C) Sc04, and D) Sc05. Black circles highlight high-probability conformer groups. Figure adapted from Ermondi *et al.,* 2023.[210]

Thus, Sc05 (especially the energy-weighted plot) shows the best although not perfect superposition with NMR conformers (Figure 46), especially in terms of $R_{gyr}$. Interestingly, the NMR-derived conformations in the polar solvent are more extended than those predicted ones by CS. In fact, none of the Sc05 conformations adopt linear shapes. A logical explanation is that NMR uses a combination of water and DMSO, providing a more polar environment and thus favoring even more extended conformations. In fact, Kihlberg and coworkers experimentally observed how the addition of water (10%) to a pure DMSO solution resulted in more folded PROTAC-1 conformations.[70] However, it is also noticeable that the mixed methods, in particular the default method (Sc01), reproduces the polarity of the NMR conformers in the best way. Overall, all the CS

strategies suggest that PROTAC-1 behaves as a chameleon, with Sc01 and Sc05 simulations being the most similar to NMR.

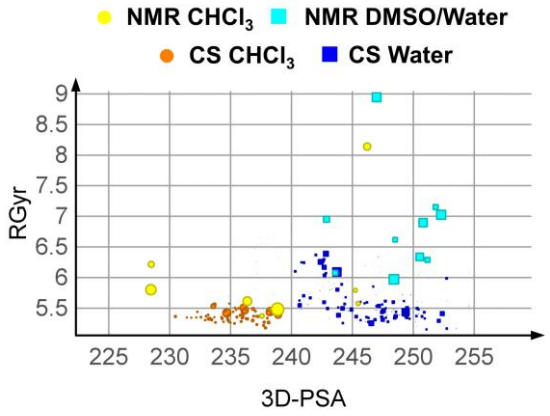



**Figure 46.** $R_{gyr}$ vs. 3D PSA comparison of NMR and Sc05 conformers (weighted scheme) in polar and nonpolar solvents. Figure adapted from Ermondi *et al.,* 2023.[210]

D. Paper VI conclusions

The growing interest in chameleonicity has created the need to develop prediction strategies. In fact, CS in implicit polar and nonpolar solvents has already been used for macrocycles.[77] However, there are no *in silico* tools available for PROTACs, even more flexible and chameleonic than macrocycles.[121] Therefore, we aimed to use the first experimentally-verified chameleonic PROTAC (PROTAC-1) to explore and benchmark CS strategies for predicting PROTAC chameleonicity. Several CS protocols were designed by modifying the algorithm (LMOD and/or MCMM), and the number of maximum steps. They were evaluated for their capacity to a) predict the chameleonicity of PROTAC-1 based on the polarity (3D PSA) and sphericity ($R_{gyr}$) of their conformations and b) obtain a property superposition with respect to the experimental NMR conformers. In particular, both standard protocols (mixed LMOD and MCMM) (Sc01-03) and MCMM algorithms (Sc04-06) (MCMM) succeeded in predicting chameleonic behavior. Moreover, Sc01 (mixed LMOD and MCMM, maximum steps $10^3$) and Sc05 (MCMM, maximum steps $10^4$) provided the best polarity and sphericity superpositions, respectively. This suggests that despite methodological differences, CS is suitable for PROTAC chameleonicity assessment. Overall, this article provides new CS protocols and strategies for the analysis of chameleonicity that can be used in future studies. However, the sensitivity of these strategies to discriminate between different candidates and the role of the IMHB in chameleonicity remain to be investigated.

E. Additional information: application of SMD to PROTAC-1

In paper VII, the application of SMD as an alternative strategy to study PROTAC chameleonicity was discussed. Consequently, we attempted to apply the same strategy

to PROTAC-1 as post-publication work. For technical details on the strategy or method, the reader is directed to paper VII.

The application of the density plot to SMD-derived conformers suggested that PROTAC-1 shows a reduction of polarity in nonpolar environments (toluene) compared to water (Figure 47A). Moreover, there is not a significant reduction in sphericity in water, which happens also in CS and NMR-derived conformations. Moreover, the presence of low-polarity conformations in toluene comes along with the formation of 1 and 2 IMHBs, absent in water (Figure 47B). As stated by Kihlberg and coworkers, the formation of IMHBs in PROTAC-1 can be a crucial factor for its permeability.[70] Overall, this short study suggests that SMD-based analysis can be used to depict the chameleonic behavior of bRo5 compounds. However, there is still room available for improvement, especially in the understanding of the specific types of IMHBs formed (future perspectives).

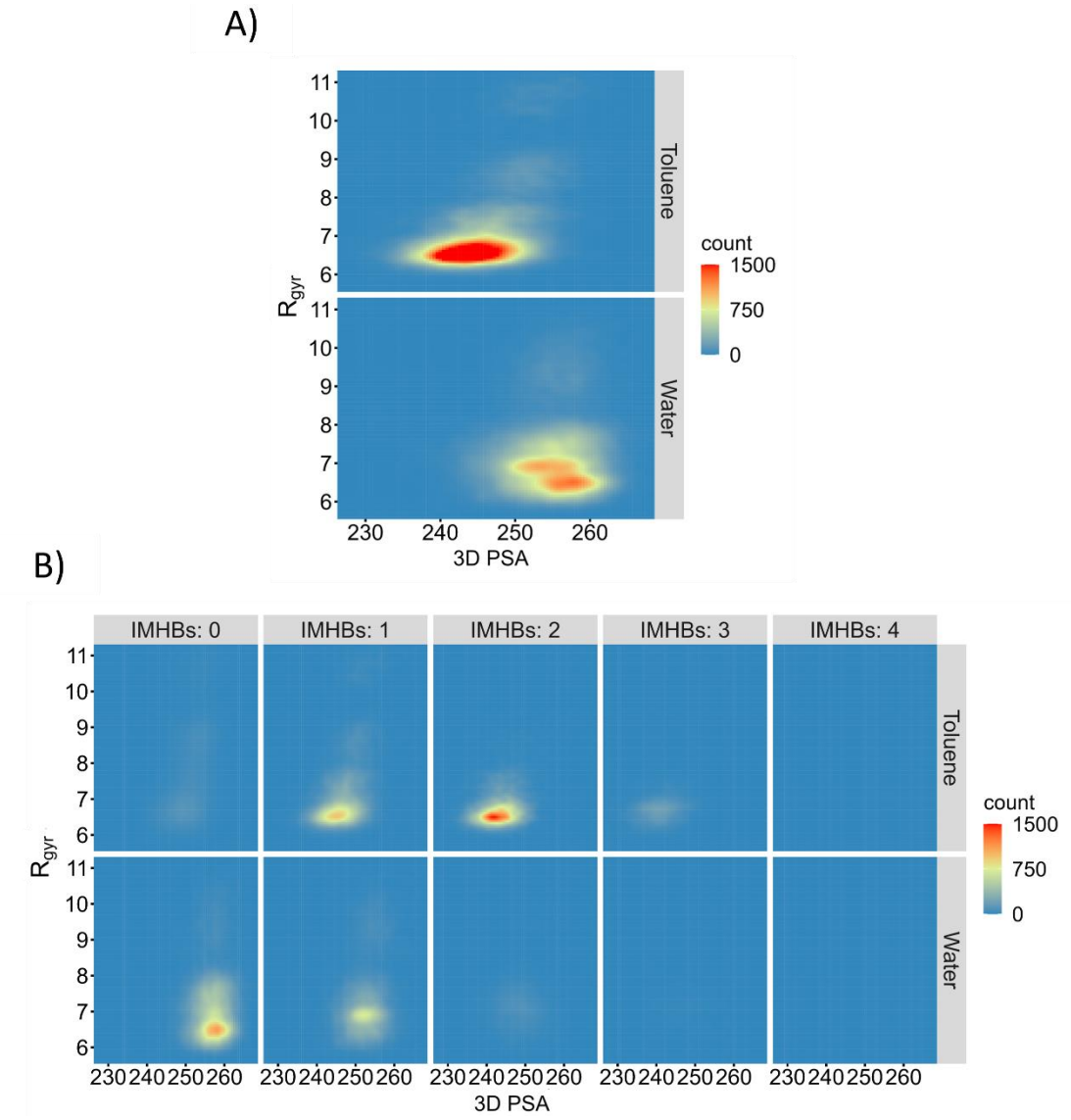



**Figure 47.** SMD Tunneling. Density plots of PROTAC-1 in A) water and toluene and then B) divided by their IMHBs.

## 9.2 Refinement of Computational Access to Molecular Physicochemical Properties: From Ro5 to bRo5 (Paper VII)[106]

Currently, CS remains the most widely used strategy to explore the conformational complexity of bRo5 compounds. In particular, its ability to reproduce experimental conformations (NMR and X-ray) and to rationalize permeability for macrocyclic and other bRo5 compounds has recently been investigated.[77,218] MD represents another useful strategy to generate environment-dependent conformations. It has mostly been used to simulate the evolution of protein-ligand complexes (PD), but its great potential to generate conformations also makes it useful for molecular property studies.[114,116,219] Up to now, a few studies have used CS and MD simultaneously to assess the conformations of bRo5 compounds in a rational way.[77] Regarding PROTACs, several studies have used MD to model the binary or ternary complexes,[220,221] while no applications have yet been reported to assess their molecular properties.

As a result, there is uncertainty about the best way to compute a conformational ensemble, find the biorelevant conformers, and determine which structural features need to be modified. In fact, up to now the chameleonicity of only one PROTAC-1 has been reproduced, using CS (Paper VI). However, the recent development of Chamelogk by our group and the availability of *ad hoc* chromatographic strategies for bRo5 compounds, provided a new opportunity to benchmark computational strategies for more than one compound and, more importantly, against proven experimental evidence, in this case, regarding all the measurable physicochemical properties (i.e., polarity, lipophilicity, chameleonicity, IMHBs). In addition, the existence of alternative strategies to the classical unbiased MD simulations made us adapt the "umbrella sampling" or SMD to the study of molecular properties. Ideally, with this strategy the molecule of interest is constrained to move in a certain direction and at a certain speed and force through the solvent, maximizing the interactions.[222] This tool is expected to be a faster tool to obtain conformations of interest compared to unbiased MD.

Therefore, three important goals of this work are a) to optimize CS and MD-based tools to model the molecular properties of structurally diverse compounds, b) to explore the pros and cons of SMD and CS and c) to provide a reasonable *in silico* strategy to assess molecular properties, confirmed with experimental data (lipophilicity, polarity, lipophilicity, chameleonicity and IMHB formation).

### A. Compounds selection

A total of three compounds belonging to different spaces were selected: (a) a Ro5 oral drug (pomalidomide, Figure 48A), (b) a bRo5 oral drug (saquinavir, Figure 48B); and (c) a non-permeable PROTAC (CMP 98, Figure 48C).

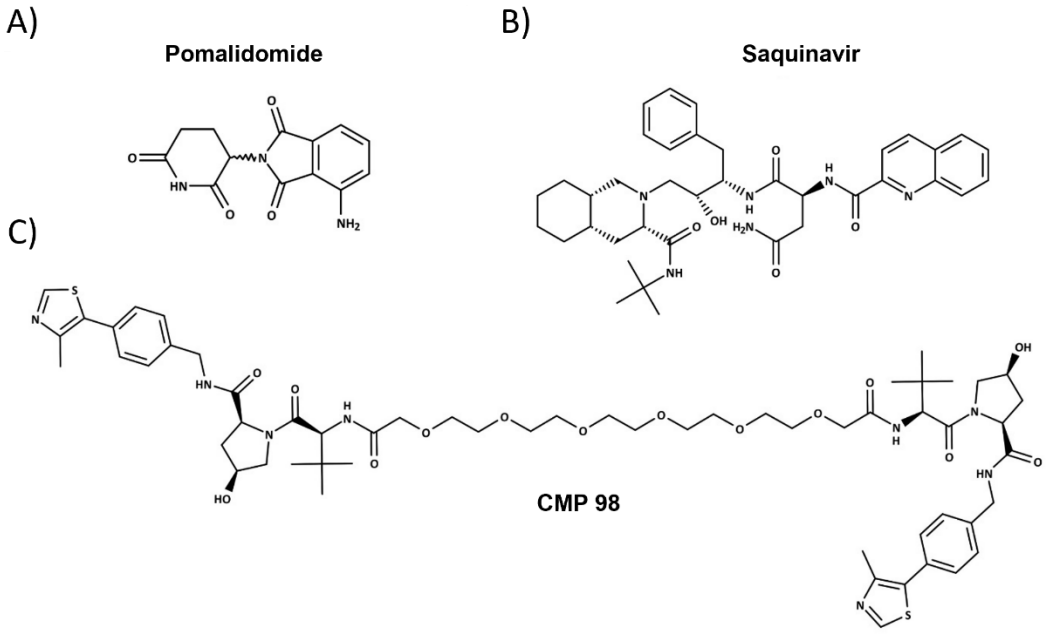



**Figure 48.** Structure of the three molecules selected: A) pomalidomide, B) saquinavir, and C) CMP 98. Figure adapted from Sebastiano *et al.,* 2022.[106]

Pomalidomide and CMP 98 are expected to be neutral at pH 7. Saquinavir has an experimental $pK_a$ of 7.1,[64] which makes it partly neutral as well. Pharmacologically, pomalidomide is an oral immunomodulatory drug for multiple myeloma.[223] Moreover, it acts as a molecular glue of two proteins (E3 ligase and POI), triggering the ubiquitin-mediated degradation of the POI. Thus, it is widely used in PROTAC technology.[224] Saquinavir is a non-macrocyclic bRo5 antiretroviral, susceptible to P-gp-mediated efflux.[127] It was chosen because of its chameleonic profile (ChameLogD and Chamelogk verified, see Chapter 8), which may contribute to a small but significant proportion of passive permeation.[64] Lastly, CMP 98 is an inactive, non-PAMPA permeable[225] and water soluble PROTAC.[63] It is an extremely flexible PROTAC, making it an ideal candidate to test the impact of flexibility on molecular properties. Overall, as suggested by the 7 2D descriptors extensively used throughout this thesis, CMP 98 is the largest (MW and nC), most flexible (PHI) and most polar compound (HBA and TPSA), followed by saquinavir and lastly by pomalidomide. In addition, the same order is maintained for hydrophobicity (NAR and nC) (Table 7).

**Table 7.** 2D molecular descriptors. Color codes account for: green (size), purple (flexibility), yellow (hydrophobicity), and blue (polarity). Table adapted from Sebastiano *et al.,* 2022.[106]

| | MW | nC | PHI | NAR | HBD | HBA | TPSA |
|---|---|---|---|---|---|---|---|
| Pomalidomide | 273 | 13 | 3 | 1 | 3 | 7 | 111 |
| Saquinavir | 671 | 38 | 12 | 3 | 6 | 11 | 167 |
| CMP 98 | 1180 | 58 | 27 | 4 | 6 | 22 | 335 |

B. Experimental physicochemical characterization

The physicochemical profiles of the three compounds were thoroughly examined by a series of experimental descriptors (Table 8) (see Methods).[95] Lipophilicity was determined in three different chromatographic systems (BRlogD, log $k_w^{IAM}$, and log $k'$ 80 PLRP-S).[82,92,97] Moreover, log $P_{tol}$ was also determined by the shake-flask method. Overall, high values suggest high lipophilicity in different environments. Two polarity indexes (EPSA and Δ log $k_w^{IAM}$) were obtained, with higher values indicating higher polarity. Additionally, an indicator of the capacity of the compounds to form IMHB (Δ log $P_{oct-tol}$) was obtained.[104,105,108] Δ log $P_{oct-tol}$ was calculated by subtracting the log $P_{oct}$ (defined by BRlogD) from the shake-flask log $P_{tol}$ value. A low value indicates that the compound has a high propensity to form IMHBs, since the formation of IMHBs is favored in toluene but not in octanol.[105] Finally, Chamelogk was also assessed as a chameleonicity indicator.[121]

**Table 8.** Experimental descriptors. Color codes account for yellow (lipophilicity), blue (polarity), cyan (IMHB-formation capacity) and black (chameleonicity). SF: shake-flask method. Table adapted from Sebastiano *et al.,* 2022.[106]

| | Lipophilicity | | | | Polarity | | IMHBs | Chameleonicity |
|---|---|---|---|---|---|---|---|---|
| | BRlogD | log $k_w^{IAM}$ | log k´80 PLRP-S | log $P_{tol}$ (SF) | Δ log $k_w^{IAM}$ | EPSA | Δ log P | Chamelogk |
| Pomalidomide | 1.41 | 0.91 | -0.65 | -0.22 | 0.64 | 82 | 1.63 | 0 |
| Saquinavir | 2.85 | 3.44 | -0.1 | 2.06 | 1.85 | 94 | 0.79 | 1.23 |
| CMP 98 | 1.38 | 2.34 | -0.84 | -0.24 | 2.1 | 103.8 | 1.62 | > 0.6 |

These data suggest that saquinavir is the most lipophilic and CMP 98 is the most polar. Moreover, saquinavir has a high tendency to form IMHBs while CMP 98 and pomalidomide do not. Notably, saquinavir is a strong chameleon, while CMP 98 behaves as a weak chameleon (Figure S17). In addition, pomalidomide does not exhibit chameleon-like behavior due to its Ro5 nature.[121] According to the experimental evidence, saquinavir has a better profile than CMP 98 to be oral. In fact, the low permeability (PAMPA) and high solubility (thermodynamic solubility) of CMP 98 is perfectly logical given its extremely high polarity.[63,225] Moreover, in our recent article on chameleonicity (see Paper V), we hypothesized that a very high chameleonicity would be required to make a compound with high polarity oral (Δ log $k_w^{IAM}$ value > 2), although a positive proof has not yet been obtained. In fact, if compared to non-oral PROTACs such as MZ1, CMP 98 is equally polar (Δ log $k_w^{IAM}$) but even less chameleonic (Chamelogk).

C. *In silico* strategies (CS and SMD)

CS and MD-based methods were used to generate environment-dependent conformers (see Methods). First, the well-performing default CS sampling method available in Maestro (Sc01 in Paper VI, mixed MCMM/LMOD algorithm, see Methods), was selected.

Water and chloroform were defined as the two solvents. Moreover, unbiased MD (hereafter referred to as MD) and steered MD (SMD)[226] were run with the CHARMM36m force field for 100 and 10 ns, respectively.[227] Notably, we observed that SMD can provide the same molecular property space in 10 ns as unbiased MD in 100 ns, and therefore MD will not be discussed in the thesis. These simulations were run in explicit solvent, using water as the polar solvent and toluene as the nonpolar solvent (see Methods). In theory, chloroform could also be used as the nonpolar phase for comparison with CS. Nevertheless, all the available experimental descriptors for lipophilicity and IMHB use toluene as the nonpolar reference (log $P_{tol}$), making toluene more suitable for comparison. Notably, future CS strategies could attempt to simulate the CS in toluene/water instead of chloroform/water. However, this option is not yet parameterized in the available software. Furthermore, CS and MD/SMD are based on different algorithms, force fields (OPLS_2005 and CHARMM36m, respectively), solvent treatment (implicit and explicit, respectively), and energy minimization (MD and SMD are not minimized). Thus, they must only be compared in terms of the final information provided.

Following the success of previous studies,[77,210] we aimed to monitor polarity (3D PSA), and the size/shape ($R_{gyr}$). Besides, IMHBs were quantified and examined in relation to the other two descriptors. However, the calculation of IMHBs also depends on several internal parameters that need to be fixed in order to obtain reproducible results. Thus, the software Chimera was used, establishing 0.4 Å and 20° as the flexible ranges for bond length and angle between HBD and HBA, respectively.[154] In addition, the donor-acceptor capacity, often referred to as the Abraham descriptors, has not been considered.[134]

D. *In silico* results (bi-property analysis)

Saquinavir and CMP 98 were plotted as a function of 3D PSA and $R_{gyr}$ (bi-descriptor analysis) for CS, and important differences were spotted (Figure 49). CMP 98 does not reach the TPSA value, which means that it is never fully exposed. Saquinavir, on the contrary, surpasses the TPSA. This suggests that saquinavir is more chameleonic.

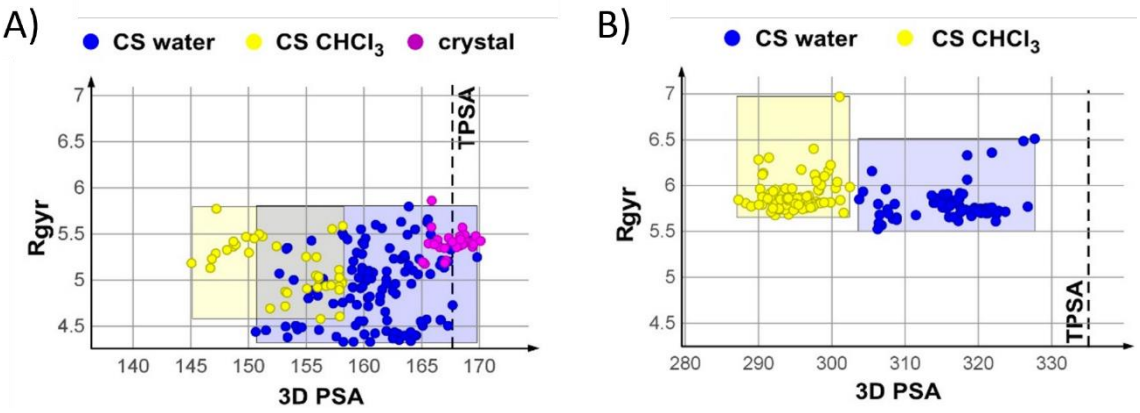


**Figure 49.** Conformational sampling: 3D PSA vs $R_{gyr}$ in water and chloroform: A) saquinavir and B) CMP 98. Individual descriptor data can be found in Tables S14 and S15. Figure adapted from Sebastiano *et al.,* 2022.[106]

Moreover, the presence of superimposable property spaces has been suggested as a factor contributing to permeability.[64] CMP 98 does not seem to have a direct superimposable space between water and chloroform (colored square in Figure 49) but saquinavir does. This may partly explain its better permeability profile (despite being mostly permeable by active transport).

SMD-derived conformers need to be analyzed using density plots, which provide information on the most populated regions. Results suggests that saquinavir displays more open and polar conformations in water (Figure 50A), and more spherical and less polar conformers in toluene (Figure 50B). CMP 98, on the other hand, behaves quite differently. In water (Figure 50C), a very dense cluster with high 3D PSA and low $R_{gyr}$ conformations is found but in toluene the conformations tend to adopt also open conformations (Figure 50D). If related to chameleonicity, both have a certain degree of chameleonicity as the conformational spaces change based on the environment. Though, saquinavir finds a completely new property space in toluene, whereas CMP 98 is superimposable to water. This data agrees with their experimental Chamelogk data (1.23 and > 0.6, respectively). Furthermore, saquinavir adopts more closed conformations in toluene which is widely known to favor permeability. CMP 98, on the contrary, would open up inside the membrane.

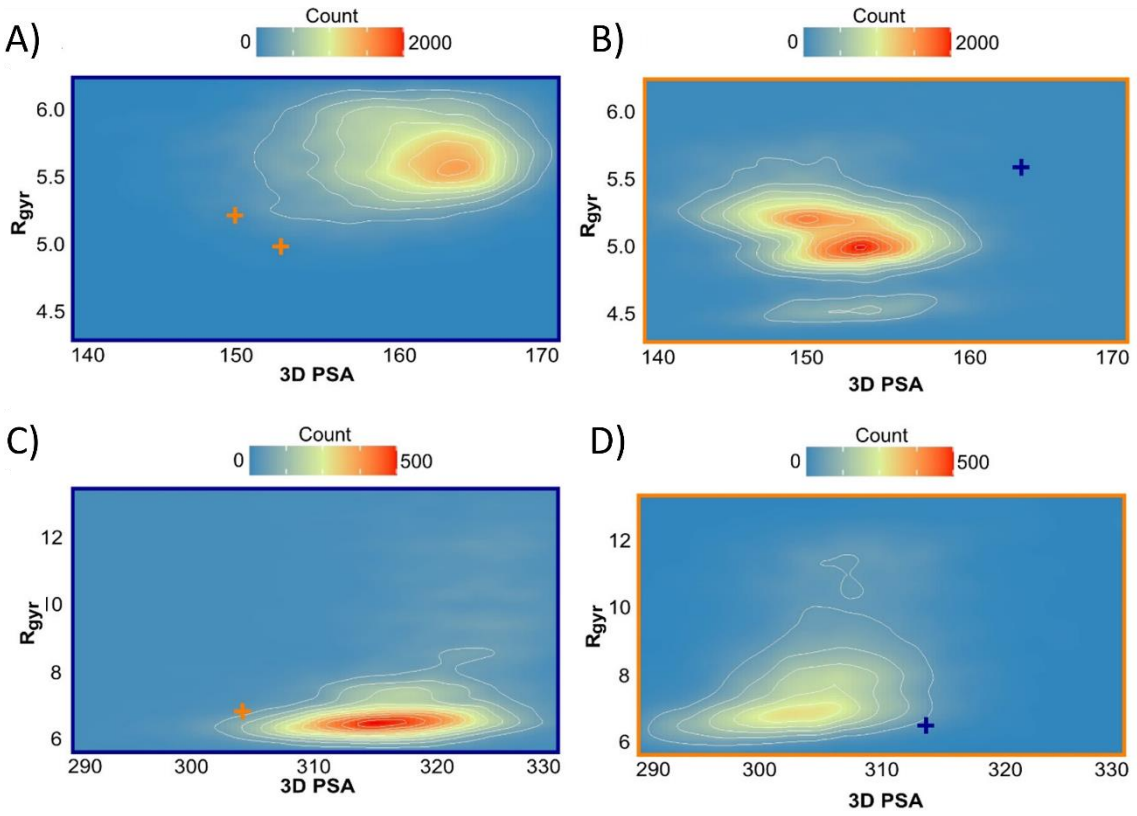


**Figure 50.** SMD Tunneling: density plots of saquinavir in water A) and toluene B) and CMP 98 in water C) and toluene D) highlighting the dispersion patterns of the generated conformers in the 2D plot of 3D PSA vs $R_{gyr}$. The color scale is expressed as conformer frequency per tile. Blue perimeter stands for water, orange for toluene, and orange/blue crosses highlight the solvent-based shift of the inner cluster. Individual descriptor data can be found in Tables S14 and S15. Figure adapted from Sebastiano *et al.,* 2022.[106]

Overall, CS and SMD bi-property plots suggest that saquinavir is less polar and more chameleonic (verified by Δ log $k_w^{IAM}$ and Chamelogk, respectively). Consequently, we attempted to examine if CS and SMD could predict the higher capacity of saquinavir to form IMHBs (experimentally verified by Δ log $P_{oct-tol}$).

No informative trends about the different IMHB profiles were obtained for CS (data not shown). However, the density plot provided valuable information for SMD (Figure 51). The conformers with the closest molecular properties to the center of each density cluster in Figure 50 were individuated and their IMHBs analyzed. The two clusters of saquinavir in toluene formed 1 and 2 IMHBs, whereas in water no IMHBs were seen (Figure 51A). This suggests that these 2 IMHBs are highly dynamic. Moreover, both clusters in toluene provide unique property regions that make the 2 IMHBs extremely relevant for chameleonicity. The clusters of CMP 98 in water and toluene revealed that 2 and 3 IMHBs were favored, respectively. However, the same two IMHBs found in water were also present in toluene. This reveals their static nature (formation independent of the solvent). Consequently, CMP 98 has only 1 dynamic IMHB, formed in toluene. In addition, this single IMHB does not provide a unique property space compared to water conformations.

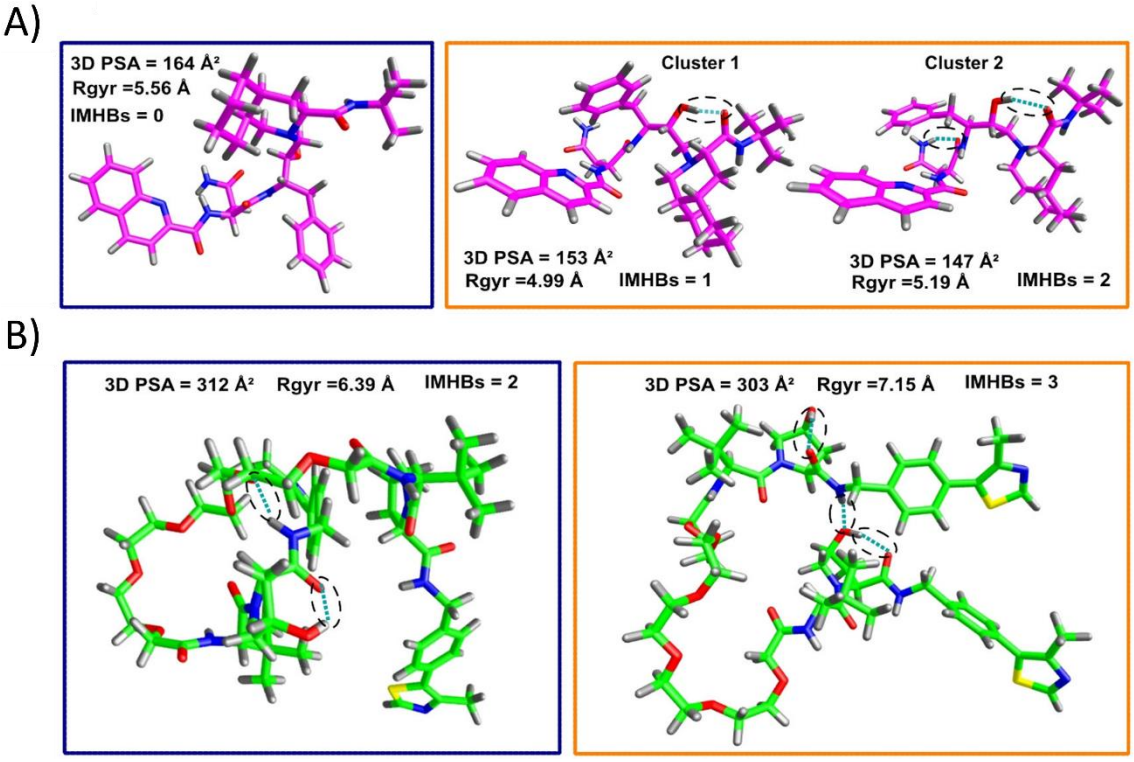


**Figure 51.** SMD Tunneling: conformer selection upon closeness to the center of the density plot (Figure 50). Blue perimeter stands for water and orange for toluene. IMHB are depicted by dashed ovals. The nIMHBs can be found in Figure S18. A) Saquinavir (purple). B) CMP 98 (green). Figure adapted from Sebastiano *et al.,* 2022.[106]

E. Agreement and missing points

Overall, the bi-descriptor property analysis of the three computational strategies (CS, MD and SMD) is in agreement with the experimental polarity and chameleonicity (Table 9). However, the quantification of the nIMHBs of CS and SMD-derived conformations (Figure S18) failed to reproduce the experimental evidence ($\Delta$ log $P_{oct\text{-}tol}$).

Moreover, while the analysis of selected CS conformations did not provide a significant explanation of the IMHB disagreement, SMD did provide useful insights. Indeed, the presence of regions of high conformer density suggested that several molecular property spaces were favored in each environment (Figure 50). Notably, in these regions saquinavir formed 1-2 dynamic IMHBs while CMP 98 formed only one. Furthermore, the dIMHBs in saquinavir provided a new property space in toluene that CMP 98 did not (Figure 50 and 51).

**Table 9.** Agreements and missing points of single and bi-descriptor strategies with experimental values.

| Generative model (polar and nonpolar) | Molecular property | Agreement |
|---|---|---|
| CS | Polarity | YES |
| | Chameleonicity | YES |
| | dIMHB formation | NO |
| MD/SMD | Polarity | YES |
| | Chameleonicity | YES |
| | dIMHB formation | YES |

Therefore, we propose that the total number of IMHBs (nIMHBs) should be replaced by the number of dynamic IMHBs between unique molecular property spaces. However, this only applies to MD-based methods, which provide comparable population densities in polar and non-polar environments. Nevertheless, we believe that rough IMHB quantifications (CS and SMD) can still be highly informative in structurally related data sets that share similar HBD and HBA moieties. Future studies will be performed to highlight this fact and to incorporate dIMHB modeling into rational design strategies.

F. Paper VII conclusions

The prediction of molecular properties is a major issue in bRo5 projects. Thus, we aimed to understand how *in silico* methods can be used in property-based designs. More specifically, in this work we have provided several milestones regarding the use of 3D molecular descriptors and the use of *in silico* strategies (CS and MD-based methods).

We have set up a steered MD method (SMD) that is more efficient than unbiased MD methods for the screening of molecular properties. Moreover, we have proposed how CS and SMD methods should be used to capture molecular property differences. In fact, the bi-descriptor analysis of these two methodologies agrees with the experimental polarity ($\Delta \log k_w^{IAM}$) and chameleonicity (Chamelogk), which are crucial for the monitoring of cell permeability and water solubility. However, the application of CS and SMD to the capacity to form of IMHBs still has some questions to be answered. First, we observed for both strategies that the simple quantification of IMHBs seems to be uninformative for structurally diverse compounds. However, SMD suggested that the quantification and evaluation of only dynamic IMHBs in polar and nonpolar solvents may be crucial to assess molecular property differences. Thus, future studies on structurally related compounds are compulsory to unravel the complexity of this topic.

## 9.3 Chapter 9 conclusions

*In this chapter, we focused on the development of new in silico strategies to study the conformational variability of bRo5 compounds in different environments (polar and non-polar) and relate them to their experimental molecular properties. For this purpose, CS was used as a straightforward method providing a general overview of the conformational space. Moreover, an SMD strategy was created to examine groups of conformations with common 3D descriptors (i.e., IMHBs). In both cases, the evaluation of the resulting conformations by polarity (3D PSA) and size/shape ($R_{gyr}$) was revealed as a valuable tool to evaluate chameleonicity.*

*The designed strategies (CS and SMD) were first applied to reproduce the chameleonicity of PROTAC-1, obtaining a good reproducibility of experimental evidence (NMR and Chamelogk). In addition, the application of both strategies to three structurally different compounds (pomalidomide, saquinavir and CMP 98) made it possible to correctly predict their polarity and chameleonicity differences (validated with experimental data).*

*However, the study of intramolecular hydrogen bonding through the count of the nIMHBs for CS and SMD-derived conformers did not agree with the experimental evidence. On the other hand, the qualitative analysis of selected populations of SMD-derived conformers allowed to shed light on the role of dynamic IMHBs on chameleonicity. However, this application needs to be confirmed on series of structurally related compounds. Overall, this chapter shows that the study of 3D conformations with in silico tools is feasible but general rules are still missing.*

# 10. General conclusions and future perspectives

New chemical modalities belonging to the bRo5 space, such as PROTACs, macrocycles and non-macrocyclic compounds are becoming popular due to their innovative mechanisms of action. However, due to their structural characteristics, they suffer from DMPK limitations that challenge their future as oral drugs. Specifically, these compounds commonly exhibit low permeability and poor solubility, ultimately resulting in low bioavailability. Presently, not only is there a lack of innovative strategies to overcome these limitations, but there is also a lack of *ad hoc* experimental and computational strategies to assess and predict their general *in vitro* ADME and physicochemical molecular properties in early drug discovery.

Therefore, the core of the thesis represented the efforts to obtain experimental and computational molecular property strategies tailored to the bRo5 space. From a drug design perspective, the obtention of good quality experimental data is an important step not only in the "hit to lead" phase, but also in the construction of *in silico* strategies to model ADME properties. Thus, the physicochemical molecular properties (i.e., polarity, lipophilicity, IMHBs) of a collection of more than 100 bRo5 and Ro5 compounds were assessed using chromatographic descriptors. This facilitated the understanding of the unique traits, strengths, and weaknesses of each bRo5 subclass (PROTACs, macrocycles, non-macrocycles). For example, macrocycles and PROTACs can exhibit excessive high lipophilicity (BRlogD) and polarity ($\Delta \log k_w^{IAM}$), respectively.

Furthermore, *in vitro* ADME properties such as thermodynamic solubility were measured. A notable correlation between solubility and lipophilicity and polarity was found, which led to the construction of ML classification models directly from molecular descriptors (Paper IV). A novel method to experimentally assess chameleonicity was developed (the Chamelogk, Paper V). Indeed, this method represents a HT strategy that deserves to be integrated in the pipeline of pharmaceutical companies.

Finally, the gained knowledge on the *in vitro* ADME and physicochemical molecular properties of bRo5 compounds during this thesis (special attention to solubility and chameleonicity), provided a golden opportunity to rationalize, for the first time, the oral F% of some bRo5 drugs based on chameleonicity, lipophilicity and polarity. As a result, the advances made in this thesis have set the basis for a correct property profiling strategy.

Relevant experimental data, such as solubility, permeability, chameleonicity, etc., became available, enabling their computational modeling. The initial approach was to select and apply a set of informative 2D descriptors in the characterization of the property space of bRo5 compounds. This knowledge was used to create the first-ever PROTAC chemical space, in which several subregions of high permeability or bioavailability were localized (Paper II and III). Moreover, the construction of a chemical space tailored to macrocycles enabled the development of predictive models for oral absorption (Paper I). Overall, these findings suggested that the use of 2D descriptor-based strategies can be useful to obtain oral bRo5 drugs. However, they need to be

complemented with information about their 3D behavior in diverse environments. Therefore, this thesis also focused on the development of *in silico* 3D strategies.

In this context, CS and MD-based strategies were optimized and adapted to the evaluation of molecular properties in polar and nonpolar environments. The scope was to mimic the behavior of the compound outside (polar) and inside (nonpolar) the cell membrane, in order to rationalize and ultimately predict experimentally verified molecular properties. The first study involved the validation of CS and SMD to reproduce the chameleonic behavior of the first NMR-proven chameleonic PROTAC (Paper VI). Moreover, both strategies were also successful in reproducing the experimentally verified differences (mainly polarity and chameleonicity) between significantly different compounds (Paper VII). Furthermore, the SMD approach showed that dynamic intramolecular hydrogen bonds can stabilize certain property spaces in an environment-dependent manner. This particular behavior, known as IMHB-mediated chameleonicity, represents a crucial feature that can be exploited in the design of bRo5 compounds.

Based on these results, our current research focuses on the study of IMHB-mediated chameleonicity. In practice, this involves assessing the Chamelogk of a series of structurally related compounds and monitoring their capacity to form dIMHBs by CS and SMD. Notably, we observed for a particular PROTAC series that chameleonicity is inversely correlated with the absence of dIMHBs (Figure S19). These initial findings suggest that chameleonicity could be intentionally induced by deliberately designing specific dIMHBs. Finally, to understand the role that chameleonicity and dIMHBs play on cell permeability, a new SMD protocol that simulates the journey of a compound through the cell membrane is being developed. In summary, the future scope is to effectively integrate chameleonicity and dIMHBs into property-based strategies.

Overall, this thesis has set the basis (method, interpretation and application) of property-based drug discovery strategies tailored to the bRo5 space.

# 11. Methods

In this section, only specific technical methodological aspects of each article will be discussed (marked throughout the text as "see Methods"). Additional information (i.e., regarding the obtention of the datasets) can be found in the original publications (see Chapter 2).

## A. Paper I

*Analysis of molecular property descriptors*

The protonation states of the compounds at pH 7.0 were adjusted using Marvin Sketch (ChemAxon, www.chemaxon.com, ver. 22.13.0, 2022),[228] and their 2D molecular descriptors were calculated using Dragon (version 7.0.10).[229] Notably, the HBDs were calculated by the software by adding up the hydrogen atoms bonded to each nitrogen and oxygen without negative charge in the molecule. In addition, the HBAs were calculated as the sum of all nitrogen, oxygen, and fluorine atoms. Moreover, the number of rotatable bonds (NRotB) and the number of aromatic rings (NAR) were calculated using DataWarrior (version 5.5.0),[230] and the logarithms of aqueous solubility (cLogS) and the octanol/water partition coefficient (cLogP) were calculated using MOE (version 2020.09).[231] Moreover, the chemical space was plotted as a PCA, an unsupervised ML algorithm that reduces the dimensionality of the data set.

*Classification of oral and parenteral macrocycles*

The samples were divided into true positives (TP) or orals, true negatives (TN) or parenterals, false positives (FP), and false negatives (FN). The performance of the models was measured with the following parameters: sensitivity, specificity, GMean, accuracy, and Cohen's kappa:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{GMean} = \sqrt{\text{Sensitivity} \cdot \text{Specificity}}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

$$\kappa = \frac{2 \cdot (\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN})}{(\text{TP}+\text{FP}) \cdot (\text{FP}+\text{TN}) + (\text{TP}+\text{FN}) \cdot (\text{FN}+\text{TN})}$$

## B. Paper II and III

The PROTAC-DB was downloaded, and the SMILES codes (untouched ionization states) were submitted to DataWarrior and Dragon following the same descriptor calculation protocol as described for Paper I.

C. Paper IV

*Solubility determination*

Thermodynamic solubility was measured in 10 mM PBS (0.15 M KCl) at pH 7, 25 °C. First, a 1−2 mg sample of each molecule was added to the same volume (1-2 mL) of the mentioned buffer to obtain 1 mg/mL solutions in the vial. The samples were then heated up to 25 °C and stirred at 500 rpm for 1 h (minimal time to ensure thermodynamic equilibrium of the less soluble compounds). After this period, the solutions or suspensions were filtered through the 0.45 µm membrane pore and diluted with buffer, 10 mM PBS (0.15 M KCl). The amount of dissolved compound in each sample was then quantified, in duplicate, by HPLC using ultraviolet (UV) spectrometric detection. Finally, the solubility value was obtained by interpolating the absorbance value on several concentration equations obtained from calibration curves. These were previously performed using 5% DMSO 10 mM PBS solutions of the compounds. For each compound, the best column and conditions were selected. More details on the quantification process can be found in the original article.

*Chromatographic descriptors determination*

For all the chromatographic descriptors, the mobile phase consisted of a solution of acetonitrile (ACN) and 20 mM ammonium acetate buffer (AAB), pH 7. Moreover, 10 µL was set as the volume of injection of compound concentrations ranging from 50 to 100 µg/mL in ACN/buffer mixtures. 1 mL/min was chosen as the isocratic flow rate and 30 °C as the oven temperature. However, each chromatographic system (RP) had its own conditions:

- BRlogD

The measurement of BRlogD as an analog of log $P_{oct}$ required the use of the XBridge Shield RP18 (130 Å, 5 µm, 5 cm × 4.6 mm) column from Waters (www.waters.com). The eluent consisted of a 60% ACN and 40% 20 mM ammonium acetate buffer (v/v).[92] Log k'60 was calculated (capacity factor k'60 = $[t_{R(60\%\ ACN)} - t_0]/t_0$) and then converted to the corresponding BRlogD value using the equation BRlogD = 3.31x + 2.79. Moreover, acetone, caffeine, ibuprofen, lidocaine, phenol, and a mixture of uracile, acetophenone, and toluene were used as gold standards.[92]

- Log $k_w^{IAM}$

This descriptor required the use of an IAM.PC.DD2 (300 Å, 10 µm, 10 cm × 4.6 mm) column from REGIS. The organic solvent of the eluent, ACN, was modified at various percentages (from 10 to 50%, v/v) and the retention times were measured. Next, the capacity factor (log k´) values were calculated as described elsewhere and the equation represented by the five was obtained. Lastly the extrapolated value at 0% ACN (100% aqueous buffer) or log $k_w^{IAM}$ was obtained. In addition, caffeine, carbamazepine, ketoprofen, theobromine, and toluene were checked daily as gold standards.[97–99]

- $\Delta \log k_w^{IAM}$

This polarity descriptor was calculated using the equation $\Delta \log k_w^{IAM} = \log k_w^{IAM} - \text{clog } k_w^{IAM}$. Clog $k_w^{IAM}$, defined as the log $k_w^{IAM}$ for neutral compounds with PSA = 0, has been correlated to BRlogD by our group with the equation clog $k_w^{IAM}$ = 0.92*BRlogD - 1.03.[97]

*$pK_a$ determination*

The ionization profile of pomalidomide, S,R,S-AHPC HCl, and S,S,S-AHPC 2HCl was measured by potentiometry. Titrations were performed on 0.15 M KCl sample solutions under a nitrogen atmosphere at 25 ± 1 °C. Moreover, standardized 0.5 M KOH and 0.5 M HCl were used as titration reagents. PROTACs were also assessed but their low solubility in solution with or without cosolvents limited the results to building blocks only.

*Calculated descriptors*

The calculated descriptors are collected in Tables S9 and S11. In particular, the predicted solubility was calculated with AdmetSAR2 (www.lmmd.ecust.edu.cn/admetsar2/), ADMETLab (www.scbdd.com), pkCSM (biosig.unimelb.edu.au/pkcsm), VolSurf+ (VS+, www.moldiscovery.com, ver. 1.1.2, 2016) and Marvin Sketch. Calculated log P values were obtained with: Swissadme (www.swissadme.ch/index.php), ACD Laboratories (www.acdlabs.com), Molinspiration (www.molinspiration.com), ADMETLab (www.admet.scbdd.com), AdmetSAR2 (www.lmmd.ecust.edu.cn/admetsar2/), pkCSM (biosig.unimelb.edu.au/pkcsm), VolSurf+, MoKa (www.moldiscovery.com, ver. 3.2.2, 2019) and Marvin Sketch.

*Generation of conformers*

For the set of 14 PROTACs, the conformations in water were obtained using the same procedure of CS and SMD as in Paper VII.

*Classification models*

The solubility classification models, based on log $k_w^{IAM}$, BRlogD, and TPSA were performed using the Weka software (version 3.8.5).[232] The supervised Random Tree and Random Forest algorithms were run using a 10-fold cross-validation (90% training set, 10% test set, repeated 10 times) with default parameters to construct accurate prediction models. The input values were low, intermediate and high solubility and the three property parameters (log $k_w^{IAM}$, BRlogD, and TPSA).

D. <u>Paper V</u>

*Chromatographic descriptors determination*

BRlogD, $\Delta \log k_w^{IAM}$ and $\log k_w^{IAM}$ were measured as described in the previous article. Log k' 80 PLRP-S and Chamelogk, on the other hand, were measured using the PLRP-S polymeric reversed-phase column (100 Å, 5 µm, 50 × 4.6 mm) from Agilent (www.agilent.com). The assessment of log k' 80 PLRP-S involved the measurement and calculation of the capacity factor of each sample at an eluent composition of 80% ACN and 20% 20 mM ammonium acetate at pH 7.0.[82] In addition, acetone, caffeine, phenol and a mixture of uracile, acetophenone and toluene were used as gold standards. Moreover, the measurement of Chamelogk required the measurement of the retention time of each compound in the dataset under six different mobile phase conditions (50 to 100% ACN) using the PLRP-S column.[121]

E. Paper VI

*Conformational sampling and calculation of 3D descriptors*

The initial 3D geometry of PROTAC-1 was created using Maestro (Schrödinger Release 2022-3: Maestro, Schrödinger, LLC, New York, NY, USA, 2021) and the structure was checked for its chiral centers and bond order. The structure was then subjected to conformational sampling using the MacroModel plugin (MacroModel, Schrödinger, LLC, New York, NY, 2023), which provided the mixed torsional/low-mode (LMOD) and Monte Carlo torsional (MMCM) sampling algorithms. Water and chloroform were selected as the two implicit environments. Additional configuration details are reported in Table S13. Molecular descriptors were then calculated with VEGA (http://www.vegazz.net/):[233] Polarity was calculated as the 3D PSA (probe radius 0 Å) and the shape and/or size as $R_{gyr}$. Lastly, IMHBs were calculated with USCF Chimera (https://www.rbvi.ucsf.edu/chimera/). The parameters 0.4 Å bond distance and 20° (HBD-HBA angle) were used. In this context, hydrogen atoms bound to nitrogen, oxygen, and sulfur were designated as donors, while nitrogen, oxygen, and sulfur atoms possessing lone pairs were identified as acceptors.[234]

F. Paper VII

The chromatographic descriptors were measured as described in Papers IV and V.

*Log $P_{tol}$ determination*

Log $P_{tol}$ was determined using a shake flask method as described by Shalaeva and coworkers.[105] Samples were dissolved in specific volumes of 2% DMSO 20 mM 0.15M KCl ammonium acetate buffer, pH 7 to obtain final compound concentrations of < 1 mg/ml. After the addition of the same volume of toluene, the samples were then vortexed for 10 min. After this step, the aqueous and organic phases were separated and quantified by HPLC for each sample.

*Conformational sampling*

The initial 3D geometries were generated using the CORINA demo (www.mnam.com/online-demos/corina_demo), directly from the SMILES codes. For pomalidomide, the S enantiomer was considered.

The refined 3D structures were then submitted to Maestro and MacroModel, as described in the previous article. In this case, the CS protocol (mixed LMOD/MCMM) with the OPLS_2005 force field was used (the remaining parameters were maintained as in Sc01). The calculated descriptors (3D PSA, $R_{gyr}$ and IMHBs) were obtained in the same way as in Paper VI.

*Molecular dynamics*

The refined 3D structures obtained from Corina (same structure files as CS) were used to generate the simulation files with CHARMM-GUI (www.charmm-gui.org/).[235] Toluene and water were selected as the nonpolar and polar environments, and two different protocols were required.

First, the CHARMM36 (force field) parameters for water were generated using the "ligand reader and modeler" functionality of CHARMM-GUI. Moreover, the periodic water boundaries, (50-70 Å), solvation box, and MD input files for an NPT ensemble at 300 K were generated using the "solution builder" function.

On the other hand, the 3D structure of toluene was generated with VMD (http://www.ks.uiuc.edu/Research/vmd/)[236] and parametrized with CHARMM-GUI for the force field and required parameters. The structure files were combined with the molecule files to build a solvent box (50-70 Å), with toluene as the solvent, using the "multicomponent assembler" function. Toluene was defined at a density of 870 g/L at RT. Moreover, the conditions were set as in water, 300 K and NPT.

Afterwards, the systems were equilibrated for 250 ps with NAMD2 (www.ks.uiuc.edu/Research/namd/, 2.13 CUDA-accelerated version)[237] and subsequently subjected to a 10 or 100 ns run on a Linux workstation (OS, CentOS7, 32GB DDR2; CPU, Xeon Octa-core 3.50 GHz, Titan XP GPU). The resulting trajectories for water and toluene simulations were analyzed with VMD.

*Steered molecular dynamics*

The implementation of a "steered" simulation required the addition of several parameters to the instructions of the MD production step: "SMD = on, SMDk = 7.0 kcal/mol/Å, SMDvel = 2e–05 Å/ts, SMDdir = 0.0, 1.0, 0.0". This sets the force constant, direction and velocity instructions, respectively. Also, the occupancy factor of the input PDB was modified to fix the position of the solvent and to allow the free movement of the atoms in the molecule of interest. Moreover, the SMD was run for 10 ns.

G. General statistical and graphical analysis

Analyses were performed using Microsoft Excel (www.microsoft.com, version 2010), Matlab (https://www.mathworks.com/products/matlab.html, 2019a), alvaDesc (https://www.alvascience.com/alvadesc/, version 2.0.0), GraphPad Prism (www.graphpad.com, version 8.0.0, 2019), and R Studio (https://www.r-project.org/, version 2022.02.3).

## 12. References


(1) Drews, J. Drug Discovery: A Historical Perspective. *Science (1979)* **2000**, *287* (5460), 1960–1964.

(2) Krishnamurti, C.; Rao, Sscc. The Isolation of Morphine by Serturner. *Indian J Anaesth* **2016**, *60* (11), 861.

(3) Chain, E.; Florey, H. W.; Gardner, A. D.; Heatley, N. G.; Jennings, M. A.; Orr-Ewing, J.; Sanders, A. G. PENICILLIN AS A CHEMOTHERAPEUTIC AGENT. *The Lancet* **1940**, *236* (6104), 226–228.

(4) Ribeiro da Cunha; Fonseca; Calado. Antibiotic Discovery: Where Have We Come from, Where Do We Go? *Antibiotics* **2019**, *8* (2), 45.

(5) Mouchlis, V. D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A. G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *Int J Mol Sci* **2021**, *22* (4), 1676.

(6) Senior, M. Fresh from the Biotech Pipeline: Fewer Approvals, but Biologics Gain Share. *Nat Biotechnol* **2023**.

(7) Mullard, A. 2021 FDA Approvals. *Nat Rev Drug Discov* **2022**, *21* (2), 83–88.

(8) Ngo, H. X.; Garneau-Tsodikova, S. What Are the Drugs of the Future? *Medchemcomm* **2018**, *9* (5), 757–758.

(9) Yang, W.; Gadgil, P.; Krishnamurthy, V. R.; Landis, M.; Mallick, P.; Patel, D.; Patel, P. J.; Reid, D. L.; Sanchez-Felix, M. The Evolving Druggability and Developability Space: Chemically Modified New Modalities and Emerging Small Molecules. *AAPS J* **2020**, *22* (2), 21.

(10) Dang, C. V.; Reddy, E. P.; Shokat, K. M.; Soucek, L. Drugging the "undruggable" Cancer Targets. *Nat Rev Cancer* **2017**, *17* (8), 502–508.

(11) Blanco, M. J.; Gardinier, K. M. New Chemical Modalities and Strategic Thinking in Early Drug Discovery. *ACS Med Chem Lett* **2020**, *11* (3), 228–231.

(12) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv Drug Deliv Rev* **1997**, *23* (1–3), 3–25.

(13) Sakamoto, K. M.; Kim, K. B.; Kumagai, A.; Mercurio, F.; Crews, C. M.; Deshaies, R. J. Protacs: Chimeric Molecules That Target Proteins to the Skp1-Cullin-F Box Complex for Ubiquitination and Degradation. *Proc Natl Acad Sci U S A* **2001**, *98* (15), 8554–8559.

(14) Burslem, G. M.; Crews, C. M. Small-Molecule Modulation of Protein Homeostasis. *Chem Rev* **2017**, *117* (17), 11269–11301.

(15) Bondeson, D. P.; Mares, A.; Smith, I. E. D.; Ko, E.; Campos, S.; Miah, A. H.; Mulholland, K. E.; Routly, N.; Buckley, D. L.; Gustafson, J. L.; Zinn, N.; Grandi, P.; Shimamura, S.; Bergamini, G.; Faelth-Savitski, M.; Bantscheff, M.; Cox, C.; Gordon, D. A.; Willard, R. R.; Flanagan, J. J.; Casillas, L. N.; Votta, B. J.; den Besten, W.; Famm, K.; Kruidenier, L.; Carter, P. S.; Harling, J. D.; Churcher, I.; Crews, C. M. Catalytic in Vivo Protein Knockdown by Small-Molecule PROTACs. *Nat Chem Biol* **2015**, *11* (8), 611–617.

(16) Graham, H. The Mechanism of Action and Clinical Value of PROTACs: A Graphical Review. *Cell Signal* **2022**, *99*, 110446.

(17) Gao, H.; Sun, X.; Rao, Y. PROTAC Technology: Opportunities and Challenges. *ACS Med Chem Lett* **2020**, *11* (3), 237–240.
(18) Chirnomas, D.; Hornberger, K. R.; Crews, C. M. Protein Degraders Enter the Clinic — a New Approach to Cancer Therapy. *Nat Rev Clin Oncol* **2023**, *20* (4), 265–278.
(19) Han, X.; Sun, Y. PROTACs: A Novel Strategy for Cancer Drug Discovery and Development. *MedComm (Beijing)* **2023**, *4* (3).
(20) Jimenez, D. G.; Sebastiano, M. R.; Caron, G.; Ermondi, G. Are We Ready to Design Oral PROTACs®? *ADMET DMPK* **2021**, *9* (4), 243–254.
(21) Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The Exploration of Macrocycles for Drug Discovery — an Underexploited Structural Class. *Nat Rev Drug Discov* **2008**, *7* (7), 608–624.
(22) Garcia Jimenez, D.; Poongavanam, V.; Kihlberg, J. Macrocycles in Drug Discovery–Learning from the Past for the Future. *J Med Chem* **2023**, *66* (8), 5377–5396.
(23) Doak, B. C.; Zheng, J.; Dobritzsch, D.; Kihlberg, J. How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets. *J Med Chem* **2016**, *59* (6), 2312–2327.
(24) Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J Med Chem* **2014**, *57* (2), 278–295.
(25) Mallinson, J.; Collins, I. Macrocycles in New Drug Discovery. *Future Med Chem* **2012**, *4* (11), 1409–1438.
(26) Tsantrizos, Y. S. The Design of a Potent Inhibitor of the Hepatitis C Virus NS3 Protease: BILN 2061—From the NMR Tube to the Clinic. *Peptide Science* **2004**, *76* (4), 309–323.
(27) Begnini, F.; Poongavanam, V.; Over, B.; Castaldo, M.; Geschwindner, S.; Johansson, P.; Tyagi, M.; Tyrchan, C.; Wissler, L.; Sjö, P.; Schiesser, S.; Kihlberg, J. Mining Natural Products for Macrocycles to Drug Difficult Targets. *J Med Chem* **2021**, *64* (2), 1054–1072.
(28) Wani, M. C.; Taylor, H. L.; Wall, M. E.; Coggon, P.; McPhail, A. T. Plant Antitumor Agents. VI. Isolation and Structure of Taxol, a Novel Antileukemic and Antitumor Agent from Taxus Brevifolia. *J Am Chem Soc* **1971**, *93* (9), 2325–2327.
(29) Kempf, D. J.; Sham, H. L.; Marsh, K. C.; Flentge, C. A.; Betebenner, D.; Green, B. E.; McDonald, E.; Vasavanonda, S.; Saldivar, A.; Wideburg, N. E.; Kati, W. M.; Ruiz, L.; Zhao, C.; Fino, L.; Patterson, J.; Molla, A.; Plattner, J. J.; Norbeck, D. W. Discovery of Ritonavir, a Potent Inhibitor of HIV Protease with High Oral Bioavailability and Clinical Efficacy. *J Med Chem* **1998**, *41* (4), 602–617.
(30) Nelson, E. Kinetics of Drug Absorption, Distribution, Metabolism, and Excretion. *J Pharm Sci* **1961**, *50* (3), 181–192.
(31) Sastry, S. V.; Nyshadham, J. R.; Fix, J. A. Recent Technological Advances in Oral Drug Delivery – a Review. *Pharm Sci Technol Today* **2000**, *3* (4), 138–145.
(32) Alqahtani, M. S.; Kazi, M.; Alsenaidy, M. A.; Ahmad, M. Z. Advances in Oral Drug Delivery. *Front Pharmacol* **2021**, *12*.
(33) Verbrugghe, M.; Verhaeghe, S.; Lauwaert, K.; Beeckman, D.; Van Hecke, A. Determinants and Associated Factors Influencing Medication Adherence and Persistence to Oral Anticancer Drugs: A Systematic Review. *Cancer Treat Rev* **2013**, *39* (6), 610–621.

(34) Abuhelwa, A. Y.; Williams, D. B.; Upton, R. N.; Foster, D. J. R. Food, Gastrointestinal PH, and Models of Oral Drug Absorption. *European Journal of Pharmaceutics and Biopharmaceutics* **2017**, *112*, 234–248.
(35) Evans, D. F.; Pye, G.; Bramley, R.; Clark, A. G.; Dyson, T. J.; Hardcastle, J. D. Measurement of Gastrointestinal PH Profiles in Normal Ambulant Human Subjects. *Gut* **1988**, *29* (8), 1035–1041.
(36) Artursson, P.; Palm, K.; Luthman, K. Caco-2 Monolayers in Experimental and Theoretical Predictions of Drug Transport. *Adv Drug Deliv Rev* **2012**, *64*, 280–289.
(37) Lin, J. H.; Yamazaki, M. Role of P-Glycoprotein in Pharmacokinetics. *Clin Pharmacokinet* **2003**, *42* (1), 59–98.
(38) Kruh, G. D.; Belinsky, M. G. The MRP Family of Drug Efflux Pumps. *Oncogene* **2003**, *22* (47), 7537–7552.
(39) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem Biol* **2014**, *21* (9), 1115–1142.
(40) Huang, W.; Lee, S. L.; Yu, L. X. Mechanistic Approaches to Predicting Oral Drug Absorption. *AAPS J* **2009**, *11* (2), 217–224.
(41) Di, L.; Kerns, E.; Carter, G. Drug-Like Property Concepts in Pharmaceutical Design. *Curr Pharm Des* **2009**, *15* (19), 2184–2194.
(42) Bunally, S. B.; Luscombe, C. N.; Young, R. J. Using Physicochemical Measurements to Influence Better Compound Design. *SLAS Discovery* **2019**, *24* (8), 791–801.
(43) Dahan, A.; Miller, J. M.; Amidon, G. L. Prediction of Solubility and Permeability Class Membership: Provisional BCS Classification of the World's Top Oral Drugs. *AAPS Journal* **2009**, *11* (4), 740–746.
(44) Berge, S. M.; Bighley, L. D.; Monkhouse, D. C. Pharmaceutical Salts. *J Pharm Sci* **1977**, *66* (1), 1–19.
(45) Pouton, C. W. Formulation of Poorly Water-Soluble Drugs for Oral Administration: Physicochemical and Physiological Issues and the Lipid Formulation Classification System. *European Journal of Pharmaceutical Sciences* **2006**, *29* (3-4 SPEC. ISS.), 278–287.
(46) Roy, K.; Kar, S.; Das, R. N. QSAR/QSPR Modeling: Introduction; 2015; pp 1–36.
(47) Kuepfer, L.; Niederalt, C.; Wendl, T.; Schlender, J.; Willmann, S.; Lippert, J.; Block, M.; Eissing, T.; Teutonico, D. Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model. *CPT Pharmacometrics Syst Pharmacol* **2016**, *5* (10), 516–531.
(48) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J Med Chem* **2002**, *45* (12), 2615–2623.
(49) Whitty, A.; Zhong, M.; Viarengo, L.; Beglov, D.; Hall, D. R.; Vajda, S. Quantifying the Chameleonic Properties of Macrocycles and Other High-Molecular-Weight Drugs. *Drug Discov Today* **2016**, *21* (5), 712–717.
(50) DeGoey, D. A.; Chen, H.-J.; Cox, P. B.; Wendt, M. D. Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection. *J Med Chem* **2018**, *61* (7), 2636–2651.
(51) Maple, H. J.; Clayden, N.; Baron, A.; Stacey, C.; Felix, R. Developing Degraders: Principles and Perspectives on Design and Chemical Space. *Medchemcomm* **2019**, *10* (10), 1755–1764.

(52) Shultz, M. D. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. *J Med Chem* **2019**, *62* (4), 1701–1714.
(53) Chris Lipinski. *Nat Rev Drug Discov* **2012**, *11* (12), 900–901.
(54) Leeson, P. D.; Springthorpe, B. The Influence of Drug-like Concepts on Decision-Making in Medicinal Chemistry. *Nat Rev Drug Discov* **2007**, *6* (11), 881–890.
(55) Matsson, P.; Doak, B. C.; Over, B.; Kihlberg, J. Cell Permeability beyond the Rule of 5. *Adv Drug Deliv Rev* **2016**, *101*, 42–61.
(56) Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical Properties As a Means of Improving Compound Disposition and Safety. *Chem Res Toxicol* **2011**, *24* (9), 1420–1456.
(57) Matsson, P.; Kihlberg, J. How Big Is Too Big for Cell Permeability? *J Med Chem* **2017**, *60* (5), 1662–1664.
(58) Nassar, A.-E. F.; Kamel, A. M.; Clarimont, C. Improving the Decision-Making Process in the Structural Modification of Drug Candidates: Enhancing Metabolic Stability. *Drug Discov Today* **2004**, *9* (23), 1020–1028.
(59) Valko, K.; Nunhuck, S.; Bevan, C.; Abraham, M. H.; Reynolds, D. P. Fast Gradient HPLC Method to Determine Compounds Binding to Human Serum Albumin. Relationships with Octanol/Water and Immobilized Artificial Membrane Lipophilicity. *J Pharm Sci* **2003**, *92* (11), 2236–2248.
(60) Shamovsky, I.; Connolly, S.; David, L.; Ivanova, S.; Nordén, B.; Springthorpe, B.; Urbahns, K. Overcoming Undesirable HERG Potency of Chemokine Receptor Antagonists Using Baseline Lipophilicity Relationships. *J Med Chem* **2008**, *51* (5), 1162–1178.
(61) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opin Drug Discov* **2010**, *5* (3), 235–248.
(62) Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G. Flexibility in Early Drug Discovery: Focus on the beyond-Rule-of-5 Chemical Space. *Drug Discov Today* **2020**, *25* (4), 621–627.
(63) García Jiménez, D.; Rossi Sebastiano, M.; Vallaro, M.; Mileo, V.; Pizzirani, D.; Moretti, E.; Ermondi, G.; Caron, G. Designing Soluble PROTACs: Strategies and Preliminary Guidelines. *J Med Chem* **2022**, *65* (19), 12639–12649.
(64) Rossi Sebastiano, M.; Doak, B. C.; Backlund, M.; Poongavanam, V.; Over, B.; Ermondi, G.; Caron, G.; Matsson, P.; Kihlberg, J. Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs beyond the Rule of 5. *J Med Chem* **2018**, *61* (9), 4189–4202.
(65) Leeson, P. D.; Young, R. J. Molecular Property Design: Does Everyone Get It? *ACS Med Chem Lett* **2015**, *6* (7), 722–725.
(66) Ermondi, G.; Garcia Jimenez, D.; Rossi Sebastiano, M.; Caron, G. Rational Control of Molecular Properties Is Mandatory to Exploit the Potential of PROTACs as Oral Drugs. *ACS Med Chem Lett* **2021**, *12* (7), 1056–1060.
(67) Carrupt, P. A.; Testa, B.; Bechalany, A.; El Tayar, N.; Descas, P.; Perrissoud, D. Morphine 6-Glucuronide and Morphine 3-Glucuronide as Molecular Chameleons with Unexpected Lipophilicity. *J Med Chem* **1991**, *34* (4), 1272–1275.
(68) Loosli, H.-R.; Kessler, H.; Oschkinat, H.; Weber, H.-P.; Petcher, T. J.; Widmer, A. Peptide Conformations. Part 31. The Conformation of Cyclosporin a in the Crystal and in Solution. *Helv Chim Acta* **1985**, *68* (3), 682–704.

(69) Alex, A.; Millan, D. S.; Perez, M.; Wakenhut, F.; Whitlock, G. A. Intramolecular Hydrogen Bonding to Improve Membrane Permeability and Absorption in beyond Rule of Five Chemical Space. *Medchemcomm* **2011**, *2* (7), 669.

(70) Atilaw, Y.; Poongavanam, V.; Svensson Nilsson, C.; Nguyen, D.; Giese, A.; Meibom, D.; Erdelyi, M.; Kihlberg, J. Solution Conformations Shed Light on PROTAC Cell Permeability. *ACS Med Chem Lett* **2021**, *12* (1), 107–114.

(71) Poongavanam, V.; Atilaw, Y.; Siegel, S.; Giese, A.; Lehmann, L.; Meibom, D.; Erdelyi, M.; Kihlberg, J. Linker-Dependent Folding Rationalizes PROTAC Cell Permeability. *J Med Chem* **2022**, *65* (19), 13029–13040.

(72) Hill, T. A.; Lohman, R.-J.; Hoang, H. N.; Nielsen, D. S.; Scully, C. C. G.; Kok, W. M.; Liu, L.; Lucke, A. J.; Stoermer, M. J.; Schroeder, C. I.; Chaousis, S.; Colless, B.; Bernhardt, P. V; Edmonds, D. J.; Griffith, D. A.; Rotter, C. J.; Ruggeri, R. B.; Price, D. A.; Liras, S.; Craik, D. J.; Fairlie, D. P. Cyclic Penta- and Hexaleucine Peptides without N-Methylation Are Orally Absorbed. *ACS Med Chem Lett* **2014**, *5* (10), 1148–1151.

(73) Bhardwaj, G.; O'Connor, J.; Rettie, S.; Huang, Y.-H.; Ramelot, T. A.; Mulligan, V. K.; Alpkilic, G. G.; Palmer, J.; Bera, A. K.; Bick, M. J.; Di Piazza, M.; Li, X.; Hosseinzadeh, P.; Craven, T. W.; Tejero, R.; Lauko, A.; Choi, R.; Glynn, C.; Dong, L.; Griffin, R.; van Voorhis, W. C.; Rodriguez, J.; Stewart, L.; Montelione, G. T.; Craik, D.; Baker, D. Accurate de Novo Design of Membrane-Traversing Macrocycles. *Cell* **2022**, *185* (19), 3520-3532.e26.

(74) Sheikh, A. Y.; Mattei, A.; Miglani Bhardwaj, R.; Hong, R. S.; Abraham, N. S.; Schneider-Rauber, G.; Engstrom, K. M.; Diwan, M.; Henry, R. F.; Gao, Y.; Juarez, V.; Jordan, E.; DeGoey, D. A.; Hutchins, C. W. Implications of the Conformationally Flexible, Macrocyclic Structure of the First-Generation, Direct-Acting Anti-Viral Paritaprevir on Its Solid Form Complexity and Chameleonic Behavior. *J Am Chem Soc* **2021**, *143* (42), 17479–17491.

(75) Kokate, A.; Li, X.; Jasti, B. Effect of Drug Lipophilicity and Ionization on Permeability Across the Buccal Mucosa: A Technical Note. *AAPS PharmSciTech* **2008**, *9* (2), 501–504.

(76) Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. The Significance of Acid/Base Properties in Drug Discovery. *Chem. Soc. Rev.* **2013**, *42* (2), 485–496.

(77) Poongavanam, V.; Danelius, E.; Peintner, S.; Alcaraz, L.; Caron, G.; Cummings, M. D.; Wlodek, S.; Erdelyi, M.; Hawkins, P. C. D.; Ermondi, G.; Kihlberg, J. Conformational Sampling of Macrocyclic Drugs in Different Environments: Can We Find the Relevant Conformations? *ACS Omega* **2018**, *3* (9), 11742–11757.

(78) Han, X.; Sun, Y. Strategies for the Discovery of Oral PROTAC Degraders Aimed at Cancer Therapy. *Cell Rep Phys Sci* **2022**, *3* (10), 101062.

(79) Desantis, J.; Mammoli, A.; Eleuteri, M.; Coletti, A.; Croci, F.; Macchiarulo, A.; Goracci, L. PROTACs Bearing Piperazine-Containing Linkers: What Effect on Their Protonation State? *RSC Adv* **2022**, *12* (34), 21968–21977.

(80) Ermondi, G.; Vallaro, M.; Saame, J.; Toom, L.; Leito, I.; Ruiz, R.; Caron, G. Rifampicin as an Example of Beyond-Rule-of-5 Compound: Ionization beyond Water and Lipophilicity beyond Octanol/Water. *European Journal of Pharmaceutical Sciences* **2021**, *161*, 105802.

(81) Vallaro, M.; Ermondi, G.; Saame, J.; Leito, I.; Caron, G. Ionization and Lipophilicity in Nonpolar Media Mimicking the Cell Membrane Interior. *Bioorg Med Chem* **2023**, *81*, 117203.
(82) Caron, G.; Vallaro, M.; Ermondi, G.; Goetz, G. H.; Abramov, Y. A.; Philippe, L.; Shalaeva, M. A Fast Chromatographic Method for Estimating Lipophilicity and Ionization in Nonpolar Membrane-Like Environment. *Mol Pharm* **2016**, *13* (3), 1100–1110.
(83) Testa, B.; Caron, G.; Crivori, P.; Rey, S.; Reist, M.; Carrupt, P. A. Lipophilicity and Related Molecular Properties as Determinants of Pharmacokinetic Behaviour. *Chimia (Aarau)* **2000**, *54* (11), 672–677.
(84) Andrés, A.; Rosés, M.; Ràfols, C.; Bosch, E.; Espinosa, S.; Segarra, V.; Huerta, J. M. Setup and Validation of Shake-Flask Procedures for the Determination of Partition Coefficients (LogD) from Low Drug Amounts. *European Journal of Pharmaceutical Sciences* **2015**, *76*, 181–191.
(85) Goetz, G. H.; Shalaeva, M. Leveraging Chromatography Based Physicochemical Properties for Efficient Drug Design. *ADMET DMPK* **2018**, *6* (2), 85–104.
(86) Soares, J. X.; Santos, Á.; Fernandes, C.; Pinto, M. M. M. Liquid Chromatography on the Different Methods for the Determination of Lipophilicity: An Essential Analytical Tool in Medicinal Chemistry. *Chemosensors* **2022**, *10* (8), 340.
(87) Avdeef, A. Assessment of Distribution-pH Profiles; 1996; pp 109–139.
(88) Valko, K.; Du, C. M.; Bevan, C. D.; Reynolds, D. P.; Abraham, M. H. Rapid-Gradient HPLC Method for Measuring Drug Interactions with Immobilized Artificial Membrane: Comparison with Other Lipophilicity Measures. *J Pharm Sci* **2000**, *89* (8), 1085–1096.
(89) Valkó, K.; Bevan, C.; Reynolds, D. Chromatographic Hydrophobicity Index by Fast-Gradient RP-HPLC: A High-Throughput Alternative to Log P/Log D. *Anal Chem* **1997**, *69* (11), 2022–2029.
(90) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F. ElogD o Ct : A Tool for Lipophilicity Determination in Drug Discovery. 2. Basic and Neutral Compounds. *J Med Chem* **2001**, *44* (15), 2490–2497.
(91) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F.; Abraham, M. H. ElogP Oct : A Tool for Lipophilicity Determination in Drug Discovery. *J Med Chem* **2000**, *43* (15), 2922–2928.
(92) Ermondi, G.; Vallaro, M.; Goetz, G.; Shalaeva, M.; Caron, G. Experimental Lipophilicity for beyond Rule of 5 Compounds. *Future Drug Discov* **2019**, *1* (1), FDD10.
(93) Dias, N. C.; Nawas, M. I.; Poole, C. F. Evaluation of a Reversed-Phase Column (Supelcosil LC-ABZ) under Isocratic and Gradient Elution Conditions for Estimating Octanol–Water Partition Coefficients. *Analyst* **2003**, *128* (5), 427–433.
(94) Caron, G.; Vallaro, M.; Ermondi, G. The Block Relevance (BR) Analysis to Aid Medicinal Chemists to Determine and Interpret Lipophilicity. *Medchemcomm* **2013**, *4* (10), 1376.
(95) Ermondi, G.; Vallaro, M.; Goetz, G.; Shalaeva, M.; Caron, G. Updating the Portfolio of Physicochemical Descriptors Related to Permeability in the beyond the Rule of 5 Chemical Space. *European Journal of Pharmaceutical Sciences* **2020**, *146*, 105274.

(96) Ong, S.; Liu, H.; Pidgeon, C. Immobilized-Artificial-Membrane Chromatography: Measurements of Membrane Partition Coefficient and Predicting Drug Membrane Permeability. *J Chromatogr A* **1996**, *728* (1–2), 113–128.
(97) Ermondi, G.; Vallaro, M.; Caron, G. Learning How to Use IAM Chromatography for Predicting Permeability. *European Journal of Pharmaceutical Sciences* **2018**, *114*, 385–390.
(98) Grumetto, L.; Carpentiero, C.; Barbato, F. Lipophilic and Electrostatic Forces Encoded in IAM-HPLC Indexes of Basic Drugs: Their Role in Membrane Partition and Their Relationships with BBB Passage Data. *European Journal of Pharmaceutical Sciences* **2012**, *45* (5), 685–692.
(99) Grumetto, L.; Russo, G.; Barbato, F. Polar Interactions Drug/Phospholipids Estimated by IAM-HPLC vs Cultured Cell Line Passage Data: Their Relationships and Comparison of Their Effectiveness in Predicting Drug Human Intestinal Absorption. *Int J Pharm* **2016**, *500* (1–2), 275–290.
(100) van Balen, G. P.; Martinet, C. a M.; Caron, G.; Bouchard, G.; Reist, M.; Carrupt, P.-A.; Fruttero, R.; Gasco, A.; Testa, B. Liposome/Water Lipophilicity: Methods, Information Content, and Pharmaceutical Applications. *Med Res Rev* **2004**, *24* (3), 299–324.
(101) Taillardat-Bertschinger, A.; Carrupt, P.-A.; Barbato, F.; Testa, B. Immobilized Artificial Membrane HPLC in Drug Research. *J Med Chem* **2003**, *46* (5), 655–665.
(102) Gramse, G.; Dols-Perez, A.; Edwards, M. A.; Fumagalli, L.; Gomila, G. Nanoscale Measurement of the Dielectric Constant of Supported Lipid Bilayers in Aqueous Solutions with Electrostatic Force Microscopy. *Biophys J* **2013**, *104* (6), 1257–1262.
(103) Li, L.; Li, C.; Zhang, Z.; Alexov, E. On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *J Chem Theory Comput* **2013**, *9* (4), 2126–2136.
(104) Caron, G.; Vallaro, M.; Ermondi, G. Log P as a Tool in Intramolecular Hydrogen Bond Considerations. *Drug Discov Today Technol* **2018**, *27*, 65–70.
(105) Shalaeva, M.; Caron, G.; Abramov, Y. A.; O'Connell, T. N.; Plummer, M. S.; Yalamanchi, G.; Farley, K. A.; Goetz, G. H.; Philippe, L.; Shapiro, M. J. Integrating Intramolecular Hydrogen Bonding (IMHB) Considerations in Drug Discovery Using ΔlogP as a Tool. *J Med Chem* **2013**, *56* (12), 4870–4879.
(106) Rossi Sebastiano, M.; Garcia Jimenez, D.; Vallaro, M.; Caron, G.; Ermondi, G. Refinement of Computational Access to Molecular Physicochemical Properties: From Ro5 to BRo5. *J Med Chem* **2022**, *65* (18), 12068–12083.
(107) Ermondi, G.; Caron, G. Block Relevance (BR) Analysis and Polarity Descriptors in Property-Based Drug Design. *ADMET DMPK* **2018**, *6* (3), 215–224.
(108) Goetz, G. H.; Philippe, L.; Shapiro, M. J. EPSA: A Novel Supercritical Fluid Chromatography Technique Enabling the Design of Permeable Cyclic Peptides. *ACS Med Chem Lett* **2014**, *5* (10), 1167–1172.
(109) Goetz, G. H.; Farrell, W.; Shalaeva, M.; Sciabola, S.; Anderson, D.; Yan, J.; Philippe, L.; Shapiro, M. J. High Throughput Method for the Indirect Detection of Intramolecular Hydrogen Bonding. *J Med Chem* **2014**, *57* (7), 2920–2929.
(110) Barbato, F.; La Rotonda, M. I.; Quaglia, F. Interactions of Nonsteroidal Antiinflammatory Drugs with Phospholipids: Comparison between

Octanol/Buffer Partition Coefficients and Chromatographic Indexes on Immobilized Artificial Membranes. *J Pharm Sci* **1997**, *86* (2), 225–229.
(111) Barbato, F.; Cappello, B.; Miro, A.; La Rotonda, M.; Quaglia, F. Chromatographic Indexes on Immobilized Artificial Membranes for the Prediction of Transdermal Transport of Drugs. *Il Farmaco* **1998**, *53* (10–11), 655–661.
(112) Ermondi, G.; Vallaro, M.; Caron, G. Degraders Early Developability Assessment: Face-to-Face with Molecular Properties. *Drug Discov Today* **2020**, *25* (9), 1585–1591.
(113) Ermondi, G.; Lavore, F.; Vallaro, M.; Tiana, G.; Vasile, F.; Caron, G. Managing Experimental 3D Structures in the Beyond-Rule-of-5 Chemical Space: The Case of Rifampicin. *Chemistry – A European Journal* **2021**, *27* (40), 10394–10404.
(114) Ono, S.; Naylor, M. R.; Townsend, C. E.; Okumura, C.; Okada, O.; Lee, H.-W.; Lokey, R. S. Cyclosporin A: Conformational Complexity and Chameleonicity. *J Chem Inf Model* **2021**, *61* (11), 5601–5613.
(115) Witek, J.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. Kinetic Models of Cyclosporin A in Polar and Apolar Environments Reveal Multiple Congruent Conformational States. *J Chem Inf Model* **2016**, *56* (8), 1547–1562.
(116) Lautz, J.; Kessler, H.; Kaptein, R.; van Gunsteren, W. F. Molecular Dynamics Simulations of Cyclosporin A: The Crystal Structure and Dynamic Modelling of a Structure in Apolar Solution Based on NMR Data. *J Comput Aided Mol Des* **1987**, *1* (3), 219–241.
(117) Erdélyi, M.; Pfeiffer, B.; Hauenstein, K.; Fohrer, J.; Gertsch, J.; Altmann, K.-H.; Carlomagno, T. Conformational Preferences of Natural and C3-Modified Epothilones in Aqueous Solution. *J Med Chem* **2008**, *51* (5), 1469–1473.
(118) Nevins, N.; Cicero, D.; Snyder, J. P. A Test of the Single-Conformation Hypothesis in the Analysis of NMR Data for Small Polar Molecules: A Force Field Comparison. *Journal of Organic Chemistry* **1999**, *64* (11), 3979–3986.
(119) Cicero, D. O.; Barbato, G.; Bazzo, R. NMR Analysis of Molecular Flexibility in Solution: A New Method for the Study of Complex Distributions of Rapidly Exchanging Conformations. Application to a 13-Residue Peptide with an 8-Residue Loop. *J Am Chem Soc* **1995**, *117* (3), 1027–1033.
(120) Danelius, E.; Poongavanam, V.; Peintner, S.; Wieske, L. H. E.; Erdélyi, M.; Kihlberg, J. Solution Conformations Explain the Chameleonic Behaviour of Macrocyclic Drugs. *Chemistry – A European Journal* **2020**, *26* (23), 5231–5244.
(121) Garcia Jimenez, D.; Vallaro, M.; Rossi Sebastiano, M.; Apprato, G.; D'Agostini, G.; Rossetti, P.; Ermondi, G.; Caron, G. Chamelogk: A Chromatographic Chameleonicity Quantifier to Design Orally Bioavailable Beyond-Rule-of-5 Drugs. *J Med Chem* **2023**, *66* (15), 10681–10693.
(122) Barrett, J. A.; Yang, W.; Skolnik, S. M.; Belliveau, L. M.; Patros, K. M. Discovery Solubility Measurement and Assessment of Small Molecules with Drug Development in Mind. *Drug Discov Today* **2022**.
(123) Saal, C.; Petereit, A. C. Optimizing Solubility: Kinetic versus Thermodynamic Solubility Temptations and Risks. *European Journal of Pharmaceutical Sciences* **2012**, *47* (3), 589–595.
(124) Palumbo, P.; Picchini, U.; Beck, B.; van Gelder, J.; Delbar, N.; DeGaetano, A. A General Approach to the Apparent Permeability Index. *J Pharmacokinet Pharmacodyn* **2008**, *35* (2), 235–248.

(125) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. *J Med Chem* **1998**, *41* (7), 1007–1010.
(126) Cantrill, C.; Chaturvedi, P.; Rynn, C.; Petrig Schaffland, J.; Walter, I.; Wittwer, M. B. Fundamental Aspects of DMPK Optimization of Targeted Protein Degraders. *Drug Discov Today* **2020**, *25* (6), 969–982.
(127) Jain, R.; Agarwal, S.; Majumdar, S.; Zhu, X.; Pal, D.; Mitra, A. K. Evasion of P-Gp Mediated Cellular Efflux and Permeability Enhancement of HIV-Protease Inhibitor Saquinavir by Prodrug Modification. *Int J Pharm* **2005**, *303* (1–2), 8–19.
(128) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry; Wiley, 2000.
(129) MORIGUCHI, I.; HIRONO, S.; NAKAGOME, I.; HIRANO, H. Comparison of Reliability of Log P Values for Drugs Calculated by Several Methods. *Chem Pharm Bull (Tokyo)* **1994**, *42* (4), 976–978.
(130) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J Pharm Sci* **1999**, *88* (8), 807–814.
(131) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J Med Chem* **2000**, *43* (20), 3714–3717.
(132) Verma, R. P.; Kurup, A.; Hansch, C. On the Role of Polarizability in QSAR. *Bioorg Med Chem* **2005**, *13* (1), 237–255.
(133) Taft, R. W.; Kamlet, M. J. The Solvatochromic Comparison Method. 2. The .Alpha.-Scale of Solvent Hydrogen-Bond Donor (HBD) Acidities. *J Am Chem Soc* **1976**, *98* (10), 2886–2894.
(134) Abraham, M. H. Hydrogen Bonding. 31. Construction of a Scale of Solute Effective or Summation Hydrogen-bond Basicity. *J Phys Org Chem* **1993**, *6* (12), 660–684.
(135) Abraham, M. H. Scales of Solute Hydrogen-Bonding: Their Construction and Application to Physicochemical and Biochemical Processes. *Chem Soc Rev* **1993**, *22* (2), 73.
(136) Rouvray, D. H.; King, R. B. *Topology in Chemistry: Discrete Mathematics of Molecules*; Horwood Publishing Series in Chemical Science; Elsevier Science, 2002.
(137) Oprea, T. I. On the Information Content of 2D and 3D Descriptors for QSAR. *J Braz Chem Soc* **2002**, *13* (6).
(138) Bahia, M. S.; Kaspi, O.; Touitou, M.; Binayev, I.; Dhail, S.; Spiegel, J.; Khazanov, N.; Yosipof, A.; Senderowitz, H. A Comparison between 2D and 3D Descriptors in QSAR Modeling Based on Bio-active Conformations. *Mol Inform* **2023**, *42* (4).
(139) Guimarães, C. R. W.; Mathiowetz, A. M.; Shalaeva, M.; Goetz, G.; Liras, S. Use of 3D Properties to Characterize Beyond Rule-of-5 Property Space for Passive Permeation. *J Chem Inf Model* **2012**, *52* (4), 882–890.
(140) Rezai, T.; Yu, B.; Millhauser, G. L.; Jacobson, M. P.; Lokey, R. S. Testing the Conformational Hypothesis of Passive Membrane Permeability Using Synthetic Cyclic Peptide Diastereomers. *J Am Chem Soc* **2006**, *128* (8), 2510–2511.
(141) Rezai, T.; Bock, J. E.; Zhou, M. V.; Kalyanaraman, C.; Lokey, R. S.; Jacobson, M. P. Conformational Flexibility, Internal Hydrogen Bonding, and Passive Membrane

Permeability: Successful in Silico Prediction of the Relative Permeabilities of Cyclic Peptides. *J Am Chem Soc* **2006**, *128* (43), 14073–14080.
(142) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J Chem Inf Model* **2012**, *52* (11), 2919–2936.
(143) Wang, Q.; Sciabola, S.; Barreiro, G.; Hou, X.; Bai, G.; Shapiro, M. J.; Koehn, F.; Villalobos, A.; Jacobson, M. P. Dihedral Angle-Based Sampling of Natural Product Polyketide Conformations: Application to Permeability Prediction. *J Chem Inf Model* **2016**, *56* (11), 2194–2206.
(144) Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. Macrocycle Conformational Sampling with MacroModel. *J Chem Inf Model* **2014**, *54* (10), 2680–2696.
(145) Olanders, G.; Alogheli, H.; Brandt, P.; Karlén, A. Conformational Analysis of Macrocycles: Comparing General and Specialized Methods. *J Comput Aided Mol Des* **2020**, *34* (3), 231–252.
(146) Jorgensen, W. L.; Tirado-Rives, J. Monte Carlo vs Molecular Dynamics for Conformational Sampling. *J Phys Chem* **1996**, *100* (34), 14508–14513.
(147) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J Med Chem* **2016**, *59* (9), 4035–4061.
(148) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol* **2011**, *9* (1), 71.
(149) Fehér, P. P.; Stirling, A. Assessment of Reactivities with Explicit and Implicit Solvent Models: QM/MM and Gas-Phase Evaluation of Three Different Ag-Catalysed Furan Ring Formation Routes. *New Journal of Chemistry* **2019**, *43* (39), 15706–15713.
(150) Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A. V. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophys J* **2015**, *108* (5), 1153–1164.
(151) Wei, G.; Sun, Y.; Zhou, Y. C.; Feig, M. Molecular Multiresolution Surfaces. *arXiv: Mathematical Physics* **2005**.
(152) Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. A Computational Pathway for Bracketing Native-like Structures for Small Alpha Helical Globular Proteins. *Physical Chemistry Chemical Physics* **2005**, *7* (11), 2364.
(153) Sauer, W. H. B.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J Chem Inf Comput Sci* **2003**, *43* (3), 987–1003.
(154) Mills, J. E. J.; Dean, P. M. Three-Dimensional Hydrogen-Bond Geometry and Probability Information from a Crystal Survey. *J Comput Aided Mol Des* **1996**, *10* (6), 607–622.
(155) Medina-Franco, J.; Martinez-Mayorga, K.; Giulianotti, M.; Houghten, R.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Current Computer Aided-Drug Design* **2008**, *4* (4), 322–333.
(156) Medina-Franco, J. L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B. I. Progress on Open Chemoinformatic Tools for Expanding and Exploring the Chemical Space. *J Comput Aided Mol Des* **2022**, *36* (5), 341–354.
(157) Karlov, D. S.; Sosnin, S.; Tetko, I. V.; Fedorov, M. V. Chemical Space Exploration Guided by Deep Neural Networks. *RSC Adv* **2019**, *9* (9), 5151–5157.

(158) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov Today* **2006**, *11* (23–24), 1046–1053.
(159) Schwartz, J.; Awale, M.; Reymond, J.-L. SMIfp (SMILES Fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules. *J Chem Inf Model* **2013**, *53* (8), 1979–1989.
(160) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J Med Chem* **2012**, *55* (7), 2932–2942.
(161) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure−Activity Relationship Analysis. *J Med Chem* **2010**, *53* (23), 8209–8223.
(162) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem Rev* **1996**, *96* (3), 1027–1044.
(163) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem Soc Rev* **1995**, *24* (4), 279.
(164) Polishchuk, P. Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. *J Chem Inf Model* **2017**, *57* (11), 2618–2639.
(165) Soares, T. A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G.-W.; Merz, K. The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. *J Chem Inf Model* **2022**, *62* (22), 5317–5320.
(166) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J Chem Inf Model* **2015**, *55* (2), 263–274.
(167) Yalkowsky, S. H.; Valvani, S. C. Solubility and Partitioning I: Solubility of Nonelectrolytes in Water. *J Pharm Sci* **1980**, *69* (8), 912–922.
(168) Dongbin, W.; Aiqian, Z.; Zhongbo, W.; Shuokui, H.; Liansheng, W. A Case Study of Logistic QSAR Modeling Methods and Robustness Tests. *Ecotoxicol Environ Saf* **2002**, *52* (2), 143–149.
(169) U.S. Food and Drug Administration: Approved Drugs. Https://Www.Accessdata.Fda.Gov/Scripts/Cder/Daf/Index.Cfm (Accessed September 1, 2022).
(170) Clinical Trials Database. National Library of Medicine (US). Https://Clinicaltrials.Gov/ (Accesed November 30, 2022).
(171) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res* **2019**, *47* (D1), D930–D940.
(172) Wishart, D. S. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res* **2006**, *34* (90001), D668–D672.
(173) Corbett, K. M.; Ford, L.; Warren, D. B.; Pouton, C. W.; Chalmers, D. K. Cyclosporin Structure and Permeability: From A to Z and Beyond. *J Med Chem* **2021**, *64* (18), 13131–13151.
(174) Maggio, E. T.; Grasso, P. Oral Delivery of Octreotide Acetate in Intravail® Improves Uptake, Half-Life, and Bioavailability over Subcutaneous Administration in Male Swiss Webster Mice. *Regul Pept* **2011**, *167* (2–3), 233–238.

(175) Rembratt, A.; Graugaard-Jensen, C.; Senderovitz, T.; Norgaard, J. P.; Djurhuus, J. C. Pharmacokinetics and Pharmacodynamics of Desmopressin Administered Orally versus Intravenously at Daytime versus Night-Time in Healthy Men 55-70 Years. *Eur J Clin Pharmacol* **2004**, *60* (6), 397–402.
(176) McHugh, M. L. Interrater Reliability: The Kappa Statistic. *Biochem Med (Zagreb)* **2012**, *22* (3), 276–282.
(177) Kenny, P. W. Hydrogen-Bond Donors in Drug Design. *J Med Chem* **2022**, *65* (21), 14261–14275.
(178) Grigalunas, M.; Brakmann, S.; Waldmann, H. Chemical Evolution of Natural Product Structure. *J Am Chem Soc* **2022**, *144* (8), 3314–3329.
(179) Stratton, C. F.; Newman, D. J.; Tan, D. S. Cheminformatic Comparison of Approved Drugs from Natural Product versus Synthetic Origins. *Bioorg Med Chem Lett* **2015**, *25* (21), 4802–4807.
(180) Ermondi, G.; Garcia-Jimenez, D.; Caron, G. Protacs and Building Blocks: The 2d Chemical Space in Very Early Drug Discovery. *Molecules* **2021**, *26* (3), 672.
(181) *National Library of Medicine (US). https://www.ncbi.nlm.nih.gov/pubmed/. (Accessed December, 2020).*
(182) Troup, R. I.; Fallan, C.; Baud, M. G. J. Current Strategies for the Design of PROTAC Linkers: A Critical Review. *Explor Target Antitumor Ther* **2020**, *1* (5), 273–312.
(183) Edmondson, S. D.; Yang, B.; Fallan, C. Proteolysis Targeting Chimeras (PROTACs)in 'beyond Rule-of-Five' Chemical Space: Recent Progress and Future Challenges. *Bioorg Med Chem Lett* **2019**, *29* (13), 1555–1564.
(184) Weng, G.; Shen, C.; Cao, D.; Gao, J.; Dong, X.; He, Q.; Yang, B.; Li, D.; Wu, J.; Hou, T. PROTAC-DB: An Online Database of PROTACs. *Nucleic Acids Res* **2021**, *49* (D1), D1381–D1387.
(185) Tyagi, M.; Begnini, F.; Poongavanam, V.; Doak, B. C.; Kihlberg, J. Drug Syntheses Beyond the Rule of 5. *Chemistry - A European Journal* **2020**, *26* (1), 49–88.
(186) Pike, A.; Williamson, B.; Harlfinger, S.; Martin, S.; McGinnity, D. F. Optimising Proteolysis-Targeting Chimeras (PROTACs) for Oral Drug Delivery: A Drug Metabolism and Pharmacokinetics Perspective. *Drug Discov Today* **2020**, *25* (10), 1793–1800.
(187) Apprato, G.; Ermondi, G.; Caron, G. The Quest for Oral PROTAC Drugs: Evaluating the Weaknesses of the Screening Pipeline. *ACS Med Chem Lett* **2023**, *14* (7), 879–883.
(188) Galdeano, C.; Gadd, M. S.; Soares, P.; Scaffidi, S.; Van Molle, I.; Birced, I.; Hewitt, S.; Dias, D. M.; Ciulli, A. Structure-Guided Design and Optimization of Small Molecules Targeting the Protein–Protein Interaction between the von Hippel–Lindau (VHL) E3 Ubiquitin Ligase and the Hypoxia Inducible Factor (HIF) Alpha Subunit with in Vitro Nanomolar Affinities. *J Med Chem* **2014**, *57* (20), 8657–8663.
(189) Frost, J.; Galdeano, C.; Soares, P.; Gadd, M. S.; Grzes, K. M.; Ellis, L.; Epemolu, O.; Shimamura, S.; Bantscheff, M.; Grandi, P.; Read, K. D.; Cantrell, D. A.; Rocha, S.; Ciulli, A. Potent and Selective Chemical Probe of Hypoxic Signalling Downstream of HIF-α Hydroxylation via VHL Inhibition. *Nat Commun* **2016**, *7* (1), 13312.
(190) Galdeano, C. Drugging the Undruggable: Targeting Challenging E3 Ligases for Personalized Medicine. *Future Med Chem* **2017**, *9* (4), 347–350.
(191) Poongavanam, V.; Kihlberg, J. PROTAC Cell Permeability and Oral Bioavailability: A Journey into Uncharted Territory. *Future Med Chem* **2022**, *14* (3), 123–126.

(192) Scott, D. E.; Rooney, T. P. C.; Bayle, E. D.; Mirza, T.; Willems, H. M. G.; Clarke, J. H.; Andrews, S. P.; Skidmore, J. Systematic Investigation of the Permeability of Androgen Receptor PROTACs. *ACS Med Chem Lett* **2020**, *11* (8), 1539–1547.
(193) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J Phys Chem A* **1998**, *102* (21), 3762–3772.
(194) Bergström, C. A. S.; Holm, R.; Jørgensen, S. A.; Andersson, S. B. E.; Artursson, P.; Beato, S.; Borde, A.; Box, K.; Brewster, M.; Dressman, J.; Feng, K.-I.; Halbert, G.; Kostewicz, E.; McAllister, M.; Muenster, U.; Thinnes, J.; Taylor, R.; Mullertz, A. Early Pharmaceutical Profiling to Predict Oral Drug Absorption: Current Status and Unmet Needs. *European Journal of Pharmaceutical Sciences* **2014**, *57*, 173–199.
(195) Cheney, M. L.; Shan, N.; Healey, E. R.; Hanna, M.; Wojtas, L.; Zaworotko, M. J.; Sava, V.; Song, S.; Sanchez-Ramos, J. R. Effects of Crystal Form on Solubility and Pharmacokinetics: A Crystal Engineering Case Study of Lamotrigine. *Cryst Growth Des* **2010**, *10* (1), 394–405.
(196) Bergström, C. A. S.; Larsson, P. Computational Prediction of Drug Solubility in Water-Based Systems: Qualitative and Quantitative Approaches Used in the Current Drug Discovery and Development Setting. *Int J Pharm* **2018**, *540* (1–2), 185–193.
(197) Elder, D. P.; Holm, R.; Diego, H. L. de. Use of Pharmaceutical Salts and Cocrystals to Address the Issue of Poor Solubility. *Int J Pharm* **2013**, *453* (1), 88–100.
(198) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes. *J Pharm Sci* **2001**, *90* (2), 234–252.
(199) Abramov, Y. A.; Sun, G.; Zeng, Q.; Zeng, Q.; Yang, M. Guiding Lead Optimization for Solubility Improvement with Physics-Based Modeling. *Mol Pharm* **2020**, *17* (2), 666–673.
(200) Ermondi, G.; Vallaro, M.; Caron, G.; Poongavanam, V.; Kihlberg, J. Solubility Prediction in the BRo5 Chemical Space: Where Are We Right Now? *ADMET DMPK* **2020**, *8* (3), 207–214.
(201) Avdeef, A.; Kansy, M. "Flexible-Acceptor" General Solubility Equation for Beyond Rule of 5 Drugs. *Mol Pharm* **2020**, *17* (10), 3930–3940.
(202) Bard, B.; Martel, S.; Carrupt, P.-A. High Throughput UV Method for the Estimation of Thermodynamic Solubility and the Determination of the Solubility in Biorelevant Media. *European Journal of Pharmaceutical Sciences* **2008**, *33* (3), 230–240.
(203) Kádár, S.; Csicsák, D.; Tőzsér, P.; Farkas, A.; Pálla, T.; Mirzahosseini, A.; Tóth, B.; Tóth, G.; Fiser, B.; Horváth, P.; Madarász, J.; Avdeef, A.; Takács-Novák, K.; Sinkó, B.; Borbás, E.; Völgyi, G. Understanding the PH Dependence of Supersaturation State—A Case Study of Telmisartan. *Pharmaceutics* **2022**, *14* (8), 1635.
(204) Zengerle, M.; Chan, K. H.; Ciulli, A. Selective Small Molecule Induced Degradation of the BET Bromodomain Protein BRD4. *ACS Chem Biol* **2015**, *10* (8), 1770–1777.
(205) Chan, K. H.; Zengerle, M.; Testa, A.; Ciulli, A. Impact of Target Warhead and Linkage Vector on Inducing Protein Degradation: Comparison of Bromodomain and Extra-Terminal (BET) Degraders Derived from Triazolodiazepine (JQ1) and Tetrahydroquinoline (I-BET726) BET Inhibitor Scaffolds. *J Med Chem* **2018**, *61* (2), 504–513.

(206) Nowak, R. P.; Deangelo, S. L.; Buckley, D.; He, Z.; Donovan, K. A.; An, J.; Safaee, N.; Jedrychowski, M. P.; Ponthier, C. M.; Ishoey, M.; Zhang, T.; Mancias, J. D.; Gray, N. S.; Bradner, J. E.; Fischer, E. S. Plasticity in Binding Confers Selectivity in Ligand-Induced Protein Degradation Article. *Nat Chem Biol* **2018**, *14* (7), 706–714.

(207) Popow, J.; Arnhof, H.; Bader, G.; Berger, H.; Ciulli, A.; Covini, D.; Dank, C.; Gmaschitz, T.; Greb, P.; Karolyi-Özguer, J.; Koegl, M.; McConnell, D. B.; Pearson, M.; Rieger, M.; Rinnenthal, J.; Roessler, V.; Schrenk, A.; Spina, M.; Steurer, S.; Trainor, N.; Traxler, E.; Wieshofer, C.; Zoephel, A.; Ettmayer, P. Highly Selective PTK2 Proteolysis Targeting Chimeras to Probe Focal Adhesion Kinase Scaffolding Functions. *J Med Chem* **2019**, *62* (5), 2508–2520.

(208) Linker, S. M.; Schellhaas, C.; Kamenik, A. S.; Veldhuizen, M. M.; Waibl, F.; Roth, H. J.; Fouché, M.; Rodde, S.; Riniker, S. Lessons for Oral Bioavailability: How Conformationally Flexible Cyclic Peptides Enter and Cross Lipid Membranes. *J Med Chem* **2023**, *66* (4), 2773–2788.

(209) Sethio, D.; Poongavanam, V.; Xiong, R.; Tyagi, M.; Duy Vo, D.; Lindh, R.; Kihlberg, J. Simulation Reveals the Chameleonic Behavior of Macrocycles. *J Chem Inf Model* **2023**, *63* (1), 138–146.

(210) Ermondi, G.; Jimenez, D. G.; Rossi Sebastiano, M.; Kihlberg, J.; Caron, G. Conformational Sampling Deciphers the Chameleonic Properties of a VHL-Based Degrader. *Pharmaceutics* **2023**, *15* (1), 272.

(211) Wang, C. K.; Swedberg, J. E.; Harvey, P. J.; Kaas, Q.; Craik, D. J. Conformational Flexibility Is a Determinant of Permeability for Cyclosporin. *J Phys Chem B* **2018**, *122* (8), 2261–2276.

(212) Amidon, G. L.; Lennernäs, H.; Shah, V. P.; Crison, J. R. A Theoretical Basis for a Biopharmaceutic Drug Classification: The Correlation of in Vitro Drug Product Dissolution and in Vivo Bioavailability. *Pharmaceutical Research: An Official Journal of the American Association of Pharmaceutical Scientists* **1995**, *12* (3), 413–420.

(213) Scott, G.; Osborne, S. A.; Greig, G.; Hartmann, S.; Ebelin, M.-E.; Burtin, P.; Rappersberger, K.; Komar, M.; Wolff, K. Pharmacokinetics of Pimecrolimus, a Novel Nonsteroid Anti-Inflammatory Drug, After Single and Multiple Oral Administration. *Clin Pharmacokinet* **2003**, *42* (14), 1305–1314.

(214) Ismailos, G.; Reppas, C.; Dressman, J. B.; Macheras, P. Unusual Solubility Behaviour of Cyclosporin A in Aqueous Media. *Journal of Pharmacy and Pharmacology* **2011**, *43* (4), 287–289.

(215) Kirchner, G. I.; Meier-Wiedenbach, I.; Manns, M. P. Clinical Pharmacokinetics of Everolimus. *Clin Pharmacokinet* **2004**, *43* (2), 83–95.

(216) He, L.; Chen, C.; Gao, G.; Xu, K.; Ma, Z. ARV-825-Induced BRD4 Protein Degradation as a Therapy for Thyroid Carcinoma. *Aging* **2020**, *12* (5), 4547–4557.

(217) *BET PROTAC | MZ1; opnme https://www.opnme.com/molecules/bet-mz-1 (accessed May 4, 2023)*.

(218) Poongavanam, V.; Atilaw, Y.; Ye, S.; Wieske, L. H. E.; Erdelyi, M.; Ermondi, G.; Caron, G.; Kihlberg, J. Predicting the Permeability of Macrocycles from Conformational Sampling – Limitations of Molecular Flexibility. *J Pharm Sci* **2021**, *110* (1), 301–313.

(219) Digiesi, V.; de la Oliva Roque, V.; Vallaro, M.; Caron, G.; Ermondi, G. Permeability Prediction in the Beyond-Rule-of 5 Chemical Space: Focus on Cyclic Hexapeptides. *European Journal of Pharmaceutics and Biopharmaceutics* **2021**, *165*, 259–270.
(220) Drummond, M. L.; Williams, C. I. In Silico Modeling of PROTAC-Mediated Ternary Complexes: Validation and Application. *J Chem Inf Model* **2019**, *59* (4), 1634–1644.
(221) Bondeson, D. P.; Smith, B. E.; Burslem, G. M.; Buhimschi, A. D.; Hines, J.; Jaime-Figueroa, S.; Wang, J.; Hamman, B. D.; Ishchenko, A.; Crews, C. M. Lessons in PROTAC Design from Selective Degradation with a Promiscuous Warhead. *Cell Chem Biol* **2018**, *25* (1), 78-87.e5.
(222) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J Comput Phys* **1977**, *23* (2), 187–199.
(223) Chanan-Khan, A. A.; Swaika, A.; Paulus, A.; Kumar, S. K.; Mikhael, J. R.; Rajkumar, S. V; Dispenzieri, A.; Lacy, M. Q. Pomalidomide: The New Immunomodulatory Agent for the Treatment of Multiple Myeloma. *Blood Cancer J* **2013**, *3* (9), e143–e143.
(224) Sasso, J. M.; Tenchov, R.; Wang, D.; Johnson, L. S.; Wang, X.; Zhou, Q. A. Molecular Glues: The Adhesive Connecting Targeted Protein Degradation to the Clinic. *Biochemistry* **2023**, *62* (3), 601–623.
(225) Klein, V. G.; Townsend, C. E.; Testa, A.; Zengerle, M.; Maniaci, C.; Hughes, S. J.; Chan, K. H.; Ciulli, A.; Lokey, R. S. Understanding and Improving the Membrane Permeability of VH032-Based PROTACs. *ACS Med Chem Lett* **2020**, *11* (9), 1732–1738.
(226) Rocco, P.; Cilurzo, F.; Minghetti, P.; Vistoli, G.; Pedretti, A. Molecular Dynamics as a Tool for in Silico Screening of Skin Permeability. *European Journal of Pharmaceutical Sciences* **2017**, *106*, 328–335.
(227) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Methods* **2017**, *14* (1), 71–73.
(228) Marvin 17.21.0, ChemAxon (Https://Www.Chemaxon.Com).
(229) *Dragon, ver. 7.0.10; Kode Cheminformatics, https://chm.kodesolutions.net/pf/dragon-7-0/, 2017*.
(230) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J Chem Inf Model* **2015**, *55* (2), 460–473.
(231) *Molecular Operating Environment (MOE); Chemical Computing Group ULC: Montreal, QC, Canada, 2020.*
(232) Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. Data Mining: Practical Machine Learning Tools and Techniques. *Data Mining: Practical Machine Learning Tools and Techniques* **2016**, 1–621.
(233) Pedretti, A.; Villa, L.; Vistoli, G. VEGA: A Versatile Program to Convert, Handle and Visualize Molecular Structure on Windows-Based PCs. *J Mol Graph Model* **2002**, *21* (1), 47–49.
(234) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25* (13), 1605–1612.

(235) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J Comput Chem* **2008**, *29* (11), 1859–1865.
(236) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14* (1), 33–38.
(237) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J Comput Chem* **2005**, *26* (16), 1781–1802.
(238) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J Chem Inf Comput Sci* **2004**, *44* (1), 266–275.
(239) Pires, D. E. V.; Blundell, T. L.; Ascher, D. B. PkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J Med Chem* **2015**, *58* (9), 4066–4072.
(240) Dong, J.; Wang, N.-N.; Yao, Z.-J.; Zhang, L.; Cheng, Y.; Ouyang, D.; Lu, A.-P.; Cao, D.-S. ADMETlab: A Platform for Systematic ADMET Evaluation Based on a Comprehensively Collected ADMET Database. *J Cheminform* **2018**, *10* (1), 29.
(241) Wang, J.; Krudy, G.; Hou, T.; Zhang, W.; Holland, G.; Xu, X. Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis. *J Chem Inf Model* **2007**, *47* (4), 1395–1404.
(242) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. Molecular Fields in Quantitative Structure–Permeation Relationships: The VolSurf Approach. *Journal of Molecular Structure: THEOCHEM* **2000**, *503* (1–2), 17–30.
(243) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J Am Chem Soc* **1996**, *118* (45), 11225–11236.

## 13. Supplementary material

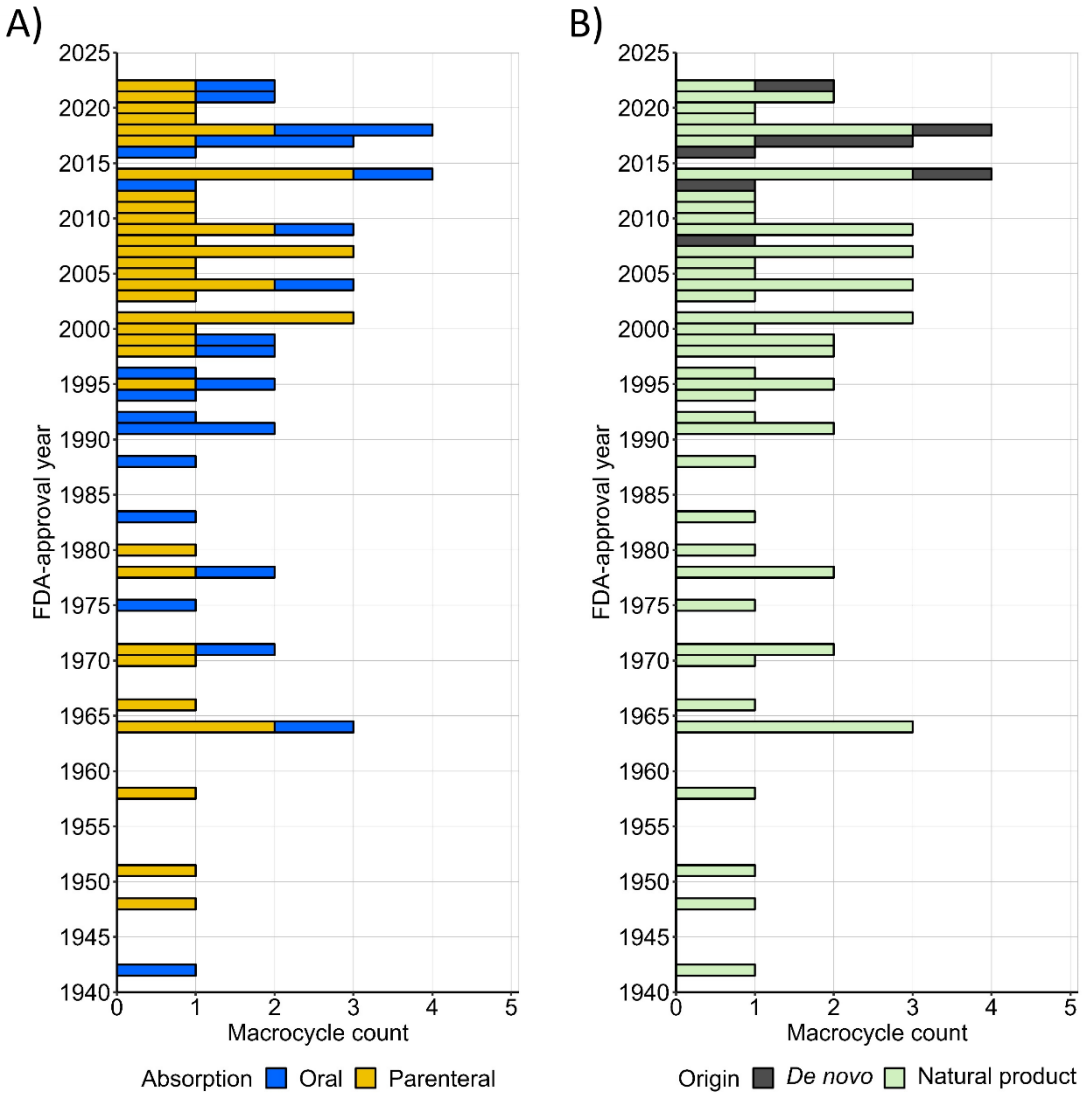



**Figure S1.** Number of macrocyclic drugs plotted by their year of approval by the FDA (n = 67, data retrieved on September 1, 2022). A) Orally absorbed drugs are indicated in blue (n = 26; 39%), while those administered parenterally are in gold (n = 41; 61%). B) Natural products and derivatives thereof are presented in light green (n = 59, 88%); *de novo* designed macrocyclic drugs are in dark gray (n = 8, 12%). Contrast agents, macrocycle-conjugated antibodies, PEG-linked macrocycles, and cyclodextrins have been excluded. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

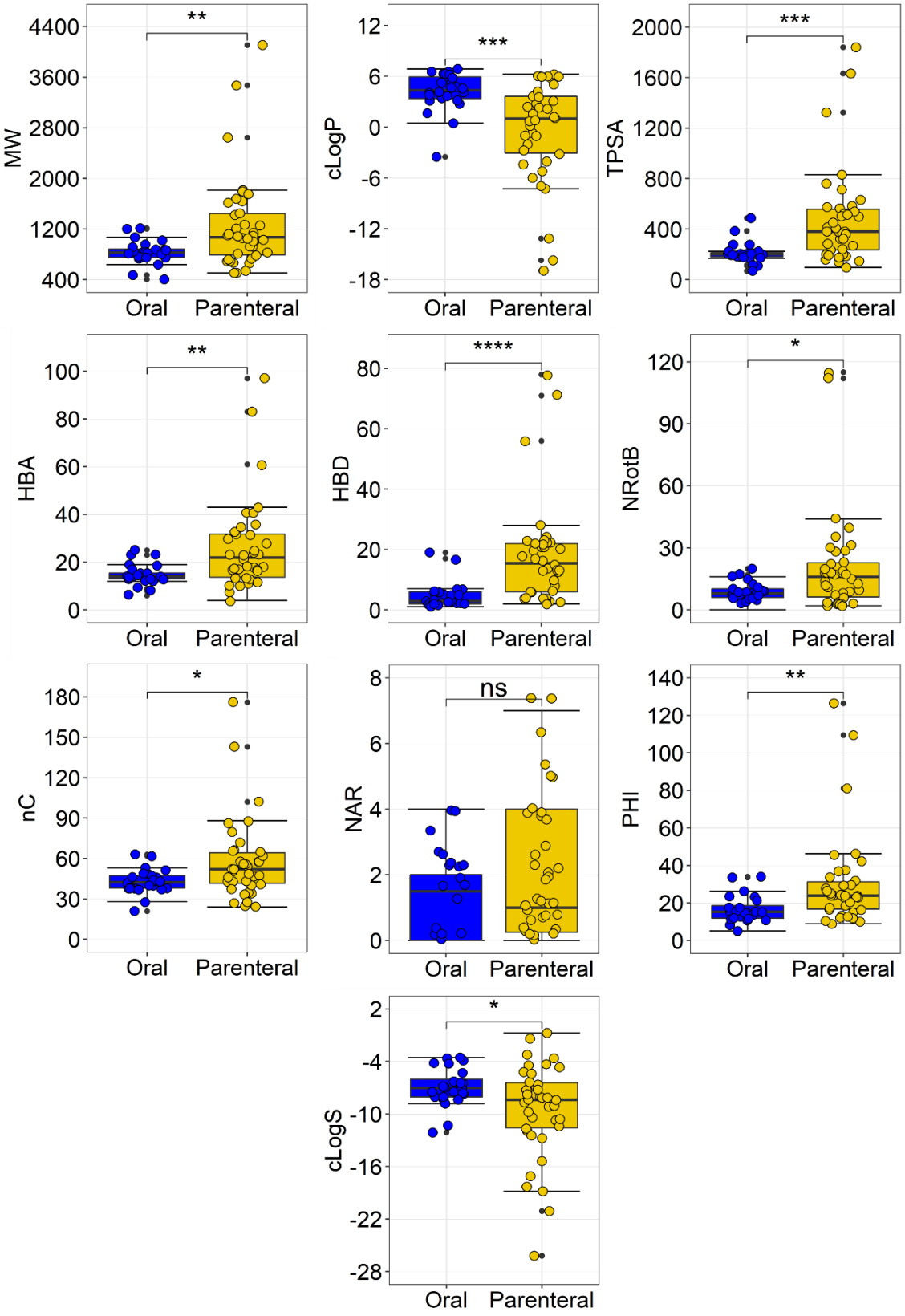

MW
4400
3600
2800
2000
1200
400
**
Oral
Parenteral
cLogP
12
6
0
-6
-12
-18
***
Oral
Parenteral
TPSA
2000
1600
1200
800
400
0
***
Oral
Parenteral
HBA
100
80
60
40
20
0
**
Oral
Parenteral
HBD
80
60
40
20
0
****
Oral
Parenteral
NRotB
120
90
60
30
0
*
Oral
Parenteral
nC
180
150
120
90
60
30
0
*
Oral
Parenteral
NAR
8
6
4
2
0
ns
Oral
Parenteral
PHI
140
120
100
80
60
40
20
0
**
Oral
Parenteral
cLogS
2
-4
-10
-16
-22
-28
*
Oral
Parenteral

**Figure S2.** Molecular property descriptors calculated at pH 7 for the macrocyclic drugs data set (n=62), split into two subsets by the route of administration. Those administered orally (n=24) are in blue, while parenterally administered are in gold. Statistical analysis was performed using Wilcoxon's non-parametric test: p-values: 0-0.0001 (****), 0.0001-0.001 (***), 0.001-0.01 (**), 0.01-0.05 (*), 0.05-1 (ns). Molecular property abbreviations are provided in the main text: Kier's flexibility index is abbreviated as PHI or (Φ). Box plots show the 50th percentiles as horizontal bars, the 25th and 75th percentiles as boxes, the 25th percentile minus 1.5 x the interquartile range and the 75th percentile plus 1.5 x the interquartile range as whiskers. Black dots represent values higher than 1.5 x the interquartile range and less than 3 x the interquartile range at either end of the box. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

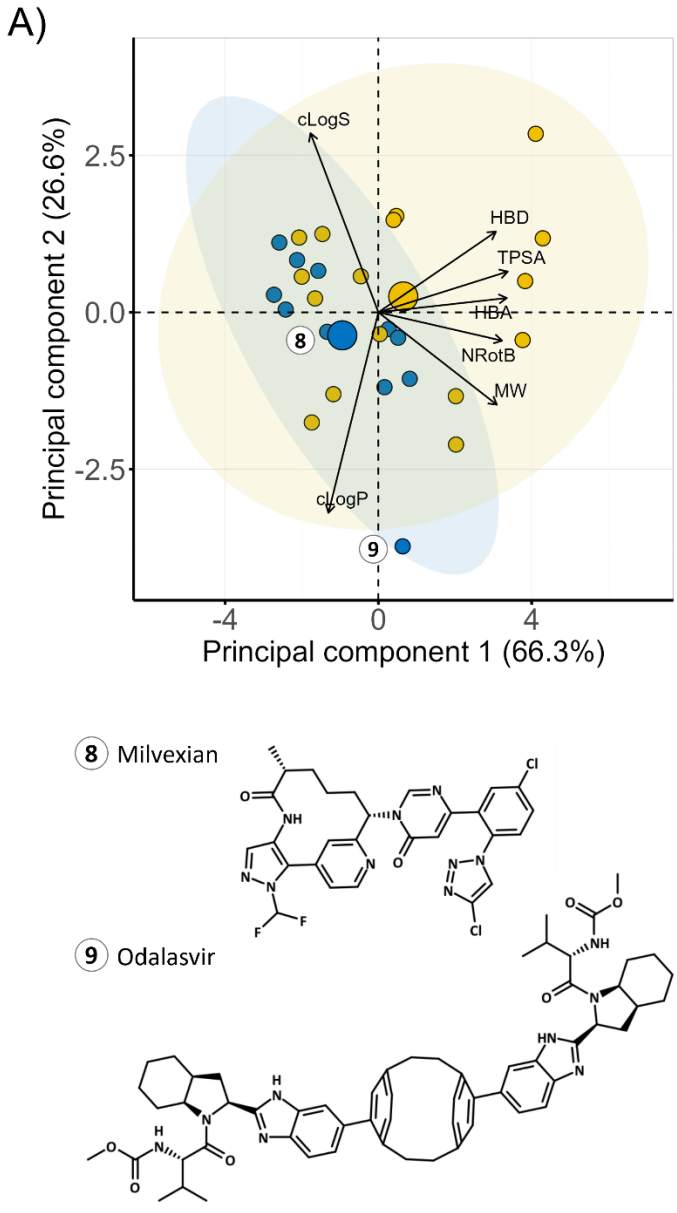


**Figure S3.** A) PCA of the chemical space of the clinical trials data set (n = 27). The PCA was based on the descriptors of Lipinski's and Veber's rules as well as cLogS calculated at pH 7.0. Five parenterals with MW > 1500 Da were excluded from the PCA to provide a better dissection of the chemical space of the orally bioavailable macrocycles. The ellipses in blue and yellow shading show the 95% confidence intervals for orally and parenterally administered macrocycles, respectively. The centroid of each class is indicated with a large circle in the color of the respective class. The contributions of individual descriptors to the PCA are indicated by the length of the arrows. The structure of the oral outlier odalasvir (9) is provided. The structure of milvexian (8), which is close to the centroid of the oral class, is given for comparison. Avasopasem manganese and motexafin gadolinium were removed due to calculation errors with metals. B) Radar plot comparing the median values for the descriptors employed in Lipinski's Ro5 and Veber's rule for the oral FDA-approved (light blue, n = 24) and clinical trial macrocyclic subsets (dark blue, n = 11). Note that HBD, HBA, and TPSA were calculated differently than in the original rules (cf. Methods). Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

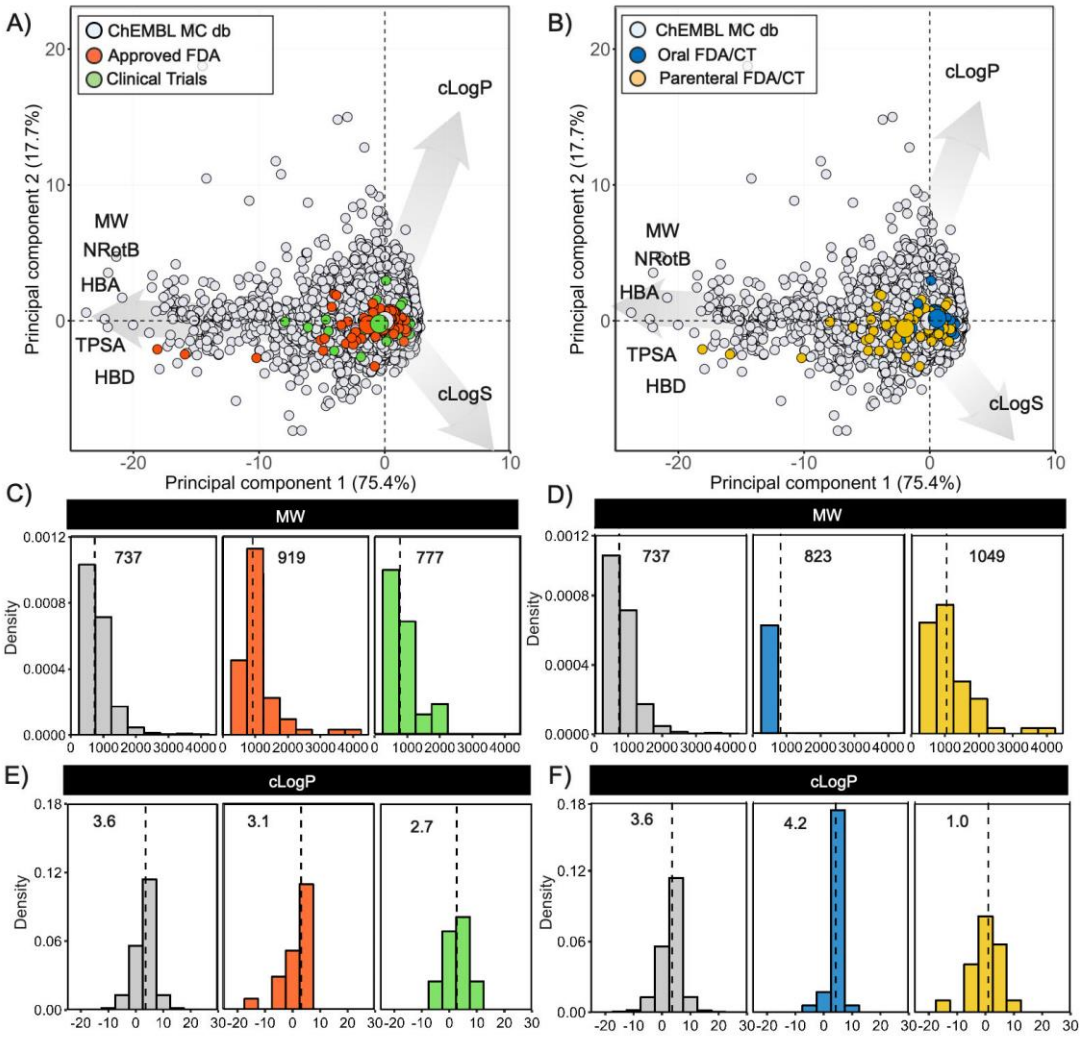


**Figure S4.** PCA comparing the chemical space of the macrocycles retrieved from ChEMBL (n = 28052, in gray) to A) the macrocyclic drugs approved by the FDA (n = 62, in red) and the macrocycles undergoing clinical trials (n = 32, in green) and to B) the combined oral (n = 35, in blue) and parenteral (n = 59, in yellow) parts of the drugs and clinical candidates data sets. The centroid of each class is indicated with a large circle in the color of the respective class. The PCA was based on the descriptors of Lipinski's and Veber's rules as well as cLogS, calculated at pH 7.0. The contributions of individual descriptors to the PCAs are indicated by the length of the arrows. Macrocycles with MW < 1500 Da are removed from the PCA for clarity. (C and E) Distribution of molecular weight (MW) and calculated lipophilicity (cLogP) for the macrocycles in the ChEMBL (n = 28052, in gray), drug (n = 62, in red), and clinical candidates (n = 32, in green) data sets. (D and F) Distribution of molecular weight (MW) and calculated lipophilicity (cLogP) for the macrocycles in the ChEMBL data set (n = 28052, in gray) and in the combined oral (n = 35, in blue) and parenteral (n = 59, in yellow) parts of the drugs and clinical candidate data sets. The median value for the descriptor is given in each panel and indicated by a dashed line. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

**Table S1.** Quality of single property models for the differentiation of oral and parenteral macrocyclic drugs in the training set and the external test set. Abbreviations: True Positive (TP), True negative (TN), False Positive (FP), False negative (FN), Sens. (Sensitivity), Spec. (Specificity), Acc. (Accuracy), Kappa (Cohen's kappa) and Geometric Mean (GMean). Positive (P) stands for "Oral" class and negative (N) stands for parenteral class. The test set was obtained from two publications in 2014. Table adapted from Garcia Jimenez *et al.,* 2023.[22]

| Data sets | Molecular property | Cut-off | Confusion matrix | | | | Sens. | Spec. | Acc. | Kappa | GMean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | TP | TN | FP | FN | | | | | |
| Training set (n=62) | MW | < 982 | 20 | 24 | 14 | 4 | 0.83 | 0.63 | 0.71 | 0.43 | 0.73 |
| | cLogP | > 2.22 | 21 | 23 | 15 | 3 | 0.88 | 0.61 | 0.71 | 0.44 | 0.73 |
| | TPSA | < 292 | 22 | 25 | 13 | 2 | 0.92 | 0.66 | 0.76 | 0.53 | 0.78 |
| | HBA | ≤ 23 | 23 | 16 | 22 | 1 | 0.96 | 0.42 | 0.63 | 0.33 | 0.64 |
| | HBD | ≤7 | 21 | 27 | 11 | 3 | 0.88 | 0.71 | 0.77 | 0.55 | 0.79 |
| | NRotB | ≤ 11 | 19 | 23 | 15 | 5 | 0.79 | 0.61 | 0.68 | 0.37 | 0.69 |
| | nC | ≤ 51 | 21 | 20 | 18 | 3 | 0.88 | 0.53 | 0.66 | 0.36 | 0.68 |
| | PHI | < 19.15 | 18 | 26 | 12 | 6 | 0.75 | 0.68 | 0.71 | 0.42 | 0.72 |
| | cLogS | > -8.95 | 22 | 17 | 21 | 2 | 0.92 | 0.45 | 0.63 | 0.32 | 0.64 |
| Test set (n=60) | MW | < 982 | 15 | 22 | 20 | 3 | 0.83 | 0.52 | 0.62 | 0.28 | 0.66 |
| | cLogP | > 2.22 | 17 | 24 | 18 | 1 | 0.94 | 0.57 | 0.68 | 0.41 | 0.73 |
| | TPSA | < 292 | 17 | 25 | 17 | 1 | 0.94 | 0.6 | 0.7 | 0.43 | 0.75 |
| | HBA | ≤ 23 | 17 | 16 | 26 | 1 | 0.94 | 0.38 | 0.55 | 0.23 | 0.6 |
| | HBD | ≤7 | 18 | 27 | 15 | 0 | 1 | 0.64 | 0.75 | 0.52 | 0.8 |
| | NRotB | ≤ 11 | 12 | 30 | 12 | 6 | 0.67 | 0.71 | 0.7 | 0.35 | 0.69 |
| | nC | ≤ 51 | 15 | 21 | 21 | 3 | 0.83 | 0.5 | 0.6 | 0.26 | 0.65 |
| | PHI | < 19.15 | 13 | 29 | 13 | 5 | 0.72 | 0.69 | 0.7 | 0.37 | 0.71 |
| | cLogS | > -8.95 | 14 | 18 | 24 | 4 | 0.78 | 0.43 | 0.53 | 0.16 | 0.58 |

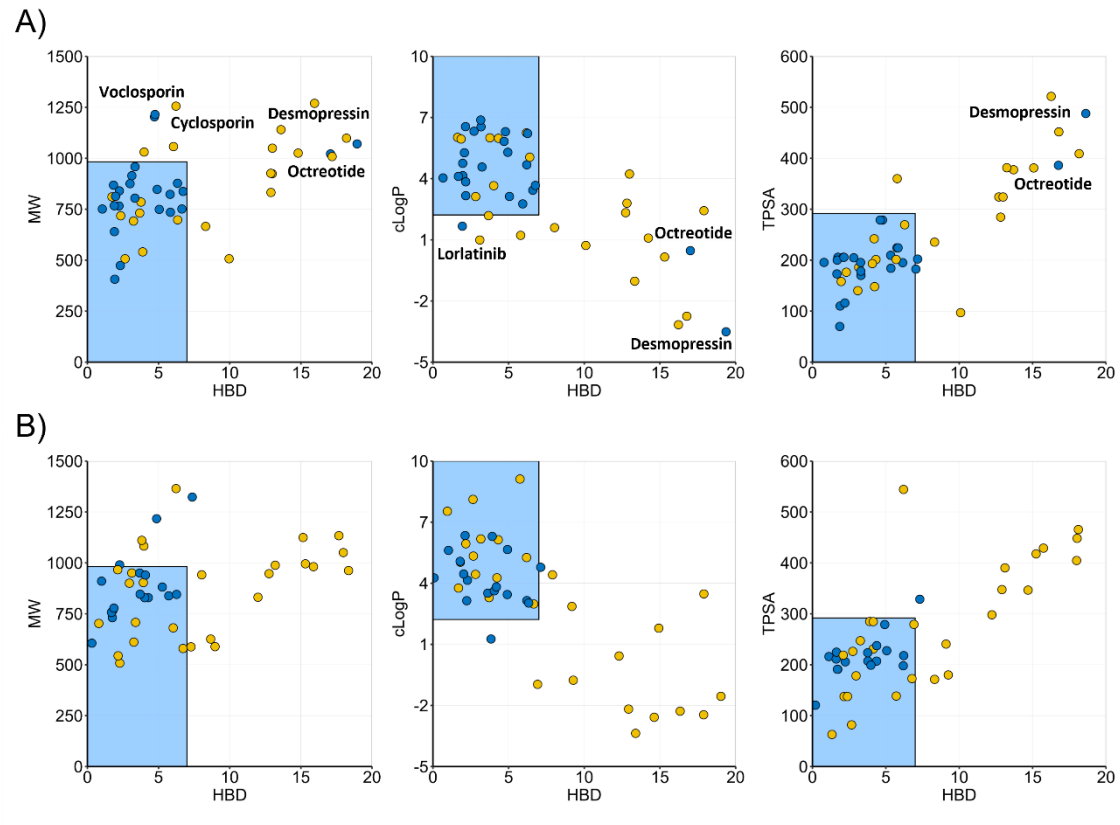



**Figure S5.** Discrimination of orally bioavailable and parenterally administered A) macrocyclic drugs and B) an external test set of macrocycles not yet approved as drugs in bi-descriptor chemical space. Oral drugs are indicated by blue circles, while parenterals are in yellow. The filled circles have been jittered slightly to avoid overlap. The blue shading marks chemical space defined by HBD ≤ 7 and one of MW < 982 Da, cLogP > 2.22, or TPSA < 292 Å$^2$. Some parenteral macrocycles are not included in the Figures which have been truncated at HBD < 20, MW < 1500 Da, −5 < cLogP < 10, and TPSA < 600 Å$^2$. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

**Table S2.** Quality of bi-property models for prediction of oral and parenteral administration of the macrocyclic drugs in the training set (n=62) and the test set (n=60). Cut-off values were selected based on the addition of two descriptors conditions (1st and 2nd cut-off). Table adapted from Garcia Jimenez *et al.,* 2023.[22]

| Data sets | 1st cut-off | 2nd cut-off | Confusion matrix | | | | Sens. | Spec. | Acc. | Kappa | GMean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | TP | TN | FP | FN | | | | | |
| Training set (n=62) | MW (< 982) | TPSA | 20 | 28 | 10 | 4 | 0.83 | 0.74 | 0.77 | 0.55 | 0.78 |
| | | cLogP | 19 | 30 | 8 | 5 | 0.79 | 0.79 | 0.79 | 0.57 | 0.79 |
| | | HBD | 20 | 30 | 8 | 4 | 0.83 | 0.79 | 0.81 | 0.6 | 0.81 |
| | cLogP (> 2.22) | MW | 19 | 30 | 8 | 5 | 0.79 | 0.79 | 0.79 | 0.57 | 0.79 |
| | | TPSA | 21 | 29 | 9 | 3 | 0.88 | 0.76 | 0.81 | 0.61 | 0.82 |
| | | HBD | 21 | 30 | 8 | 3 | 0.88 | 0.79 | 0.82 | 0.64 | 0.83 |
| | HBD (≤ 7) | MW | 20 | 30 | 8 | 4 | 0.83 | 0.79 | 0.81 | 0.6 | 0.81 |
| | | cLogP | 21 | 30 | 8 | 3 | 0.88 | 0.79 | 0.82 | 0.64 | 0.83 |
| | | TPSA | 22 | 28 | 10 | 2 | 0.92 | 0.74 | 0.81 | 0.62 | 0.82 |
| | TPSA (< 292) | MW | 20 | 28 | 10 | 4 | 0.83 | 0.74 | 0.77 | 0.55 | 0.78 |
| | | cLogP | 21 | 29 | 9 | 3 | 0.88 | 0.76 | 0.81 | 0.61 | 0.82 |
| | | HBD | 22 | 28 | 10 | 2 | 0.92 | 0.74 | 0.81 | 0.62 | 0.82 |
| Test set (n=60) | MW (< 982) | TPSA | 15 | 27 | 15 | 3 | 0.83 | 0.64 | 0.7 | 0.4 | 0.73 |
| | | cLogP | 14 | 29 | 13 | 4 | 0.78 | 0.69 | 0.72 | 0.41 | 0.73 |
| | | HBD | 15 | 30 | 12 | 3 | 0.83 | 0.71 | 0.75 | 0.48 | 0.77 |
| | cLogP (> 2.22) | MW | 14 | 29 | 13 | 4 | 0.78 | 0.69 | 0.72 | 0.41 | 0.73 |
| | | TPSA | 16 | 27 | 15 | 2 | 0.89 | 0.64 | 0.72 | 0.44 | 0.76 |
| | | HBD | 17 | 28 | 14 | 1 | 0.94 | 0.67 | 0.75 | 0.51 | 0.79 |
| | HBD (≤ 7) | MW | 15 | 30 | 12 | 3 | 0.83 | 0.71 | 0.75 | 0.48 | 0.77 |
| | | cLogP | 17 | 28 | 14 | 1 | 0.94 | 0.67 | 0.75 | 0.51 | 0.79 |
| | | TPSA | 17 | 28 | 14 | 1 | 0.94 | 0.67 | 0.75 | 0.51 | 0.79 |
| | TPSA (< 292) | MW | 15 | 27 | 15 | 3 | 0.83 | 0.64 | 0.7 | 0.4 | 0.73 |
| | | cLogP | 16 | 27 | 15 | 2 | 0.89 | 0.64 | 0.72 | 0.44 | 0.76 |
| | | HBD | 17 | 28 | 14 | 1 | 0.94 | 0.67 | 0.75 | 0.51 | 0.79 |

Abbreviations: True Positive (TP), True negative (TN), False Positive (FP), False negative (FN), Sens. (Sensitivity), Spec. (Specificity), Acc. (Accuracy), Kappa (Cohen's kappa) and Geometric Mean (GMean). Positive (P) stands for “Oral” class and negative (N) stands for parenteral class. The test set was obtained from two publications in 2014.

**Table S3.** Bi-descriptor models for prediction of oral bioavailability applied to the macrocycles in the clinical trial set (n=32). Predictions were made using HBD $\leq$ 7 as the first cut-off for all three models, in combination with one of TPSA < 292 Å$^2$, cLogP > 2.22 or MW < 982 Da. The best model is highlighted in bold. Figure adapted from Garcia Jimenez *et al.,* 2023.[22]

| Model descriptor | HBD and TPSA | **HBD and cLogP[b]** | HBD and MW |
|---|---|---|---|
| Accuracy | 0.6 | **0.75** | 0.63 |
| Sensitivity | 1 | **0.91** | 0.82 |
| Specificity | 0.43 | **0.67** | 0.52 |
| Kappa | 0.34 | **0.51** | 0.29 |

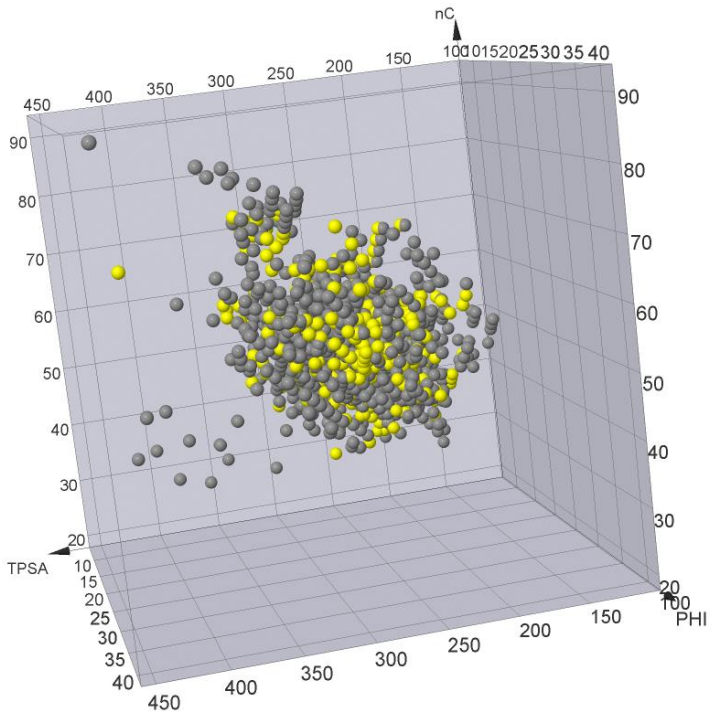



**Figure S6.** Databases overlap (PROTAC-DB in grey, PROTACpedia in yellow). Figure adapted from Ermondi *et al.,* 2021.[180]

**Table S4.** 2D descriptors for the PROTACs and their building blocks. Table adapted from Ermondi *et al.,* 2021.[180]

| Name | E3 Ligand | Warhead | Linker | PROTACs |
|---|---|---|---|---|
| MW | 462 | 479 | 215 | 984 |
| nC | 24 | 24 | 11 | 50 |
| NAR | 2 | 3 | 0 | 5 |
| PHI | 8 | 7 | 9 | 18 |
| HBD | 3 | 3 | 1 | 4 |
| HBA | 8 | 8 | 4 | 18 |
| TPSA | 113 | 109 | 55 | 242 |

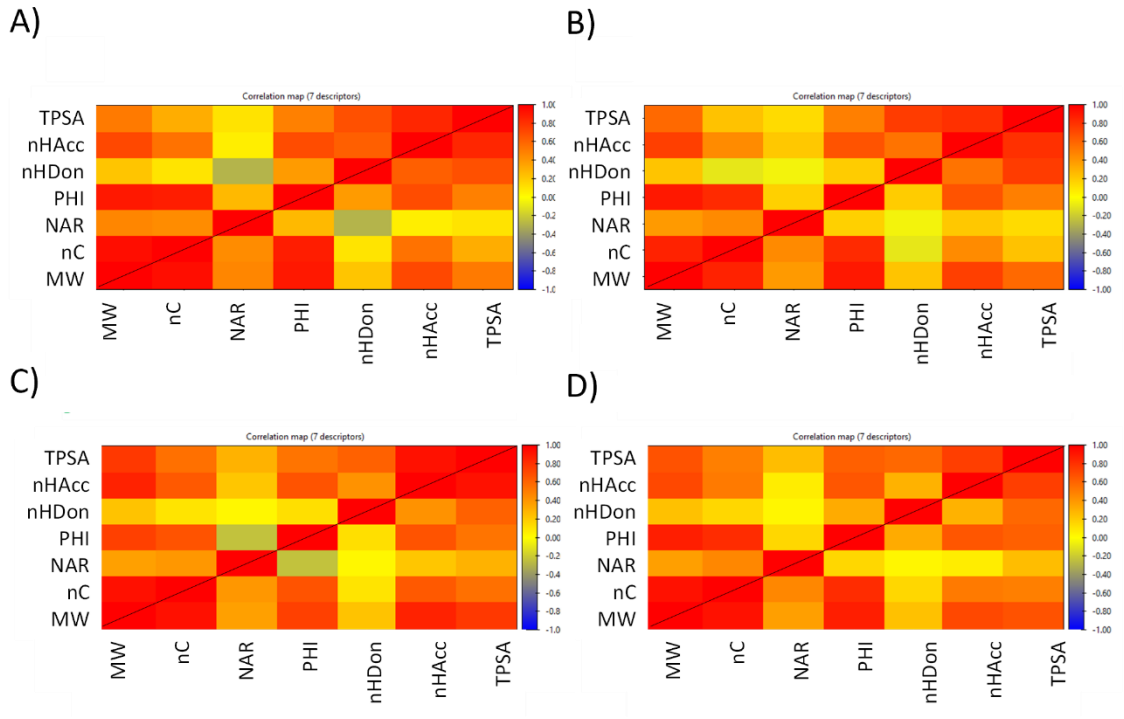



**Figure S7.** 2D Correlation matrices. A) E3 ligands, B) warhead, C) linkers and D) PROTACs. Figure adapted from Ermondi *et al.,* 2021.[180] HBA and HBD were named as nHAcc and nHDon in the original publication, respectively. Figure adapted from Ermondi *et al.,* 2021.[180]

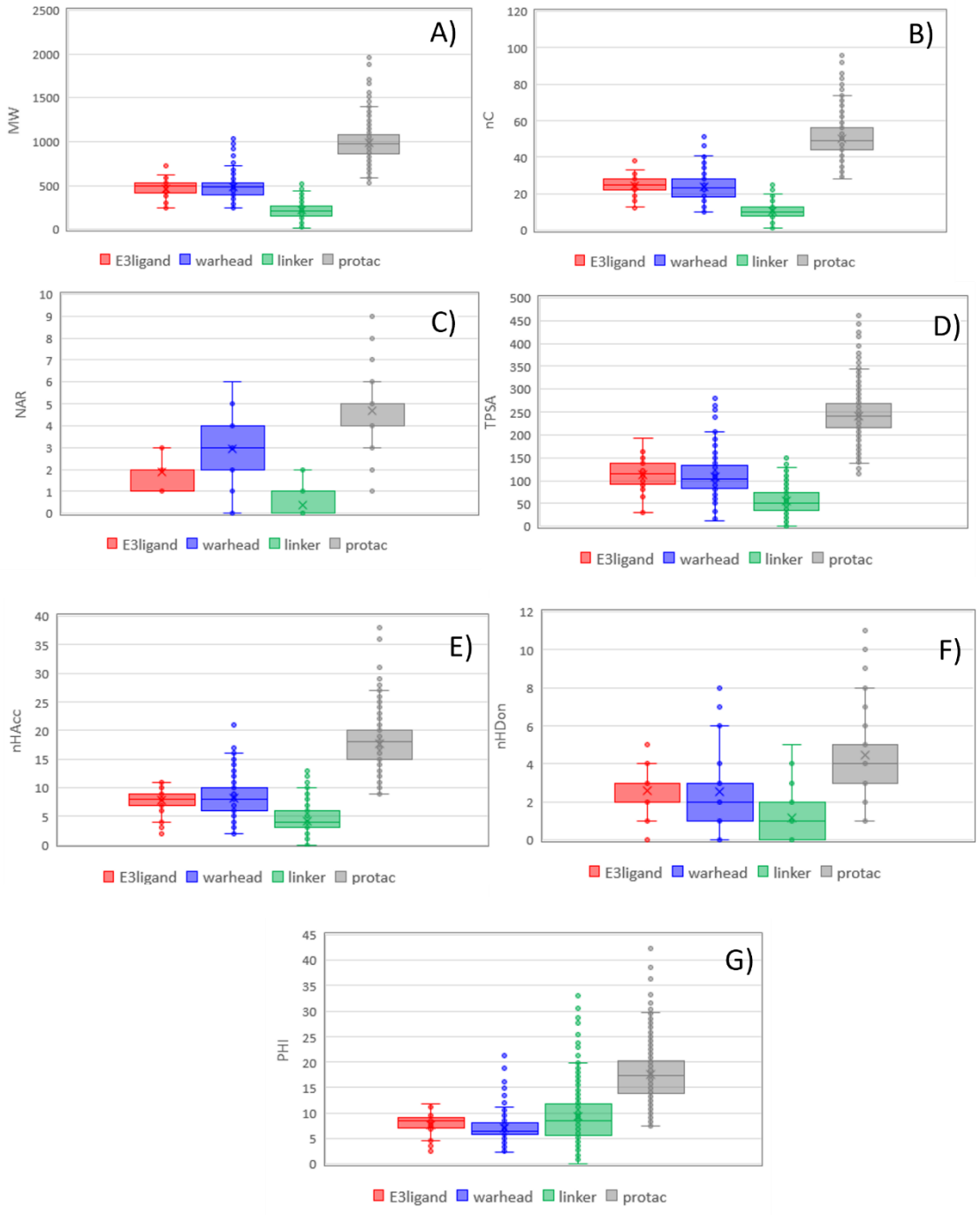



**Figure S8.** Box plot of the 7 descriptors for PROTACs and their building blocks. A) MW, B) nC, C) NAR D) TPSA E) HBA F) HBD and G) PHI. HBA and HBD were named as nHAcc and nHDon in the original publication, respectively. Figure adapted from Ermondi *et al.,* 2021.[180]

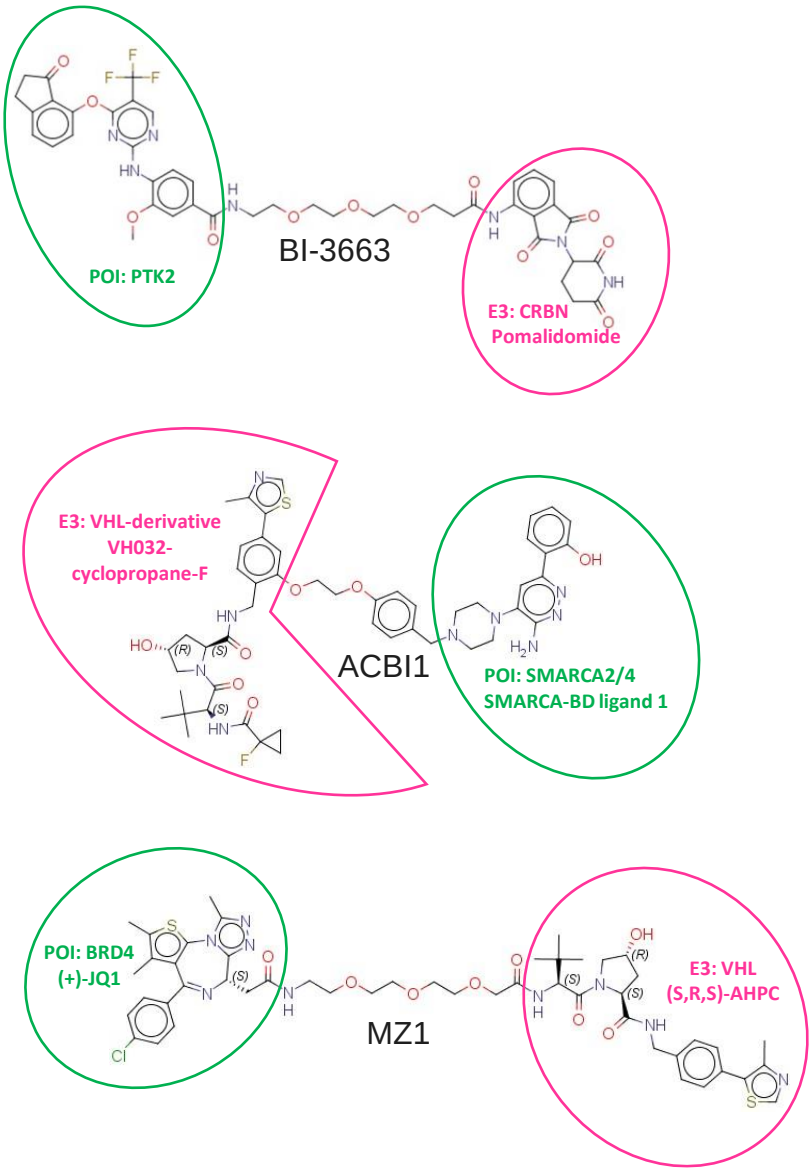


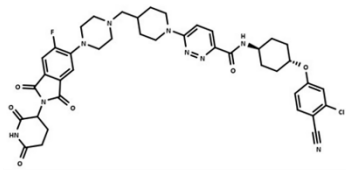

**Figure S9.** Investigated PROTAC chemical structures. Figure adapted from Ermondi *et al.,* 2020 and 2021.[112,180]

**Table S5.** PROTACs reported in the permeability subregions of the PROTAC chemical space. Permeability values and their assessment methods. Caco-2: (AB: apical-basal, BA: basal-apical). Table adapted from Garcia Jimenez *et al.,* 2021.[20]

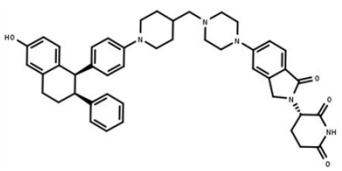
| Name | Permeability ($10^{-6}$ cm/s) | Permeability experiment | Permeability classification | Source |
|---|---|---|---|---|
| ACBI1 | 2 | Caco-2 | High | opnMe[112] |
| BI-0319 | 0.2 | Caco-2 | Low | opnMe[112] |
| BI-3663 | 15.85 | Caco-2 | High | opnMe[112] |
| BI-4206 | 0.4 | Caco-2 | Low | opnMe[112] |
| PROTAC-1 | 1.41 | PAMPA | High | Kihlberg[70] |
| PROTAC-14 | 1.7 (AB), 14.1 (BA) | Caco-2 | High | Skidmore[192] |
| MZ4 | 0.60 | PAMPA | Low | Lokey[225] |
| MZ1 | 0.03 | PAMPA | Low | Lokey[225] |
| MZ3 | 0.01 | PAMPA | Low | Lokey[225] |
| MZP-61 | 0.30 | PAMPA | Low | Lokey[225] |
| MZP-55 | 0.16 | PAMPA | Low | Lokey[225] |
| CM09 | 0.01 | PAMPA | Low | Lokey[225] |
| CM 11 | 0.01 | PAMPA | Low | Lokey[225] |
| CMP 98 | 0.00 | PAMPA | Low | Lokey[225] |
| AT3 | 0.01 | PAMPA | Low | Lokey[225] |
| AT6 | 0.00 | PAMPA | Low | Lokey[225] |
| AT2 | 0.00 | PAMPA | Low | Lokey[225] |

A)

**ARV-110**

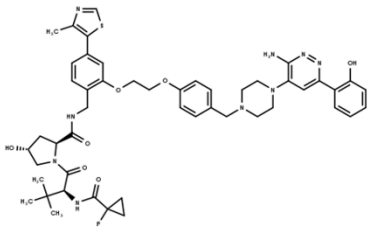


**ARV-471**

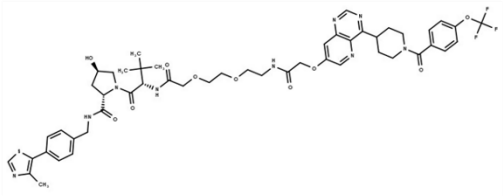


B)

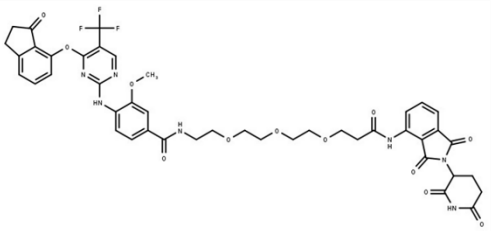
**ACBI1**



**PROTAC-1**



**BI-3663**



**PROTAC-14**

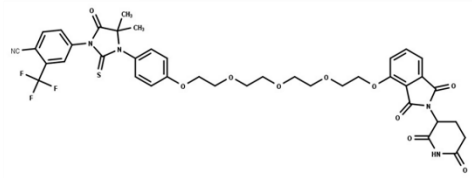


**Figure S10.** Structures of A) bioavailable and B) permeable PROTACs. Figure adapted from Garcia Jimenez *et al.,* 2021.[20]

**Table S6.** List of the studied PROTACs and their predicted constitutive building blocks (information obtained directly from the PROTAC-DB: http://cadd.zju.edu.cn/protacdb/). Building blocks are presented as the fundamental ideal constituents of the PROTAC, regardless of the real building blocks structures used for the synthetic approach. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| PROTACs | DB-ID | Warhead | Linker | E3 ligand |
|---|---|---|---|---|
| ACBI1 | 798 | Undefined | 1-ethoxy-4-methylbenzene | VH101: (S,R,S) VH032-cyclopropane-F |
| ARV-825 | 329 | Birabresib (OTX-015) | 1-ethoxy-2-(2-ethoxyethoxy)ethane | pomalidomide |
| BI-0319 | 1656 | BI-4464 | 1-(2-ethoxyethoxy)-2-methoxyethane | VH032: (S,R,S)-AHPC |
| BI-3663 | 1655 | BI-4464 | 3-[2-(2-ethoxyethoxy)ethoxy]propanal | Pomalidomide |
| BI-4206 | | BI-4464 | 1-(2-ethoxyethoxy)-2-methoxyethane | VH032, negative control: (S,S,S)-AHPC |
| BRD9 degrader-1 | 1187 | GSK-39 | N-(4-acetamidobutyl)-2-hydroxyacetamide | Thalidomide |
| BSJ-03-123 | 240 | Palbociclib | N-[2-[2-(2-ethoxyethoxy)ethoxy]ethyl]-2-hydroxyacetamide | Thalidomide |
| *Cis*ACBI1 | | Undefined | 1-ethoxy-4-methylbenzene | VH101, negative control :(S,S,S) VH032-cyclopropane-F |
| *Cis*MZ1 | | JQ1 | 2-[2-(2-methoxyethoxy)ethoxy]ethanamine | VH032, negative control: (S,S,S)-AHPC |
| CM 11 | 998 | VH032: (S,R,S)-AHPC | 1-methoxy-2-[2-[2-[2-(2-methoxyethoxy)ethoxy]ethoxy]ethoxy]ethane | VH032: (S,R,S)-AHPC |
| CMP 98 | | VH032, negative control: (S,S,S)-AHPC | 1-methoxy-2-[2-[2-[2-(2-methoxyethoxy)ethoxy]ethoxy]ethoxy]ethane | VH032, negative control: (S,S,S)-AHPC |
| CRBN-6-5-5-VHL | 176 | Pomalidomide | 1-(5-butoxypentoxy)hexane | VH032: (S,R,S)-AHPC |
| dBET57 | 342 | JQ1 | Ethanamine | Pomalidomide |
| Gefitinib-based PROTAC 3 | 276 | Gefitinib | 1-(2-ethoxyethoxy)pentane | VH032: (S,R,S)-AHPC |
| Mcl1 degrader-1 | 500 | Undefined | hexan-1-amine | Pomalidomide |
| MD-224 | 50 | MI-1242 | pent-1-yne | Lenalidomide |
| MZ1 | 335 | JQ1 | 2-[2-(2-methoxyethoxy)ethoxy]ethanamine | VH032: (S,R,S)-AHPC |
| MZP-54 | 1652 | I-BET726 | 2-[2-(2-methoxyethoxy)ethoxy]ethanamine | VH032: (S,R,S)-AHPC |
| THAL-SNS-032 | 799 | SNS-032 | N-{2-[2-(2-ethoxyethoxy)ethoxy]ethyl}acetamide | Pomalidomide |
| VZ185 | 22 | BI-7273 | Pentane | VH101 |
| ZXH-3-26 | 350 | JQ1 | Pentan-1-amine | Pomalidomide |

**Table S7.** Ionization states for the PROTAC data set calculated at pH 7. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| PROTACs | Predicted Marvin $pK_a$ (1-14) | Predicted ionization state at pH 7 |
|---|---|---|
| **ACBI1** | 2.61 (b), 5.39 (b), 7.27 (b), 8.48 (a), 11.05 (a) | 32% (N), 62% (+) |
| **ARV-825** | 4.16 (b), 11.59 (a) | N |
| **BI-0319** | 2.68 (b), 11.46(a) | N |
| **BI-3663** | 11.23 (a), 11.84 (a) | N |
| **BI-4206** | 2.68 (b), 11.46(a) | N |
| **BRD9 degrader-1** | 7.86 (b), 11.81 (a) | 12% (N), 88% (+) |
| **BSJ-03-123** | 3.19 (b), 7.38 (b), 11.15 (a), 11.79 (a) | 29% (N), 70% (+) |
| ***Cis*ACBI1** | 2.61 (b), 5.39 (b), 7.27 (b), 8.48 (a), 11.05 (a) | 32% (N), 62% (+) |
| ***Cis*MZ1** | 2.6 (b), 4.32 (b) | N |
| **CM 11** | 2.35 (b),2.95 (b) | N |
| **CMP 98** | 2.35 (b),2.95 (b) | N |
| **CRBN-6-5-5-VHL** | 2.06 (b), 2.79 (b), 11.61 (a) | N |
| **dBET57** | 4.21(b), 11.61 (a) | N |
| **Gefitinib-based PROTAC 3** | 2.65 (b), 4.66 (b) | N |
| **Mcl1 degrader-1** | 2.17(b), 11.61(a) | N |
| **MD-224** | 9.02 (b), 11.56 (a) | 99% (+) |
| **MZ1** | 2.6 (b), 4.32 (b) | N |
| **MZP-54** | 3.44 (b), 3.57 (b) | N |
| **THAL-SNS-032** | 7.11 (b), 8.02 (a), 11.61 (a) | 36% (N), 54% (+) |
| **VZ185** | 2.60 (b), 3.71 (b), 7.99 (b), 11.05 (a) | 9% (N), 88% (+) |
| **ZXH-3-26** | 2.27(b), 3.96(b), 11.61 (a) | N |

**Table S8.** Experimental solubility values for the PROTAC data set. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| Molecule | Solubility (mg/mL) | Solubility log S (mol/l) |
|---|---|---|
| ACBI1 | < 6.80 $x10^{-4}$ | <- 6.14 |
| ARV-825 | < 2.90 $x10^{-4}$ | < - 6.5 |
| BI-0319 | 2.78 $x10^{-3}$ | -5.58 |
| BI-3663 | 0.01 | -5.16 |
| BI-4206 | 6.06 $x10^{-4}$ | -6.24 |
| BRD9 degrader-1 | 0.60 | -3.18 |
| BSJ-03-123 | 0.02 | -4.75 |
| *Cis*ACBI1 | < 8.17 $x10^{-4}$ | < -6.06 |
| *Cis*MZ1 | 0.05 | -4.30 |
| CM 11 | 0.73 | -3.20 |
| CMP 98 | 0.39 | -3.48 |
| CRBN-6-5-5-VHL | 1.19 $x10^{-3}$ | -5.91 |
| dBET57 | 0.02 | -4.52 |
| Gefitinib-based PROTAC 3 | 1.14 $x10^{-3}$ | -5.91 |
| Mcl1 degrader-1 | < 7.93 $x10^{-4}$ | < -6.06 |
| MD-224 | < 2.89 $x10^{-4}$ | < -6.64 |
| MZ1 | 0.04 | -4.42 |
| MZP-54 | 5.28 $x10^{-4}$ | -6.29 |
| THAL-SNS-032 | 0.05 | -4.28 |
| VZ185 | 0.05 | -4.30 |
| ZXH-3-26 | 0.00 | -5.53 |

**Table S9.** 2D *in silico* log P descriptors. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| Tool name | log P method | Tool name | log P method | Structure imput |
|---|---|---|---|---|
| **Marvin** | Chemaxon method: Atom-based (AlogP) | **SwissADME** | iLOGP: Physics-based method.<br>XLOGP: Atom-based.<br>WLOGP: Atom-based.<br>MLOGP: Chemical descriptor based.<br>SILICOS-IT: Hybrid method (Atom based- chemical descriptor based). | 2D |
| **pkCSM** | Distance-based graph structural signatures | **MoKa** | Unknown | |
| **Scbdd** | Chemical descriptor-based | **ACD Labs** | Fragment group-based:<br>- Classic: Principal of isolating carbons.<br>- GALAS: Similarity-based.<br>- Consensus: Mixed approach. | |
| **AdmetSAR2** | AlogP | **Molinspiration** | Fragment group-based | |
| **Volsurf pH 7.5** | 3D molecular field-based | | | 3D |

**Table S10.** Linear correlation of computed log P with experimental solubility. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

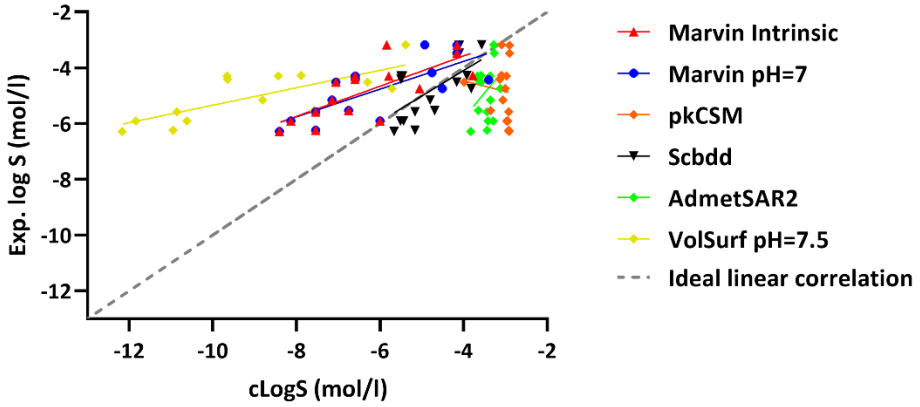
| *In silico predictors* | *log P (scbdd)* | *logD pH 7.4 (scbdd)* | *LogP (ACD)* | *Log D (pH=7.4) (ACD)* | *ALOGP (ADMETSAR2)* | *LOG P n-Oct VolSurf+* |
|---|---|---|---|---|---|---|
| ***R square*** | 0.01 | 0.01 | 0.69 | 0.63 | 0.51 | 0.59 |
| ***Equation*** | Y = -0.09X - 4.56 | Y = -0.09X - 4.56 | Y = -0.49X - 3.57 | Y = -0.43X - 3.47 | Y = -0.43X - 2.74 | Y = -0.38X - 3.33 |

| *In silico predictors* | *log P (Marvin)* | *miLogP (molinspiration)* | *LogD7.4 MoKa* | *logP MoKa* | *LogD7.5 VolSurf+* | *LOG P n-Oct VolSurf+* |
|---|---|---|---|---|---|---|
| ***R square*** | 0.69 | 0.63 | 0.61 | 0.61 | 0.53 | 0.59 |
| ***Equation*** | Y = -0.55X - 3.83 | Y = -0.48X - 3.25 | Y = -0.33X - 3.69 | Y = -0.40X - 3.31 | Y = -0.29X - 3.82 | Y = -0.38X - 3.33 |

| *SWISSADME in silico predictors* | *Log Po/w (iLOGP) (Swissadme)* | *Log Po/w (XLOGP3) (Swissadme)* | *Log Po/w (WLOGP) (Swissadme)* | *Log Po/w (MLOGP) (Swissadme)* | *Log Po/w (SILICOS-IT) (Swissadme)* |
|---|---|---|---|---|---|
| ***R square*** | 0.01 | 0.43 | 0.45 | 0.28 | 0.04 |
| ***Equation*** | Y = -0.07X - 4.39 | Y = -0.40X - 3.11 | Y = -0.34X - 3.47 | Y = -0.34X - 4.89 | Y = -0.10X - 4.04 |

**Table S11.** 2D *in silico* log S descriptors. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| Method | Principle | Source | Structure imput |
|---|---|---|---|
| ● Marvin pH dependent | Atom-based contribution with 2 correction factors: hydrophobic carbon count and square of molecular weight[238]. | https://chemaxon.com/products/marvin | 2D |
| ● Marvin intrinsic | | | |
| ● pkCSM | Distance-graph based structural signatures[239] | http://biosig.unimelb.edu.au/pkcsm/prediction | |
| ● Scbdd | Model based on a similarity engine[240] | http://www.scbdd.com/ | |
| ● AdmetSAR2 | Solvent accessible surface-based[241] | http://lmmd.ecust.edu.cn/admetsar2/ | |
| ● VolSurf pH 7.5 | 3D molecular field-based[242] | https://www.moldiscovery.com/software/vsplus/ | 3D |




| In silico predictor | R square | Equation |
|---|---|---|
| Marvin Intrinsic | 0.56 | Y = 0.54X - 1.40 |
| Marvin pH = 7 | 0.57 | Y = 0.50X - 1.79 |
| pkCSM | 0.01 | Y = -0.37X - 5.97 |
| Scbdd | 0.42 | Y = 0.91X - 0.45 |
| AdmetSAR2 | 0.11 | Y = 1.81X + 1.41 |
| VolSurf pH=7.5 | 0.57 | Y = 0.31X - 2.21 |

**Figure S11.** Calculated versus experimental solubility for 16 PROTACs. (b) Linear regression of solubility predictors with solubility. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

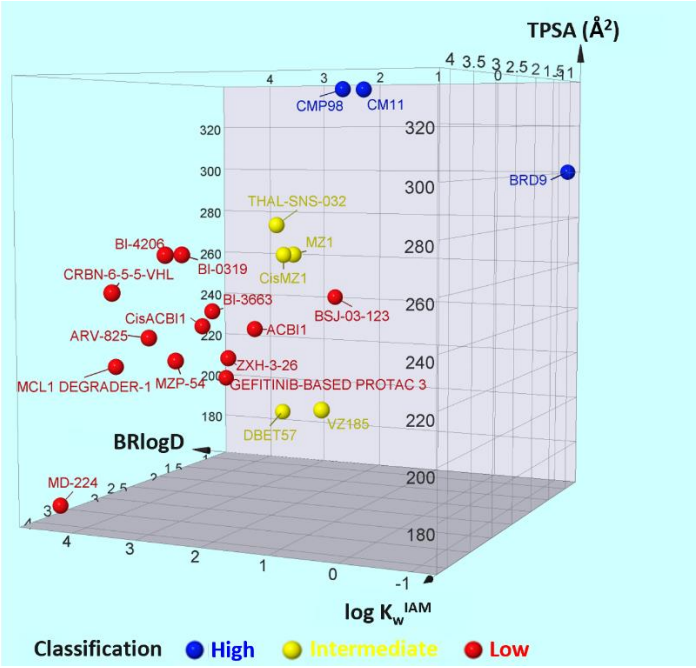



**Figure S12.** PROTAC solubility distribution colored by the GSK classification: low (< 30 µM), intermediate (30-200 µM) or highly soluble molecules (> 200 µM). PROTACs without ND solubility (Red values from Table 3) were included in the low solubility class. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

**A) Random Tree**

| | |
|---|---|
| **Correctly Classified Instances** | **13 (87%)** |
| **Incorrectly Classified Instances** | **2 (13%)** |
| **Total Number of Instances** | **15** |

| TP Rate | FP Rate | Precision | Recall | F-Measure | Matthews Correlation Coefficient (MCC) | Class |
|---|---|---|---|---|---|---|
| 0.667 | 0.083 | **0.667** | 0.667 | 0.667 | 0.583 | high |
| 0.800 | 0.100 | **0.800** | 0.800 | 0.800 | 0.700 | intermediate |
| 1.000 | 0.000 | **1.000** | 1.000 | 1.000 | 1.000 | low |

| Confusion Matrix | | | |
|---|---|---|---|
| a | b | c | Classified as |
| 2 | 1 | 0 | a = high |
| 1 | 4 | 0 | b = intermediate |
| 0 | 0 | 7 | c = low |

**B) Random Forest**

| | |
|---|---|
| **Correctly Classified Instances** | **13 (87%)** |
| **Incorrectly Classified Instances** | **2 (13%)** |
| **Total Number of Instances** | **15** |

| TP Rate | FP Rate | Precision | Recall | F-Measure | Matthews Correlation Coefficient (MCC) | Class |
|---|---|---|---|---|---|---|
| 0.667 | 0.083 | **0.667** | 0.667 | 0.667 | 0.583 | high |
| 0.800 | 0.100 | **0.800** | 0.800 | 0.800 | 0.700 | intermediate |
| 1.000 | 0.000 | **1.000** | 1.000 | 1.000 | 1.000 | low |

| Confusion Matrix | | | |
|---|---|---|---|
| a | b | c | Classified as |
| 2 | 1 | 0 | a = high |
| 1 | 4 | 0 | b = intermediate |
| 0 | 0 | 7 | c = low |

**Figure S13.** ML solubility classification models. A) Random Tree B) Random Forest. Figure adapted from Garcia Jimenez *et al.,* 2022.[63]

**Table S12.** Experimental descriptors and calculated TPSA for the explored building blocks. ND: not determined. Table adapted from Garcia Jimenez *et al.,* 2022.[63]

| Molecule | BRlogD | Log $k_w^{IAM}$ | TPSA | Δ log $k_w^{IAM}$ | Experimental $pK_a$ (2-12) | Ionization state at pH 7 |
|---|---|---|---|---|---|---|
| I-BET726 | 0.30 | 1.88 | 69.64 | 2.64 | ND | ND |
| JQ1 (carboxylic acid) | 0.08 | 0.80 | 108.61 | 1.76 | ND | ND |
| PROTAC BET-binding moiety 2 | -0.03 | 0.53 | 134.91 | 1.59 | ND | ND |
| BI-4464 | 1.79 | 1.84 | 105.68 | 1.23 | ND | ND |
| Pomalidomide | 1.41 | 0.91 | 111.26 | 0.64 | Not ionizable | Neutral |
| *S,R,S*-AHPC HCl | 0.56 | 1.83 | 136.79 | 2.34 | 3.02 (b), 7.59 (b) | + |
| *S,S,S*-AHPC 2HCl | 0.91 | 1.86 | 136.79 | 2.05 | 2.79 (b), 7.48 (b) | + |

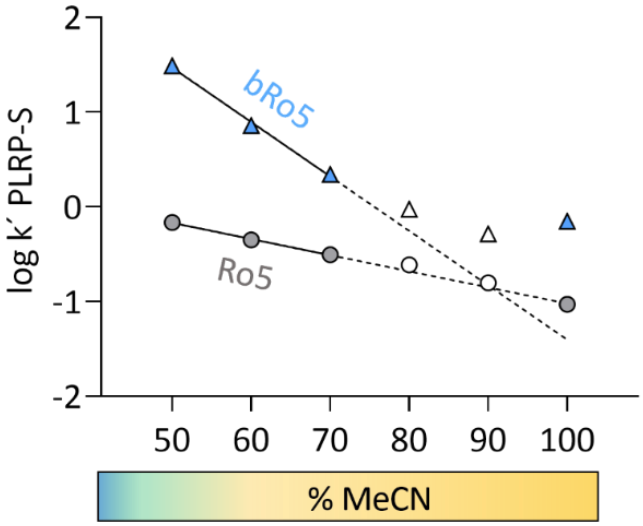



**Figure S14**. Log k' variation of Ro5 (pomalidomide) and bRo5 (CsA) compounds in the PLRP-S system. The solid trend lines represent the equations obtained from experimental values at 50, 60 and 70% of MeCN. The dashed lines represent the extrapolation of the equation for higher MeCN percentages. White-colored symbols reflect the deviation from the equation. Figure adapted from Garcia Jimenez *et al.,* 2023.[121]

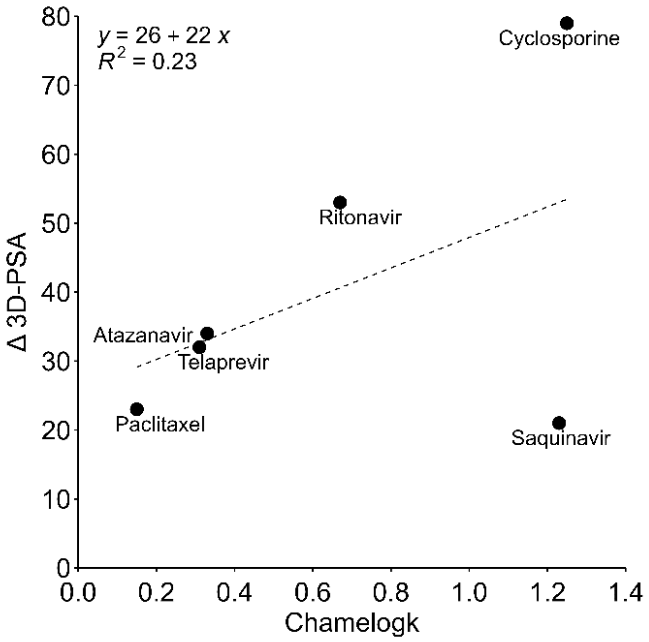



**Figure S15**. Chameleonicity equivalence between Chamelogk and Δ 3D-PSA. The dashed line represents the linear regression. Figure adapted from Garcia Jimenez *et al.,* 2023.[121]

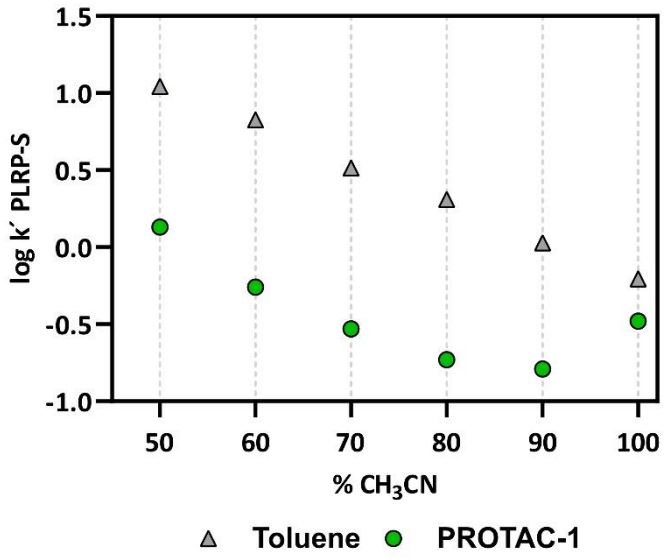



**Figure S16**. Chamelogk plot of PROTAC-1 and toluene (reference Ro5 compound). Figure adapted from Ermondi *et al.,* 2023.[210]

**Table S13.** Detailed description of the protocols used for CS and post CS minimization. In order to maximize comparability of the results with NAMFIS, the same software was used (Schrodinger). In addition, since NAMFIS was based on Monte Carlo simulations, a MC-based algorithm derivative was tested: the mixed torsional/low-mode sampling (MCMM/LMOD) (Sc01-Sc03). This methodology combines Monte Carlo methods (MCMM), which generate conformations by randomly adjusting rotatable bonds, and combinations of low-frequency vibrational modes (LMOD).[145] Moreover, the pure Monte Carlo torsional sampling (MCMM) was also tested (Sc04-Sc06). Furthermore, the force field OPLS3e[243] was selected to simplify the CS search. Finally, the number of steps was tested from $10^3$ to $10^5$. Table adapted from Ermondi *et al.,* 2023.[210]

| CS protocols | | | | |
|---|---|---|---|---|
| **Software** | **Protocol label** | **Algorithm** | **Max. Number steps** | **Force field; Solvation** |
| Schrodinger | Sc01 | Mixed MCMM/LMOD | $10^3$ | OPLS3e; GB/SA |
| | Sc02 | | $10^4$ | |
| | Sc03 | | $10^5$ | |
| | Sc04 | MCMM | $10^3$ | |
| | Sc05 | | $10^4$ | |
| | Sc06 | | $10^5$ | |

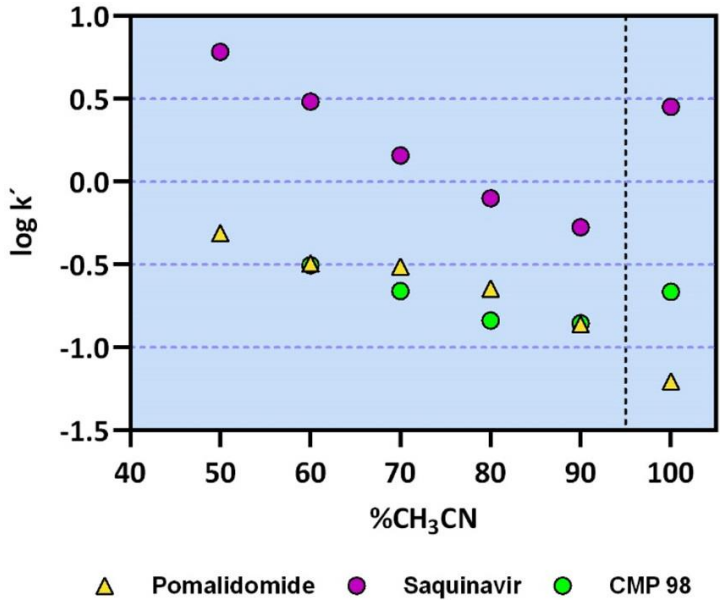



**Figure S17.** Pomalidomide (yellow), CMP 98 (green), and saquinavir (violet) behavior in the PLRP-S system. The gray dashed line at 95% $CH_3CN$ highlights the slope change Figure adapted from Sebastiano et al., 2022.[106]

**Table S14**. Statistical 3D PSA values for the generated conformers with the three methods. p: polar, *np: nonpolar. Table adapted from Sebastiano et al., 2022.[106]

| | | Polar solvent | | | | Non-polar solvent | | | | Difference |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Compound | Max | Min | Average | Median | Max | Min | Average | Median | Δ Max p - Min np* |
| CS | Pomalidomide | 114.2 | 112.8 | 113.7 | 113.8 | 113.5 | 112.7 | 113.2 | 113.4 | 1.5 |
| | Saquinavir | 169.8 | 150.6 | 161.1 | 160.9 | 158.2 | 145.1 | 153.3 | 154.2 | 24.7 |
| | CMP 98 | 327.7 | 303.8 | 316.3 | 317.4 | 302.5 | 287.3 | 294.8 | 294.8 | 40.4 |
| MD | Pomalidomide | 118.0 | 107.3 | 113.1 | 113.1 | 118.5 | 106.7 | 112.4 | 112.4 | 11.3 |
| | Saquinavir | 172.9 | 145.3 | 163.1 | 163.7 | 170.5 | 142.6 | 157.0 | 157.0 | 30.3 |
| | CMP 98 | 337.5 | 299.6 | 321.5 | 321.7 | 325.8 | 284.0 | 305.3 | 305.2 | 53.4 |
| SMD | Pomalidomide | 117.6 | 107.4 | 113.1 | 113.2 | 118.1 | 106.8 | 112.4 | 112.4 | 10.8 |
| | Saquinavir | 173.3 | 142.6 | 160.5 | 161.0 | 165.2 | 137.3 | 151.9 | 151.9 | 36.0 |
| | CMP 98 | 338.3 | 293.5 | 317.2 | 317.2 | 329.8 | 280.7 | 304.7 | 304.9 | 57.7 |

**Table S15**. Statistical $R_{gyr}$ values for the generated conformers with the three methods. p: polar, *np: nonpolar. Table adapted from Sebastiano et al., 2022.[106]

| | | Polar solvent | | | | Non-polar solvent | | | | Difference |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Compound | Max | Min | Average | Median | Max | Min | Average | Median | Δ Max p - Min np* |
| CS | Pomalidomide | 3.3 | 3.2 | 3.2 | 3.3 | 3.2 | 3.2 | 3.2 | 3.2 | 0.2 |
| | Saquinavir | 5.8 | 4.3 | 4.9 | 5.0 | 5.8 | 4.6 | 5.1 | 5.1 | 1.2 |
| | CMP 98 | 6.5 | 5.5 | 5.8 | 5.8 | 7.0 | 5.7 | 5.9 | 5.8 | 0.8 |
| MD | Pomalidomide | 3.4 | 3.1 | 3.3 | 3.3 | 3.4 | 3.2 | 3.3 | 3.3 | 0.3 |
| | Saquinavir | 6.3 | 4.8 | 5.6 | 5.5 | 6.0 | 4.8 | 5.3 | 5.3 | 1.5 |
| | CMP 98 | 13.3 | 6.0 | 8.8 | 8.3 | 13.2 | 5.7 | 8.0 | 7.9 | 7.7 |
| SMD | Pomalidomide | 3.4 | 3.2 | 3.3 | 3.3 | 3.4 | 3.2 | 3.3 | 3.3 | 0.3 |
| | Saquinavir | 6.3 | 4.8 | 5.6 | 5.6 | 6.2 | 4.3 | 5.1 | 5.1 | 1.9 |
| | CMP 98 | 13.3 | 5.7 | 7.3 | 6.7 | 13.3 | 5.8 | 8.0 | 7.6 | 7.6 |

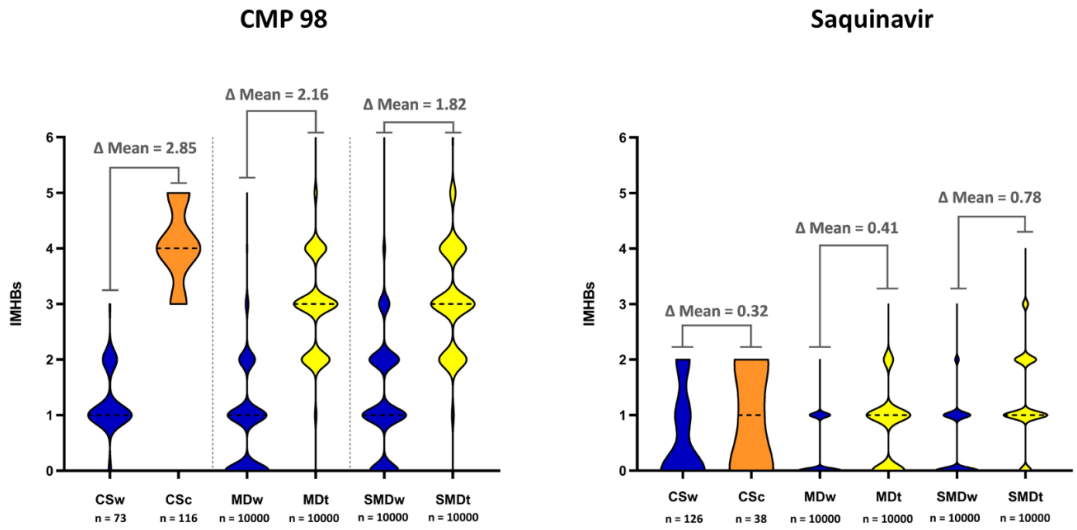



**Figure S18**. IMHB quantification of the generated conformers by CS in water/chloroform (CSw/ CSc), molecular dynamics in water/toluene (MDw/ MDt) and SMD tunneling in water/toluene (SMDw/ SMDt). Δ Mean is the difference of the average of IMHBs in nonpolar solvent and water. Median values are presented as black dashed lines. Figure adapted from Sebastiano et al., 2022.[106]

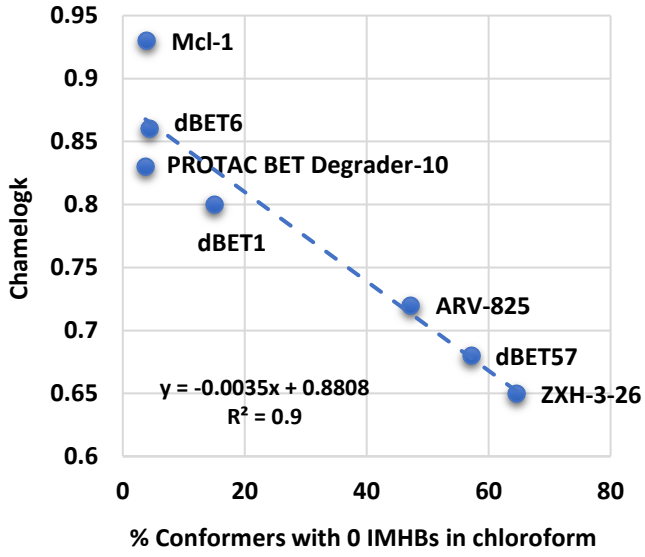



**Figure S19**. Relationship between chameleonicity and the predicted absence of IMHBs in chloroform (nonpolar solvent) using CS, for structurally related compounds. Unpublished content.